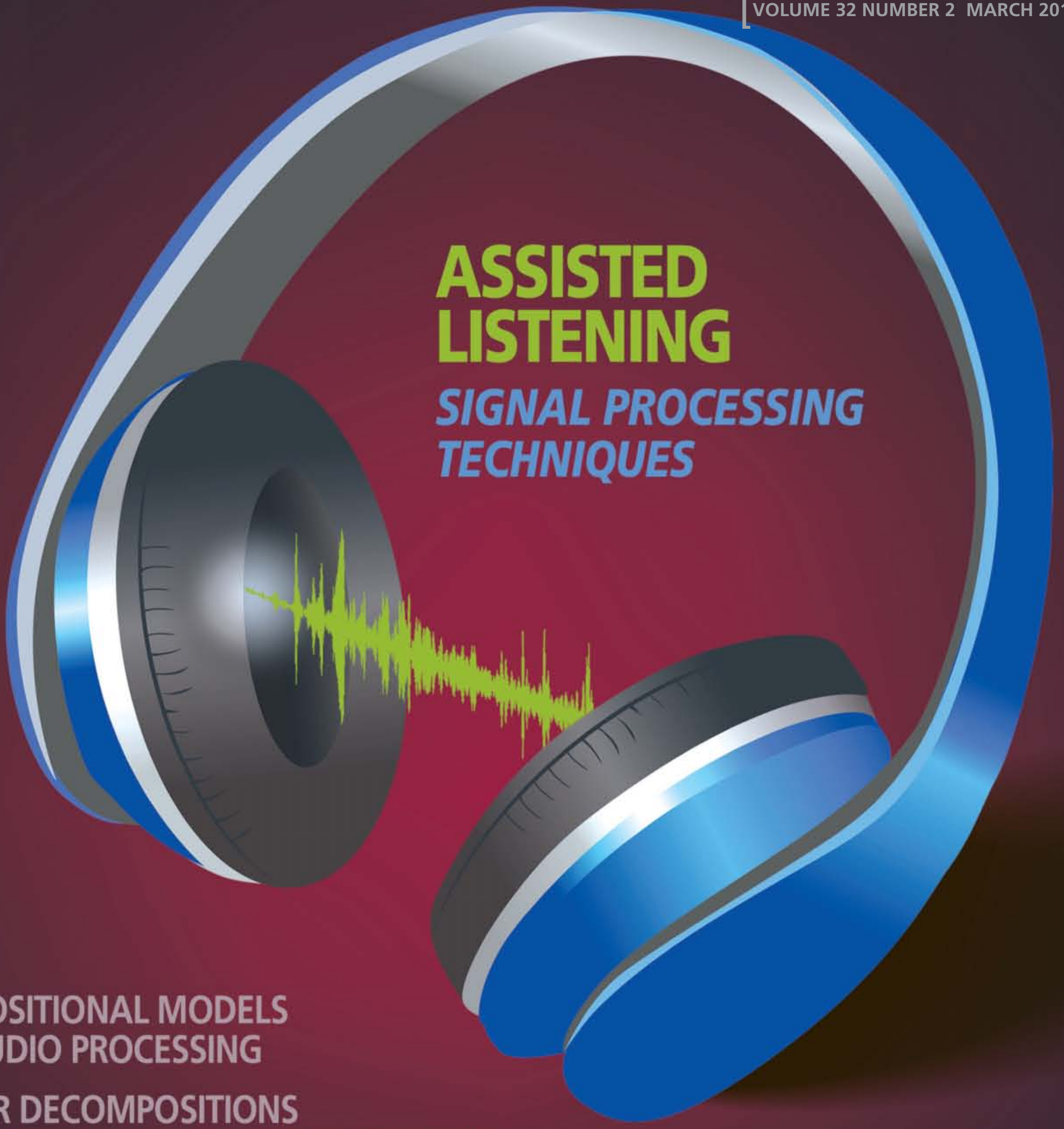
IEEE
SignalProcessing
MAGAZINE

[VOLUME 32 NUMBER 2 MARCH 2015]



ASSISTED LISTENING

SIGNAL PROCESSING TECHNIQUES

COMPOSITIONAL MODELS FOR AUDIO PROCESSING

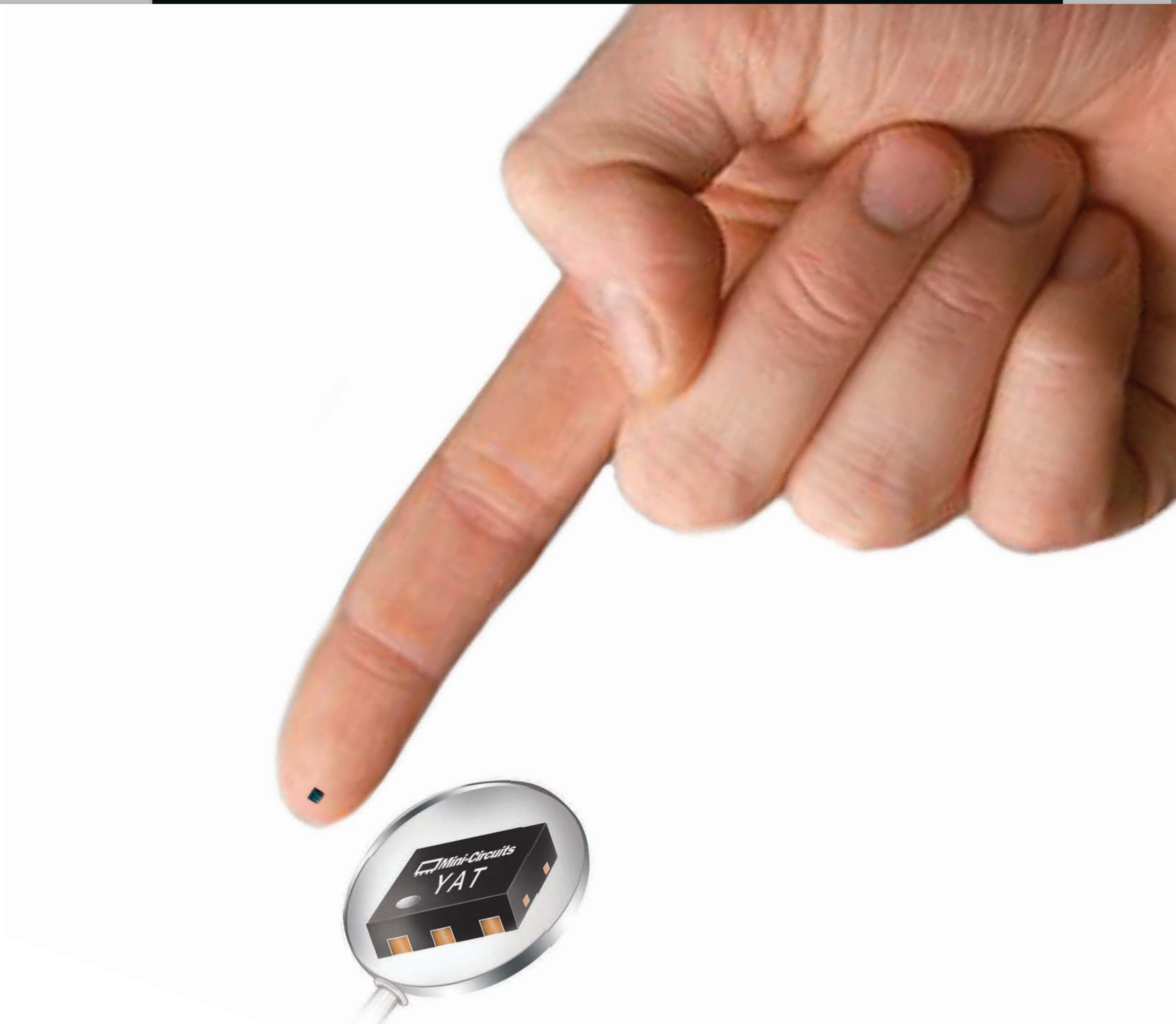
TENSOR DECOMPOSITIONS FOR SIGNAL PROCESSING APPLICATIONS

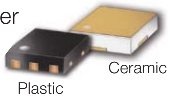
A MEDICAL SENSOR REVOLUTION

CRAMÉR–RAO BOUND ANALOG OF BAYES' RULE

IEEE Signal Processing Society®

IEEE


Mini-Circuits
YAT


Ceramic
Plastic

[CONTENTS]

[VOLUME 32 NUMBER 2]

[SPECIAL SECTION—SIGNAL PROCESSING TECHNIQUES FOR ASSISTED LISTENING]

Digital Object Identifier 10.1109/MSP.2014.2372532

[FEATURES]

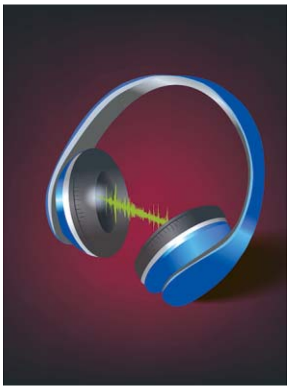
[COLUMNS]

[DEPARTMENTS]

[IEEE SIGNAL PROCESSING magazine]

[COVER]

HEADSET IMAGE LICENSED BY INGRAM PUBLISHING,
SOUND WAVE LICENSED BY GRAPHIC STOCK



IEEE prohibits discrimination, harassment, and bullying. For more information, visit http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

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 US$20.00 (first copy only), nonmembers US$201.00 per copy. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright Law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA; 2) pre-1978 articles without fee. Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. For all other copying, reprint, or republication permission, write to IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854 USA. Copyright ©2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. 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.

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.

Digital Object Identifier 10.1109/MSP.2014.2372533

SUSTAINABLE FORESTRY INITIATIVE Certified Chain of Custody At Least 25%Certified Forest Content www.sfiprogram.org SFI-01042

[from the **EDITOR**]

Min Wu
Editor-in-Chief
minwu@umd.edu

Sharing Signal Processing with the World

I am writing this editorial for the March issue of *IEEE Signal Processing Magazine* (*SPM*) as 2014 comes to a close. My son's elementary school class just learned about the Jewish holiday of Hanukkah, the Muslim holiday of Ramadan, and the African-American celebration of Kwanzaa. This was in addition to the Thanksgiving and Christmas holidays that students are already keenly aware of. The school encourages parents to share any major holidays that their families celebrate as part of a cultural education for global citizenship. I volunteered to teach my son's class about the Lunar New Year celebrated by Chinese and several other Asian ethnic groups. Indeed, wherever we are and whatever ethnic roots we have, we are all proud of our cultural heritage. Through celebrations, not only do we enjoy this important time with our families and friends, but more importantly, we pass the cultural assets onto the next generation and share our cultures with the world.

This pride and desire to share are also common in our professional lives. Many professional groups, including the IEEE, have public outreach efforts to raise awareness of their (respective) professions and to attract more young people to join. One recent high-profile effort is CODE.org, which aims at expanding participation in computer science by making it available in elementary, middle, and high schools (known as K–12). With hands-on participation by celebrities and even U.S. President Barack Obama, this nonprofit organization developed accessible means to demystify computer programming. Within just a year from its launch, CODE.org reportedly prepared 3,000 new teachers in K–12 schools, brought an introductory course to 4 million students in 90,000 classrooms, and had tens of millions of people try an hour of programming. Even my son in elementary school proudly brought home a certificate that declared he completed an hour of coding!

This is one of many successful efforts by the computer science community in bringing excitement and the "cool factor" to the public as well as in attracting funding agencies' support. What can we learn from their efforts to advocate our field, i.e., to explain what signal processing is and to share the far-reaching contributions of signal processing with the world?

The leadership of the IEEE Signal Processing Society (SPS) has been working on this for a number of years. The most recent effort is an outreach video series led by SPS President-Elect Rabab Ward. The first video is now available on YouTube and shows the ubiquitous contributions of signal processing in our everyday life [1]. This 2-minute video uses multimedia to visualize "Signal Processing Inside," a notion coined in the September 2004 editorial by SPS Past President K.J. Ray Liu (who was editor-in-chief of *SPM* at the time). Check it out, and please share this cool video with your schools, colleagues, friends, and families.

Now comes the harder part: how can we go further to explain in accessible terms and engaging styles what signals and signal processing are? Published over a decade ago, the book *Engineering Our Digital Future: The Infinity Project* by Orsak et al. offered a unique curriculum for high school students and college freshmen to learn about digital technologies. Authored by active volunteers in the SPS community, it covered the creation, storage, and communications of various modalities of signals. Since then, digital cameras, broadband communications, and online platforms have become affordable and ubiquitous to everyone including kids and senior citizens. These advances have lowered the entry point for the general public to relate and appreciate signal processing technologies, but perhaps not through a systematic curriculum and hundreds of textbook pages.

Could and should *SPM*—known for its fine tutorials—fill in this gap to bring short stand-alone tutorials accessible to a broader audience (in addition to serving its traditional readership)? Such articles may supplement overview videos to raise awareness and the visibility of signal processing; they might serve as a bridge to invite interested students, teachers, and professionals to explore in-depth articles in the magazine (as well as the SigView online tutorials highlighted by SPS President Alex Acero in the January 2015 issue of *SPM*).

To quote Nobel Laureate Richard Feynman, the author of *The Feynman Lectures on Physics,* "If you can't explain something to a six-year-old, you really don't understand it yourself." Perhaps six-year-old readers are on the other extreme from the expert audience to which many of our authors are accustomed. As a compromise, how about explaining signal processing to a sixth grader? I invite you, our readers, to join our editorial team for this exercise, as we explore new opportunities to share signal processing with the world.

REFERENCES

[1] IEEE SPS. "What is signal processing?" [Online]. Available: https://www.youtube.com/watch?v=EErkgr1MWw0

[2] G. C. Orsak, S. L. Wood, S. C. Douglas, D. C. Muson, J. R. Treichler, R. A. Athale, and M. W. Yoder, *Engineering Our Digital Future: The Infinity Project.* Englewood Cliffs, NJ: Prentice Hall, 2003.

Digital Object Identifier 10.1109/MSP.2014.2378014
Date of publication: 12 February 2015

Mini-Circuits®
TSS

Alex Acero
2014–2015 SPS President
a.acero@ieee.org

The IEEE Signal Processing Cup: A Competition for Undergraduate Students

Signal processing is becoming part of the curriculum in many undergraduate engineering programs. To leverage and reinforce this, the IEEE Signal Processing Society (SPS) has created the IEEE Signal Processing Cup, a competition that provides undergraduate students with the opportunity to form teams and work together to solve a challenging and interesting real-world problem using signal processing techniques and methods.

The theme of the inaugural 2014 competition was "Image Restoration/Superresolution for Single-Particle Analysis." Each team participating in the competition was composed of one faculty member (who is the supervisor of the team members), at most one graduate student (who will assist the supervisor), and at least three but no more than ten undergraduates. At least three of the undergraduate team members had to be either IEEE SPS members or student members.

Twelve teams from all over the world submitted their work. Three of these teams were selected to present their work at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2014. I was part of the panel that judged the three projects, and I was pleasantly surprised by the work of these talented and energetic undergraduates. I could envision several future start-ups based on all three projects.

All three projects were very good, but we had the difficult job of ranking them in first, second, and third place as follows.

- *First place*: EPOCH (Anik Khan, Forsad Al Hossain, Tawab Ullas, Md. Abu Rayhan, and Mohammad Ariful Haque) from Bangladesh University of Engineering and Technology

> I WAS PART OF THE PANEL THAT JUDGED THE THREE PROJECTS, AND I WAS PLEASANTLY SURPRISED BY THE WORK OF THESE TALENTED AND ENERGETIC UNDERGRADUATES.

- *Second place*: NtUeLsA (Kai-Wen Liang, Yen-Chen Wu, Guan-Lin Chao, Kuan-Hao Huang, Shao-Hua Sun, Ming-Jen Yang, Po-Wen Hsiao, Ti-Fen Pan, Yi-Ching Chiu, Wei-Chih Tu, and Shao-Yi Chien) from National Taiwan University
- *Third place*: Uchihas (Emroz Khan, Shiekh Zia Uddin, Mukhlasur Rahman Tanvir, and Md. Kamrul Hasan) from Bangladesh University of Engineering and Technology.

Teams were awarded SPS-sponsored prizes in the amounts of US$5,000, US$2,500, and US$1,000 for first, second, and third place, respectively.

Each team invited to ICASSP2014 had their travel expenses supported by the SPS. Each team member was offered up to US$1,200 for continental travel, or US$1,700 for intercontinental travel, with at most three people from each team supported. We are very grateful to the Bioimaging and Signal Processing Technical Committee for proposing the theme for the first Signal Processing Cup and steering this competition and to Mathworks for providing complimentary licenses of MATLAB and selected toolboxes.

Following the success of the first competition, we will be hosting another competition at ICASSP2015 with the theme of "Heart Rate Monitoring During Physical Exercise Using Wrist-Type Photoplethysmographic Signals," and we expect just as much interest!

The students's feedback was extremely positive: they told us that they learned much about signal processing and working on a team, and they also relayed that this was a tremendously enriching experience for them. I encourage professors to help put together more teams for the Signal Processing Cup and undergraduate students to participate in this unique opportunity.

A Acero

Digital Object Identifier 10.1109/MSP.2014.2378016
Date of publication: 12 February 2015

Top Downloads in IEEE *Xplore*

The "Reader's Choice" column in *IEEE Signal Processing Magazine* 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 is based on download data through June 2014. 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).

TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	ABSTRACT	RANK IN IEEE TOP 100						*N* TIMES IN TOP 100 (SINCE JAN 2011)
		JUN 2014	MAY 2014	APR 2014	MAR 2014	FEB 2014	JAN 2014	
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 article 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 algorithm and compared with the standard EKF.	12	15	**9**	**9**	10	31	39
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.	**13**	33	46	31	42	24	21
AN INTRODUCTION TO COMPRESSIVE SAMPLING Candes, E.J.; Wakin, M.B. *IEEE Signal Processing Magazine* vol. 25, no. 2, Mar. 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.	35	31	27	21	19	**14**	41
IMAGE QUALITY ASSESSMENT FOR FAKE BIOMETRIC DETECTION: APPLICATION TO IRIS, FINGERPRINT, AND FACE RECOGNITION Galbally, J; Marcel, S.; Fierrez, J. *IEEE Transactions on Image Processing* vol. 23, no. 2, 2014, pp. 710–724	This paper uses 25 general image quality features extracted from the authentication image to distinguish between legitimate and imposter samples for fingerprint, iris, and two-dimensional face biometrics.	**50**				74	**50**	3
NEW CHALLENGES FOR IMAGE PROCESSING RESEARCH Pappas, T.N. *IEEE Transactions on Image Processing* vol. 20, no. 12, 2011, p. 3321	The editor-in-chief of *IEEE Transactions on Image Processing* addresses the direction of the journal and image processing.	**77**						1
SCALING UP MIMO: OPPORTUNITIES AND CHALLENGES WITH VERY LARGE ARRAYS Rusek, F.; Persson, D.; Lau, B.K.; Larsson, E.G.; Marzetta, T.L.; Edfors, O.; Tufvesson, F. *IEEE Signal Processing Magazine* vol. 30, no. 1, 2013, pp. 40–60	The more antennas the transmitter/receiver is equipped with and the more degrees of freedom that the propagation channel can provide, the better the performance in terms of data rate or link reliability. This article quantifies the reliability and achievable rates.	80	**77**	93	78		82	15

Digital Object Identifier 10.1109/MSP.2014.2374977
Date of publication: 12 February 2015

		RANK IN IEEE TOP 100						*N* TIMES
TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	**ABSTRACT**	**JUN 2014**	**MAY 2014**	**APR 2014**	**MAR 2014**	**FEB 2014**	**JAN 2014**	**IN TOP 100 (SINCE JAN 2011)**
IMAGE SUPER-RESOLUTION VIA SPARSE REPRESENTATION Yang, J.; 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.	84	81		55	92	**27**	12
K-SVD: AN ALGORITHM FOR DESIGNING OVERCOMPLETE DICTIONARIES FOR SPARSE REPRESENTATION Aharon, M.; Elad, M.; Bruckstein, A. *IEEE Transactions on Signal Processing* vol. 54. no. 11, 2006, pp. 4311–4322	K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data in a computationally efficient manner.	**85**						1
AN OVERVIEW OF MASSIVE MIMO: BENEFITS AND CHALLENGES Lu, L.; Li, G.Y.; Swindlehurst, A.L.; Ashikhmin, A.; Zhang, R. *IEEE Journal on Selected Topics in Signal Processing* vol. 8, no. 5, 2014, pp. 742–758	Equipping cellular base stations with a very large number of antennas, potentially allows for orders of magnitude improvement in spectral and energy efficiency. This paper presents an extensive overview and analysis of massive MIMO systems.	**89**						1
TENSORS: A BRIEF INTRODUCTION Comon, P. *IEEE Signal Processing Magazine* vol. 31, no. 3, 2014, pp. 44–53	This article explains the different properties of tensors and matrices. In particular the canonical polyadic tensor decomposition and singular-value matrix decomposition.	97		**23**				2
THE PAST, PRESENT, AND THE FUTURE OF UNDERWATER ACOUSTIC SIGNAL PROCESSING Vaccaro, R.J. *IEEE Signal Processing Magazine* vol. 15, no. 4, 1998, pp. 21–51	A collection of articles by members of the Underwater Acoustic Signal Processing Technical Committee ranging from 1960s history to future applications including synthetic aperture sonar.		**89**					1
SPARSE REPRESENTATION FOR BRAIN SIGNAL PROCESSING: A TUTORIAL ON METHODS AND APPLICATIONS Li, Y.; Yu, Z.L.; Bi, N.; Xu, Y.; Gu, Z.; Amari, S.I. *IEEE Signal Processing Magazine* vol. 31, no. 3, 2014, pp. 96–106	Formulates the task of blind source separation of brain signals and other brain signal processing problems as an underdetermined linear model and solves via sparse representations. Includes applications such as BSS and EEG inverse imaging, feature selection, and classification.			**72**				1

[SP]

SPS Fellows and Award Winners Recognized

In this column of *IEEE Signal Processing Magazine*, 51 IEEE Signal Processing Society (SPS) members are recognized as Fellows, and award recipients are announced.

51 SPS MEMBERS ELEVATED TO FELLOW

Each year, the IEEE Board of Directors confers the grade of Fellow on up to one-tenth of 1% of the Members. To qualify for consideration, an individual must have been a Member, normally for five years or more, and a Senior Member at the time for nomination to Fellow. The grade of Fellow recognizes unusual distinction in IEEE's designated fields.

The SPS congratulates the following 51 SPS members who were recognized with the grade of Fellow as of 1 January 2015:

Jean Armstrong, Melbourne, Australia: For contributions to the theory and application of orthogonal frequency division multiplexing in wireless and optical communications.

Kristine Bell, Reston, Virginia, United States: For contributions to statistical signal processing with radar and sonar applications.

Ewert Bengtsson, Uppsala, Sweden: For contributions to quantitative microscopy and biomedical image analysis.

Daniel Bliss, Tempe, Arizona: For contributions to adaptive sensor systems in radar and communications.

Christian Cachin, Ruschlikon, Switzerland: For contributions to steganography and secure distributed systems.

Joseph Cavallaro, Houston, Texas, United States: For contributions to VLSI architectures and algorithms for signal processing and wireless communications.

Biao Chen, Syracuse, New York, United States: For contributions to decentralized signal processing in sensor networks and interference management of wireless networks.

Israel Cohen, Haifa, Israel: For contributions to the theory and application of speech enhancement.

*Iain Colling*s, Epping, Australia: For contributions to multiple user and multiple antenna wireless communication systems.

Michael Davies, Edinburgh, United Kingdom: For contributions to sparse representations in signal processing and compressed sensing.

Mérouane Debbah, Gif-sur-Yvette, France: For contributions to the theory and application of signal processing in wireless networks.

Lieven De Lathauwer, Leuven, Belgium: For contributions to signal processing algorithms using tensor decompositions.

Gordon Frazer, Edinburgh, Australia: For contributions to advanced over-the-horizon radar.

Pascale Fung, Clear Water Bay, Hong Kong: For contributions to human–machine interactions.

Xiqi Gao, Nanjing, China: For contributions to broadband wireless communications and multirate signal processing.

Monisha Ghosh, Melville, New York, United States: For contributions to cognitive radio and signal processing for communication systems.

S. Gunasekaran, Rome, New York, United States: For contributions to high-performance computer vision algorithms for airborne applications.

K.V.S. Hari, Bangalore, India: For contributions to high-resolution signal parameter estimation.

Zhihai He, Columbia, Missouri, United States: For contributions to video communication and visual sensing technologies.

Jianying Hu, Yorktown Heights, New York, United States: For contributions to pattern recognition in business and health analytics and document analysis.

Hong Jiang, Santa Clara, California, United States: For leadership in parallel multimedia computing architectures and systems.

Tzyy-Ping Jung, La Jolla, California, United States: For contributions to blind source separation for biomedical applications.

Simon King, Edinburgh, United Kingdom: For contributions to text-to-speech synthesis and speech technology.

Stefanos Kollias, Athens, Greece: For contributions to intelligent systems for multimedia content analysis and human–machine interaction.

Deepa Kundur, Toronto, Canada: For contributions to signal processing techniques for multimedia and cybersecurity.

Edmund Lam, Pokfulam, Hong Kong: For contributions to modeling and computational algorithms in imaging applications.

Henry Leung, Calgary, Canada: For contributions to chaotic communications and nonlinear signal processing.

Zicheng Liu, Redmond, Washington: For contributions to visual processing for multimedia interaction.

David Love, West Lafayette, Indiana, United States: For contributions to feedback-adaptive wireless communication systems.

Detlev Marpe, Berlin, Germany: For contributions to video coding research and standardization.

Teresa Pace, Orlando, Florida, United States: For contributions to image and signal processing algorithms for sensor systems.

Mark Plumbley, London, United Kingdom: For contributions to latent variable analysis.

Markus Rupp, Wien, Austria: For contributions to adaptive filters and communication technologies.

Robert Safranek, Warren, New Jersey, United States: For contributions to

Digital Object Identifier 10.1109/MSP.2014.2378191
Date of publication: 12 February 2015

perceptual image and video compression and quality.

Paris Smaragdis, Urbana, Illinois, United States: For contributions to audio source separation and audio processing.

Hing Cheung So, Kowloon, China: For contributions to spectral analysis and source localization.

Eckehard Steinbach, Munich, Germany: For contributions to visual and haptic communications.

Wonyong Sung, Seoul, South Korea: For contributions to real-time signal processing systems.

Johan Suykens, Leuven, Belgium: For developing the least squares support vector machines.

Dacheng Tao, Sydney, Australia: For contributions to pattern recognition and visual analytics.

David Taubman, Sydney, Australia: For contributions to image and video communications.

James Truchard, Austin, Texas, United States: For leadership in instrumentation and computing for signal processing.

Vesa Valimaki, Espoo, Finland: For contributions to synthesis and processing of audio signals.

An-Yeu (Andy) Wu, Taipei, Taiwan: For contributions to digital signal processing algorithms and VLSI designs for communication IC/SoC.

Hsiao-Chun Wu, Baton Rouge, Louisiana, United States: For contributions to digital video broadcasting and wireless systems.

Isao Yamada, Tokyo, Japan: For contributions to inverse problems and learning in signal processing.

Liuqing Yang, Fort Collins, Colorado, United States: For contributions to theory and practice of ultrawideband communications.

Aylin Yener, University Park, Pennsylvania, United States: For contributions to wireless communication theory and wireless information security.

Moti Yung, New York, New York, United States: For contributions to cryptography.

Wei Zhang, Sydney, Australia: For contributions to cognitive radio communications.

Haitao Zheng, Santa Barbara, California, United States: For contributions to dynamic spectrum access and cognitive radio networks.

2014 IEEE SPS AWARDS PRESENTED IN BRISBANE, AUSTRALIA

The IEEE SPS congratulates the following SPS members who will receive the Society's prestigious awards during ICASSP 2015 in Brisbane, Australia, 19–24 April 2015.

The Society Award honors outstanding technical contributions in a field within the scope of the IEEE SPS and outstanding leadership within that field. The Society Award comprises a plaque, a certificate, and a monetary award of US$2,500. It is the highest-level award bestowed by the IEEE SPS. This year's recipient is K.J. Ray Liu, "for influential technical contributions and profound leadership impact."

The IEEE Signal Processing Magazine Best Paper Award honors the author(s) of an article of exceptional merit and broad interest on a subject related to the Society's technical scope and appearing in the Society's magazine. The prize comprises US$500 per author (up to a maximum of US$1,500 per award) and a certificate. In the event that there are more than three authors, the maximum prize shall be divided equally among all authors and each shall receive a certificate. This year, the IEEE Signal Processing Magazine Best Paper Award recipients are Sergios Theodoridis, Konstantinos Slavakis, and Isao Yamada for their article "Adaptive Learning in a World of Projections: A Unifying Framework for Linear and Nonlinear Classification and Regression Tasks," published in *IEEE Signal Processing Magazine*, vol. 28, no. 1, Jan. 2011.

The IEEE Signal Processing Magazine Best Column Award honors the author(s) of a column of exceptional merit and broad interest on a subject related to the Society's technical scope and appearing in the Society's magazine. The prize shall consist of US$500 per author (up to a maximum of US$1,500 per award) and a certificate. In the event that there are more than three authors, the maximum prize shall be divided equally among all authors and each shall receive a certificate. This year, the IEEE Signal Processing Magazine Best Column Award recipients are Göran Bergqvist and Erik G. Larsson for their article "The Higher-Order Singular Value Decomposition: Theory and an Application," published in *IEEE Signal Processing Magazine*, vol. 27, no. 3, May 2010.

Two Technical Achievement Awards will be presented this year. Moeness G. Amin will receive the award "for fundamental contributions to signal processing algorithms for communications, satellite navigations, and radar imaging." Richard G. Baraniuk will be recognized "for contributions to the theory and applications of sparsity and compressive sensing." The Technical Achievement Award honors a person who, over a period of years, has made outstanding technical contributions to theory and/or practice in technical areas within the scope of the Society, as demonstrated by publications, patents, or recognized impact on this field. The prize for the award is US$1,500, a plaque, and a certificate.

The Meritorious Service Award is presented this year to V. John Mathews "for exemplary service to and leadership in the IEEE Signal Processing Society." The award comprises a plaque and a certificate; judging is based on dedication, effort, and contributions to the Society.

The SPS Education Award honors educators who have made pioneering and significant contributions to signal processing education. Judging is based on a career of meritorious achievement in signal processing education as exemplified by writing of scholarly books and texts, course materials, and papers on education; inspirational and innovative teaching; and creativity in the development of new curricula and methodology. The award comprises a plaque, a monetary award of US$1,500, and a certificate. The recipient of the SPS Education Award is Sergios Theodoridis, "for sustained contributions to education in the area of machine learning for signal processing."

The Sustained Impact Paper Award honors the author(s) of a journal article of broad interest that has had sustained impact over many years on a subject related to the Society's technical scope. The prize consists of US$500 per author (up to a maximum of US$1,500 per award) and a certificate. In the event that there are more than three authors, the maximum prize shall be divided equally among all authors and each shall receive a certificate. To be eligible for consideration, a paper must have appeared in one of the

IEEE SPS transactions or *IEEE Journal of Selected Topics in Signal Processing*, in an issue predating the Spring Awards Board meeting by at least ten years (typically held in conjunction with ICASSP). The recipients of the first Sustained Impact Paper Award are:

- Stephane G. Mallat and Zhifeng Zhang, "Matching Pursuits with Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, Dec. 1993.

Six Best Paper Awards will be awarded, honoring the author(s) of a paper of exceptional merit dealing with a subject related to the Society's technical scope and appearing in one of the Society's transactions, irrespective of the author's age. The prize is US$500 per author (up to a maximum of US$1,500 per award) and a certificate. Eligibility is based on a five-year window preceding the year of election, and judging is based on general quality, originality, subject matter, and timeliness. Up to six Best Paper Awards may be presented each year. This year, the awardees are:

- Namrata Vaswani and Wei Lu, "Modified-CS: Modifying Compressive Sensing for Problems with Partially Known Support," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, Sept. 2010.
- Hiroshi Sawada, Shoko Araki, and Shoji Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, Mar. 2011.
- Alexey Ozerov and Cédric Févotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, Mar. 2010.
- *Stefania Sardellitti, Massimiliano Giona, and Sergio Barbarossa*, "Fast Distributed Average Consensus Algorithms Based on Advection-Diffusion Processes," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, Feb. 2010.
- Federico S. Cattivelli and Ali H. Sayed, "Diffusion LMS Strategies for Distributed Estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, Mar. 2010.
- Rony Ferzli and Lina J. Karam, "A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, Apr. 2009.

The Young Author Best Paper Award honors the author(s) of an especially meritorious paper dealing with a subject related to the Society's technical scope and appearing in one of the Society's transactions and who, upon date of submission of the paper, is fewer than 30 years of age. The prize is US$500 per author (up to a maximum of US$1,500 per award) and a certificate. Eligibility is based on a three-year window preceding the year of election, and judging is based on general quality, originality, subject matter, and timeliness. Five Young Author Best Paper Awards are being presented this year:

- Tomáš Filler and Jan Judas, for the paper coauthored with Jessica Fridrich, "Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, Sept. 2011.
- Jort F. Gemmeke, for the paper coauthored with Tuomas Virtanen and Antti Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, Sept. 2011.
- Daniele Giacobello, for the paper coauthored with Mads Græsbøll Christensen, Manohar N. Murthi, Søren Holdt Jensen, and Marc Moonen, "Sparse Linear Prediction and Its Applications to Speech Processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, July 2012.
- Tiangao Gou and Chenwei Wang, for the paper coauthored with Syed A. Jafar, "Aiming Perfectly in the Dark-Blind Interference Alignment Through Staggered Antenna Switching," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, June 2011.
- Meisam Razaviyayn, for the paper coauthored with Gennady Lyubeznik and Zhi-Quan Luo, "On the Degrees of Freedom Achievable Through Interference Alignment in a MIMO Interference Channel," *IEEE Transactions on Signal Processing,* vol. 60, no. 2, Feb. 2012.

The IEEE Signal Processing Letters Best Paper Award honors the author(s) of a letter article of exceptional merit and broad interest on a subject related to the Society's technical scope and appearing in *IEEE Signal Processing Letters*. The prize shall consist of US$500 per author (up to a maximum of US$1,500 per award) and a certificate. To be eligible for consideration, an article must have appeared in *IEEE Signal Processing Letters* in an issue predating the Spring Awards Board meeting by five years (typically held in conjunction with ICASSP). Judging shall be on the basis of the technical novelty, the research significance of the work, quality, and effectiveness in presenting subjects in an area of high impact to the Society's members. The recipients of the IEEE Signal Processing Letters Best Paper Award are

- Emanuël A.P. Habets, Sharon Gannot, and Israel Cohen, "Late Reverberant Spectral Variance Estimation Based on a Statistical Model," *IEEE Signal Processing Letters*, vol. 16, no. 9, Sept. 2009.

2014 CHAPTER OF THE YEAR AWARD

The IEEE SPS Malaysia Chapter has been selected as the fourth recipient of the 2014 Chapter of the Year Award. The award is presented annually to a Chapter that has provided its membership with the highest quality of programs, activities, and services. The SPS Malaysia Chapter will receive a certificate and a monetary award of US$1,000 to support local Chapter activities. The Chapter will publish an article in a future issue of *IEEE Inside Signal Processing e-Newsletter*.

SPS MEMBERS RECEIVE 2015 IEEE AWARDS

The following SPS members will receive 2015 IEEE Technical Field Awards.

(continued on page 15)

John Edwards

Signal Processing Drives a Medical Sensor Revolution

Sensor technology's impact on health care is growing rapidly. New applications are appearing almost daily. Wireless sensors are now used in an ever-growing number ways, such as monitoring glucose levels in diabetics, recording and tracking heart irregularities, and diagnosing infectious diseases. Linking sensors to mobile phones has made wearable sensors a reality, allowing individuals to monitor not only chronic diseases but also their lifestyle activities.

"There's a lot happening in health monitoring," says Andreas Spanias, a professor in the Arizona State University School of Electrical, Computer, and Energy Engineering. "Integrated sensors, on mobile phones, for example, can monitor vital signs, such as heart rates, breathing activity, oxygenation, and blood pressure," says Spanias, who is also the founder and director of the university's Sensor Signal and Information Processing (SenSIP) Center, a National Science Foundation (NSF) Industry and University Cooperative Research Center (NSF/UCRC).

Spanias says that signal processing is essential to optimal sensor operation and performance. "Our industry collaborators build inexpensive sensors, and signal processing improves precision and event detection using machine learning and fusion," he notes. "Even if the data is noisy or contains artifacts, signal processing can reduce noise effects. Signal processing algorithms, for example, will make wireless health monitoring more accurate and reliable. Signal processing makes it possible to use data from several sensors and combine the information appropriately to maximize the probability of correct detection."

Signal processing algorithms are likely to become even more essential to wireless health-care sensor development in the years to come. The technology is now entering a new phase made possible by the development of microscopic nanosensors and nanorobots designed for insertion into bodily tissues and the bloodstream. "With so many sensors in the body, and the large volumes of data they will be transmitting, how do you fish out the information that you need?" Spanias asks. "Signal processing and biomedical informatics will have a big role in that area, and algorithms will enable reliable prediction of disease and incentivize healthy lifestyles."

[LINKING SENSORS TO MOBILE PHONES HAS MADE WEARABLE SENSORS A REALITY, ALLOWING INDIVIDUALS TO MONITOR NOT ONLY CHRONIC DISEASES, BUT ALSO THEIR LIFESTYLE ACTIVITIES.]

INTRABODY NETWORKS

A system of wirelessly networked intrabody sensors and actuators could lead to revolutionary new applications in health-care monitoring, potentially creating innovative approaches to the treatment of an almost endless number of diseases, both major and minor. Yet an important obstacle to the development of reliable intrabody sensor/actuator networks is the fact that most health-care sensor network research to date has focused on communication along the body surface via devices linked through traditional electromagnetic radio-frequency (RF) transmissions. Such technology has significant limitations for intrabody system developers, however, due to the physical nature of propagation within the human body, which is composed primarily of water, a medium through which RF electromagnetic waves do not easily propagate.

Researchers at Northeastern University in Boston, in collaboration with researchers at the University of Catania and the Sapienza University of Rome, are hoping that by taking a novel approach to wireless sensor communication—ultrasonic networking technology—they can make intrabody sensors and actuators an accurate and reliable technology.

The researchers are currently pursuing a closed-loop combination of mathematical modeling, simulation, and experimental evaluation to determine the practicality of using ultrasonic networking in human tissues. "A major challenge is creating a waveform that's resistant to the effects of multipath and scattering," says team member Tommaso Melodia, an associate professor in Northeastern University's Department of Electrical and Computer Engineering.

The magnitude and direction of a reflected wave depends on the orientation of the boundary surface as well as on the acoustic impedance of the tissue. Scattered reflections happen whenever an acoustic wave encounters an object that's relatively small in relation to its wavelength or meets a tissue with an irregular surface. "At the receiver, basically, you receive a combination of multiple replicas of the same signal," Melodia says. "You have to create a receiver that can differentiate between these various signals; it basically needs to be able to record the original signal from multiple replicas that it's receiving."

Digital Object Identifier 10.1109/MSP.2014.2373231
Date of publication: 12 February 2015

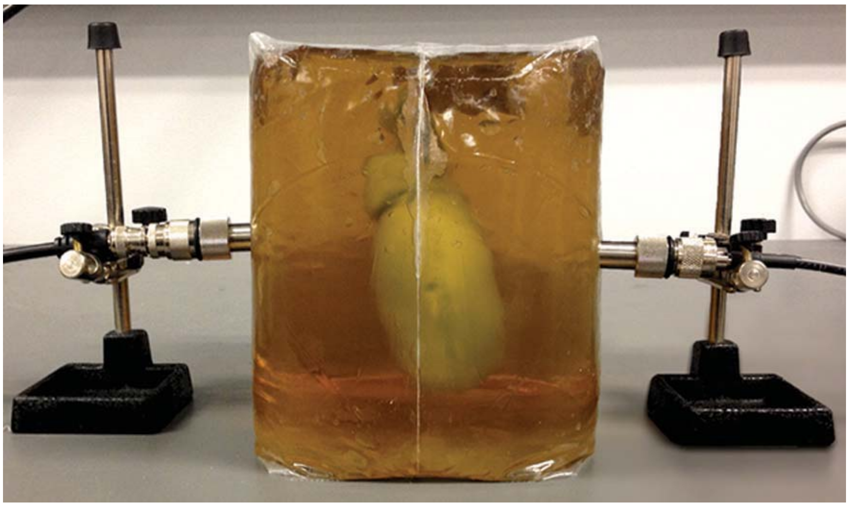
1053-5888/15©2015IEEE

Signal processing is also essential for creating receivers that can cope with an onslaught of data coming in from large numbers of ultrasonic sensors floating inside a body. "Basically, solving some mathematical optimization problems gives us the best way to share the channel between different devices that are trying to transmit at the same time," Melodia explains.

Early in their investigation, the researchers proposed basing their ultrasonic intrabody network on ultrasonic wideband (UsWB), a relatively new multipath-resilient physical and medium access control (MAC) layer integrated protocol. UsWB is the only MAC protocol specifically designed for ultrasonic intrabody sensor networks, while a wide variety of MAC protocols designed for traditional RF-based wireless networks are currently available.

According to the researchers, UsWB is based on the concept of transmitting short carrierless ultrasonic pulses following a pseudorandom adaptive time-hopping pattern, featuring a superimposed adaptive spreading code. After testing the protocol, the researchers were able to show that UsWB enables nodes to flexibly trade data rate performance for power consumption while allowing multiple concurrent sensors to coexist by dynamically adapting their transmission rate to channel and interference conditions.

Recently, the researchers compared the performance of UsWB with a pair of existing MAC protocols originally designed for use with RF-based wireless networks: ALOHA and carrier sense multiple access (CSMA). Their tests showed that UsWB generally outperforms ALOHA in terms of throughput, although CSMA can achieve comparable performance under certain kinds of setups. Additionally, according to the researchers, ALOHA and CSMA both exhibit very high packet drop rates compared to UsWB, which always keeps the packet drop rate below a given threshold. The tests also found that UsWB performs better than either ALOHA or CSMA in terms of short-term fairness, average packet delay, and delay variation. Finally, the researchers discovered that CSMA has the highest energy consumption per bit, due to long idle listening times. UsWB's bit cost, on the other hand, is the lowest and can be further reduced by trading throughput for energy consumption through energy-minimizing rate adaptation, the researchers say.



[FIG1] A synthetic kidney was used in preliminary tests at Northeastern University to evaluate various ultrasonic communication technologies and configurations. (Photo credit: Northeastern University.)

Testing various intrabody network system technologies and configurations required creating an environment that mimicked real-world conditions. "We used some devices known as medical phantoms," Melodia says. For this particular project, the phantoms were synthetic devices designed to produce the basic ultrasound characteristics of real tissue. "We've used mostly kidneys so far," Melodia says. "We have a synthetic kidney that we transmit information through" (Figure 1). The next step is building a miniaturized prototype of the transceiver. "We have a prototype that works well, but it's big," Melodia says. "We want to build a miniaturized platform that will be able to do what we're doing now, but be much smaller and can be implantable.

Melodia predicts that an intrabody network system could become available for general use within a decade. "It will take a lot of work, but that seems like a realistic possibility," he says.

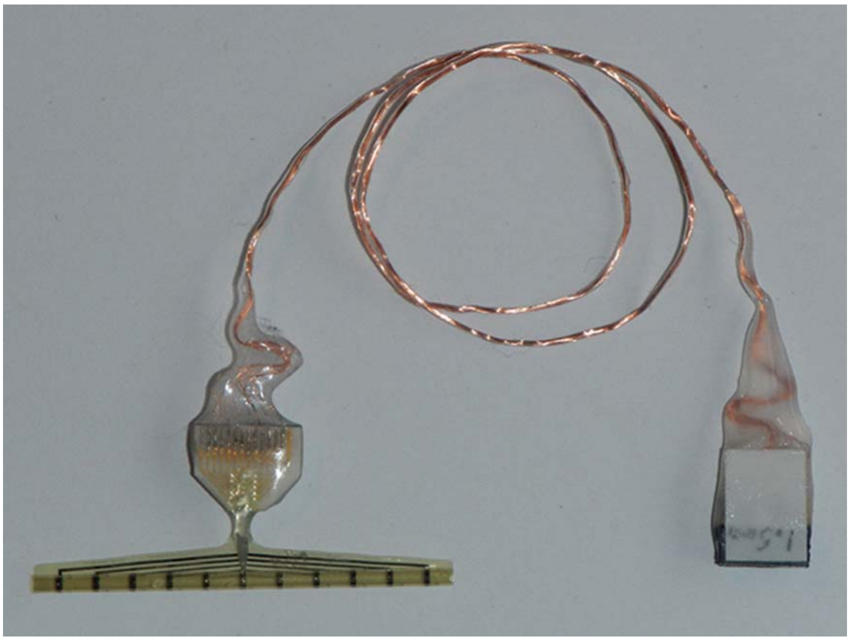
NEURAL RECORDING SENSORS

A research group in the University of Bath's Centre for Advanced Sensor Technologies (CAST) is investigating the use of implantable devices for electroneurogram (ENG) signal recording, potentially increasing the quantity and quality of received information. An ENG is used to visualize directly recorded electrical activity of neurons in the central nervous system (brain and spinal cord) or the peripheral nervous system.

Reliably collecting neural data is a goal that has eluded numerous researchers for many years. "Nerves tend to come in bundles of hundreds or thousands and carry neural traffic to different destinations to and from the central nervous system," says John Taylor, CAST's head and a professor in the University of Bath's Department of Electronic and Electrical Engineering. Identifying individual pathways and the traffic on them is difficult. "Several years ago we invented a technique called velocity selective recording (VSR) that, we believe, goes some way to solving this problem," Taylor continues.

Working with project collaborators, including University College London, the University of Cambridge, the University of Freiburg, and Aalborg University, the Bath researchers developed a range of implantable electrodes and amplifiers to test a technique that Taylor says is essentially a simple signal processing concept.

"Our recording technique provides real-time velocity spectral analysis of activity on a nerve," Taylor says. In practice,



[FIG2] A multielectrode nerve cuff used for velocity-selective recordings made by Martin Schuettler, a senior scientist at the University of Freiburg and chief technology officer of CorTek, a Freiuburg, Germany-based developer of a neurotechnological platform for measuring and stimulating of brain activity. (Photo credit: Martin Schuettler.)

such a nerve might contain hundreds or thousands of individual fibers [axons], with signals propagating over a wide range of velocities (up to 120 m/s in humans) in both directions. "This (technique) should be useful because neural propagation velocity and fiber diameter are generally related, so an analysis of activity by velocity and direction is equivalent to knowing the diameters of the nerves that are excited at that time," Taylor says. "Anatomy then allows us to link this to function."

In the future, Taylor says these function-specific signals might be used to design systems for controlling neuroprosthetic devices, such as providing a neural stimulator with a feedback loop for bladder control to treat urinary incontinence. "Currently available methods to provide the information we seek tend to rely on fairly classical pattern processing methods, such as clustering and principal component analysis under the generic title of 'spike sorting'," Taylor says. Such methods tend to be computationally intensive and therefore unattractive for implantation. "In addition, some form of training is generally required and may be impossible or impracticable," he adds. By contrast the signal processing required for VSR is computationally simple, power efficient and lends itself to real-time working.

Taylor says that signal processing is a key building block in the group's research. "This is because surgical considerations impose strict limits on the size and complexity of our implanted devices and hence on the sensitivity and resolution of our basic signal acquisition capability," he explains. "Signal processing can compensate for this and is used wherever possible for filtering noise, performing spectral analysis of waveforms, and ultimately for decoding the impulses that we record from the nervous system."

According to Taylor, VSR requires multiple samples of the composite propagating neural signal. Such samples are typically provided by a multielectrode cuff (MEC) placed around the nerve (Figure 2). The MEC, which is an insulating cuff typically 2–3 cm in length and containing 10–12 electrodes, is an extension of the traditional tripolar type of nerve cuff that has been implanted in many patients successfully for several decades, Taylor says.

The samples are identical but delayed by a period that depends on both the cuff geometry and the propagation velocity of the signal. To construct the velocity spectrum from this data an operation called "delay-and-add" is applied. The operation adds artificial delays that cancel the natural delays in each channel before finally adding all the signals together. "When the artificial delays are equal to the naturally-occurring ones, the spectral output passes through a peak (local maximum) indicating the presence of an excited population of axons at that velocity," Taylor says. "This is the simplest approach to VSR and the resulting spectrum is called the intrinsic velocity spectrum (IVS)." The method, he notes, is closely related to various beamforming algorithms employed in radio and radar antenna systems.

Unfortunately the method achieves relatively poor velocity selectivity, Taylor says. It has particular difficulty in distinguishing closely spaced velocity peaks. Various additional techniques have been developed to improve the velocity selectivity including the use of bandpass filters and time delay neural networks (TDNNs), Taylor explains.

One of the biggest limitations inherent in existing neural signal processors is the requirement to build complex statistical models. "These models are not only computationally expensive to produce but also require a good deal of time to 'learn' as they become patient-specific," Taylor says. "To overcome these limitations we considered an entirely different signal processing approach, based on conduction velocity instead of pattern shape."

Noise poses another challenge. "The signals we record are from biological sources and so are often very noisy," Taylor remarks. "It is not uncommon for the signal-to-noise ratio (SNR) to be less than 0 dB, and so innovative methods must be developed to extract information." Coupled with the requirements for real-time operation and good long-term stability, the challenges are not insignificant.

Recording neural activity from an intact nerve represents another highly challenging task, due to the poorly understood nature of the electrode-tissue interface and the associated problems of handling very small signals. "We have worked extensively to improve electrode and amplifier designs so as to stabilize the electrode characteristics and maximize the possible recorded SNR," Taylor says.

Taylor notes that the group's signal processing algorithms are still incomplete. "So far, we have been recording and analyzing electrically evoked ENG—neural signals produced by electrical stimulation," he says. "This is an interesting and useful exercise, but is an approximation in several ways to natural neural recording."

According to Taylor, the amplitude and SNR of the recorded signals are much larger than in comparable natural signals. Additionally, information such as pressure or joint angle are encoded in neural firing rates, so identifying the source and direction of a neural signal is only part of the overall package necessary to create a complete recording system.

"The impulses generated by electrical stimulation—compound action potentials (CAPs)—are synchronized to the stimulating pulse, so their arrival times are predictable," Taylor says. "The signal processing algorithms required to interpret them are, therefore, essentially time invariant and therefore relatively simple." Taylor notes that the researchers have recently begun modifying their VSR algorithms to include time dependence, including the ability to identify not just the velocity and direction of neural traffic but also the number of impulses in a particular velocity band arriving per second. "This has required the inclusion of statistical methods in our algorithms that we refer to as velocity spectral density (VSD)," he says.

THE RESEARCHERS ARE NOW LOOKING TO EXTEND THEIR WORK, WHICH TO DATE HAS INCLUDED ONLY SINGLE ACUTE EXPERIMENTS, TO EXTENSIVE LONG-TERM CHRONIC STUDIES IN NONHUMAN MODELS.

The researchers are now looking to extend their work, which to date has included only single acute experiments, to extensive long-term chronic studies in nonhuman models. "To achieve this, we must overcome the surgical, mechanical, and electrical challenges that are associated with long-term implantation of electronic devices," Taylor says. "New methods will need to be devised to handle communications and power concerns."

Since the project is still in a developmental stage, seeking commercial interest would be premature, Taylor says. "However we have good links with the United Kingdom's largest commercial manufacturer of implanted medical devices, indeed the only company licensed to produce implantable electronics in this country, and they are aware of and interested in our project," he says. "However, before giving it to a company for development, we have still to prove conclusively that VSR is clinically useful."

Yet Taylor is optimistic that the research will ultimately lead to a widely used medical technology. "We have tested the method in animals, and our results are quite promising so far, although we feel we are still a long way from a human implant that could be generally adopted," he says.

AUTHOR

John Edwards (jedwards@johnedwardsmedia.com) is a technology writer based in the Phoenix, Arizona, area.

[SP]

[society NEWS]

(continued from page 11)

The IEEE James L. Flanagan Speech and Audio Processing Award will be presented to Stephen John Young "for pioneering contributions to the theory and practice of automatic speech recognition and statistical spoken dialogue systems." This award was founded and is sponsored by the IEEE SPS.

The IEEE Fourier Award for Signal Processing will be presented to Georgios B. Giannakis "for contributions to the theory and practice of statistical signal processing and its applications to wireless communications."

The IEEE Donald O. Pederson Award in Solid-State Circuits will be presented to Robert Whitlock Adams "for contributions to noise-shaping data converter circuits, digital signal processing, and log-domain analog filters."

IEEE medals are the highest honor of awards presented by the IEEE. The medals will be presented at the 2015 IEEE Honors Ceremony at ICASSP in Brisbane, Australia. Three SPS members were awarded with IEEE medals for 2015:

The IEEE Edison Medal recognizes a career of meritorious achievement in electrical science, electrical engineering, or the electrical arts. James Julius Spilker will be honored "for contributions to the technology and implementation of civilian GPS navigation systems."

The IEEE Jack S. Kilby Signal Processing Medal, awarded for outstanding achievements in signal processing, was presented to Harry L. Van Trees "for fundamental contributions to detections, estimation, and modulation theory; sensor array processing; and Bayesian bounds."

The IEEE James H. Mulligan, Jr. Education Medal, distributed for a career of outstanding contributions to education in the fields of interest of the IEEE, was awarded to Richard Gordon Baraniuk "for fundamental contributions to open educational resources for electrical engineering and beyond."

[SP]

Sven Nordholm, Walter Kellermann,
Simon Doclo, Vesa Välimäki,
Shoji Makino, and John R. Hershey

Signal Processing Techniques for Assisted Listening

Natural hearing is a desirable goal in many electronic communication applications, such as hearing aids, audio conferencing, gaming, and virtual reality applications. The era of low-power, high-complexity electronics supports the implementation of computationally complex algorithms as needed to provide a more natural listening environment for the advanced augmentation of virtual reality and natural content.

As such, assisted listening techniques provide the means to communicate audio information from devices to human listeners. The main objective is to provide the user with a listening experience through the device that resembles natural hearing of the sound information. Prominent applications include virtual augmented audio, hearing aids, and cochlear implants. But the same techniques are also applicable in other communication applications such as monaural voice communication, where additional spatial information can greatly enhance the listening experience. In the realm of hearing aids, current devices aim to be more natural both for the hearing impaired and profoundly deaf by using new processing techniques. They even promise to enhance the listening experience of so-called normal hearing users. Along with increasingly affordable and growing computer power, a large variety of elaborate algorithms for overlayed audio or so-called augmented audio continuously strive toward new applications in gaming and telepresence to provide a "being there" experience.

The articles in this special issue of *IEEE Signal Processing Magazine* (*SPM*) focus on three main aspects of signal processing in this domain: audio enhancement, presentation/rendering, and evaluation. To limit the scope in this special issue, machine-learning techniques have been excluded. While it is understood that future systems for assisted listening will greatly be influenced by machine-learning-based algorithms, another special issue dedicated to this development can already be envisioned.

[THE MAIN OBJECTIVE OF ASSISTED LISTENING TECHNIQUES IS TO PROVIDE THE USER WITH A LISTENING EXPERIENCE THROUGH THE DEVICE THAT RESEMBLES NATURAL HEARING OF THE SOUND INFORMATION.]

Audio signal enhancement, particularly of speech signals, has a long research tradition and still dominates the scene, and, consequently, is also the main topic of this special issue. Techniques for single-channel and multichannel signal enhancement techniques play a preeminent role in telecommunication, hearing aids, and augmented headsets. Accordingly, fundamental problems and state-of-the-art techniques are presented in the article "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices" by Doclo et al. Beyond the description of the generic algorithms, this article emphasizes the specific problems and solutions for hearing aids and headsets addressing both the signal acquisition and the binaural rendering aspect.

As a special technique for capturing and describing the spatial information relevant for assisted listening, multichannel techniques that estimate the direct path information and suppress a combination of reverberation and diffuse noise are provided in the article "Parametric Spatial Sound Processing" by Kowalczyk et al.

Two articles provide overviews on highly relevant aspects of single-channel enhancement techniques: "Optimizing Speech Intelligibility in a Noisy Environment" by Kleijn et al. focuses on techniques for improving speech intelligibility using perceptual criteria and auditory modeling, and "Phase Processing for Single-Channel Speech Enhancement" by Gerkmann et al. provides a survey of techniques that utilize both amplitude and phase information for speech enhancement.

Processing and coding signals for cochlear implants is addressed in the article "Sound Coding in Cochlear Implants" by Wouters et al. This article describes signal processing techniques used in cochlear implants to map the information extracted from an audio signal onto cochlea excitation that a profoundly deaf person can understand.

Rendering of audio aims at providing an immersive, undisturbed listening experience for recorded information via loudspeakers or headsets with typical applications demanding high-quality sound reproduction, such as, e.g., home theaters, gaming, or telepresence systems. Betlehem et al. provide an overview of techniques to deliver audio information to multiple listeners via loudspeakers in their article "Personal Sound Zones." These techniques also have applications in providing audio in public areas without disturbing the surroundings. Then the natural sound in the environment is augmented by the rendered audio. A similar concept to augment outside information using personal headsets is presented by Välimäki et al. in "Assisted Listening Using a Headset," which also

Digital Object Identifier 10.1109/MSP.2014.2375972
Date of publication: 12 February 2015

reviews audio enhancement techniques for music listening in a noisy environment. A third article in this area, “Natural Sound Rendering for Headphones,” by Sunder et al., is an overview of techniques for rendering via headsets for applications in three-dimensional audio.

Finding methods for the evaluation and prediction of speech and audio quality is a central task for anyone working in audio signal processing. As subjective evaluations are resource intense and time consuming, it is highly desirable to find objective methods that closely match subjective measures. Objective methods provide instant feedback and results become reproducible. In their article “Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices,” Falk et al. provide an overview of algorithms for objective quality and intelligibility evaluation for hearing aids and cochlear implants.

AUDIO SIGNAL ENHANCEMENT, PARTICULARLY OF SPEECH SIGNALS, HAS A LONG RESEARCH TRADITION AND STILL DOMINATES THE SCENE, AND, CONSEQUENTLY, IS ALSO THE MAIN TOPIC OF THIS SPECIAL ISSUE.

We sincerely thank all of the authors for their high-quality contributions and are grateful for the reviewers for their invaluable help in selecting and improving the articles in this special issue. We also thank Fulvio Gini, special issues area editor, and Abdelhak Zoubir, *SPM’s* past-editor-in-chief, for their constant support, patience, and guidance in the process of outlining, soliciting, and reviewing the selected articles. Our appreciation also goes to Rebecca Wollman for her administrative guidance in the process.

We hope that you will find this special issue useful and inspiring for your work!

[SP]

[Simon Doclo, Walter Kellermann, Shoji Makino, and Sven Nordholm]

Multichannel Signal Enhancement Algorithms for Assisted Listening Devices

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

[Exploiting spatial diversity using multiple microphones]

In everyday environments, we are frequently immersed by unwanted acoustic noise and interference while we want to listen to acoustic signals, most often speech. Technology for assisted listening is then desired to increase the efficiency of speech communication, reduce listener fatigue, or just allow for enjoying undisturbed sounds (e.g., music). For people with normal hearing, assisted listening devices (ALDs) mainly aim to achieve hearing protection or increase listening comfort; however, for hearing-impaired individuals, as the most prominent user group so far, further progress of assisted listening technology is crucial for better inclusion into our world of pervasive acoustic communication.

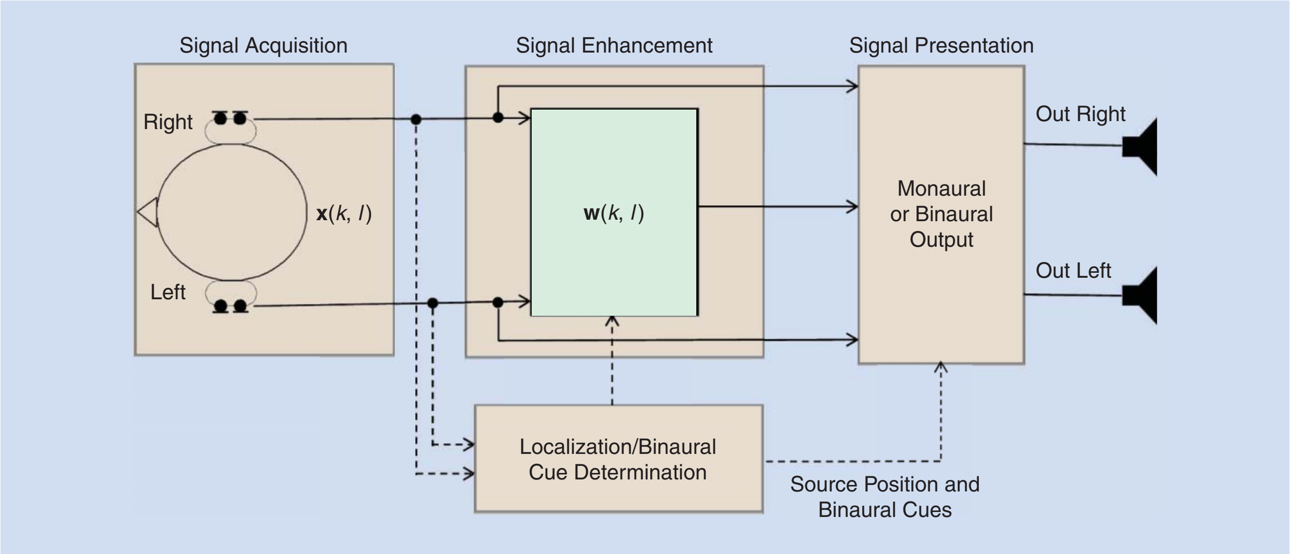
Digital Object Identifier 10.1109/MSP.2014.2366780
Date of publication: 12 February 2015

MOTIVATION

The essential functionality of ALDs comprises three steps (see Figure 1): acquiring the signals of interest, enhancing desired and removing undesired components from the acquired signals, and presenting the enhanced signal(s) to the listener.

Given the acquired microphone signals, the efficiency of such devices is largely determined by the performance of the signal processing algorithms for signal enhancement and presentation. Considering that multiple microphones are now common in many

1053-5888/15©2015 European Union


Signal Acquisition
Signal Enhancement
Signal Presentation
Right
Left
x(k, l)
w(k, l)
Monaural or Binaural Output
Out Right
Out Left
Localization/Binaural Cue Determination
Source Position and Binaural Cues

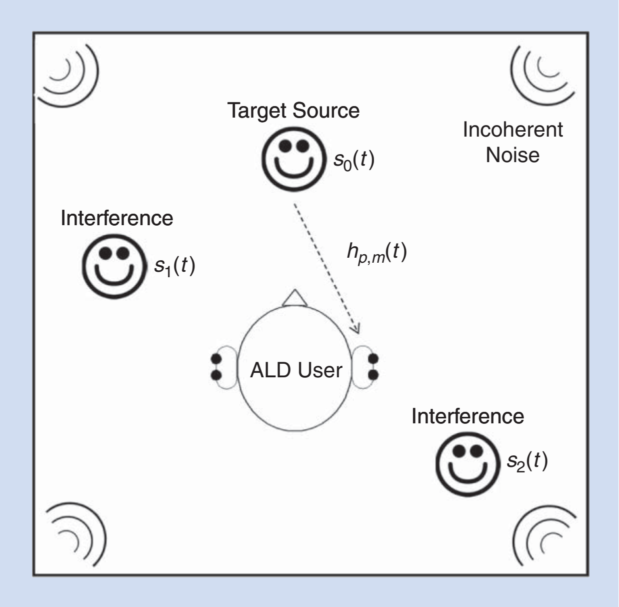
[FIG1] The main processing blocks in an ALD.

listening devices (e.g., hearing aids or mobile phones) and allow to exploit the spatial diversity in addition to the spectrotemporal diversity, multichannel algorithms appear to be decisive for current and future ALDs. Moreover, in contrast to single-microphone signal enhancement algorithms, which have not been shown to improve speech intelligibility but may reduce, e.g., the listening effort, multimicrophone signal enhancement algorithms are capable of increasing speech intelligibility [1], especially when the sound sources have different spatial characteristics.

Although microphone array signal processing, e.g., for teleconferencing systems, is a well-established field dealing with similar problems and signals [2], the problem setting for ALDs exhibits a number of distinctive features. First, the microphone placement is typically constrained by the fact that the devices should be inconspicuously placed at the user's head and should capture the relevant spatial information of the sound sources. Moreover, while all signal enhancement algorithms ideally aim to remove the undesired components and leave the desired components undistorted, the compromises need to be chosen differently depending on the application domain: for ALDs, distortion of the desired signal or annoying noise artifacts will typically be penalized more than a higher level of residual undistorted noise, and the balance between reduced listener fatigue, increased speech intelligibility, and subjective quality plays an even greater role than in other speech communication devices. Finally, for binaural systems that are expected to dominate the future markets, preservation of the critical binaural cues as necessary for a correct spatial perception is crucial [3], not just for the desired signal, but also for the residual noise and interferers.

SCOPE

In this article, we will discuss several algorithms for multimicrophone signal enhancement and presentation that are suitable for ALDs. The considered acoustic scenario is defined by a single source of interest (target source) at any point in time, while multiple interfering point sources (e.g., competing speakers) and additional incoherent noise (e.g., sensor noise, diffuse background noise) may be active simultaneously (see Figure 2). It is assumed that some knowledge is available to distinguish the target source from the interfering sources once they are sufficiently enhanced or separated. Bearing in mind that the wearers of ALDs may move their heads, the relative positions of both the target source as well as the interfering sources must be considered as time-varying, so that source localization and tracking is required.



[FIG2] A scenario with the target source $s_0(t)$, point-like interferers $s_p(t)$, incoherent noise sources, and microphones at the user's head.

The fundamental concept of all considered multimicrophone algorithms relies on spatial and/or spectrotemporal diversity, i.e., the desired components should be separated from the undesired components in the spatial and/or time-frequency domain. The algorithms hence correspond to spatial filtering (often termed *beamforming*) and filtering in the time-frequency domain, respectively. In addition to exploiting the statistics of the available observations, the optimum filter design should also use available prior knowledge, e.g., the estimated or assumed position of the target source. This implies that, in this article, blind source separation (BSS) algorithms [4] are only considered in forms that allow the inclusion of such prior knowledge. Aside from some target-related knowledge, we assume natural unpredictable scenarios that may be arbitrarily complex and time-varying. This implies that the filters must be estimated from currently available observations and cannot be learned in advance, thus algorithms that are based on trained models (e.g., using nonnegative matrix factorization) are not considered in this article. In addition, in time-varying environments, the estimation of the spatial and spectrotemporal information from short observation intervals is of crucial importance, so we will focus on techniques exploiting second-order statistics, keeping the variance of the estimated quantities small.

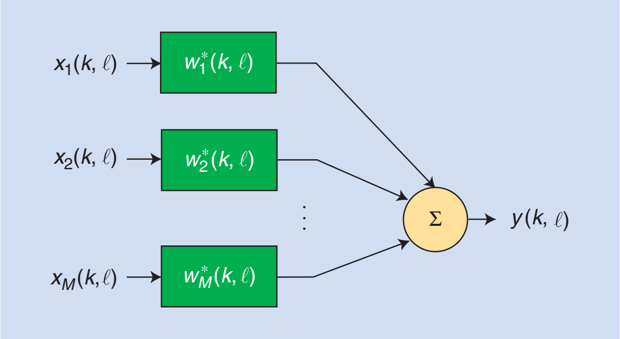
> FOR HEARING-IMPAIRED INDIVIDUALS, AS THE MOST PROMINENT USER GROUP SO FAR, FURTHER PROGRESS OF ASSISTED LISTENING TECHNOLOGY IS CRUCIAL FOR BETTER INCLUSION INTO OUR WORLD OF PERVASIVE ACOUSTIC COMMUNICATION.

SIGNAL MODEL

According to the acoustic scenario in Figure 2, we consider P point sources $s_p(t)$, with t as the discrete time index, in a noise field of unknown coherence, which are recorded by an array of M microphones. The target source is denoted by $s_0(t)$. Assuming the acoustic paths between the sources and the microphones to be linear and time-invariant, the mth microphone signal $x_m(t)$ is given by the convolutive mixing model

$$x_m(t) = \sum_{p=0}^{P-1} h_{p,m}(t) * s_p(t) + n_m(t), \; m = 1 \ldots M, \qquad (1)$$

where $n_m(t)$ denotes the noise component in the mth microphone signal, $h_{p,m}(t)$ is the room impulse response (RIR) between the pth source and the mth microphone, and $*$ denotes convolution. Typically, the signals are processed in the short-time Fourier transform (STFT) domain, i.e.,



[FIG3] The filter-and-sum structure.

$$x_m(k, \ell) = \sum_{p=0}^{P-1} h_{p,m}(k) s_p(k, \ell) + n_m(k, \ell), \; m = 1 \ldots M, \qquad (2)$$

where $x_m(k, \ell), s_p(k, \ell)$ and $n_m(k, \ell)$ denote the STFTs of the respective time-domain signals, with ℓ representing the frame index and k representing the frequency bin index, and where $h_{p,m}(k)$ denotes the acoustic transfer function (ATF) between the pth source and the mth microphone. Note that (2) is strictly speaking only valid for frames that are significantly longer than the RIR length. When this is not the case, a convolutive transfer function model should be used. For conciseness, we omit the dependency on the indices k and ℓ in the remainder of this article. In vector form, the equation set (2) can be written as

$$\mathbf{x} = \mathbf{h}_0 s_0 + \sum_{p=1}^{P-1} \mathbf{h}_p s_p + \mathbf{n} = \mathbf{h}_0 s_0 + \mathbf{v}, \qquad (3)$$

with $\mathbf{x} = [x_1 \;\cdots\; x_M]^T$, and $\mathbf{n}$ and $\mathbf{h}_p$ defined similarly, and $\mathbf{h}_0$ denoting the ATF of the target source. This signal model will form the basis for the subsequent description of the main signal processing tasks with ALDs, i.e., source localization, signal enhancement, and signal presentation.

SIGNAL ACQUISITION

For ALDs in realistic acoustic environments, the ATFs include the microphone characteristics, room acoustics, and filtering effects due to the user's head. The diffraction and reflection properties of the user's head, pinna, and torso are described by the so-called head-related transfer function (HRTF), which is the frequency- and angle-dependent transfer function between a sound source and the user's ear drum in an anechoic environment [5]. The pair of left and right HRTFs contain the so-called binaural cues of a sound source: the interaural time difference (ITD) and the interaural level difference (ILD), which are resulting from the time difference of arrival (TDOA) between both ears and the acoustic head shadow, respectively. In contrast to point sources, the spatial characteristics of incoherent noise can not be properly described by the ITD and ILD, but rather by the interaural coherence (IC) [5]. Binaural cues play a major role in spatial awareness, i.e., for source localization and for determining the spaciousness of auditory objects, and are important for speech intelligibility due to binaural unmasking, e.g., [5].

For capturing the relevant spatial information and binaural cues of the sound sources, in principle, at least two microphones are required, which are preferably mounted on both sides of the head. Ideally, the microphones are placed as close as possible to the corresponding loudspeakers that present the signals to the ear drums to allow the recreation of the authentic spatial impression for the listener. In typical ALDs today, two or three microphones are available on each side of the head, with spacings ranging from 7 mm to 15 mm. Since the positions of the microphones do not coincide with the ear drum, and the acoustic path between the loudspeaker and the ear drum differs from the HRTF, the overall response of the device should be equalized to match the open-ear HRTF [3].

SOURCE LOCALIZATION

The objective of source localization is to estimate the position or the direction of arrival (DOA) of the target source (and possibly the interfering sources), be it for supporting signal extraction or for furnishing signal presentation algorithms with spatial information.

SIGNAL EXTRACTION

The main task is to extract from the given recordings an undistorted version of the target source while all undesired components are suppressed. Two generic approaches can be used to achieve this:

- One can aim at separating all point sources and then pick the target source based on additional knowledge.
- One can directly use the additional knowledge to extract the target source only.

Intuitively, the second approach promises a lower overall algorithmic complexity for a desired performance, as it essentially requires only to separate the target source from all other sources, and obviously avoids the complexity of estimating the potentially large number of irrelevant sources in a given acoustic scene. In addition, the first approach may be limited to setups where the number of microphones is larger than the number of point sources.

Signal extraction is typically achieved using a filter-and-sum structure, depicted in Figure 3, where each microphone signal x_m is passed through a linear filter w_m^* and the outputs are summed. The output signal y is then given in the STFT domain by

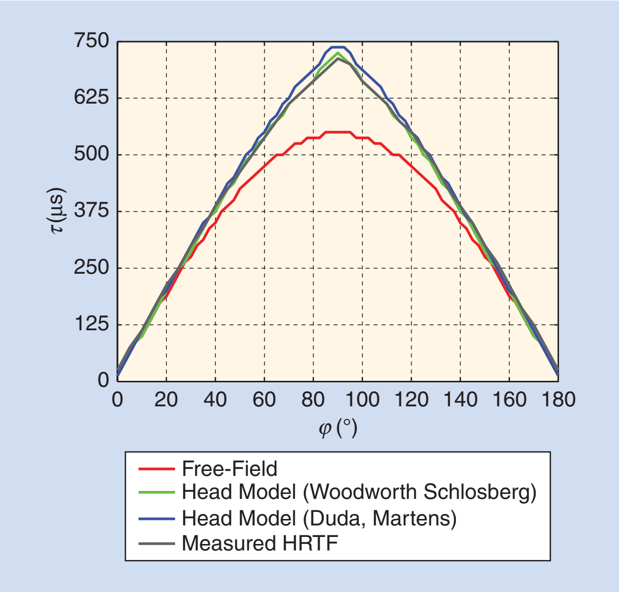
$$y = \sum_{m=1}^{M} w_m^* x_m = \mathbf{w}^H \mathbf{x} \qquad (4)$$

with $\mathbf{w}^H = [w_1^* \;\; w_2^* \;\; \dots \;\; w_M^*]$. The time-domain output signal may then be computed using the inverse STFT.

While, in principle, additional knowledge may describe source characteristics in both the time-frequency domain or the spatial domain, in this article we will mainly consider additional knowledge in the spatial domain, assuming that the sources are physically located at different positions. Typical prior spatial knowledge is then given by, e.g., the estimated or assumed DOA of the target source relative to the head. With this spatial information, we can support signal extraction algorithms, e.g., a beamformer pointing toward a given DOA or BSS algorithm exploiting the target DOA. These algorithms will be covered in more detail in the sections "Data-Independent Beamforming" and "Statistically Optimum Signal Extraction."

SIGNAL PRESENTATION

After extracting the target source, the enhanced signal is to be presented to the listener, where we need to distinguish between monaural and binaural systems. For a monaural ALD, i.e., a single device on one ear, it seems obvious to just feed the enhanced signal to the loudspeaker of this device. For a binaural ALD, i.e., a system jointly considering and processing the microphone signals of both ears, different signals can be presented to the left and the right ear. This can generate an important binaural advantage since the auditory system can exploit binaural cues and the signal processing algorithms can use information from all microphones on both devices [6, ch. 14]. On the other hand, in a bilateral system where both devices work independently, this potential is not fully exploited since not all microphone signals from both devices are combined. To exploit the full potential of binaural processing, both devices need to cooperate with each other and exchange information or signals, e.g., through a wireless link.



[FIG4] TDOAs for different azimuthal directions θ (0° = front, 180° = back) based on free-field assumption, measured HRTFs and two head models, respectively.

Besides signal extraction, a second major task should be achieved in binaural ALDs: the auditory impression of the acoustic scene, i.e., the spatial perception of the target source, the residual interfering sources and noise, should be preserved. This can be achieved either by so-called binaural rendering of the monaural output signal of the signal extraction algorithm, or by directly incorporating the desired binaural cues into the spatial filter design. These algorithms will be covered in more detail in the section "Presentation of the Enhanced Signals."

SOURCE LOCALIZATION

In principle, any source localization algorithm that can handle multiple nonstationary wideband sources can be used for ALDs [6, ch. 6]. This includes direct methods based on steered-response power (SRP) [2, ch. 8] or subspace methods [Multiple Signal Classification (MUSIC)] [7] and the large and popular class of indirect two-step methods based on TDOA estimation and a subsequent geometric inference of the source position. The latter class comprises cross-correlation-based [8] and cross-relation-based algorithms, e.g., [9] and [10].

The main difference of using these algorithms for ALDs compared to their conventional use results from the fact that the microphones are typically mounted close to the user's head. Therefore, the propagation paths of a point source to the different microphones can not be simply modeled by the free-field TDOA, but the filtering effects of the head should be taken into account. As HRTFs vary between individuals, the results produced by source localization algorithms will always suffer from some uncertainty if the individual HRTFs and the microphone topology are not exactly known. This is especially true for binaural systems, where the relative microphone positions are user dependent and not fixed. However, useful approximations can be employed, which are, e.g., based on spherical head models [11] or measured HRTFs. The TDOAs for different source directions based on the free-field assumption, measured HRTFs, and typical head models is depicted in Figure 4. Alternatively, for binaural systems, computational auditory scene analysis (CASA) algorithms [12] can be used for localizing multiple sources, e.g., incorporating a probabilistic model of the binaural ILD and ITD cues [13].

THE FUNDAMENTAL CONCEPT OF ALL CONSIDERED MULTIMICROPHONE ALGORITHMS RELIES ON SPATIAL AND/OR SPECTROTEMPORAL DIVERSITY.

Given the microphone topology, cross-correlation-based algorithms such as the generalized cross-correlation with phase transform (GCC-PHAT) [8] can be used to localize a single source for ALDs when the head filtering effects are taken into account. However, when multiple sound sources are present, identifying the correct source-specific TDOAs typically becomes very difficult [14]. Generalizations of the GCC, such as SRP-PHAT [2, ch. 8], coherently add up signals originating from a certain point in space to estimate the source likelihood at this position. While conceptually suited for an arbitrary number of microphones and sources, they involve considerable computational complexity for sufficient spatial resolution and are inherently sensitive to reverberation.

More general cross-relation-based algorithms, e.g., [9] and [10], aim at system identification via cross-relation and are naturally suited for identifying relative head-related impulse responses (HRIRs) from the source to the different microphones, delivering TDOA information as long as the direct path can be detected in the identified relative impulse responses. While the adaptive eigenvalue decomposition method in [9] is able to identify relative HRIRs only for a single source while exploiting nonstationarity, the BSS-based method in [10] can robustly localize multiple sources even in noisy and moderately reverberant environments.

Finally, subspace-based source localization algorithms such as MUSIC [7] are in principle also suitable for arbitrary numbers of microphones and sources (assuming the number of sources is known). As they essentially estimate the source positions using the eigenvectors corresponding to the largest eigenvalues of a spatial covariance matrix, the estimates for this covariance matrix must be sufficiently reliable for every frequency bin. Since subspace-based algorithms are separating the signal and noise subspace, where the noise needs to be white or whitened, this is typically difficult to achieve for wideband nonstationary sources in time-varying environments where only short observation intervals can be considered.

DATA-INDEPENDENT BEAMFORMING

A simple but popular way for enhancing the target source in ALDs is data-independent beamforming, where the filters $\mathbf{w}$ in (4) are designed to enhance sources arriving from the (estimated or assumed) target DOA and suppress sources not arriving from this DOA, but do not account for the statistics of the microphone signals. Various data-independent beamformers include delay-and-sum beamformers and superdirective or differential beamformers [2, ch. 2], [15]. For the design of such beamformers, the target DOA and the complete microphone topology need to be known. Data-independent beamformers have mainly been used for monaural devices [16], where robustness against microphone mismatch is crucial due to the closely spaced microphones [17], [18]. For binaural devices, data-independent beamformers have also been proposed, which, however, suffer from spatial aliasing due to the distance between the microphones and require consideration of the head filtering effects, e.g., [19].

STATISTICALLY OPTIMUM SIGNAL EXTRACTION

In contrast to data-independent beamformers, data-dependent signal enhancement methods exploit both the spectrotemporal as well as the spatial information of the microphone signals to extract the target source s_0 (or a filtered version of it) from all interferers and noise [20], possibly equalizing the reverberation effect caused by the ATFs' $\mathbf{h}_0$. Since the filters adapt to the current statistics of the typically nonstationary signals, this will be treated as an optimum multichannel filtering problem in the sequel.

Relying on estimates of either the interference and noise statistics or the target source statistics, two main classes of supervised optimum multichannel filtering will be discussed in the sections "Minimum Variance Distortionless Response Beamformer" and "Multichannel Wiener Filtering." In addition, BSS algorithms, in particular the variants exploiting target-related prior information for constraining the optimization problem to explicitly separate the target source, will be considered in the section "Blind Source Separation." Techniques for estimating the required second-order statistics will be presented in the section "Estimation of Interference and Noise Statistics."

MINIMUM VARIANCE DISTORTIONLESS RESPONSE BEAMFORMER

The minimum variance distortionless response (MVDR) beamformer is a special case of a linearly constrained minimum variance (LCMV) beamformer [20], [21], where the power of the output signal is minimized subject to a single constraint assuring an undistorted response for the target source (or a filtered version of it). Different versions of the MVDR beamformer exist, either using the complete target ATF, the direct path of the ATF, or the relative transfer functions (RTFs). In practice, the MVDR beamformer is often implemented using a so-called generalized sidelobe canceler (GSC) structure [22]–[25].

DERIVATION OF THE MVDR BEAMFORMER

The power spectral density (PSD) of the filter-and-sum beamformer output signal y is given by

$$E\{|y|^2\} = E\{\mathbf{w}^H \mathbf{x}\mathbf{x}^H \mathbf{w}\} = \mathbf{w}^H \boldsymbol{\Phi}_{\mathrm{xx}} \mathbf{w}, \qquad (5)$$

where $\boldsymbol{\Phi}_{\mathrm{xx}} \triangleq E\{\mathbf{x}\mathbf{x}^H\}$ denotes the crosspower spectral density matrix of the observed microphone signals. The distortionless response constraint requires that the desired component in the output signal y_{s_0} is equal to the target signal s_0, i.e.,

$$y_{s_0} = \mathbf{w}^H \mathbf{h}_0 s_0 \overset{!}{=} s_0. \qquad (6)$$

Hence, by solving the constrained minimization problem

$$\min_{\mathbf{w}} \mathbf{w}^H \boldsymbol{\Phi}_{\mathrm{xx}} \mathbf{w}, \text{ subject to } \mathbf{w}^H \mathbf{h}_0 = 1, \qquad (7)$$

we obtain the MVDR filter [20], [21]

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathrm{xx}}^{-1} \mathbf{h}_0}{\mathbf{h}_0^H \boldsymbol{\Phi}_{\mathrm{xx}}^{-1} \mathbf{h}_0}. \qquad (8)$$

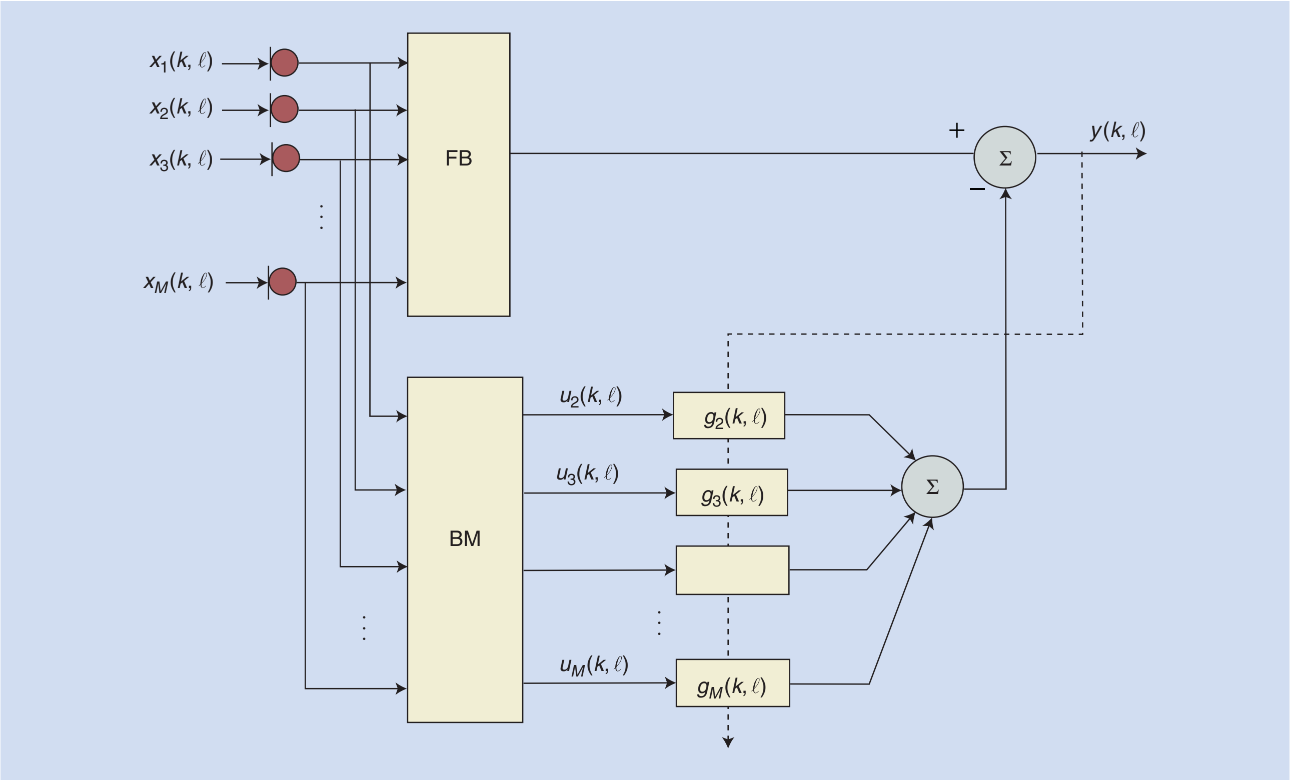
By assuming the target source, the interfering sources and the noise to be mutually uncorrelated and of zero mean, the crosspower spectral density matrix $\boldsymbol{\Phi}_{\mathrm{xx}}$ can be written using (3) as

$$\boldsymbol{\Phi}_{\mathrm{xx}} = \phi_{s_0 s_0} \mathbf{h}_0 \mathbf{h}_0^H + \boldsymbol{\Phi}_{\mathrm{vv}}, \qquad (9)$$

where $\boldsymbol{\Phi}_{\mathrm{vv}} \triangleq E\{\mathbf{v}\mathbf{v}^H\}$ denotes the crosspower spectral density matrix of the interference and noise components and $\phi_{s_0 s_0} = E\{|s_0|^2\}$. Using (9), it can be shown that the MVDR filter in (8) can be written as [20]

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathrm{vv}}^{-1} \mathbf{h}_0}{\mathbf{h}_0^H \boldsymbol{\Phi}_{\mathrm{vv}}^{-1} \mathbf{h}_0}. \qquad (10)$$

As can be seen, the MVDR filter is solely determined by the crosspower spectral density matrix of the observations and the ATFs $\mathbf{h}_0$. However, due to the high order and the typically time-varying nature of the corresponding RIRs $h_{0,m}(t)$, blindly identifying these impulse responses is generally difficult if at all possible. Hence, instead of using the complete RIRs, one can consider only the direct path of the RIRs (corresponding to the free-field HRIR for the estimated or assumed target DOA), which may, however, lead to target signal distortion, or one can use the so-called RTFs.



[FIG5] The GSC implementation of an MVDR beamformer

MVDR USING RTFs

By constraining the desired component in the output signal to be equal to the speech component at an arbitrarily chosen reference microphone r [24], the constraint in (6) becomes

$$y_{s_0} = \mathbf{w}^H \mathbf{h}_0 s_0 \overset{!}{=} h_{0,r} s_0, \tag{11}$$

which is equivalent to $\mathbf{w}^H \tilde{\mathbf{h}}_0 = 1$, where the RTF $\tilde{\mathbf{h}}_0$ is defined as

$$\tilde{\mathbf{h}}_0 \triangleq \frac{\mathbf{h}_0}{h_{0,r}} = \left[\frac{h_{0,1}}{h_{0,r}} \;\; \frac{h_{0,2}}{h_{0,r}} \;\; \cdots \;\; 1 \;\; \cdots \frac{h_{0,M}}{h_{0,r}}\right]^T. \tag{12}$$

By substituting the ATFs $\mathbf{h}_0$ with the RTFs $\tilde{\mathbf{h}}_0$ in (8) and (10), the modifed MVDR filter is obtained as

$$\tilde{\mathbf{w}}_{\mathrm{MVDR}} = h_{0,r}^{*} \mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{\Phi}_{\mathrm{xx}}^{-1} \tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^H \mathbf{\Phi}_{\mathrm{xx}}^{-1} \tilde{\mathbf{h}}_0} = \frac{\mathbf{\Phi}_{\mathrm{vv}}^{-1} \tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^H \mathbf{\Phi}_{\mathrm{vv}}^{-1} \tilde{\mathbf{h}}_0}. \tag{13}$$

Note that blind identification of RTFs is significantly easier than blind identification of ATFs. When noise and interference are absent, this can simply be achieved by dividing the crosspower spectral densities of the microphone signals. When noise and/or interference are present, methods exploiting the nonstationarity of speech or based on the generalized eigenvalue decomposition have been proposed, e.g., [24] and [25].

GSC

The constrained optimization problem of the MVDR beamformer in (7) can be transformed into an unconstrained optimization problem, leading to the highly popular GSC structure [22]–[25], consisting of three main blocks (see Figure 5): 1) a fixed beamformer (FB), ensuring the fulfillment of the constraint in (6) or (11), 2) a blocking matrix (BM), creating so-called noise references u_m, and 3) a multichannel interference canceler g_m, minimizing the residual interference and noise in the output of the FB that is correlated with the noise references. If the target signal leaks into the noise references due to a mismatched BM (e.g., caused by RTF estimation errors or by DOA errors, microphone mismatch, and reverberation when using free-field HRIRs), the target signal will be partially canceled as well. To mitigate this target signal cancellation, the interference canceler is typically adapted only during periods when the target source is inactive; see, e.g., [23]. Moreover, several techniques have been proposed to reduce the speech leakage components in the noise references, e.g., [24], [25], and/or limit the distorting effect of the remaining speech leakage [23], [26], [27], e.g., by imposing a quadratic inequality constraint or by using the so-called speech-distortion-regularized GSC [27].

APPLICATION IN ALDs

The GSC or one of its more robust variants can be considered as the current state-of-the-art solution for monaural hearing devices with an end-fire microphone array configuration, e.g., [28]–[30]. A very popular variant is the adaptive directional microphone (ADM) [15], [28], [29], where the fixed beamformer and the BM are differential beamformers forming a front- and back-oriented cardioid pattern, and an adaptive scalar minimizes the energy arriving from the back hemisphere. A two-microphone implementation was indeed shown to achieve a considerable speech intelligibility improvement for hearing aid users (about 3.4 dB improvement for three babble noise sources) [29].

MULTICHANNEL WIENER FILTER

The second popular class of multichannel signal enhancement techniques is associated with the multichannel Wiener filter (MWF), e.g., [2, ch. 3, 6, 14], [27], [31]. It produces a minimum mean square error (MMSE) estimate of either the target source [2, ch. 3], the speech component at an arbitrarily chosen microphone [2, ch. 6,14], [31], or a reference speech signal [2, ch. 14], [27]. To trade off speech distortion and noise reduction, the so-called speech-distortion-weighted MWF was introduced [27], [31].

Similarly to the MVDR using RTFs, the MWF neither requires a priori information about the microphone configuration nor the position of the target source, making it an appealing approach from a robustness point of view. On the other hand, relying on the second-order statistics of the desired and undesired signal components implies that, for the assumed nonstationary processes, these statistics must be estimated with sufficient accuracy at all times; cf. the section "Estimation of Interference and Noise Statistics."

MMSE ESTIMATION FOR THE MWF

The MWF aims to extract the target source by minimizing the mean square error (MSE) between the (unknown) source signal s_0 and the beamformer output, i.e.,

$$\mathbf{w}_{\mathrm{MWF}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|s_0 - y|^2\} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|s_0 - \mathbf{w}^H \mathbf{x}|^2\}. \tag{14}$$

Assuming the target source and the interfering sources and noise to be uncorrelated, the solution of (14) is given by

$$\mathbf{w}_{\mathrm{MWF}} = \mathbf{\Phi}_{\mathrm{xx}}^{-1} E\{\mathbf{x} s_0^{*}\} = \mathbf{\Phi}_{\mathrm{xx}}^{-1} \mathbf{h}_0 \phi_{s_0 s_0}, \tag{15}$$

requiring the ATFs $\mathbf{h}_0$ and the target source PSD $\phi_{s_0 s_0}$ to be estimated, which is a nontrivial task. However, similarly to the MVDR using RTFs, we can also design an MWF aiming at extracting the speech component at an arbitrarily chosen reference microphone r by

$$\tilde{\mathbf{w}}_{\mathrm{MWF}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|h_{0,r} s_0 - \mathbf{w}^H \mathbf{x}|^2\}, \tag{16}$$

which yields

$$\tilde{\mathbf{w}}_{\mathrm{MWF}} = \mathbf{\Phi}_{\mathrm{xx}}^{-1} E\{\mathbf{x} h_{0,r}^{*} s_0^{*}\} = \left(\phi_{s_0 s_0} \mathbf{h}_0 \mathbf{h}_0^H + \mathbf{\Phi}_{\mathrm{vv}}\right)^{-1} \phi_{s_0 s_0} \mathbf{h}_0 h_{0,r}^{*}. \tag{17}$$

Although it appears that the ATFs and the target source PSD are required to compute (17), the (rank-1) crosspower spectral density matrix $\phi_{s_0 s_0}\mathbf{h}_0\mathbf{h}_0^H$ can be estimated from the second-order statistics of the microphone signals; cf. the section "Estimation of Interference and Noise Statistics."

SPEECH-DISTORTION-WEIGHTED MWF

The MMSE criterion in (16) can be easily generalized to allow for a tradeoff between noise reduction and speech distortion [27], [31] by introducing a weighting factor $\mu \in [0, \infty]$:

$$\tilde{\mathbf{w}}_{\text{SDW}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|h_{0,r}s_0 - \mathbf{w}^H\mathbf{h}_0 s_0|^2\} + \mu E\{|\mathbf{w}^H\mathbf{v}|^2\}, \quad (18)$$

which is referred to as the speech-distortion-weighted MWF (SDW-MWF). The solution of (18) is given by

$$\tilde{\mathbf{w}}_{\text{SDW}} = (\phi_{s_0 s_0}\mathbf{h}_0\mathbf{h}_0^H + \mu\boldsymbol{\Phi}_{\mathbf{vv}})^{-1}\phi_{s_0 s_0}\mathbf{h}_0 h_{0,r}^*. \quad (19)$$

The smaller the factor μ is chosen, the smaller the resulting speech distortion. If $\mu = 1$, the MMSE criterion (16) is obtained. If $\mu > 1$, the residual noise level will be reduced at the expense of increased speech distortion.

RELATIONSHIP BETWEEN MWF AND MVDR

It is interesting to note that the MWF can be decomposed as an MVDR beamformer, exploiting the spatial information of the target and interfering sources, followed by a single-channel Wiener filter (SWF) [2, ch. 3], [32], i.e.,

$$\tilde{\mathbf{w}}_{\text{SDW}} = \underbrace{\frac{\phi_{y_s y_s}}{\phi_{y_s y_s} + \mu\phi_{y_v y_v}}}_{\text{SDW}-\text{SWF postfilter}} \times \underbrace{\frac{\boldsymbol{\Phi}_{\mathbf{vv}}^{-1}\tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^H\boldsymbol{\Phi}_{\mathbf{vv}}^{-1}\tilde{\mathbf{h}}_0}}_{\text{MVDR beamformer}}, \quad (20)$$

where $\phi_{y_s y_s}$ and $\phi_{y_v y_v}$ denote the PSDs of the desired and undesired components at the output of the MVDR beamformer $\tilde{\mathbf{w}}_{\text{MVDR}}$ using RTFs.

APPLICATION IN ALDs

In [1], a three-microphone MWF implementation for a monaural hearing device was evaluated at different test sites and compared with other single- and multimicrophone noise reduction techniques. In this study it was shown that overall the MWF achieved the largest speech intelligibility improvements (up to 7 dB), even in highly reverberant environments.

BLIND SOURCE SEPARATION

Generalizing the approach of extracting a single desired source, BSS algorithms aim at extracting multiple sources from observed mixtures without requiring prior knowledge on the positions of the sources and the microphones, spatiotemporal signal statistics, or the mixing system. Moreover, they do not need any reference information on the activity of the sources in the spectrotemporal domain. On the other hand, they do require knowledge on the total number of sources and can only separate sources that can be modeled as point sources. Considering time-varying mixing systems, we disregard approaches that perform BSS based on learning from a large amount of data and focus on independent component analysis (ICA)-based methods that are—similar to adaptive filtering approaches—suited to time-varying acoustic scenes [4], [33]–[35].

For the following, we rewrite the STFT signal model in (3) as

$$\mathbf{x} = \sum_{p=0}^{P-1}\mathbf{h}_p s_p + \mathbf{n} = \mathbf{Hs} + \mathbf{n}, \quad (21)$$

describing M noisy observations $\mathbf{x}$ of the convolutive mixture of P point sources s_p. To obtain estimates of the original sources s_p, a linear demixing/separation system $\mathbf{W}$ is applied, consisting of $M \times P$ filters with frequency response w_{mp}, $m = 0, \ldots, M-1$, $p = 0, \ldots, P-1$. The P separated signals y_q, stacked in the vector $\mathbf{y}$, are then obtained as

$$\mathbf{y} = \mathbf{W}^H\mathbf{x} = \mathbf{W}^H\mathbf{Hs} + \mathbf{W}^H\mathbf{n}. \quad (22)$$

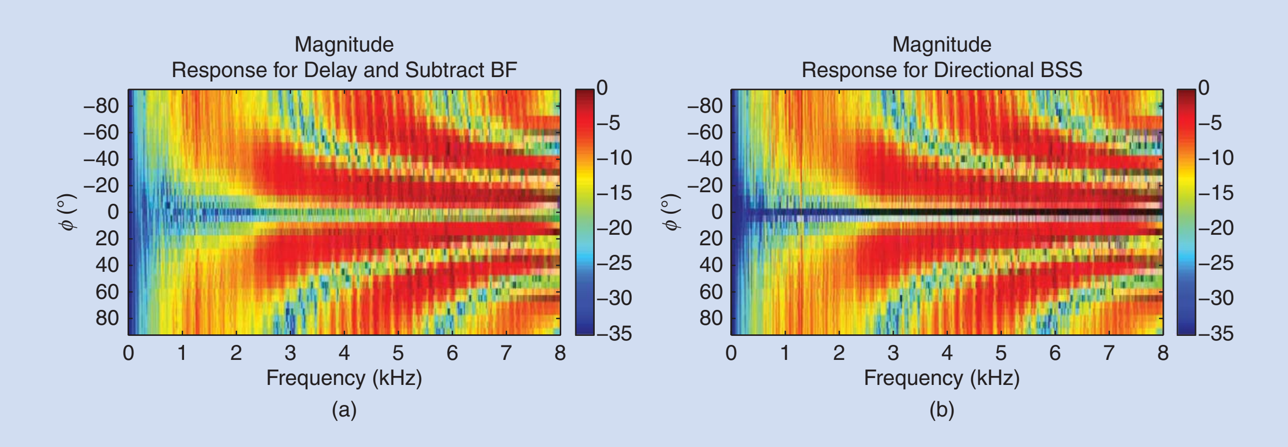
Known methods for identifying optimum demixing filters $\mathbf{W}$ are based on the assumption that the signals to be separated are mutually statistically independent and that enforcing statistically independent outputs y_q of the demixing system yields good estimates of the desired separated source signals s_p. For the mostly assumed case where the number of microphones is larger than or equal to the number of sources $(M \geq P)$, an appropriate generic cost function $\mathcal{J}(\ell)$ for frame ℓ, describing an estimate of the Kullback–Leibler divergence between the joint probability density function (pdf) of the output signals y_q and the desired independent outputs, can be formulated as [4, ch. 4]:

$$\mathcal{J}_{\text{ICA}}(\ell) = \sum_{\lambda=0}^{\infty}\beta(\lambda,\ell)\frac{1}{K}\sum_{\kappa=0}^{K-1}\log\frac{\hat{p}_{\mathbf{y},PL}(\mathbf{y}(\kappa,\lambda))}{\prod_{q=0}^{P-1}\hat{p}_{y_q,L}(y_q(\kappa,\lambda))}, \quad (23)$$

where $\hat{p}_{y_q,L}(y_q(\kappa,\lambda))$ denotes an estimate for the L-variate pdf of a segment of length L of the qth output signal y_q, and $\hat{p}_{\mathbf{y},PL}(\mathbf{y}(\kappa,\lambda))$ denotes an estimate for the PL-variate joint pdf for all P output signals. Averaging over K frames accounts for the nonstationarity of the data, while the windowing function $\beta(\ell,\lambda)$ describes the weight of a block average at time λ for the cost function at time ℓ, in a similar way as for recursive least squares adaptation. Forming gradients of this cost function, or simplified versions, with respect to the demixing matrix $\mathbf{W}$ allows for maximization of statistical independency with respect to individual data frames (online adaptation, $K = 1, \beta(\lambda,\ell) = 0$ for $\lambda \neq \ell$), as well as for an entire recording (offline adaptation, $K > 1, \beta = \text{constant}$) [35].

It should be noted that using the statistical independence assumption only, the separation system $\mathbf{W}$ can at best be obtained up to a linear filtering uncertainty and a permutation of the outputs, and thus cannot itself identify the inverse mixing system which would solve the deconvolution problem and perfectly dereverberate the source signals [36].

Numerous algorithms have been proposed for ICA of convolutive mixtures, which are often categorized as either time-domain or frequency-domain algorithms. Time-domain algorithms estimate the demixing system $\mathbf{W}$ as finite impulse response filters [35], whereas frequency-domain algorithms



[FIG6] Interference cancelation in a reverberant environment obtained by (a) null-steering beamformer and (b) ICA ($T_{60} = 300$ ms, interfering point source at 0° at a distance of 1.1 m of a two-microphone array with spacing $d = 15$ cm).

formulate the demixing problem as a scalar source separation problem for each frequency independently (instantaneous ICA), and implement scalar ICA algorithms for each STFT bin [33], [36], [37].

Similar to adaptive filtering, where time-domain approaches imply a significantly higher computational complexity for obtaining a similar performance as frequency-domain approaches, frequency-domain implementations of ICA (FD-ICA) are computationally more attractive. On the other hand, if these are straightforwardly formulated as independent ICA problems in each STFT bin, the resulting demixing system does not perform a linear but a circular convolution, which is inadequate for demixing a linear mixing system [35]. As an immediate consequence, the so-called internal permutation and scaling problems result: as the outputs of any unconstrained ICA system are only determined up to an unknown scaling factor and a permutation of their order, for FD-ICA the order and the scaling of the outputs may be different for each STFT bin. Therefore the outputs of the scalar ICA units have to be realigned so that for a given output channel y_q all frequency bins belong to the same source [37] and are properly scaled, e.g., by minimizing the average power difference of the outputs y_q relative to the inputs x_m (minimum distortion principle) [38].

In the acoustic signal extraction context, the mechanism of BSS based on ICA has been shown to be equivalent to a set of P adaptive beamformers, each of which aims to extract one source by suppressing all other sources, thereby exploiting the spatial diversity of the microphone signals [39]. Note that for adaptive beamforming, the DOA or the RTFs of the target source should be known, and that it can adapt the required statistics only with given source activity information, while ICA does not need such information.

APPLICATION IN ALDs

To illustrate the spatial filtering capacity of ICA, Figure 6 depicts the overall transfer function $\mathbf{W}^H\mathbf{H}$ from a given source position in a reverberant environment for a null-steering (delay-and-subtract) beamformer and one output channel of an ICA system, thereby demonstrating the actual interference suppression performance in a reverberant environment [40]. The improved spatial null achieved by ICA confirms the hypothesis that, due to capturing all correlated components belonging to the same source in the same output, ICA does not only suppress the direct path but also reflections of an interfering source, e.g., [40]. Nevertheless, one has to bear in mind that the suppression of reflections results from a compromise in the spatial directivity, which a null-steering beamformer cannot offer. Obviously, using the same number of microphones, ICA cannot use more spatial degrees of freedom than a supervised beamformer, and therefore the spatial selectivity of ICA remains limited to what an optimum and ideally controlled beamformer can achieve, as long as it uses the same statistics for determining its parameters [39].

The fact that ICA does not require prior knowledge about source positions, microphone topology, and source activity, and can adapt well during the activity of multiple sources, renders it a highly attractive method for ALDs in complex acoustic environments with unpredictable interference and noise, and usually unknown source and microphone topologies. Unfortunately, however, ICA systems that can robustly and quickly separate more than three sources in real-world environments have not been presented yet, so that scenarios with an unknown number of interferers cannot be handled by such a generic ICA system.

ESTIMATION OF INTERFERENCE AND NOISE STATISTICS

The performance of the signal extraction algorithms discussed in the sections "MVDR Beamformer" and "Multichannel Wiener Filter" critically depends on the estimates of the statistics of the desired and the undesired signal components, respectively. When implementing these algorithms, it is typically assumed that there is a domain where either the desired or the undesired components can be observed alone. While in selected cases, stationarity assumptions may hold reliably to justify a predetermined estimate [41], it must usually be assumed that the statistics of both the desired and the undesired components vary in an unpredictable way and call for instantaneous estimates.

In the spectrotemporal domain, voice activity detection and speech presence probability estimation typically aim at identifying regions in the STFT domain where only undesired components are present, e.g., [6, ch. 5], [42]. Obviously, this is very difficult for the given scenario with interfering speech sources that naturally occupy the same frequency range and whose temporal activity pattern is generally not known, especially if their signal level is comparable to the level of the target source in any of the microphone signals. Therefore, the desired and undesired components can usually only be separated along the time axis. For example, for computing the MWF according to (19), it is typically assumed that the interference and noise can be observed during noise-only periods, so that with the assumed uncorrelatedness of noise and desired speech, the crosspower spectral density matrix $\phi_{s_0s_0}\mathbf{h}_0\mathbf{h}_0^H$ can be estimated as

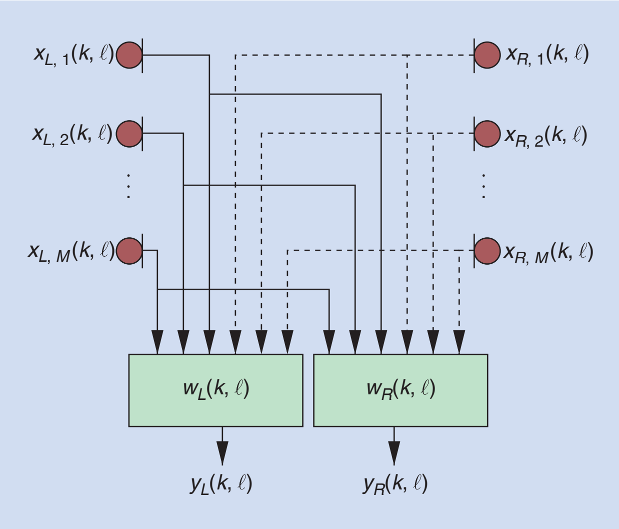
$$\phi_{s_0s_0}\mathbf{h}_0\mathbf{h}_0^H \approx \mathbf{\Phi}_{\mathrm{xx}} - \mathbf{\Phi}_{\mathrm{vv}}, \qquad (24)$$

where $\mathbf{\Phi}_{\mathrm{xx}}$ is estimated continually and $\mathbf{\Phi}_{\mathrm{vv}}$ during periods of interference and noise only. As a fundamental problem, however, all these methods still suffer from the fact that the interference and noise estimates cannot be updated while the target source is active, so that they are prone to failure with nonstationary noise and interference, such as human speakers.

On the other hand, in the spatial domain, reference information for all the interference and noise components can be obtained by suppressing the target source. Here, the spatial selectivity allowed by the microphone array topology constitutes the main limitation. Exploiting the spatial domain for obtaining interference and noise reference information is an inherent feature of the GSC (cf. the section "MVDR Beamformer"), where the BM aims to suppress the target source. For moving sources and multipath propagation scenarios, robust adaptation schemes for the BM have already been proposed, e.g., [23]. These concepts still require knowledge about the activity of the target source, as the BM should only be adapted when the target source is dominant. If the DOA of the target source is known, its activity can be monitored by directing both a delay-and-sum beamformer and a delay-and-subtract beamformer in this direction and inferring the activity from the ratio of its output powers, see, e.g., [23]. However, these noise estimates will still be suboptimal if the BM could not be updated while the target source changed its position relative to the microphones on the user's head or the acoustic environment changed.

More recently, a constrained BSS scheme has been proposed to identify the filters of two-channel blocking matrices [40], which does not need source activity information and continuously delivers up-to-date estimates for noise and interference. For this, the cost function in (23) is complemented by a quadratic constraint for one output (here y_p) steering a null toward the target source:

$$\mathcal{J}_{\mathrm{C}}(\mathbf{W}) = \| \mathbf{w}_p^H \mathbf{d} \|_2^2, \qquad (25)$$



[FIG7] **The general binaural processing scheme.**

where $\mathbf{w}_p$ denotes the vector of demixing filters in $\mathbf{W}$ which produce the output y_p, and $\mathbf{d}$ denotes the steering vector corresponding to the DOA of the direct path of the target source. This yields the constrained ICA cost function

$$\mathcal{J}_{\mathrm{C-ICA}}(\mathbf{W}) = \mathcal{J}_{\mathrm{ICA}}(\mathbf{W}) + \eta \mathcal{J}_{\mathrm{C}}(\mathbf{W}), \qquad (26)$$

whose minimization suppresses the target in one output channel and thereby provides a reference for all other sources and noise of unknown coherence. The weight is typically chosen as $\eta \approx 0.5 \ldots 0.8$ with larger values required if interfering sources are close to the target source. It should be noted that, although the constraint captures only the direct path, constrained ICA will intrinsically also aim at suppressing all correlated components, i.e., reflections of the target source, in the same output, thereby providing an advantage over a delay-and-subtract beamformer as shown in Figure 6. As the most attractive advantage, however, the fundamental concept of ICA assures a continuous update of the noise estimate without the need of estimating the activity of the involved sources. Recently, it was also shown that this concept can be generalized to identify all RTFs required for the BM of a GSC with an arbitrary number of constraints [43].

PRESENTATION OF THE ENHANCED SIGNALS

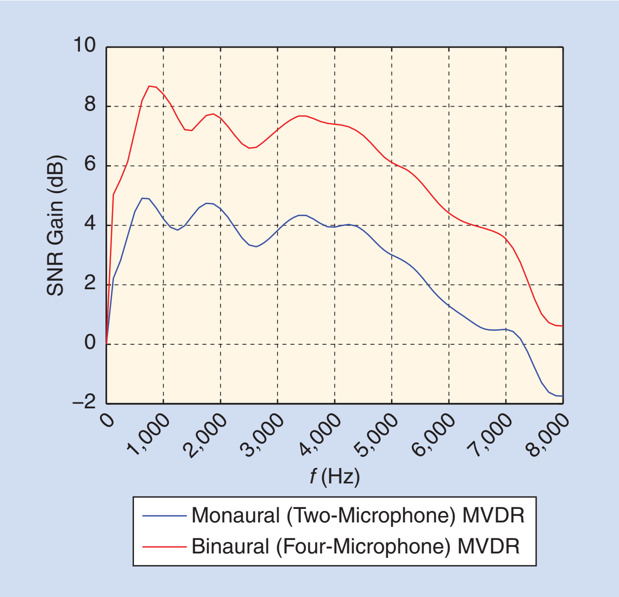
After extracting the target source using data-independent beamforming or statistically optimum filtering (cf. the sections "Data-Independent Beamforming" and "Statistically Optimum Signal Extraction"), the enhanced signal needs to be presented to the listener. While microphone placement is important to maintain a close relationship to the individual HRTFs, we also need to distinguish between a monaural system, i.e., a single device on one ear, and a binaural system, i.e., a system jointly processing signals, at both ears. While for a monaural system it seems obvious to just feed the enhanced signal to the loudspeaker of this device, for a binaural system different output signals y_L and y_R can be generated and presented to the left and the right ear (cf. Figure 7).

In a bilateral system, i.e., a set of two independently operating monaural systems, each device uses its own microphone signals and optimizes its filter coefficients independently, which may lead to a distortion of the binaural cues and hence the localization ability [44]. To achieve true binaural processing, both devices need to cooperate with each other and exchange information or signals, e.g., through a wireless link. Currently, the first commercial systems reach the market which exchange microphone signals in full-duplex mode. These systems pave the way to future implementations of fully fledged binaural multimicrophone signal extraction algorithms, where microphone signals from both devices are processed and combined in each device. The gain in noise reduction performance of a binaural over a monaural system is exemplarily shown for an MVDR beamformer in Figure 8.

TO EXPLOIT THE FULL POTENTIAL OF BINAURAL PROCESSING, BOTH DEVICES NEED TO COOPERATE WITH EACH OTHER AND EXCHANGE INFORMATION OR SIGNALS.

The objective of a binaural speech enhancement algorithm is not only to selectively extract the target source and to suppress interfering sources and background noise, but also to preserve the auditory perception of the complete acoustic scene. This can be achieved by preserving the binaural cues, i.e., ITD, ILD, and IC, of the target source and the residual interfering sources and background noise. In addition to monaural cues, these binaural cues play a major role in spatial awareness and localization and are very important for speech intelligibility due to binaural unmasking, e.g., [5].

All discussed signal enhancement algorithms in the sections “Data-Independent Beamforming” and “Statistically Optimum Signal Extraction” essentially generate a single-channel output signal. Since in a binaural system two output signals (i.e., one for each ear) are required, this single-channel output signal can either be binauralized, e.g., using binaural spectral postfiltering techniques [19], [45], [46] or by mixing the output signal with scaled (noisy) microphone signals [47], [48], or two different complex-valued spatial filters can be optimized, where the desired binaural cues are directly incorporated into the spatial filter design, e.g., [48]–[50]. Although the latter paradigm allows for more degrees of freedom to achieve noise reduction, there is typically a tradeoff between noise reduction performance and binaural cue preservation.



[FIG8] The SNR gain of a monaural and a binaural MVDR beamformer (diffuse noise field).

In binaural spectral postfiltering techniques, the same real-valued gain is applied to one microphone signal of each device, where a gain close to one is applied when the STFT bin should be retained (target source), and a gain close to zero is applied when the STFT bin should be suppressed (interfering source or background noise). This spectral gain can, e.g., be computed by comparing the estimated binaural cues with the expected cues of the target source or based on the temporal fluctuations of the ITD [45]. Other commonly used approaches compute the spectral gain based on the output signal of a data-independent or statistically optimum spatial filter (e.g., MVDR beamformer, BSS) [19], [46]. Although binaural spectral postfiltering techniques preserve the binaural cues of all sound sources, in essence, they can be viewed as single-channel noise reduction techniques, hence typically introducing speech distortion and exhibiting single-channel noise reduction artifacts (e.g., musical noise), especially at low input SNRs.

The MVDR beamformer (using RTFs) and the MWF can be straightforwardly extended into a binaural version producing two output signals, by estimating the speech component in two reference microphone signals, i.e., one on each hearing aid [48]. In [48] and [44], it was shown both analytically and using subjective listening experiments that the binaural MWF preserves the binaural cues of the target source but distorts the binaural cues of interferers and noise, such that all components are perceived as coming from the direction of the target source. Clearly, this is undesired and, in some situations (e.g., traffic), even dangerous. To optimally benefit from binaural unmasking and to optimize the spatial awareness of the hearing aid user, several extensions for the binaural MWF and the MVDR beamformer have been proposed, which aim at also preserving the binaural cues of the residual noise component by including cue preservation terms in the binaural cost function, e.g., [48]–[50]. These include either RTF preservation or interference rejection constraints for directional interfering sources [48], [49], or IC preservation constraints for diffuse noise [50]. Another approach is partial noise estimation, which corresponds to mixing the binaural outputs with scaled versions of the noisy reference microphone signals [48].

APPLICATION IN ALDs

In [30] and [44], the performance of the binaural MWF and some of its extensions has been perceptually evaluated, both in terms of speech intelligibility and localization performance. First, it was shown that the binaural MWF achieved significant speech intelligibility improvements compared to the bilateral MWF and the bilateral ADM. This demonstrates that transmitting and processing microphone signals from both devices can result in a significant gain in noise reduction, especially when multiple interfering sources are present. Second, using a localization experiment in the frontal horizontal hemisphere, it was shown that using the binaural MWF with partial noise estimation it is possible to preserve spatial awareness without significantly affecting speech intelligibility.

SUMMARY AND OUTLOOK

In this article, we have presented an overview of several multimicrophone signal enhancement algorithms for ALDs and have addressed other important issues, such as microphone placement and binaural signal presentation. Using appropriate processing with multiple microphones in a binaural ALD allows both speech intelligibility improvement as well as a preservation of the auditory perception of the acoustic scene.

Future work in this area will focus both on algorithmic aspects and a better integration of psychoacoustics. On the algorithmic side, more accurate and robust estimation and careful exploitation of comprehensive spatiotemporal signal statistics for all relevant sources in highly time-varying scenarios will be necessary to allow for the ultimate desired binaural presentation. The learning of acoustic scenarios and source characteristics can certainly be expected to contribute to reaching this goal. Optimum distribution of the computational load over the available computing hardware via bit rate-constrained "body area networks" will constitute another challenge to algorithm developers. On the psychoacoustic side, ideally, meaningful criteria are desirable that can directly be integrated into the cost functions to allow perceptually optimum signal processing at any given time instant. This may start from incorporating general knowledge about well-known noise masking effects combined with knowledge on the relative importance of certain binaural cues as used already in audio coding and reach to more powerful, yet unknown models for human hearing. For each individual, it should be merged with knowledge about possible hearing impairments or personal listening preferences, i.e., a so-called auditory consumer profile. One may speculate that with suitable user interfaces, the traditional fitting procedures will be replaced by training procedures supervised by the user and even the cost functions for optimizing the multichannel filtering will be as individual as the users themselves. All of these developments will certainly benefit from the integration into handy, but powerful personal computing platforms that are already emerging.

[ALL OF THESE DEVELOPMENTS WILL CERTAINLY BENEFIT FROM THE INTEGRATION INTO HANDY, BUT POWERFUL PERSONAL COMPUTING PLATFORMS THAT ARE ALREADY EMERGING.]

ACKNOWLEDGMENTS

This research was supported in part by the Research Unit FOR 1732 "Individualized Hearing Acoustics" and the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG), and by the European Union under contract FP6004171 (HEARCOM).

AUTHORS

Simon Doclo (simon.doclo@uni-oldenburg.de) received the Ph.D. degree in applied sciences from Katholieke Universiteit Leuven, Belgium, in 2003. Since 2009, he has been a full professor at the University of Oldenburg, Germany, and a scientific advisor for the Fraunhofer Institute for Digital Media Technology. His research activities center on signal processing for acoustical and biomedical applications. He received the EURASIP Signal Processing Best Paper Award in 2003 and the IEEE Signal Processing Society (SPS) Best Paper Award in 2008. He was member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and is an associate editor of the EURASIP *Journal on Advances in Signal Processing*.

Walter Kellermann (wk@LNT.de) received the Dipl.-Ing. degree in electrical engineering in 1983 and the Dr.-Ing. degree in 1988. From 1989 to 1990, he was a postdoctoral member of technical staff at AT&T Bell Laboratories. He has been a professor of communications at the University of Erlangen-Nuremberg, Germany, since 1999. He (co)authored 16 book chapters and more than 200 refereed papers and was corecipient of several best paper awards for IEEE publications. He served as an associate editor and as a guest editor to various journals and was a general chair for several international conferences. His service to the IEEE Signal Processing Society (SPS) includes Distinguished Lecturer, chair of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing, and member-at-large for the SPS Board of Governors. He is an IEEE Fellow.

Shoji Makino (maki@tara.tsukuba.ac.jp) received the Ph.D. degree from Tohoku University, Japan, in 1993. Since 2009, he has been a professor at the University of Tsukuba. His research interests include acoustic signal processing for speech and audio applications. He received the ICA Unsupervised Learning Pioneer Award and the IEEE Machine Learning for Signal Processing Competition Award. He has served on the IEEE Signal Processing Society (SPS) Technical Directions Board, Awards Board, and Conference Board. He was associate editor of *IEEE Transactions on Speech and Audio Processing*. He is the chair of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing. He is an IEEE Distinguished Lecturer and an IEEE Fellow.

Sven Nordholm (s.nordholm@curtin.edu.au) received the Ph.D. degree in signal processing from Lund University, Sweden, in 1992. Since 1999, he has been a full professor at Curtin University, Perth, Australia, and an advisor for Hearmore and Sensear. His research activities focus on signal processing and communication in acoustic media. He founded Sensear in 2006 and is also its technology inventor. He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and is an associate editor of *Journal of the Franklin Institute*.

REFERENCES

[1] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 2054–2063, Mar. 2010.

[2] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001.

[3] J. Blauert, *The Technology of Binaural Listening*. New York: Springer, 2013.

[4] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. New York: Springer, 2007.

[5] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localisation*. Cambridge, MA: MIT, 1997.

[6] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*. Hoboken, NJ: Wiley, 2008.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, pp. 276–280, Mar. 1986.

[8] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[9] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[10] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[11] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 3048–3058, 1998.

[12] D. L. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York: IEEE Press/Wiley-Interscience, 2007.

[13] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[14] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 8, pp. 1479–1489, 2008.

[15] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Commun.*, vol. 20, no. 3–4, pp. 229–240, 1996.

[16] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 3138–3148, May 1996.

[17] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[18] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 77–80.

[19] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Processing*, vol. 2006, no. 1, pp. 175–175, Jan. 2006.

[20] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[21] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 7, pp. 1408–1418, Aug. 1969.

[22] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[23] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[24] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non-stationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[25] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 1, pp. 206–218, Jan. 2011.

[26] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[27] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[28] F. L. Luo, J. Y. Yang, C. Pavlovic, and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1583–1590, 2002.

[29] J. B. Maj, L. Royackers, M. Moonen, and J. Wouters, "Comparison of adaptive noise reduction algorithms in dual microphone hearing aids," *Speech Commun.*, vol. 48, no. 8, pp. 957–960, Aug. 2006.

[30] B. Cornelis, M. Moonen, and J. Wouters, "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel Wiener filtering based noise reduction," *J. Acoust. Soc. Amer.*, vol. 131, no. 6, pp. 4743–4755, June 2012.

[31] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7–8, pp. 636–656, July–Aug. 2007.

[32] L. W. Brooks and I. S. Reed, "Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter, and the Wiener filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 8, no. 5, pp. 690–692, 1972.

[33] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[34] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[35] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[36] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, nos. 1–3, pp. 21–34, Nov. 1998.

[37] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[38] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Independent Component Analysis and Signal Separation*, Dec. 2001, pp. 722–727.

[39] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1157–1166, Nov. 2003.

[40] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP J. Appl. Signal Processing*, vol. 26, 2014. Doi:10.1186/1687-6180-2014-26. [Online]. Available: http://asp.eurasipjournals.com/content/2014/1/26

[41] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: An analytical evaluation," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 241–252, May 1999.

[42] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 5, pp. 910–919, July 2008.

[43] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2013.

[44] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 484–497, July 2008.

[45] G. Grimm, V. Hohmann, and B. Kollmeier, "Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-feedback noise reduction scheme," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 17, no. 7, pp. 1408–1419, Sept. 2009.

[46] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Comput. Speech Lang. (CSL)*, vol. 27, no. 3, pp. 726–745, May 2012.

[47] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-array hearing aids with binaural output–Part II: A two-microphone adaptive system," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[48] B. Cornelis, S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural multi-microphone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[49] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012, pp. 117–120.

[50] D. Marquardt, V. Hohmann, and S. Doclo, "Perceptually motivated coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Florence, Italy, May 2014, pp. 3688–3692.

[Konrad Kowalczyk, Oliver Thiergart, Maja Taseska, Giovanni Del Galdo, Ville Pulkki, and Emanuël A.P. Habets]

Parametric Spatial Sound Processing

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

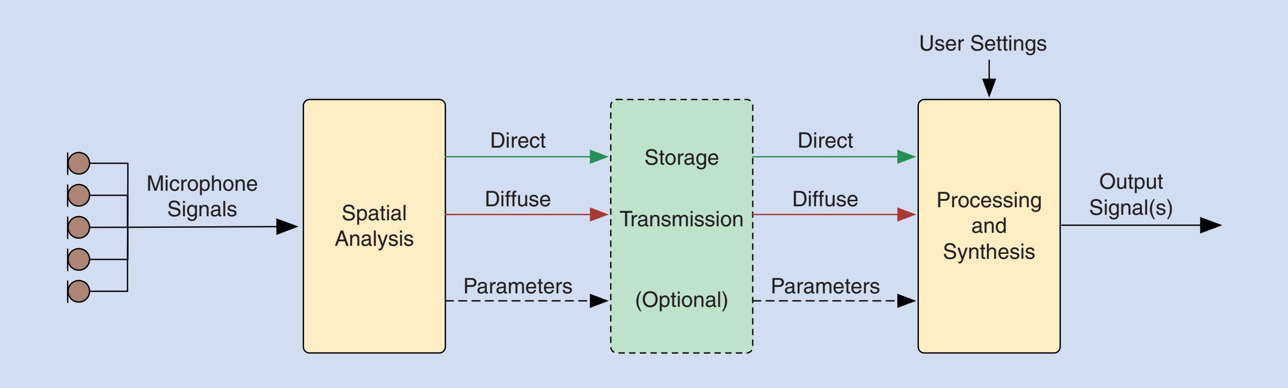
[A flexible and efficient solution to sound scene acquisition, modification, and reproduction]

Flexible and efficient spatial sound acquisition and subsequent processing are of paramount importance in communication and assisted listening devices such as mobile phones, hearing aids, smart TVs, and emerging wearable devices (e.g., smart watches and glasses). In application scenarios where the number of sound sources quickly varies, sources move, and nonstationary noise and reverberation are commonly encountered, it remains a challenge to capture sounds in such a way that they can be reproduced with a high and invariable sound quality. In addition, the objective in terms of what needs to be captured, and how it should be reproduced, depends on the application and on the user's preferences. Parametric spatial sound processing has been around for two decades and provides a flexible and efficient solution to capture, code, and transmit, as well as manipulate and reproduce spatial sounds.

Instrumental to this type of processing is a parametric model that can describe a sound field in a compact and general way. In most cases, the sound field can be decomposed into a direct sound component and a diffuse sound component. These two components together with parametric side information such as the direction-of-arrival (DOA) of the direct sound component or the position of the sound source, provide a perceptually motivated description of the acoustic scene [1]–[3]. In this article, we provide an overview of recent advances in spatial sound capturing, manipulation, and reproduction based on such parametric descriptions of the sound field. In particular, we focus on two established parametric descriptions presented in a unified way and show how the signals and parameters can be obtained using multiple microphones. Once the sound field is analyzed, the sound scene can be transmitted, manipulated, and synthesized depending on the application. For example,

Digital Object Identifier 10.1109/MSP.2014.2369531
Date of publication: 12 February 2015

1053-5888/15©2015IEEE



[FIG1] A high-level overview of the parametric spatial sound processing scheme.

sounds can be extracted from a specific direction or from a specific arbitrary two-dimensional or even three-dimensional region of interest. Furthermore, the sound scene can be manipulated to create an acoustic zoom effect in which direct sounds within the listening angular range are amplified depending on the zoom factor, while other sounds are suppressed. In addition, the signals and parameters can be used to create surround sound signals. As the manipulation and synthesis are highly application dependent, we focus in this article on three illustrative assisted listening applications: spatial audio communication, virtual classroom, and binaural hearing aids.

INTRODUCTION

Communication and assisted listening devices commonly use multiple microphones to create one or more signals, the content of which highly depends on the application. For example, when smart glasses are used to record a video, the microphones can be used to create a surround sound recording that consists of multiple audio signals. A compact yet accurate representation of the sound field at the recording position makes it possible to render the sound field on an arbitrary reproduction setup in a different location. On the other hand, when the device is used in hands-free or speech recognition mode, the microphones can be used to extract the user's speech while reducing background noise and interfering sounds. In the last few decades, sophisticated solutions for these applications were developed.

Spatial recordings are commonly made using specific microphone setups. For instance, there are several stereo recording techniques in which different positioning of the microphones of the same or different types (e.g., cardioid or omnidirectional microphones) is exploited to make a stereo recording that can be reproduced using loudspeakers. When more loudspeakers are available for spatial sound rendering, the microphone recordings are often specifically mixed for a given reproduction setup. These classical techniques do not provide the flexibility required in many modern applications where the reproduction setup is not known in advance. Signal enhancement, on the other hand, is commonly achieved by filtering, and subsequently summing the available microphone signals. Classical spatial filters often require information on the second-order statistics (SOS) of the desired and undesired signals (cf. [4] and [5]). For real-time applications, the SOS need to be estimated online, and the quality of the output signal highly depends on the accuracy of these estimates. To date, major challenges remain, such as:

1) achieving a sufficiently fast response to changes in the sound scene (such as moving and emerging sources) and to changes in the acoustic conditions
2) providing sufficient flexibility in terms of spatial selectivity
3) ensuring a high-quality output signal at all times
4) providing solutions with a manageable computational complexity.

Although the use of multiple microphones provides, at least in theory, a major advantage over a single microphone, the adoption of multimicrophone techniques in practical systems has not been particularly popular until very recently. Possible reasons for this could be that in real-life scenarios, these techniques provided insufficient improvement over single-microphone techniques, while significantly increasing the computational complexity, the system calibration effort, and the manufacturing costs. In the last few years, the smartphone and hearing aid industries made a significant step forward in using multiple microphones, which has recently become a standard for these devices.

Parametric spatial sound processing provides a unified solution to both the spatial recording and signal enhancement problems, as well as to other challenging sound processing tasks such as adding virtual sound sources to the sound scene. As illustrated in Figure 1, the parametric processing is performed in two successive steps that can be completed on the same device or on different devices. In the first step, the sound field is analyzed in narrow frequency bands using multiple microphones to obtain a compact and perceptually meaningful description of the sound field in terms of direct and diffuse sound components and some parametric information (e.g., DOAs and positions). In the second step, the input signals and possibly the parameters are modified, and one or more output signals are synthesized. The modification and synthesis can be user, application, or scenario dependent. Parametric spatial sound processing is also common in audio coding (cf. [6]) where parametric information is extracted directly from the loudspeaker channels instead of the microphone signals.

The described scheme also allows for an efficient transmission of sound scenes to the far-end side [1], [7] for loudspeaker reproduction with arbitrary setups or for binaural reproduction [8]. Hence, instead of transmitting many microphone signals and carrying out the entire processing at the receiving side, only two signals (i.e., the direct and diffuse signals) need to be transmitted together with the parametric information. These two signals enable synthesis of the output signals on the receiving side for the reproduction system at hand, and additionally allow the listener to arbitrarily adjust the spatial responses. Note that in the considered approach, the same audio and parametric side information is sent, irrespective of the number of loudspeakers used for reproduction.

As an alternative to the classical filters used for signal enhancement, where an enhanced signal is created as a weighted sum of the available microphone signals, an enhanced signal can be created by using the direct and diffuse sound components and the parametric information. This approach can be seen as a generalization of the parametric filters used in [9]–[12] where the filters are calculated based on instantaneous estimates of an underlying parametric sound field model. As will be discussed later in this article, these parameters are typically estimated in narrow frequency bands, and their accuracy depends on the resolution of the time-frequency transform and the geometry of the microphone array. If accurate parameter estimates with a sufficiently high time-frequency resolution are available, parametric filters can quickly adapt to changes in the acoustic scene. The parametric filters have been applied to various challenging acoustic signal processing problems related to assisted listening, such as directional filtering [10], dereverberation [11], and acoustic zooming [13]. Parametric filtering approaches have been used also in the context of binaural hearing aids [14], [15].

PARAMETRIC SOUND FIELD MODELS

BACKGROUND

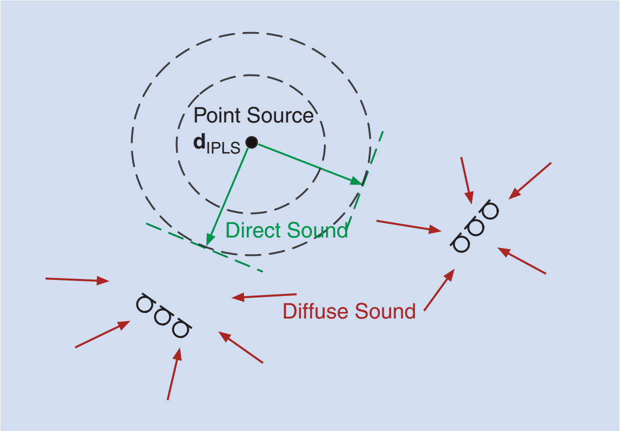
Many parametric models have originally been developed with the aim to subsequently capture, transmit, and reproduce high-quality spatial audio; examples include directional audio coding (DirAC) [1], microphone front ends for spatial audio coders [16], and high angular resolution plane wave expansion (HARPEX) [17]. These models were developed based on observations about the human perception of spatial sound, aiming to recreate perceptually important spatial audio attributes for the listener. For example, in the basic form of DirAC [1], the model parameters are the DOA of the direct sound and the diffuseness that is directly related to the power ratio between the direct signal power and the diffuse signal power. Using a pressure signal and this parametric information, a direct signal and a diffuse signal could be reconstructed at the far-end side. The direct signal is attributed to a single plane wave at each frequency, and the diffuse signal is attributed to spatially extended sound sources, concurrent sound sources (e.g., applause from an audience or cafeteria noise), and room reverberation that occurs due to multipath acoustic wave propagation when sound is captured in an enclosed environment. A similar sound field model that consists of the direct and diffuse sound has been applied in spatial audio scene coding (SASC) [2] and in [3] for sound reproduction with arbitrary reproduction systems and for sound scene manipulations. On the other hand, in [16] the model parameters include the interchannel level difference and the interchannel coherence [18] that were estimated using two microphones and were previously used in various spatial audio coders [6]. These model parameters are sent to the far-end side together with a so-called downmix signal to generate multiple loudspeaker channels for sound reproduction. In this case, the downmix signal and parameters are compatible with those used in different spatial audio coders. In contrast to DirAC and SASC, HARPEX assumes that the direct signal at a particular frequency is composed only of two plane waves.

Besides offering a compact and flexible way to transmit and reproduce high-quality spatial audio, independent of the reproduction setup, parametric processing is highly attractive for sound scene manipulations and signal enhancement. The extracted model parameters can be used to compute parametric filters that can, for instance, achieve directional filtering [10] and dereverberation [11]. The parametric filters represent spectral gains applied to a reference microphone signal, and can in principle provide arbitrary directivity patterns that can adapt quickly to the acoustic scene provided that the sound field analysis is performed with a sufficiently high time-frequency resolution. For this purpose, the short-time Fourier transform (STFT) is considered a good choice as it often offers a sufficiently sparse signal representation to assume a single dominant directional wave in each time-frequency bin. For instance, the assumption that the source spectra are sufficiently sparse is commonly made in speech signal processing [19]. The sources that exhibit sufficiently small spectrotemporal overlap fulfill the so-called W-disjoint orthogonality condition. This assumption is, however, violated when concurrent sound sources with comparable powers are active in one frequency band. Another family of parametric approaches emerged within the area of computational auditory scene analysis [20], where the auditory cues are utilized for instance to derive time-frequency masks that can be used to separate different source signals from the captured sound.

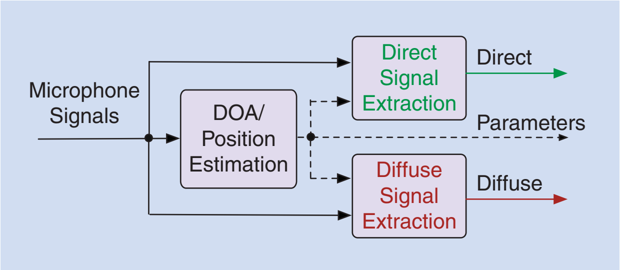
Clearly, the choice of an underlying parametric model depends on the specific application and on the way the extracted parameters and the available audio signals are used to generate the desired output. In this article, we focus on geometry-based parametric models that take into account both direct and diffuse sound components, allowing for high-quality spatial sound acquisition, which can be subsequently used both for transmission and reproduction purposes, as well as to derive flexible parametric filters for sound scene manipulation and signal enhancement for assisted listening.

GEOMETRIC MODELS

In the following, we consider the time-frequency domain with k and n denoting the frequency and time indices, respectively. For each (k, n), we assume that the sound field is a superposition of a single spherical wave and a diffuse sound field. The spherical wave models the direct sound of the point-source in a reverberant environment, while the diffuse field models room reverberation and spatially extended sound sources. As shown in Figure 2, the



[FIG2] A geometric sound field model: the direct sound emitted by a point source arrives at the array with a certain DOA, and the point-source position can be estimated when the DOA estimates from at least two arrays are available.



[FIG3] A block diagram for spatial analysis.

spherical wave is emitted by an isotropic point-like source (IPLS) located at a time-frequency-dependent position $\mathbf{d}_\mathrm{IPLS}(k,n)$. The magnitude of the pressure of the spherical wave is inversely proportional to the distance traveled, which is known in physics as the inverse distance law. The diffuse sound is assumed to be spatially isotropic and homogenous, which means that diffuse sound arrives from all directions with equal power and that its power is position independent. Finally, it is assumed that the direct sound and diffuse sound are uncorrelated.

The direct and diffuse sounds are captured with one or more microphone arrays (depending on the application) that are located in the far field of the sound sources. Therefore, at the microphone array(s), the spherical wave can be approximated by a plane wave arriving from direction $\theta(k,n)$. In the following, we will differentiate between two related geometrical models: the DOA-based model and the position-based model. In the DOA-based model, the DOA and direct sound are estimated with a single microphone array, while in the position-based model, the position of the IPLS is estimated using at least two spatially distributed arrays, and the sound is captured using one or more microphones.

Under the aforementioned assumptions, the signals received at the omnidirectional microphones of an M-element microphone array can be written as

$$\mathbf{x}(k,n) = \mathbf{x}_\mathrm{s}(k,n) + \mathbf{x}_\mathrm{d}(k,n) + \mathbf{x}_\mathrm{n}(k,n), \qquad (1)$$

where the vector $\mathbf{x}(k,n) = [X(k,n,\mathbf{d}_1),\ldots,X(k,n,\mathbf{d}_M)]^\mathrm{T}$ contains the M microphone signals in the time-frequency domain, where $\mathbf{d}_{1\ldots M}$ are the microphone positions. Without loss of generality, the first microphone located at $\mathbf{d}_1$ is used as a reference microphone. The vector $\mathbf{x}_s(k,n) = [X_s(k,n,\mathbf{d}_1),\ldots,X_s(k,n,\mathbf{d}_M)]^\mathrm{T}$ is the captured direct sound at the different microphones and $\mathbf{x}_\mathrm{d}(k,n) = [X_\mathrm{d}(k,n,\mathbf{d}_1),\ldots,X_\mathrm{d}(k,n,\mathbf{d}_M)]^\mathrm{T}$ is the captured diffuse sound. Furthermore, $\mathbf{x}_\mathrm{n}(k,n)$ contains the slowly time-varying noise signals (for example, the microphone self-noise). The direct sound at the different microphones can be related to the direct sound at the reference microphone via the array propagation vector $\mathbf{g}(k,\theta)$, which can be expressed as

$$\mathbf{x}_\mathrm{s}(k,n) = \mathbf{g}(k,\theta)X_\mathrm{s}(k,n,\mathbf{d}_1). \qquad (2)$$

The mth element of the array propagation vector $\mathbf{g}(k,\theta) = [g(k,n,\mathbf{d}_1),\ldots,g(k,n,\mathbf{d}_M)]^\mathrm{T}$ is the relative transfer function of the direct sound from the mth to the first microphone, which depends on the DOA $\theta(k,n)$ of the direct sound from the point of view of the array. For instance, for a uniform linear array of omnidirectional microphones $g(k,n,\mathbf{d}_m) = \exp\{j\kappa\|\mathbf{d}_m - \mathbf{d}_1\|\sin\theta\}$ where j denotes the imaginary unit, κ is the wavenumber, and $\|\mathbf{d}_m - \mathbf{d}_1\|$ is the distance between positions $\mathbf{d}_m$ and $\mathbf{d}_1$.

In this article, we will demonstrate how this geometric model can be effectively utilized to support a number of assisted listening applications. In the considered applications, the desired output signal of a loudspeaker (or headphone) channel $Y_i(k,n)$ is given as a weighted sum of the direct and diffuse sound at the reference microphone, i.e.,

$$\begin{aligned} Y_i(k,n) &= G_i(k,n)X_\mathrm{s}(k,n,\mathbf{d}_1) + Q_i(k)X_\mathrm{d}(k,n,\mathbf{d}_1) && (3a)\\ &= Y_{\mathrm{s},i}(k,n) + Y_{\mathrm{d},i}(k,n), && (3b)\end{aligned}$$

where i is the index of the output channel, and $G_i(k,n)$ and $Q_i(k)$ are the application-dependent weights. It is important to note that $G_i(k,n)$ depends on the DOA $\theta(k,n)$ of the direct sound or on the position $\mathbf{d}_\mathrm{IPLS}(k,n)$. To synthesize a desired output signal two steps are required: 1) extract the direct and diffuse sound components and estimate the parameters (i.e., DOAs or positions), and 2) determine the weights $G_i(k,n)$ and $Q_i(k)$ using the estimated parameters and application-specific requirements. The first step is commonly referred to as the *spatial analysis* and is discussed next. In this article, the second step is referred to as the *application-specific synthesis*.

SPATIAL ANALYSIS

To facilitate flexible sound field manipulation with high-quality audio signals, it is crucial to accurately estimate the components describing the sound field, specifically the direct and diffuse sound components, as well as the DOAs or positions. Such spatial analysis based on the microphone signals is depicted in Figure 3. The direct and diffuse sound components can be estimated using single-channel or multichannel filters. To compute these filters, we may exploit knowledge about the DOA estimate of the direct sound or compute additional parameters as discussed in the following.

SIGNAL EXTRACTION

SINGLE-CHANNEL FILTERS

A computationally efficient estimation of the direct and the diffuse components is possible using single-channel filters. Such processing is applied for instance in DirAC [1], where the direct and diffuse signals are estimated by applying a spectral gain to a single microphone signal. The direct sound is then estimated as

$$\hat{X}_{\mathrm{s}}(k,n,\mathbf{d}_1) = W_{\mathrm{s}}(k,n)X(k,n,\mathbf{d}_1), \tag{4}$$

where $W_{\mathrm{s}}(k,n)$ is a single-channel filter, which is multiplied with the reference microphone signal to obtain the direct sound at $\mathbf{d}_1$. An optimal filter $W_{\mathrm{s}}(k,n)$ can be found, for instance, by minimizing the mean-squared error between the true and estimated direct sound, which yields the well-known Wiener filter (WF). If we assume no microphone noise, the WF for extracting the direct sound is given by $W_{\mathrm{s}}(k,n) = 1 - \Psi(k,n)$. Here, $\Psi(k,n)$ is the diffuseness, which is defined as

$$\Psi(k,n) = \frac{1}{1+\mathrm{SDR}(k,n)}, \tag{5}$$

where $\mathrm{SDR}(k,n)$ is the signal-to-diffuse ratio (SDR) (power ratio of the direct sound and the diffuse sound). The diffuseness is bounded between zero and one, and describes how diffuse the sound field is at the recording position. For a purely diffuse field, the SDR is zero leading to the maximum diffuseness $\Psi(k,n) = 1$. In this case, the WF, $W_{\mathrm{s}}(k,n)$, equals zero and thus, the estimated direct sound in (4) equals zero as well. In contrast, when the direct sound is strong compared to the diffuse sound, the SDR is high and the diffuseness in (5) approaches zero. In this case, the WF $W_{\mathrm{s}}(k,n)$ approaches one and thus, the estimated direct sound in (4) is extracted as the microphone signal. The SDR or diffuseness, required to compute the WF, is estimated using multiple microphones as will be explained in the section "Parameter Estimation."

The diffuse sound $X_{\mathrm{d}}(k,n,\mathbf{d}_1)$ can be estimated in the same way as the direct sound. In this case, the optimal filter is found by minimizing the mean-squared error between the true and estimated diffuse sound. The resulting WF is given by $W_{\mathrm{d}}(k,n) = \Psi(k,n)$. Instead of using the WF, the square root of the WF is often applied to estimate the direct sound and diffuse sound (cf. [1]). In the absence of sensor noise, the total power of the estimated direct and diffuse sound components is then equal to the total power of the received direct and diffuse sound components.

In general, extracting the direct and diffuse signals with single-channel filters has several limitations:

1) Although the required SDR or diffuseness are estimated using multiple microphones (as will be discussed later), only a single microphone signal is utilized for the filtering. Hence, the available spatial information is not fully exploited.

2) The temporal resolution of single-channel filters may be insufficient in practice to accurately follow rapid changes in the sound scene. This can cause leakage of the direct sound into the estimated diffuse sound.

3) The WFs defined earlier do not guarantee a distortionless response for the estimated direct and diffuse sounds, i.e., they may alter the direct and diffuse sounds, respectively.

4) Since the noise, such as the microphone self-noise or the background noise, is typically not considered when computing the filters, it may leak into the estimated signals and deteriorate the sound quality.

Limitations 1 and 4 are demonstrated in Figure 4(a), (b), and (d), where the spectrograms of the input (reference microphone) signal and both extracted components for the noise only (before time frame 75), castanet sound (between time frame 75 and time frame 150), and speech (latter frames) are shown. The noise is clearly visible in the estimated diffuse sound and slightly visible in the estimated direct sound. Furthermore, the onsets of the castanets leak into the estimated diffuse signal, while the reverberant sound from the castanets and the speech leaks into the estimated direct signal.

MULTICHANNEL FILTERS

Many limitations of single-channel filters can be overcome by using multichannel filters. In this case, the direct and diffuse signals are estimated via a weighted sum of multiple microphone signals. The direct sound is estimated with

$$\hat{X}_{\mathrm{s}}(k,n,\mathbf{d}_1) = \mathbf{w}_{\mathrm{s}}^{\mathrm{H}}(k,n)\,\mathbf{x}(k,n), \tag{6}$$

where $\mathbf{w}_{\mathrm{s}}(k,n)$ is a complex weight vector containing the filter weights for the M microphones and $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose. A filter $\mathbf{w}_{\mathrm{s}}(k,n)$ can be found for instance by minimizing the mean-squared error between the true and estimated direct sound, similarly as in the single-channel case. Alternatively, the filter weights can be found by minimizing the diffuse sound and noise at the filter output while providing a distortionless response for the direct sound, which assures that the direct sound is not altered by the filter. This filter is referred to as the linearly constrained minimum variance (LCMV) [21] filter, which can be obtained by solving

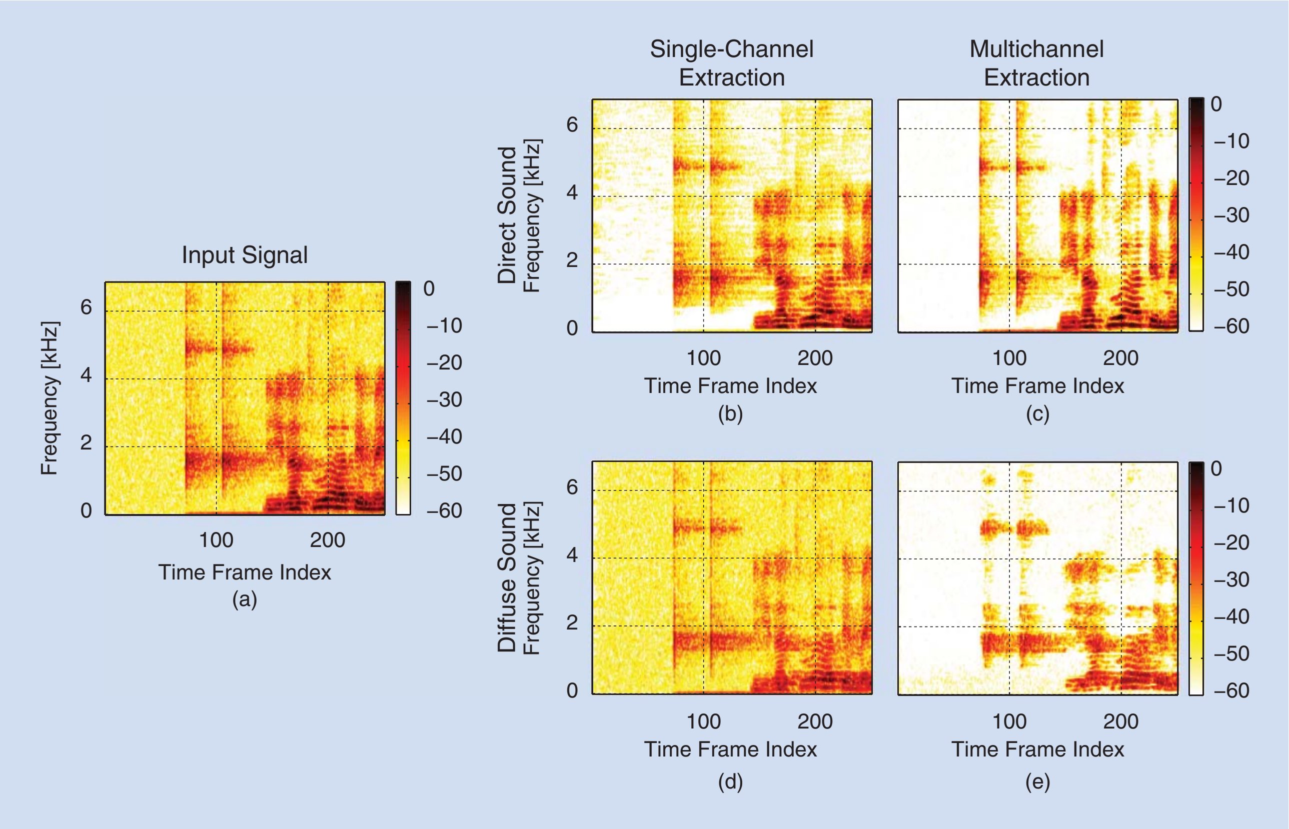
$$\begin{aligned} \mathbf{w}_{\mathrm{s}}(k,n) = {} & \underset{\mathbf{w}}{\operatorname{argmin}}\; \mathbf{w}^{\mathrm{H}}[\boldsymbol{\Phi}_{\mathrm{d}}(k,n) + \boldsymbol{\Phi}_{\mathrm{n}}(k)]\mathbf{w} \\ & \text{subject to } \mathbf{w}^{\mathrm{H}}(k,n)\,\mathbf{g}(k,\theta) = 1, \end{aligned} \tag{7}$$

where the propagation vector $\mathbf{g}(k,\theta)$ depends on the array geometry and DOA $\theta(k,n)$ of the direct sound. Here, $\boldsymbol{\Phi}_{\mathrm{d}}(k,n)$ is the power spectral density (PSD) matrix of the diffuse sound, which can be written using the aforementioned assumptions as

$$\boldsymbol{\Phi}_{\mathrm{d}}(k,n) = \mathrm{E}\{\mathbf{x}_{\mathrm{d}}(k,n)\,\mathbf{x}_{\mathrm{d}}^{\mathrm{H}}(k,n)\} \tag{8a}$$

$$= \phi_{\mathrm{d}}(k,n)\,\boldsymbol{\Gamma}_{\mathrm{d}}(k), \tag{8b}$$

where $\phi_{\mathrm{d}}(k,n)$ is the power of the diffuse sound and $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$ is the diffuse sound coherence matrix. The (m',m)th element of $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$ is the spatial coherence between the signals received at microphones m and m', which is known a priori when assuming a specific diffuse field characteristic. For instance, for a spherically isotropic diffuse field and omnidirectional microphones, the spatial coherence is a sinc function depending on



[FIG4] Spectograms of (a) the input signal, (b) the direct signal estimated using a single-channel filter, (c) the direct signal estimated using a multichannel filter, (d) the diffuse signal estimated using a single-channel filter, and (e) the diffuse signal estimated using a multichannel filter.

the microphone spacing and frequency [22]. Therefore, $\mathbf{\Phi}_{\mathrm{d}}(k,n)$ in (7) can be computed with (8b) when the diffuse sound power $\phi_{\mathrm{d}}(k,n)$ is known. The PSD matrix of the noise $\mathbf{\Phi}_{\mathrm{n}}(k)$ in (7) is commonly estimated during silence, i.e., when the sources are inactive, assuming that the noise is stationary. The estimation of $\phi_{\mathrm{d}}(k,n)$ and $\mathbf{\Phi}_{\mathrm{n}}(k)$ is explained in more detail in the next section. Note that the filter $\mathbf{w}_{\mathrm{s}}(k,n)$ is recomputed for each time-frequency bin with the geometric parameters estimated for that bin. The solution is computationally feasible since there exists a closed-form solution to the optimization problem in (7) [21].

To estimate the diffuse sound $\hat{X}_{\mathrm{d}}(k,n,\mathbf{d}_1)$, a multichannel filter that suppresses the direct sound and minimizes the noise while capturing the diffuse sound can be applied. Such a filter can be obtained by solving

$$\mathbf{w}_{\mathrm{d}}(k,n) = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}}\mathbf{\Phi}_{\mathrm{n}}(k)\mathbf{w} \text{ subject to}$$
$$\mathbf{w}^{\mathrm{H}}(k,n)\mathbf{g}(k,\theta) = 0 \text{ and } \mathbf{w}^{\mathrm{H}}(k,n)\mathbf{a}(k,n) = 1. \qquad (9)$$

The first linear constraint ensures that the direct sound is strongly suppressed by the filter. The second linear constraint ensures that we capture the diffuse sound as desired. Note that there exist different definitions for the vector $\mathbf{a}(k,n)$. In [23], $\mathbf{a}(k,n)$ corresponds to the propagation vector of a notional plane wave arriving from a direction $\theta_0(k,n)$, which is far away from the DOA $\theta(k,n)$ of the direct sound. With this definition, $\mathbf{w}_{\mathrm{d}}(k,n)$ represents a multichannel filter that captures the diffuse sound mainly from direction $\theta_0(k,n)$, while attenuating the direct sound from direction $\theta(k,n)$. In [24], $\mathbf{a}(k,n)$ corresponds to the mean relative transfer function of the diffuse sound between the array microphones. With this approach, $\mathbf{w}_{\mathrm{d}}(k,n)$ represents a multichannel filter that captures the diffuse sound from all directions except for the direction $\theta(k,n)$ from which the direct sound arrives. Note that the optimization problem (9) has a closed-form solution [21], which can be computed when the DOA $\theta(k,n)$ of the direct sound is known.

Figure 4(c) and (e) depict the spectrograms of the direct sound and diffuse sound that were extracted using the multichannel LCMV filters for the example scenario consisting of noise, castanets, and speech. As can be observed, the direct sound extracted using the multichannel filter is less noisy and contains less diffuse sound compared to the direct sound extracted using the single-channel filter. Moreover, the diffuse sound extracted using the multichannel filer contains no onsets of the direct sound (clearly visible for the onsets of the castanets in time frames 75–150) and a significantly reduced noise level. As expected, the multichannel filters provide more accurate decomposition of the sound field into a direct and a diffuse signal component. The estimation accuracy strongly influences the performance of the discussed parametric processing approaches.

PARAMETER ESTIMATION

For the computation of the filters described in the previous section, the required parameters need to be estimated. In single-channel extraction, one parameter needs to be estimated, specifically the signal-to-diffuse ratio $\mathrm{SDR}(k, n)$ or the diffuseness $\Psi(k, n)$. In the case of multichannel signal extraction, the required parameters include the DOA $\theta(k, n)$ of the direct sound, the diffuse sound power $\phi_{\mathrm{d}}(k, n)$, and the PSD matrix $\mathbf{\Phi}_{\mathrm{n}}(k)$ of slowly time-varying noise. In addition, the DOA or the position of the direct sound sources, respectively, are required to control the application-specific processing and synthesis. It should be noted that the quality of the extracted and synthesized sounds is largely influenced by the accuracy of the estimated parameters.

The estimation of the DOA of a direct sound component is a well-addressed topic in literature and different approaches for this task are available. Common approaches to estimate the DOAs in the different frequency bands are ESPRIT and root MUSIC (cf. [21] and the references therein).

For estimating the SDR, two different approaches are common in practice, depending on which microphone array geometry is used. For linear microphone arrays, the SDR is typically estimated based on the spatial coherence between the signals of two array microphones [25]. The spatial coherence is given by the normalized cross-correlation between two microphone signals in the frequency domain. When the direct sound is strong compared to the diffuse sound (i.e, the SDR is high), the microphone signals are strongly correlated (i.e., the spatial coherence is high). On the other hand, when the diffuse sound is strong compared to the direct sound (i.e., the SDR is low), the microphone signals are less correlated.

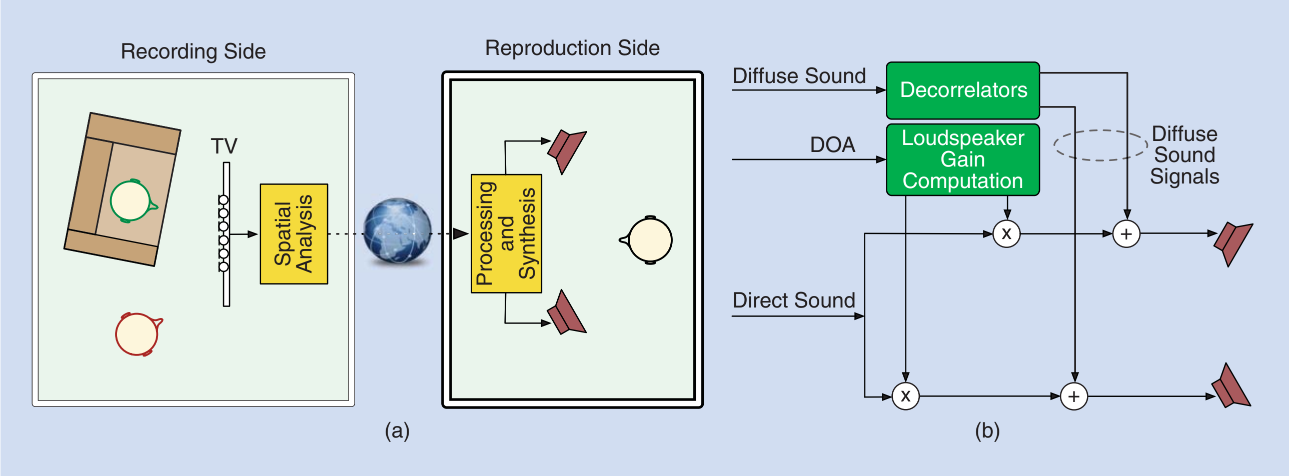
Alternatively, when a planar microphone array is used, the SDR can be estimated based on the so-called active sound intensity vector [26]. This vector points in the direction in which the acoustic energy flows. When only the direct sound arriving at the array from a specific DOA is present, the intensity vector constantly points in this direction and does not change its direction unless the sound source moves. In contrast, when the sound field is entirely diffuse, the intensity vector fluctuates quickly over time and points towards random directions as the diffuse sound is arriving from all directions. Thus, the temporal variation of the intensity vector can be used as a measure for the SDR and diffuseness, respectively [26]. Note that, as in [1], the inverse direction of the intensity vector can also be used to estimate the DOA of the direct sound. The intensity vector can be determined from an omnidirectional pressure signal and the particle velocity vector as described in [26], where the later signals can be computed from the planar microphone array as explained, for instance, in [11].

Various approaches have been described in the literature to estimate the slowly time-varying noise PSD matrix $\mathbf{\Phi}_{\mathrm{n}}(k)$. Assuming that the noise is stationary, which is a reasonable assumption in many applications (e.g., when the noise represents microphone self-noise or a stationary background noise), the noise PSD matrix can be estimated from the microphone signals during periods where only the noise is present in the microphone signals, which can be detected using a voice activity detector. To estimate the diffuse power $\phi_{\mathrm{d}}(k, n)$, we employ the spatial filter $\mathbf{w}_{\mathrm{d}}(k, n)$ in (9) that provides an estimate of the diffuse sound $X_{\mathrm{d}}(k, n, \mathbf{d}_1)$. Computing the mean power of $\hat{X}_{\mathrm{d}}(k, n, \mathbf{d}_1)$ yields an estimate of the diffuse power.

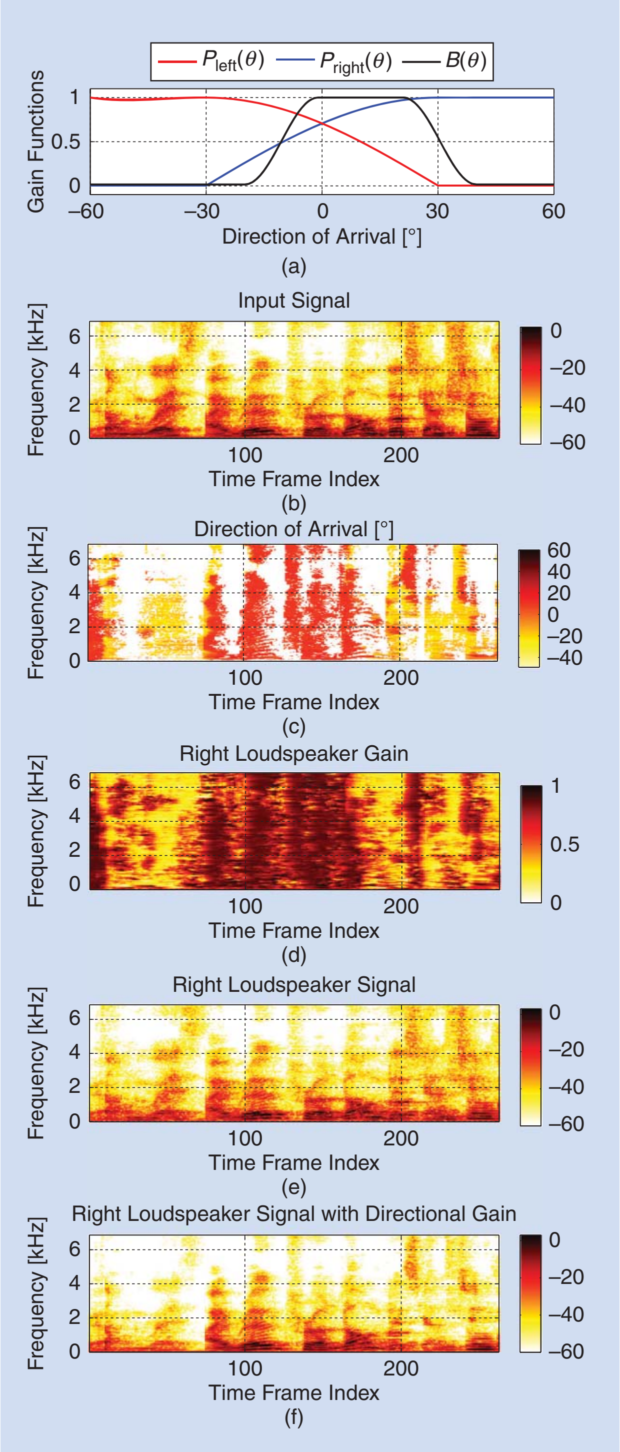
Finally, note that for some applications, such as the virtual classroom application described in the next section, the estimation of the IPLS positions from which the direct sounds originate may also be required to perform the application-specific synthesis. To determine the IPLS positions, the DOAs at different positions in the room are estimated using multiple distributed microphone arrays. The IPLS position can then be determined by triangulating the estimated DOAs, as done in [27] and illustrated in Figure 2.

APPLICATION-SPECIFIC SYNTHESIS

The compact description of the sound field in terms of a direct signal component, a diffuse signal component, and sound field parameters, as shown in Figure 1, can contribute to assisted listening in a variety of applications. While the spatial analysis yielding estimates of the model parameters and the direct and diffuse signal components at a reference microphone is application independent, the processing and synthesis is application dependent. For this



[FIG5] Spatial audio communication application: (a) communication scenario and (b) rendering of the loudspeaker signals.



[FIG6] The results in a communication scenario: (a) applied gain functions, (b) spectrogram of the input signal, (c) estimated directions of arrival, (d) gains applied to the direct sound for the right loudspeaker channel, (e) spectrogram of the right loudspeaker signal, and (f) spectrogram of the right loudspeaker signal after applying $B(\theta)$ defined in (a).

purpose, we adjust the gains $G_i(k,n)$ and $Q_i(k)$ in (3) depending on the application and as desired by the user. For spatial audio rendering, $G_i(k,n)$ and $Q_i(k)$ are used to generate the different output channels for a given reproduction setup, whereas for signal enhancement applications, $G_i(k,n)$ and $Q_i(k)$ are used to realize parametric filters that extract a signal of the desired sound source while reducing undesired and diffuse sounds. In all cases, the gains are computed using the estimated sound field parameters, and are used to obtain a weighted sum of the estimated direct and diffuse components, as given by (3). In the following, we present an overview of different applications in which the output signals are obtained using this approach.

SPATIAL AUDIO COMMUNICATION

Using spatial audio communication, we can allow participants in different locations to communicate with each other in a natural way. The sound acquisition and reproduction should provide good speech intelligibility, as well as a natural and immersive sound. Spatial cues are highly beneficial for understanding speech of a desired talker in multitalker and adverse listening situations [18]. Therefore, accurate spatial sound reproduction is expected to enable the human brain to better segregate spatially distributed sounds, which in turn could lead to better speech intelligibility. In addition, flexible spatial selectivity offered by adjusting the time-frequency dependent gains of the transmitted signals based on the geometric side information, enables the listener to focus even better on one or more talkers. These two features make the parametric methods particularly suited to immersive audio-video teleconferencing, where hands-free communication is typically desired. In hands-free communication (that is without any tethered microphones), the main challenge is to ensure the high quality of the reproduced audio signals captured from distance, and to recreate plausible spatial cues at the listeners ears. Note that for full-duplex communication, multichannel acoustic echo control would additionally be required to remove the acoustic coupling between the loudspeakers and the microphones [5]. However, the acoustic echo cancelation problem is beyond the scope of this article.

Let us consider such a teleconferencing scenario with two active talkers at the recording side, as illustrated in Figure 5. The goal is to recreate the spatial cues from the recording side at the listener side over an arbitrary, user-defined multichannel loudspeaker setup. At the recording side, one of the talkers is sitting on a couch located in front of a TV screen at a distance of 1.5 m and angle 10° with respect to array broadside direction, while the other is located to the left (at –20°) at roughly the same distance. The TV has a built-in camera and is equipped with a six-element linear array with inter-microphone spacing of 2.5 cm that captures the reverberant speech and noise (with SNR = 45 dB); the reverberation time is 350 ms. At the reproduction side, the ith loudspeaker signal is obtained as a weighted sum of the direct and diffuse signals, as given by (3). To recreate the original spatial impression of the recording side (without additional sound scene manipulation), the following gains suffice $G_i(k,n) = P_i(k,n,\theta)$ and $Q_i(k) = 1$, where $P_i(k,n,\theta)$ is the panning gain for reproducing the direct sound from the correct direction, which depends on the selected panning scheme and the loudspeaker setup. As an example, the vector-base amplitude panning (VBAP) [28] gain factors for a stereo reproduction system with loudspeakers positioned at $\pm 30^\circ$ are depicted in Figure 6(a). To reproduce the diffuse sound, the signals $Y_{\mathrm{d},i}(k,n)$ are decorrelated such that $Y_{\mathrm{d},i}(k,n)$ and $Y_{\mathrm{d},j}(k,n)$ for $i \neq j$ are uncorrelated [29]. Note that the less correlation between the loudspeaker channels, the more enveloping the perceived sound is. The described processing for synthesizing the loudspeaker signals is depicted in Figure 5(b).

When sound scene manipulation, such as directional filtering [10] and dereverberation [11], is also desired, an additional gain $B(k,n,\theta)$ can be applied to modify the direct signal. In this case, the ith loudspeaker channel gain for the direct sound can be expressed as

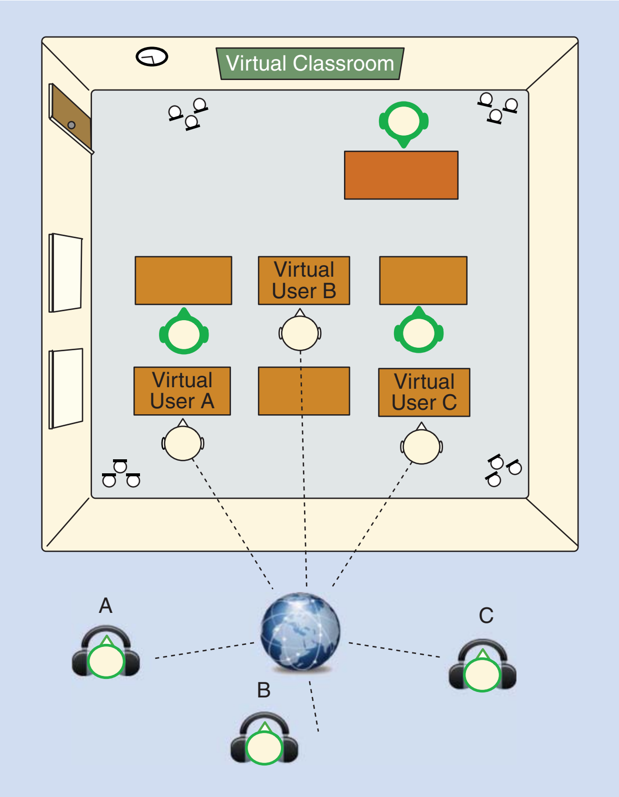
$$G_i(k,n) = P_i(k,n,\theta) B(k,n,\theta), \qquad (10)$$

where $B(k,n,\theta)$ is the desired gain for the sound arriving from $\theta(k,n)$. In principle, $B(k,n,\theta)$ can be defined freely to provide any desired directivity pattern; an example directivity gain function is shown in Figure 6(a). In addition, the diffuse sound gain $Q_i(k)$ can be adjusted to control the level of reproduced ambient sound. For instance, dereverberation is achieved by selecting $Q_i(k) < 1$.

The results for the considered teleconferencing scenario are illustrated in Figure 6. Depicted in Figure 6(a)–(c) are the gain functions, the spectrogram of an input signal, and the DOAs estimated using ESPRIT. Figure 6(d) and (e) illustrate the spatial reproduction and depict the panning gains $P_{\mathrm{right}}(k,n,\theta)$ used for the right loudspeaker and the spectrogram of the resulting signal. Lower weights can be observed when the source on the left side is active than for the source in the right, which is expected from the panning curve $P_{\mathrm{right}}(k,n,\theta)$ depicted in Figure 6(a). Note that the exact values for the respective DOAs should be $P_{\mathrm{right}} = 0.26$ for $-20°$ and $P_{\mathrm{right}} = 0.86$ for $10°$. Next we illustrate an example of sound scene manipulation. If the listener prefers to extract the signal of the talker sitting on a sofa, while reducing the other talker, a suitable gain function $B(k,n,\theta)$ can be designed to preserve the sounds coming from the sofa and attenuate sounds arriving from other directions; an example of such a gain function is shown in Figure 6(a). Additionally, setting the diffuse gain to a low value, for example $Q_i(k) = 0.25$, reduces the power level of the diffuse sound, thereby increasing the SDR during reproduction. The spectrogram of the manipulated output signal is shown in Figure 6(f), where the power of the interfering talker and reverberation are significantly reduced.

VIRTUAL CLASSROOM

The geometric model with IPLS positions as parametric information can facilitate assisted listening by creating binaural signals for any desired position in the acquired sound scene, regardless of where the microphone arrays are located. Let us consider the virtual classroom scenario in Figure 7 as an example, although the same concept also applies to other applications such as teleconference systems in dedicated rooms, assisted listening in museums, augmented reality, and many others. A teacher tutors in a typical classroom environment, where only some students are physically present, while the rest participates in the class remotely, for example, from home. As illustrated in Figure 7, the sound scene is captured using several distributed microphone arrays, with known positions. The goal is to assist a remote student to virtually participate in a class from his preferred position, for instance close to the teacher, in between the teacher and another student involved in the discussion, or at his favorite desk, by synthesizing the binaural signals for the desired virtual listener (VL) location $\mathbf{d}_{\mathrm{VL}}$. These binaural signals are generated at the reproduction side based on the received audio and position information, such that the student could listen to the synthesized sound over headphones on a laptop or any mobile device that can play multimedia content.



[FIG7] A virtual classroom scenario.

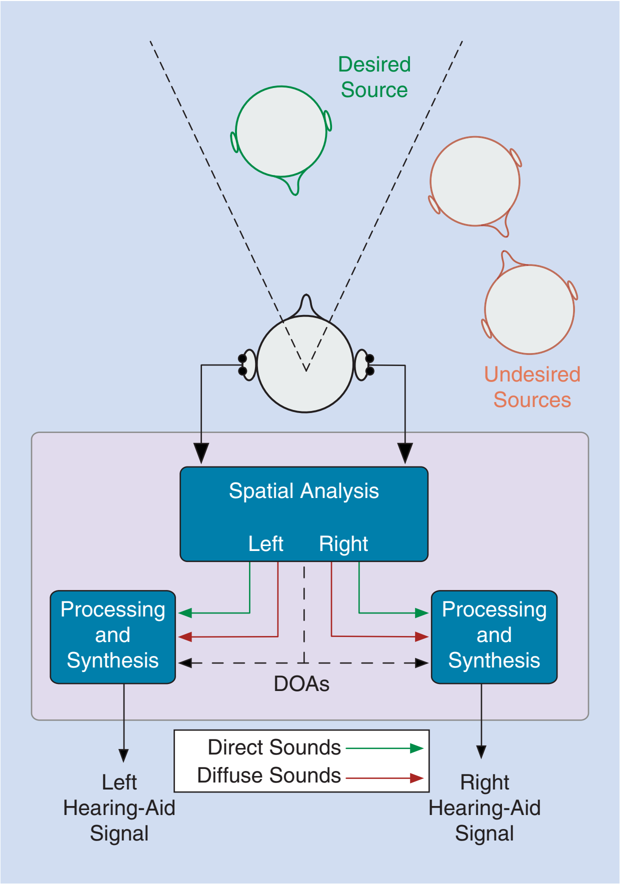
The processing to achieve this goal is in essence similar to that utilized in the virtual microphone (VM) technique [12], [27], [30], where the goal was to generate the signal of a VM that sounds perceptually similar to the signal that would be recorded with a physical microphone located at the same position. The technique has been shown successful in synthesizing the VM signals in arbitrary positions in a room [27], [30]. However, in the virtual classroom application, instead of generating the signals of nonexisting microphones with physical characteristics, we directly aim to generate the binaural signals for headphone reproduction. The overall gain for the direct sound in the ith channel can be divided into three components:

$$G_i(k,n) = D_{\mathrm{s}}(k,n) H_{\mathrm{HRTF},i}(k,n) B(k,n,\mathbf{d}_{\mathrm{IPLS}}). \qquad (11)$$

The first gain $D_{\mathrm{s}}(k,n)$ is a factor compensating for the wave propagation from $\mathbf{d}_{\mathrm{IPLS}}$ to the VL position $\mathbf{d}_{\mathrm{VL}}$, and from $\mathbf{d}_{\mathrm{IPLS}}$ to $\mathbf{d}_1$ for the direct signal estimated at the reference microphone position $\mathbf{d}_1$. As in [27], the real factors are typically applied which compensate for the amplitude change following the $1/r$ law, where r is the propagated distance. The second gain $H_{\mathrm{HRTF},i}(k,n)$ is a complex head-related transfer function (HRTF) for the left or right ear, $i \in \{\mathrm{left}, \mathrm{right}\}$, respectively, which depends on the DOA $\theta_{\mathrm{VL}}(k,n)$ with respect to the position and look direction of the VL. Apart from creating a plausible feeling of being present in the actual classroom, the user-defined spatial selectivity can be achieved with the third gain $B(k,n,\mathbf{d}_{\mathrm{IPLS}})$, which enables the amplification or attenuation of directional sounds emitted from $\mathbf{d}_{\mathrm{IPLS}}$ as desired. In principle, any desired spatial selectivity function $B(k,n,\mathbf{d})$ can be defined. For instance, a spatial spot can be defined at a teacher's desk or in front of a blackboard to assist the student in better hearing the teacher's voice. Such a gain function for a circular spot centered around $\mathbf{d}_{\mathrm{spot}}$ with a 1 m radius could be defined as

$$B(k,n,\mathbf{d}_{\mathrm{IPLS}}) = \begin{cases} 1 & r < 1; \\ \dfrac{1}{r^{\alpha}} & \text{otherwise,} \end{cases} \qquad (12)$$

where $r = \| \mathbf{d}_{\mathrm{spot}} - \mathbf{d}_{\mathrm{IPLS}}(k,n) \|$ and α controls the spatial selectivity for the sources located outside the spot. In addition, the gain $Q_i(k) \in [0,1]$ applied to the diffuse component enables the student to control the level of the ambient sound. The output diffuse signals $Y_{\mathrm{d},i}(k,n)$ for the left and right headphone channel are decorrelated such that the coherence between $Y_{\mathrm{d,left}}(k,n)$ and $Y_{\mathrm{d,right}}(k,n)$ corresponds to the target coherence in binaural hearing [18], [29]. Finally, it should be noted that since the propagation compensation and the spatial selectivity gains are typically real factors, the phase of the direct and diffuse components are equal to those observed at the reference microphone. However, the complex HRTFs that dependent on the DOAs at the virtual listening position ensure that the spatial cues are correct.



[FIG8] **A general parametric spatial sound processing scheme for binaural hearing aids.**

BINAURAL HEARING AIDS

Developments in acoustic signal processing and psychoacoustics have lead to the advancement of digital hearing aids that were first developed in the 1990s. The early devices included the unilateral (i.e., single-ear) and bilateral hearing aids, where two independent unilateral hearing aids are used for the left and right ears, respectively. More recently binaural hearing aids, in which signals and parameters can be exchanged between the left and right hearing aid, have been brought to the market. Binaural hearing aids are advantageous compared to unilateral and bilateral hearing aids as they can further improve speech intelligibility in difficult listening situations, improve the ability to localize sounds, and decrease listening fatigue. Besides dynamic range compression and feedback cancelation, wind and ambient noise reduction, dereverberation and directional filtering are important features of state-of-the-art hearing aids.

Let us consider a situation in which we have one desired talker in front and two interfering talkers at the right side of the hearing-aids user, as illustrated in Figure 8. In such a situation, directional filtering allows a hearing-aid user to perceive sounds arriving from the front more clearly than the sounds from the sides. In addition, one can aim at reducing the amount of diffuse sounds such that the SDR increases.

While many state-of-the-art directional filtering techniques for hearing aids are based on classical differential array processing, some parametric spatial sound processing techniques have been proposed. In [14], the left and right microphone signals were jointly analyzed in the time-frequency domain to determine: 1) the interaural phase difference and interaural level difference that strongly depend on the DOA of the direct sound, and 2) the interaural coherence that measures the degree of diffuseness. Based on these parameters, three gains were computed related to the degree of diffuseness, signal-to-interference ratio, and direction of the sound. Finally, real-valued gains for the left and right microphones were determined based on these gains to reduce reverberation and interfering sounds. According to the authors of [14], the quality of the signal was good but the speech intelligibility improvement for a single interfering talker was unsatisfactory. In [15], the authors used two microphones at each side and adopted the DOA-based geometric model. The DOAs were estimated at low frequencies using the microphones at the left and respectively right side, and at high frequencies using the intermicrophone level differences. Finally, the signal of a single microphone positioned at the left and right, respectively, was modified based on the DOA estimates and degree of diffuseness. The evaluation of different setups with one desired talker and one interfering talker demonstrated that an improvement in the speech reception threshold (SRT) between 4 and 24 dB could be obtained.

In Figure 8, a general parametric spatial sound processing scheme is illustrated, where spatial analysis provides the DOA estimates, and the direct and diffuse sound estimates for the left and right ear are obtained using different (left or right) reference microphones. The left (and right) output signal can then be computed using (3) with $G_i(k,n) = B(k,n,\theta)H_{\text{ex}}(k)$ for $i \in \{\text{left, right}\}$, where $B(k,n,\theta)$ defines the desired spatial response that depends on the listening mode, $H_{\text{ex}}(k)$ helps to externalize sounds, and $Q_i(k) = c(k)H_{\text{ex}}(k)$ with $0 \leq c(k) < 1$ is a constant used to reduce the diffuse sound and hence increase the SDR at the output. At the cost of an increase in computational complexity and memory use, the proposed scheme can fully exploit all microphones.

While many more examples can be found in the literature, it can readily been seen that the parametric spatial sound processing, using either geometrically or psychoacoustically motivated parametric models, provides a flexible and efficient way to achieve directional filtering. The limited improvement in terms of the SRT reported in [14] could be related to the inherent tradeoff between interference reduction and speech distortion found in most single-channel processing techniques. Further research is required to develop robust and efficient parameter estimators for this application and to study the impact on the SRT. More advanced schemes to modify the spatial response and the DOAs based on the listening mode and the listening situation could be realized using the processing scheme depicted in Figure 8.

CONCLUSIONS

Parametric models have been shown to provide an efficient way to describe sound scenes. While in earlier work multiple microphones were only used to estimate the geometric model parameters, in more recent work it has been shown that they can also be used to estimate the direct and diffuse sound components. As the latter estimates are more accurate than single-channel estimates, the sound quality of the overall system is increased, for instance, by avoiding decorrelating the direct sound that may partially leak into the diffuse sound estimate in single-channel extraction. Depending on the application, the estimated components and parameters can be manipulated before computing one or more output signals by mixing the components together based on the parametric side information. In a spatial audio communication scenario in which the direct and diffuse signals as well as the parameters are transmitted to the far-end side, it is possible to determine at the receiver side which sounds to extract and how to accurately reproduce the recorded spatial sounds over loudspeakers or headphones. By using the position-based model, we have shown how binaural signals can be synthesized at the receiver side that correspond to a desired listening position on the recording side. Finally, we have described how parametric spatial sound processing can be applied to binaural hearing aids to achieve both directional filtering and dereverberation.

To date, the majority of the geometric models assume that at most one direct sound is active per time-frequency. Extensions of these models are currently under development where multiple direct sound components plus diffuse sound components coexist in a single time-frequency instance [23]. Preliminary results have shown that this model can help to further improve the spatial selectivity and sound quality.

We hope that by presenting this unified perspective on parametric spatial sound processing we can help readers to approach other problems encountered in assisted listening from this perspective and to help highlight relations between a family of approaches that may initially seem divergent.

ACKNOWLDGEMENTS

This work has received funding from the European Community's Seventh Framework (FP7/2007-2013) under grant agreement ICT-2011-287760, from the European Research Council under the European Community's Seventh Framework (FP7/2007-2013)/ERC grant agreement 240453, and from the Academy of Finland.

AUTHORS

Konrad Kowalczyk (konrad.kowalczyk@iis.fraunhofer.de) received the B.Eng. and M.Sc. degrees in telecommunications from AGH University of Science and Technology, Krakow, Poland, in 2005 and the Ph.D. degree in electronics and electrical engineering from Queens University, Belfast, United Kingdom, in 2009. From 2009 until 2011, he was a postdoctoral research fellow at the Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander-University Erlangen-Nürnberg, Germany. In 2012, he joined Fraunhofer Institue for Integrated Circuits IIS as an associate researcher for communication acoustics and spatial audio processing. His main research interests include virtual acoustics, sound field analysis, spatial audio, signal enhancement, and array signal processing.

Oliver Thiergart (oliver.thiergart@iis.fraunhofer.de) studied media technology at Ilmenau University of Technology (TUI), Germany, and received his Dipl.-Ing. (M.Sc.) degree in 2008. In 2008, he was with the Fraunhofer Institute for Digital Media Technology IDMT in Ilmenau where he worked on sound field analysis with microphone arrays. He then joined the Audio Department of the Fraunhofer Institute for Integrated Circuits IIS in Erlangen, Germany, where he worked on spatial audio analysis and reproduction. In 2011, he became a member of the International Audio Laboratories Erlangen where he is currently pursuing a Ph.D. degree in the field of parametric spatial sound processing.

Maja Taseska (maja.taseska@audiolabs-erlangen.de) received her B.Sc. degree in electrical engineering at Jacobs University, Bremen, Germany, in 2010, and her M.Sc. degree at the Friedrich-Alexander-University Erlangen-Nürnberg, Germany, in 2012. She then joined the International Audio Laboratories Erlangen, where she is currently pursuing a Ph.D. degree in the field of informed spatial filtering. Her current research interests include informed spatial filtering, source localization and tracking, blind source separation, and noise reduction.

Giovanni Del Galdo (giovanni.delgaldo@iis.fraunhofer.de) studied telecommunications engineering at Politecnico di

Milano, Italy. In 2007, he received his doctoral degree from Technische Universität Ilmenau on the topic of channel modeling for mobile communications. He then joined Fraunhofer Institute for Integrated Circuits IIS working on audio watermarking and parametric representations of spatial sound. In 2012, he was appointed full professor at TU Ilmenau in the research area of wireless distribution systems and digital broadcasting. His current research interests include the analysis, modeling, and manipulation of multidimensional signals, over-the-air testing for terrestrial and satellite communication systems, and sparsity-promoting reconstruction methods.

Ville Pulkki (Ville.Pulkki@aalto.fi) has been working in the field of audio since 1995. In his Ph.D. thesis (2001), he developed a method to position virtual sources for three-dimensional loudspeaker setups after researching the method using psychoacoustic listening tests and binaural computational models of human hearing. Later he worked on the reproduction of recorded spatial sound scenarios, on the measurement of head-related acoustics and on the measurement of room acoustics with laser-induced pressure pulses. Currently he holds a tenure-track assistant professor position in Aalto University and runs a research group with 14 researchers. He is a fellow of the Audio Engineering Society (AES) and received the AES Publication Award. He has also received the Samuel L. Warner Memorial Medal from the Society of Motion Picture and Television Engineers.

Emanuël A.P. Habets (e.habets@ieee.org) is an associate professor at the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg and Fraunhofer IIS), and head of the Spatial Audio Research Group at Fraunhofer IIS, Germany. He received the Ph.D. degree in electrical engineering from the Technische Universiteit Eindhoven, The Netherlands, in 2007. From 2007 until 2009, he was a postdoctoral fellow at the Technion–Israel Institute of Technology and at the Bar-Ilan University, Israel. From 2009 until 2010, he was a research fellow at Imperial College London, United Kingdom. Currently, he is an associate editor of *IEEE Signal Processing Letters*, a member of the IEEE Signal Processing Society (SPS) Technical Committee on Audio and Acoustic Signal Processing, a member of the IEEE SPS Standing Committee on Industry Digital Signal Processing Technology, and has been a guest editor of *IEEE Journal of Selected Topics in Signal Processing*. He is the recipient, with I. Cohen and S. Gannot, of the 2014 IEEE SPS Signal Processing Letters Best Paper Award. He is a Senior Member of the IEEE.

REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[2] M. Goodwin and J.-M. Jot, "Spatial audio scene coding," in *Proc. Audio Engineering Society Convention 125*, Oct. 2008.

[3] Z. Fejzo, S. Hastings, J. D. Johnston, and J.-M. Jot, "Beyond coding: Reproduction of direct and diffuse sound in multiple environments," in *Proc. Audio Engineering Society Convention 129*, Nov. 2010.

[4] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 47, pp. 945–978.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[6] C. Faller, "Parametric coding of spatial audio," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2004.

[7] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," *J. Audio Eng. Soc.*, vol. 59, no. 12, pp. 924–935, 2011.

[8] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics, (WASPAA'09)*, Oct. 2009, pp. 337–340.

[9] I. Tashev, M. Seltzer, and A. Acero, "Microphone array for headset with spatial noise suppressor," in *Proc. 9th Int. Workshop Acoustic, Echo, Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005, pp. 29–32.

[10] M. Kallinger, G. Del Galdo, F. Kuech, D. Mahne, and R. Schultz-Amling, "Spatial filtering using directional audio coding parameters," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 2009, pp. 217–220.

[11] M. Kallinger, G. Del Galdo, F. Kuech, and O. Thiergart, "Dereverberation in the spatial audio coding domain," in *Proc. Audio Engineering Society Convention 130*, London, U.K., May 2011.

[12] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. Hands-Free Speech Communication Microphone Arrays (HSCMA)*, Edinburgh, U.K., May 2011, pp. 185–190.

[13] R. Schultz-Amling, F. Kuech, O. Thiergart, and M. Kallinger, "Acoustical zooming based on a parametric sound field representation," in *Proc. Audio Engineering Society Convention 128*, London, U.K., May 2010.

[14] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, no. 1–2, pp. 111–138, Jan. 2003.

[15] J. Ahonen, V. Sivonen, and V. Pulkki, "Parametric spatial sound processing applied to bilateral hearing aids," in *Proc. Audio Engineering Society Conf.: 45th Int. Conf. Applications Time-Frequency Processing Audio*, Mar. 2012.

[16] C. Faller, "Microphone front-ends for spatial audio coders," in *Proc. Audio Engineering Society Convention 125*, San Francisco, CA, Oct. 2008.

[17] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. 2nd Int. Symp. Ambisonics Spherical Acoustics*, May 2010.

[18] J. Blauert, Ed., *Communication Acoustics*. New York: Springer, 2005, vol. 1.

[19] S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 2002, vol. 1, pp. 529–532.

[20] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. New York: Springer, 2008.

[21] H. L. van Trees, *Detection, Estimation, and Modulation Theory*, vol. IV, *Optimum Array Processing*. New York: Wiley, Apr. 2002.

[22] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin: Springer, 2001, ch. 4, pp. 61–85.

[23] O. Thiergart, M. Taseska, and E. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Processing*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[24] O. Thiergart and E. A. P. Habets, "Extracting reverberant sound using a linearly constrained minimum variance spatial filter," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 630–634, May 2014.

[25] O. Thiergart, G. D. Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J. Acoust. Soc. Amer.*, vol. 132, no. 4, pp. 2337–2346, 2012.

[26] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *J. Acoust. Soc. Amer.*, vol. 131, no. 3, pp. 2141–2151, Mar. 2012.

[27] O. Thiergart, G. Del Galdo, M. Taseska, and E. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 12, pp. 2583–2594, Dec. 2013.

[28] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[29] V. Pulkki and J. Merimaa, "Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20, Feb. 2006.

[30] K. Kowalczyk, O. Thiergart, A. Craciun, and E. A. P. Habets, "Sound acquisition in noisy and reverberant environments using virtual microphones," in *Proc. 2013 IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA)*, Oct. 2013.

[SP]

[W. Bastiaan Kleijn, João B. Crespo, Richard C. Hendriks,
Petko N. Petkov, Bastian Sauert, and Peter Vary]

Optimizing Speech Intelligibility in a Noisy Environment

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

[A unified view]

Modern communication technology facilitates communication *from* anywhere *to* anywhere. As a result, low speech intelligibility has become a common problem, which is exacerbated by the lack of feedback to the talker about the rendering environment. In recent years, a range of algorithms has been developed to enhance the intelligibility of speech rendered in a noisy environment. We describe methods for intelligibility enhancement from a unified vantage point. Before one defines a measure of intelligibility, the level of abstraction of the representation must be selected. For example, intelligibility can be measured on the message, the sequence of words spoken, the sequence of sounds, or a sequence of states of the auditory system. Natural measures of intelligibility defined at the message level are mutual information and the hit-or-miss criterion. The direct evaluation of high-level measures requires quantitative knowledge of human cognitive processing. Lower-level measures can be derived from higher-level measures by making restrictive assumptions. We discuss the implementation and performance of some specific enhancement systems in detail, including speech intelligibility index (SII)-based systems and systems aimed at enhancing the sound-field where it is perceived by the listener. We conclude with a discussion of the current state of the field and open problems.

INTRODUCTION

Humans adapt their speech to the physical environment. Based on the facial expression of a listener, a talker may repeat or reformulate the message. A noisy environment gives rise to the Lombard effect, e.g., [1], an involuntary change in the speech characteristics that makes speech more intelligible.

In modern communication systems, the speaker often has little or no awareness of the physical environment in which the speech is rendered. This is perhaps most obvious for current-generation speech synthesis, which produces speech without

Digital Object Identifier 10.1109/MSP.2014.2365594
Date of publication: 12 February 2015

1053-5888/15©2015 European Union

consideration of the rendering environment. It is also a major factor in human-to-human communications as communication technology degrades or severs the auditory and visual links between the speaker and the environment. For example, an announcer at a railway station generally receives little visual or auditory feedback. Similarly, a phone user lacks information about the rendering environment, even less so if effective noise-suppression technology is used.

The lack of feedback, together with the recent ability to communicate from anywhere to anywhere, often leads to low intelligibility. Phone booths are a relic of the past: the mobile phone is expected to function in any environment, whether it is a car, a cafeteria, or a windstorm. Thus, there is a strong motivation for algorithms that can improve the intelligibility of speech rendered in a noisy environment.

IN RECENT YEARS, A RANGE OF ALGORITHMS HAS BEEN DEVELOPED TO ENHANCE THE INTELLIGIBILITY OF SPEECH RENDERED IN A NOISY ENVIRONMENT. WE DESCRIBE METHODS FOR INTELLIGIBILITY ENHANCEMENT FROM A UNIFIED VANTAGE POINT.

Ever since the early work of Griffiths [2] and Niederjohn and Grotelueschen [3], researchers have attempted to create processing methods that increase the intelligibility of speech in a noisy environment. Driven by the rapid growth of mobile telephony, research efforts on intelligibility in noise have increased significantly in the last five years. The result is that it is now possible to significantly increase the intelligibility of speech in noise, e.g., [4]– [11]. Approaches to intelligibility enhancement are increasingly based on the mathematical optimization of quantitative measures that are hypothesized to represent intelligibility accurately. First introduced by [2], the optimization approach has been used in numerous recent studies, starting with [12]. The optimization criteria vary widely as the signal processing algorithms are derived from different viewpoints and with different computational and delay constraints. Criteria used include the probability of correct phoneme recognition [11], auditory models [6], [13], [14], the articulation index [2], the SII [4], [8], mutual information [15], and sound-field distortion [16].

In this tutorial, we describe a range of methods for intelligibility enhancement from a unified vantage point, delineating the similarities and dissimilarities between the various approaches. In contrast to the broad overview of human and algorithmic modifications that affect intelligibility in [7], our discussion focuses on the definition and use of quantitative measures of intelligibility, showing that many of these measures can be derived from the same basic principle.

MEASURES OF INTELLIGIBILITY

In this section, we first discuss how to define a quantitative measure of intelligibility. We then discuss practical measures of intelligibility.

DEFINING INTELLIGIBILITY

The word *intelligibility* expresses a qualitative measure of whether a conveyed message is interpreted correctly by a human listener. To define quantitative instrumental measures of intelligibility, we must select a level of abstraction. That is, we must decide if we measure intelligibility on the sequence of words spoken, on the sequence of sounds, on a sequence of states of the auditory system, or on the acoustic signal waveform. A word sequence is an example of a description at a high level of abstraction, whereas a signal waveform is a description at a low level of abstraction.

The higher the level of abstraction, the more fundamental the measure of intelligibility: the objective of speech is to convey a message and not to convey a sequence of sounds. A particular measure will be useful for enhancement at its own level of abstraction and below. Consider an intelligibility measure operating at the word sequence level. It can be used to evaluate which of a set of sentence formulations with similar meaning is more intelligible. It can also be used to evaluate if a particular spectral modification (e.g., a particular filtering operation) makes speech more intelligible.

The generality of high-level measures has a cost: we must map the observations into a sequence at that high abstraction level. For acoustic observations and a measure operating at the word-sequence level, this requires a robust model of hearing that maps the observed acoustic signal into a word sequence. Therefore, although it cannot optimize linguistic formulations, an intelligibility measure operating on a sequence of auditory states may be attractive when optimizing a spectral modification of the signal.

While illusive in practical measurements, the message itself, a random variable that we denote as $\mathbf{M}$, can be used to define the most basic measure of intelligibility. (To aid clarity, we will write random variables as bold-face characters and their realizations as regular characters.) In the following, we will show how such a basic measure can be used to derive measures that have been derived earlier on a heuristic basis. To facilitate our reasoning, we will be opportunistic and sometimes describe the messages as countable, which is consistent with the notion that a message is a discrete word sequence, and at other times as continuous, which is consistent with the notion that articulation is continuously variable. To avoid confusion, we add a breve, as in $\check{\mathbf{M}}$, whenever messages are considered countable.

A natural measure of intelligibility is the mutual information between the message conveyed by the talker $\check{\mathbf{M}}_T$ and the message interpreted by the listener $\check{\mathbf{M}}_L$:

$$I(\check{\mathbf{M}}_L; \check{\mathbf{M}}_T) = \sum_{\check{M}_L, \check{M}_T} p_{LT}(\check{M}_L, \check{M}_T) \log \frac{p_{L|T}(\check{M}_L \,|\, \check{M}_T)}{p_L(\check{M}_L)}, \qquad (1)$$

where we used the simplified notation $p_{LT} = p_{\check{\mathbf{M}}_L \check{\mathbf{M}}_T}$ and $p_{L|T} = p_{\check{\mathbf{M}}_L | \check{\mathbf{M}}_T}$ for the joint and conditional probabilities and use the same convention for the marginal probabilities of the conveyed and received messages and p_T and p_L.

We can reformulate the criterion (1) as a measure of distortion $D(\check{\mathbf{M}}_L, \check{\mathbf{M}}_T)$ that is a functional of $p_{L|T}$. Mutual information is nonnegative and cannot be larger than the entropy $H(\check{\mathbf{M}}_T)$. Thus, the difference $D(\check{\mathbf{M}}_L, \check{\mathbf{M}}_T) = H(\check{\mathbf{M}}_T) - I(\check{\mathbf{M}}_L; \check{\mathbf{M}}_T)$ is nonnegative and can be interpreted as a distortion. It can be written as a general distortion measure operating on $p_{L|T}$ for a given talker message distribution p_T:

$$D(\check{\mathbf{M}}_L, \check{\mathbf{M}}_T) = \sum_{\check{M}_T} p_T(\check{M}_T) \sum_{\check{M}_L} d(p_{L|T}(\check{M}_L \,|\, \check{M}_T)), \qquad (2)$$

where d is a nonnegative function of $p_{L|T}(\check{M}_L \,|\, \check{M}_T)$. For the mutual information based distortion measure $d(p_{L|T}(\check{M}_L \,|\, \check{M}_T)) = p_{L|T}(\check{M}_L \,|\, \check{M}_T) \log(p_L(\check{M}_L)/p_{L,T}(\check{M}_L, \check{M}_T))$, where we note that the argument of the logarithm can be written in terms of $p_{L|T}(\check{M}_L \,|\, \check{M}_T)$ and the given $p_T(\check{M}_T)$ only. The intelligibility enhancement problem is to find the $p_{L|T}$ that minimizes the distortion (2) subject to the constraints set by the scenario.

An alternative to the mutual information based distortion measure can be based on the hit-or-miss distortion, $d(p_{L|T}(\check{M}_L \,|\, \check{M}_T)) = p_{L|T}(\check{M}_L \,|\, \check{M}_T)(1 - \delta_{\check{M}_L, \check{M}_T})$, where $\delta_{\check{M}_L, \check{M}_T}$ is a Kronecker delta function. In this case (2) becomes

$$D(\check{\mathbf{M}}_L, \check{\mathbf{M}}_T) = 1 - \sum_{\check{M}_T} p_{LT}(\check{M}_T, \check{M}_T) = 1 - E_T[p_{L|T}(\check{\mathbf{M}}_T \,|\, \check{\mathbf{M}}_T)]. \quad (3)$$

The conditional probability $p_{L|T}(\check{M}_T \,|\, \check{M}_T)$ in (3) corresponds to the probability that the message is interpreted correctly. Thus, an alternative to maximizing the mutual information of the conveyed and received message is to maximize the expected probability of correct message interpretation, $\mathrm{E}_T[p_{L|T}(\check{\mathbf{M}}_T \,|\, \check{\mathbf{M}}_T)]$, where the expectation is over the conveyed messages, $\check{\mathbf{M}}_T$. We will discuss the practical use of this high-level measure in the section "Measures Operating on a Word Sequence."

While the measures (1) and (3) are general, they cannot be used directly. Either the description of the message or the human cognitive system must be approximated such that the measures can be applied to observable signals. The paradigm shows where such approximations are made, but it does not show their quantitative impact. Thus, experiments must be used to verify the validity of the resulting system.

Next, we consider how to derive a low-level, acoustics-based measure from a high-level, message-based measure. For this it is convenient to consider the message as a continuous variable. A conveyed speech message $\mathbf{M}_T$ is rendered in the form of an acoustic signal, which we represent by an acoustic sequence $\mathbf{a}_T$. The sequence $\mathbf{a}_T$ can, for example, consist of signal samples or short-term spectral descriptions, such as cepstral vectors. This sequence is rendered in a noisy environment and the listener observes a corrupted sequence $\mathbf{a}_L$, which is then interpreted as a message $\mathbf{M}_L$. The communication process thus forms a Markov chain $\mathbf{M}_T \to \mathbf{a}_T \to \mathbf{a}_L \to \mathbf{M}_L$. It is natural that environmental noise makes the mapping $\mathbf{a}_T \to \mathbf{a}_L$ stochastic.

Upon reflection, it is clear that the mappings $\mathbf{M}_T \to \mathbf{a}_T$ and $\mathbf{a}_L \to \mathbf{M}_L$ are also stochastic: a message is generally not formulated and never articulated in precisely the same manner, and the interpretation of the acoustic sequence $\mathbf{a}_L$ is subject to random variations during the human cognitive process. Anticipating the discussions in the section "Measures Operating on a Word Sequence," it can be argued that these variations are captured by the statistical modeling of modern automatic speech recognition (ASR) algorithms. If we assume the message formulation is perfect, a simple but effective model of the production and interpretation processes is that they are subject to additive noise components [15], which we will refer to as, respectively, *production noise* and *interpretation noise*. For example, variability in articulation across different persons may be approximated as additive noise in a representation based on cepstral or log spectral vectors.

For convenience let us define auxiliary bijective mappings $M_T \leftrightarrow s_T$ and $M_L \leftrightarrow s_L$, where s_T and s_L are realizations of random acoustic sequences. We have

$$\begin{aligned} \mathbf{a}_T &= \mathbf{s}_T + \mathbf{v}_T \\ \mathbf{a}_L &= \mathbf{a}_T + \mathbf{v}_E \\ \mathbf{s}_L &= \mathbf{a}_L + \mathbf{v}_L, \end{aligned} \qquad (4)$$

where $\mathbf{v}_T$, $\mathbf{v}_E$, and $\mathbf{v}_L$ are additive noise processes, modeling the production noise, environmental noise, and interpretation noise, respectively. Note that the system model differs from the standard system model in communication theory, which does not include production noise and interpretation noise.

To facilitate analysis, let us assume the sequences $\mathbf{s}_T$, $\mathbf{v}_T$, $\mathbf{v}_E$, and $\mathbf{v}_L$ to be jointly Gaussian processes. Furthermore, we denote by ρ_{sa} the correlation coefficient of (the samples of the) processes $\mathbf{s}$ and $\mathbf{a}$ and write $\rho_0 = \rho_{\mathrm{s}_T \mathrm{a}_T} \rho_{\mathrm{a}_L \mathrm{s}_L}$. Let us first consider the case where the signals are white. Exploiting that mutual information is invariant under reparametrization of the marginal variables, it is then easy to see that [15]

$$I(\mathbf{M}_L; \mathbf{M}_T) = I(\mathbf{s}_T; \mathbf{s}_L) = -\frac{1}{2} \log \frac{(1 - \rho_0^2)\,\xi + 1}{\xi + 1}, \qquad (5)$$

where $\xi = (\sigma_{\mathrm{a}_T}^2 / \sigma_{\mathrm{v}_E}^2)$ is the signal-to-noise ratio (SNR) of the acoustic channel $\mathbf{a}_T \to \mathbf{a}_L$, and $\sigma_{\mathrm{a}_T}^2$ and $\sigma_{\mathrm{v}_E}^2$ are the variances of processes $\mathbf{a}_T$ and $\mathbf{v}_E$, respectively. An important and intuitive conclusion that can be drawn from (5) is that if the environmental noise variance is small compared to the production and interpretation noise variances, then the mutual information between talker and listener is not affected significantly by the environmental noise.

The spectral coloring of the acoustic content can be accounted for by splitting the signal into spectral bands such that each band can be approximated as white. If we assume the signals to be stationary, the frequency bands are independent and the mutual information can be written as the sum of the mutual informations in the bands

$$I(\mathbf{M}_L; \mathbf{M}_T) = -\frac{1}{2} \sum_i \log \frac{(1 - \rho_{0,i}^2)\,\xi_i + 1}{\xi_i + 1}, \qquad (6)$$

where i is the band index and where $\xi_i = (\sigma_{\mathrm{a}_{T,i}}^2 / \sigma_{\mathrm{v}_{E,i}}^2)$ is the SNR of the acoustic channel in band i. Note that the SNR in (6) is computed on whichever representation is used for the acoustic features. Also note that the variances $\sigma_{\mathrm{a}_{T,i}}^2$ and $\sigma_{\mathrm{v}_{E,i}}^2$ are generally unknown and must be estimated in practice. For example, if the acoustic features are based on the short-time discrete Fourier transform (DFT) coefficients, variance estimation can be based on the short-time DFT periodogram, i.e., $\hat{\sigma}^2_{\mathrm{a}_{T,i}} = |a_{T,i}|^2$ having a variance of $E[|\mathbf{a}_{T,i}|^2]^2$. The low-level measure (6) can then be used directly to optimize speech intelligibility [15].

THE LACK OF FEEDBACK, TOGETHER WITH THE RECENT ABILITY TO COMMUNICATE FROM ANYWHERE TO ANYWHERE, OFTEN LEADS TO LOW INTELLIGIBILITY.

The frequency resolution of the human auditory system decreases with frequency, which reduces the mutual information from that obtained with (6) for a uniform high resolution. An improved model of information transfer is obtained by assuming that the signal is represented with one independent component per equivalent rectangular bandwidth (ERB), which is consistent with studies on intelligibility [17]. We show in the section "Measures Operating on Spectral Band Powers" that this approach provides an information-theoretical justification of the well-known SII [18], a low-level measure of intelligibility.

PRACTICAL MEASURES OF INTELLIGIBILITY

Existing practical measures of intelligibility generally operate at the word-sequence level, at the level of a sequence of auditory states, or at the level of short-term spectra. We discuss these classes next and end with a discussion of the constraints that must be imposed on the optimization.

MEASURES OPERATING ON A WORD SEQUENCE

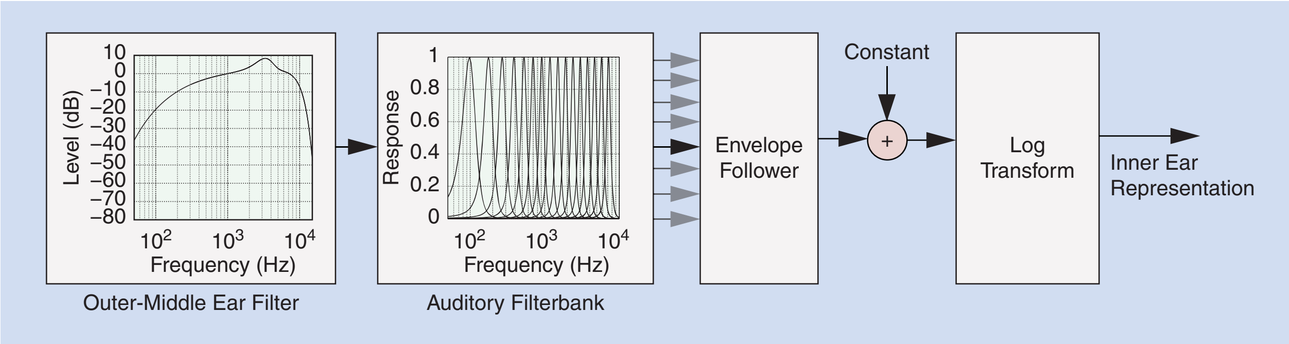
In the section "Defining Intelligibility," we discussed that the expected probability of correct interpretation of the message, $E_T[p_{L|T}(\check{\mathbf{M}}_T|\check{\mathbf{M}}_T)]$, is a reasonable measure of intelligibility. This measure can be approximated as $\overline{p_{L|T}(\check{M}_T|\check{M}_T)}$ on real-world data, where the overbar indicates averaging over realizations $\check{M}_T$. If the averaging is done in time, i.e., over segments of a single larger message (e.g., words), then this operation assumes ergodicity. The measure is easily evaluated in a test with human test subjects, where $p_{L|T}(\check{M}_T|\check{M}_T)$ can be estimated using histograms. A machine-based quantitative measure requires a mapping from any particular acoustic observation a_L to a message $\check{M}_L$ that captures the probabilistic nature of this mapping as performed by humans. As will be discussed in the section "Word-Sequence Probability-Based Enhancement," the standard approach to ASR computes the probability of the observations given a message (word, or word sequence). The basic assumption for machine-based intelligibility enhancement is then that the trend of ASR word probability in noise tracks the trend of human recognition performance in noise sufficiently well for the modification parameters that are optimized. Experiments confirmed this hypothesis [11], [19] for a particular set of practical systems.

MEASURES OPERATING ON A SEQUENCE OF AUDITORY STATES

It is advantageous to minimize the delay and computational requirements of the intelligibility measure, particularly if the types of modification are restricted. Let us assume that the modification is a spectral modification, that the word sequence and speaking rate are fixed, and that the highest intelligibility is achieved by the original speech without environmental noise. (The latter assumption is an additional simplification required for this approach.) Then it is natural to use a distortion measure operating on the sequence of auditory states as a measure of intelligibility. Such measures can exploit that quantitative knowledge of the auditory periphery has increased significantly in the last three decades (e.g., [20]).

The straight comparison of the auditory states of the conveyed and received signal ignores the production noise $\mathbf{v}_T$ of (4). That is, the auditory model does not weigh signal components according to their relevance in terms of precision of signal production. However, the auditory model precision of a speech component may form a reasonable match to the precision of speech production, simplifying the introduction of production noise.

Although auditory models differ in exactly how the inner ear representation is obtained, they follow in many cases a similar strategy for modeling the auditory system. In Figure 1, we outline the basic building blocks of the psychoacoustic model presented in [21], which is simple but representative of many other models, such as [20]. The first stage of the auditory model consists of a filter that mimics the frequency characteristics of the outer and



[FIG1] The basic structure of the auditory model presented in [21].

middle ear. This filter is cascaded with an auditory filter bank that models processing at the level of the basilar membrane in the cochlea. Subsequently, the envelope of each of the outputs of the auditory filters is obtained, which simulates the transduction of the inner hair cells. To model an absolute hearing threshold, a constant is added to each envelope. In the current context, this threshold corresponds to an interpretation noise. In the final stage, a log transform is used to model the loudness dependent compression of the auditory filter bank outputs by the outer hair cells. An important difference between the model from [21] and the more advanced model presented in [20] is the logarithmic transform, which is a simplification of the adaptation loops that are used in [20]. The simplification particularly affects the output near transitions where the gain of adaptation loops changes.

By applying an auditory model to the acoustic sequences a_T and a_L and comparing the results, a distortion measure can be obtained. Mutual information is a natural measure for this purpose, but, to our best knowledge, it has not been applied to the auditory representation for intelligibility enhancement. Note that while mutual information is not affected by smooth invertible mappings, auditory representations likely are not smooth mappings from features such as cepstra, or line spectral frequencies. This suggests that it may be essential to consider the detailed behavior of more sophisticated auditory models.

In the literature, various measures have been used to compare the auditory representations of a_T and a_L. In [14], it was shown that an ℓ_1 criterion leads to a mathematically tractable method and to provide good results for intelligibility enhancement. Reference [13] uses a similar auditory model for the so-called glimpse proportion measure of intelligibility: rather than comparing a_T and a_L directly, it compares the auditory representation of the a_T with the auditory representation of the environmental noise v_E. The glimpse proportion approach computes the proportion of signal blocks where the auditory representation of the signal is louder than the noise. In more recent work on the glimpse proportion, a sigmoidal function is applied to the difference of the auditory signal and noise representations [6], [22]. The method provides good intelligibility enhancement [6], [22], [23]. Both the ℓ_1 criterion and glimpse proportion approaches do not explicitly consider the information conveyed in a particular signal component, which should, at least in principle, be a disadvantage compared to mutual information-based approaches.

MEASURES OPERATING ON SPECTRAL BAND POWERS

The mutual information between $\mathbf{M}_L$ and $\mathbf{M}_T$ (6) can be seen to correspond to a classic view of intelligibility based on band powers of the auditory filter bank [17], [18], [24]–[27], by writing it as

$$I(\mathbf{M}_L;\mathbf{M}_T)=\sum_i \tilde{I}_i\, A_i(\xi_i) \qquad (7)$$

$$\tilde{I}_i=-\frac{1}{2}\log\left(1-\rho_{0,i}^2\right) \qquad (8)$$

$$A_i(\xi_i)=\frac{\log\dfrac{(1-\rho_{0,i}^2)\,\xi_i+1}{\xi_i+1}}{\log\left(1-\rho_{0,i}^2\right)}. \qquad (9)$$

The maximum mutual information is attained at high SNR and is $\sum_i \tilde{I}_i$. Defining $I_i=\tilde{I}_i/\sum_j \tilde{I}_j$ and normalizing (7) accordingly, we recognize I_i as the so-called band-importance function and $A_i(\xi_i)$ as the so-called weighting function or band-audibility function. The formulation (7) forms the basis of speech intelligibility measures such as the SII [18] and the extended SII [27]. These measures are descendants of the so-called articulation index [24], [25], a measure that predates information theory. In this classic view, I_i characterizes the importance of frequency band i and the factor A_i is a weighting function that indicates what fraction of the information is delivered to the listener. The information-theory derived form of A_i shown in (9) describes a sigmoidal function that approximates the definition of A_i in the SII. [Equation (9) neglects the threshold of hearing, the effect of high loudness, and the self-masking of noise.] Our derivation of the band importance function I_i of (8) makes its dependency on the production and interpretation noise explicit. If the relative variances of the production and interpretation noise of a band are low (high production and interpretation SNR; $\rho_{0,i}$ approaches one), that band is important for intelligibility. In the SII definition, the values of I_i are set empirically. As is shown in [15], the differences between the formulas for the classic approach and the aforementioned information-theoretical derivation are well within the precision of the original heuristic derivation of the classic view. The classic SII has proven to be highly correlated with speech intelligibility in many conditions and has been used as a basis for speech intelligibility enhancement [4], [8], [12], [28]. It is discussed in additional detail in the section "SII-Based Enhancement."

CONSTRAINTS ON OPTIMIZATION

In most cases, the optimization must be performed subject to one or more constraints. Important constraints are the speech-like nature of the output, the signal power, and system delay. Additional constraints may be required. For instance, for a given message M_T (and speaking rate), a longer word sequence will likely be more intelligible than a short one, thus making a length constraint natural.

The speech-like nature, or the speech quality, of the enhanced output may require an explicit constraint. However, in most practical systems the speech-like nature is enforced implicitly by either the modification strategy, or the optimization criterion, or both. Modification strategies such as slowly varying spectral shaping facilitate speech-like output only. The maximum probability of correct phoneme recognition is an example of a criterion that favors signal features that resemble those of clean speech.

Signal power is a natural constraint. The unconstrained optimization of signal spectral modifications may lead to an unbounded increase of the signal power if the reduction in recognition performance of the human auditory system for loud sounds is not considered. Thus, a power constraint must be applied to prevent hearing injuries and loudspeaker damage. Approximations to perceived loudness, either in the form of an analytic expression, or in the form of an algorithm, may also be used as constraints.

The system delay must be constrained in real-time systems. This may prevent the usage of particular distortion measures and modification operators.

SIGNAL PROCESSING APPROACHES

In this section, the focus is on creating practical enhancement systems. We start with a discussion of various modifications that can be made and then discuss three approaches to enhancement and their performance. Specific applications are described in "Making Mobile Phones More Intelligible" and "Making It Work for Hearing Instruments."

THE CLASSIC SII HAS PROVEN TO BE HIGHLY CORRELATED WITH SPEECH INTELLIGIBILITY IN MANY CONDITIONS AND HAS BEEN USED AS A BASIS FOR SPEECH INTELLIGIBILITY ENHANCEMENT.

SPEECH MODIFICATIONS

The basic paradigm of intelligibility enhancement discussed in this article is to select a modification operation to be used for preprocessing the signal and a measure of intelligibility, and then to adjust the parameters of the modification operation to maximize the measure. We discuss the classes of modifications that have been used or can be used and report on current knowledge about their effectiveness.

Enhancement operators can be classified according to a number of criteria. Operators can be classified generically as time-varying or time-invariant and as linear or nonlinear. Most intelligibility enhancement operators are time-invariant and nonlinear. However, low-level operators that use a linear filtering of the signal [8] have been used and perform well (if the filter is adapted, the operator is nonlinear).

Additional classifications can be made based on the specific processing performed on the message. Depending on the abstraction level where a modification takes place, we identify lexical (high level), prosodic (midlevel), and spectral and temporal (low level) modifications. In accordance with the Markov chain model of the communication process, presented in the section "Defining Intelligibility," a high-level modification affects the message representation at the lower levels. The operator can be independent or dependent on the environmental disturbance, i.e., it can be nonadaptive or adaptive. Finally, depending on the origin of a modification there are 1) mimicking strategies, i.e., modifications that attempt to mimic modifications used consciously or subconsciously by humans producing speech in adverse conditions, and 2) rational strategies based on, e.g., expert insight in the human auditory periphery and in cognition [3] or of the sound field [16], [29].

MAKING MOBILE PHONES MORE INTELLIGIBLE

Mobile telephony is often conducted in the presence of acoustical background noise such as traffic or babble noise. In this situation, the listener perceives a mixture of clean speech and environmental noise from the near-end side, which generally leads to an increased listening effort and possibly to reduced speech intelligibility. As the noise signal generally cannot be changed, the manipulation of the far-end signal is the only way to effectively improve speech intelligibility and to ease listening effort for the near-end listener.

In the mobile phone application, the algorithmic delay of the processing is crucial since the allowed round-trip delay of the communication system is limited. This places a severe constraint on the modification operator. Furthermore, the restrictions of the microloudspeakers of mobile phones need to be considered. The maximum thermal load of the microloudspeaker constitutes a major limitation, which can be taken into account with a constraint on the total audio power. Finally, the ear of the near-end listener is usually next to the loudspeaker and must be protected from damage and pain. This can be ensured by power limitations for the critical bands.

In unpublished work of the Listening Talker (LISTA) project (http://listeningtalker.org), 44 possible modifications were identified. This includes the modification strategies used in essentially all existing intelligibility enhancement systems. The effectiveness of some of the listed modifications on the intelligibility in noisy environments is reviewed in [7] and [9].

As is discussed in [7] and [9], mimicking strategies such as pitch modification, vowel space adjustment, and uniform speaking rate reduction do not improve intelligibility consistently when applied to natural speech. This outcome suggests that such modifications may have an auxiliary role or may be the result of physical limitations in the speech production mechanism. Other mimicking candidate modifications include changing the relative duration of phonetic units and shortening units that are more sensitive to energetic masking in favor of more robust units. As of now, no conclusions can be drawn about the benefit from such modifications. In the remainder of this section we focus on rational strategies.

Lexical speech modifications consist of, among others: 1) repetition to provide additional cues and 2) rephrasing to increase correct recognition probability as a result of better noise robustness and/or higher predictability. While repetition does not facilitate intelligibility optimization, rephrasing provides an intuitive and attractive modification class. The section "Measures of Intelligibility" discussed high-level modification measures that can, at least in principle, be used for this purpose. A practical rephrasing approach is presented in [19]: rather than comparing the measures directly, the method compares the sensitivity to noise addition of each formulation, according to the probability of correct recognition. The approach does not consider the predictability of the formulation, which is a major factor in intelligibility. An indirect indication of the expected gain from increasing the predictability of a formulation, e.g., by vocabulary size reduction, can be obtained by comparing the outcomes of intelligibility evaluations using closed-set [14] and open-set vocabulary bases [9]. The considerably higher intelligibility gain for closed-set evaluation suggests that it is feasible to design a modification system achieving intelligibility gain by improving the predictability of the formulation.

Low-level modifications do not require knowledge of the intended message transcription. These can be subdivided into spectral, temporal, and spatial signal modifications as well as combinations thereof.

Straightforward spectral shaping is employed in [8] and [12]. This modification facilitates both low complexity and a high intelligibility gain, e.g., [9], making these approaches particularly suitable for application in mobile telephony.

[THE MAXIMUM PROBABILITY OF CORRECT PHONEME RECOGNITION IS AN EXAMPLE OF A CRITERION THAT FAVORS SIGNAL FEATURES THAT RESEMBLE THOSE OF CLEAN SPEECH.]

Spectrotemporal energy redistribution is considered in [6], where the glimpse proportion is optimized. The use of a genetic algorithm to perform the optimization makes this method interesting primarily from a theoretical perspective. A low-complexity approach with high intelligibility gain that performs spectrotemporal energy redistribution by optimizing a perceptual distortion measure is presented in [14].

A particular class of spectrotemporal energy redistribution is obtained with dynamic range compression. This approach can either be nonadaptive or adaptive. In a large-scale subjective evaluation of proposed speech modification systems [9], most of the entries that incorporated dynamic range compression, including those related to the descriptions in [5], [23], [28], performed well.

Intelligibility can also be enhanced by controlling the spatial sound field near the ear with a multitude of remote loudspeakers. As discussed in more detail in the section "Enhancement over Multiple Spatial Points," if users are wearing microphones near their ears, reverberation and cross-talk between different messages can be reduced by feedback [16]. The goal is that only the desired signal is present at the ear of a user. If microphones are further from the ears of the listeners, the emerging field of multizone audio rendering becomes relevant, e.g., [29].

INTELLIGIBILITY ENHANCEMENT SYSTEMS

This section describes three practical methods for intelligibility optimization approaches. The described approaches are based on different principles.

SII-BASED ENHANCEMENT

State-of-the-art systems have been developed based on the decomposition into band-importance and band-audibility functions [4], [8], [28]. We provided a recent perspective on this decomposition in the section "Measures Operating on Spectral Band Powers." This section describes implementations that closely follow the SII standard.

The computation of the SII [18] uses a carefully calibrated specification of the speech spectrum $\sigma^2_{\mathbf{a}_{T,i}}$ and the noise spectrum $\sigma^2_{\mathbf{v}_{E,i}}$ (where i is a critical or third-octave band index) as measured over an entire utterance, including minor pauses. The approach accounts for both the hearing threshold and the loss of intelligibility at very high presentation (loudness) levels, using information stored in tables. For an acoustic time-domain speech signal $\mathbf{a}_T$, the equivalent speech spectrum level in dB, commonly denoted as E_i, is computed as

$$E_i = 10\log_{10}\left(\frac{\sigma^2_{\mathbf{a}_{T,i}}}{f_{\Delta,i}}\right) - 10\log_{10}(\sigma_0^2), \qquad (10)$$

where σ_0^2 denotes the digital reference power per hertz corresponding to the reference sound pressure of 20 μ Pa and $f_{\Delta,i}$ is the frequency bandwidth of the ith subband in hertz. The equivalent disturbance spectrum level, D_i, is computed in three steps: first the calibration (10) is applied, and then the threshold of hearing and instantaneous masking are accounted for. In [4] the threshold of hearing and in [8] both the threshold of hearing and instantaneous masking are neglected.

The band-audibility function of the SII also accounts for the decrease in intelligibility at high presentation (loudness) levels, which is not accounted for in (9). Consequently, it depends on both the SNR in the band and the absolute presentation level E_i. The band-audibility function is identical for different bands and

MAKING IT WORK FOR HEARING INSTRUMENTS

Hearing instruments aim to compensate for a hearing loss. Typically, this is done by amplifying a sound recording, followed by dynamic range compression to ensure the signal remains within the audible and comfortable range. Environmental noise degrades intelligibility for hearing instrument users in two ways. A first degradation is due to noise recorded by the microphones. To decrease the impact of this noise, noise reduction is applied to the recorded signal prior to amplification for hearing loss compensation.

A second degradation depends on the fitting: the user may experience direct environmental noise, leaking through the hearing instrument vent. This leakage degrades the intelligibility and can be overcome by processing the signal with the application of a speech intelligibility enhancement algorithm before play-out as discussed in the article.

Adopting the concept of interpretation noise, the patient's hearing loss can be measured and modeled by the noise process $\mathbf{v}_L$. The environmental noise that reaches the ear through the hearing instrument vent can be modeled by the process $\mathbf{v}_E$ of (4). Dynamic range compression can be taken into account by expressing the desired output range in terms of (frequency-dependent) absolute power constraints. Given this model, the hearing instrument can be optimized using one of the measures discussed in the section "Measures Operating on Spectral Band Powers" in a constrained fashion. The resulting integrated solution compares favorably with an ad hoc concatenation of processing steps, facilitates a conceptual understanding of the hearing impairment, and is likely to lead to an effective control of the instrument.

we denote it for a band i as $A(E_i, D_i)$. Let us define the piecewise linear sigmoid

$$\mathcal{S}(x; \beta_1, \beta_2) = (\max(\min(x, \beta_2), \beta_1) - \beta_1)/(\beta_2 - \beta_1),$$

which has a range $[0, 1]$. The band audibility function of the SII is factorized into two factors: the first factor accounts for the instantaneous masking and the second factor accounts for high presentation levels:

$$A(E_i, D_i) = \mathcal{S}(E_i; D_i - 15, D_i + 15) \\ \mathcal{S}(-E_i; -U_i - 170, -U_i - 10), \qquad (11)$$

where U_i is the standard speech level at normal voicing effort (provided in a table in the standard). The heuristic factor $\mathcal{S}(E_i; D_i - 15, D_i + 15)$ assumes that speech signals 15 dB below the disturbance level are fully masked, and speech signals 15 dB above the disturbance level are not masked, which leads to a curve similar to the result derived in (9).

The SII is a refined and normalized version of (7) that accounts for decreased intelligibility at high presentation levels

$$\mathrm{SII} = \sum_i I_i\, A(E_i, D_i). \qquad (12)$$

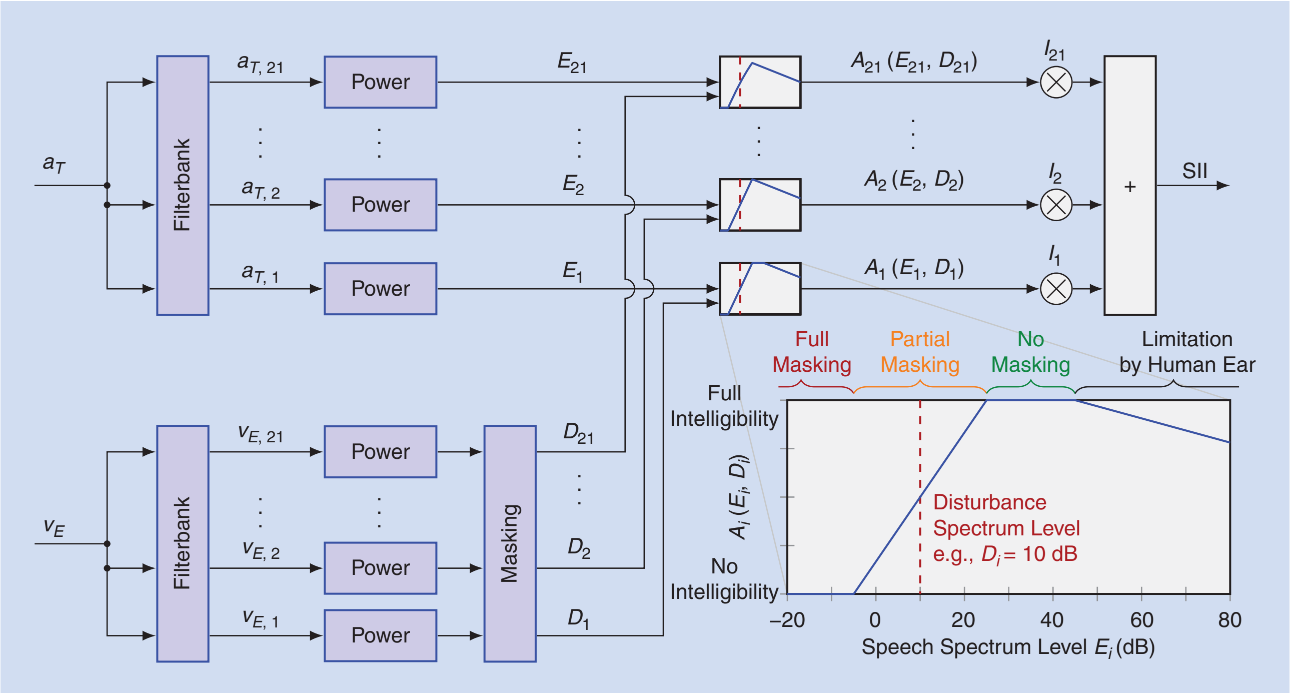
The band-importance function I_i in the SII is specified by a table that is based on fitting to a database. Figure 2 illustrates the computation of the SII. The suppression of the audibility function at high presentation levels is clearly shown in the panel showing the audibility function (11).

The measure (12) can be used to optimize a modification operator that shapes the spectrum. As the intelligibility decreases both at high and low presentation levels, the SII criterion can, in principle, be optimized without constraint. It is seen from (12) that if there is no global constraint, each frequency band can be optimized independently. The resulting solutions are not necessarily unique because of the form of $\mathcal{S}$. It is natural to select the solution that has the lowest power but does not reduce the speech power in any band. For low absolute noise levels, where the solution is not limited by the second factor in (11), the solution for the gain is [4]

$$g_i = \max(D_i + 15, E_i) - E_i, \qquad (13)$$

where the shaping gain g_i for band i is given in dB. In (13) the original equivalent speech spectrum level is E_i and the modified speech has equivalent speech spectrum level $g_i + E_i$.

As was discussed in the section "Constraints on Optimization," it is common to constrain the overall loudspeaker signal power in practical applications. The optimization of (12) subject to a power constraint was studied in [4] and [8]. To facilitate analysis, the two approaches use approximations of (12). Although the approximations are different, both neglect the second factor in (11) and start from $A(E_i, D_i) \approx \mathcal{S}(E_i; D_i - 15, D_i + 15)$. Reference [4] simplifies $A(E_i, D_i)$ further by removing the lower bound on the sigmoid and writing $A(E_i, D_i) \approx (1/2) + \min(E_i - D_i, 15)/30$. Reference [8], on the other hand, makes the approximation $A(E_i, D_i) \approx 10^{E_i/10}/(10^{E_i/10} + 10^{D_i/10})$, which is a differentiable function. When writing the above expressions for the modified speech, the audibility-function approximations are concave functions of the (linear) spectral gain $10^{g_i/10}$. Optimizing the approximations subject to linear constraints on $10^{g_i/10}$ form



[FIG2] The computation of the SII.

straightforward optimization problems that can be solved using the Karush–Kuhn–Tucker conditions. The resulting analytic solutions are easy to implement. The later work of [8] models $A(E_i, D_i)$ more accurately at low SNR values and provides improved performance over the original work of [4] under low SNR conditions.

The discussion in this section assumed stationarity. Time variation can be accounted for by recursive updating of the equivalent spectrum levels E_i and D_i and periodically recomputing the gains g_i [4]. This is consistent with the SII update described in [27], which uses frequency-dependent temporal windows.

WORD-SEQUENCE PROBABILITY-BASED ENHANCEMENT

The section "Defining Intelligibility" identified the suitability of the expected probability of correct message recognition as a measure for optimizing intelligibility at a high level of abstraction. We noted in the section "Measures Operating on a Word Sequence" that, under an ergodicity assumption, the expectation over messages can be approximated by averaging over time. Optimizing a measure derived from the probability of correct recognition under a power constraint has been shown to provide significant intelligibility gain assuming that accurate sound segmentation information and an appropriate acoustic speech model are available [11]. We emphasize that the method assumes that ASR word probability tracks the human recognition performance, which was found to be true in [11] but is not guaranteed. Here we provide more detail about this approach.

TO MAKE HIGH-LEVEL MACHINE-BASED OPTIMIZATION FEASIBLE IN PRACTICE, WE CAN REPRESENT THE MESSAGE AT THE PHONEME LEVEL.

To make high-level machine-based optimization feasible in practice, we can represent the message at the phoneme level. This means we refine our Markov chain to include an intermediate level. The chain now becomes $\mathbf{M}_T \to \mathbf{u}_T \to \mathbf{a}_T \to \mathbf{a}_L \to \mathbf{u}_L \to \mathbf{M}_L$, where $\mathbf{u}_T$ and $\mathbf{u}_L$ denote the talker and lister phoneme sequences, respectively. By first performing time alignment of a sequence of acoustic features vectors $\mathbf{a}_T$ and a sequence of phonemes $\mathbf{u}_T$ by means of an ASR engine, a practical intelligibility enhancement approach can be defined. The ASR speech model can then be used to provide the probability densities that characterize clean speech sounds in the acoustic feature space.

To enhance intelligibility, we want to find the parameters C^* of our speech modification scheme that maximize the average probability that the listener interpreted phoneme sequence $\mathbf{u}_L$ is the talker-generated sequence $\mathbf{u}_T$:

$$C^* = \underset{C}{\operatorname{argmax}} \overline{p_{\mathbf{u}_L|\mathbf{u}_T}(u_T\,|\,u_T, C)}, \tag{14}$$

where the subscripts of the density label the density it represents. Note that the densities are consistent with the models shown in (4).

Simplifications were introduced in [11] to make the optimization tractable. It was tacitly assumed that the message is accurately represented by the phonemes and production noise was not formally considered. It was also assumed that v_E (the representation of the noise) can be approximated as deterministic, which is reasonable for typical acoustic signal representations and stationary noise. The only remaining uncertainty is due to the interpretation noise in the mapping from a_L to u_L. In an ASR system based on an HMM, this is modeled by the observation noise. Equation (14) can now be approximated by

$$C^* \approx \underset{C}{\operatorname{argmax}} \overline{p_{\mathbf{u}_L|\hat{\mathbf{a}}_L}(u_T\,|\,\hat{a}_L(u_T, C))} \tag{15}$$

$$= \underset{C}{\operatorname{argmax}} \overline{p_{\hat{\mathbf{a}}_L|\mathbf{u}_L}(\hat{a}_L\,|\,u_T, C)p_{\mathbf{u}_L}(u_T)} \\ \overline{\left(\textstyle\sum_{u'_T} p_{\hat{\mathbf{a}}_L|\mathbf{u}_L}(\hat{a}_L\,|\,u'_T, C)p_{\mathbf{u}_L}(u'_T)\right)^{-1}}, \tag{16}$$

where we used Bayes' rule and where $\hat{a}_L(\mathbf{u}_T, C)$, abbreviated to $\hat{\mathbf{a}}_L$, is the set of acoustic features observed by the listener, which is modeled as a deterministic function of the talker phoneme sequence u_T and the speech modification parameters C. The first term of (16) is the likelihood of the talker phoneme sequence for the observed features $\hat{\mathbf{a}}_L$, the second term is the a priori probability that the phoneme sequence $\mathbf{u}_T$ is decoded by the listener, and the third term is the inverse a priori probability of the listener-observed features. Optimization of the likelihood term only reduces complexity and provides good results [11].

The theory is simplest to implement if the sequences are considered stationary. The averaging of (16) over long time intervals (multiple sentences) is then preferred. In a practical implementation, shortcuts may have to be made due to requirements on delay and complexity and because the stationarity assumption may not be sufficiently accurate.

A system-level perspective of the proposed approach is shown in Figure 3. In [11], the approach was validated for a combination of two modifications: prosody-affecting phoneme gain adjustment and a spectral modification redistributing the signal energy across frequency bands. The method compared favorably to a method based on the optimization of a measure operating on a sequence of auditory states [14], discussed in the section "Measures Operating on a Sequence of Auditory States." Results reported in [9] suggest that using the full Bayesian approach rather than optimizing only the likelihood component of (16) improves performance.

In text-to-speech applications it may be possible to select from a set of phrases to convey a particular message. The measure given in (16) has also been used to determine the optimal phrasing of utterances [19]. This study indicates that maximizing the probability of correct interpretation of the phoneme sequence increases intelligibility. Considering prior information on the predictability of various formulations is expected to further enhance performance.

ENHANCEMENT OVER MULTIPLE SPATIAL POINTS

We have considered preprocessing techniques that do not consider the spatial aspects of the rendering scenario. In this section, we show that spatial aspects can also be exploited to enhance intelligibility. In announcement scenarios in public spaces such as airports, train stations, or shopping malls, environmental noise and reverberation contribute to a reduced intelligibility for the listeners. If different messages are communicated to different spatial regions, acoustic leakage between regions [16] exacerbates the problem. The impact on intelligibility is particularly large for hearing-impaired persons.

IN ANNOUNCEMENT SCENARIOS IN PUBLIC SPACES SUCH AS AIRPORTS, TRAIN STATIONS, OR SHOPPING MALLS, ENVIRONMENTAL NOISE AND REVERBERATION CONTRIBUTE TO A REDUCED INTELLIGIBILITY FOR THE LISTENERS.

Consider a scenario in a public environment where N messages are conveyed via the public address (PA) system to N listeners wearing a hearing instrument. A possibility is to downstream the corresponding signals directly to the listeners, but listeners often wear an open fit (nonoccluded) hearing instrument, where the direct signal also is mixed in at the eardrum. Instead of using direct downlink connections, it is possible to preprocess all speech signals jointly at the PA system so as to minimize the expected distortion at the eardrums of the listeners. The distortion measure can be based on any (mathematically well-behaved) model for speech quality or intelligibility, such as some of the models discussed in the section "Practical Measures of Intelligibility."

Let $\mathbf{a}_T = [a_{T,1}, a_{T,2}, \ldots, a_{T,N}]^T$, $\tilde{\mathbf{a}}_T$ and $\mathbf{a}_L$ (defined similarly) be the (complex-valued) short-time DFT coefficients of the source speech signals, enhanced signals (at the PA system), and received signals at the listeners, respectively. The signals $\mathbf{a}_L$ are captured by the microphones of the hearing instruments. For simplicity, we neglect production and interpretation noises of the section "Defining Intelligibility" and assume that degradations are purely acoustical and consist of noise, reverberation, and cross-talk between messages. It is easy to see that if we use stacked-vector notation for the signals $a_{T,i}$ and $a_{L,i}$, $i = 1, 2, \ldots, N$, upon preprocessing, all effects can be included in the affine signal model given by [16]

$$\mathbf{a}_L = \mathbf{H}_E \tilde{\mathbf{a}}_T + \mathbf{v}_E, \qquad (17)$$

where the channel matrix $\mathbf{H}_E$ collects all reverberation and cross-talk transfer coefficients between production and reception points, and $\mathbf{v}_E$ is additive noise in the environment.

Consider also a distortion measure $d(\mathbf{a}_T, \mathbf{a}_L)$, smooth (continuously differentiable) as a function of $\mathbf{a}_L$, which quantifies the distortion between the reference produced coefficients $\mathbf{a}_T$ and what is eventually listened to, $\mathbf{a}_L$. Our aim is to find the modification $\mathbf{a}_T \mapsto \tilde{\mathbf{a}}_T$ that minimizes the expected distortion according to d, jointly for all talker-listener points, i.e., we want to solve the optimization problem

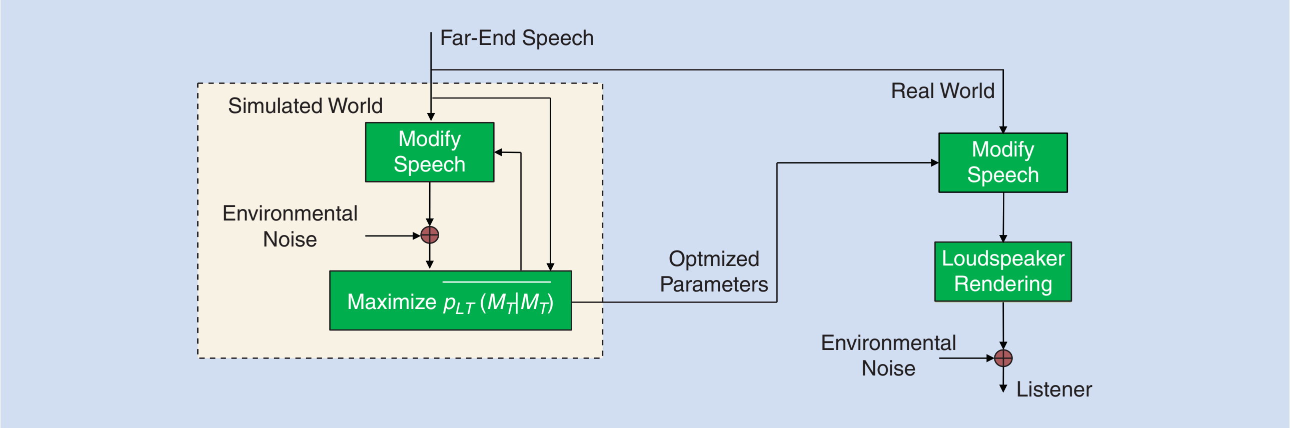
$$\underset{\tilde{\mathbf{a}}_T}{\text{minimize}}\ \mathrm{E}\,[d(\mathbf{a}_T, \mathbf{H}_E \tilde{\mathbf{a}}_T + \mathbf{v}_E)], \qquad (18)$$

where the expectation is taken only over the acoustic disturbances $\mathbf{H}_E$, $\mathbf{v}_E$, since we have direct access to the speech of the talker $\mathbf{a}_T$ and therefore take it to be deterministic.

Generic necessary conditions can be derived for solving (18) in terms of a functional description of the distortion measure d. The conditions are [16]

$$E\left[\mathbf{H}_E^{\mathrm{H}} \frac{\partial d}{\partial \mathbf{a}_L^*}(\mathbf{a}_T, \mathbf{H}_E \tilde{\mathbf{a}}_T + \mathbf{v}_E)\right] = 0, \qquad (19)$$

where $(\cdot)^{\mathrm{H}}$ is the Hermitian transpose, and $(\partial/\partial v^*) \equiv (1/2)((\partial/\partial v_{\Re}) - (1/j)(\partial/\partial v_{\Im}))$ is a complex differential operator, expressed in terms of the real differential operators $(\partial/\partial v_{\Re})$ and $(\partial/\partial v_{\Im})$, in Hessian (vertical) notation, with respect to the real and imaginary components of the variable v, respectively. The meaning of (19) is that, for optimality, it is required to choose the preprocessed speech $\tilde{\mathbf{a}}_T$ such as to make the complex gradient of the distortion measure with respect to the listener DFT bins in all zones orthogonal to all columns of the channel matrix $\mathbf{H}_E$.



[FIG3] The intelligibility enhancement using a phoneme-level measure.

To demonstrate the use of the optimality conditions (19), let us consider the simple ℓ_2 distortion measure given by

$$d(a_T, a_L) = \|a_L - a_T\|^2, \qquad (20)$$

where $\|\cdot\|$ is the ℓ_2 norm. In this case, (18) is a convex optimization problem, so that (19) are also sufficient conditions. By using the optimality conditions (19) under the assumption that H_E and v_E are uncorrelated, and including the hybrid deterministic-stochastic model for H_E introduced in [16], where the early response is described solely by a deterministic direct path and the late response is modeled by an exponentially fading stochastic process, the preprocessing algorithm is derived as

$$\tilde{a}_T = (D^H D + \Lambda)^{-1} D^H a_T, \qquad (21)$$

where D is a matrix collecting direct path responses of the channel, and Λ is a diagonal matrix collecting diffuse reverberation response channel energies. Note that in the case of low reverberation, $\Lambda \to 0$, the scheme (21) reduces to a conventional acoustic cross-talk canceler [30], $\tilde{a}_T = D^{-1} a_T$, which by compensating for the direct paths of the channel H_E, makes the cross-signals cancel out at the listeners. We thus conclude that optimization-based multipoint preprocessing enhancement as formulated in (18) leads to acoustic cross-talk cancelation, when applied to the ℓ_2 distortion measure (20).

[THE TECHNICAL OUTCOMES WILL LIKELY BECOME AN INTEGRAL PART OF SPEECH-RENDERING DEVICES IN THE NEAR FUTURE, LEADING TO IMPROVED COMMUNICATION AMONG HUMANS AND FROM MACHINES TO HUMANS.]

CONCLUSIONS AND OPEN PROBLEMS

Modern speech communication often leads to the signal being rendered by a machine in a noisy environment. In these circumstances, communication benefits from methods that make speech more intelligible in noise, particularly if the enhancement can adapt to the scenario at hand. This requires quantitative models of the communication process and distortion measures.

The use of a distortion measure facilitates the formulation of convergent algorithms and generally reduces the need for ad hoc solutions. Measures formulated at a high level of abstraction, such as (1) and (3) apply, at least in principle, to all communication tasks. However, when these high-level measures are applied to specific tasks assumptions must be made, either for the signal or for a model of the human cognitive system (e.g., by an ASR system), or both. Thus, optimization of any measure can never replace the need of extensive real-world testing to verify the performance of an intelligibility-enhancement system for the task at hand.

At first sight, the intelligibility-enhancement problem resembles the standard problem of transmission over a noisy channel. However, we have shown that the unprecise nature of the human production and interpretation must be accounted for. When that is done, standardized measures for intelligibility, which have a long history and were derived heuristically, are found to be consistent with communication theory.

While the field of intelligibility enhancement has developed rapidly, opportunities for significant improvement remain. Careful accounting for time-domain masking may improve performance. Methods developed for scenarios with additive noise only must be extended to include reverberation. Refining methods that perform spectral shaping to include range compression may increase their performance. For methods based on mutual information, the effect of time and frequency dependencies must be considered. Studies to determine the best representation (e.g., cepstra or DFT coefficients) and the determination and usage of appropriate noise distributions for the model likely will lead to improvement. The determination of a word choice for a message that is more robust to noise is an essentially unsolved task.

Although major challenges remain, the field of intelligibility enhancement has made major strides in recent years. The technical outcomes will likely become an integral part of speech-rendering devices in the near future, leading to improved communication among humans and from machines to humans.

AUTHORS

W. Bastiaan Kleijn (bastiaan.kleijn@ecs.vuw.ac.nz) received the Ph.D. degree in electrical engineering from Delft University of Technology, The Netherlands (TU Delft); an M.S.E.E. degree from Stanford University; and a Ph.D. degree in soil science and an M.Sc. degree in physics from the University of California, Riverside. He is a professor at Victoria University of Wellington, New Zealand, and TU Delft, The Netherlands (part-time). He was a professor and head of the Sound and Image Processing Laboratory at The Royal Institute of Technology (KTH), Stockholm, Sweden, from 1996 until 2010 and a founder of Global IP Solutions, a company that provided the original audio technology to Skype and was later acquired by Google. Before 1996, he was with the Research Division of AT&T Bell Laboratories in Murray Hill, New Jersey. He is an IEEE Fellow.

João B. Crespo (j.b.farinhapereiracrespo@student.vu.nl.) is a Ph.D. student in the Circuits and Systems Group of Delft University of Technology, The Netherlands. In 2009, he received his M.Sc. degree in electrical engineering from the Technical University of Lisbon, Portugal. During the last year of his M.Sc. studies, he was an exchange student at the Information and Communication Theory Group of Delft University of Technology. In 2010–2011, he worked at ExSilent B.V., The Netherlands, as a digital signal processing developer. His areas of interest include audio and speech processing, auditory perception, and information theory.

Richard C. Hendriks (R.C.Hendriks@tudelft.nl) obtained the M.Sc. and Ph.D. degrees (both cum laude) in electrical engineering from Delft University of Technology, The Netherlands, in 2003 and 2008, respectively. He was a Ph.D. researcher (2003–2007) and a postdoctoral researcher (2007–2010) at Delft University of Technology. In 2005, he was a visiting researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Germany, and in 2008–2009 he was a visiting researcher at Oticon A/S, Denmark. He is an assistant professor at Delft University of Technology. His main research interests include intelligibility improvement and digital speech processing in general.

Petko N. Petkov (petkov@kth.se) received the B.Sc. degree in communication engineering from the Technical University of Sofia, Bulgaria, and the M.Sc. and Ph.D degrees in electrical engineering from The Royal Institute of Technology (KTH) Stockholm, Sweden. He was a research and development engineer with Global IP Solutions from 2006 to 2007. He is currently with the Speech Technology Group, Cambridge Research Laboratory, Toshiba, working on speech intelligibility enhancement.

Bastian Sauert (bastian.sauert@head-acoustics.de) obtained both the Dipl.-Ing. and Dr.-Ing. degrees from RWTH Aachen University, Germany. In 2014, he joined HEAD acoustics, Herzogenrath, Germany. He was a researcher at the Institute of Communication Systems and Data Processing of RWTH Aachen University, Germany, where he studied the enhancement of speech intelligibility for listeners in a noisy environment. His focus was on optimizing objective speech intelligibility measures in noise with special consideration of the application in mobile phones. His main research interests are speech/audio processing, including noise suppression and near-end listening enhancement, as well as speech quality estimation.

Peter Vary (vary@rwth-aachen.de) received the Dipl.-Ing. degree in electrical engineering from the University of Darmstadt, Germany, in 1972 and the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Germany, in 1978. In 1980, he joined Philips Communication Industries, Nuremberg, Germany, where he became the head of the Digital Signal Processing Group. Since 1988, he has been a professor at RWTH Aachen University, Germany, and head of the Institute of Communication Systems and Data Processing. His main research interests are speech coding, joint source-channel coding, error concealment, and speech enhancement including noise suppression, acoustic echo cancellation, and artificial wideband extension. He is a Fellow of the IEEE.

REFERENCES

[1] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2059–2069, 2010.

[2] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Am.*, vol. 43, no. 1, pp. 81–86, 1968.

[3] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 24, no. 4, pp. 277–282, 1976.

[4] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachbericht-Sprachkommunikation*, 2010.

[5] T. C. Zorilc, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, Portland, OR, 2012, pp. 635–638.

[6] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, Portland, OR, 2012, pp. 955–958.

[7] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Elsevier Comput. Speech Lang.*, vol. 28, no. 2, pp. 543–571, 2014.

[8] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Lett.*, vol. 20, no. 3, pp. 225–228, 2013.

[9] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane challenge," in *Proc. Interspeech*, Lyon, France, 2013, pp. 3552–3556.

[10] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Commun.*, vol. 55, no. 4, pp. 572–585, 2013.

[11] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 5, pp. 1035–1045, 2013.

[12] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index," in *EURASIP European Signal Processing Conf. (EUSIPCO)*, 2009, vol. 17, pp. 1844–1848.

[13] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[14] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 858–872, 2014.

[15] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, Mar. 2015.

[16] J. B. Crespo and R. C. Hendriks, "Multizone speech reinforcement," *IEEE/ACM Trans. Audio, Speech, Lang. Processing*, vol. 22, no. 1, pp. 54–66, 2014.

[17] J. Allen, "How do humans process and recognize speech?" *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[18] *Methods for the Calculation of the Speech Intelligibility Index*, ANSI S3.5-1997.

[19] M. Zhang, P. N. Petkov, and W. B. Kleijn, "Rephrasing-based speech intelligibility enhancement," in *Proc. Interspeech*, Aug. 2013, pp. 3587–3591.

[20] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[21] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1553–1564, 2012.

[22] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Aug. 2011, pp. 1837–1840.

[23] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the Glimpse Proportion," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 665–686, 2014.

[24] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[25] K. D. Kryter, "Methods for the calculation and use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[26] G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, "A frequency importance function for continuous discourse," *J. Acoust. Soc. Amer.*, vol. 81, no. 4, pp. 1130–1138, 1987.

[27] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[28] H. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," in *Proc. Interspeech*, 2013, pp. 3577–3581.

[29] S. Elliott, J. Cheer, J.-W. Choi, and Y. Kim, "Robustness and regularization of personal audio systems," *IEEE Trans. Speech, Audio Lang. Processing*, vol. 20, pp. 2123–2133, Sept. 2012.

[30] D. B. Ward and G. W. Elko, "Virtual sound using loudspeakers: robust acoustic crosstalk cancellation," in *Acoustics Signal Processing for Telecom*, S. L. Gay and J. Benesty, Eds. Boston, MA: Kluwer Academic, 2000, ch. 14.

[SP]

[Timo Gerkmann, Martin Krawczyk-Becker, and Jonathan Le Roux]

Phase Processing for Single-Channel Speech Enhancement

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

[History and recent advances]

With the advancement of technology, both assisted listening devices and speech communication devices are becoming more portable and also more frequently used. As a consequence, users of devices such as hearing aids, cochlear implants, and mobile telephones, expect their devices to work robustly anywhere and at any time. This holds in particular for challenging noisy environments like a cafeteria, a restaurant, a subway, a factory, or in traffic. One way to making assisted listening devices robust to noise is to apply speech enhancement algorithms. To improve the corrupted speech, spatial diversity can be exploited by a constructive combination of microphone signals (so-called beamforming), and by exploiting the different spectrotemporal properties of speech and noise. Here, we focus on single-channel speech enhancement algorithms which rely on spectrotemporal properties. On the one hand, these algorithms can be employed when the miniaturization of devices only allows for using a single microphone. On the other hand, when multiple microphones are available, single-channel algorithms can be employed as a postprocessor at the output of a beamformer. To exploit the short-term stationary properties of natural sounds, many of these approaches process the signal in a time-frequency representation, most frequently the short-time discrete Fourier transform (STFT) domain. In this domain, the coefficients of the signal are complex-valued, and can therefore be represented by their absolute value (referred to in the literature both as STFT magnitude and STFT amplitude) and their phase. While the modeling and processing of the STFT magnitude has been the center of interest in the past three decades, phase has been largely ignored.

In this article, we review the role of phase processing for speech enhancement in the context of assisted listening and speech communication devices. We explain why most of the research conducted in this field used to focus on estimating spectral magnitudes in the STFT domain, and why recently phase processing is attracting increasing interest in the speech

Digital Object Identifier 10.1109/MSP.2014.2369251
Date of publication: 12 February 2015

enhancement community. Furthermore, we review both early and recent methods for phase processing in speech enhancement. We aim to show that phase processing is an exciting field of research with the potential to make assisted listening and speech communication devices more robust in acoustically challenging environments.

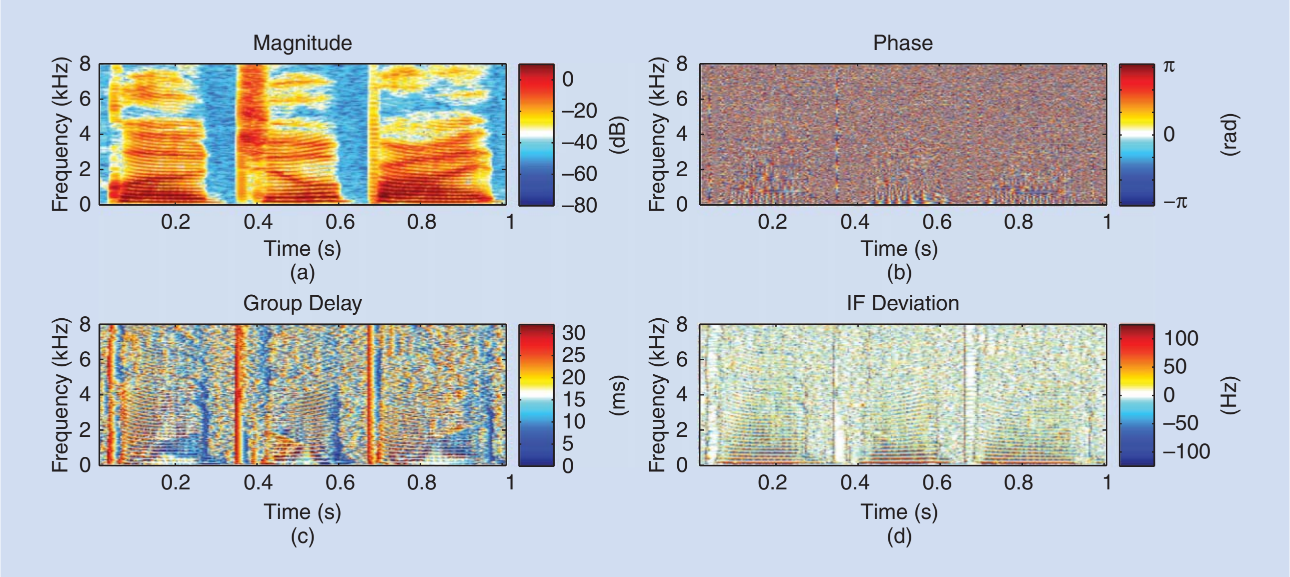
WITH THE ADVANCEMENT OF TECHNOLOGY, BOTH ASSISTED LISTENING DEVICES AND SPEECH COMMUNICATION DEVICES ARE BECOMING MORE PORTABLE AND ALSO MORE FREQUENTLY USED.

INTRODUCTION

Let us first consider the common speech enhancement setup consisting of STFT analysis, spectral modification, and subsequent inverse STFT (iSTFT) resynthesis. The analyzed digital signal $x(n)$, with time index n, is chopped into L segments with a length of N samples, overlapping by $N-R$ samples, where R denotes the segment shift. Each segment ℓ is multiplied with the appropriately shifted analysis window $w_a(n-\ell R)$ and transformed into the frequency domain by applying the discrete Fourier transform (DFT), yielding the complex-valued STFT coefficients $X_{k,\ell} \in \mathbb{C}$ for every segment ℓ and frequency band k. To compactly describe this procedure, we define the STFT operator: $X = \text{STFT}(\boldsymbol{x})$. Here, $\boldsymbol{x}$ is a vector containing the complete time-domain signal $x(n)$ and X is an $N \times L$ matrix of all $X_{k,\ell}$, which we will refer to as the *spectrogram*. Since we are interested in real-valued acoustic signals, we consider only complex symmetric spectrograms $X \in \mathcal{S} \subset \mathbb{C}^{N \times L}$, where $\mathcal{S}$ denotes the subset of spectrograms for which $X_{N-k,\ell} = \overline{X}_{k,\ell}$ for all ℓ and k, with $\overline{X}$ being the complex conjugate of X.

After some processing, such as magnitude improvement, is applied on the STFT coefficients, a modified spectrogram $\widetilde{X}$ is obtained. From $\widetilde{X}$ a time-domain signal can be resynthesized through an iSTFT operation, denoted by $\tilde{\boldsymbol{x}} = \text{iSTFT}(\widetilde{X})$. For this, the inverse DFT of the STFT coefficients is computed and each segment is multiplied by a synthesis window $w_s(n-\ell R)$; the windowed segments are then overlapped and added to obtain the modified time-domain signal. A final renormalization step is performed to ensure that, if no processing is applied to the spectral coefficients, there is perfect reconstruction of the input signal, i.e., $\text{iSTFT}(\text{STFT}(\boldsymbol{x})) = \boldsymbol{x}$. The renormalization term, equal to $\sum_{q=-\infty}^{+\infty} w_a(n+qR)\, w_s(n+qR)$, is R-periodic and can be included in the synthesis window. A common choice for both $w_a(n)$ and $w_s(n)$ is the square-root Hann window, which for overlaps such that $N/R \in \mathbb{N}$ (e.g., 50%, 75%, etc.) only requires normalization by a scalar. If the spectrogram is modified, using the same window for synthesis as for analysis can be shown to lead to a resynthesized signal whose spectrogram is closest to $\widetilde{X}$ in the least-squares sense [1]. This fact will turn out to be important for the iterative phase estimation approaches discussed later.

Until recently, in STFT-based speech enhancement, the focus was on modifying only the magnitude of the STFT components, because it was generally considered that most of the insight about the structure of the signal could be obtained from the magnitude, while little information could be obtained from the phase component. This would seem to be substantiated by Figure 1 when considering only (a) and (b), where the STFT magnitude (a) and STFT phase (b) of a clean speech excerpt are depicted. In contrast to the magnitude spectrogram, the phase spectrogram appears to show only little temporal and spectral regularities. There are nonetheless distinct structures inherent to the spectral phase, but they are hidden to a great extent because the phase is



(a)

(b)

(c)

(d)

[FIG1] (a) Magnitude spectrogram, (b) phase spectrogram, (c) group delay, and (d) IF deviation of the utterance ”glowed jewel-bright” using a segment length of 32 ms and a shift of 4 ms.

wrapped to its principle value, i.e., $-\pi \leq \phi^{X}_{k,\ell} = \angle X_{k,\ell} \leq \pi$. To reveal these structures, alternative representations have been proposed, which consider phase relations between neighboring time-frequency points instead of absolute phases. Two examples of such representations are depicted in Figure 1(c) and (d). In (c), the negative derivative of the phase along frequency, known as the *group delay*, is shown. It has been shown to be a useful tool for speech enhancement, e.g., by Yegnanarayana and Murthy [2]. Besides the group delay, the derivative of the phase along time, i.e., the instantaneous frequency (IF), also unveils structures in the spectral phase. For an improved visualization, in (d), we do not show the IF, but rather its deviation from the respective center frequency in Hz, which reduces wrapping along frequency [3], [4]. It is interesting to remark that the temporal as well as the spectral derivatives of the phase both reveal structures similar to those in the magnitude spectrogram in (a). Please note that both phase transformations are invertible and thus carry the same information as the phase itself. No additional prior knowledge has been injected.

The observed structures in the spectral phase can well be explained by employing models of the underlying signal, e.g., by sinusoidal models in the case of voiced speech [5]. Besides the structures in the phase that are caused by signal characteristics, neighboring time-frequency points also show dependencies due to the STFT analysis: first, because of the finite length of the segments, neighboring frequency bands are not independent; second, successive segments overlap and hence share partly the same signal information. This introduces particular spectrotemporal relations between STFT coefficients within and across frames of the spectrogram, regardless of the signal. If the spectrogram is modified, these relations are not guaranteed to be maintained and the modified spectrogram $\widetilde{X}$ may not correspond to the STFT of any time-domain signal anymore. As a consequence, the resynthesized signal may have a spectrogram $\mathcal{G}(\widetilde{X})$, where

$$\mathcal{G}(\widetilde{X}) := \mathrm{STFT}(\mathrm{iSTFT}(\widetilde{X})), \tag{1}$$

which is different from the desired spectrogram $\widetilde{X}$, as illustrated in Figure 2. Such spectrograms are called *inconsistent*, while *consistent* spectrograms verify $\mathcal{G}(X) = X$ and can be obtained from a time-domain signal.

Since the majority of speech enhancement approaches only modify the magnitude, the mismatch between the enhanced magnitude and the degraded phase will most likely lead to an inconsistent spectrogram. This implies that even if the estimated magnitudes $|\widetilde{X}|$ are optimal with respect to some objective function, the magnitude spectrogram of the synthesized time-domain signal is not, as $|\mathcal{G}(\widetilde{X})| \neq |\widetilde{X}|$ (where $|\cdot|$ denotes the element-wise absolute value). To maintain consistency, and thus also optimality, the STFT phase has to be taken into account as well.

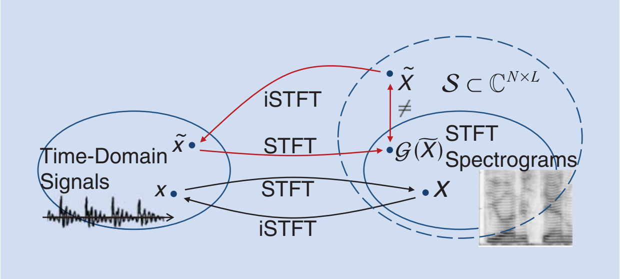
As a final illustration emphasizing the power of phase, it is interesting to remark that, from a particular magnitude spectrogram, it is possible to reconstruct virtually any time-domain signal with a carefully crafted phase. For instance, one can derive a magnitude spectrogram from that of a speech signal such that it yields either a speech signal similar to the original or a piece of rock music, depending on the choice of the phase. The point here is to exploit the inconsistency between magnitude and phase to make contributions of neighboring frames cancel each other just enough to reconstruct the energy profile of the target sound. Reconstruction is thus done up to a scaling factor, and quality is good albeit limited by dynamic range issues. An audio demonstration is available in http://www.jonathanleroux.org/research/LeRoux2011ASJ03_sound_transfer.html.

SPEECH ENHANCEMENT IN THE STFT DOMAIN

Speech enhancement is a field of research with a long-standing history. In this section, we will wrap up the different fields of research that have led to remarkable progress over the years. For a more detailed treatment and references to the original publications, see [6].

In the STFT domain, noisy spectral coefficients can, for instance, be improved using spectral subtraction or using minimum mean squared error (MMSE) estimators of the clean speech spectral coefficients [6, Ch. 4]. Examples of the latter are the Wiener filter as an estimator of the complex speech coefficients and the short-time spectral amplitude estimator [7]. These MMSE estimators are driven by estimates of the speech and noise power spectral densities (PSDs). The noise PSDs can be estimated in speech pauses as signaled by a voice activity detector, by searching for spectral minima in each subband, or based on the speech presence probability [6, Ch. 6]. With the noise PSD at hand, the speech PSD can be estimated by subtracting the noise PSD from the periodogram of the noisy signal. This has been shown to be the maximum likelihood (ML) optimal estimator of the clean speech PSD when considering isolated and independent time-frequency points and complex Gaussian distributed speech and noise coefficients [6, Sec. 4.2]. To reduce outliers, the ML speech PSD estimate is often smoothed, for instance, using the decision-directed approach [7] or more advanced smoothing techniques [6, Ch. 7].

Over the years, many improvements have been proposed resulting in a considerable progress thanks to better statistical models of speech and noise [6, Ch. 3], improved estimation of speech and noise PSDs [6, Ch. 6 and 7], combination with speech presence probability estimators [6, Ch. 5], and integration of perceptual models [6, Sec. 2.3.3]. Recent years have seen an explosion of interest in data-driven methods, with model-based approaches



[FIG2] An illustration of the notion of consistency.

such as nonnegative matrix factorization, hidden Markov models, and discriminative approaches such as deep neural networks. However, mainstream approaches have tended to ignore the phase, mainly due to the difficulty of modeling it and the lack of clarity about its importance, as discussed next.

RISE, DECLINE, AND RENAISSANCE OF PHASE PROCESSING FOR SPEECH ENHANCEMENT

The first proposals for noise reduction in the STFT domain arose in the late 1970s. While the spectral subtraction approaches only modified the spectral magnitudes, the role of the STFT phase was also actively researched at the time. In particular, several authors investigated conditions under which a signal is uniquely specified by only its phase or only its magnitude and proposed iterative algorithms for signal reconstruction from either one or the other (e.g., [1], [8], and references therein). For minimum or maximum phase systems, log-magnitude and phase are related through the Hilbert transform, meaning that only the spectral phase (or only the spectral magnitude) is required to reconstruct the entire signal. But the constraint of purely minimum or maximum phase is too restrictive for real audio signals, and Quatieri [8] showed that more constraints are needed for mixed-phase signals. For instance, imposing a causality or a finite-length constraint on the signal and specifying a few samples of the phase or the signal itself is in some cases sufficient to uniquely characterize the entire phase function from only the magnitude. Quatieri [8] also showed how to exploit such constraints to estimate a signal from its spectral magnitude: assuming some time-domain samples are known, and starting with an initial phase estimate and the known spectral magnitude, the signal is transformed to the time domain, where the given set of known samples is used to replace the corresponding time-domain samples. Then the time-domain signal is transformed back to the frequency domain, where the resulting magnitude is replaced by the known magnitude. This procedure is repeated for a certain number of iterations. In the case of the STFT domain, the correlation between overlapping short-time analysis segments can be exploited to derive similar iterative algorithms that do not require time-domain samples to be known. A popular example of such methods is that of Griffin and Lim (GL) [1], which we describe in more detail later along with more recent approaches. While algorithms such as GL can also be employed with magnitudes that are estimated rather than measured from an actual signal, the quality of the synthesized speech and the estimated phase strongly depends on the accuracy of the estimated speech spectral magnitudes and artifacts such as echo, smearing, and modulations may occur [9].

To explore the relevance of phase estimation for speech enhancement, Wang and Lim [10] performed listening experiments where the magnitude of a noisy speech signal at a certain signal-to-noise ratio (SNR) was combined with the phase of the same speech signal but distorted by noise at a different SNR. Listeners were asked to compare this artificial test stimulus to a noisy reference speech signal and to set the SNR of the reference such that the perceived quality was the same for the reference and the test stimulus. The result of this experiment was that the SNR gain obtained by mixing noisy magnitudes with a less distorted phase resulted in typical SNR improvements of 1 dB or less. Hence, Wang and Lim concluded that improving phase was not critical in speech enhancement [10]. Similarly, Vary [11] showed that only for local SNRs below 6 dB a certain roughness could be perceived if the noisy phase was kept unchanged. Finally, Ephraim and Malah [7] investigated the role of phase improvement from a statistical perspective: they showed that, under a zero-mean circular Gaussian speech and noise model and assuming that time-frequency points are mutually independent given the speech and noise PSDs, the MMSE estimate of the complex exponential of the speech phase has an argument equal to the noisy phase. Also, for more general models for the speech magnitudes with the same circularity assumption, it has been shown that the noisy phase is the ML optimal estimator of the clean speech phase, e.g., [12]. Note, however, that the independence assumption does not hold in general, and especially not for overlapping STFT frames, where part of the relationship is actually deterministic.

As a consequence of these observations, subsequent research in speech enhancement focused mainly on improving magnitude estimation, while phase estimation received far less attention for the next two decades. Even methods that considered phase, either by use of complex domain models, or by integrating out phase in log-magnitude-based models in a sophisticated way [13], ultimately used the noisy phase because of similar circularity assumptions.

However, as the performance of magnitude-only methods can only go so far without considering phase, and with the increase in computational power of assisted listening and speech communication devices, all options for improvements are back on the table. Therefore, researchers started reinvestigating the role of the STFT phase for speech intelligibility and quality [14], [15]. For instance, Kazama et al. [14] investigated the influence of the STFT segment length on the role of phase for speech intelligibility for a segment overlap of 50%. They found that, while for signal segments between 4 ms and 64 ms the STFT magnitude spectrum is more important than the phase spectrum, for segments shorter than 2 ms and segments longer than 128 ms, the phase spectrum is more important. These results are consistent with Wang and Lim's earlier conclusions [10]. To focus on practical applications, Paliwal et al. [15] investigated signal segments of 32 ms length, but in contrast to Wang and Lim [10] and Kazama et al. [14], they used a segment overlap of 7/8th instead of 1/2 in the STFT analysis, and they also zero-padded the time segments before computing the Fourier transform. With this increased redundancy in the STFT, the performance of existing magnitude-based speech enhancement can be significantly improved [15] if combined with enhanced phases. For instance, Paliwal et. al [15, case 4] report an improvement of 0.2 points of the mean opinion score (MOS) predicted by the instrumental "perceptual evaluation of speech quality" (PESQ) measure for white Gaussian noise at an SNR of 0 dB when combining an MMSE estimate of the clean speech magnitude with the oracle clean speech phase in a perfectly reconstructing STFT framework.

Paliwal et al.'s research confirmed the importance of developing and improving phase processing algorithms. This has recently been the focus of research by multiple groups. We now survey the main directions that have been investigated so far: better and faster phase estimation from magnitude, modeling of the signal phase, group delay and transient processing, and joint estimation of phase and magnitude.

ITERATIVE ALGORITHMS FOR PHASE ESTIMATION

Among the first proposals for phase estimation are iterative approaches, which aim at estimating a time-domain signal whose STFT magnitude is as close as possible to a target one [1], [8]. Indeed, if the STFT magnitude of two signals are close, the signals will in general be perceptually close as well. Thus, finding a signal whose STFT magnitude is close to a target one is considered a valid goal when looking to obtain a signal that "sounds" like that target magnitude. This motivated intense research on algorithms to estimate signals (or equivalently a corresponding phase) given target magnitudes, with applications such as speech enhancement or timescale modification. In the case of speech enhancement, the magnitude is typically obtained through one of the many magnitude estimation algorithms mentioned earlier, while some estimate of the phase, such as that of the noisy mixture, may further be exploited for initialization or as side information.

[FINDING A SIGNAL WHOSE STFT MAGNITUDE IS CLOSE TO A TARGET ONE IS CONSIDERED A VALID GOAL WHEN LOOKING TO OBTAIN A SIGNAL THAT "SOUNDS" LIKE THAT TARGET MAGNITUDE.]

The most well known and fundamental of these approaches is that of Griffin and Lim [1], which consists in applying STFT synthesis and analysis iteratively while retaining information about the updated phases and replacing the updated magnitudes by the given ones. This exploits correlations between neighboring STFT frames to lead to an estimate of the spectral phases and the time-domain signal.

Given a target magnitude spectrogram A, Griffin and Lim formulated the problem as that of estimating a real-valued time-domain signal x such that the magnitude of its STFT X is closest to A in the least-squares sense, i.e., estimating a signal x which minimizes the squared distance

$$d(x,A) = \sum_{k,\ell} \| X_{k,\ell} | - A_{k,\ell} |^2. \qquad (2)$$

They proposed an iterative procedure which can be proven to minimize, at least locally, this distance. Starting from an initial signal estimate $x^{(0)}$ such as random noise, iterate the following computations: compute the STFT $X^{(i)}$ of the signal estimate $x^{(i)}$ at step i; compute the phase estimate $\phi^{(i)}$ as the phase of $X^{(i)}$, $\phi^{(i)} = \angle X^{(i)}$; compute the signal estimate $x^{(i+1)}$ at step $i+1$ as the iSTFT of $Ae^{j\phi^{(i)}}$. Using the operator $\mathcal{G}$ defined in (1), this can be reformulated as

$$\phi^{(i+1)} = \angle \mathcal{G}(Ae^{j\phi^{(i)}}). \qquad (3)$$

This procedure can be proven to be nonincreasing as well for a measure of inconsistency of the spectrogram $Ae^{j\phi^{(i)}}$ defined directly in the time-frequency domain:

$$\mathcal{I}(\phi) = \| \mathcal{G}(Ae^{j\phi}) - Ae^{j\phi} \|_2^2. \qquad (4)$$

Indeed, one can easily show that $d(x^{(i+1)}, A) \leq \mathcal{I}(\phi^{(i)}) \leq d(x^{(i)}, A)$. Interestingly, if only parts of the phase are updated according to (3), the nondecreasing property still holds for $\mathcal{I}(\phi)$, but whether it still does for $d(x, A)$ has not been established.

Due to the extreme simplicity of its implementation and to its perceptually relatively good results, GL was used as the standard benchmark and a starting point for multiple extensions in the three decades that have followed, even after better and only marginally more involved algorithms had been devised. Most of the algorithms that have been developed since attempted to fix GL's issues, of which there are several: first, convergence typically requires many iterations; second, GL does not provide a good initial estimate, starting from random phases with no considerations for cross-frame dependencies; third, the updates rely on computing STFTs, which are computationally costly even when implemented using fast Fourier transforms (FFTs); fourth, the updates are typically performed on whole frames, without emphasis on local regularities; and finally, the original version of GL processes signals in batch mode.

On this last point, it is interesting to note that Griffin and Lim did actually hint at how to modify their algorithm to use it for online applications. They described briefly in [1] and with more details in [16] how to sequentially update the phase using "cascaded processors" that each take care of one iteration; their particular proposal however still incurs an algorithmic delay of I times the window length if performing I iterations. In [16], Griffin also presented several methods that he referred to as "sequential estimation methods": these only incur a single frame delay and could thus be used for online application, the best performing one being reported as on par with batch GL.

While one can already see in Griffin's account [16] several elements to modify GL into an algorithm that can lead to high quality reconstruction in a real-time setting, such as sliding-block analysis across the signal and the use of windows that compensate for partially reconstructed frames, these ideas seem to have gone largely unnoticed and it is not until much later that they were incorporated into more refined methods. Beauregard, Zhu, and Wyse proposed consecutively two algorithms for real-time signal reconstruction from STFT magnitude, the real-time iterative spectrogram inversion (RTISI) algorithm and RTISI with look ahead (RTISI-LA) [17]. RTISI aims at improving the original batch GL in two respects: allowing for online implementation, and generating better initial phase estimates. The algorithm considers the frames sequentially in order, and at frame ℓ, it only uses information from the current frame's magnitude and the previous overlapping frames. The initial phase estimate $\phi_\ell^{(0)}$ for frame ℓ is obtained as the phase of the partial reconstruction from the previous frames, windowed by an analysis window, which already ensures some consistency between the phases of the current and previous frames. An iterative procedure similar to GL is then applied, limited to the current frame's phase: at each iteration, frame ℓ's contribution to the signal is obtained by the inverse DFT of the phase $\phi_\ell^{(i)}$ combined with the target magnitude; frame ℓ's contribution is then combined by overlap-add to the contribution of the previous frames, leading to a signal estimate for frame ℓ; the phase $\phi_\ell^{(i+1)}$ is estimated as the phase of this signal estimate to which the analysis window is applied.

RTISI does lead to better results than GL for the first few iterations, but it quickly reaches a plateau and is ultimately significantly outperformed by GL. This is mainly due to the fact that RTISI does not consider information from future frames at all, even though the contribution of these future frames will later on be added to that of the past and current frames, effectively altering the estimation performed earlier. Its authors thus proposed an extension to RTISI including an M frame look-ahead, RTISI-LA. Instead of considering only the current frame as active, RTISI-LA performs GL-type updates on the phases in a block of multiple frames. The contribution of future frames outside the block is discarded during the updates, because the absence of a reliable phase estimate for them is regarded as likely to make their contribution more of a disturbance than a useful clue. This creates an asymmetry, which Zhu et al. [17] proposed to partially compensate by using asymmetric analysis windows with a reverse effect. Although the procedure relies on heuristic considerations, the authors show that it leads to much better performance than GL for a given number of iterations per block.

While RTISI and RTISI-LA were successful in overcoming GL's issues regarding online processing and poor initialization, they did not tackle the problems of heavy reliance on costly FFT computations and lack of care for local regularities in the time-frequency domain. Solving these problems was difficult in the context of classical approaches relying on enforcing constraints both in the time-frequency domain (to impose a given magnitude) and the time domain (to ensure that magnitude and phase are consistent), because they inherently had to go back and forth between the two domains, processing whole frames at a time. A solution was proposed by Le Roux et al. [18], whose key idea was to bypass the time domain altogether and reformulate the problem within the time-frequency domain. The standard operation of classical iterative approaches, i.e., computing the STFT of the signal obtained by iSTFT from a given spectrogram, can indeed be considered as a linear operator in the time-frequency domain. Le Roux et al. noticed that the result of that operation at each time-frequency bin can be well approximated by a local weighted sum (LWS) with complex coefficients on a small neighborhood of that bin in the original spectrogram. While the very small number of terms in the sum does not suffice to reduce the complexity of the operation compared to using FFTs, the locality of the sum opens the door to selectively updating certain time-frequency bins, as well as to immediately propagating the updated value for a bin in the computations of its neighbors' updates. Taking advantage of the sparseness of natural sound signals, Le Roux et al. showed in particular that focusing first on updating only the bins with high energy not only reduced greatly the complexity of each iteration, but also could lead to better initializations, the high energy regions serving as anchors for lower energy ones. While the LWS algorithm was originally proposed as an extension to GL for batch-mode computations, the authors later showed that it could be effectively used in online mode as well in combination with RTISI-LA [19]. Interestingly, a different prioritization of the updates based on energy, at the frame level instead of the bin level, was also successfully used by Gnann and Spiertz to improve RTISI-LA [20].

Recently, several authors investigated signal reconstruction from magnitudes with specific task-related side information. Those developed in the context of source separation are of particular interest to this article. Gunawan and Sen [21] proposed the multiple input spectrogram inversion (MISI) algorithm to reconstruct multiple signals from their magnitude spectrograms and their mixture signal. The phase of the mixture signal acts as very powerful side information, which can be exploited by imposing that the reconstructed complex spectrograms add up to the mixture complex spectrogram when estimating their phases, leading to much better reconstruction quality than in situations where the mixture signal is not available. Sturmel and Daudet's partitioned phase retrieval (PPR) method [9] also handles the reconstruction of multiple sources. Their proposal was to reconstruct the phase of the magnitude spectrogram obtained by Wiener filtering by applying a GL-like algorithm, which keeps the mixture phase in high SNR regions as a good estimate for the corresponding source and only updates the phase in low- to mid-SNR regions. Both methods, however, only modify the phase of the sources, and thus implicitly assume that the input magnitude spectrograms are close to the true source spectrograms, which is not realistic in general in the context of blind or semiblind source separation. Sturmel and Daudet proposed to extend MISI to allow for modifications of both the magnitude and phase, leading to the informed source separation using iterative reconstruction (ISSIR) method [22], and showed that it is efficient in the context of informed source separation where a quantized version of the oracle magnitude spectrograms is available. Methods to jointly estimate phase and magnitude for blind source separation and speech enhancement will be presented later.

SINUSOIDAL MODEL-BASED PHASE ESTIMATION

In contrast to the iterative approaches presented in the previous section, sinusoidal model-based phase estimation [4] does not require estimates of the clean speech spectral magnitudes. Instead, the clean spectral phase is estimated using only an estimate of the fundamental frequency, which can be obtained from the degraded signal. However, since usage of the sinusoidal model is reasonable only for voiced sounds, these approaches do not provide valid spectral phase estimates for unvoiced sounds, like fricatives or plosives.

For a single sinusoid, $\sin(\Omega n + \varphi)$, with normalized angular frequency Ω, the phase difference between two samples $n_2 = n_1 + R$ is given by $\Delta\phi = \phi(n_2) - \phi(n_1) = \Omega R$. For a harmonic signal, H sinusoids at integer multiples of the normalized angular fundamental frequency Ω_0, i.e., $\Omega^h = (h+1)\Omega_0 \in [0, 2\pi)$, are present at the same time:

$$s(n) = \sum_{h=0}^{H-1} A^h(n) \cos(\Omega^h(n) \cdot n + \varphi^h), \tag{5}$$

with real-valued amplitude A^h and initial time-domain phase φ^h for harmonic component h. Due to the fixed relation between the frequencies, (5) is also referred to as the *harmonic model*, which is a special case of the more general sinusoidal model. The harmonic frequencies and amplitudes are assumed to be slowly changing over time with respect to the length N of an STFT signal segment and we define $A_\ell^h = A^h(\ell R + N/2)$ and $\Omega_\ell^h = \Omega^h(\ell R + N/2)$ as the representative harmonic amplitudes and frequencies for the ℓth signal segment.

In speech enhancement, the sinusoidal model has, for instance, been employed in [23], where the model parameters are iteratively estimated from a noisy observation in the STFT domain, and the enhanced signal is synthesized using (5). In the absence of noise, synthesis results are reported to be almost indistinguishable from the clean speech signal, underlining the capability of (5) to accurately model voiced human speech. In contrast to [23], we now discuss how the sinusoidal model (5) can be employed to directly reconstruct the STFT phase. If the frequency resolution of the STFT is high enough to resolve the harmonic frequencies Ω^h in (5), in each frequency band k only a single harmonic component is dominant. The normalized angular frequency Ω_ℓ^h of the harmonic that dominates frequency band k is denoted as

$$\widetilde{\Omega}_{k,\ell} = \underset{\Omega_\ell^h}{\mathrm{argmin}}\{|2\pi k/N - \Omega_\ell^h|\}, \qquad (6)$$

i.e., the harmonic frequency that is closest to the center frequency $2\pi k/N$ of the kth frequency band. Interpreting the STFT of a signal as the output of a complex filter bank subsampled by the hop size R, the spectral phase $\phi_{k,\ell}^{\mathrm{S}}$ changes from segment to segment according to

$$\phi_{k,\ell}^{\mathrm{S}} = \underset{2\pi}{\mathrm{mod}}\,(\phi_{k,\ell-1}^{\mathrm{S}} + \widetilde{\Omega}_{k,\ell} R) = \underset{2\pi}{\mathrm{mod}}(\phi_{k,\ell-1}^{\mathrm{S}} + \Delta\phi_{k,\ell}^{\mathrm{S}}), \qquad (7)$$

where the modulo operator $\underset{2\pi}{\mathrm{mod}}(\cdot)$ wraps the phase to values between 0 and 2π.

When the clean signal $s(n)$ is deteriorated by noise, the spectral phases and thus the temporal phase differences $\Delta\phi_{k,\ell}^{\mathrm{S}}$ are deteriorated as well. With an estimate of the fundamental frequency at hand, however, the temporal phase relations in each band can be restored using (7) recursively from segment to segment.

Almost 50 years ago, a similar approach for the propagation of the spectral phase along time was taken in the phase vocoder [5] for time-scaling or pitch-shifting of acoustic signals. The temporal STFT phase difference is modified according to

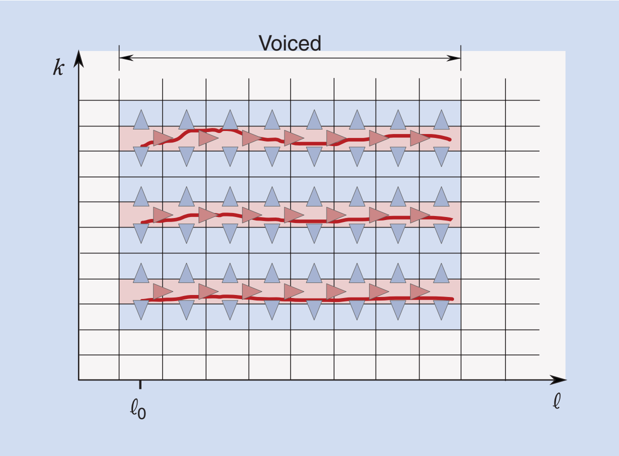
$$\hat{\phi}_{k,\ell}^{\mathrm{S}} = \hat{\phi}_{k,\ell-1}^{\mathrm{S}} + \alpha\Delta\phi_{k,\ell}^{\mathrm{S}}, \qquad (8)$$

where in this context, $\Delta\phi_{k,\ell}^{S}$ is often referred to as the IF. By scaling $\Delta\phi_{k,\ell}^{S}$ with the positive real-valued factor α, the IF of the signal component is either increased $(\alpha > 1)$ or decreased $(\alpha < 1)$. Comparing (7) to (8), the phase estimation along time for speech enhancement can be expressed in terms of a phase vocoder with a scaling factor of $\alpha = 1$. However, the application is completely different: instead of deliberately modifying the original phase, the clean speech phase is estimated from a noisy observation. It is worth noting that for the original phase vocoder, in contrast to phase estimation in speech enhancement, no fundamental frequency estimate is needed, as the phase difference $\Delta\phi_{k,\ell}^{\mathrm{S}} = \phi_{k,\ell}^{\mathrm{S}} - \phi_{k,\ell-1}^{\mathrm{S}}$ can be taken directly from the clean original signal.

For an accurate estimation of the clean spectral phase along segments using (7) a proper initialization is necessary [4]. In voiced sounds, the bands between spectral harmonics contain only little signal energy and, in the presence of noise, these bands are likely to be dominated by the noise component, i.e., $\phi_{k,\ell}^{\mathrm{Y}} \approx \phi_{k,\ell}^{\mathrm{N}}$, where $\phi_{k,\ell}^{\mathrm{Y}}$ and $\phi_{k,\ell}^{\mathrm{N}}$ are the spectral phases of the noisy mixture and the noise, respectively. Even though the phase might be set consistent within each band, the spectral relations across frequency bands are distorted already at the initialization stage. Directly applying (7) to every frequency band therefore does not necessarily yield phase estimates that could be employed for phase-based speech enhancement [4].

In the phase vocoder, this problem can be alleviated by aligning phases of neighboring frequency bands relative to each other, which is known as *phase locking*, e.g., [24]. There, the phase is evolved along time only in frequency bands that directly contain harmonic components. The phase in the surrounding bands, which are dominated by the same harmonic, is then set relative to the modified phase. For this, the spectral phase relations of the original signal are imposed on the modified phase spectrum.

In the context of speech enhancement, the same principle has been incorporated to improve the estimation of the clean speech spectral phase [4]. However, since only a noisy signal is observed, the clean speech phase relations across frequency bands are not readily available. To overcome this limitation, again the sinusoidal model is employed. The spectrum of a harmonic signal segment is given by the cyclic convolution of a comb-function with the transfer function of the analysis window, which causes spectral leakage. The spectral leakage induces relations not only between the amplitudes, but also between the phases of neighboring bands. It can be shown that phases of bands that are dominated by the same



[FIG3] Symbolic spectrogram illustrating the sinusoidal model-based phase estimation [4]. Starting from the noisy phase at the onset of a voiced sound in segment ℓ_0, in bands containing harmonic components (red) the phase is estimated along segments. Based on the temporal estimates, the spectral phase of bands between the harmonics (blue) is then inferred across frequency.

THE PHASE OF TRANSIENT SOUNDS IS NOT ONLY RELEVANT FOR DETECTION, BUT ALSO FOR THE REDUCTION OF TRANSIENT NOISE.

harmonic are directly related to each other through the phase response of the analysis window ϕ_k^{W}; see, e.g., [4] for more details. Accordingly, starting from a phase estimate at a band that contains a spectral harmonic, possibly obtained using (7), the phase of the surrounding bands can be inferred by accounting for the phase shift introduced by the analysis window. For this, only the fundamental frequency and the phase response ϕ_k^{W} are required, of which the latter can be obtained offline either from the window's discrete-time Fourier transform (DTFT) or from its DFT with a large amount of zero padding. The complete setup of [4] is illustrated in Figure 3.

It can be argued that for speech enhancement, the phase reconstruction across frequency bands between harmonics is more important than the temporal reconstruction on the harmonics: on the one hand, the local SNR in bands that directly contain harmonics is rather large for many realistic SNR situations, i.e., $\phi_{k,\ell}^{\mathrm{Y}} \approx \phi_{k,\ell}^{\mathrm{S}}$. Thus, the temporal alignment of the harmonic components is maintained rather well in the noisy signal. Further, the noisy phase $\phi_{k,\ell}^{\mathrm{Y}}$ in these bands typically yields a good starting point for the phase reconstruction along frequency. On the other hand, frequency bands between harmonics are likely to be dominated by the noise, i.e., $\phi_{k,\ell}^{\mathrm{Y}} \approx \phi_{k,\ell}^{\mathrm{N}}$, and the clean phase relations across bands are strongly disturbed. Here, the possible benefit of the phase reconstruction is much larger.

Even though the employed model is simple and limited to purely voiced speech sounds, the obtained phase estimates yield valuable information about the clean speech signal that can be employed for advanced speech enhancement algorithms. Interestingly, even the sole enhancement of the spectral phase can lead to a considerable reduction of noise between harmonic components of voiced speech after overlap-add [4]. This is because the speech components of successive segments are adding up constructively after the phase modifications, while the noise components suffer from destructive interference, since the phase relations of the noise have been destroyed. However, speech distortions are also introduced, which are substantially reduced when the estimated phase is combined with an enhanced magnitude, as, e.g., in [25]. Besides its value for signal reconstruction, the estimated phase can also be utilized as additional information for phase-aware magnitude estimation [25] and even for the estimation of clean speech complex coefficients [12], which will be discussed in more detail later.

GROUP DELAY AND TRANSIENT PROCESSING

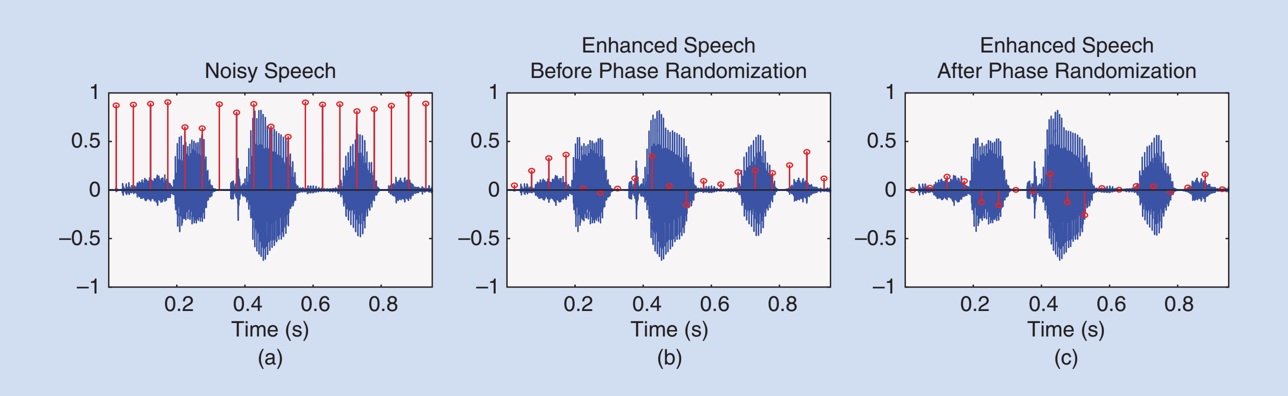
Structures in the phase are not limited to voiced sounds, but are also present for other sounds, like impulses or transients. These structures are well captured by the group delay, which can be seen in Figure 1(c), rendering it a useful representation for phase processing. For example, the group delay has been employed to facilitate clean speech phase estimation in phase-sensitive noise reduction [26]. It can be shown geometrically that if the spectral magnitudes of speech and noise are known, only two possible combinations of phase values remain, both of which perfectly explain the observed spectral coefficients of the mixture. In [26] (and the references therein), Mowlaee and Saedi proposed to solve this ambiguity by choosing the phase combination that minimizes a function of the group delay.

Besides phase estimation, the group delay has successfully been employed for the detection of transients sounds, such as sounds of short duration and speech onsets. To illustrate the role of the phase for transient sounds, let us consider a single impulse as the simplest example. The DFT of such a pulse is $Ae^{-\mathrm{j}2\pi\frac{n_0 k}{N}}$, where n_0 is the shift of the peak relative to the beginning of the current segment and A denotes the spectral magnitude. Hence, we observe a linear phase with a constant slope of $-2\pi\,(n_0/N)$. For impulsive signals, we accordingly expect a phase difference across frequency bands that is approximately constant, i.e., a constant group delay. That this is the case also for real speech sounds can be seen in Figure 1(c), where transient sounds show vertical lines with almost equal group delay.

For the detection of impulsive sounds, in [27] a linearity index $\mathrm{LI}_\phi(k)$ is defined, which measures the deviation of the observed phase difference across frequencies to the one that is expected for an impulse at n_0, i.e., $-2\pi(n_0/N)$. The observed phase differences are weighted with the spectral magnitude and averaged over frequency to obtain an estimate of the time domain offset n_0. Only if $\mathrm{LI}_\phi(k)$ is close to zero, i.e., the observed phase fits well to the expected linear phase, an impulsive sound is detected. The detection can be made either at a segment level or for each time-frequency point separately. While the former states if an impulsive sound is present in the current signal segment or not, the latter allows to localize frequency regions that are dominated by an impulsive sound, such as a narrowband onset.

Apart from the group delay, the IF, which corresponds to the temporal derivative of the phase, has also been employed for the detection of transient sounds, e.g., in [28] and the references therein. For steady-state signals, like voiced sounds, the IF is changing only slowly over time, due to the temporal correlation of the overlapping segments. When a transient is encountered, however, the most current segment differs significantly from previous segments and thus the IF also changes abruptly. This can be observed in Figure 1(d), where at speech onsets thin vertical lines appear in the IF deviation. Hence, the change of the IF from segment to segment—and its distribution—allow for the detection of transient sounds, such as note onsets [28].

The phase of transient sounds is not only relevant for detection, but also for the reduction of transient noise. In low SNR time-frequency regions, the observed noisy phase is close to the approximately linear phase of the transient noise. This can lead to artifacts in the enhanced signal if only the spectral magnitude is improved and the noisy phase is used for signal reconstruction: usage of the phase of the transient noise reshapes the enhanced time-domain signal in an uncontrolled way, such that it may again depict an undesired transient behavior. Even for a perfect magnitude estimate, the interfering noise is not perfectly suppressed if the phase



[FIG4] **(a) Speech degraded by a click train. (b) Signal obtained by combination of the clean speech spectral magnitude with the noisy phase. (c) Signal after supplemental phase randomization. Samples that contain a click are highlighted in red.**

is not processed alongside. To illustrate this, let us consider a speech signal degraded by an impulse train with a period length of T_0, which is nonzero every $N_0 = T_0 f_s$ samples. In Figure 4, the noisy signal (a) is presented together with the result obtained when combining the true clean speech STFT magnitudes with the noisy phase (b). Even though the clean magnitude is employed, which represents the best possible result for phase-blind magnitude enhancement, the time-domain signal still depicts residual impulses, which are caused by the noisy phase. In regions where the enhanced spectral magnitude is close to zero, i.e., in speech absence, the phase is not relevant and the peaks are well suppressed. During speech presence, however, the spectral magnitude is nonzero and the phase becomes important. Accordingly, the residual impulses are most prominent in regions with some speech energy at low local SNRs, where the noisy phase is close to the phase of the impulsive noise.

Recently, Sugiyama and Miyahara proposed the concept of phase randomization to overcome this issue; see, e.g., [27] and references therein. First, time-frequency points that are dominated by speech are identified by finding spectral peaks in the noisy signal. These peaks are excluded from the phase randomization to avoid speech distortions. To further narrow down time-frequency regions where randomization of the spectral phase is sensible, phase-based transient detection can be employed as well [27]. Then, the spectral phase in bins classified as dominated by transient noise is randomized by adding a phase term that is uniformly distributed between $-\pi$ and π. In this way, the approximately linear phase of the dominant noise component is neutralized. The effect of phase randomization is depicted in Figure 4(c), where a perfect magnitude estimate is combined with the modified phase for signal reconstruction. It can be seen that the residual peaks that are present when the noisy phase is employed are strongly attenuated, showing that phase randomization can indeed lead to a considerable increase of noise reduction, especially in low local SNRs. It is interesting to note that while the previously described iterative and sinusoidal model-based approaches aim at estimating the phase of the clean speech signal, the phase randomization approach merely aims at reducing the impact of the phase of the noise on the enhanced speech signal. Although the presented example is just a simple toy experiment, it still highlights the potential of phase randomization toward an improved suppression of transient noise, which has also been observed for real-world impulsive noise, like tapping noise on a touchscreen [27].

RELATION BETWEEN PHASE- AND MAGNITUDE ESTIMATION

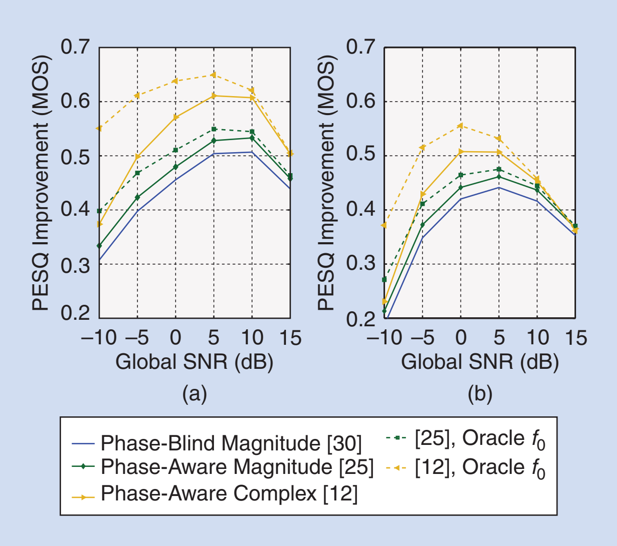
So far, we have discussed phase estimation using iterative approaches, sinusoidal model-based approaches, and group delay approaches; we now address the question of how STFT phase estimation can best be employed to improve speech enhancement. The most obvious way to do this is to combine enhanced speech spectral magnitudes in the STFT domain with the estimated or reconstructed STFT phases. It is interesting to note that Wang and Lim [10] already stated that obtaining a more accurate phase estimate than the noisy phase is not worth the effort "if the estimate is used to reconstruct a signal by combining it with an independently estimated magnitude [...]. However, if a significantly different approach is used to exploit the phase information such as using the phase estimate to further improve the magnitude estimate, then a more accurate estimation of phase may be important" [10]. However, at that point it was not clear how a phase estimate could be employed to improve magnitude estimation.

Gerkmann and Krawczyk [25] derived an MMSE estimator of the spectral magnitude when an estimate of the clean speech phase is available, referred to as *phase-sensitive* or *phase-aware* magnitude estimation. They were able to show that the information of the speech spectral phase can be employed to derive an improved magnitude estimator that is capable of reducing noise outliers that are not tracked by the noise PSD estimator. In babble noise, in a blind setup, the PESQ MOS can be improved by 0.25 points in voiced speech at 0 dB input SNR [25]. Further experimental results are given in the following section.

Instead of estimating phase and magnitude separately, one may argue that they should ideally be jointly estimated. The first step in this direction was proposed by Le Roux and Vincent [29] and references therein in the context of Wiener filtering for speech enhancement. As a classical Wiener filter only changes the magnitudes in the STFT domain, the modified spectrum $\widetilde{X}$ is inconsistent, meaning that $\mathrm{STFT}(\mathrm{iSTFT}(\widetilde{X})) \neq \widetilde{X}$. In contrast to this, in [29] the relationship between STFT coefficients across time and frequency is taken into account, leading to the consistent Wiener filter [29], which modifies both the magnitude and the phase of the noisy observation to obtain the separated speech. Wiener filter optimization is formulated as a maximum a posteriori problem under Gaussian assumptions, and a consistency-enforcing term is added either through a hard constraint or a soft penalty. Optimization is respectively performed directly on the signal in the time domain or jointly on phase and magnitude in the complex time-frequency domain, through a conjugate gradient method with a well-chosen preconditioner. Thanks to this joint optimization, the consistent Wiener filter was shown to lead to an improved separation performance compared to the classical Wiener filter and other methods that attempt to use phase information in combination with variance estimates [9], [21], [22], in an oracle scenario as well as in a blind scenario where the speech spectrum is obtained by spectral subtraction from a stationary estimate of the noise spectrum.

To combine phase-sensitive magnitude estimation and iterative approaches, Mowlaee and Saeidi [26] proposed placing the phase-sensitive magnitude estimator into the loop of an iterative approach that enforces consistency. Starting with an initial group-delay-based phase estimate, they proposed to estimate the clean speech spectral magnitude using a phase-sensitive magnitude estimator similar to [25]. After computing the iSTFT and the STFT they reestimated the clean speech phase, and from this reestimate the magnitudes. With this approach, convergence is reached after only few iterations.

WHEN AN INITIAL PHASE ESTIMATE IS ALSO EMPLOYED AS UNCERTAIN PRIOR INFORMATION WHEN IMPROVING THE SPECTRAL PHASE AS PROPOSED IN THE PHASE-AWARE COMPLEX ESTIMATOR CUP, THE PERFORMANCE CAN BE IMPROVED FURTHER.



[FIG5] The PESQ improvement over the noisy input. The results are averaged over four noise types. Evaluated (a) on voiced speech and (b) on the entire signal.

Another way to jointly estimate magnitudes and phases is to derive a joint MMSE estimator of magnitudes and phases directly in the STFT domain when an uncertain initial phase estimate is available. This phase-aware complex estimator is referred to as the *complex estimator with uncertain phase* (*CUP*) [12]. The initial phase estimate can be obtained by an estimator based on signal characteristics, such as the sinusoidal model-based approach [4]. Using this joint MMSE estimator [12], no STFT iterations are required. The resulting magnitude estimate is a nonlinear tradeoff between a phase-blind and a phase-aware magnitude estimator, while the resulting phase is a tradeoff between the noisy phase and the initial phase estimate. These tradeoffs are controlled by the uncertainty of the initial phase estimate, avoid processing artifacts, and lead to an improvement in predicted speech quality [12]. Experimental results for the CUP estimator are given in the following section.

EXPERIMENTAL RESULTS

In this section, we demonstrate the potential of phase processing to improve speech enhancement algorithms. To focus only on the differences due to the incorporation of the spectral phases, we choose algorithms that employ the same statistical models and PSD estimators: for the estimation of the noise PSD we choose the speech presence probability-based estimator with fixed priors (see [6, Sec. 6.3] and references therein) while for the speech PSD we choose the decision-directed approach [7]. We assume a complex Gaussian distribution for the noise STFT coefficients and a heavy-tailed χ-distribution for the speech magnitudes. Furthermore, we use an MMSE estimate of the square root of the magnitudes to incorporate the compressive character of the human auditory system. These models are employed in the phase-blind magnitude estimator [30], the phase-aware magnitude estimator [25], and the phase-aware CUP [12]. We use a sampling rate of 8 kHz and 32 ms spectral analysis windows with 7/8th overlap to facilitate phase estimation. To assess the speech quality, we employ PESQ as an instrumental measure that has been originally proposed for speech coding applications but has been show to correlate with subjective listening tests also for enhanced speech signals. The results are averaged over pink noise modulated at 0.5 Hz, stationary pink noise, babble noise, and factory noise, where the latter three are obtained from the NOISEX-92 database. To have a fair balance between male and female speakers, per noise type, the first 100 male and the first 100 female utterances from dialect region 6 of the Texas Instruments and Massachusetts Institute of Technology (TIMIT) training database are employed. The initial phase estimate is obtained based on a sinusoidal model [4], which only yields a phase estimate in voiced speech. The fundamental frequency is estimated using PEFAC from the voicebox toolkit (http://www.

ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html). Because with [4] we only have a phase estimate in voiced sounds, we show the improvement in voiced segments alongside the overall improvement for entire utterances in Figure 5. When the fundamental frequency estimator detects unvoiced speech segments, the estimators fall back to a phase-blind estimation. Thus, if evaluated over entire signals, the results of the phase-aware estimators will get closer to the phase-blind approaches while the general trends remain.

A PROMISING APPROACH FOR PERFORMANCE IMPROVEMENT IS TO JOIN THE DIFFERENT TYPES OF PHASE PROCESSING APPROACHES, SUCH AS BY INCLUDING MORE EXPLICIT SIGNAL MODELS INTO ITERATIVE PHASE ESTIMATION APPROACHES OR VICE VERSA.

It can be seen that employing phase information to improve magnitude estimation [25] can indeed improve PESQ. The dominant benefit of the phase-aware magnitude estimators is that the phase provides additional information to distinguish between noise outliers and speech. Thus, the stronger the outliers after processing with phase-blind approaches, the larger the potential benefit of phase-aware processing. While here we show the average result over four noise types, a consistent improvement for the tested nonstationary noise types has been observed. While in stationary pink noise the PESQ scores are virtually unchanged, the largest improvements are achieved in babble. This is because babble bursts are often of high energy and may result in large outliers in phase-blind magnitude estimation that can be reduced by exploiting the additional information in the phase.

When an initial phase estimate is also employed as uncertain prior information when improving the spectral phase as proposed in the phase-aware complex estimator CUP [12], the performance can be improved further. The CUP estimator [12] employs the probability of a signal segment being voiced to control the certainty of the initial phase estimate. In unvoiced speech, the uncertainty is largest, effectively resulting in a phase-blind estimator. Therefore, again, we can only expect a PESQ improvement in voiced speech. Compared to phase-blind magnitude estimation [30] in voiced speech and at an input SNR of 0 dB, an improvement in PESQ by 0.12 points is achieved when all parameters are blindly estimated, while 0.18 points are gained with an oracle fundamental frequency. Considering that the improvement of the phase-blind estimator improves PESQ by 0.46 points, the additional improvement of 0.18 points by incorporating phase information in voiced speech is remarkable (factor 1.4), and demonstrates the potential of phase processing for the improvement of speech enhancement algorithms. While the average improvements using phase processing are still moderate, in specific scenarios, e.g., in voiced sounds or impulsive noise, phase processing can help to reduce noise more effectively than using phase-blind approaches. Audio examples can be found at www.speech.uni-oldenburg.de/pasp.html.

FUTURE DIRECTIONS

While the majority of single-channel STFT domain speech enhancement algorithms only address the modification of STFT magnitudes, in this article we reviewed methods that also involve STFT phase modifications. We showed that phase estimation could be done mainly based on models of the signal or by exploiting redundancy in the STFT representation. Examples for model-based algorithms are sinusoidal model-based approaches, and approaches that employ the group delay. By contrast, iterative approaches mainly rely on the spectrotemporal correlations introduced by the redundancy of the STFT representation with overlapping signal segments. While the results of the instrumental evaluations indicate that a sophisticated utilization of phase information can lead to improvements in speech quality, for a conclusive assessment, formal listening tests are required, rendering the subjective evaluation of particularly promising phase-aware algorithms a necessity for future research.

Despite recent advances, there are still many open issues in phase processing. For instance, similar to magnitude estimation, phase estimation is still difficult in very low SNRs. A promising approach for performance improvement is to join the different types of phase processing approaches, such as by including more explicit signal models into iterative phase estimation approaches or vice versa. A first step in this direction is presented in [26]. As another example, while the consistent Wiener filter only exploits the phase structure of the STFT representation, an exciting challenge going forward is to integrate models of the phase structure of the signal itself into a joint optimization framework.

Modern machine-learning approaches such as deep neural networks, which have proven to be very successful in improving speech recognition performance, have recently been shown to lead to state-of-the-art performance for speech enhancement using a magnitude-based approach. The natural next step is to extend their use to phase estimation to further improve performance. On top of the fact that they are data driven, which reduces the necessity for modeling assumptions that may be inaccurate, a great advantage of such methods over the iterative approaches for phase estimation presented here or approaches based on nonnegative matrix factorization or Gaussian mixture models, is that they can typically be efficiently evaluated at test time.

Indeed, striving for fast, lightweight algorithms is critical in the context of assisted listening and speech communication devices, where special requirements with respect to complexity and latency persist. While more and more computational power will be available with improved technology, for economic reasons as well as to limit power consumption, it is always of interest to keep the complexity as low as possible. Thus, more research in reducing complexity remains of interest. Complexity reduction could be obtained, for instance, by decreasing the overlap of the STFT analysis, but its impact on performance of phase estimation algorithms is not well studied. On the other hand, the lower bound on the latency of the algorithms is dominated by the window lengths in

STFT analysis and synthesis. Further research could therefore also address phase estimation using low latency filter banks.

After many years in the shadow of magnitude-centric speech enhancement, phase-aware signal processing is now burgeoning and expanding quickly: with still many aspects to explore, it is an exciting area of research that is likely to lead to important breakthroughs and push speech processing forward. Supplemental material and further references can be found at www.speech.uni-oldenburg.de/pasp.html.

ACKNOWLEDGMENT

This work was supported by grant GE2538/2-1 of the German Research Foundation.

AUTHORS

Timo Gerkmann (timo.gerkmann@uni-oldenburg.de) received his Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology in 2004 and 2010 from the Ruhr-Universität Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, New Jersey, United States. From 2010 to 2011, he was a postdoctoral researcher at the Royal Institute of Technology, Stockholm, Sweden. Since 2011, he has been a professor for speech signal processing at the University of Oldenburg, Germany. His main research interests are digital speech and audio processing, including speech enhancement, dereverberation, modeling of speech signals, speech recognition, and hearing devices.

Martin Krawczyk-Becker (martin.krawczyk-becker@uni-oldenburg.de) studied electrical engineering and information technology at the Ruhr-Universität Bochum, Germany. His major was communication technology with a focus on audio processing, and he received his Dipl.-Ing. degree in August 2011. From January 2010 to July 2010, he was with Siemens Corporate Research in Princeton, New Jersey, United States. Since November 2011, he has been pursuing his Ph.D. degree in the field of speech enhancement and noise reduction at the University of Oldenburg, Germany.

Jonathan Le Roux (leroux@merl.com) completed his B.Sc. and M.Sc. degrees in mathematics at the Ecole Normale Supérieure, Paris, France, and his Ph.D. degree at the University of Tokyo, Japan, and the Université Pierre et Marie Curie, Paris, France. He is a principal research scientist at Mitsubishi Electric Research Laboratories in Cambridge, Massachusetts, United States, and was previously a postdoctoral researcher at Nippon Telegraph and Telephone Communication Science Laboratories. His research interests are in signal processing and machine learning applied to speech and audio. He is a Senior Member of the IEEE and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee.

REFERENCES

[1] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[2] B. Yegnanarayana and H. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2281–2289, Sept. 1992.

[3] A. P. Stark and K. K. Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. ISCA Interspeech*, 2008, pp. 2602–2605.

[4] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Processing*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[5] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, no. 9, pp. 1493–1509, 1966.

[6] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-art*. San Rafael, CA: Morgan & Claypool, Feb. 2013.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[8] T. F. Quatieri, "Phase estimation with application to speech analysis-synthesis," Ph.D. dissertation, Massachusetts Inst. Technol., 1979.

[9] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. ICASSP*, Mar. 2012, pp. 101–104.

[10] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[11] P. Vary, "Noise suppression by spectral magnitude estimation – mechanism and theoretical limits," *Elsevier Signal Process.*, vol. 8, pp. 387–400, May 1985.

[12] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Processing*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[13] J. R. Hershey, S. J. Rennie, and J. Le Roux, "Factorial models for noise robust speech recognition," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. Hoboken, NJ: Wiley, 2012, ch. 12.

[14] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term Fourier spectrum for speech intelligibility." *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 1432–1439, Mar. 2010.

[15] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Elsevier Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[16] D. W. Griffin, "Signal estimation from modified short-time Fourier transform magnitude," Master's thesis, *Dept. Electr. Eng. and Computer Sci.*, Massachusetts Inst. Technol., Dec. 1983.

[17] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[18] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop Statistical Perceptual Audition (SAPA)*, Sept. 2008, pp. 23–28.

[19] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Phase initialization schemes for faster spectrogram-consistency-based signal reconstruction," in *Proc. Acoustical Society Japan Autumn Meeting*, paper no. 3-10-3, Sept. 2010.

[20] V. Gnann and M. Spiertz, "Improving RTISI phase estimation with energy order and phase unwrapping," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Sept. 2010.

[21] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 421–424, May 2010.

[22] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 1, pp. 178–185, Jan. 2013.

[23] J. Jensen and J. H. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 731–740, Oct. 2001.

[24] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[25] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, Feb. 2013.

[26] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[27] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE WASPAA*, Oct 2013, pp. 1–4.

[28] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[29] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[30] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 4037–4040.

[SP]

[Jan Wouters, Hugh J. McDermott, and Tom Francart]

Sound Coding in Cochlear Implants

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

[From electric pulses to hearing]

Cochlear implantation is a life-changing intervention for people with a severe hearing impairment [1]. For most cochlear implant (CI) users, speech intelligibility is satisfactory in quiet environments. Although modern CIs provide up to 22 stimulation channels, information transfer is still limited for the perception of fine spectrotemporal details in many types of sound. These details contribute to the perception of music and speech in common listening situations, such as where background noise is present. Over the past several decades, many different sound processing strategies have been developed to provide more details about acoustic signals to CI users. In this article, progress in sound coding for CIs is reviewed. Starting from a basic strategy, the current commercially most-used signal processing schemes are discussed, as well as recent developments in coding strategies that aim to improve auditory perception. This article focuses particularly on the stimulation strategies, which convert sound signals into patterns of nerve stimulation. The neurophysiological rationale behind some of these strategies is discussed and aspects of CI performance that require further improvement are identified.

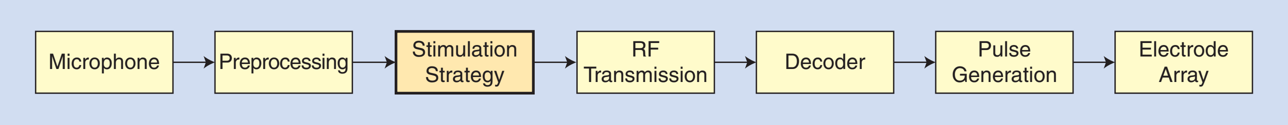
INTRODUCTION

The CI is the most successful man-made interface to the human neural system; i.e., a machine–brain interface. The auditory nerve is stimulated electrically, which leads to a partial restoration of auditory perception for people who have a severe hearing impairment. The understanding of speech by CI recipients in quiet environments can be very good, but is considerably worse than that of normal-hearing (NH) listeners in realistic listening situations. Typically, the presence of background noise greatly reduces the performance of CI systems. For example, the signal-to-noise ratio required for many CI users to attain 50% speech understanding is about 15 dB higher than that of NH listeners.

Currently, more than 300,000 people worldwide with severe hearing impairment, of whom 80,000 are young children, have received CIs. The proportion of children with CIs (two years old and younger) is increasing due to the increasing deployment of neonatal hearing-screening programs in many countries. Early implantation can give profoundly deaf children access to important information to process auditory signals and master spoken

Digital Object Identifier 10.1109/MSP.2014.2371671
Date of publication: 12 February 2015

1053-5888/15©2015IEEE



[FIG1] A block diagram of a complete CI system.

language skills at a young age. In many countries, a single CI is reimbursed by health insurance organizations, and in some countries, the cost of a second CI is also reimbursed, primarily for children. About 80% of normally developing, severely hearing-impaired children with a CI eventually participate in the mainstream educational system.

Apart from the technological and surgical progress that has made cochlear implantation the success it is today, the preformed cochlear duct and the ease of surgical access via the middle ear have played a role in its proliferation and progress. How CIs work has been described before in several articles; e.g., in [2]–[4]. This article focuses on a review of stimulation strategies. These are the techniques that convert sound signals picked up by a microphone into patterns of electric stimuli that activate the auditory nerve. The remainder of this section provides a short overview on how we hear and how a CI works.

In the normal auditory system, sound is captured and transmitted by the outer ear, predominantly the pinna (external ear) and ear canal, and then transformed in the middle ear (via the ossicles—small bones that have a mechanical impedance-matching function) to movement of the fluids and membranes in the cochlea, or inner ear. The cochlea has a spiral structure typically about 10 mm wide and 5 mm high. Within the cochlea, there are numerous transducer structures—the inner and outer hair cells—which have stereocilia that are deflected in response to incoming sound waves. In a healthy ear, movement of the stereocilia of inner hair cells leads to streams of action potentials in the auditory nerve fibers. This electrical activity has patterns with temporal and tonotopic characteristics that ultimately enable identification and interpretation of sounds, including music, speech, and language, at higher neural levels [5]. Temporal information about sound signals is carried through the precise timing of action potentials both within and between nerve fibers, whereas spectral information is represented mainly in the spatial distribution of activity across the neural population; the latter is referred to as the *tonotopic* organization of auditory nerve.

The most common cause of deafness is damage to or loss of the stereocilia and hair cells, resulting from infections, trauma, exposure to high levels of noise, side effects of certain drugs, and a range of physiological disorders. Hearing impairment may be acquired by adults who previously had normal hearing, or it may be present at birth. In many cases, the degree of hearing loss becomes progressively worse over time. When the hair cells are absent or extensively damaged, the transduction of the acoustically induced motion in the cochlea to neural action potentials is disrupted. If the resulting hearing loss is severe, the amplification that can be provided by acoustic hearing aids may be insufficient to restore satisfactory perception of sounds.

A CI bypasses the deficient transducer structures and produces action potentials at the auditory nerve sites (or the residual neurons, depending on the degree and type of pathology) using direct electrical stimulation. Most of today's CI systems have an external and an internal part. The external part consists of a behind-the-ear (BTE) device connected to an external transmission coil, which provides a radio-frequency (RF) link to a matching coil in the internal part, the implant. The implant consists of a miniature enclosure containing electronics connected to a number of electrodes. There are one or more reference electrodes on the enclosure or on a separate lead, and there is an array of multiple intracochlear electrodes, between 12 and 22 depending on the manufacturer and implant type. The stimulation currents flow between selected electrodes to activate the neural structures near the electrode-neuron interface. The electrode array is surgically inserted into the cochlea. Implantation of the complete internal system takes approximately three hours.

As illustrated in Figure 1, sound is captured in the external BTE device by a microphone system (one or more microphones). Preprocessing is applied, for example, to optimize the input dynamic range relative to input signal levels and to adjust the spectrum shape using a pre-emphasis filter. In some systems, there is also fixed or adaptive beamforming or other types of noise-reduction processing that typically exploit the differences between signals obtained from several microphones to enhance desired sounds while suppressing competing noise. The stimulation "strategy" refers to the transformation of the input sound signal into a pattern of electrical pulses. Digital specifications of the required stimulation patterns produced by the stimulation strategy are coded in the transcutaneous RF transmission. The RF signal also provides power to the internal part. The specifications of the stimulation are decoded from the RF signal. The electronics of the implant include one or more current source(s) to deliver the electrical stimulation pattern to the electrode channels. A channel is defined as a set of two or more electrodes with currents flowing between them. The term *monopolar stimulation* is used to describe current passing between an intracochlear electrode and a remote reference electrode, whereas *bipolar* refers to stimulation current passing between two intracochlear electrodes. The implant also has measurement amplifiers on-chip for the recording of evoked neural activity from nonstimulating electrodes via outward telemetry.

A few weeks after implantation and at regular intervals thereafter, stimulation levels are adjusted ("fitted") to the individual patient. In each fitting session, a patient-specific "map" is set up containing all stimulation parameters. For each channel, minimal levels of stimulation (min) and levels of maximal comfortable loudness (max) are determined. In some cases, the shape of the growth function between min and max that converts the input acoustic levels to electric stimulation levels is also determined. During a fitting session, impedances of the stimulation channels can be measured (which may lead to deactivation of some electrodes if faults are detected) and parameters of the preprocessing stage can be adjusted.

Today's CIs have a high power consumption compared to hearing aids, which means that the batteries largely determine the size of the BTE sound processor, making it cumbersome and unsightly for users. This also means that users need to replace batteries often, typically every day with rechargeable cells and every two days for primary cells, which may be expensive and inconvenient. Therefore, extensive research and development is currently devoted to reducing power consumption. Another major comfort improvement would be a totally implantable CI. The major challenge of a totally implantable system is the capture of airborne target sound with microphones and accelerometers, while suppressing the high levels of unwanted noise emanating from inside the human body.

A major technical and basic scientific challenge, and the subject of this article, is the translation of the captured sounds, particularly speech or music, to electrical stimulation patterns across the intracochlear channels to optimize auditory perception and interpretation. Historically, the objective of CIs has mainly been to improve speech intelligibility. Speech intelligibility is determined by spectral and temporal characteristics of the acoustic signal. The spectral information is coarsely coded through multichannel representation following the auditory system's natural tonotopic organization; i.e., acoustic spectral information is normally represented from low to high frequency in a corresponding spatial progression within the cochlea. Temporal speech information is commonly classified into three categories:

- the speech envelope, defined as the fluctuations in overall amplitude at rates between 2 and 20 Hz
- the periodicity from around 50 to 500 Hz, usually due to the fundamental frequency (F0)
- temporal fine structure (TFS).

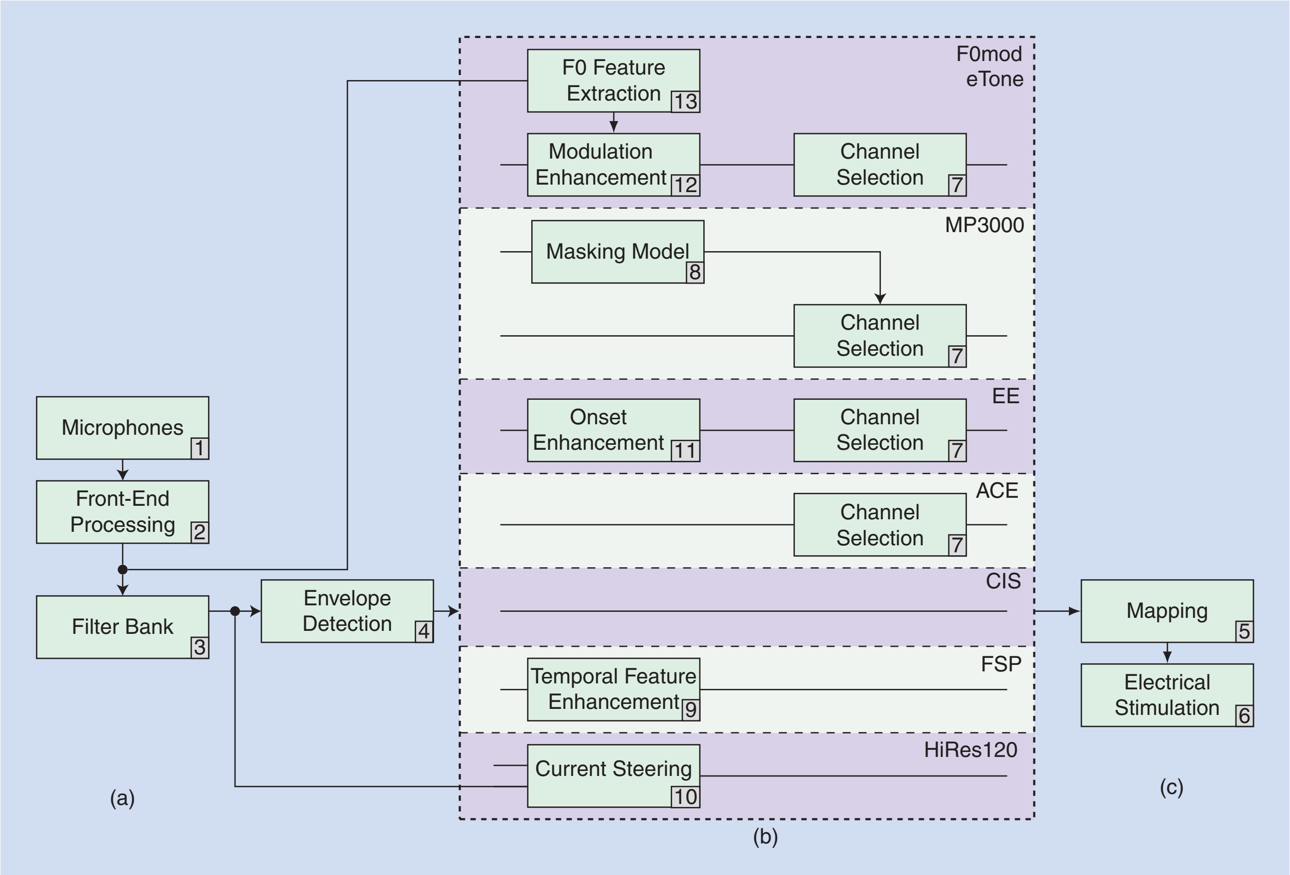
TFS can be defined as the variations in wave shape within single periods of periodic sounds, or over short time intervals of aperiodic ones. It has dominant fluctuation rates from around 500 Hz to 10 kHz. Alternatively, from a perceptual point of view, TFS can be defined as the fast fluctuations in a signal that can be used by NH listeners to perceive pitch, to localize sounds, and to binaurally segregate different sound sources. The fine structure is modulated in amplitude by the temporal envelope and periodicity. For speech sounds, F0 is the frequency at which the vocal cords vibrate. Recently the transmission of F0 information, related to pitch perception, has attracted a lot of interest because of the need to improve perception of music and tonal languages with CIs.

It is not easy to define pitch. It is defined by the American National Standards Institute (1994) as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low." From a musical point of view, it can be defined as "that attribute of sensation whose variation is associated with musical melodies." For periodic sounds, pitch is the perceptual counterpart of the fundamental frequency (F0), leading to the alternative definition that "a sound has a certain pitch if it can be reliably matched by adjusting the frequency of a pure tone of arbitrary amplitude" [6]. While F0 is a purely physical signal attribute, i.e., the frequency of the first harmonic of a complex tone, pitch is a perceptual attribute that arises after processing in the brain and can not always be easily linked to physical signal attributes. Typical relevant signals that elicit a pitch percept are spoken vowels and sustained sounds produced by musical instruments. Aperiodic sounds can also elicit a pitch percept, but it is less well-defined.

In the normal auditory system, pitch is determined by three different physical cues: 1) place of stimulation in the cochlea, 2) TFS, and 3) periodicity. The cochlea is tonotopically organized, so sounds with different spectral content will activate distinct neural populations, leading to different percepts. In the case of a simple sinusoid, there is a one-to-one relation between frequency and place of stimulation. For harmonic sounds, the situation is more complicated: the place of stimulation of the lowest harmonic still has a one-to-one relationship with F0, but the higher harmonics do not by themselves directly code F0. The spectral pitch mechanism is not very sensitive to small changes in F0, and the change in percept associated with a pure change in spectral pitch has been reported to correspond more to a change in timbre than a change in pitch [6]. Timbre, also called *tone color, tone quality*, or *brightness,* is the quality of a sound that distinguishes different types of sound production, such as voices or musical instruments. The American Standards Association (1960) defines timbre by exclusion as "that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar."

The second pitch-related cue, TFS, can yield a strong and tonal pitch percept when individual harmonics are coded by discrete neural populations and their frequency is lower than the maximal frequency to which neurons can phase-lock (around 1,500 Hz); i.e., the neural action potentials tend to occur during a particular phase of the oscillation. When multiple harmonics excite the same hair cells and therefore neurons, information is carried mainly by the aggregate stimulation pattern. This is likely to happen at higher frequencies because harmonics of a given F0 are spaced linearly in frequency whereas the auditory periphery is organized logarithmically. This leads to unavailability of the TFS of individual harmonics. However, the auditory system can still make use of a third physical cue: the periodicity of the combined harmonics, which corresponds to the F0. Perception of periodicity is limited to around 300–500 Hz. Periodicity pitch is weak compared to TFS pitch. For good pitch perception across a wide variety of types of sound, all three cues are needed.

Pitch perception with CIs is extremely poor. This is due both to limitations at the interface with electrical stimulation (spread of excitation) and to imprecise coding of temporal cues. The large spread of excitation in the cochlea and the small number of channels to code the low frequencies with electrical stimulation reduces the spectral resolution and therefore the precision of spectral pitch. Another limitation with electrical stimulation is the inability of CI users to perceive TFS. Therefore the only remaining mechanism is periodicity pitch perception, which is much weaker than TFS pitch and limited by the maximum frequency



[FIG2] **A block diagram of all monaural strategies discussed in this article. Common elements are shown in (a) and (c), while strategy-specific elements are shown in (b).**

at which pitch changes are perceived, around 300 Hz. Furthermore, temporal envelope fluctuations are not always accurately coded by current sound processing strategies.

Currently, an increasing number of people are being implanted bilaterally, especially children. Also, due to relaxed implantation criteria, an increasing number of people can make use of bimodal stimulation. These CI recipients have residual hearing in the nonimplanted ear, which can be aided with an acoustic hearing instrument. Listeners with bilateral CIs or using bimodal stimulation can potentially perceive interaural time differences (ITDs). Therefore another topic of intensive research is binaural hearing and the preservation of binaural cues in applications with bilateral and bimodal devices. ITDs, the difference in arrival time between the ears, are important binaural cues for NH listeners to localize sound sources and to separate multiple sound sources such as speech and noise. The latter is called binaural unmasking. ITDs range from 0 µs for sounds in front to around 700 µs for sounds from the side of the head. NH listeners can use ongoing temporal cues that are present in both the fine structure and the envelope of sound signals, and temporal cues in the onset of signals.

BASIC STIMULATION STRATEGIES

Historically, the first main types of stimulation strategies can be classified as feature extraction strategies. In such strategies, estimates of F0 and formants F1 and F2 of speech signals are calculated in real time. Formants are peaks in the spectral envelope corresponding to resonances of the vocal tract. Formants are used by the auditory system to identify sounds such as vowels. The formant information is used predominantly to stimulate channels corresponding to F1 and F2. The F0 is used to control the pulse rate. The outcomes in speech understanding of these schemes are, on average, lower than those of more recent schemes, and therefore they are not normally used any more in commercial processors [2[, [4].

A simple strategy, widely used in CI signal processing, is continuous interleaved sampling (CIS); see Figure 2 and [7]. CIS is based on a running spectral analysis of the preprocessed digital input sound signal performed by a bank of band-pass filters or a fast Fourier transform (FFT). The filter bank has an overall bandwidth from approximately 100 to 8,000 Hz, and the number of filters usually equals the number of stimulation channels at the electrode array-neuron interface. The filters have partially overlapping frequency responses and bandwidths that generally become broader with increasing frequency. Each filter is assigned to (at least) one intracochlear electrode following the frequency-place tonotopic organization of the cochlea. Although the correspondence of signal frequencies and filter bank outputs to depth of electrode insertion follows the tonotopy, the signal is not necessarily delivered to the normal anatomical or neurophysiological place because generally electrode arrays do not allow insertion beyond the anatomical position corresponding to acoustic frequencies lower than 500–1,000 Hz. However, studies have shown that with time of use of the CI, cortical plasticity can partly compensate for this mismatch. Also, manufacturers have recently introduced CI systems with electrode arrays that allow deeper insertion depths to facilitate more apical stimulation.

After the filter bank, the magnitude of the envelope in each channel is determined (block 4 in Figure 2), for instance with an envelope detector using rectification or using a Hilbert transformation followed by low-pass filtering. The filter cut-off frequency should at least comprise the modulation frequencies below 20 Hz to preserve the speech envelope information. Typical cut-off frequencies are between 125 and 300 Hz. When spectral estimates are obtained via an FFT, magnitudes corresponding to each of the electrodes are obtained from the allocated FFT bins, summing the powers across adjacent FFT bins depending on the filter bandwidths.

The stimulation levels are related to the magnitudes of the band-limited input signals by user-specific functions. The output of the envelope detector is transformed to a value between the min and max levels according to a nonlinear compression function because the electrical stimulation dynamic range (≈ 10 dB) is much smaller than the input dynamic range of the preprocessor (block 5 in Figure 2). This mapping is patient specific because min and max can vary widely across patients, stimulation channels, and electrode configurations (due to the status of the neural structures at the electrode-neuron interface and higher-level neural structures). Next, these transformed magnitudes modulate carrier waves of electrical pulses. Commonly, symmetric biphasic pulse trains are used in commercial CIs, and magnitude is coded by varying the pulse amplitude and/or the pulse width.

For practical reasons (many CIs have only one current source) but also for limiting across-channel interactions, pulsatile stimuli are used in an interleaved stimulation scheme (i.e., only one pulse is delivered at any time). Furthermore, all channels are activated in a temporally nonoverlapping sequence, and a fixed stimulation carrier rate is used [typically 500–2,000 pulses per second (pps)], with the total pulse rate equal to the number of active channels times the channel rate. The latter has no relationship with auditory neurophysiology, as neural fibers do not fire at fixed rates and stimulation rates are generally far higher than neural spike rates. However, it is simple from a signal processing point of view and provides most CI recipients with adequate perception of sounds.

This strategy can faithfully represent the temporal speech envelope in the electrical stimulation patterns, leading to effective transmission of envelope information, which is a necessary condition for speech perception. CIS was described by Wilson et al. in 1991 [7]. Essentially the same sound processing scheme, albeit with a relatively low stimulation rate (around 300 pps), was previously used in an earlier French CI system [8].

In general, the evaluation (and comparison) of strategies is mainly based on behavioral performance measures on identification and discrimination tasks related to speech understanding, music and tone perception, directional hearing, sound quality, and preference measures. Right now, no validated model of these measures, nor objective neurophysiological markers, exists for electrical stimulation. So behavioral tests are the reference evaluation approach.

In the following sections, a range of stimulation strategies for CI sound coding is described. Along with a description of the technical features of each strategy, we highlight the rationale behind the strategy, where one can be identified. We also review selected published outcomes for speech understanding and, if relevant and available, also for music or tone perception. Some of these schemes are widely used in commercial processors while others are experimental and still in development.

SOUND PROCESSING STRATEGIES IMPLEMENTED IN COMMERCIAL SOUND PROCESSORS

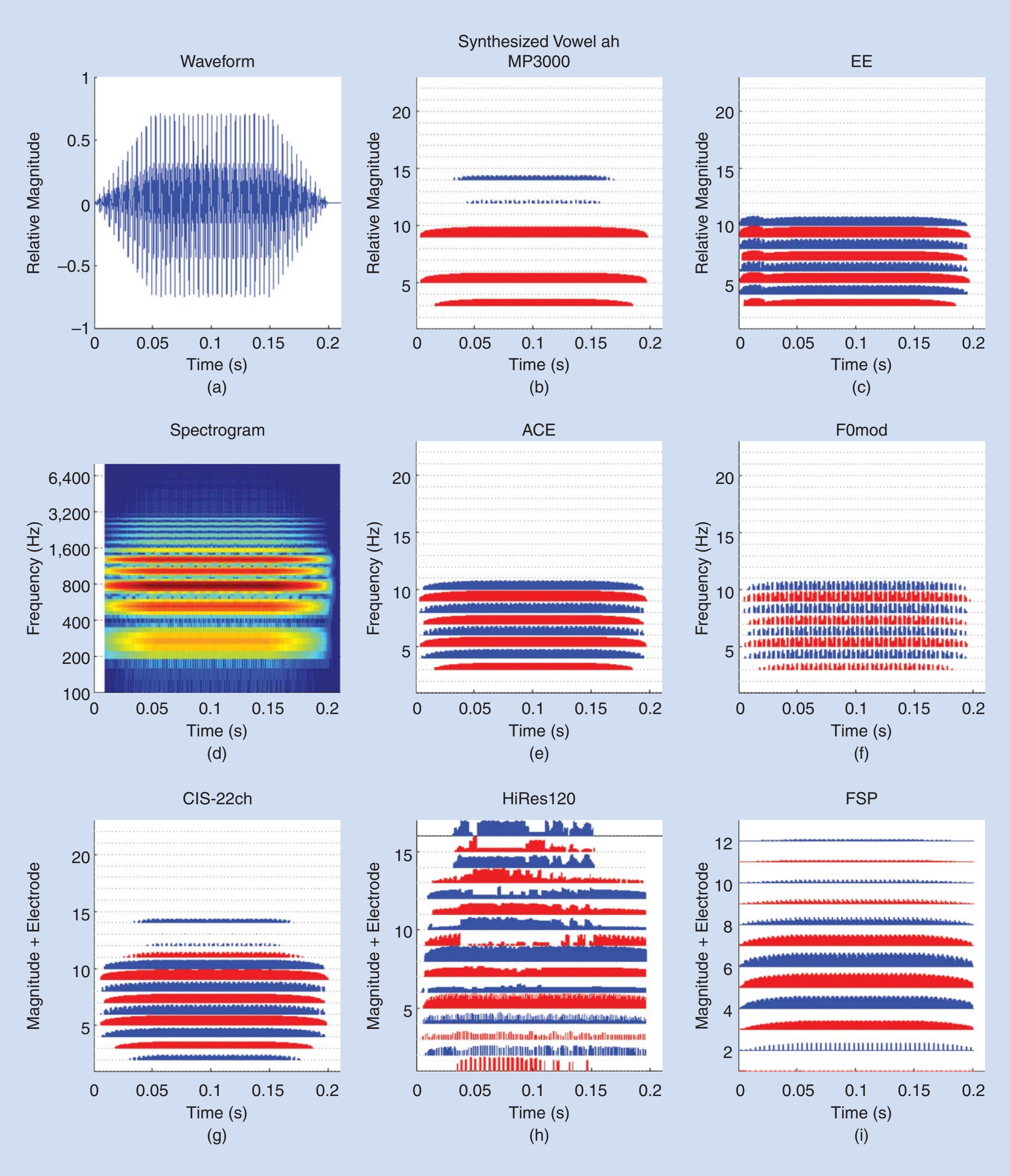
Since the introduction of the first stimulation strategies in commercial multichannel CIs over 30 years ago, a number of diverse sound processing strategies have been devised and evaluated. These strategies focus on better spectral representation, better distribution of stimulation across channels, and better temporal representation of the input signal. The four most commonly used strategies are described: 1) advanced combination encoder (ACE) with channel selection based on spectral features; 2) MP3000 (named after the MP3 digital audio format) with channel selection and stimulation based on spectral masking; 3); fine structure processing (FSP) based on enhancement of temporal features; and 4) HiRes120 (high resolution) with temporal feature enhancement and current steering to improve the spatial precision of stimulus delivery.

An overall outline of the sound processing steps for the different stimulation strategies, with common and differentiating parts, is shown in Figure 2. The outputs of the strategies are shown as electrodograms. An electrodogram is similar to a spectrogram, but the vertical axis indicates channel number rather than frequency, and biphasic current pulses are represented as vertical lines with amplitudes between 0 (min level of map) and 1 (max level of map). Electrodograms are shown of the synthesized vowel *ah* (Figure 3), a naturally spoken sentence in quiet taken from the HINT corpus (Figure 4), a selected word from the same sentence (Figure 5), and the same sentence in steady noise with a speech-weighted spectrum at a signal-to-noise ratio of 10 dB (Figure 6). The base stimulation rate per channel for ACE/CIS was 900 pps, for FSP 1,500 pps, and for HiRes120 1,856 pps.

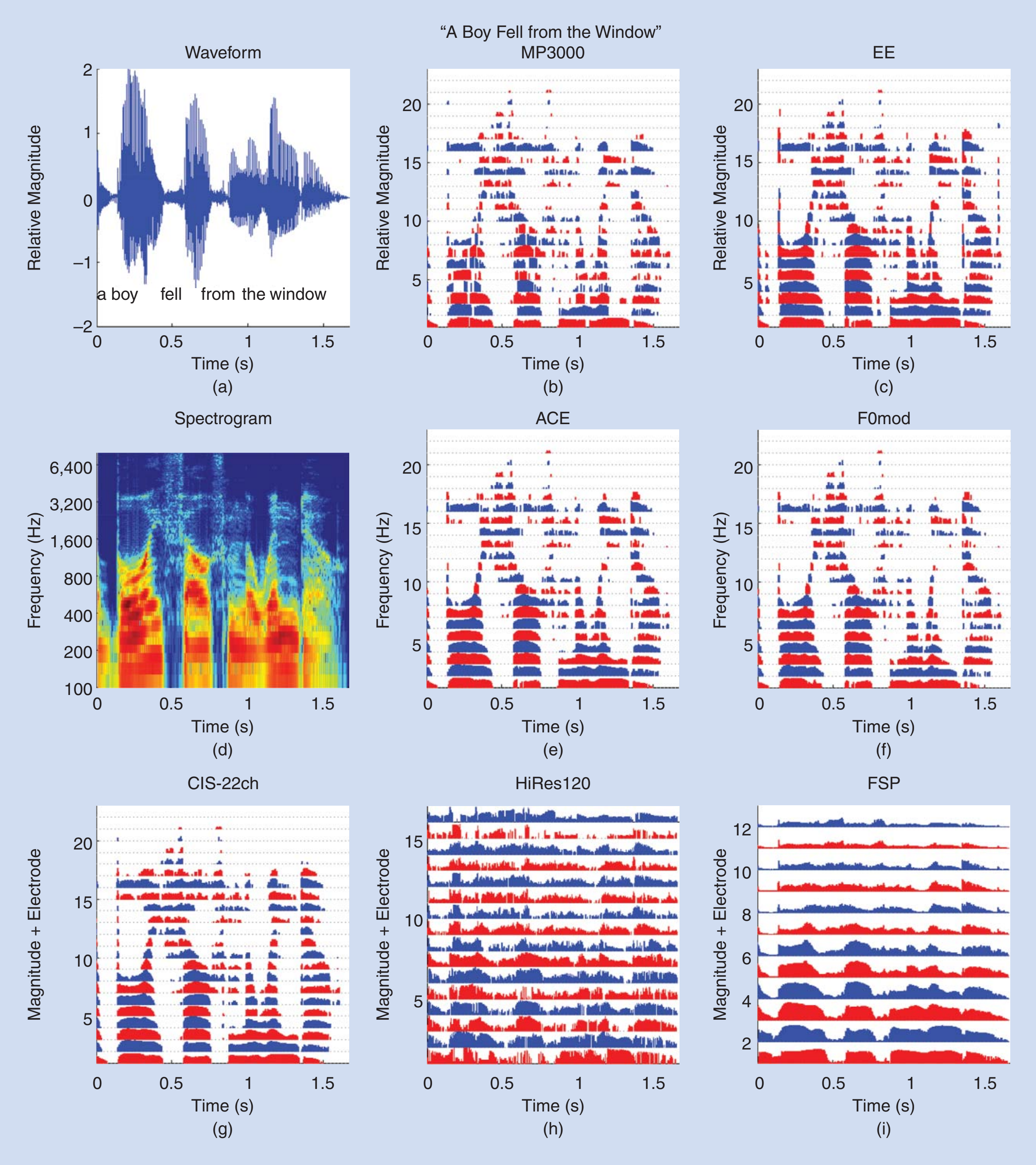
Four manufacturers of CI systems are on the international market (with implementations of strategies described in this review): Cochlear (ACE, MP3000), Advanced Bionics (HiRes120), Med-El (FSP), and Oticon Medical (formerly Neurelec).

ACE

ACE is the sound processing scheme currently used by most recipients of CI systems manufactured by Cochlear. It is functionally very similar to the spectral maxima sound processor (SMSP) and the Speak scheme [9] used with previous models of Cochlear CIs. The original development of the SMSP arose from the observation

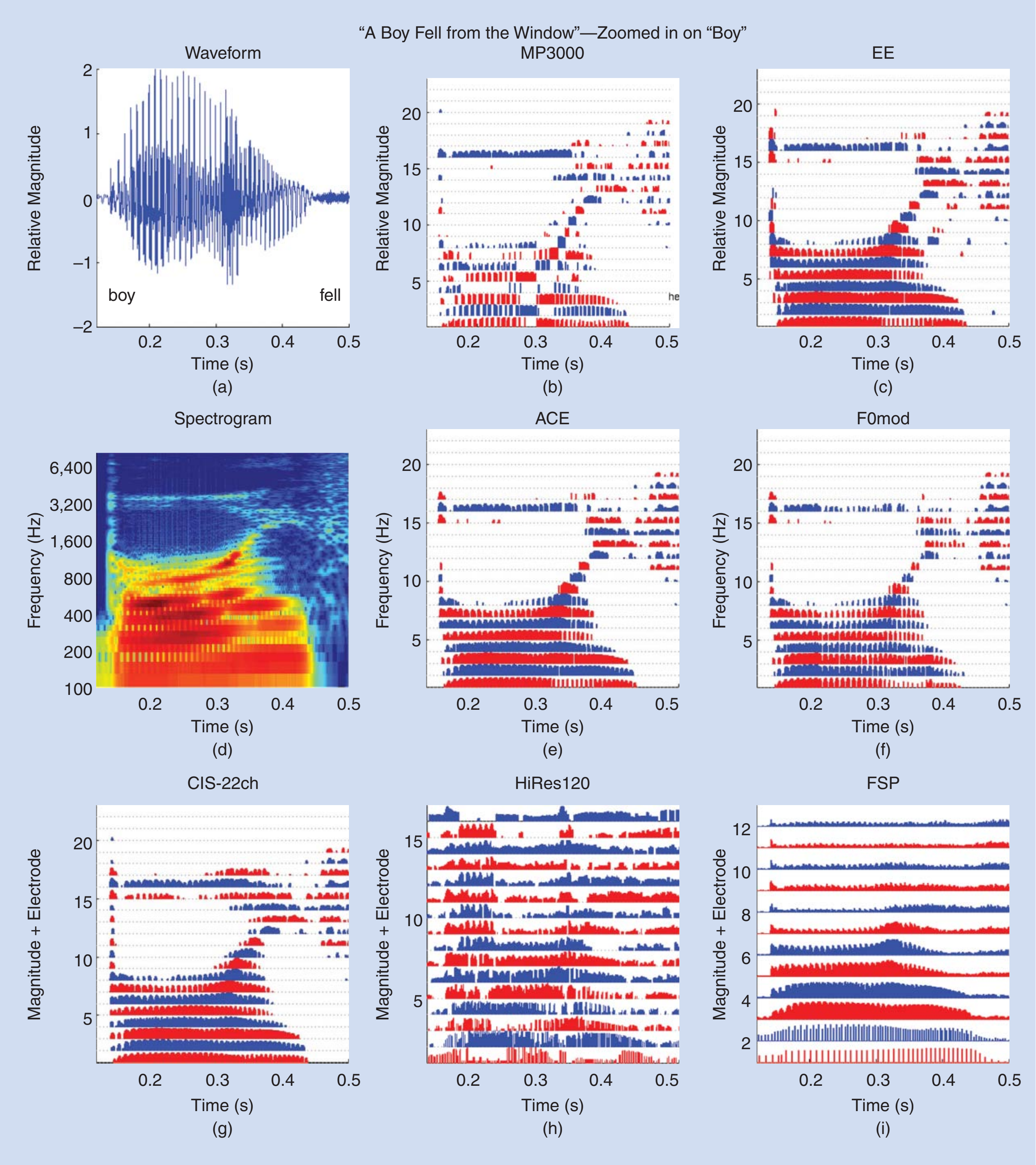


[FIG3] (a) Waveform, (d) spectrogram, and (b)–(c) and (e)–(i) electrodograms for a synthesized vowel with F0=100 Hz, and formant frequencies 700, 1,220, and 2,600 Hz. The signal was presented at an average root-mean-square (RMS) level of 60 dB sound pressure level (SPL). For the electrodograms, the vertical axis indicates the channel, and the height of each vertical line represents the magnitude of the pulse. The magnitude is expressed in different units for different strategies. The red and blue colors serve to visually distinguish adjacent channels and have no additional meaning. For the (g) CIS, (e) ACE, (b) MP3000, (c) EE, and (f) F0mod strategies, the channel magnitudes are shown between 0 and 1 before compression. For (h) HiRes120, the current was normalized by dividing by the maximum current, and normalized values below 0.1 were set to 0. HiRes120 uses simultaneous stimulation of adjacent electrodes to generate virtual channels, which is hard to distinguish on the current plot. For (i) FSP, the channel magnitudes between 0 and 1 are shown, which are linearly mapped to current, and multipulse sequences have been replaced by single pulse sequences for clarity.



[FIG4] (a) Waveform, (d) spectrogram, and (b)–(c) and (e)–(i) electrodograms of the sentence "A boy fell from the window" from the HINT corpus, uttered by a male speaker. All parameters are identical to those of Figure 3.

that sound processing schemes based on the presentation of selected acoustic features of speech signals were technically and perceptually limited. As mentioned previously, most of those schemes provided CI users with partial information primarily about the two lowest speech formants (F1, F2) and the fundamental frequency (F0). While those schemes enabled many recipients to understand speech adequately in favorable listening conditions, performance was degraded by even moderate levels of background noise. This was mainly because of the technical difficulty of estimating parameters corresponding to the selected speech features in real time when the signal-to-noise ratio is poor. The SMSP and its successor schemes, Speak and ACE (as well as closely related schemes provided by other CI companies), attempt to provide CI users with information about salient aspects of the acoustic



[FIG5] (a) Waveform, (d) spectrogram, and (b)–(c) and (e)–(i) electrodograms of the sentence "A boy fell from the window" from the HINT corpus, uttered by a male speaker, but zoomed in on the word "boy."

spectral shape without explicitly estimating speech features. Indeed, there is no inherent assumption that the sound signals processed for CI recipients contain any speech.

ACE has many signal processing modules in common with CIS and almost all other current CI processing schemes (blocks 1–6 in Figure 2). However, the major distinction with CIS is that on each stimulation cycle, only a subset of the available electrodes is selected. This is indicated by the "channel selection" block (block 7) in Figure 2. The subset comprises the n type of processing scheme is sometimes referred to as n-of-m. In cochlear CI systems, typically eight electrodes from the available set of 22 are selected for stimulation at a rate of 900 pps per electrode, although stimulation parameter values can be varied to optimize performance for individual recipients.

Figures 4–6 show that ACE represents some speech formant peaks and formant trajectories (i.e., changes in formant frequency over time) more distinctly than CIS, particularly when background noise is present. Because frequency bands containing relatively low signal levels are not represented in the stimulation pattern, ACE can enhance certain spectral features when perceived by CI users. This may be one reason that several studies of speech understanding have demonstrated slightly higher scores for ACE than CIS. For example, Skinner et al. [10] reported that CI listeners in two separate comparison studies scored about six to nine percentage points higher, on average, in sentence tests when using ACE rather than CIS.

FSP

Although most CI users obtain good performance with sound processing schemes such as ACE and CIS, unfortunately intelligibility of speech in competing noise is often unsatisfactory, and essential components of musical sounds—particularly pitch—are poorly perceived. Part of the reason may be the lack of TFS in the stimulation patterns. In general, TFS is characterized by the rapid amplitude variations within each of the band-pass filters that implement the initial spectral analysis of sound signals. In contrast, only the slowly varying envelope of the band-limited signals is used to modulate stimulation levels in schemes such as ACE and CIS.

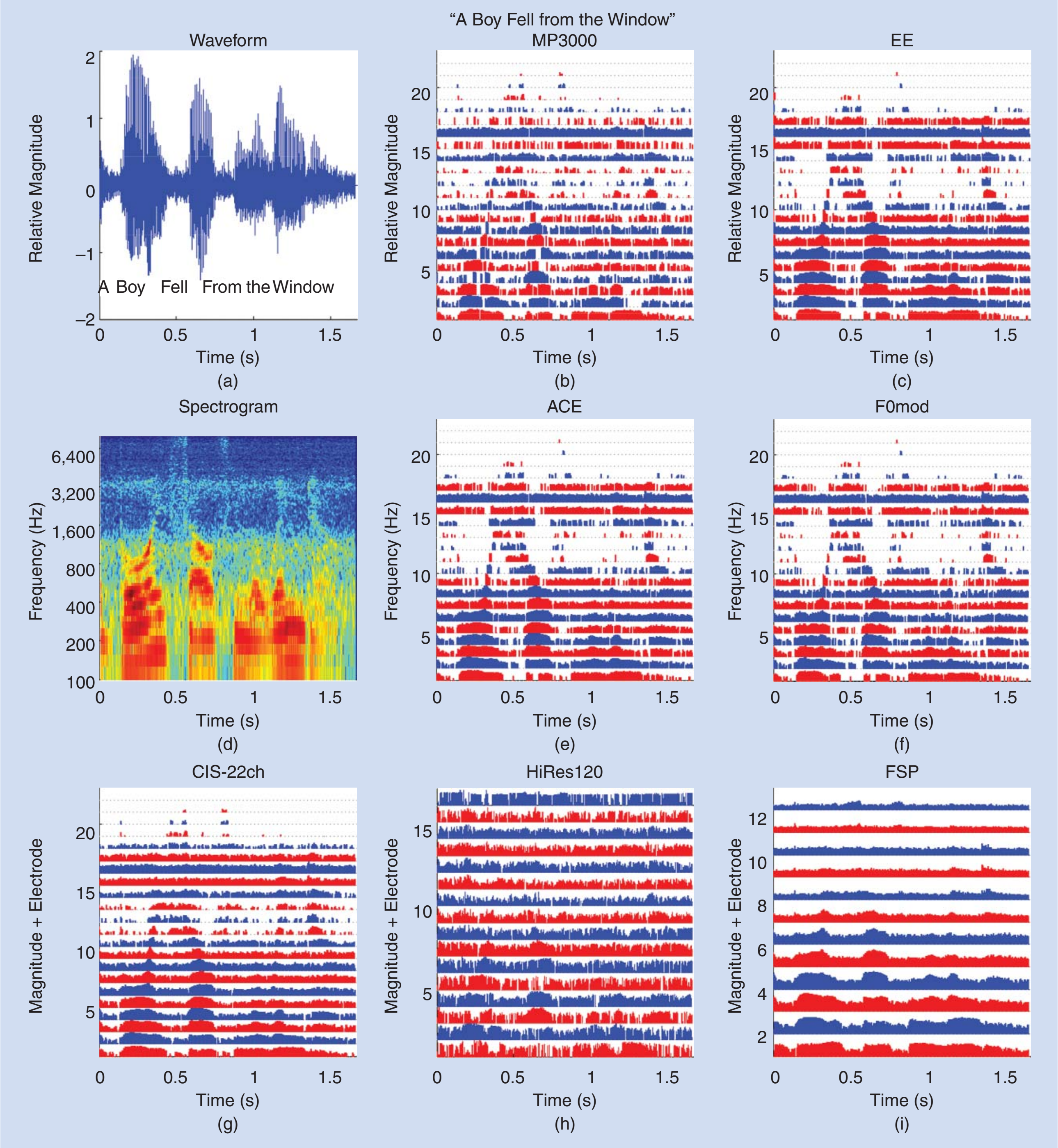
In the quest for improved CI sound processing, numerous attempts have been made to introduce TFS cues explicitly. One such scheme, currently the default in systems manufactured by Med-El, is known as FineHearing Technology. The aim of FineHearing Technology is to represent TFS information present in the lowest frequencies of the input sound signals by delivering bursts of stimulus pulses on one or several of the corresponding CI electrodes. These bursts can consist of one or more stimulation pulses and are derived indirectly from the band-limited acoustic signals. Each burst is triggered by a positive zero-crossing in the bandpass-filtered waveform, while stimulus pulses within the burst are delivered at a constant, high rate that depends on user-specific settings (typically 5,000–10,000 pps). The duration and amplitude-envelope modulation of each burst are predetermined to approximate the filtered acoustic waveforms after half-wave rectification. These bursts contain information about the TFS in the lower frequency bands that is not available in the envelope of those signals, potentially leading to improved perception for CI users. In essence, FineHearing Technology uses variable-rate coding to provide additional information about the TFS of the signal. Med-El has released the FSP, FS4, and FS4-p coding strategies. These strategies differ mainly in the frequency range across which TFS is presented. While in FSP, TFS is represented for frequencies up to 350–500 Hz; in FS4 and FS4-p, TFS is presented for frequencies up to 750–950 Hz. To faithfully represent F0, these strategies cover an input frequency range from 100–8,500 Hz by default, which differs from the CIS strategies from Med-El (250–8,500 Hz). The FSP coding strategy is illustrated in Figures 3–6, where TFS pulse patterns are delivered by the two most apical electrodes while the remaining electrodes convey CIS-like pulse trains.

Several of the coding strategies available in the Med-El system have been compared in a number of studies. Most published studies evaluating the perception of CI recipients when using FSP relative to other sound processing schemes (e.g., CIS) are difficult to interpret. In some cases, the sound-processor hardware and settings such as the input frequency range were altered at the same time as the processing algorithm was changed. In one study of 46 experienced CI users where such differences were explicitly taken into account, no significant differences were found between FSP and a variant of CIS in speech perception tests, although the participants' subjective preferences generally favored FSP [11]. Moreover, it should be noted that in some experiments the fitting of the CI system to recipients was not altered when changing from CIS to FSP. The study by Riss et al. [12] seems to indicate that at least some of the short-term improvements that have been reported with FSP can be attributed to the extended frequency range. As studies with the newer FS4 and FS4-p strategies are ongoing, further research is needed to quantify perceptual outcomes more thoroughly.

HiRes120

Another sound processing scheme designed to enhance delivery of TFS information to CI recipients is used in systems manufactured by Advanced Bionics. Known as HiRes120, this scheme applies a technique to identify the dominant spectral peak within each of the band-pass filters that perform the spectral analysis of incoming sounds. The frequency of each spectral peak is used to control a synthetic modulator such that the modulations contain temporal information derived from each frequency band that is not present in the amplitude envelope of the band-limited signals [13]. These modulations are combined with the corresponding envelope levels and then sampled in synchrony with the pulses delivered to the electrodes. The typical pulse rate on each electrode is about 2,000 pps. At the same time, the estimated peak frequency within each of the analysis filters is used to control the relative currents of pulses delivered simultaneously on two adjacent electrodes that are allocated to the filter. There are 16 intracochlear electrodes in the Advanced Bionics implant, and therefore 15 paired electrodes can be allocated to the filters. By varying the relative currents on the electrode pairs, so-called virtual channels are created, and it is assumed that the site of maximal neural activity can be steered with finer spatial resolution than is possible when the electrodes are activated one at a time. With HiRes120, eight different ratios of current are implemented, leading to eight virtual channels per adjacent pair of physical electrodes. HiRes120 is claimed to provide improvements over sound processing schemes such as CIS in both temporal and spatial resolution of the stimulation patterns. The main differences between these stimulation schemes are most clearly visible in the electrodograms of Figures 3 and 5. Additionally, a graphical representation of the virtual channels is shown in Figure 7.

HiRes, which is a CIS-like strategy without current-steering, has been compared with HiRes120 in various studies (e.g., [14] and the references therein) using measures of speech perception in quiet and in noise, and music perception. There were no clear significant effects of the processing strategy on any of the speech and music perception abilities nor on temporal modulation



[FIG6] (a) Waveform, (d) spectrogram, and (b)–(c) and (e)–(i) electrodograms of the sentence "A boy fell from the window" from the HINT corpus, uttered by a male speaker, but with noise added at an SNR of 10 dB.

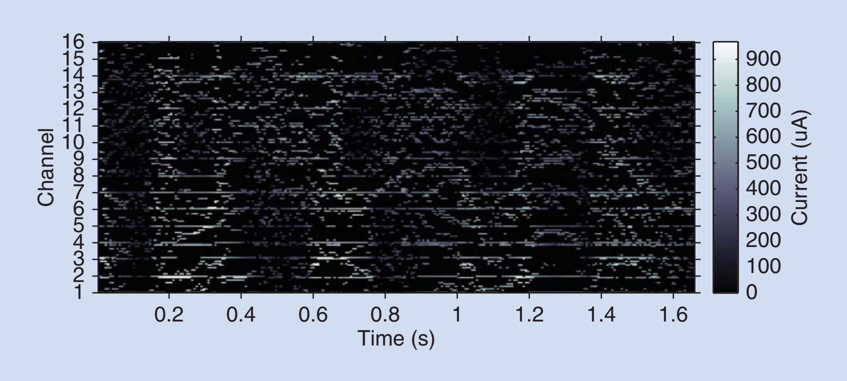
detection. Furthermore, experience with the strategies did not seem to play a significant role. For some psychophysical measures differences were observed, but with varying results for HiRes120. Further research is needed to investigate the impact on more ecologically relevant outcome measures.

For all CI sound processing strategies, the information throughput at the electrode-neural interface may be a fundamental limitation restricting improvements in perceptual performance. The limited perceptual effects of introducing explicit information about the fine structure of acoustic signals in some CI sound processing schemes such as HiRes120 and FSP may be a consequence of this "bottleneck" at the electrode-neural interface. In particular, if the spatial extent of the neural population activated by each electrode is broad and the populations associated with each electrode partially overlap, then temporal information from closely spaced electrodes will generally be combined at the neural level. Psychophysical studies have reported evidence that temporal patterns from nearby electrodes cannot be completely resolved by most CI recipients. This suggests that sound processing schemes like HiRes120 and FSP, which use very different approaches but rely on providing independent channels of information across adjacent electrodes, may result in only limited benefits [15]. More carefully controlled studies of CI recipients' listening experiences using schemes such as HiRes120 and FSP over an extended time are needed to determine specifically whether provision of fine-structure information by these schemes is perceptually beneficial.

MP3000

The MP3000 strategy is based on the ACE scheme but uses a psychoacoustic masking model with the aim of improving sound perception for CI users based on more perceptually relevant channel selection. The masking model attempts to select the perceptually most important spectral components in the coding of any given input audio signal. The rationale for this development was that it should not be necessary to code sounds in parts of the spectrum that are masked. This approach reduces the spread of excitation and can lead to a more precise representation of the spectrum, which in turn could lead to improved speech intelligibility. Processing techniques based on auditory masking are widely used in common audio and music data-compression algorithms. These techniques also compress the audio signals by selecting only a subset of the frequency bands at a time. A well-known example is the MP3 compression algorithm. In principle, the n-of-m speech coding strategies such as ACE are similar to these data reduction or compression algorithms.

In MP3000, an additional processing stage is introduced between the envelope estimation and the channel selection modules (see Figure 2, block 8). The psychoacoustic masking model used is derived from a body of data from psychoacoustic measurements in human auditory perception, such as studies on absolute thresholds of hearing and simultaneous masking [5]. For each sound, the envelopes of each channel of the filter bank are inputs to the psychoacoustic model, and masking spread functions with three parameters (peak amplitude or attenuation, high- and low-frequency slope) are calculated. The masked threshold is calculated for each channel selected. The overall masked threshold from all channels is approximated by a nonlinear superposition of the separate masked thresholds [16]. Subsequently, the n channels with highest levels relative to an estimate of the spread of masking are selected in each stimulation cycle. This selection of stimulation channels can be significantly different from the ACE standard scheme where only the n channels (typically n = 8) with the highest envelope magnitudes are selected. This is clearly visible in Figure 3, where in channel 14 a formant is coded with MP3000 that is not coded by ACE.



[FIG7] A virtual channel plot for the sentence "A boy fell from the window" processed by HiRes120. Color intensity indicates current. Integer numbers indicate "real" channels.

MP3000 has been implemented and evaluated in a within-subject repeated measures design with 221 subjects using an ABABA-design with "A" for ACE and "B" for MP3000. With a fixed pulse rate per channel, no significant difference was found for speech intelligibility and strategy preference between MP3000 (four to six spectral maxima selected) and ACE (eight to ten spectral maxima selected). The best results were found for MP3000 with six spectral maxima, leading to an increase in battery life of about 24% relative to ACE [17]. Thus when a lower number of stimulation channels is selected in each cycle, resulting in a lower overall stimulation rate, MP3000 has advantages. However, overall subject preferences were equally distributed between the two strategies, and additional parameters have to be fitted in the MP3000 mapping sessions.

EXPERIMENTAL PROCESSING STRATEGIES

In this section, some experimental stimulation strategies are briefly discussed to demonstrate the current limitations and opportunities with CI stimulation. Most of these strategies have been or are being considered for implementation in commercial speech processors for CIs. The following sections concern loudness-based strategies, envelope enhancement based on a neural model, enhancement of periodicity modulation, and bilateral stimulation strategies. The loudness-based strategies are not shown in Figure 2. They can be added onto any strategy by adding an extra block before the mapping block (5). The bilateral strategies are not shown for reasons of clarity.

LOUDNESS-BASED STRATEGIES

A distinctive approach to sound processing for CIs has been explored in a range of experimental schemes with the broad aim to improve the experience of loudness by CI recipients when listening to sounds with widely varying acoustic characteristics. Psychophysical studies have shown that CI users generally do not experience the loudness of sounds in the same way as listeners with NH, particularly when the spectral content and level of sound signals change over time.

In one such scheme, known as SpeL (for "Specific Loudness"), the initial stages of sound processing are based on a running spectral analysis and the distribution of current levels across electrodes is determined such that the loudness experienced by the CI user is similar to that experienced by an average listener with NH. Preliminary perceptual studies with CI recipients using SpeL confirmed that the relation between loudness and the level and bandwidth of sounds was closer to normal [18].

More recently, SCORE ("Stimulus Control to Optimize Recipient Experience"), a simplified version of SpeL, was developed that uses the estimated specific loudness function to calculate the total loudness of sound signals in real time. Tests of speech recognition with SCORE showed small but significant improvements over ACE. Tests with an extended version for CI recipients who use an acoustic hearing aid in the nonimplanted ear (SCORE bimodal), suggested that it may improve speech recognition and the ability of users to localize sounds, presumably because the loudness differences between ears that carry information about the direction of a sound source are conveyed more consistently (cf. [19] and the references therein).

ENVELOPE ENHANCEMENT

In a CI, the electrical stimulation directly generates action potentials in auditory neurons, predominantly bypassing any remaining hair cells and synapse function. The synapse normally demonstrates neural short-term adaptation [20], i.e., an increased firing rate at the onsets of sounds. This short-term adaptation acts as an across-channel phonological timing cue [20] and, with conventional schemes such as CIS, is not present in the electrically stimulated auditory nerve as in the normal auditory system. Furthermore, recent studies have demonstrated that the transient parts of the speech envelope carry information that is important for speech intelligibility in NH listeners.

Based on this rationale and former investigations, the enhanced envelope (EE) strategy was developed and its feasibility studied for applications in auditory prostheses. In this approach, an additional processing stage is introduced after the envelope detection stage (see Figure 2, block 11) wherein peaks, as a model for the short-term adaptation and dependent on the onset rise time, are added at the onsets in the envelope. This scheme is complementary to the main structure of ACE or CIS.

The EE algorithm was evaluated with CI users and all listeners demonstrated an immediate benefit with EE relative to ACE [21]. The advantage of this enhanced envelope coding is due to the emphasis of across-channel temporal coherence in the coded speech signal. This temporal marker is an important attribute for speech understanding in adverse listening situations and for sound source segregation; see also the electrodograms in Figures 4–6. The onset enhancement is particularly noticeable for the "b" sound in the word "boy" in Figure 5.

PERIODICITY MODULATION ENHANCEMENT

From psychophysical studies it is known that periodicity cues are better perceived when modulation depth is high and modulations are synchronized across channels to some extent [15]. This is probably due to spread of excitation: electrodes close together stimulate overlapping populations of neurons, which therefore receive the aggregate stimulation pattern of multiple electrodes. So if modulations are not synchronized across electrodes, the modulation depth at the neural level may be severely reduced.

From the electrodograms in Figures 4–6, it is clear that with most commercial strategies temporal modulations are not well coded. In some channels, modulation depth is quite shallow and the desynchronization across channels combined with spread of excitation leads to reduced modulation depth or even spurious modulations in the aggregate pattern that will be received by the auditory nerve fibers.

To improve this, several strategies have been developed (e.g., [22]–[25] and the references therein). While the signal processing to achieve it may differ, these strategies either expand modulation depth or remove existing modulations and explicitly modulate the envelope at the rate of F0. As an example, in the following, the F0 modulation (F0mod) and eTone strategies are briefly described.

The F0mod strategy is a simple example of a periodicity enhancement strategy based on the ACE strategy. For each frame of samples it estimates F0 and voicing probability using an autocorrelation approach. If a frame is unvoiced, ACE processing is applied. If a frame is voiced, all channels are modulated synchronously using a sinusoidal modulator constructed based on the F0 estimate. The block diagram and the output of F0mod are shown in Figure 2 and Figures 4–6, respectively.

The eTone strategy [23] is based on the same principles but includes an F0 estimator based on harmonic sieves, which is very precise and robust to noise, and the modulated envelope is mixed with the original envelope with a ratio depending on an estimate of harmonicity of that particular channel. Modulations are synchronous across channels and an exponential decay modulation shape is used.

The F0mod and eTone strategies were evaluated for music perception and speech recognition, and with tonal languages in which pitch determines the lexical meaning of certain phonemes (see [22]–[25] and the references therein). While periodicity enhancement strategies can clearly improve periodicity pitch perception, performance is still well below that of NH listeners. For good pitch perception, listeners need access to all three physical cues (see the "Introduction" section) and spectral (place) and temporal cues need to be consistent. There are no current CI strategies that make this possible, and we hypothesize that with the current electrode design and stimulation paradigm it is not possible to provide sufficiently place-specific stimulation to achieve performance similar to NH. Note that for a good representation of temporal information, good place specificity is required as well: when a population of neurons is stimulated by information from several channels due to spread of excitation, the aggregate pattern will be coded.

BILATERAL STRATEGIES

In various studies with controlled stimulation in laboratory conditions, it has been found that bilateral CI users can be sensitive to ITDs [26]. ITD thresholds, i.e., the smallest ITD that can be detected, vary widely across subjects, with the best thresholds around 50–100 µs and, in the worst case, no ITD sensitivity at all. However, with commercial sound processors, subjects hardly use ITDs in ecologically relevant tasks such as sound source localization. This is at least partly due to poor coding of temporal cues by current commercial sound processing strategies. The delay and spectral characteristics of the processing paths can be very different for the left and right CI devices, and certainly for bimodal systems where a CI is used in one ear and an acoustic hearing aid in the other ear. This may lead to nonsynchronous and noncoordinated (across channels) left and right auditory stimulation.

One of the first strategies developed to improve ITD coding with bilateral CIs is the peak derived timing (PDT) strategy [26]. It operates by synchronizing stimulation pulses from the CI with amplitude peaks in the fine structure of the signals in the different channels of the filter bank. In this manner, fine pulse timing cues are transmitted, in contrast with CIS-type sound processing techniques that provide only envelope information with fixed stimulation rates. Evaluations are reported in [22].

As bilateral CI strategies like PDT can introduce temporal patterns that are not synchronized with the acoustic signal, the modulation enhancement strategy (MEnS) was proposed [27] for bimodal stimulation. A deeply modulated envelope is imposed on all frequency channels simultaneously, explicitly synchronized with peaks in the acoustic signal. Improved ITD thresholds and improved lateralization (the extent to which the sound image can be moved to either side of the head by only changing ITD) were found with MEnS compared to ACE.

While some improvements in ITD perception have been obtained in laboratory tests with experimental strategies, the same caveats hold as with the pitch strategies described in the section “Periodicity Modulation Enhancement”: performance is much poorer than with NH. It should be noted, however, that thus far only acute experiments have been performed, while listeners potentially need long-term exposure to the novel stimulation paradigm to learn to use the binaural timing cues provided.

GENERAL DISCUSSION

In this article, a tutorial of CI stimulation strategies was presented, together with a review of concepts and rationales of different standard and experimental processing schemes. Some of the newer schemes have demonstrated significant improvements in the understanding of speech and perception of other types of sound. Although each of these strategies may lead to only a small benefit, it is plausible that appreciably larger benefits may be obtained when they are combined. Furthermore, some signal-processing approaches introduce speech enhancements in noisy conditions at the cost of significant signal distortions. These distortions may be detrimental for sound quality when appraised by listeners with normal or impaired acoustic hearing, but are hardly noticeable by most CI recipients. This is an opportunity for further improvements in auditory perception for CI users.

However, the broad neural excitation profiles inherent to present-day electrode array technology and electrical stimulation parameters most probably limit the potential for improvement. The number of independent information transmission channels is still very small because of both technical and perceptual/neural sensitivity limitations. Not all CI users can discriminate all channels, but even if all actual and virtual stimulation channels and electrodes may be perceptually discriminated from each other, this does not imply that channels can be resolved, nor that different channels can effectively convey independent information.

It has become clear that some temporal aspects of the input sound, such as the speech envelope and partly periodicity can be transmitted faithfully by CIs. However, the TFS and F0 are not adequately represented in present-day CI processors and are therefore presented to the auditory neural system only imprecisely.

Auditory perception results can be spectacular for many CI recipients in quiet environments, particularly in early-implanted deaf children when neural plasticity can fully play its role and in adults with a largely intact neural periphery. However, hearing in realistic adverse listening situations, as well as music perception and sound source localization are still major challenges for sound coding and electrical stimulation in CIs. Also, a wide variation in outcomes is observed across CI users. A significant proportion of recipients receive a limited benefit from their CI, at least in terms of speech understanding. Some investigations indicate that a better individual fitting of the stimulation parameters (the map) may result in substantial improvement, be it by better selection of active channels [28] or by development of closed-loop automatic fitting paradigms [29].

Another important factor is the neural survival at the electrode-neuron interface in the auditory periphery, which may be improved by application of drugs such as neurotrophins. Future research will include a greater focus on the combination of nonstandard pulse waveforms [30], new stimulation modes to reduce across-channel interactions, and improved electrode designs. These approaches may result in the provision of more independent information channels in future CI systems.

DECLARATION OF INTEREST

We have been involved in the development of the following strategies: SMSP, ACE, Speak, SpeL, SCORE (Hugh J. McDermott); F0mod, MEnS, SCORE, (Jan Wouters and Tom Francart); and EE (Jan Wouters).

ACKNOWLEDGMENTS

We thank Bas van Dijk (Cochlear), Leonid Litvak (Advanced Bionics), Peter Nopp (Med-El), and Dirk Meister (Med-El) for providing the data used to generate the electrodogram plots, information on the strategies, and insightful comments on the respective parts of the manuscript. We apologize for not being able to include all significant references that we wanted due to space constraints. We are grateful to Jonas Vanthornhout for his assistance with the figures. The Bionics Institute acknowledges the support it receives from the Victorian Government, Australia, through its Operational Infrastructure Support Program.

AUTHORS

Jan Wouters (jan.wouters@med.kuleuven.be) obtained M.S. and Ph.D. degrees in physics from the University of Leuven, KU Leuven, Belgium, in 1982 and 1989, respectively, with an intermission for officer military service. From 1989 to 1992, he was a postdoctoral research fellow with the National Fund for Scientific Research (FWO) at the Institute of Nuclear Physics (UCL Louvain-la-Neuve) and at NASA Goddard Space Flight Center (United States). Since 1993, he has been a professor in the Department of Neurosciences of the KU Leuven (full professor since 2005) where he teaches five physics and audiology courses. His research focuses on audiology, the auditory system, and auditory prostheses. He has authored approximately 240 articles in international peer-reviewed journals and is an associate editor of three international journals, president of the European Federation of Audiology Societies, president of the Belgian Audiology Society, and secretary-general of the International Collegium of Rehabilitative Audiology.

Hugh J. McDermott (hmcdermott@bionicsinstitute.org) is the deputy director of the Bionics Institute of Australia and a professorial fellow of the University of Melbourne. He is a biomedical engineer and Fellow of the IEEE. He was elected a fellow of the Acoustical Society of America in 2002 for signal processing that improves speech recognition with cochlear implants (CIs). For over 30 years, he has contributed directly to the design, development, and evaluation of CIs, hearing aids, and neurostimulation devices. He was named as an inventor on 19 patent families and has ten patent applications currently being processed. He has authored more than 105 journal articles, seven book chapters, and over 100 additional publications. In 2009, he was awarded the first Callier Prize in Communication Disorders, a biennial award from the University of Texas, Dallas, for leadership that has fostered scientific advances and significant developments in the diagnosis and treatment of communication disorders.

Tom Francart (tom.francart@med.kuleuven.be) received the M.S. and Ph.D. degrees in engineering from the University of Leuven, KU Leuven, Belgium, in 2004 and 2008, respectively. He has been a research professor in the research group ExpORL, Department of Neurosciences, KU Leuven since 2013. His research interests include sound processing for auditory prostheses, binaural hearing, and objective measures of hearing.

REFERENCES

[1] G. O'Donoghue, "Cochlear implants—Science, serendipity, and success." *N. Engl. J. Med.*, vol. 369, no. 13, pp. 1190–1193, Sept. 2013.

[2] P. Loizou, "Mimicking the human ear," *IEEE Signal Process. Mag.*, vol. 15, no. 5, pp. 101–130, Sept. 1998.

[3] M. Dorman and B. Wilson, "The design and function of cochlear implants," *Am. Sci.*, vol. 92, no. 5, pp. 436-445, Sept. 2004.

[4] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: system design, integration, and evaluation." *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 115–42, Jan. 2008.

[5] B. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Bingley, U.K.: Emerald, 2013.

[6] C. Plack, A. Oxenham, and R. Fay, *Pitch: Neural Coding and Perception*. New York: Springer, 2005.

[7] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants." *Nature*, vol. 352, no. 6332, pp. 236–238, July 1991.

[8] C.-H. Chouard, C. Fugain, B. Meyer, and H. Lacombe, "Long-term results of the multichannel cochlear implant," *Ann. N. Y. Acad. Sci.*, vol. 405, no. 1, pp. 387–411, June 1983.

[9] P. Seligman and H. McDermott, "Architecture of the Spectra 22 speech processor," *Ann. Otol. Rhinol. Laryngol. Suppl.*, vol. 104, Suppl. 166, pp. 139–141, 1995.

[10] M. W. Skinner, L. A. Whitford, K. L. Plant, C. Psarros, and T. A. Holden, "Speech recognition with the Nucleus 24 SPEAK, ACE, and CIS speechcoding strategies in newly implanted adults," *Ear Hear.*, vol. 23, no. 3, pp. 207–223, June 2002.

[11] J. Müller, S. Brill, R. Hagen, A. Moeltner, S.-J. Brockmeier, T. Stark, S. Helbig, J. Maurer, T. Zahnert, C. Zierhofer, P. Nopp, and I. Anderson, "Clinical trial results with the MED-EL fine structure processing coding strategy in experienced cochlear implant users." *ORL J. Otorhinolaryngol. Relat. Spec.*, vol. 74, no. 4, pp. 185–198, Jan. 2012.

[12] D. Riss, J.-S. Hamzavi, A. Selberherr, A. Kaider, M. Blineder, V. Starlinger, W. Gstoettner, and C. Arnoldner, "Envelope versus fine structure speech coding strategy: A crossover study." *Otol. Neurotol.*, vol. 32, no. 7, pp. 1094–1101, Sept. 2011.

[13] W. Nogueira, L. M. Litvak, B. Edler, J. Ostermann, and A. Büchner, "Signal processing strategies for cochlear implants using current steering," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–21, Nov. 2009.

[14] G. S. Donaldson, P. K. Dawson, and L. Z. Borden, "Within-subjects comparison of the HiRes and Fidelity120 speech processing strategies: speech perception and its relation to place-pitch sensitivity," *Ear Hear.*, vol. 32, no. 2, pp. 238–250, 2011.

[15] C. M. McKay and H. J. McDermott, "The perception of temporal patterns for electrical stimulation presented at one or two intracochlear sites." *J. Acoust. Soc. Amer.*, vol. 100, no. 2 (Pt. 1), pp. 1081–1092, Aug. 1996.

[16] W. Nogueira, A. Büchner, T. Lenarz, and B. Edler, "A psychoacoustic n-of-m-type speech coding strategy for cochlear implants," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, pp. 3044–3059, 2005.

[17] A. Buechner, A. Beynon, W. Szyfter, K. Niemczyk, U. Hoppe, M. Hey, J. Brokx, J. Eyles, P. Van de Heyning, G. Paludetti, A. Zarowski, N. Quaranta, T. Wesarg, J. Festen, H. Olze, I. Dhooge, J. Müller-Deile, A. Ramos, S. Roman, J.-P. Piron, D. Cuda, S. Burdo, W. Grolman, S. R. Vaillard, A. Huarte, B. Frachet, C. Morera, L. Garcia-Ibáñez, D. Abels, M. Walger, J. Müller-Mazotta, C. A. Leone, B. Meyer, N. Dillier, T. Steffens, A. Gentine, M. Mazzoli, G. Rypkema, M. Killian, and G. Smoorenburg, "Clinical evaluation of cochlear implant sound coding taking into account conjectural masking functions, MP3000." *Cochlear Implants Int.*, vol. 12, no. 4, pp. 194–204, Nov. 2011.

[18] H. J. McDermott, C. M. McKay, L. M. Richardson, and K. R. Henshall, "Application of loudness models to sound processing for cochlear implants," *J. Acoust. Soc. Amer.*, vol. 114, no. 4 (Pt. 1), pp. 2190–2197, 2003.

[19] T. Francart and H. J. McDermott, "Speech perception and localisation with SCORE bimodal: a loudness normalisation strategy for combined cochlear implant and hearing aid stimulation," *PLoS One*, vol. 7, no. 10, p. e45385, Oct. 2012.

[20] B. Delgutte, "Auditory neural processing of speech," in *Handbook of Phonetic Science*, W. J. Hardcastle and J. Laver, Eds. Oxford: Blackwell, 1997, pp. 507–538.

[21] R. Koning and J. Wouters, "Speech onset enhancement improves intelligibility in adverse listening conditions in cochlear implants," *Hear. Res.*, to be published.

[22] A. E. Vandali, C. M. Sucher, D. J. Tsang, C. M. McKay, J. W. Chew, and H. McDermott, "Pitch ranking ability of cochlear implant recipients: a comparison of sound-processing strategies." *J Acoust Soc Am*, vol. 117, no. 5, pp. 3126–3138, 2005.

[23] A. E. Vandali and R. J. M. van Hoesel, "Development of a temporal fundamental frequency coding strategy for cochlear implants." *J. Acoust. Soc. Amer.*, vol. 129, no. 6, pp. 4023–4036, June 2011.

[24] T. Francart, A. Osses, and J. Wouters, "Speech perception with F0mod, a cochlear implant pitch coding strategy," *Int. J. Audiol.*, to be published. >

[25] M. Milczynski, J. E. Chang, J. Wouters, and A. van Wieringen, "Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues," *Hear. Res.*, vol. 285, no. 1–2, pp. 1–12, Mar. 2012.

[26] R. J. van Hoesel and R. S. Tyler, "Speech perception, localization, and lateralization with bilateral cochlear implants." *J. Acoust. Soc. Amer.*, vol. 113, no. 3, pp. 1617–1630, 2003.

[27] T. Francart, A. Lenssen, and J. Wouters, "Modulation enhancement in the electrical signal improves perception of interaural time differences with bimodal stimulation," *J. Assoc. Res. Otolaryngol.*, vol. 15, no. 4, pp. 633–647, Aug. 2014.

[28] S. N. Garadat, T. A. Zwolan, and B. E. Pfingst, "Using temporal modulation sensitivity to select stimulation sites for processor MAPs in cochlear implant listeners," *Audiol. Neurotol.*, vol. 18, no. 4, pp. 247–260, Jan. 2013.

[29] M. Mc Laughlin, T. Lu, A. Dimitrijevic, and F.-G. Zeng, "Towards a closed-loop cochlear implant system: application of embedded monitoring of peripheral and central neural activity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 4, pp. 443–454, July 2012.

[30] J. A. Undurraga, R. P. Carlyon, J. Wouters, and A. van Wieringen, "The polarity sensitivity of the electrically stimulated human auditory nerve measured at the level of the brainstem," *J. Assoc. Res. Otolaryngol.*, vol. 14, no. 3, pp. 359–377, June 2013.

Terence Betlehem, Wen Zhang, Mark A. Poletti, and Thushara D. Abhayapala

Personal Sound Zones

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

Delivering interface-free audio to multiple listeners

Sound rendering is increasingly being required to extend over certain regions of space for multiple listeners, known as *personal sound zones*, with minimum interference to listeners in other regions. In this article, we present a systematic overview of the major challenges that have to be dealt with for multizone sound control in a room. Sound control over multiple zones is formulated as an optimization problem, and a unified framework is presented to compare two state-of-the-art sound control techniques. While conventional techniques have been focusing on point-to-point audio processing, we introduce a wave-domain sound field representation and active room compensation for sound pressure control over a region of space. The design of directional loudspeakers is presented and the advantages of using arrays of directional sources are illustrated for sound reproduction, such as better control of sound fields over wide areas and reduced total number of loudspeaker units, thus making it particularly suitable for establishing personal sound zones.

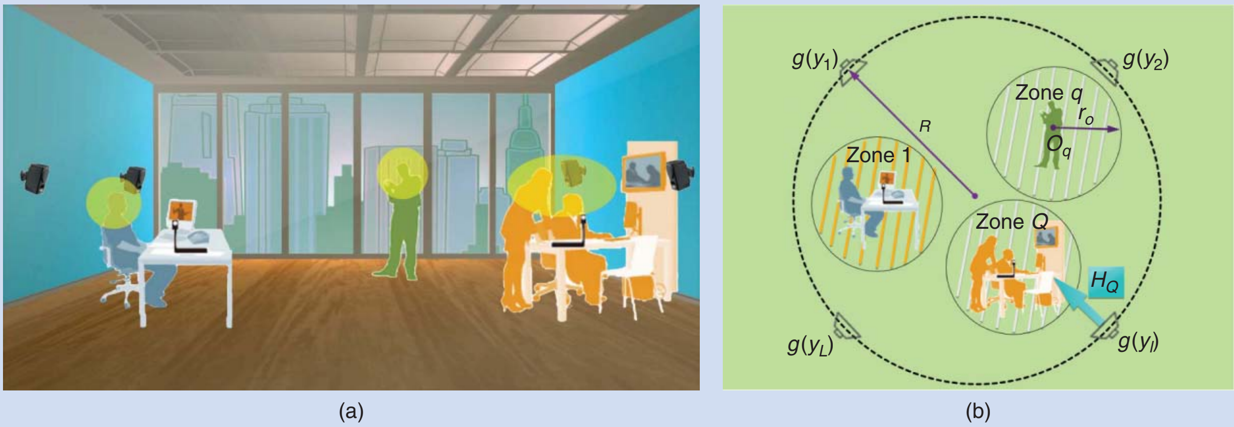
INTRODUCTION

Sound recording and sound reproduction are becoming increasingly ubiquitous in our daily lives. The ultimate goal of sound reproduction is to recreate the full richness of a sound field including not only the sound content but also the spatial properties to give the listener full knowledge about both the sound source and acoustic environment. Spatial sound reproduction technologies so far have made tremendous progress in reproducing sound fields over fairly large regions of space using an array of loudspeakers. This introduces the idea of establishing personal sound zones, whereby interface-free audio is delivered to multiple listeners in the same environment without physical isolation or the use of headphones (Figure 1). This concept has recently drawn attention due to a whole range of audio applications, from controlling sound radiation from a personal audio device to creating individual sound zones in all kinds of enclosures (such as shared offices, passenger cars, and exhibition centers) and generating quiet zones in noisy environments.

The first known demonstration of reproducing a sound field within a given region of space was conducted by Camras at the Illinois Institute of Technology in 1967, where loudspeakers were distributed on the surface enclosing the selected region to control sound radiation, and the listeners could move freely within the recreated environment [1]. The well-known ambisonics [2], wave field synthesis [3], and higher-order spherical harmonics-based techniques [4] were developed separately for more advanced spatial sound field reproduction over a large region of space. Druyvesteyn and Garas [5] first proposed the

Digital Object Identifier 10.1109/MSP.2014.2360707
Date of publication: 12 February 2015

1053-5888/15©2015IEEE



(a)

(b)

[FIG1] (a) An illustration of personal sound zones in an office environment. (b) A loudspeaker array is used to create multiple sound zones for multiple listeners.

concept of a personal sound zone, i.e., reproducing sound within a desired region of space with a reduced sound level elsewhere. Microsoft researchers later demonstrated their "Personal Audio Space" project at Microsoft Research TechFest 2007, where a linear loudspeaker array consisting of 16 drivers was used to enhance the sound in one area while canceling sound waves in another area within the same physical space. By stepping even a few paces outside the target region, users reported that they could not hear the reproduced music. Researchers further extended this concept to develop personal audio for personal computers and televisions [6], as well as for mobile devices [7] and automobile cabins [8].

A way to create personal sound zones is to formulate a multizone sound control problem within the same physical space as illustrated in Figure 1. Here, multiple microphones and loudspeakers are used to control the reproduced sound fields. A preference is to use a single array of loudspeakers rather than separate arrays for each zone. This improves freedom and flexibility, allowing sound zones to be positioned dynamically and listeners to freely move between zones. When the system is implemented in reverberant enclosures, loudspeaker designs and audio processing are two key aspects to control sound radiation and to deal with the complexity and uncertainty associated with sound field reproduction. This article aims at reviewing these techniques to support the goal of establishing personal sound zones.

MULTIZONE SOUND CONTROL

In a general formulation, sound fields are produced over Q sound zones. Here M pressure controlling microphones are placed within each zone so that the zone sound fields are controlled by a total of QM matching points. The sound pressures measured at the microphone positions in each zone q are represented as a vector $\boldsymbol{p}_q = [p(\boldsymbol{x}_{q,1}, \omega), \ldots, p(\boldsymbol{x}_{q,M}, \omega)]^T$ and given by

$$\boldsymbol{p}_q = \boldsymbol{H}_q \boldsymbol{g}, \qquad (1)$$

where $\boldsymbol{g} = [g(\boldsymbol{y}_1, \omega), \ldots, g(\boldsymbol{y}_L, \omega)]^T$ denotes the vector of loudspeaker driving signals at a given frequency ω to create personal audio sound scenes and $\boldsymbol{H}_q$ represents a matrix of acoustic transfer functions (or acoustic impedances) between the loudspeaker drivers and the microphones in zone q. Sound control techniques can broadly be classified into two categories, acoustic contrast control (ACC) and pressure matching (PM), and we consider each in turn.

ACOUSTIC CONTRAST CONTROL

Choi and Kim [9] first formulated the personal audio problem by creating two kinds of sound zones: the bright zone within which we want to reproduce certain sounds with high acoustic energy, and the dark zone (or the quiet zone) within which the acoustic energy is kept at a low level. The principle of ACC is to maximize the contrast in the acoustic energy between the bright zone and the dark zone. Among the Q sound zones, we specify the first zone as the bright zone and the remaining $Q-1$ zones as the dark zones. The acoustic energy in the bright zone is defined from the sound pressures measured at the M matching points, that is $E_{\mathrm{b}} = \|\boldsymbol{p}_{\mathrm{b}}\|^2 = \|\boldsymbol{H}_{\mathrm{b}}\boldsymbol{g}\|^2$ with $\boldsymbol{H}_{\mathrm{b}} = \boldsymbol{H}_1$ and $\|\cdot\|$ denoting the ℓ_2 norm. Similarly, the acoustic energy in the dark zones is represented as $E_{\mathrm{d}} = \|\boldsymbol{p}_{\mathrm{d}}\|^2 = \|\boldsymbol{H}_{\mathrm{d}}\boldsymbol{g}\|^2$ with $\boldsymbol{H}_{\mathrm{d}} = [\boldsymbol{H}_2^H, \ldots, \boldsymbol{H}_Q^H]^H$ and $(\cdot)^H$ represents the Hermitian transpose.

In [9], the acoustic contrast, defined as a ratio between the average acoustic potential energy density produced in the bright zone to that in the dark zones, is maximized. The acoustic contrast maximizing method may perform well over the dark zones but may be unrobust to providing the desired maximum energy in the bright zone. To ensure the sound energy within different zones are optimized simultaneously, the problem can be reformulated as maximizing the acoustic energy in the bright zone with the constraint that the energy in the dark zone is limited to a very small value D_0. In addition, a limit is imposed on the loudspeaker power consumption, i.e., $\|\boldsymbol{g}\|^2 \leq E_0$, also known as the *array effort*. These constraints ensure that sound leakage outside the Q zones is not excessive and also that realized loudspeaker weights are chosen to ensure the implementation is robust to driver positioning errors and changes in the acoustic environment. The ACC problem can then be posed as

$$\max_{\boldsymbol{g}} \| H_{\mathrm{b}} \boldsymbol{g} \|^2 \tag{2a}$$

$$\text{subject to} \| H_{\mathrm{d}} \boldsymbol{g} \|^2 \le D_0 \tag{2b}$$

$$\| \boldsymbol{g} \|^2 \le E_0 . \tag{2c}$$

The objective and the constraints are summarized into a single objective function represented using the Lagrangian [10],

$$\max_{\boldsymbol{g}} L_c(\boldsymbol{g}) = \| H_{\mathrm{b}} \boldsymbol{g} \|^2 - \lambda_1 (\| H_{\mathrm{d}} \boldsymbol{g} \|^2 - D_0) - \lambda_2 (\| \boldsymbol{g} \|^2 - E_0), \quad \lambda_1, \lambda_2 \ge 0, \tag{3}$$

where λ_1 and λ_2 are Lagrange multipliers to adjust the relative importance of each condition (2b) and (2c). The solution that maximizes the Lagrangian is obtained by taking the derivative of L_c with respect to $\boldsymbol{g}$ and equating it to zero, and is written as

$$\lambda_1 [H_{\mathrm{d}}^H H_{\mathrm{d}} + \frac{\lambda_2}{\lambda_1} I] \boldsymbol{g} = [H_{\mathrm{b}}^H H_{\mathrm{b}}] \boldsymbol{g}, \tag{4}$$

which is recognized as a generalized eigenvector problem. The optimum source strength vector $\boldsymbol{g}_c$ is set as the eigenvector corresponding to the maximum eigenvalue of the matrix $[H_{\mathrm{d}}^H H_{\mathrm{d}} + (\lambda_2/\lambda_1) I]^{-1} [H_{\mathrm{b}}^H H_{\mathrm{b}}]$. The ratio of Lagrange multipliers $\lambda = \lambda_2/\lambda_1$ determines the tradeoff between the performance and array effort and must be chosen iteratively for the constraint on the array effort to be satisfied. The formulation in (4) yields essentially the same answer as that in [8], or the so-called indirect formulation in [10], which diagonally loads the matrix $H_{\mathrm{d}}^H H_{\mathrm{d}}$ before inverting it to improve the matrix condition number.

The formulation adopted here leads to a straightforward way for demonstrating the connection between the ACC method and the PM method, which will be explained next.

PRESSURE MATCHING

The PM method aims to reproduce a desired sound field in the bright zone at full strength, while producing silence in other zones. The idea comes from the traditional crosstalk-cancelation problem, where small regions of personal audio are created by controlling the pressure at discrete spatial points (microphone or listener positions). Multizone sound control is an extension of the traditional approach with a sufficiently dense distribution of matching points within all the zones. Given a target sound field $\boldsymbol{p}_{\mathrm{des}}$ to be reproduced in the bright zone, the robust PM formulation can be written using an ℓ_2 PM objective along with the constraints on the sound energy in the dark zones and the array effort constraint,

$$\min_{\boldsymbol{g}} \| H_{\mathrm{b}} \boldsymbol{g} - \boldsymbol{p}_{\mathrm{des}} \|^2 \tag{5a}$$

$$\text{subject to} \| H_{\mathrm{d}} \boldsymbol{g} \|^2 \le D_0 \tag{5b}$$

$$\| \boldsymbol{g} \|^2 \le E_0 . \tag{5c}$$

The problem can then be written as a Lagrangian cost function,

$$\min_{\boldsymbol{g}} L_p(\boldsymbol{g}) = \| H_{\mathrm{b}} \boldsymbol{g} - \boldsymbol{p}_{\mathrm{des}} \|^2 + \lambda_1 (\| H_{\mathrm{d}} \boldsymbol{g} \|^2 - D_0) + \lambda_2 (\| \boldsymbol{g} \|^2 - E_0), \quad \lambda_1, \lambda_2 \ge 0, \tag{6}$$

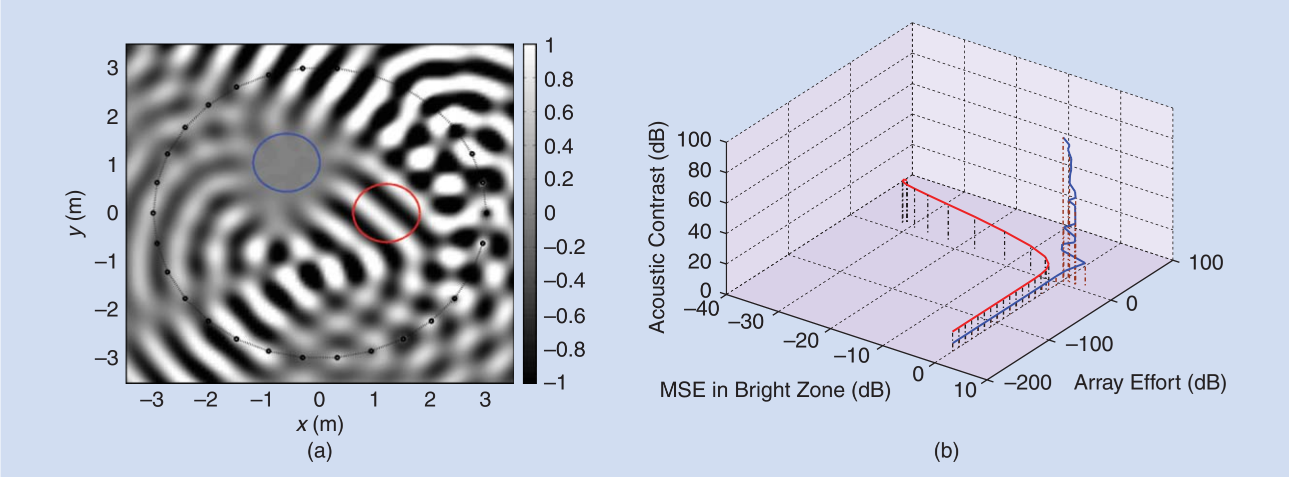
where again λ_1 and λ_2 are Lagrange multipliers. The solution that minimizes L_p is obtained by setting the derivative of L_p with respect to $\boldsymbol{g}$ to zero and is written as

$$[H_{\mathrm{b}}^H H_{\mathrm{b}} + \lambda_1 H_{\mathrm{d}}^H H_{\mathrm{d}} + \lambda_2 I] \boldsymbol{g} = H_{\mathrm{b}}^H \boldsymbol{p}_{\mathrm{des}} . \tag{7}$$

Equation (7) may be solved using an interior point algorithm to choose appropriate values of λ_1 and λ_2 to satisfy the constraints [11]. A simpler formulation is to set the parameter $\lambda_1 = 1$, which implies applying equal effort to matching the pressure in the bright zone and minimizing the energy in the dark zone. This gives the original formulation of multizone sound control as in [12] but has an added robustness constraint on the array effort, that is $\boldsymbol{g}_p = [H_{\mathrm{b}}^H H_{\mathrm{b}} + H_{\mathrm{d}}^H H_{\mathrm{d}} + \lambda_2 I]^{-1} H_{\mathrm{b}}^H \boldsymbol{p}_{\mathrm{des}}$. This solution is also identical to that of the ACC method provided that 1) the choice of target pressures in the bright zone is an ACC solution, $\boldsymbol{p}_{\mathrm{des}} = H_{\mathrm{b}} \boldsymbol{g}_{\mathrm{c}}$ and 2) identical constraints in E_0 and D_0 are met. This demonstrates that the formulation in the PM approach to sound field reproduction subsumes the ACC problem. Chang and Jacobsen [13] investigated a combined solution of these two techniques, $\boldsymbol{g}_{\mathrm{cb}} = [(1-\kappa) H_{\mathrm{b}}^H H_{\mathrm{b}} + \kappa H_{\mathrm{d}}^H H_{\mathrm{d}}]^{-1} (1-\kappa) H_{\mathrm{b}}^H \boldsymbol{p}_{\mathrm{des}}$, which is actually same as the one presented in (7) with the regularization term omitted. The tuning parameter κ is equivalent to the tuning parameter λ_1. The design has been shown effective for reproducing plane wave sound fields at frequencies even above the Nyquist frequency with good contrast control, thus providing the potential to reduced the number of loudspeakers required and increase the zone sizes and upper operating frequencies using the PM method.

The PM approach gives an explicit solution to obtain the loudspeaker driving signals and does not require solving an eigenvector problem, as is required in the case of acoustic contrast optimization. PM is especially suitable for the situation that different constraints are imposed on each sound zone when the listeners require different quality of listening experiences. However a series of Lagrange multipliers need to be determined, and a generalized eigenvalue solution is no longer possible. Instead convex-optimization methods like the interior-point method should be used [11]. The PM approach also imposes an objective on the phase of reproduced sound fields within the bright zone, and thus provides a better holographic image compared to the contrast control method. Figure 2(b) demonstrates that the ACC method always maintains a high level of contrast between the bright and dark zone using a small array effort, but a high reproduction error also indicates that the reproduced sound field may swirl around the listener in different directions. On the other hand, the PM approach achieves small reproduction error while higher contrast may be obtained by choosing an appropriate desired sound field. Preliminary perceptual tests were found to generally agree with the simulation results however the source signal itself significantly affects the system performance [14].

While the least squares solutions in the frequency domain seems to provide a great deal of simplicity and flexibility, the



[FIG2] A plane wave of 500 Hz from 45° is reproduced in the bright zone (red circle) using PM while deadening the sound in the dark zone (blue circle) using 30 loudspeakers placed on a circle of radius $R = 3$ m, and each zone is of radius $r = 0.6$ m as shown in (a). (b) The acoustic contrast versus the array effort and the mean-square reproduction error in the bright zone using the ACC method (blue line) and the PM method (red line).

positions of the loudspeakers and the matching points within sound zones must be chosen judiciously for good reproduction performance. Representing sound fields in the wave domain or mode domain as in (S1) in "Wave-Domain Sound Field Representation" can provide physical insights into these critical issues [15]. Dimensionality analysis tells us that for PM over Q sound zones, the number of loudspeakers required is determined by the upper frequency or wave number k of operation, the number of sound zones, and the size of each sound zone [15]. Here we assume that each sound zone is a circle or sphere of radius r_0 located at the origin O_q as shown in Figure 1, although without loss of generality each sound zone could be of arbitrary shape. The minimum number L is about $Q(2kr_0+1)$ for two-dimensional (2-D) reproduction and $Q(kr_0+1)^2$ for three-dimensional (3-D) reproduction, respectively [4].

DISCUSSION

Practical Implementation

When a small number of loudspeakers are used, for example, three speakers used in a mobile device, current personal audio systems can only achieve limited performance, i.e., ~10 dB contrast level between bright and dark zones [7]. An array of nine sources has been implemented for personal audio systems in televisions and personal computers, in an anechoic chamber achieving over 19 dB contrast under ideal conditions [6]. However, in terms of practical implementation in a car cabin, Cheer et al. [8] demonstrated that the optimized level of acoustic contrast obtained from the ACC method may not be achieved because of errors and uncertainties and the least-squares-based PM approach may provide a more robust solution. In addition, multizone reproduction is fundamentally constrained whenever attempting to reproduce a sound field in the bright zone that is directed to or obscured by another zone. This is known as the *occlusion problem* [11], [12].

Loudspeaker Positions

Using the compressive sensing idea, the formulation of multizone sound field reproduction can be regularized with the ℓ_1 norm of

WAVE-DOMAIN SOUND FIELD REPRESENTATION

The Helmholtz wave equation can be solved to express any sound field as a weighted sum of basis functions,

$$p(\mathbf{x},\omega)=\sum_{n=1}^{\infty}\alpha_n(\omega)\beta_n(\mathbf{x},\omega), \qquad \text{(S1)}$$

where $\alpha_n(w)$ are sound field coefficients corresponding to mode index n, $\beta_n(\mathbf{x},\omega)$ are basis functions with the orthogonality property

$$\langle \beta_n,\beta_m\rangle \triangleq \int_{\mathbb{C}} \beta_n^*(\mathbf{x},\omega)\beta_m(\mathbf{x},\omega)d\mathbf{x}=\xi_n(w)\delta_{nm}.$$

The sound field within a control region $\mathbb{C}$ can be represented using a finite number of basis functions, i.e., $n\in[1,N]$ and $\xi_n(w)=\langle\beta_n,\beta_n\rangle$ is the strength of each mode over the control zone.

The modal basis functions for source distributions and sound fields expressed in cylindrical coordinates and spherical coordinates can be written as [17]

$$p_{2\mathrm{D}}(\mathbf{x},\omega)=\sum_{\nu=-N}^{N}\alpha_\nu(\omega)\mathcal{J}_\nu^{(2\mathrm{D})}(kr)\exp(i\nu\phi) \qquad \text{(S2a)}$$

$$p_{3\mathrm{D}}(\mathbf{x},\omega)=\sum_{\nu=0}^{N}\sum_{\mu=-\nu}^{\nu}\alpha_\nu^\mu(\omega)\mathcal{J}_\nu^{(3\mathrm{D})}(kr)Y_\nu^\mu(\theta,\phi), \qquad \text{(S2b)}$$

where $\exp(\cdot)$ and $Y_\nu^\mu(\cdot)$ are complex exponentials and spherical harmonics, respectively and $\mathcal{J}_\nu^{(2\mathrm{D})}(kr)$ and $\mathcal{J}_\nu^{(3\mathrm{D})}(kr)$ are functions representing the 2-D and 3-D mode amplitudes at radius r, respectively. Given the radius of the control region r_0, wave number k, and the truncation number $N\approx kr_0$ [4], we have the following dimensionality results: $N_{2D}=2kr_0+1$ and $N_{3D}=(kr_0+1)^2$. This gives the Nyquist sampling condition for a uniform circular array $(M\geq N_{2D})$ and a spherical array $(M\geq N_{3D})$, respectively.

the loudspeaker weights and solved using the least-absolute shrinkage and selection operator [16]. The assumption here is that the desired sound field can be reproduced by a few loudspeakers, which are placed close to the direction of the virtual source and are sparsely distributed in space. This can produce low sound levels outside the bright zones and hence can reduce the interference to the dark zone.

Further Remarks

While the reproduction error has been widely used to quantify the performance of sound field rendering methods, a planar wavefront may be reproduced whose direction of propagation does not coincide with the desired direction, which may give a high reproduction error. In [18], the cost function of the ACC method is refined to optimize the extent to which a sound field resembles a plane wave. A constraint is imposed on the plane-wave energy within the bright zone over a range of incoming directions, thus optimizing the spatial aspects of the sound field for ACC. Simulation results demonstrate that a circular array of 48 equally spaced loudspeakers produces consistently high contrast and a planar target sound zone of radius 0.15 m for frequencies up to 7 kHz.

ACTIVE ROOM COMPENSATION

One challenge in the personal audio problem is room reverberation. A strong wall reflection may ruin the personal audio listening experience [14]. Room reverberation can be corrected by using active room compensation, provided the acoustic transfer function (ATF) matrices are determined. For static room environments these matrices may be premeasured but for time-varying environments they must be determined adaptively. In this section, methods for determining and correcting for these matrices to compensate for room responses over spatial regions are described.

The room compensation approaches described here are more robust at low frequencies. At high frequencies, a reverberant sound field is diffuse. Compensation is extremely sensitive to small changes within the room and cannot be practically compensated for without very fast filter adaptation. Personal sound systems may not be able to compensate for these variations. Instead, diffuse components may be treated as noise and the system made robust to them.

We summarize the advances made for the case of a single zone with the ATF matrix, $\boldsymbol{H} \equiv \boldsymbol{H}_1$, using wave-domain or modal-space processing. These approaches demonstrate the challenges inherent in applying room compensation to the multizone problem. We also review a crosstalk-cancelation approach to the multizone case that utilizes impulse response reshaping.

MODAL-SPACE PROCESSING

Based on the wave-domain sound field representation (S1), the sound field $p(\boldsymbol{x},\omega)$ can be expressed as in (3). The ATF $H_\ell(\boldsymbol{x},\omega)$ from each loudspeaker ℓ to a point $\boldsymbol{x}$ inside the sound control zone can also be parameterized as

$$H_\ell^{(2D)}(\boldsymbol{x},\omega) = \sum_{\nu=-N}^{N} \gamma_{\nu\ell}(\omega)\mathcal{J}_\nu^{(2D)}(kr)\exp(i\nu\phi), \tag{8a}$$

$$H_\ell^{(3D)}(\boldsymbol{x},\omega) = \sum_{\nu=0}^{N}\sum_{\mu=-\nu}^{\nu} \gamma_{\nu\ell}^{\mu}(\omega)\mathcal{J}_\nu^{(3D)}(kr)Y_\nu^{\mu}(\theta,\phi), \tag{8b}$$

where $\gamma_{n\ell}(w)$ and $\gamma_{\nu\ell}^{\mu}(\omega)$ are ATF coefficients. The sound pressure vector $\boldsymbol{p}$ and ATF matrix $\boldsymbol{H}$ can then be written in matrix form

$$\boldsymbol{p} = \boldsymbol{B}\boldsymbol{\alpha}, \tag{9a}$$

$$\boldsymbol{H} = \boldsymbol{B}\boldsymbol{\Gamma}, \tag{9b}$$

where $\boldsymbol{B}$ is the $M \times \mathcal{N}$ matrix of basis functions evaluated at each of the M microphone positions defined $[\boldsymbol{B}]_{mn} = \beta_n(\boldsymbol{x}_m,\omega)$, $\boldsymbol{\alpha}$ is an M-long vector of sound field coefficients, $\boldsymbol{\Gamma}$ is the $\mathcal{N} \times L$ matrix of the ATF coefficients defined $[\Gamma]_{n\ell} = \gamma_{n\ell}$, and $\mathcal{N}$ is either $\mathcal{N}_{2D}$ or $\mathcal{N}_{3D}$. The PM problem of (5a) in the mode domain becomes $\boldsymbol{\Gamma g} = \boldsymbol{\alpha}_{\text{des}}$, where $\boldsymbol{\alpha}_{\text{des}}$ is the $\mathcal{N}$-long vector of coefficients for the desired sound field. The compensation problem can then be solved in offline manner by determining the least-squares solution [19].

An adaptive mode-domain approach was devised in [20]. The ATF matrix can be further parameterized

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{J}\boldsymbol{\Gamma}, \tag{10}$$

where $\boldsymbol{U}$ is a tall Vandermonde matrix (2-D) or spherical harmonic matrix (3-D) with the property that $\boldsymbol{U}^H\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{J}$ is a diagonal matrix of the mode amplitudes at the microphone positions. The vector of microphones' signals $\boldsymbol{p} = \boldsymbol{Hg}$ are hence transformed into mode-domain coefficients through $\boldsymbol{\alpha} = \boldsymbol{J}^{-1}\boldsymbol{U}^H\boldsymbol{p}$. For modest levels of room reverberation, $\boldsymbol{\Gamma}$ can be expressed as the sum of an anechoic room component and a small reverberant component. By approximating the reverberation as small, a simple iterative procedure for choosing $\boldsymbol{g}$ to drive $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha}_{\text{des}}$ can be formulated. Reverberant compensation methods [19], [20] may have difficulties in practice with preringing artefacts, but these artefacts may be reduced by using more advanced multiple-input, multiple-output polynomial filter designs [21].

ACTIVE LISTENING ROOM COMPENSATION WITH WAVE-DOMAIN ADAPTIVE FILTERING

Active listening room compensation can be used to make a reverberant room problem look like an anechoic room problem [22]. By applying a compensation filter matrix to the input loudspeaker signals, the uncompensated anechoic-room driving signals can then be used. The essence of the problem as depicted in Figure 3 is to minimize the error energy $\boldsymbol{e}^H\boldsymbol{e}$, where

$$\boldsymbol{e} = \boldsymbol{H}_0\boldsymbol{g} - \boldsymbol{HCg},$$

$\boldsymbol{H}_0$ is the anechoic-room ATF matrix, and $\boldsymbol{C}$ is an $L \times L$ compensation filter matrix. This effectively chooses the filter matrix $\boldsymbol{C}$ to drive the net transfer function matrix $\boldsymbol{HC}$ to the anechoic-room ATF matrix $\boldsymbol{H}_0$.

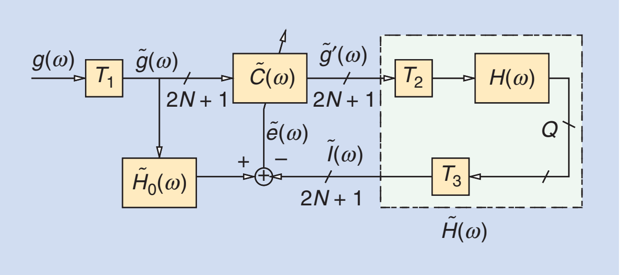
In massive multichannel problems for which the number of loudspeakers L and microphones M are large, the resultant matrices are large and may have issues with computational requirements (for filtered x-RLS) and convergence rates (for filtered x-LMS). Poor convergence can be solved using eigenspace adaptive filtering [22] by performing a generalized singular value decomposition (SVD) to diagonalize the system. Unfortunately the SVD still incurs a high computational cost.

Fortunately, the problem can be solved computationally and efficiently by using a wave-domain approach. If the microphones are arranged over a uniform circular array of radius r and the sources are arranged over a concentric uniform circular array, then the anechoic-room ATF matrix may be parameterized

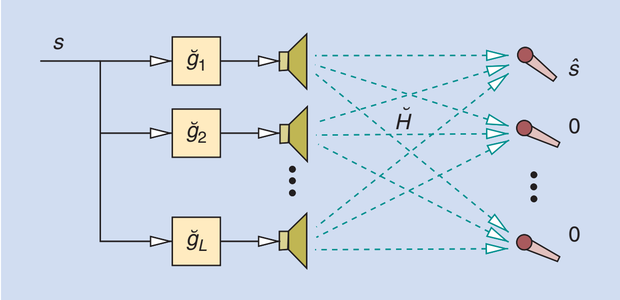
$$H_0 = UJ\underbrace{K^H V^H}_{\Gamma_0}, \tag{11}$$

where Γ_0 is a matrix of ATF coefficients corresponding to the anechoic room, K is a diagonal matrix of Hankel functions, and V is a tall Vandermonde matrix (2-D) or a spherical harmonic matrix (3-D). Matrix V possesses the property $V^H V = I$, provided that at least one loudspeaker is present for each mode to be controlled, i.e., $L \geq N_{2D}$ or $L \geq N_{3D}$.

The wave-domain adaptive filtering (WDAF) approach transforms the signals at the microphones and the loudspeaker signals into the wave domain through the transform $\mathcal{T}_1$ and $\mathcal{T}_3$, then adaptively calculates the mode-domain compensation signals $\tilde{C}(w)$, and transforms the compensated loudspeaker signals back using the inverse transform $\mathcal{T}_2$ as depicted in Figure 3. If the compensation filter matrix $\tilde{C}(w)$ is forced to be diagonal, then each of its diagonal entries can be determined from decoupled adaptive filters. This would explicitly solve the problems of computational complexity that appeared in multipoint compensation techniques. While it is straightforward to choose $\mathcal{T}_1$ and $\mathcal{T}_3$ to do so, in reality $\mathcal{T}_2$ cannot always be chosen without a priori knowledge of the ATF matrix. However, [22] and [23] show that the system can be partially diagonalized by choosing $\mathcal{T}_1 = V^H$, $\mathcal{T}_2 = V$, and $\mathcal{T}_3 = U^H$.



[FIG3] The listening room compensation using WDAF. The free-field transformed loudspeaker excitation signals $\tilde{\boldsymbol{g}}$ are used in a reverberant room with the filter matrix $\tilde{C}$ to compensate for the ATFs in matrix $\boldsymbol{H}$.



[FIG4] Crosstalk cancelation for delivering a time-domain signal $\boldsymbol{s}$ to the top microphone while canceling the signals at the remaining $Q-1$ microphones.

SYSTEM IDENTIFICATION OF THE ATF MATRIX

The ATFs change in a room as people move about and as temperature changes. Since active room compensation in particular is sensitive to this phenomenon, it is better if the ATFs are determined adaptively. Similar to active listening room compensation, this task can be performed efficiently in the wave domain where transforms are used to part-diagonalize the reverberant-room ATF matrix [23].

The advantages of WDAF and the mode-domain approaches are that 1) sound pressure is controlled over the entire control region and not just at specific points and 2) they represent the problem with a reduced number $N_{2D} < M$ (or $N_{3D} < M$) of parameters, which reduces the complexity and reduces the correlation in the elements of the ATF matrix since the filters are part decoupled. This helps speed the convergence of adaptive filtering.

Since many more microphones and loudspeakers are required for a 3-D control zone, active room compensation is more practically deployed in 2-D scenarios. However, 2-D compensation cannot satisfactorily correct for roof and floor reflections, so sound absorbers must be employed to eliminate these effects.

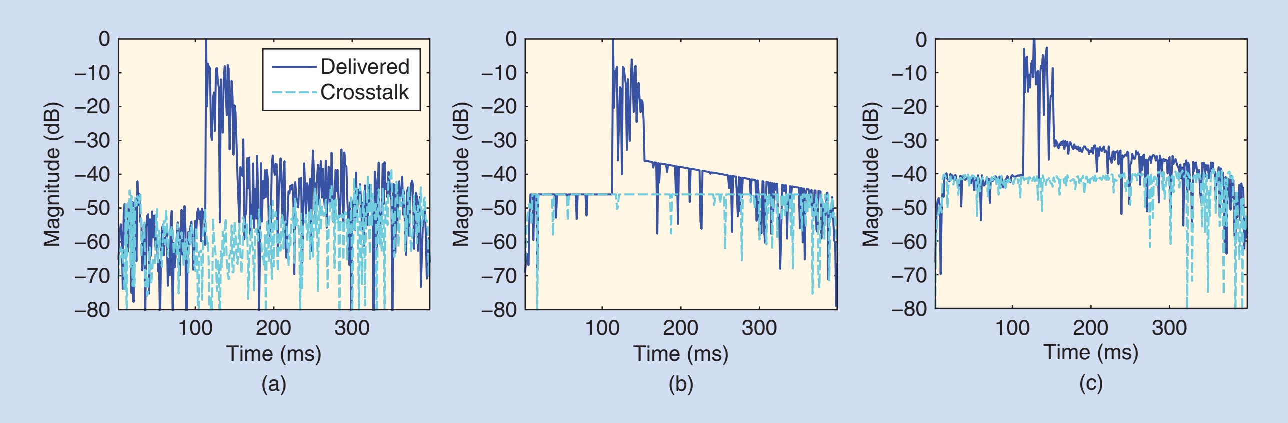
IMPULSE RESPONSE RESHAPING

Multiple listening zones may also be achieved by using crosstalk cancelation. Here, each of Q signal is delivered to a listening position while canceling the crosstalk paths to the remaining $Q-1$ positions using L loudspeakers and, for monaural signals, $M = 1$ microphone in each zone. As shown in Figure 4, this problem is solved by implementing crosstalk-cancelation filters. The basic idea of the impulse response reshaping approach is that fully equalizing the delivered paths is unnecessary. It is more robust and efficient to reshape these impulse responses.

Using impulse response reshaping, the early reflections of the delivered paths are reinforced while late reverberation and crosstalk are minimized [25]. Here, by defining windows on these desired and undesired ATF components $\boldsymbol{w}_q^{(\mathrm{d})}$ and $\boldsymbol{w}_q^{(\mathrm{u})}$ respectively in each zone q, the ratio of undesired-to-desired components is minimized

$$\min_{\breve{\boldsymbol{g}}} \log \frac{\| W_{\mathrm{u}} \breve{\boldsymbol{r}} \|_{p_u}}{\| W_{\mathrm{d}} \breve{\boldsymbol{r}} \|_{p_d}}, \tag{12}$$

where $\breve{\boldsymbol{r}}$ represents the global impulse response given a concatenated vector of crosstalk cancelation filters $\breve{\boldsymbol{g}} \triangleq [\breve{\boldsymbol{g}}_1^T, \ldots, \breve{\boldsymbol{g}}_L^T]^T$ and a block-Toeplitz matrix $\breve{\boldsymbol{H}}$ representing the room impulse responses, i.e., $\breve{\boldsymbol{r}} = \breve{\boldsymbol{H}}\breve{\boldsymbol{g}}$, $W_{\mathrm{u}} \triangleq \mathrm{Diag}(\boldsymbol{w}_1^{(\mathrm{u})}, \ldots, \boldsymbol{w}_Q^{(\mathrm{u})})$, and $W_{\mathrm{d}} \triangleq \mathrm{Diag}(\boldsymbol{w}_1^{(\mathrm{d})}, \ldots, \boldsymbol{w}_Q^{(\mathrm{d})})$. Different p_d- and p_u-norms may be chosen for the desired and undesired components, but it has been shown to be perceptually favorable to choose norms that approximate the



[FIG5] **The shortening of impulse responses to 50 ms in a room of reverberation time 250 ms using (a) relaxed multichannel least squares, (b) the relaxed minimax approach in [24], and (c) the ratio optimization approach of [25].**

infinite norm. Equation (12) can be solved analytically for the $p_u = p_d = 2$ case where it reduces to a generalized Rayleigh quotient. In general, (12) is solved using the steepest descent methods [25]. A relaxed multichannel approach using least squares [26] and minimax metrics [24] may include regularizations to reduce the array effort below that of the ratio-based approach in [25]. These approaches are compared in Figure 5 for simulation with $L = 3$ and $Q = 2$ in a room with a reverberation time of 250 ms using only 150 ms-long reshaping filters. The ratio-based approach shown is for $p_u = 10$, $p_d = 10$, and 1,000 steepest descent iterations.

Impulse response reshaping, in principle, can be applied to the PM- and modal-space approaches of creating personal sound zones. More robust and efficient filters can be obtained than equalization by canceling the undesirable late reverberation while leaving in some beneficial early reflections. Unfortunately this problem must be formulated in the time domain, which results in a computationally intractable massive multichannel problem. The future development of lower-complexity convex optimization algorithms may permit practical solutions to these large problems.

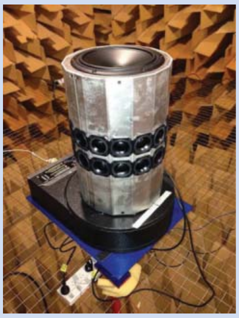
DIRECTIONAL SOURCES

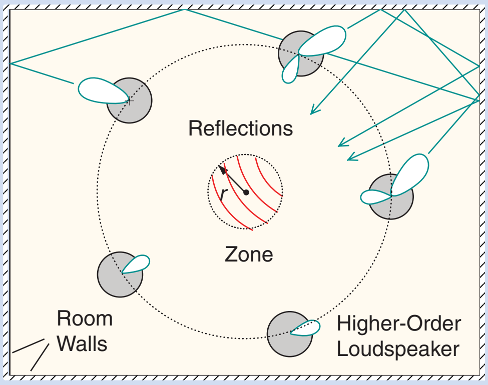
The use of directional sources can provide advantages over conventional loudspeakers, whose directivity is omnidirectional at low frequencies and is not typically controllable. Directional sources that provide multiple modes of sound radiation can be used with active compensation to produce sound arriving from angles where there are no sources by reflecting sound from room surfaces and can also cancel unwanted reverberation (Figure 6).

In a multilistener situation, a single directional loudspeaker can reduce unwanted radiation of sound to other listeners by maximizing the direct sound to the intended recipient relative to the reverberant field. A loudspeaker with directivity D and radiating acoustic power W in an ideal Sabinian space produces a direct sound intensity $I_{\text{dir}} = WD/(4\pi r^2)$ and a reverberant sound intensity of $I_{\text{rev}} = 4W/R'$, where $R' = S\epsilon/(1-\epsilon)$ is the room constant, S the room surface area, and ϵ the mean absorption coefficient of the room surfaces. The direct to reverberant intensity ratio is thus

$$\text{DRR} = \frac{DR'}{4\pi r^2}. \tag{13}$$

Increasing the directivity then allows the direct sound at the listener to be increased relative to the reverberant sound. Equivalently, the reverberant field is reduced by $1/DRR$.


(a)



(b)

[FIG6] **A demonstration of the higher-order loudspeaker in (a) a cylindrical baffle and (b) the schematic plot of its behavior.**

Standard loudspeakers typically have insufficient directivity to provide a significant enhancement of direct sound in a reverberant space. High directivity can be achieved using traditional array techniques such as delay and sum beamforming, but the array size must be large at low frequencies to achieve significant directivity. For practical use, superdirectional arrays are required, which achieve higher directivities than an array with uniform amplitude weightings [27]. Superdirectivity can be achieved using linear differential arrays, where the transducer weights have alternating signs, or by using circular and spherical arrays, where the weights are obtained from trigonometric or spherical harmonic functions, respectively. Such loudspeakers are termed *higher-order sources* (*HOSs*) and can produce multiple radiation patterns that are described by cylindrical or spherical harmonics.

Because superdirectional arrays are compact relative to their directivity, they may be built into a single unit, and we therefore assume here that a directional source is a single unit, typically of similar dimension to a standard loudspeaker. This section considers the design of directional loudspeakers and their application to maximum directivity and then focuses on the advantages of using arrays of sources, which allow greater control of sound fields over wide areas and are particular suitable for establishing personal sound zones.

SPHERICAL ARRAYS

The sound field produced by an arbitrary source of maximum radius a positioned at the origin and radiating a complex frequency $\exp(i\omega t)$ is represented in the wave domain as in (S2b) [17]

$$p(r,\theta,\phi,w) = \sum_{\nu=0}^{N}\sum_{\mu=-\nu}^{\nu} \alpha_\nu^\mu(w) h_\nu^{(2)}(kr) Y_\nu^\mu(\theta,\phi), r \geq a, \qquad (14)$$

where $h_\nu^{(2)}(kr)$ is the spherical Hankel function of the second kind, i.e., the radial function to represent the mode amplitude at r and $\alpha_\nu^\mu(w)$ are sound field coefficients. Similar to the dimensionality analysis in the wave domain, we will assume that the directivity of the source can be described by a maximum order N so that $\nu \in [0, N]$.

The most direct method for constructing a loudspeaker that can produce a controllable directivity is to mount a number of drivers in a spherical baffle of radius a [28]. The general behavior of such a source is most simply explained by deriving the sound field due to a sphere with arbitrary surface velocity

$$v(\theta_s,\phi_s,t,w) = e^{i\omega t}\sum_{\nu=0}^{N}\sum_{\mu=-\nu}^{\nu} \zeta_\nu^\mu(w) Y_\nu^\mu(\theta_s,\phi_s), \qquad (15)$$

where (θ_s,ϕ_s) is the driver position on the sphere. The exterior field has the general form of (14). The expansion coefficients are found by calculating the radial velocity for the general case, and requiring that they equal (15), i.e.,

$$\alpha_\nu^\mu(w) = -i\rho c \frac{\zeta_\nu^\mu(w)}{h_\nu'^{(2)}(ka)}$$

and the sound field, including the effect of mass-controlled drivers, is

$$p(r,\theta,\phi,t,w) = -\frac{i\rho c e^{i\omega t}}{k}\sum_{\nu=0}^{N}\sum_{\mu=-\nu}^{\nu} \zeta_\nu^\mu(w)\frac{h_\nu^{(2)}(kr)}{h_\nu'^{(2)}(ka)} \times Y_\nu^\mu(\theta,\phi), r \geq a.$$

Hence, each coefficient of the surface velocity produces a corresponding mode of radiation whose polar response is governed by a spherical harmonic.

The normalized magnitude of the mode responses for orders 0–5 are shown in Figure 7(a). For all modes greater than order $\nu = 0$, the response reduces at low frequencies. All modes of order ν become active at a frequency approximately given by $ka = \nu$ or

$$f = \frac{\nu c}{2\pi a}. \qquad (16)$$

This means that it is not possible to create high-order directivities at low frequencies. The spherical loudspeaker is omnidirectional at low frequencies and can produce increasing directivities as more modes become active above frequencies given by (16).

In practice, the surface velocity in (15) must be approximated using a discrete array of L_0 drivers positioned on the sphere. Ideally the drivers are positioned so that they are spaced equally from each other which produces the most robust approximation to the integration over the sphere required to approximate each spherical harmonic. This is possible if the drivers are placed in the center of the faces of platonic solids, allowing up to 20 drivers (for the icosahedron). Higher numbers of drivers can be used using numerically optimized integration nodes for the sphere.

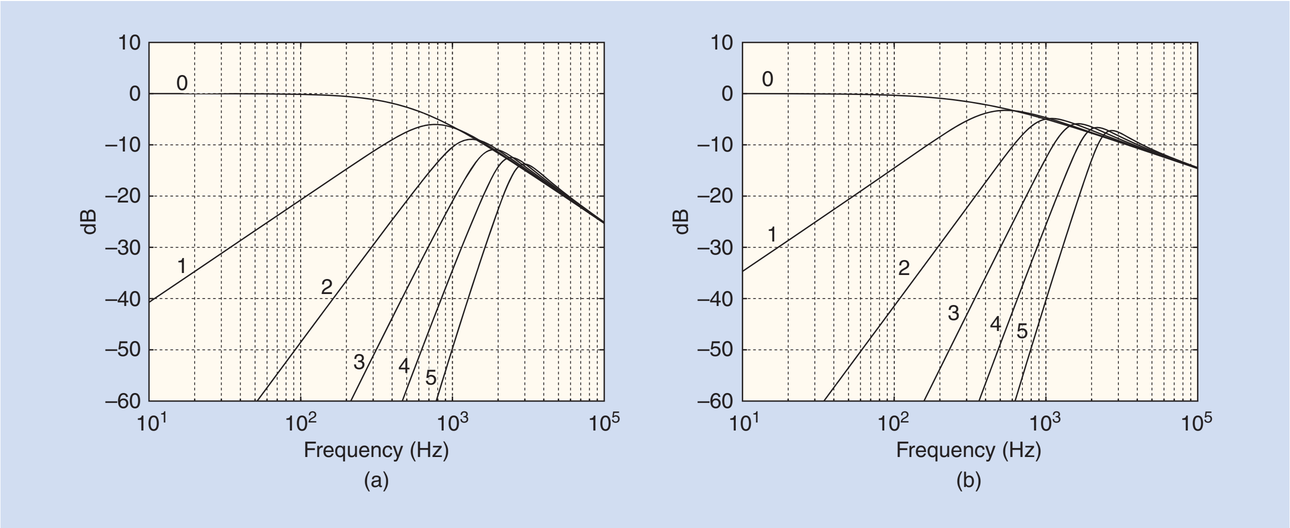
A simple way to model the discrete approximation is to assume each driver is a point source. The sound field due to a point source on a sphere then models a single driver, and the sound fields due to L_0 point sources allows the calculation of the total field. However, this approach ignores the directivity of each driver, which becomes significant at high frequencies. A more accurate model of the drivers that is mathematically tractable is to model each one as a spherical cap vibrating radially [28].

The sampling of the sphere means that the spherical loudspeaker is unable to generate spherical harmonic terms above the spatial Nyquist frequency of the array. This may be derived by noting that there are a total of $\mathcal{N} = (N+1)^2$ spherical harmonics up to order N. Controlling this number of modes using L_0 loudspeakers is possible for $L_0 \geq \mathcal{N}$. At a given frequency, the maximum-mode order that can be radiated is $N = ka$. Hence, the spatial Nyquist frequency is

$$f_{\mathrm{Nyq,3D}} = \frac{c(\sqrt{L_0}-1)}{2\pi a}. \qquad (17)$$

The number of drivers required for a sphere of radius a to produce Nth-order directional responses up to a frequency f is given by

$$L_{3D} = \left(\frac{2\pi a f}{c}+1\right)^2.$$



[FIG7] The normalized magnitude of the mode responses of (a) a spherical source and (b) a cylindrical source for orders 0–5.

For example, a third-order speaker with radius $a = 0.1$ m and a Nyquist frequency of 4 kHz would require 70 drivers. This is a large number of drivers, and motivates the investigation of simpler approaches such as cylindrical and line arrays.

CYLINDRICAL ARRAYS

A simpler approach may be taken if the directional loudspeaker is only required to produce directivity in a 2-D plane. This is commonly the case for sound reproduction in the home, where stereo and 5.1-surround formats are ubiquitous. A circular array requires fewer drivers than a spherical array for the same spatial Nyquist frequency. To see this, consider a sphere where L_0 drivers are placed on the equator instead of equally spaced around the sphere. This arrangement allows for the generation of sectorial spherical harmonics, where $\nu = |\mu|$, which produce radiation with lobes only in the (x, y) plane. The driver spacing is now $2\pi a/L_0$ and the spatial Nyquist frequency is

$$f_{\text{Nyq},2D} = \frac{c\,(L_0 - 1)}{4\pi a}. \qquad (18)$$

The number of drivers for a given 2-D spatial frequency is

$$L_{2D} = \frac{4\pi a f}{c} + 1.$$

Comparing (18) with (17), the 2-D Nyquist frequency can be much higher than the 3-D Nyquist frequency for the same number of drivers. The limitation of the circular array is that the transducer layout does not provide sufficient vertical directivity at high frequencies, and the source begins to produce unwanted radiation lobes in elevation. To reduce these lobes, the transducers must either have greater aperture in elevation or a line array must be used to control the vertical directivity. Since a line array is more effective when mounted on a cylinder than on a sphere, a practical alternative to the spherical array for the 2-D case is a cylindrical baffle in which multiple circular arrays are mounted (Figure 6). Such a geometry can still use fewer transducers than the spherical case, for the same spatial Nyquist frequency.

The radiation of sound for the cylindrical case can be approximated by assuming that the cylinder is infinite and that each driver is represented as a surface velocity distribution in height z and azimuth angle ϕ [29]. Its produced mode responses are shown in Figure 7(b). The responses are similar to those for the spherical source, and the activation frequencies are the same. The limitation of this analysis is that, in practice, a truncated cylinder must be used leading to variations of the mode response magnitude around the infinite cylinder values due to diffraction from the ends of the cylinder.

LINE ARRAYS

The simplest array for providing high directivity is a line array, which produces an axisymmetric polar response. While this does not provide the full control of 3-D or 2-D radiation that the spherical and cylindrical arrays do, it is sufficient for maximizing the direct to reverberant ratio. It has the same limitation as the circular and spherical arrays, that is difficult to create high-order responses at low frequencies. However, the line array allows an order N response to be produced using $L_0 = N + 1$ transducers as opposed to $(N + 1)^2$ using a spherical array or $2N + 1$ for a circular array (assuming no vertical directivity control). The maximum directivity produced in 3-D is [30]

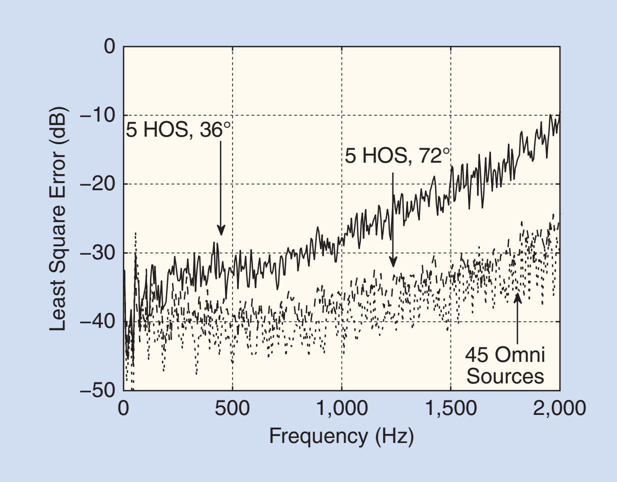
$$D = (N + 1)^2.$$

An order N loudspeaker with this directivity will produce the maximum direct to reverberant ratio for an on-axis listener. The simplest case, $N = 1$, results in a polar response $p(\theta) = 0.25 + 0.75\cos(\theta)$, which has a directivity of four [7]. The first-order response can be implemented using $N = 2$ coupled or uncoupled drivers, or more simply, using a single driver and controlling the radiation from the rear of the driver, although the directivity results can be less accurate with frequency [7].

ARRAYS OF DIRECTIONAL SOURCES

If multiple directional loudspeakers are available, then it becomes possible to create multiple zones of sound. Multizone reproduction requires a large number of monopole loudspeakers. The use of directional sources allows the production of multizone fields using significantly fewer loudspeaker units. In effect, a large number of drivers are grouped into a small number of physical devices to allow the creation of complex sound fields.

It has been shown that an array of L Nth order sources operating in free-field conditions has a spatial Nyquist frequency of approximately $2N$ times that of the same geometry monopole array [31]. Results better than free-field can be achieved in a reverberant room by using the techniques discussed in [32]. In this case, the directional sources are able to exploit room reflections to provide directions of arrival other than those directly from the sources. The use of L HOSs, each of which can produce up to order N responses, can produce a similar accuracy of a reconstructed field to $L(2N+1)$ monopole loudspeakers in the 2-D case, and $L(N+1)^2$ loudspeakers in the 3-D case. For example, Figure 8 shows the sound field reproduction error achieved using a circular array of five higher-order loudspeakers in comparison with an array of 45 monopole sources. For a virtual source angle of 72° (the desired source position is equal to the first loudspeaker position), the error is similar to that produced by the monopole sources. At the angle of 36° (the desired source halfway between two loudspeakers), the error is about 10 dB higher than the monopole case but still reasonably accurate, particularly at low frequencies. Reproduction has been achieved over a 1-m diameter using only five loudspeaker units with room dereverberation. The simulation is limited to 2-kHz bandwidth for computational complexity reasons. The worst-case reproduction error will be below -10 dB up to around 3 kHz. The bandwidth and reproduction radius of accurate reproduction can be extended by using more sources and higher orders, creating sufficient space for multiple listeners listening to independent sound fields.



[FIG8] The least squares error of reproduction as a function of frequency for an array of five fourth-order sources at 36° exactly between a pair of loudspeakers (dashed) and 72° coinciding with a loudspeaker (dashed), and a circular array of 45 omnidirectional line sources (unbroken) in a 2-D rectangular room of dimensions 6.4 × 5 m and with wall reflection coefficients of 0.7.

The use of HOSs can be viewed as an optimization problem with a constraint on the total number of loudspeaker units in the room. The only way to improve reproduction in such a case is to add capability to the existing loudspeakers. HOSs offer a practical approach to providing the control of the high-spatial-dimension sound fields that are required for creating multiple personal sound zones. For example, the reproduction of sound in Q zones of radius r_0, up to a spatial frequency $k_{\max}$, using L HOSs requires a maximum order per source of

$$N = \left\lceil \frac{Q(k_{\max} r_0 + 0.5)}{L} - 0.5 \right\rceil. \qquad (19)$$

For 8 kHz reproduction over regions of radius 0.2 m, the order is $N = 10$ for $L = 10$ sources and $N = 6$ for $L = 15$ sources. Such numbers are achievable in moderate- to large-sized rooms.

SUMMARY AND FUTURE OPPORTUNITIES

In this article, we presented, according to our involvement and insights, the audio processing and loudspeaker design aspects that support the goal of establishing personal sound zones. The problems that have been explored include multizone sound control, wave-domain active room compensation, and directional loudspeaker design, which allow for sound control over spatial regions. A high-performance personal audio system would likely address many of these aspects in its design. In sound field control, interference mitigation and room compensation robust to changes and uncertainty in the acoustic environment remain as challenging problems. Yet future opportunities exist in 1) higher-order surround sound using an array of directional sources and wave-domain active room compensation to perform multizone sound control in reverberant enclosures and 2) personal audio devices using multiple sensors to establish personal sound zones by efficiently canceling crosstalk and using distributed beamforming.

AUTHORS

Terence Betlehem (Terence.Betlehem@callaghaninnovation.govt.nz) received the B.S., B.E., and Ph.D. degrees in telecommunications engineering from the Australian National University (ANU) in 1998, 2000, and 2005 respectively. From 2005 to 2006, he was a research fellow at the ANU Research School of Information Sciences and Engineering and a visiting researcher at National Information and Communication Technology Australia working in the areas of spatial signal processing and wireless channel modeling. Since 2007, he has worked at Callaghan Innovation (formerly Industrial Research Limited) in Lower Hutt, New Zealand, in the areas of spatial audio and wireless communications, where he is currently a senior research engineer. His research interests

include active room compensation, microphone array processing, and room acoustic modeling.

Wen Zhang (wen.zhang@anu.edu.au) received the M.E. and Ph.D. degrees in electrical engineering from the Australian National University in 2005 and 2010, respectively. She worked as a Commonwealth Scientific and Industrial Research Organization (CSIRO) Office of the Chief Executive postdoctoral fellow at the CSIRO Process Science and Engineering Division in Sydney from 2010 to 2012. She is currently a research fellow at the College of Engineering and Computer Science at the Australian National University. Her research interests include spatial audio, source separation and localization, and active noise cancelation. She is currently an affiliate member of the Audio and Acoustics Signal Processing Technical Committee of the IEEE Signal Processing Society.

Mark A. Poletti (Mark.Poletti@callaghaninnovation.govt.nz) received an M.S. degree in physics at the University of Auckland in 1984, then worked at the Acoustics Research Centre at Auckland University for five years, where he was involved in acoustics testing and signal processing research. In 1989, he joined the Department of Scientific and Industrial Research communications group in Lower Hutt, New Zealand. This group became a part of Industrial Research Limited in 1992, which became Callaghan Innovation in 2013, where he is currently employed. In the 1990s, he developed the Variable Room Acoustic System (now called the Constellation System) for the electroacoustic enhancement of room acoustics. This work was the topic of his Ph.D. dissertation. His current research interests include electronic enhancement of room acoustics, holographic sound recording and reproduction systems using higher-order loudspeakers, and virtual acoustics systems.

Thushara D. Abhayapala (thushara.abhyapala@anu.edu.au) received the B.E. degree in interdisciplinary systems engineering and the Ph.D. degree in telecommunications engineering from the Australian National University (ANU), Canberra, in 1994 and 1999, respectively. Since December 1999, he has been a faculty member at ANU. His research interests are in the areas of spatial audio and acoustics signal processing and array signal processing. He is an associate editor of *IEEE/ACM Transactions in Audio, Speech, and Language Processing*. He is a member of the Audio and Acoustic Signal Processing Technical Committee (2011–2015) of the IEEE Signal Processing Society.

REFERENCES

[1] M. Camras, "Approach to recreating a sound field," *J. Acoust. Soc. Amer.*, vol. 43, no. 6, pp. 1425–1431, 1967.

[2] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc.*, vol. 33, no. 11, pp. 859–871, 1985.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2764–2778, 1993.

[4] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[5] W. F. Druyvesteyn and J. Garas, "Personal sound," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 685–701, 1997.

[6] J.-H. Chang, C.-H. Lee, J.-Y. Park, and Y.-H. Kim, "A realization of sound focused personal audio system using acoustic contrast control," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2091–2097, 2009.

[7] S. J. Elliott, J. Cheer, H. Murfet, and K. R. Holland, "Minimally radiating sources for personal audio," *J. Acoust. Soc. Amer.*, vol. 128, no. 4, pp. 1721–1728, 2010.

[8] J. Cheer, S. J. Elliott, and M. F. S. Gálvez, "Design and implementation of a car cabin personal audio system," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 414–424, 2013.

[9] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1695–1700, 2002.

[10] S. J. Elliott, J. Cheer, J.-W. Choi, and Y.-H. Kim, "Robustness and regularization of personal audio systems," *IEEE Trans. Audio Speech Lang. Processing*, vol. 20, no. 7, pp. 2123–2133, 2012.

[11] T. Betlehem and P. D. Teal, "A constrained optimization approach for multizone surround sound," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Prague, Czech Republic, May 2011, pp. 437–440.

[12] M. A. Poletti, "An investigation of 2D multizone surround sound systems," in *Proc. 125th Audio Engineering Soc. Convention*, San Francisco, CA, Oct. 2008, pp. 1–9.

[13] J.-H. Chang and F. Jacobsen, "Sound field control with a circular double-layer array of loudspeaker," *J. Acoust. Soc. Amer.*, vol. 131, no. 6, pp. 4518–4525, 2012.

[14] M. Olik, J. Francombe, P. Coleman, P. J. Jackson, M. Olsen, M. Møller, R. Mason, and S. Bech, "A comparative performance study of sound zoning methods in a reflective environment," in *Proc. 52th AES Conf. Sound Field Control*, Guildford, U.K., Sept. 2013, pp. 1–10.

[15] Y. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Trans. Audio Speech Lang. Processing*, vol. 19, no. 6, pp. 1711–1720, 2011.

[16] N. Radmanesh and I. S. Burnett, "Generation of isolated wideband sound field using a combined two-stage Lasso-LS algorithm," *IEEE Trans. Audio Speech Lang. Processing*, vol. 21, no. 2, pp. 378–387, 2013.

[17] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. San Diego, CA: Academic, 1999.

[18] P. Coleman, P. Jackson, M. Olik, and J. A. Pederson, "Optimizing the planarity of sound zones," in *Proc. 52nd Audio Engineering Society Int. Conf.*, Guildford, U.K., Sept. 2013, pp. 1–10.

[19] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2100–2111, 2005.

[20] D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Multi-channel adaptive room equalization and echo suppression in sound field reproduction," *IEEE/ACM Trans. Audio Speech Lang. Processing*, vol. 22, no. 10, pp. 1522–1532, 2014.

[21] L.-J. Brannmark, A. Bahne, and A. Ahlen, "Compensation of loudspeaker-room response in a robust MIMO control framework," *IEEE Trans. Audio Speech Lang. Processing*, vol. 21, no. 6, pp. 1201–1216, 2013.

[22] S. Spors, H. Buchner, R. Rabenstein, and W. Herbordt, "Active listening room compensation for massive multichannel sound reproduction systems," *J. Acoust. Soc. Amer.*, vol. 122, no. 1, pp. 354–369, 2007.

[23] M. Schneider and W. Kellermann, "Adaptive listening room equalization using a scalable filtering structure in the wave domain," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Kyoto, Japan, May 2012, pp. 13–16.

[24] T. Betlehem, P. Teal, and Y. Hioka, "Efficient crosstalk canceler design with impulse response shortening filters," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 393–396.

[25] J. O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. Audio Speech Lang. Processing*, vol. 20, no. 6, pp. 1829–1842, 2012.

[26] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Int. Workshop on Acoustic Signal Enhancement*, Tel Aviv, Israel, Aug. 2010.

[27] G. W. Elko, "Differential microphone arrays," in *Audio Signal Processing for Next-Generation Multimedia*. Norwell, MA: Kluwer, 2004, pp. 11–65.

[28] B. Rafaely and D. Khaykin, "Optimal model-based beamforming and independent steering for spherical loudspeaker arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2234–2238, 2011.

[29] M. Poletti and T. Betlehem, "Design of a prototype variable directivity loudspeaker," in *Proc. 52th AES Conf. Sound Field Control*, Guildford, U.K., Sept. 2013, pp. 1–10.

[30] A. T. Parsons, "Maximum directivity proof for three-dimensional arrays," *J. Acoust. Soc. Amer.*, vol. 82, no. 1, pp. 179–182, 1987.

[31] M. A. Poletti and T. D. Abhayapala, "Spatial sound reproduction systems using higher order loudspeakers," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Prague, Czech Republic, May 2011, pp. 57–60.

[32] T. Betlehem and M. A. Poletti, "Two dimensional sound field reproduction using higher-order sources to exploit room reflections," *J. Acoust. Soc. Amer.*, vol. 135, no. 4, pp. 1820–1833, 2014.

[Vesa Välimäki, Andreas Franck, Jussi Rämö, Hannes Gamper, and Lauri Savioja]

Assisted Listening Using a Headset

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

[Enhancing audio perception in real, augmented, and virtual environments]

Historically, headphones have mainly been used for analytic listening in music production and in homes. During the last decade, with the boom of dedicated music players and mobile phones, the everyday use of light headphones has become highly popular. Current headphones are also paving the way for more sophisticated assisted listening devices. Today, active noise control (ANC), equalization techniques, and a hear-through function are already a standard part of many headphones that people commonly use while traveling. It is not difficult to predict that, in the near future, a headset will be a "hearing aid for those with normal hearing," which can improve listening conditions for example in a noisy environment.

Additionally, mobile augmented reality has become a hot topic [1], and new products such as the Google Glass will make it more common. On the audio side of augmented reality systems, mixing of the ambient and reproduced sounds will be an essential feature. Augmented reality headsets may also serve as the main user interface for the disappearing computer in the future, when visual displays and tangible keyboards vanish.

This article gives an overview of various signal processing techniques needed in assisted listening. The basic use case and various others are described in Figure 1, which shows how headphone listening can be extended by incorporating external microphones and some signal processing. Assisted listening in heavy background noise environments, such as in an airplane, can be implemented using ANC [2].

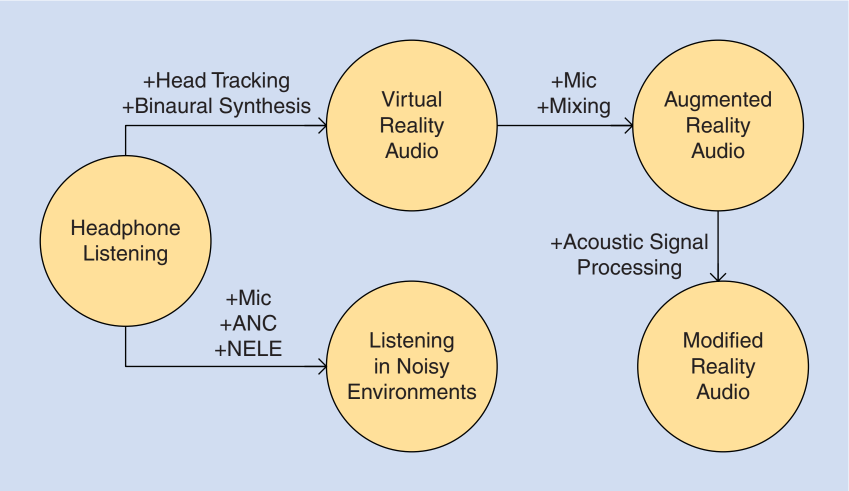
HEADPHONE LISTENING IN A NOISY ENVIRONMENT

When headphones are used in a noisy environment, their design goals are somewhat different than those of a conventional pair of high-fidelity (hi-fi) headphones. The most important feature of mobile headphones is the noise isolation capability, which can be passive or enhanced with ANC. Furthermore, the headphone frequency response can be designed to manage noisy environments, typically by boosting the bass end of the response, as natural ambient sounds have most of their energy at low frequencies (such as bus or airplane noise) and headphones usually attenuate low frequencies the least.

Digital Object Identifier 10.1109/MSP.2014.2369191
Date of publication: 12 February 2015

1053-5888/15©2015IEEE

Figure 2 shows measured isolation curves of different types of headphones, where the black solid line is the isolation curve of an open-back circum-aural (CA) hi-fi headphone, the green dashed-dotted line is the isolation of a closed-back supra-aural (SA) headphone, the red dashed line shows the passive isolation curve of an in-ear (IE) headphone, and the purple dashed line shows the active isolation of the same IE headphone, i.e., when the ANC is turned on.



[FIG1] The main use cases of a headset are natural listening, listening in noise, virtual reality, augmented reality, and modified reality.

ANC actively reduces the ambient noise leaked into the ear canal by introducing an antinoise signal. Ideally, the antinoise signal has the same magnitude and opposite phase as the noise signal inside the ear canal. The two basic operation principles of ANC are feedforward and feedback control [2]. The main difference between these two types is the position where the ambient noise is captured. Feedforward ANC typically uses two microphones: an external microphone capturing ambient noise, and an error microphone inside the headset, which is used to adaptively tune the ANC filtering. A feedback ANC uses only the internal error microphone both for capturing ambient sounds and for adapting the ANC system. Moreover, the feedback and feedforward structures can also be used simultaneously.

IN THE NEAR FUTURE, A HEADSET WILL BE A "HEARING AID FOR THOSE WITH NORMAL HEARING," WHICH CAN IMPROVE LISTENING CONDITIONS FOR EXAMPLE IN A NOISY ENVIRONMENT.

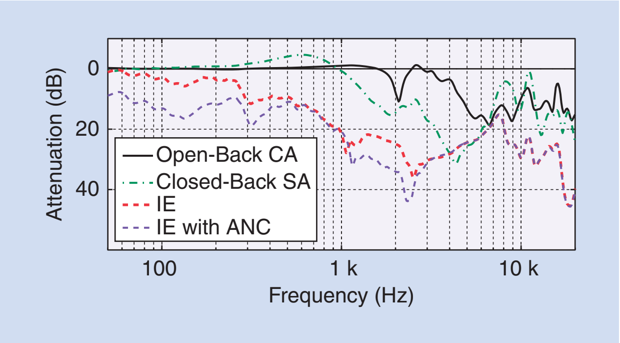
At its most basic, the frequency response equalization can be a bass boost to compensate for the auditory masking occurring at low frequencies. However, more intelligent systems also exist, e.g., for speech in mobile communication applications in noise [3], for car audio [5], and for headphone listening in noisy environments [4]. The goal in [3] is to maximize the speech intelligibility index by enhancing the clean far-end speech signal for the near-end listener who is situated in a noisy environment, hence the abbreviation NELE, which stands for "near-end listening enhancement." NELE techniques typically exploit psychoacoustically justifiable spectral resolution, such as Bark or equivalent rectangular bandwidth (ERB) scale, and an auditory masking model.

A similar idea is suggested in [4], where a clean wideband music signal is enhanced based on the content of the music, the characteristics of the headphones, and the ambient noise around the user. This system uses a microphone outside the headset to register the ambient noise, as shown in Figure 3. The characteristics of the headphones, which are measured beforehand, are used to estimate the level of the music and noise at the eardrum. These levels depend on the frequency and isolation responses of the headphones, respectively. The music and ambient noise are analyzed in Bark bands. The goal is to estimate the masking threshold from the captured ambient noise signal, and then to enhance only those frequency bands of the music signal which are below the masking threshold (completely masked) or just above it (partially masked), as shown in Figure 4. Using the unmasking process, the timbral balance of the music is enhanced in the presence of ambient noise and at the same time the volume increase is minimized, since only those bands that actually need amplification are boosted [4].

VIRTUAL REALITY LISTENING

Virtual auditory content can be displayed spatially, e.g., for guidance or navigation, by processing the sound so that it mimics the experience of listening to a real, physical sound source. Rendering a virtual sound source in such a way is referred to as *binaural synthesis* [6].

Binaural synthesis works by providing localization cues to the listener's ears [7], [8]. The delay and attenuation due to sound propagation, resulting in time and level differences between the ear signals, are the two main cues. They are termed the *interaural time difference* (*ITD*) and *interaural level difference* (*ILD*), respectively, and depend on the relative direction and, for nearby

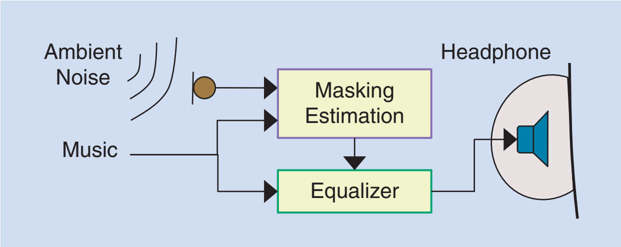


[FIG2] Measured isolation curves of different headphones.

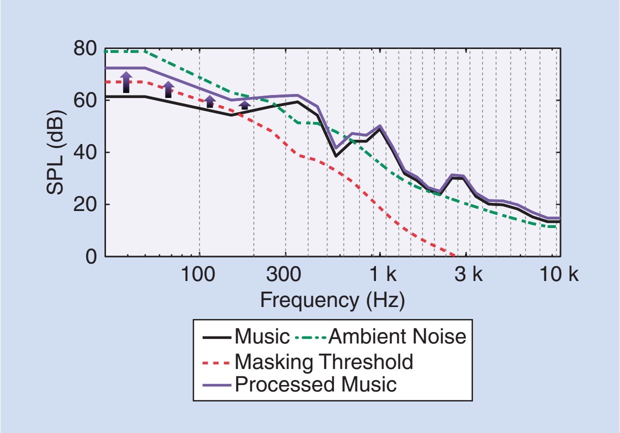
sources, also on the distance of the source. Spectral cues (SCs), caused by the reflection or diffraction by the listener's torso and pinnae, provide additional localization information. The ITD, ILD, and SCs are conveniently represented by head-related transfer functions (HRTFs), which model the free-field acoustic transfer functions between the position of the source in space and the listener's ears [9]. Thus, in its elementary form, binaural synthesis is performed by filtering a source signal with the respective HRTFs for the left and right ears.

[BINAURAL SYNTHESIS IS PERFORMED BY FILTERING A SOURCE SIGNAL WITH THE RESPECTIVE HRTFS FOR THE LEFT AND RIGHT EARS.]

Adding room reflections and reverberation to binaural synthesis improves the naturalness, source localization properties, and externalization [6], [8], [10]. Discrete room reflections improve the perception of both direction and distance of a sound source. They can be incorporated into binaural synthesis by rendering them as additional virtual sources obtained from a geometric model, e.g., the mirror-image source model. This, however, significantly increases reproduction complexity. Late diffuse reverberation mainly contributes to the perception of distance, because the ratio between direct and reverberant energy decreases with increasing source distance [8]. A recent review paper explains various artificial reverberation techniques [11].



[FIG3] The unmasking of an audio signal disturbed by ambient noise can be implemented with adaptive equalization, which uses the external microphone and knowledge of the characteristics of the headphones [4].



[FIG4] An example of the unmasking process in Bark bands, where the black line is the energy of the music, the green dash-dotted line is the energy of the ambient noise, the red dashed line is the estimated masking threshold, and the purple line is the spectrum of the unmasked music. The vertical dashed lines show the Bark band edges.

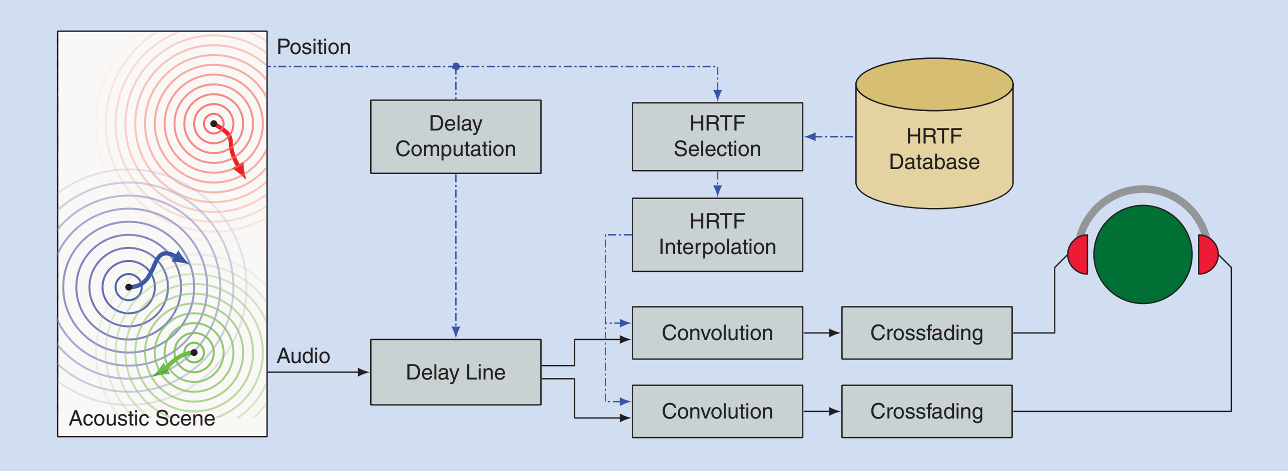
In a natural listening environment, dynamic head movements contribute significantly to source localization. To exploit these so-called dynamic binaural cues, a binaural synthesis system for virtual reality listening must provide several functionalities. First, the position and orientation of the listener's head have to be determined continuously, e.g., by using a head tracker. Second, the synthesis has to be adapted dynamically by updating the HRTFs according to the relative position of the virtual sources. Finally, the overall latency of the reproduction system must satisfy perceptual limits to provide the intended localization and a plausible listener's experience. The situation becomes more complicated if the spatial audio content is to be broadcasted as the standards for spatial audio are still under development.

HEAD TRACKING

In many applications, it is essential that the orientation of the user's head is known. This is achieved by head-tracking techniques. Once the listener's orientation is known, it is possible to use HRTFs to project the sound sources in correct directions. This is especially important in interactive applications in which the soundscape should remain stable even when the user turns his/her head. Without head tracking, the soundscape moves with respect to the user's head, breaking the illusion of virtual sources. Head tracking also helps in reducing reversals in sound source localization [10].

The same head-tracking techniques can be used both for visual head-mounted displays and headphones for audio reproduction. For the visual domain, one of the first head-tracking systems was presented by Ivan Sutherland in a virtual reality installation in 1968 [12]. This early system utilized mechanical and ultrasound tracked techniques. Jens Blauert, the pioneer of head tracking for audio reproduction over headphones, presented in his patent in 1973 several alternative techniques for head tracking, including the use of mechanical levers as well as magnetic and gyroscopic control arrangements [13]. Even today, most of those techniques are utilized in practice as one can buy tracking systems that are based on electromagnetic, inertial, or computer-vision techniques, or in their combination, such as in the new Oculus Rift (Crystal Cove prototype) virtual reality headset.

The category of computer-vision-based head tracking contains two different approaches. The most common one uses external infrared light emitters and reflective markers attached to the user. A less intrusive technique is the use of regular cameras without any special markers or light source, but it is more challenging to implement reliably. While both of these techniques can provide accurate and wireless tracking, they require the line-of-sight to the user, which is not needed with electromagnetic or inertial tracking. Smartphones, with their embedded video camera,



[FIG5] The signal flow of a dynamic binaural synthesis system for multiple sound sources.

provide the required hardware for computer-vision-based head tracking for mobile applications.

DYNAMIC BINAURAL SYNTHESIS

HRTFs can be obtained by measuring the acoustic path from a point in space to the ear entrances of a test subject or dummy head. Alternatively, HRTFs of an individual can be estimated from anthropometric data or by using listening tests. These methods are reviewed in the article "Natural Sound Rendering for Headphones" on page 98 of this special issue of *IEEE Signal Processing Magazine* [14]. Existing and upcoming standards for broadcasting spatial audio, e.g., MPEG Spatial Audio Object Coding or MPEG-H 3D Audio address the need for suitable HRTFs by either providing interfaces to supply individualized data or by transmitting predefined data sets in the encoded bit stream.

HRTF measurements are typically performed at discrete locations on a spherical grid centered around the test subject's head. Using such measurements directly for binaural synthesis would impose the same discrete grid on the virtual source positions. However, HRTFs at nonmeasured positions can be estimated from available measurements via interpolation, allowing to place and move virtual sources freely inside the measurement grid. For hi-fi rendering, the measurement grid should cover all or most of the sphere surrounding the listener and have a spatial resolution of 5–15° in elevation and 4–5° in azimuth, with fewer measurement points required toward extreme elevations [15].

Different types of preprocessing, either in the time or the frequency domain, are typically applied to the measured HRTF data [16]. Equalization techniques such as free-field or diffuse-field equalization compensate for the response of the measurement or reproduction system. Smoothing of HRTF data decreases perceptually irrelevant fluctuations, thus reducing the complexity of the frequency responses, enabling more efficient filtering and smoother interpolation between HRTFs.

Several approaches for HRTF interpolation have been proposed in the literature, including linear interpolation of neighboring HRTFs, spherical splines, and spherical harmonics. The advantage of linear interpolation over more sophisticated approaches is the reduced complexity in terms of implementation and computation, which can be a decisive factor in real-time applications. Linear interpolation is typically performed via a weighted combination of a subset of measured HRTFs lying close to the desired spatial location.

Publicly available HRTF databases are typically measured at locations on the surface of a sphere, based on the assumption that HRTFs are distance-independent further than about 1 m from the head of the listener [17]. For HRTFs measured on a sphere, the measurement points can be grouped into nonoverlapping triangles via triangulation. The interpolation is then performed by combining the HRTFs forming the triangle enclosing the location to be estimated. For measurement points obtained at various distances, triangulation yields a mesh of nonoverlapping tetrahedra. To estimate the HRTFs at a nonmeasured location, the HRTFs forming a tetrahedron enclosing the location to be estimated are interpolated. The weights for interpolating HRTFs forming a triangle or tetrahedron can be calculated from barycentric coordinates [18].

Once a suitable subset has been determined and the interpolation weights have been calculated, the actual interpolation is performed. A direct weighted addition of the selected HRTFs, which is equivalent to a linear combination of the corresponding impulse responses due to the linearity of the Fourier transform, typically leads to severe comb-filtering artifacts. This is due to the combination of transfer functions with different phases. Several approaches have been proposed to overcome this problem. A typical signal flow for dynamic synthesis, which contains the basic building blocks for interpolation and application of HRTF filters, is depicted in Figure 5. The main functionalities are the handling of time delays, interpolation of frequency responses, convolution with the source signals, and crossfading to enable smooth transitions between different HRTFs.

The separate handling of time delays, which are either extracted from the HRTF data set in a preprocessing step or from geometrical models, e.g., a spherical head model [19], yields several advantages. First, it lowers the required filter orders of the HRTFs, reducing computational and memory requirements. Second, the phase differences between neighboring HRTFs are significantly reduced, which alleviates the comb-filtering artifacts during interpolation. Finally, the perceptual limits for the ITD necessitate variable time delays with subsample accuracy, which are best implemented using fractional-delay filtering techniques, e.g., [20].

Several strategies exist to interpolate the HRTF responses (without delays). Interpolating the magnitude and phase responses separately preserves the complex-valued responses of HRTF filters [6]. Other approaches make use of the physical properties of (delay-compensated) HRTFs, which closely resemble minimum-phase systems [21], or the limited perceptual relevance of the phase [19], [22]. Interpolation of the magnitude responses followed by minimum-phase reconstruction is proposed in [6] and [16]. Another method is to interpolate only the HRTF magnitudes [19].

Filtering of HRTFs can be performed either by linear convolution in the time domain, or by frequency-domain fast convolution techniques. While the latter is significantly more efficient than linear convolution for all but the lowest filter orders, it introduces an additional blocking latency in the order of the HRTF filter length, which can be critical for assisted listening, e.g., hear-through applications. Partitioned convolution techniques [23], [24] enable advantageous tradeoffs between the efficiency of fast convolution and system latency.

HRTF crossfading, which is also denoted as *commutation* [6], refers to the gradual transition between interpolated HRTFs. It reduces audible artifacts that are caused by the exchange of filter coefficients. Thus, crossfading is typically performed at a much higher time resolution than HRTF interpolation. The choice of the crossfading algorithm tightly depends on the convolution method used for HRTF filtering. In case of linear convolution, it can be efficiently implemented by a linear interpolation of the finite impulse reponse (FIR) filter coefficients. In contrast, integrating crossfading with frequency-domain convolution is more difficult due to block-based operation. A typical solution is to perform two convolution processes in parallel and to crossfade the filtered signals in the time domain. A technique that combines crossfading with frequency-domain and partitioned convolution to avoid the complexity of two separate filtering processes is proposed in [24].

AUDIO-AUGMENTED REALITY

Audio-augmented reality refers to a system with which the user hears simultaneously both the synthetic and the ambient sounds around her/him. In addition to the requirements of regular headphone or virtual-reality listening, a hear-through mode is now needed [25], [26].

The hear-through mode is trivial in open and bone-conduction headphones, which do not block the ear canal [27]. Then the user will always hear the ambient sounds without extra attenuation. However, other types of headphones, such as closed and IE headphones, block the ear canal and suppress outside sounds. The hear-through mode must compensate for this attenuation so that the environmental sounds could be heard in a natural way. As seen in Figure 2, in closed-back and IE headphones, the attenuation at low frequencies is not dramatic, but at frequencies higher than 1 kHz it can be remarkable, such as more than 20 dB. This corresponds to a severe acoustic isolation of the headphone user, similar to that observed with hearing protectors.

A hear-through system is usually based on an external microphone [25]. The ambient sound signal captured by the microphone is filtered and sent to the earpiece with an appropriate gain. The aim of the filtering and the amplification is to cancel the attenuation caused by the headphone itself. Thus, the filter is usually of high-pass type, because low frequencies leak to the ear without being much damped.

An additional constraint in a hear-through system is its latency, or the time delay between the leaked and processed sound [25]. It is inevitable that some delay is caused by the analog-to-digital and digital-to-analog conversions and the processing itself, which the microphone signal undergoes. This delay can be, e.g., 1 ms. When the delayed and processed sound are added to the leaked sound at the ear, a comb-filtering effect can color ambient sounds, which is disturbing. The disturbance is strongest when a notch of the comb filter occurs at the frequency range where the leaked and processed sound are equally loud [25]. This corresponds to a 6 dB attenuation in both direct sound and the processed sound. For this reason, slightly surprisingly, a colorless hear-though system is easiest to implement for headphones that attenuate outside sounds well, because then most of the ambient sound can come through microphones and processing.

[AN ADDITIONAL CONSTRAINT IN A HEAR-THROUGH SYSTEM IS ITS LATENCY, OR THE TIME DELAY BETWEEN THE LEAKED AND PROCESSED SOUND.]

ALL-PASS HEAR-THROUGH DESIGN

We describe briefly a method to design a hear-through system based on the all-pass principle [28]. The method takes as its input the impulse response corresponding to the isolation transfer function of the headset. It can be measured using a dummy head with headphones and by playing a sinusoidal sweep signal from a loudspeaker. Additionally, it is necessary to know the latency of the acoustic signal processing system from the microphone input to the earpiece output, which is easy to measure. Furthermore, it is important to account for the magnitude and group-delay of the earpiece response, but here we assume it to be flat and delay-free.

The beginning of the impulse response is given as the input to the all-pass filter design method, which completes it so that the overall system is all-pass [28]. Figure 6(a) shows an example where the given sequence is the beginning of the isolation impulse response, which corresponds to the low-pass filter response in Figure 6(b). When a truncated impulse response of an all-pass filter is combined with it, the overall magnitude response becomes flat, as shown in Figure 6(b). In practice, the headphone itself produces the given sequence while the external microphone signal is filtered with an FIR filter having the allpass tail impulse response.

ASSISTED LISTENING APPLICATIONS

We limit the scope of assisted listening to such applications that help listening in a noisy environment or that employ augmented and modified reality technologies. Nevertheless, by far the most common application of assisted listening technology is the hearing aid. In this context, a hearing aid can be interpreted as a modified reality system that enhances ambient sounds mainly by amplifying them in a desired manner. However, the focus of this article is on other modern assisted listening applications, which are aimed at people with normal or nearly normal hearing.

LISTENING TO MUSIC

In a noisy environment, everyone suffers from hearing problems caused by the auditory masking effect. A typical situation is listening to music or a movie soundtrack in a vehicle, such as in a bus or in an airplane. ANC and NELE methods help to cope with such situations, as discussed previously.

There are additionally other situations where music listening can be enjoyed better by using augmented reality technology. Examples of systems, which are easy to implement, include the silent disco and the silent concert. Both applications require the use of headphones. In a silent disco, music played by the DJ is delivered to the audience via a wireless network or FM radio. Everyone can then decide whether to listen to the music or not and can also adjust the volume to her/his liking. The actual sonic environment in the disco is therefore fairly quiet, with noises mainly coming from conversations and dancing. The silent concert is a similar concept but with the important difference that singers and acoustic musical instruments will be heard also without headphones. However, a proper mix of the music can only be enjoyed through the headphones.

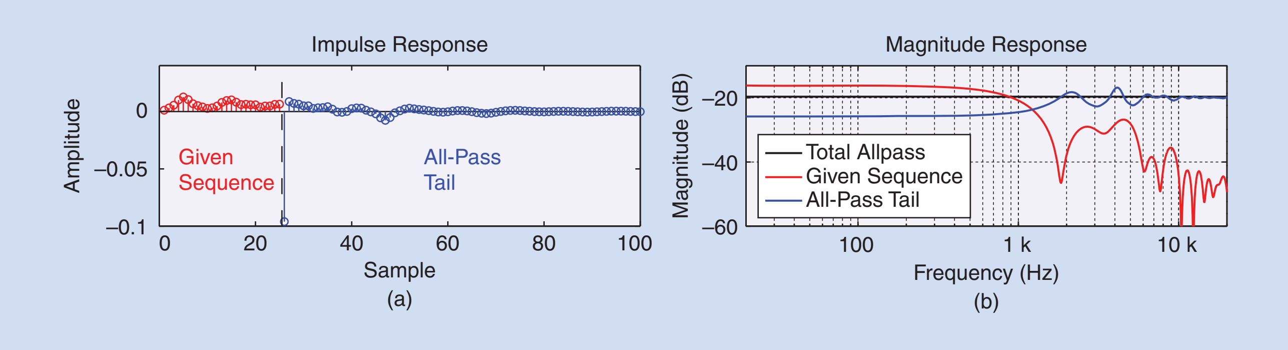
In music festivals, the listening position of most people is far from optimal, and assisted listening technology can enhance the experience. Larsen et al. have built and tested a system in which the music from the stage is transmitted via FM radio to mobile devices of the listeners to be played through headphones [29]. The usability of such a system is critical with respect to the delay such that the actual audio from the public address system is heard simultaneously with the content transmitted via radio. Thus, a localization method and an intelligent delay control, based on the distance from the stage, are required. Overall, the opinions of test users were positive.

The LiveEQ application described by Rämö et al. captures ambient sounds around the user, provides a user-controllable graphic equalization, and plays back the equalized ambient signal to the user with headphones [30]. This modified reality hear-through system can be used for example in a loud concert to limit the noise exposure caused by the live music. If the headset attenuates the ambient sound well, the user can mix her/his own version of the music by boosting selected frequency ranges with the real-time equalizer. Still, the sound pressure level of the mixed music can be lower than that of the original live music.

AUDIO-AUGMENTED REALITY

In audio-augmented reality, the real soundscape and virtual auditory events mix seamlessly together [31]. In such applications sound is typically used to deliver information that assists the user in performing certain tasks, or to enhance the perception of the environment. A major advantage of auditory over graphical display is that the user can perceive acoustic information from any direction without being required to turn toward the acoustic source. Therefore, an important application area of audio-augmented reality are scenarios where the user cannot or should not look at a display to obtain information, due to the user's vision being either impaired or occupied with a primary task, e.g, while walking or driving. Similarly, audio-augmented reality may convey information about the immediate surroundings that lie outside the user's field of view, for instance, the approaching of a quiet electric car.

Most often, the application scenarios are mobile such that the user can freely move around and the augmented audio content is determined based on the user's location. One typical application is navigation in which the user always gets accurate spatialized instructions on how to proceed to achieve the target location. This same concept is valid in a wide range of use scenarios covering, e.g., walking in a city, driving a car, or taxiing an airplane at an airport. However, the group of people that benefits most from audio in navigation are the visually impaired [32], especially in



[FIG6] An example all-pass filter design for acoustic transparency [28]: (a) The engineered impulse response contains the leaked sound in the beginning and a designed all-pass tail, and (b) their combination has a flat magnitude response.

short-distance navigation and object avoidance in which spatialized sound can play a crucial role to help people advance safely.

A typical use case of audio-augmented reality is a museum guide [33], [34]. With the augmented reality techniques it is possible to have each piece at an exhibition to act as a virtual sound source such that a visitor can hear the attraction introduce itself. Another interesting application domain for augmented reality is gaming. An example of an audio-only augmented reality game is *Guided by Voices*, where an overlay of virtual objects and game characters onto the real world is created entirely using sound [35].

USABILITY ISSUES

Although the presented techniques have many attractive applications, they are not completely free of usability problems. The observed sound quality and naturalness are on a very high level, but user comfort should be improved [36]. Tikander has shown that the user's own sounds are challenging, too, since current audio-augmented reality systems aim at natural hear-through of ambient sounds [36]. For example, when the user reads aloud or eats crispbread, the technology alters the experience such that it may cause annoyance. This is caused by the blocking of the ear canals—the occlusion effect—and the inability of the techniques to alter bone-conducted sounds, such as the user's own voice.

Another related question is social acceptability. When one is wearing headphones, others often assume that she/he is not listening, although with augmented reality headphones the case might be the opposite, and one's listening can actually be more intense than without the headphones.

CONCLUSIONS

This article has reviewed signal processing methods and applications related to assisted listening. Headphones are commonly used with a mobile phone or another portable device, and the ambient noise disturbs listening by masking some of the audio content. Active noise control helps to improve the attenuation of noises while NELE methods improve the audibility and intelligibility by modifying the audio signal itself. An unmasking method developed for headphone listening was described, which estimates the levels of music and noise in the ear by accounting for the attenuation characteristics of the headphone. It then computes a masking threshold from the noise signal and compares the spectrum of the music signal against it. An adaptive equalizing filter is then adjusted to boost those frequencies in the music, which would otherwise be completely or partially masked by the noise.

Virtual reality audio is an extension of regular headphone listening in which the user can hear transmitted or recorded sounds seemingly from his environment. This is achieved by using head-tracking and binaural synthesis techniques, which help to keep the virtual sources at their prescribed locations even when the user is moving. Interpolation techniques and a complete signal processing system to implement time-varying binaural synthesis were discussed.

Audio-augmented reality mixes real and reproduced sounds, requiring external microphones and a hear-through function to cancel the attenuation caused by the headset. A new method for designing a filter to achieve colorless, or allpass-type, hear-through system was described.

While there are many audio applications for virtual and augmented reality, such as navigation tools and museum guides, several relevant applications belong to the category of modified reality. These systems reproduce through the headset a processed version of the ambient sound field, such as in enhanced concert applications or in the LiveEQ system, which provides a real-time equalizer to concert audiences. The increase in computational power of mobile devices will enable even more advanced new devices and applications, such as adaptive intelligent headphones that will observe the environment continuously and modify the audio mix and content to deliver the most relevant information to the user according to her/his personal preferences.

[THE INCREASE IN COMPUTATIONAL POWER OF MOBILE DEVICES WILL ENABLE EVEN MORE ADVANCED NEW DEVICES AND APPLICATIONS.]

AUTHORS

Vesa Välimäki (vesa.valimaki@aalto.fi) received the M.Sc. and the doctor of science in technology degrees from the Helsinki University of Technology, Espoo, Finland, in 1992 and 1995, respectively. He is a professor of audio signal processing in the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. In 2008–2009, he was a visiting scholar at Stanford University. His research interests include headphone audio, digital filters, audio effects processing, sound synthesis, and acoustics of musical instruments. He is an associate member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He is a Fellow of the IEEE and of the Audio Engineering Society.

Andreas Franck (A.Franck@soton.ac.uk) received the diploma degree in computer science and the Ph.D. degree in electrical engineering, both from the Ilmenau University of Technology, Germany. Since 2004, he has been with the Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau, Germany. In 2014, he joined the Institute of Sound and Vibration Research, University of Southampton, United Kingdom, as a postdoctoral research fellow. The work for this article was performed while he was with Fraunhofer IDMT. His research interests include spatial and object-based audio, in particular efficient sound reproduction algorithms, variable digital filters, and fast convolution techniques.

Jussi Rämö (jussi.ramo@aalto.fi) received the M.Sc. degree in communication engineering in 2009 from the Helsinki University of Technology, Finland, and the doctor of science in technology degree from Aalto University, Finland, in 2014. His major in both degrees was acoustics and audio signal processing. Since 2009, he has worked as a researcher in the

Department of Signal Processing and Acoustics at Aalto University, Espoo, Finland. His research interests include sound reproduction, headphone audio, and digital filtering. He was a member of the organizing committee of the 2013 Audio Engineering Society 51st International Conference on Loudspeakers and Headphones, Helsinki, Finland.

Hannes Gamper (hannes.gamper@aalto.fi) received his Ph.D. degree in media technology from Aalto University, Espoo, Finland, in 2014. His doctoral research focused on enabling technologies for audio-augmented reality. In 2012, he was a visiting scholar at the Human Interface Technology Laboratory (HIT Lab) in Christchurch, New Zealand. He currently works as a postdoctoral researcher at Microsoft Research in Redmond, Washington, United States, but the work reported here was conducted outside of Microsoft Research. His research interests include binaural modeling, and the analysis, synthesis, and perception of spatial sound.

Lauri Savioja (lauri.savioja@aalto.fi) received the M.Sc. (Tech.) and D.Sc. (Tech.) degrees in computer science from the Helsinki University of Technology (TKK), Espoo, Finland, in 1991 and 1999, respectively. The topic of his doctoral thesis was room acoustic modeling. He worked at the TKK Laboratory of Telecommunications Sxoftware and Multimedia as a researcher, lecturer, and professor from 1995 until the formation of the Aalto University, where he is currently a professor and heads the Department of Media Technology in the School of Science. His research interests include room acoustics, virtual reality, and parallel computing.

REFERENCES

[1] J. White, D. C. Schmidt, and M. Golparvar-Fard, “Applications of augmented reality,” *Proc. IEEE*, vol. 102, no. 2, pp. 120–123, Feb. 2014.

[2] B. Rafaely, “Active noise reducing headset—an overview,” in *Proc. Int. Congress Exhibition Noise Control Engineering (Internoise)*, The Hague, The Netherlands, Aug. 2001.

[3] C. H. Taal, J. Jensen, and A. Leijon, “On optimal linear filtering of speech for near-end listening enhancement,” *IEEE Signal Processing Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[4] J. Rämö, V. Välimäki, and M. Tikander, “Perceptual headphone equalization for mitigation of ambient noise,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP-13)*, Vancouver, Canada, May 2013, pp. 724–728.

[5] M. Christoph, “Noise dependent equalization control,” in *Proc. Audio Engineering Society 48th Int. Conf. Automotive Audio*, Munich, Germany, Sept. 2012.

[6] J.-M. Jot, V. Larcher, and O. Warusfel, “Digital signal processing issues in the context of binaural and transaural stereophony,” in *Proc. Audio Engineering Society 98th Conv.*, Paris, France, Feb. 1995.

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

[8] V. R. Algazi and R. O. Duda, “Headphone-based spatial sound,” *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 33–42, Jan. 2011.

[9] F. L. Wightman and D. J. Kistler, “Headphone simulation of free-field listening. I: Stimulus synthesis,” *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 858–867, 1989.

[10] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[11] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, “Fifty years of artificial reverberation,” *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 20, no. 5, pp. 1421–1448, July 2012.

[12] I. E. Sutherland, “A head-mounted three dimensional display,” in *Proc. Fall Joint Computer Conf.*, New York, NY, 1968, pp. 757–764.

[13] J. Blauert, G. Boerger, and P. Laws, “Method and equipment to avoid the localization shifts of the auditory events caused by head movements with earphones on,” DE 2331619.0, U.S. patent 3,962,543, Eugen Beyer Elektrotechnische Fabrik, Heilbronn, Germany, 1973.

[14] K. Sunder, J. He, E.-L. Tan, and W.-S. Gan, “Natural sound rendering for headphones,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 98–111, Mar. 2015.

[15] W. Zhang, M. Zhang, R. Kennedy, and T. Abhayapala, “On high-resolution head-related transfer function measurements: An efficient sampling scheme,” *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 20, no. 2, pp. 575–584, Feb. 2012.

[16] J. Huopaniemi and J. O. Smith III, “Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters,” in *Proc. Audio Engineering Society 16th Int. Conf.*, Rovaniemi, Finland, Mar. 1999, pp. 301–312.

[17] D. S. Brungart, “Near-field virtual audio displays,” *Presence: Teleoper. Virtual Environ.*, vol. 11, no. 1, pp. 93–106, 2002.

[18] H. Gamper, “Head-related transfer function interpolation in azimuth, elevation, and distance,” *J. Acoust. Soc. Amer.*, vol. 134, pp. EL547–554, Dec. 2013.

[19] D. N. Zotkin, R. Duraiswami, and L. S. Davis, “Rendering localized spatial audio in a virtual auditory space,” *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.

[20] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, “Splitting the unit delay—Tools for fractional delay filter design,” *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 30–60, Jan. 1996.

[21] J. Nam, M. Kolar, and J. S. Abel, “On the minimum-phase nature of head-related transfer functions,” in *Proc. Audio Engineering Society 125th Conv.*, San Francisco, CA, 2008.

[22] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, “Sensitivity of human subjects to head-related transfer-function phase spectra,” *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2821–2840, May 1999.

[23] F. Wefers and M. Vorländer, “Optimal filter partitions for real-time FIR filtering using uniformly partitioned FFT-based convolution in the frequency-domain,” in *Proc. 14th Int. Conf. Digital Audio Effects*, Paris, France, Sept. 2011, pp. 155–161.

[24] A. Franck, “Efficient frequency-domain filter crossfading for fast convolution with application to binaural synthesis,” in *Proc. Audio Engineering Society 55th Int. Conf.*, Helsinki, Finland, Aug. 2014.

[25] J. Rämö and V. Välimäki, “Digital augmented reality audio headset,” *J. Electr. Comput. Eng.*, vol. 2012, Oct. 2012.

[26] P. F. Hoffmann, F. Christensen, and D. Hammershøi, “Insert earphone calibration for hear-through options,” in *Proc. Audio Engineering Soc. 51st Int. Conf. Loudspeakers and Headphones*, Helsinki, Finland, Aug. 2013.

[27] R. W. Lindeman, H. Noma, and P. G. de Barros, “Hear-through and mic-through augmented reality: using bone conduction to display spatialized audio,” in *Proc. IEEE and ACM Int. Symp. Mixed and Augmented Reality*, Nara, Japan, Nov. 2007, pp. 173–176.

[28] J. Rämö and V. Välimäki, “An allpass hear-through headset,” in *Proc. EUSIPCO*, Lisbon, Portugal, Sept. 2014.

[29] J. E. Larsen, A. Stopczynski, J. Larsen, C. Vesterskov, P. Krogsgaard, and T. Sondrup, “Augmenting the sound experience at music festivals using mobile phones,” in *Proc. 16th Int. Conf. Intelligent User Interfaces* (IUI-11), New York, NY, 2011, pp. 383–386.

[30] J. Rämö, V. Välimäki, and M. Tikander, “Live sound equalization and attenuation with a headset,” in *Proc. Audio Engineering Society 51st Int. Conf. Loudspeakers and Headphones*, Helsinki, Finland, Aug. 2013.

[31] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, “Augmented reality audio for mobile and wearable appliances,” *J. Audio Eng. Soc.*, vol. 52, pp. 618–639, June 2004.

[32] B. F. G. Katz, S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, M. Auvray, P. Truillet, M. Denis, S. Thorpe, and C. Jouffrais, “NAVIG: Augmented reality guidance system for the visually impaired,” *Virtual Reality*, vol. 16, no. 4, pp. 253–269, 2012.

[33] A. Zimmermann and A. Lorenz, “LISTEN: a user-adaptive audio-augmented museum guide,” *User Model. User-Adapt. Interact.*, vol. 18, no. 5, pp. 389–416, 2008.

[34] A. Damala, T. Schuchert, I. Rodriguez, J. Moragues, K. Gilleade, and N. Stojanovic, “Exploring the affective museum visiting experience: adaptive augmented reality (A2R) and cultural heritage,” *Int. J. Heritage Digital Era*, vol. 2, no. 1, pp. 117–142, 2013.

[35] K. Lyons, M. Gandy, and T. Starner, “Guided by voices: An audio augmented reality system,” in *Proc. Int. Conf. Auditory Display* (ICAD), Atlanta, GA, Apr. 2000, pp. 57–62.

[36] M. Tikander, “Usability issues in listening to natural sounds with an augmented reality audio headset,” *J. Audio Eng. Soc.*, vol. 57, no. 6, pp. 430–441, June 2009.

[SP]

Kaushik Sunder, Jianjun He, Ee-Leng Tan, and Woon-Seng Gan

Natural Sound Rendering for Headphones

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

Integration of signal processing techniques

With the strong growth of assistive and personal listening devices, natural sound rendering over headphones is becoming a necessity for prolonged listening in multimedia and virtual reality applications. The aim of natural sound rendering is to naturally recreate the sound scenes with the spatial and timbral quality as natural as possible, so as to achieve a truly immersive listening experience. However, rendering natural sound over headphones encounters many challenges. This tutorial article presents signal processing techniques to tackle these challenges to assist human listening.

INTRODUCTION

Sound is an inherent part of our everyday lives for information, communication, and interaction. Sound improves situational awareness by providing feedback for actions and situations that are out of the view of the listener. An advantage of sound is that multiple sound sources can be perceived from any location around the head in the three-dimensional (3-D) space [1]. The role of natural 3-D sound, or spatial sound, in high-stress applications, like flight navigation and communication systems, is indisputable [1]. Naturally rendered sound has also been proven to be beneficial in personal route guidance for visually impaired people and in medical therapy for patients [1]. Last but not least, the ever-growing market of consumer electronics calls for natural sound rendering for digital media, such as movies, games, and augmented, virtual reality applications like teleconferencing.

In most of these applications, listening is seldom from the physical sound sources but is instead from playback devices, such as headphones or loudspeakers. Headphones, by virtue of their convenience and portability, are typically chosen as the preferred playback device, especially for personal listening. Therefore, to assist headphone listening, it is critical for the sound to be rendered in a way that listeners can perceive it as natural as possible. In this context, natural sound rendering essentially refers to rendering of the original sound scene using headphones to create an immersive listening experience and the sensation of "being there" at the venue of the acoustic event. To achieve natural sound rendering, the virtual sound rendered should exactly emulate all the spatial cues of the

Digital Object Identifier 10.1109/MSP.2014.2372062
Date of publication: 12 February 2015

1053-5888/15©2015IEEE

original sound scene, as well as the individual spectral characteristics of the listener's ears. In this article, we mainly consider the most widely used channel-based audio as the input signals for the natural sound rendering system, though some of the signal processing techniques discussed could also be used in other audio formats, such as object-based format and ambisonics [2], [3].

In recent years, the design criteria for commercial headphones have undergone significant development. At Harman International Industries, Olive et al. investigated the best target responses for designing headphones based on the listener's preference for the most natural sound [4]. Creating realistic surround sound in headphones has become a common pursuit of many headphone technologies such as Dolby, DTS, etc. Furthermore, a personalized listening experience and incorporation of the information of listening environment have also been trends in the headphone industry. These trends in headphones share one common objective—to render natural sound in headphones.

CHALLENGES

The listening process in humans can generally be considered as a source-medium-receiver model, as stated by Begault [1]. This model is used in this article to highlight the differences between natural listening in a real environment and listening via headphones. In natural listening, we listen to the physical sound sources in a particular acoustic space, with the sound waves undergoing diffraction, interference with different parts of our morphology (torso, head, and pinna) before reaching the eardrum. This information of sound wave propagation can be encapsulated in spatial digital filters termed *head-related transfer functions* (*HRTFs*) [1]. Listeners also get valuable interaural cues for sound localization with head movements. However, headphone listening is inherently different from natural listening as the sources we are listening to are no longer physical sound sources but are recorded and edited sound materials. These differences between natural and headphone listening lead to various challenges in rendering natural sound over headphones, which can be broadly classified into categories from the perspectives of source, medium, and receiver, as described next.

SOURCE

The sound scenes rendered for headphone listening should comprise not only the individual sound sources but also the features of the sound environment. Listeners usually perceive these sound sources to be directional, i.e., coming from certain directions. Moreover, in most of the digital media content, the sound environment is usually perceived by the listener to be diffuse (partially). This perceptual difference between the sound sources and the sound environment requires them to be considered separately in natural sound rendering [2]. Though there are other formats that can represent the sound scenes (e.g., object based, ambisonics), the convention for today's digital media is still primarily a channel-based format. Hence, the focus of this article lies in the rendering of channel-based audio, where sound source and environment signals are mixed in each channel [2]. In channel-based signals, where only the sound mixtures are available (assuming one mixture in every channel), it is necessary to extract the source signals and environment signals, which can be quite challenging. Furthermore, most of the traditional recordings are processed and mixed for optimal playback over loudspeakers rather than headphones. Direct playback of such recordings over headphones results in an unnatural listening experience, which is mainly due to the loss of crosstalk, and localization issues.

> TO ACHIEVE NATURAL SOUND RENDERING, THE VIRTUAL SOUND RENDERED SHOULD EXACTLY EMULATE ALL THE SPATIAL CUES OF THE ORIGINAL SOUND SCENE, AS WELL AS THE INDIVIDUAL SPECTRAL CHARACTERISTICS OF THE LISTENER'S EARS.

MEDIUM

Headphone listening does not satisfy free-air listening conditions as in natural listening. Since the headphone transfer function (HPTF) is not flat, equalization of the headphone is necessary. However, this equalization is tedious and challenging as the headphone response is highly dependent on the individual anthropometrical features and also varies with repositioning.

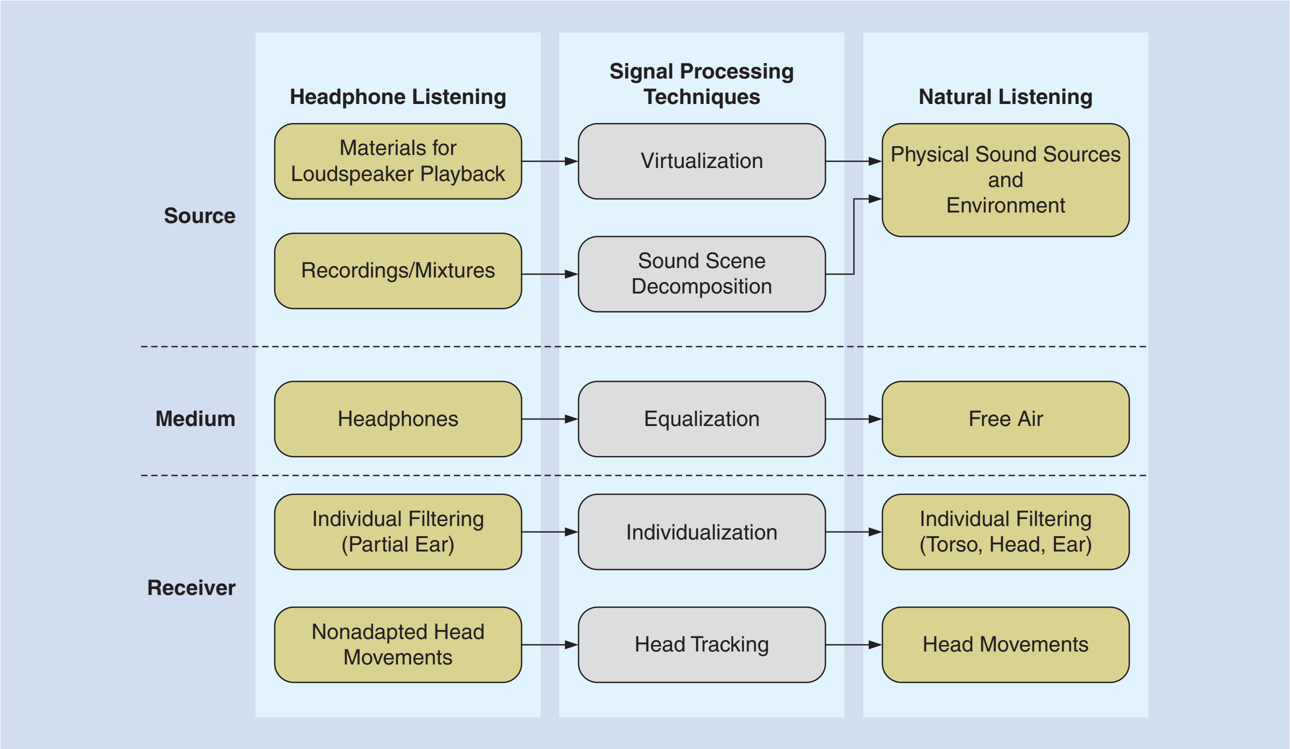
RECEIVER

The omission of listener's individualized filtering with the outer ear in headphone listening often leads to coloration and localization inaccuracies. These individualized characteristics of the listener are lost when the sound content is recorded or synthesized nonindividually, i.e., the subject in the listening is different from the subject in the recording or synthesis. Furthermore, the sound in headphone listening is not adapted to the listener's head movements, which departs from a natural listening experience.

SIGNAL PROCESSING TECHNIQUES

To tackle the aforementioned challenges and enhance natural sound rendering over headphones, digital signal processing techniques are commonly used. In Figure 1, we summarize the differences between natural listening and headphone listening and introduce the following corresponding signal processing techniques to tackle these challenges:

- *Virtualization*: to match the desired playback for the digital media content
- *Sound scene decomposition using blind source separation* (*BSS*) *and primary-ambient extraction* (*PAE*): to optimally facilitate the separate rendering of sound sources and sound environment
- *Individualization of HRTF*: to compensate for the lost or altered individual filtering of the sound in headphone listening
- *Equalization*: to preserve the original timbral quality of the source and alleviate the adverse effect of the inherent headphone response


Headphone Listening
Signal Processing Techniques
Natural Listening
Source
Materials for Loudspeaker Playback
Virtualization
Physical Sound Sources and Environment
Recordings/Mixtures
Sound Scene Decomposition
Medium
Headphones
Equalization
Free Air
Receiver
Individual Filtering (Partial Ear)
Individualization
Individual Filtering (Torso, Head, Ear)
Nonadapted Head Movements
Head Tracking
Head Movements

[FIG1] A summary of the differences between natural listening and headphone listening and the corresponding signal processing techniques to solve these challenges for natural sound rendering. The main challenges and their corresponding signal processing techniques in each category (source, medium, and receiver) are highlighted and their interactions (not shown here) are further discussed in the article.

- *Head tracking*: to adapt to the dynamic head movements of the listener.

The following sections describe in detail the virtualization and its interaction with head tracking, sound scene decomposition, individualization, and equalization. These signal processing techniques are integrated and evaluated using subjective tests.

VIRTUALIZATION

In digital media, sound is typically mixed for loudspeaker playback rather than headphone playback. The spatial sound to be rendered naturally over headphones should emulate the natural propagation of the acoustic waves emanating from the loudspeaker to the eardrum of the listener. To emulate stereo or surround sound loudspeaker rendering over headphones, virtualization techniques based on HRTFs corresponding to the loudspeaker positions are commonly used. Given these acoustic transfer functions (i.e., HRTFs), the virtualization technique is applicable to any multichannel loudspeaker setup, be it stereo, 5.1, 7.1, 22.2, or even loudspeaker arrays in wavefield synthesis. As shown in Figure 2(a), for every desired loudspeaker position, the signal in the mth channel $x_m(n)$ is filtered with the corresponding HRTF $h_{xmL}(n)$, $h_{xmR}(n)$, and summed before being routed to the left and right ears [1], [5], respectively, as:

$$\begin{aligned} y_L(n) &= \sum_{m=1}^{M} h_{xmL}(n) * x_m(n), \\ y_R(n) &= \sum_{m=1}^{M} h_{xmR}(n) * x_m(n), \end{aligned} \tag{1}$$

where * denotes convolution and M is the total number of channels. When the HRTFs are directly applied to multichannel loudspeaker signals, the rendered sound scenes in headphone playback suffer from inaccurate virtual source directions, lack of depth, and reduced image width [5], [6].

To solve these problems in virtualization of multichannel loudspeaker signals and achieve a faithful reproduction of the sound scenes, the HRTFs should be applied to the individual source signals that are usually extracted (using BSS and PAE) from the loudspeaker signals (i.e., mixtures). In this virtualization [as shown in Figure 2(b)], the sources are rendered directly using the HRTFs of the corresponding source directions $h_{skL}(n)$, $h_{skR}(n)$

$$\begin{aligned} y_L(n) &= \sum_{k=1}^{K} h_{skL}(n) * s_k(n) + a_L(n), \\ y_R(n) &= \sum_{k=1}^{K} h_{skR}(n) * s_k(n) + a_R(n), \end{aligned} \tag{2}$$

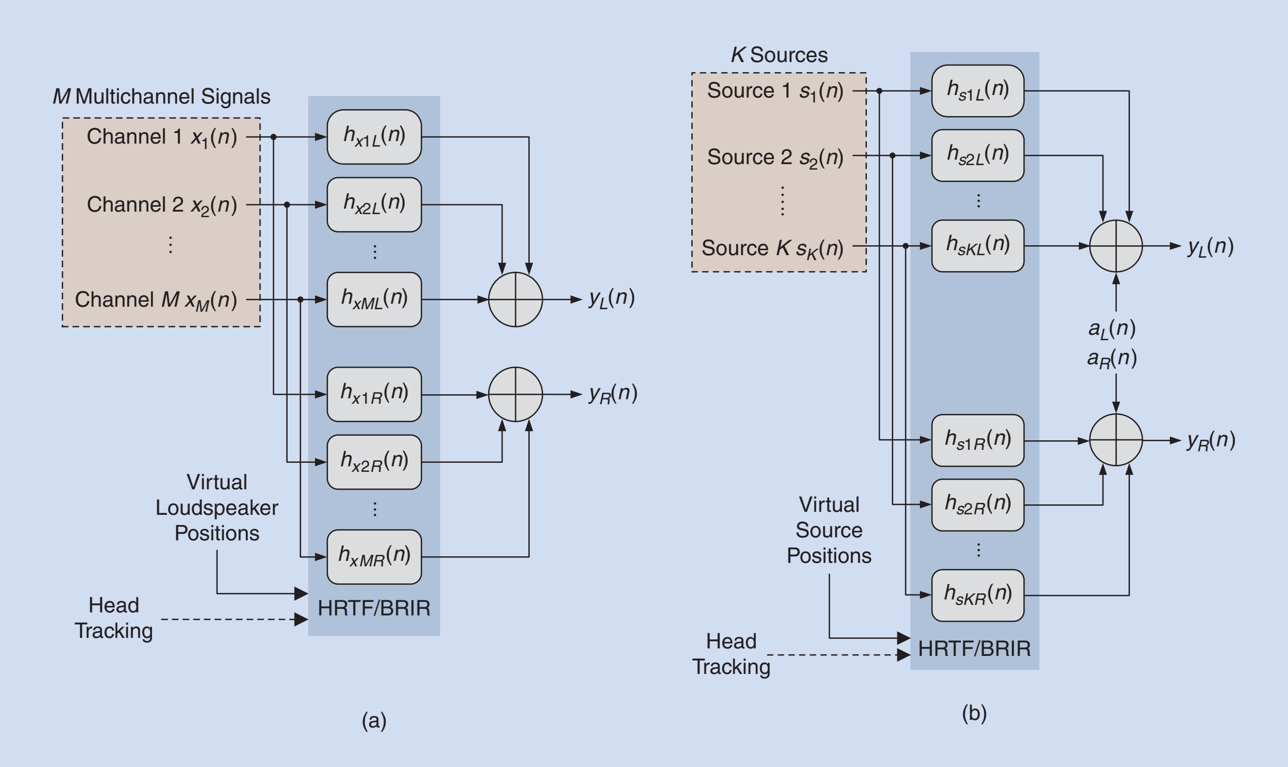
where K is the total number of sources, $s_k(n)$ is the kth source in the multichannel signal, and the environment signals $a_L(n)$, $a_R(n)$ are the rendered signals representing the sound environment perceived by two ears. To render the acoustics of the environment, the environment signals can be either synthesized according to the sound environment [7] or extracted from the mixtures. Techniques like decorrelation [5], [8] and artificial reverberation [9] are commonly employed to render the environment signals to create a more diffuse and natural sound environment.

Furthermore, adding the reverberation of sources (or the loudspeaker signals in virtualization of multichannel loudspeaker signals) can also improve the realism of the reproduced sound scene [10]. Therefore, in virtualization, it is quite common to use BRIRs [1], [5] that encapsulate HRTFs and reverberation. Accordingly, selecting the correct amount of early reflections as well as late reverberation is critical to recreate a faithful sound environment [1]. In general, the BRIR that matches the sound environment of the scene or BRIR of a mixing studio are considered to be more suitable [4]. As discussed in the section "Challenges," natural sound rendering requires the accurate reproduction of both the sound sources and the sound environment. Compared to the virtualization of multichannel loudspeaker signals [Figure 2(a)], the latter technique of virtualizing the source and environment signals [Figure 2(b)] is more desirable as it is closer to natural listening [6], [8], [9]. These virtualization techniques can also be incorporated into spatial audio coding systems, such as binaural cue coding [11], spatial audio scene coding [5], and directional audio coding [3].

In virtualization, the directions of the sources [or the loudspeakers in virtualization of multichannel loudspeaker signals as in Figure 2(a)] have to be calibrated according to the head movements (as in natural listening). To fulfill this need, the HRTFs/BRIRs in the virtualization are updated on the fly based on these head movements, which are often tracked by a sensor (e.g., accelerometer, gyroscope, camera, etc.). The latency between the head tracking and sound rendering should be such that the localization accuracy is not affected [12]. When incorporated in the virtualization process, such a head-tracking system can provide useful dynamic cues to resolve the localization conflicts [1] and enhance natural sound rendering [10], [12]. It shall be noted that head tracking is more critical for the directional sources but less important for the diffuse signals like environment signals and late reverberation [12]. This is because the perception of diffuse signals is less affected by head movements.

A PERSONALIZED LISTENING EXPERIENCE AND INCORPORATING THE INFORMATION OF LISTENING ENVIRONMENT HAVE ALSO BEEN TRENDS IN THE HEADPHONE INDUSTRY. THESE TRENDS SHARE ONE COMMON OBJECTIVE—TO RENDER NATURAL SOUND IN HEADPHONES.

Recreating the perception of distance of the sources close to natural listening is another critical aspect in virtualization for natural sound rendering. However, the challenges in simulating accurate distance perception are numerous. Human beings' ability to accurately estimate these distances has long been known to be poorer compared to our ability to estimate directions, even in the physical listening space [1]. Virtual listening through headphones



[FIG2] Virtualization of (a) multichannel loudspeaker signals $x_m(n)$ [5], and (b) multiple sources $s_k(n)$ and environment signals $a_L(n), a_R(n)$. $y_L(n), y_R(n)$ is the signal sent to the left and right ear, respectively. Note that head tracking can be used to update the selected directions of HRTFs/binaural room impulse responses (BRIRs).

[TABLE 1] AN OVERVIEW OF TYPICAL TECHNIQUES IN BSS.

OBJECTIVE: TO EXTRACT $K(K>2)$ SOURCES FROM M MIXTURES

CASE		TYPICAL TECHNIQUES
DETERMINED: $K=M$		ICA [14]
OVERDETERMINED: $K<M$		ICA WITH PCA OR LS [14]
UNDERDETERMINED: $K>M$	$M>2$	ICA WITH SPARSE SOLUTIONS [14], [15]
	$M=2$	TIME-FREQUENCY MASKING [16]
	$M=1$	NMF [17], [18]; CASA [19]

further hinders the distance perception as it leads to inside-the-head localization (IHL) of sound [1]. IHL of sound is caused by several factors, such as the use of nonindividualized HRTFs, absence of equalization, lack of reverberation, and impedance mismatch due to the presence of headphones [1], [13]. The presence of individualized HRTFs, equalization, and reverberation can improve the externalization of sound but does not ensure accurate distance perception [1].The direct-to-reverberation energy ratio is found to be the most critical cue for absolute distance perception, even though the intensity, loudness, and binaural cues can provide relative cues for distance perception [1]. Since reverberation is an essential cue for both distance perception and perception of a real environment context, a veridical simulation of the reverberation is highly imperative for natural sound rendering [1]. However, accurate simulation of distance perception is challenging since reverberation entirely depends on the room characteristics. The correct amount of reverberation to be added to simulate distance perception in a particular room can be obtained only by carrying out acoustical measurements.

SOUND SCENE DECOMPOSITION USING BSS AND PAE

To achieve natural sound rendering in headphones, two important constituents of the sound scenes are required in the virtualization: the individual sound sources and characteristics of the sound environment. However, this information is usually not directly available to the end user. One has to work with the existing digital media content that is available, i.e., the mastered mix distributed in channel-based formats (e.g., stereo, 5.1 surround sound). Therefore, to facilitate natural sound rendering, it is necessary to extract the sound sources and/or sound environment from their mixtures. In this section, we discuss two types of techniques applied in sound scene decomposition: BSS and PAE.

DECOMPOSITION USING BSS

Extracting the sound sources from the mixtures, often referred to as *BSS*, has been extensively studied in the last few decades. The basic mixing model in BSS can be considered as anechoic mixing, where the sources $s_k(n)$ in each mixture $x_m(n)$ have different gains g_{mk} and delays τ_{mk}. Hence, the anechoic mixing is formulated as follows:

$$x_m(n)=\sum_{k=1}^{K} g_{mk}s_k(n-\tau_{mk})+e_m(n), \quad \forall m\in\{1,2,\ldots,M\}, \tag{3}$$

where $e_m(n)$ is the noise in each mixture, which is usually neglected for most cases. Note that estimating the number of sources is quite challenging and it is usually assumed to be known in advance [14]. This formulation can be simplified to represent instantaneous mixing by ignoring the delays, or can be extended to reverberant mixing by including multiple paths between each source and mixture. An overview of the typical techniques applied in BSS is listed in Table 1.

Based on the statistical independence and non-Gaussianity of the sources, independent component analysis (ICA) algorithms have been the most widely used techniques in BSS to separate the sources from mixtures in the determined case, where the numbers of mixtures and sources are equal [14]. In the overdetermined case, where there are more mixtures than sources, ICA is combined with principal component analysis (PCA) to reduce the dimension of the mixtures, or combined with least-squares (LS) to minimize the overall mean-square error (MSE) [14]. In practice, the underdetermined case is the most common, where there are fewer mixtures than sources. For the underdetermined BSS, sparse representations of the sources are usually employed to increase the likelihood of sources to be disjoint [15]. The most challenging underdetermined BSS is when the number of mixtures is two or lesser, i.e., in stereo and mono signals.

Stereo signals (i.e., $M=2$), being one of the most widely used audio format, have been the focus in BSS. Many of these BSS techniques can be considered as time-frequency masking and usually assume one dominant source in one time-frequency bin of the stereo signal [16]. In these time-frequency masking-based approaches, a histogram for all possible directions of the sources is constructed, based on the range of the bin-wise amplitude and phase differences between the two channels. The directions, which appear as peaks in the histogram, are selected as source directions. These selected source directions are then used to classify the time-frequency bins and to construct the mask. For every time-frequency bin (n,l), the kth source at mth channel $\hat{S}_{mk}(n,l)$ is estimated as:

$$\hat{S}_{mk}(n,l)=\Psi_{mk}(n,l)X_m(n,l), \tag{4}$$

where the mask and the mth mixture are represented by $\Psi_{mk}(n,l)$ and $X_m(n,l)$, respectively.

In the case of single-channel (or mono) signals, the separation is even more challenging since there is no interchannel information. Hence, there is a need to look into the inherent physical or perceptual properties of the sound sources. Nonnegative matrix factorization (NMF)-based approaches have been extensively studied and applied in single-channel BSS in recent years. The key idea of NMF is to formulate an atom-based representation of the sound scene [17], where the atoms have repetitive and nondestructive spectral structures. NMF usually expresses the magnitude (or power) spectrogram of the mixture as a product of the atoms and time varying nonnegative weights in an unsupervised manner. These atoms, after being multiplied with their corresponding weights, can be considered as potential components of sources [18]. Another technique applied in single-channel BSS is the computational auditory scene analysis (CASA) that simulates the segregation and grouping mechanism of the human auditory system [19] on the model-based representation (monaural case) of the auditory scenes. An important aspect worth considering is the directions of the extracted sources, which can usually come as a by-product in multichannel BSS. In single-channel BSS, this information of source directions has to be provided separately.

DECOMPOSITION USING PAE

In most sound scenes, the mixture comprises not only the dry sources but also the reverberation and ambient sound, which are contributed by the acoustics of the surrounding space. Therefore, the mixing model of the sources in BSS usually does not match with the actual sound scenes. In this article, we refer to the dominant sources as *primary* (or direct) components, while the signals contributed by the sound environment are referred to as *ambient* (or diffuse) components. The primary and ambient components are perceived to be directional and diffuse, respectively. Different rendering methods should be applied to the primary and ambient components [6], [7] due to their perceptual differences. Therefore, rendering of natural sound scenes requires the decomposition of the mixtures into primary and ambient components [6], [7], [9]. Since stereo is still the most widely used format for digital media content, our discussion on the decomposition using PAE is focused on stereo signals ($M = 2$).

In PAE, we often follow some intuitive signal models as discussed in [3], [5], [7], [8], and [20]. In the mth channel, the mixture $x_m(n)$ is assumed to be the sum of the primary component $p_m(n)$ and ambient component $a_m(n)$, i.e., $x_m(n) = p_m(n) + a_m(n)$. The discrimination of directional primary components and diffuse ambient components is mainly based on their interchannel correlations, where the primary and ambient components in the two channels are assumed to be correlated and uncorrelated, respectively. In the basic mixing model for PAE, the primary components are assumed to be amplitude panned, while the ambient components are of approximately equal levels in all channels.

Based on these assumptions, various approaches are proposed in PAE for stereo signals. Similar to BSS, time-frequency masking approaches are introduced to extract ambient components $\hat{A}_m(n,l)$ [7], [20] and these approaches can be generalized as

$$\hat{A}_m(n,l) = X_m(n,l)\Psi_A(n,l), \tag{5}$$

where $0 \leq \Psi_A(n,l) \leq 1$ is the real-valued ambient mask at the time-frequency bin (n,l). Time-frequency bins having high interchannel correlation are considered to be primary components (or mostly primary components in the soft masking case), whereas low correlation bins are more likely to be ambient components.

Several linear estimation-based PAE approaches were also introduced [21], which exploits the differences between the two channels of the stereo signal to perform the PAE, including PCA-based approaches [20] and LS-based approaches. In these approaches, the extracted primary components $\hat{p}_0(n), \hat{p}_1(n)$ and ambient components $\hat{a}_0(n), \hat{a}_1(n)$ are expressed as weighted sums of the mixtures:

$$\begin{bmatrix} \hat{p}_0(n) \\ \hat{p}_1(n) \\ \hat{a}_0(n) \\ \hat{a}_1(n) \end{bmatrix} = \begin{bmatrix} w_{\mathrm{P}0,0} & w_{\mathrm{P}0,1} \\ w_{\mathrm{P}1,0} & w_{\mathrm{P}1,1} \\ w_{\mathrm{A}0,0} & w_{\mathrm{A}0,1} \\ w_{\mathrm{A}1,0} & w_{\mathrm{A}1,1} \end{bmatrix} \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}. \tag{6}$$

The solutions for the weights in (6) are derived based on different performance-related criteria [21]. More specifically, PCA extracts the primary components having maximum variance and extracts the ambient components having minimum variance with the constraint that the primary and ambient components are uncorrelated, while LS extracts these components having minimum MSE. Based on the study in [21], it is recommended that PCA-based approaches should be used for signals that contain dominant primary components (e.g., gaming), while LS-based approaches are preferred for signals that contain a balanced mix of primary and ambient components (e.g., movies). In addition, to deal with more complex types of input signals that do not fit into the basic mixing model, other techniques have also been introduced, such as time shifting to compensate for time differences [22] and adaptive frequency bin partitioning for multiple sources in primary components [23]. Furthermore, though it is possible to extend the framework of PAE from stereo signals to multichannel signals, e.g., [24], more comprehensive studies on PAE for multichannel signals are required.

A COMPARISON BETWEEN BSS AND PAE

Both BSS and PAE are extensively applied in sound scene decomposition—a comparison between these approaches is summarized in Table 2. The common objective of BSS and PAE is to extract useful information (mainly the sound sources and their directions) about the original sound scene from the mixtures, and to use this information to facilitate natural sound rendering. There are three common characteristics in BSS and PAE. First, only the mixtures are available and usually no other prior information is given. Second, the extraction of the specific components from the mixtures is based on certain signal models. Third, both techniques require objective and subjective evaluation.

As discussed earlier, the applications of different signal models in BSS and PAE lead to different techniques. In BSS, the mixtures are considered as the sums of multiple sources, and the independence among the sources is one of the most important characteristics. In contrast, the mixing model in PAE is based on human perception of directional sources (primary components) and diffuse sound environment (ambient components). The perceptual difference between primary and ambient components is due to the directivity of these components which can be characterized by their correlations. The applications that adopted BSS and PAE also have distinct differences. BSS is commonly used in speech and music applications, where the clarity of the sources is usually more important than the effect of the environment. On the other hand, PAE is more suited for the reproduction of movie and gaming sound content, where the

[TABLE 2] COMPARISON BETWEEN BSS AND PAE IN SOUND SCENE DECOMPOSITION.

	BSS	PAE
OBJECTIVE	TO OBTAIN USEFUL INFORMATION ABOUT THE ORIGINAL SOUND SCENE FROM GIVEN MIXTURES AND FACILITATE NATURAL SOUND RENDERING.	
COMMON CHARACTERISTICS	USUALLY NO PRIOR INFORMATION, ONLY MIXTURES ■ BASED ON CERTAIN SIGNAL MODELS ■ REQUIRE OBJECTIVE AS WELL AS SUBJECTIVE EVALUATION	
BASIC MIXING MODEL	SUMS OF MULTIPLE SOURCES (INDEPENDENT, NON-GAUSSIAN, ETC.)	PRIMARY COMPONENTS (HIGHLY CORRELATED) AND AMBIENT COMPONENTS (UNCORRELATED)
TECHNIQUES	ICA [14], SPARSE SOLUTIONS [15], TIME-FREQUENCY MASKING [16], NMF [17], [18], CASA [19], ETC.	PCA [20], LS [8], [21], TIME-FREQUENCY MASKING [7], [20], TIME/PHASE-SHIFTING [22], [23], ETC.
TYPICAL APPLICATIONS	SPEECH, MUSIC	MOVIE, GAMING
RELATED APPLICATIONS	SPEECH ENHANCEMENT, NOISE REDUCTION, SPEECH RECOGNITION, MUSIC CLASSIFICATION	SOUND REPRODUCTION, SOUND LOCALIZATION, CODING
LIMITATIONS	■ SMALL NUMBER OF SOURCES ■ SPARSENESS/DISJOINT ■ NO/SIMPLE ENVIRONMENT	■ SMALL NUMBER OF SOURCES ■ SPARSENESS/DISJOINT ■ LOW AMBIENT POWER ■ PRIMARY AMBIENT COMPONENTS UNCORRELATED

ambient components also contribute significantly to the naturalness and immersiveness of the sound scenes. Subjective experiments revealed that BSS- and PAE-based headphone rendering can improve the externalization and enlarge the sound stage with minimal coloration [6].

Despite the recent advances in BSS and PAE, the challenges due to the complexity and uncertainty of the sound scenes still remain to be resolved. One common challenge in both BSS and PAE is the increasing number of audio sources in the sound scenes, while only a limited number of mixtures (i.e., channels) are available. In certain time-frequency representations, the sparse solutions in BSS and PAE would require the sources to be sparse and disjoint [15]. Considering the diversity of audio signals, finding a robust sparse representation for different types of audio signals is extremely difficult. The recorded or postprocessed source signals might even be filtered due to physical or equivalently simulated propagation and reflections. Moreover, the audio signals coming from adverse environmental conditions (including reverberation and strong ambient sound) usually degrade the performance of the decomposition. These difficulties can be addressed by studying the features of the resulting signals and by obtaining more prior information on the sources, the sound environment, the mixing process [18], and combining auditory with visual information of the scene.

INDIVIDUALIZATION OF HRTF

Binaural technology is the most promising solution for delivering spatial audio in headphones, as it is the closest to natural listening. Unlike conventional microphone recordings, which are meant for loudspeaker playback, the binaural signals are recorded or synthesized at the ears of the listener. In a binaural audio system, the spatial encoding (i.e., HRTFs) should encapsulate all the spectral features due to the interaction of the acoustic wave with the listener's morphology (torso, head, and pinna). The pinna, which is also considered as the acoustic fingerprint, embeds the most idiosyncratic spectral features into HRTFs, which are essential for accurate perception of the sound [Figure 3(a)]. Thus, the HRTF features of the individuals are extremely unique, as shown in Figure 3(c). Often the HRTFs used for virtualization are nonindividualized HRTFs, typically measured on a dummy's head, since they are easily accessible.

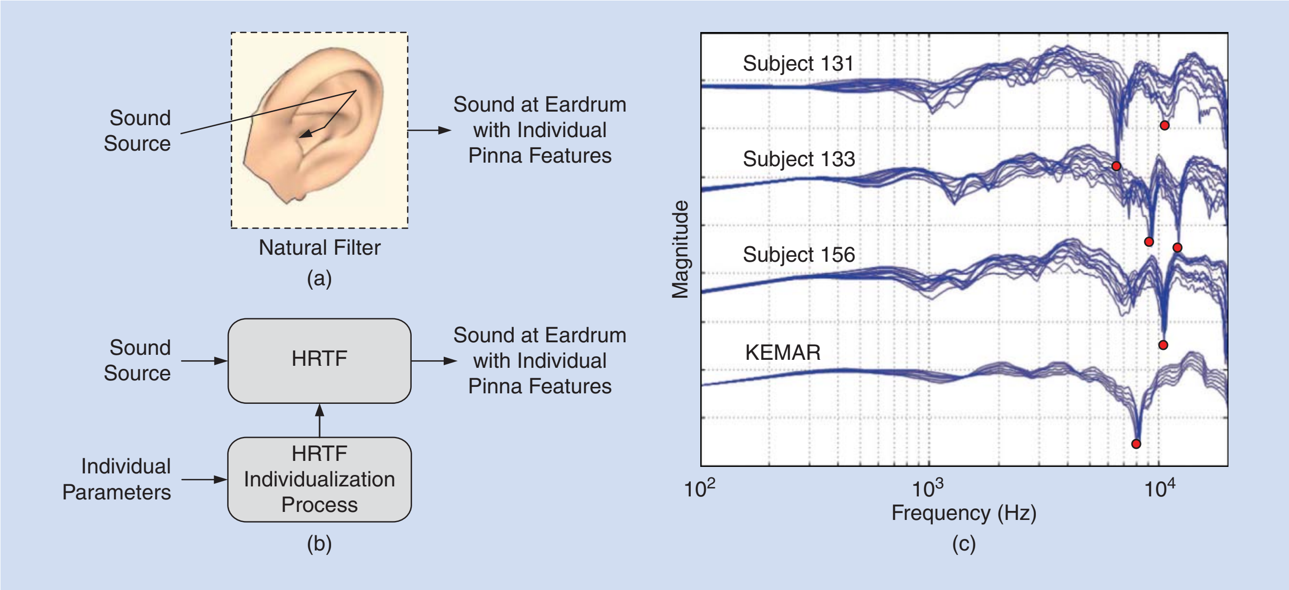
However, the use of nonindividualized HRTFs leads to several artefacts like IHL, localization inaccuracies in perceiving elevation, and front–back, up–down reversals. Additionally, subjects display poor angular resolution and sometimes find it difficult to pinpoint the exact location of the auditory image in the case of using nonindividualized HRTFs. Thus, individualization of the HRTFs [Figure 3(b)] plays a critical role to create an immersive experience closest to the natural listening experience. There are various individualization techniques to obtain the individualized HRTFs from acoustical measurements, anthropometric features of the listener, customizing generic HRTFs with perceptual feedback or frontal projection of sound, as summarized in Table 3.

ACOUSTICAL MEASUREMENTS

The most straightforward individualization technique is to actually measure the individualized HRTFs for every listener at different sound positions [25], [26]. This is the most ideal solution but it is extremely tedious and involves highly precise measurements. These measurements also require the subjects to remain motionless for long periods, which may cause the subjects fatigue. Zotkin et al. developed a fast HRTF measurement system using the technique of reciprocity, where a microspeaker is placed into the ear and several microphones are placed around the listener [13]. Other researchers developed a continuous 3-D azimuth acquisition system to measure the HRTFs using a multichannel adaptive filtering technique [27]. However, all these techniques to acoustically measure the individual HRTFs require a large amount of resources and expensive setups.

ANTHROPOMETRIC DATA

Individualized HRTFs can also be modeled as weighted sums of basis functions, which can be performed either in the frequency or spatial domain. The basis functions are usually common to all individuals and the individualization information is often conveyed by the weights. The HRTFs are essentially expressed as weighted sums of a set of eigenvectors, which can be derived from PCA or ICA [26], [13]. The individual weights are derived from the anthropometric parameters that are captured by optical descriptors, which can be derived from direct measurements, pictures, or a 3-D mesh of the morphology [13]. The solution to the problem of diffraction of an acoustic wave with the listener's body results in individual HRTFs. This solution

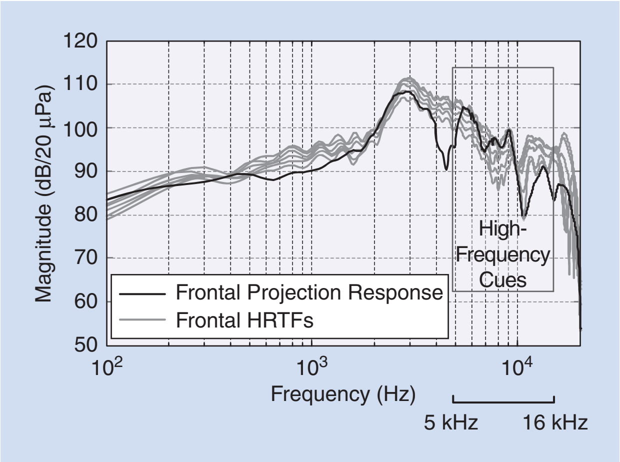


[FIG3] (a) Human ears act as a natural filter in physical listening. (b) The natural HRTF filter is modeled by a digital filter using various individualization techniques. (c) Note the vast variation of the HRTF spectrum at high frequencies of the various subjects taken from the Center for Image Processing and Integrated Computing (CIPIC) database and the Massachusetts Institute of Technology's Knowles Electronic Manikin for Acoustic Research (KEMAR) dummy head database [26]. This is due to the idiosyncratic nature of the pinna.

[TABLE 3] A COMPARISON OF THE VARIOUS HRTF INDIVIDUALIZATION TECHNIQUES.

HOW TO OBTAIN INDIVIDUAL FEATURES	TECHNIQUES	PROS	CONS	PERFORMANCE AND REMARKS
ACOUSTICAL MEASUREMENTS	INDIVIDUAL MEASUREMENTS [25], IRCAM FRANCE, CIPIC, UNIVERSITY OF MARYLAND, TOHOKU UNIVERSITY, NAGOYA UNIVERSITY AUSTRIAN ACADEMY OF SCIENCES [26]	IDEAL, ACCURATE	REQUIRES HIGH PRECISION; TEDIOUS; IMPRACTICAL FOR EVERY LISTENER	REFERENCE FOR INDIVIDUALIZATION TECHNIQUES
ANTHROPOMETRIC DATA	OPTICAL DESCRIPTORS: 3-D MESH, 2-D PICTURES [13] ANALYTICAL OR NUMERICAL SOLUTIONS: PCA + MULTIPLE LINEAR REGRESSION [26] FEM, BEM [26], [13], MULTIWAY ARRAY ANALYSIS [28], ARTIFICIAL NEURAL NETWORK [26] STRUCTURAL MODEL OF HRTFs [13], HRTF DATABASE MATCHING [30]	BASED ON ACOUSTIC PRINCIPLES; STUDIES THE EFFECTS OF INDEPENDENT ELEMENTS OF THE MORPHOLOGY	NEED A LARGE DATABASE; TEDIOUS; REQUIRES HIGH-RESOLUTION IMAGING; EXPENSIVE EQUIPMENT; QUALIFIED USERS	USES THE CORRELATION BETWEEN INDIVIDUAL HRTF AND ANTHROPOMETRIC DATA
LISTENING/TRAINING	SELECTION FROM NONINDIVIDUALIZED HRTF [13], FREQUENCY SCALING [31] TUNE MAGNITUDE SPECTRUM [13], ACTIVE SENSORY TUNING [26], PCA WEIGHT TUNING [32] SELECT CEPSTRUM PARAMETERS [34]	EASY TO IMPLEMENT; DIRECTLY RELATES TO PERCEPTION	TAKES TIME; REQUIRES REGULAR TRAINING; CAUSES FATIGUE	OBTAINS THE BEST HRTFs PERCEPTUALLY
PLAYBACK MODE	FRONTAL PROJECTION HEADPHONE [33]	NO ADDITIONAL MEASUREMENT, LISTENING TRAINING	NEW STRUCTURE; NOT APPLICABLE TO NORMAL HEADPHONES; TYPE-2 EQUALIZATION	AUTOMATIC CUSTOMIZATION, REDUCED FRONT–BACK CONFUSIONS
NONINDIVIDUALIZED HRTF	GENERALIZED HRTF [1]	EASY TO IMPLEMENT	NOT ACCURATE; POOR LOCALIZATION	NOT AN INDIVIDUALIZATION TECHNIQUE

may be obtained by analytical or numerical methods, such as the boundary element method (BEM) or the finite element method (FEM) [13], [26]. Other methods used include multiple linear regressions [26], multiway array analysis [28], and artificial neural networks [26]. The inputs to these methods can be a simple geometrical primitive [29] (e.g., a sphere, cylinder, or an ellipsoid), a 3-D mesh obtained from a magnetic resonance imaging (MRI) machine or laser scanner or a set of two-dimensional (2-D) images [13]. An important advantage of these techniques is that the relative effects of a particular morphological element (e.g., torso, head, and pinna) and their variation with size, location, and shape can be independently investigated [13]. Another technique used a simple customization technique, where an HRTF is selected by matching certain anthropometric



[FIG4] A comparison of the frontal projection headphone response and the frontal directional HRTFs measured on a dummy's head. (Figure used courtesy of [33].)

parameters [30]. One of the major challenges today to numerically model the HRTF is the very high resolution of imaging techniques required for accurate prediction of HRTFs at high frequencies. The required resolution of the mesh imaging depends on the shortest wavelength, which is around 17 mm at 20 kHz [13]. Moreover, obtaining these optical descriptors demands for the use of extremely expensive laser, MRI scanners, and also requires highly skilled, qualified users.

PERCEPTUAL FEEDBACK

Several attempts have been carried out to personalize HRTF from a generic HRTF database using perceptual feedback. Subjects select the HRTFs through listening tests, where they choose the HRTFs based on the correct perception of frontal sources and reduced front–back reversals [13]. Listeners can also adapt to the nonindividualized HRTF by modifying the HRTFs to suit his or her perception. Middlebrooks observed that the peaks and notches of HRTFs are frequency shifted for different individuals and that the extent of the shift is related to the size of pinna [31]. Listeners often tune the spectrum until they achieve a satisfactory and natural spatialization [13]. Other techniques involve active sensory tuning [26] and tuning the PCA weights [32] to individualize the HRTFs. These perceptual-based methods are much simpler in terms of the required resources and effort compared to the individualization methods using acoustical measurements or anthropometric data. However, these listening sessions can sometimes be quite long and result in listener fatigue.

FRONTAL PROJECTION PLAYBACK

More recently, a study by Sunder et al. [33] customized the nonindividualized HRTFs using a frontal projection headphone. Unlike side projection of sound in conventional headphones, a frontal projection headphone projects the sound from the front to emulate the playback from a physical set of loudspeakers. By projecting the sound from the front, the idiosyncratic frontal pinna spectral cues of the listener are captured inherently during the playback [33]. It is found that the idiosyncratic high-frequency pinna cues captured in the frontal projection headphones response match well with the frontal HRTF cues, giving it a better frontal perception (as shown in Figure 4). The authors of [33] reported that the front–back reversals were reduced by almost 50% [33] using the frontal projection headphone, thus improving the veracity of the 3-D audio. The advantage of this technique is that it does not require any measurements, training, or the anthropometric data of the listener. However, the frontal projection individualization technique has been limited to only the horizontal plane and also requires a special kind of headphone equalization (Type-2).

As discussed previously, head tracking is important in the virtualization process. It was found that head tracking, when used with nonindividualized HRTFs, can improve the localization [10]. However, head tracking primarily helps in reducing the front–back confusions and has minimal effect in reducing the elevation localization errors, IHL [10], and coloration caused by nonindividualized HRTFs. Since individualization of HRTFs can alleviate some of these limitations, it is suggested that head tracking be used with individualized rendering.

In summary, there is a noticeable trend to achieve more and more accurate individualization with lesser data, complexity, and effort. However, the effect of individualization of HRTFs can be hindered by the presence of the headphones. Hence, the headphones have to be compensated to ensure that the spectrum at the eardrum has only the individualized HRTF features. Additionally, equalization of the binaural recording itself may be necessary in certain applications (e.g., musical recordings). The challenges and methods of equalization for both binaural and stereo recordings are explained in the next section.

EQUALIZATION

Headphones are not acoustically transparent as they not only color the sound that is played from the headphone but also affect the free-air characteristics at the ear. Typically the HPTF comprises the headphones transducer response and the acoustic coupling between the headphones and the listener's ears. To compensate for the headphone response, the HPTF is first measured at the same point where the recording was carried out at the blocked ear canal or at the eardrum [35]. The binaural recording is then deconvolved with the HPTF to eliminate the effect of the recording microphones and the headphone. This type of direct equalization is also known as the *nondecoupled* mode of equalization (Table 4) [36]. This method is often used when the HPTF is measured with the same measurement setup as the recording and particularly works well when the HPTF measurement and recording are carried out on the same dummy's head.

It is observed that, in the absence of headphone equalization, the front–back reversals are increased and the elevation localization is distorted [1], [13], [26]. Thus, headphone equalization is critical to create a convincing perception of virtual sound sources. However, headphone equalization is challenging since the HPTF depends on individual morphology (headphone–ear coupling). Researchers have also reported that the use of nonindividualized equalization can reduce the externalization and the effect can be

[TABLE 4] EQUALIZATION TECHNIQUES FOR DIFFERENT PLAYBACK MODES (BINAURAL, STEREOPHONY).

MODE OF EQUALIZATION	AIM	TYPES OF EQUALIZATION AND TARGET RESPONSE	CHARACTERISTICS
NONDECOUPLED (BINAURAL)	SPECTRUM AT EARDRUM IS THE INDIVIDUAL HRTF FEATURES	CONVENTIONAL EQUALIZATION (FLAT TARGET RESPONSE)	■ FOR CONVENTIONAL HEADPHONES. THE SPECTRUM AT THE EARDRUM HAS INDIVIDUAL FEATURES (IF INDIVIDUALIZED HRTF IS USED) ■ DEPENDENT ON THE INDIVIDUAL'S UNIQUE PINNA FEATURES
		TYPE-2 EQUALIZATION [33]	■ FOR FRONTAL PROJECTION HEADPHONES. THE SPECTRUM AT EARDRUM AUTOMATICALLY MODELS THE INDIVIDUAL PINNA SPECTRAL CUES ■ REMOVES ONLY THE DISTORTION DUE TO THE HEADPHONE EMITTER ■ INDEPENDENT OF THE IDIOSYNCRATIC FEATURES OF THE EAR
DECOUPLED (BINAURAL, STEREOPHONY)	EMULATE THE MOST NATURAL REPRODUCTION CLOSER TO THE PERCEPTION IN A REFERENCE FIELD	FF EQUALIZATION [38]	■ TARGET RESPONSE IS THE FF RESPONSE CORRESPONDING TO THE FRONTAL INCIDENCE
		DF EQUALIZATION [38]	■ TARGET RESPONSE IS THE DF RESPONSE ■ LESSER INTERINDIVIDUAL VARIABILITY
		DF TARGET RESPONSE BASED ON MØLLER [38]	■ TARGET RESPONSE BASED ON AVERAGE OF HRTFS BETWEEN ± 45 DEGREES AZIMUTH AND ELEVATION WITH UNEQUAL WEIGHTING
		DF TARGET RESPONSE BASED ON LORHO [4]	■ REDUCED A 3-KHZ PEAK FROM ABOUT 12 DB TO 3 DB OF DF RESPONSE
		RR_G AND RR1_G [4]	■ RR_G: BASED ON THE IMPULSE RESPONSE OF HARMAN REFERENCE LISTENING ROOM ■ RR1_G HAS LESSER BASS AND TREBLE

as critical as the use of nonindividualized HRTFs [13]. Thus, equalization using individual HPTFs is strongly recommended. Another difficulty in carrying out accurate headphone equalization is the variability of the HPTFs with repositioning. The effect of repositioning of headphones is lower at low frequencies but displays high standard deviations up to 10 dB at high frequencies [37]. Kulkarni et al. [37] observed that equalization based on a single measurement may become worse than no equalization at all. The positional dependency has no specific solution and its effect can only be reduced by taking the average of a number of trials as a representative HPTF [37]. Thus, to create a convincing immersive sound environment, use of individualized HRTFs and individualized equalization is entailed, which may not be viable all the time. To reduce the dependency on individualized equalization, Sunder et al. [33] designed a Type-2 equalization technique for the playback through frontal projection headphone, which is independent of the headphone-ear coupling. Unlike the conventional equalization technique, Type-2 equalization compensates only for the distortion due to the emitter, thereby preserving the individual pinna cues due to frontal projection.

The other type of equalization is the "decoupled" equalization technique, and it is the most commonly used method of equalization for rendering music. In this technique, the binaural recording [(BIR) or HRTFs] as well as the headphone are equalized using a reference sound field (REF) (e.g., FF, DF, etc.) [36]. If the REF of the recording environment is well known and reproduced reliably, this method of equalization can result in a very natural perception of sound similar to the nondecoupled equalization technique. This method of equalization is mainly carried out to make the binaural recordings compatible with stereophonic (conventional microphone) recordings in terms of timbral quality.

If the recording is binaural, then a reference field equalized binaural recording (BIR/REF) achieves a sound quality equivalent to a conventional microphone recording. When the equalized recording is played from a reference field equalized headphone (HPTF/REF), the perceived timbre of the spatial sound would be as natural as the original binaural recording. Individualized binaural recordings are thus necessary to experience the true immersiveness of sound without any timbral coloration and spatial degradation. Note that for rendering conventional stereo recorded music, it is sufficient to carry out just the headphone equalization using an appropriate reference field. Some of the commonly used reference fields are:

- *Free-field (FF) equalization*: With the aim to replicate the ear signals produced by frontal loudspeakers, the target response of FF equalization is the HRTF of frontal incidence. Hammershoi et al. proposed an FF equalization curve, which has additional high frequency energy above 3 kHz to approximate listening to stereo loudspeakers in the FF [4]. A FF equalized headphone can reproduce a frontal sound with natural sound quality but colors the sound that originates from other directions. Moreover, it is important to note that there are large interindividual variations in the FF equalization filters [38].
- *Diffuse-field (DF) equalization*: In this case, the target response for equalization is the DF response, i.e., the average of the HRTFs of all measured directions in horizontal plane. The interindividual variations are reduced drastically due to the averaging effect [38]. Thus, the DF target response can be achieved universally over a great number of individuals. Møller [35] identified certain headphones which are already DF equalized and recommended such type of headphones for stereo listening.
- *Other target responses*: A typical listening room is not completely diffuse but it can be considered somewhere between a FF and a DF. Møller [38] illustrated other alternative target responses which are partially diffuse by applying unequal weighting to different directions within ± 45° azimuth and elevation. Other researchers also modified the DF equalization filters with the help of certain parametric filters and found that the subjects generally preferred the target response with a 3 kHz peak lower in amplitude than in the DF response for both music and speech [4]. Recent

experiments [4], [38] showed that listeners prefer other alternative target responses more than the conventional FF and DF equalizations. Examples of these preferred target curves include RR_G and RR1_G proposed by Olive et al. [4] based on the impulse response of the loudspeaker system in the Harman Reference rooms.

Ideally, the best reference field that preserves the true quality of the recording would be the field where the recording is carried out. Furthermore, the choice of headphones can also greatly affect the transparency of the binaural rendering even with the correct headphone equalization. The external ear is unhindered in the natural listening conditions, where the sound pressures at the ear are governed by free-air characteristics. With headphones placed over the ear, the pressure characteristics of the sound arriving at the eardrum are greatly affected compared to the free-air characteristics due to the interaction between the external ear and the headphone enclosure. The closer the coupling characteristic of the headphones with that of the free-air, the more accurate and transparent is the reproduced sound. Møller [35] defined the effect of the headphone for a binaural recording at the blocked ear canal in terms of the electrical transmission gain, G:

$$G = \left(\frac{1}{\mathrm{MPTF} \cdot \mathrm{HPTF}}\right) \cdot \mathrm{PDR}, \tag{7}$$

where MPTF is the transfer function of the recording microphone, and PDR is the pressure division ratio. PDR is defined as the ratio of the equivalent Thévenin impedances when the ear is in free-air to the case when the headphone is placed on the ear, and is given as [35]:

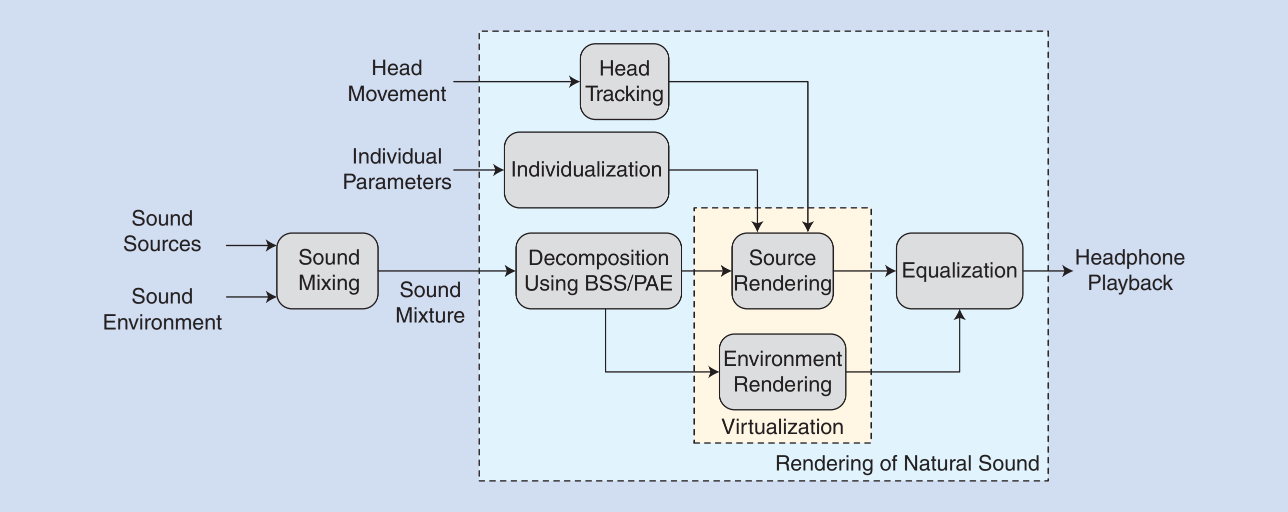
$$\mathrm{PDR} = \frac{Z_{\text{earcanal}} + Z_{\text{headphones}}}{Z_{\text{earcanal}} + Z_{\text{radiation}}}, \tag{8}$$

where Z_{earcanal} and $Z_{\text{headphones}}$ are the input impedances of the ear canal and the impedance of the headphone, respectively; $Z_{\text{radiation}}$ is the free-air radiation impedance as seen from the ear canal. The PDR reduces to unity when the pressures in the free-air and with headphones become equal. Such headphones are defined as FEC (free-air equivalent coupling) headphones, which are also sometimes called *open headphones* [35]. Open headphones are different from the commercially available "open–back headphones." Most of the commercially available headphones have less than ideal FEC characteristics [35]. It is important to note that the FEC condition for the headphone is necessary only for binaural recordings made at the blocked ear canal, which is also the most common technique for individualized binaural recording [35]. In such a case, headphone equalization alone is sufficient to achieve auralization transparency. To summarize, equalization (both recording and playback) and individualization play a critical role in the natural rendering of sound of any formats (binaural or stereo) over headphones.

BY PROJECTING THE SOUND FROM THE FRONT, THE IDIOSYNCRATIC FRONTAL PINNA SPECTRAL CUES OF THE LISTENER ARE CAPTURED INHERENTLY DURING THE PLAYBACK.

INTEGRATION OF NATURAL SOUND RENDERING TECHNIQUES

An integration of these signal processing techniques for natural sound rendering reviewed in this article is depicted in Figure 5. The original sound sources along with their environmental information are represented as a sound mixture after the mixing process. The sound scenes from the mix are then decomposed into primary components (sources) and/or ambient components (environment) using BSS and/or PAE. The extracted primary components, which are basically directional sound sources as perceived by the listener, can be rendered using (individualized) HRTFs [1]. Ambient components



[FIG5] The natural sound rendering system for headphones: an integration of all the signal processing techniques reviewed in this article.

are rendered in a manner so as to recreate a natural sound environment. Modeling the acoustics of the natural sound environment by adding the correct amount of early reflections and reverberation also helps in enhancing the perception of the sound environment as well as veridical distance, which is critical for natural listening. Moreover, a suitable individualization technique has to be applied to the directional sources such that the rendered sound scenes played over headphones are maximally tailored for the individual listener. Meanwhile, the use of a robust equalization technique can significantly reduce the adverse coloration of the source. Finally, the influence of the head movements on the rendered sound can be taken into account by incorporating head tracking in virtualization.

[IN GENERAL, NATURAL SOUND RENDERING REQUIRES BOTH THE SPATIAL AND TIMBRAL QUALITY OF THE REPRODUCED SOUND TO BE REALISTIC.]

In general, natural sound rendering requires both the spatial and timbral quality of the reproduced sound to be realistic. For digital media content that contains plenty of spatial cues (e.g., movies, games), all five techniques reviewed are important in creating a sense of immersiveness. For other content, where the timbral quality is of utmost importance (e.g., music recordings), a subset of the techniques (e.g., individualization, equalization) are sufficient.

SUBJECTIVE EXPERIMENTS

Subjective experiments were carried out to validate the reviewed natural sound rendering system by comparing it with the conventional stereo playback system. A total of 18 subjects (15 males and three females), who were all between 20 and 30 years old, participated in this listening experiment. None of the subjects reported any hearing loss. The test was conducted in a semianechoic listening room at Nanyang Technological University (NTU) in Singapore. The two systems of headphone listening tested in this experiment were:

- *Conventional stereo system*: The materials are directly played back over headphones without any processing.
- *Natural sound rendering system*: The signal processing techniques introduced in the article were applied to the audio content.

In this study, we chose PAE as the sound scene decomposition method since our primary interest lies in movie and gaming audio content that contains the individual sound sources and the sound environment [21]. Individualization is carried out by frontal projection headphones since it inherently embeds the personal pinna cues during playback and does not require any individual acoustical experiments, anthropometric data, or training [33]. To fully exploit the frontal projection in the natural sound rendering, we have developed a new four-emitter headphone [39] that houses a frontal emitter and a conventional side emitter in each ear cup of the headphone [33]. In the virtualization process, the frontal emitters are used to render the directional sources, while all the emitters (both frontal and side) are used to render the sound environment. Type-2 equalization is applied to the frontal emitters for source rendering [33], and DF equalization is used to render environment signals over all the emitters. Head tracking has not been incorporated in this system.

The stimuli used in this experiment were binaural (a motorcycle in a storm and a bee at a waterfall), movies (*Brave, Prometheus*), and gaming tracks (*Battlefield 3*), which contain numerous spatial cues. Each track was played back using the two headphone playback systems tested in this article. The tracks corresponding to the two systems were named "A" and "B" and played back in a random order. The listening tests were conducted in a double-blind manner, where both the experimenter and the subjects were unaware of the order of the stimuli. In this experiment, four audio quality measures were considered to evaluate the performance of the two systems. Their descriptions are:

1) *Sense of direction*: How clear or distinct are the perceived directions of the sound objects?

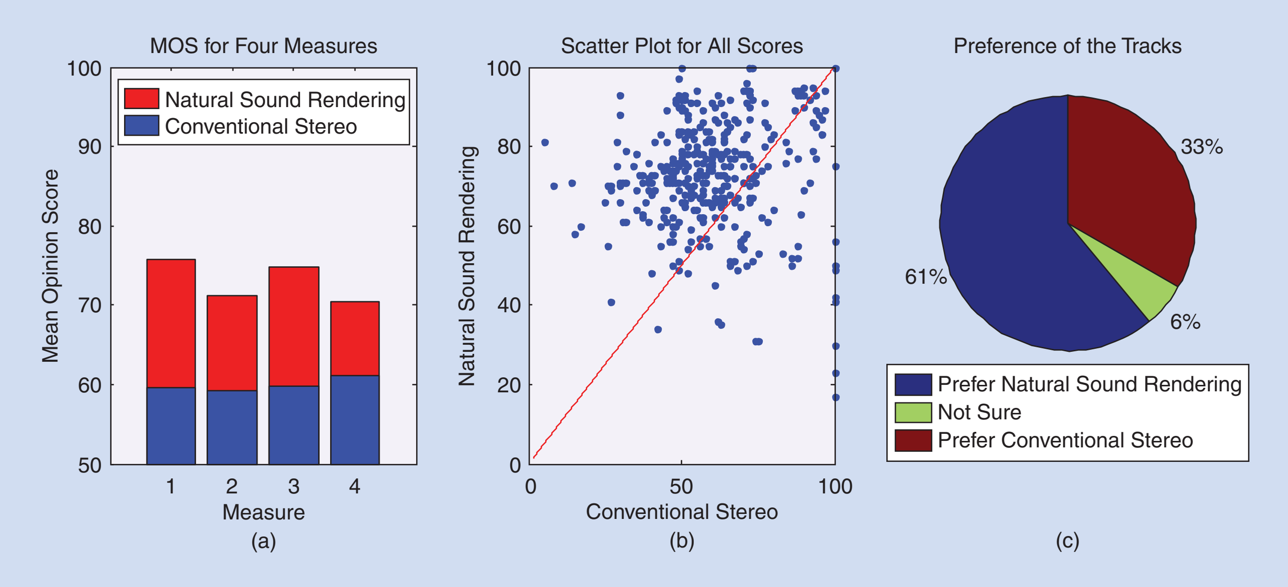
2) *Externalization*: How clear is the stimulus perceived outside the head?

3) *Ambience*: How clear and natural is the perceived ambience of the sound environment?

4) *Timbral quality*: How realistic is the timbral quality of the sound?

Subjects were asked to give scores for the four measures for each of the two tracks "A" and "B." The scores were based on a 0–100 scale where subjects rated 0–20 (Bad), 21–40 (Poor), 41–60 (Fair), 61–80 (Good), and 81–100 (Excellent). Finally, the subjects were also required to indicate their overall preference for the two tracks by selecting one of the following three choices: "Prefer A," "Not sure," or "Prefer B." To carry out this experiment, a graphical user interface was created, which randomized the order of the stimuli and automatically stored the responses of the subjects in a file.

The responses of the subjects were analyzed for both sound rendering systems. Figure 6 shows the overall comparison between the two systems in terms of the mean opinion score (MOS), scatter plot, and the overall preference of the subjects. In (a), the MOS of the four measures for the two systems were computed across all 18 subjects and five stimuli. While the MOS for the conventional stereo system for all the measures were around 60, the natural sound rendering system performed much better with an MOS of over 70. An analysis of variance (ANOVA) was conducted to generalize these results to the whole population of listeners. The p-values were found to be very small ($<< 0.01$) for all measures, indicating that the improved performance of the natural sound rendering system over the conventional stereo system is statistically significant. The scatter plot in Figure 6(b) implies that most of the subjects gave a higher score for the natural sound rendering system for all the four measures. The overall preference of the subjects across all the five tracks is shown in Figure 6(c). The



[FIG6] Results of the subjective experiments: (a) MOS, (b) scatter plot, and (c) overall preference.

pie chart suggests that 61% of the subjects preferred the natural sound rendering, while only 33% preferred the conventional stereo rendering.

To sum up the subjective test results, we found that the natural sound rendering system using the various signal processing techniques explained in this article enhances the listening experience compared to a conventional stereo system. Additionally, the presence of head tracking in the system will only improve the natural sound rendering as observed in several studies [10].

CONCLUSIONS AND FUTURE TRENDS

With the advent of low cost, low power, small form factor, and high-speed multicore embedded processor, we can now implement the aforementioned signal processing techniques in real time and embed processors into the headphone design. However, various implementation issues regarding the computation cost of sound scene decomposition, HRTF/BRIR filtering in virtualization, and equalization as well as the latency in head tracking should be carefully considered. One example of such a natural sound rendering system is the four-emitter 3-D audio headphone [39] developed at the Digital Signal Processing Lab at NTU. This system has been psychophysically validated and found to perform much better than the conventional stereo headphone playback system.

Besides the five types of techniques discussed in this article, there have been other efforts to enhance the natural experience of headphone listening. To enable the natural pass through of the sound from outside world without coloration, headphones can be designed with suitable acoustically transparent materials. When this is not effective, microphones integrated into headphones and associated signal processing techniques, such as equalization, and active noise control, are employed. The headphones with built-in microphones open a new dimension to augment the listening experience with the physical world.

THE FUTURE OF HEADPHONES FOR ASSISTIVE LISTENING APPLICATIONS WOULD BE WHERE LISTENERS CANNOT DIFFERENTIATE BETWEEN THE VIRTUAL ACOUSTIC SPACE CREATED FROM HEADPHONE PLAYBACK AND THE REAL ACOUSTIC SPACE.

The future of headphones for assistive listening applications would be where listeners cannot differentiate between the virtual acoustic space created from headphone playback and the real acoustic space. This would require a combined effort from the whole audio community—from the headphone manufacturers and sound engineers to audio scientists. More information about the content production has to be distributed from the content developers to the end user to enhance the extraction process. Moreover, obtaining and exploiting every individual's anthropometrical feature or hearing profile is crucial for a natural listening experience. Finally, with more sensors, such as global positioning systems, gyroscopes, and microphones that can be integrated into headphones, future headphones are becoming more content-, location-, and listener-aware, and hence more intelligent and assistive.

ACKNOWLEDGMENTS

This work is supported by the Singapore National Research Foundation Proof-of-Concept program under grant NRF 2011 NRF-POC001-033. We thank the guest editors and reviewers for their constructive comments and suggestions.

AUTHORS

Kaushik Sunder (KAUSHIK1@e.ntu.edu.sg) received his B.Tech degree in electrical and electronics engineering from the National Institute of Technology Karnataka, Surathkal, India, in 2011. He is currently pursuing his Ph.D. degree in electrical and electronics engineering at Nanyang Technological University, Singapore. His research interest includes spatial audio, psychoacoustics, and music signal processing.

Jianjun He (JHE007@e.ntu.edu.sg) received his B.Eng. degree in automation from Nanjing University of Posts and Telecommunications, China, in 2011 and is currently pursuing his Ph.D. degree in electrical and electronic engineering at Nanyang Technological University, Singapore. His research interests include audio and acoustic signal processing, three-dimensional audio, psychoacoustics, active noise control, and emerging audio and speech applications.

Ee-Leng Tan (ETanEL@ntu.edu.sg) received his B.Eng. (first class honors) and Ph.D. degrees in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2003 and 2012, respectively. Currently, he is with NTU as a research fellow. His research interests include image/audio processing and real-time digital signal processing.

Woon-Seng Gan (ewsgan@ntu.edu.sg) received his B.Eng. (first class honors) and Ph.D. degrees in electrical and electronic engineering from the University of Strathclyde, United Kingdom, in 1989 and 1993, respectively. He is currently an associate professor and the head of the Information Engineering Division, School of Electrical and Electronic Engineering at Nanyang Technological University. His research interests span a wide and related areas of adaptive signal processing, active noise control, and directional sound system.

REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Cambridge, MA: AP Professional, 2000.

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, Sept. 2013.

[3] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[4] S. Olive, T. Welti, and E. McMullin, "Listener preferences for different headphone target response curves," in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, May 2013, pp. 1–12.

[5] M. M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," in *Proc. 123rd Audio Engineering Society Convention*, New York, Oct. 2007, pp. 1–12.

[6] J. Breebaart and E. Schuijers, "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 8, pp. 1503–1511, Nov. 2008.

[7] C. Avendano and J. M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740–749, July 2004.

[8] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064, Nov. 2006.

[9] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Proc. 128th Audio Engineering Society Convention*, London, UK, May 2010, pp. 1–14.

[10] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[11] C. Faller and F. Baumgarte, "Binaural cue coding—Part II: Schemes and applications," *IEEE Trans. Speech Audio, Speech, Lang. Processing*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[12] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Processing Mag.*, vol. 28, no. 1, pp. 33–42, Jan. 2011.

[13] R. Nicol, *Binaural Technology*. New York: Audio Engineering Society, Inc., 2010.

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ: Wiley, 2004.

[15] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, June 2010.

[16] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[17] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. thesis, Tampere Univ. of Technology, 2006.

[18] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Mag.*, vol. 31, no. 3, pp. 107–115, 2014.

[19] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[20] J. Merimaa, M. M. Goodwin, and J. M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. 123rd Audio Engineering Society Convention*, New York, Oct. 2007, pp. 1–15.

[21] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, Lang. Processing,* vol. 22, no. 2, pp. 505–517, 2014.

[22] J. He, E. L. Tan, and W. S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Canada, May 2013, pp. 266–270.

[23] J. He, E. L. Tan, and W. S. Gan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 2892–2896.

[24] J. Thompson, B. Smith, A. Warner, and J. M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations," in *Proc. 133rd Audio Engineering Society Convention*, San Francisco, CA, 2012, pp. 1–15.

[25] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–321, May 1995.

[26] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: A review," in *Virtual Reality,* R. Shumaker, Ed. New York: Springer, 2007, pp. 397–407.

[27] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* Oct. 2009, pp. 325–328.

[28] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "HRTF customization using multiway array analysis," in *Proc. 18th European Signal Processing Conf. (EUSIPCO'10)*, Aalborg, Denmark, Aug. 2010, pp. 229–233.

[29] R. O. Duda, V. R. Algazi, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Proc. 113th Audio Engineering Society Convention*, Los Angeles, CA, Oct. 2002, pp. 1–18.

[30] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, Oct. 2003, pp. 157–160.

[31] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1480–1492, Sept. 1999.

[32] K. J. Fink and L. Ray, "Tuning principal component weights to individualize HRTFs," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Mar. 2012, pp. 389–392.

[33] K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 989–1000, Dec. 2013.

[34] R. Nicol, V. Lemaire, A. Bondu, and S. Busson, "Looking for a relevant similarity criterion for HRTF clustering: A comparative study," in *Proc. 120th Audio Engineering Society Convention*, Paris, France, May 2006, pp. 1–14.

[35] H. Møller, D. Hammershoi, C. B. Jensen, and M. F. Sorensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217, Apr. 1995.

[36] V. Larcher, J. M. Jot, and G. Vandernoot, "Equalization methods in binaural technology," in *Proc. 105th Audio Engineering Society Convention*, San Francisco, CA, Sept. 1998, pp. 1–29.

[37] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Amer.*, vol. 107, no. 2, pp. 1071–1074, Feb. 2000.

[38] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218–232, Apr. 1995.

[39] W. S. Gan and E. L. Tan, "Listening device and accompanying signal processing method," U.S. Patent 2014/0153765 A1, 2014.

Tiago H. Falk, Vijay Parsa, João F. Santos, Kathryn Arehart, Oldooz Hazrati, Rainer Huber, James M. Kates, and Susan Scollie

Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

Advantages and limitations of existing tools

This article presents an overview of 12 existing objective speech quality and intelligibility prediction tools. Two classes of algorithms are presented—intrusive and nonintrusive—with the former requiring the use of a reference signal, while the latter does not. Investigated metrics include both those developed for normal hearing (NH) listeners, as well as those tailored particularly for hearing impaired (HI) listeners who are users of assistive listening devices [i.e., hearing aids (HAs) and cochlear implants (CIs)]. Representative examples of those optimized for HI listeners include the speech-to-reverberation modulation energy ratio (SRMR), tailored to HAs (SRMR-HA) and to CIs (SRMR-CI); the modulation spectrum area (ModA); the HA speech quality (HASQI) and perception indices (HASPI); and the perception-model-based quality prediction method for hearing impairments (PEMO-Q-HI). The objective metrics are tested on three subjectively rated speech data sets covering reverberation-alone, noise-alone, and reverberation-plus-noise degradation conditions, as well as degradations resultant from nonlinear frequency compression and different speech enhancement strategies. The advantages and limitations of each measure are highlighted and recommendations are given for suggested uses of the different tools under specific environmental and processing conditions.

INTRODUCTION

According to 2005 estimates from the World Health Organization, 278 million people worldwide had moderate to profound hearing loss in one or both ears. Depending on the degree of hearing

Digital Object Identifier 10.1109/MSP.2014.2358871
Date of publication: 12 February 2015

impairment, these subjects can become candidates for HA or CI devices. Recently, a number of factors, such as aging population, enlargement of candidacy criteria, and technological advances have drawn great attention to HA and CI research and development. For users of such assistive listening devices, however, environmental distortions, such as reverberation and additive noise (and their combined effects) significantly degrade speech intelligibility and reduce perceived quality to unacceptable levels [1]. As such, current research has focused on the development of speech enhancement techniques (e.g., noise suppression, feedback cancellation) to meet this demand. To assure that the developed algorithms are behaving as expected, quality and intelligibility monitoring must be performed.

Traditionally, subjective tests have been used to assure that acceptable levels of speech quality and intelligibility are attained. For CI devices, two approaches are commonly taken. The first makes use of vocoded speech to simulate CI hearing and presents vocoded speech to NH listeners for identification. The second approach is more direct and presents degraded (or enhanced) speech stimuli directly to HI CI users for analysis (e.g., [1]). For HA users, this latter approach has been commonly used to investigate the effects of various HA signal processing techniques, such as noise suppression and feedback cancellation, on the perceived speech quality. Subjective testing, however, is laborious, time-consuming, and expensive. As such, automated, repeatable, fast, and cost-effective objective quality/intelligibility monitoring tools need to be developed, thus replacing the listeners with an auditory-inspired computational algorithm.

Reliable objective quality/intelligibility measurement tools can play key roles in the development, fitting, and online processing of different assistive listening devices. In the development stage, for example, different processing algorithms can be optimized to improve the final perceived speech quality/intelligibility. Wide dynamic-range compression algorithms have been developed to improve the audibility of low-intensity speech sounds. It is well known, however, that the time-varying gain changes can introduce unwanted nonlinear distortions. As such, objective tools provide a means of evaluating the tradeoffs between audibility and distortion, thus allowing for optimal parameters to be set. Moreover, for HA fitting, objective measures can be used to provide presettings tailored to the individual hearing loss, thus providing more effective starting points for the adjustment of the HA. Furthermore, the settings that provide optimum intelligibility may not be the ones that result in maximum quality, thus toggling between settings based on an intelligibility and on a quality index can provide a meaningful comparison for the HA user. Finally, objective tools can be used in the real-time adaptation of, e.g., speech enhancement algorithms (i.e., model-in-the-loop), such that the processing guarantees optimal quality/intelligibility as the user moves from one (noisy/reverberant) environment to another.

Signal-based objective metrics can be classified as intrusive or nonintrusive, depending on the need for a reference signal or not, respectively. While significant research and standardization efforts have been placed in developing objective measures for telephone speech with NH listeners [2], only a small number of objective measurement tools targeted toward CI/HA users have been developed. Given the rapidly aging population and the projected increase of hearing loss that comes with growing older, it is of great importance that the advantages and drawbacks of existing tools be characterized, as well as compared to each other on data sets collected under different practical experimental conditions.

In this article, we present several existing tools that have been recently developed for users of assistive listening devices; seven of the investigated tools belong to the intrusive class and five are nonintrusive. All the metrics were evaluated on the same data sets comprising speech processed under different complex listening conditions, such as noise, reverberation, noise-plus-reverberation, as well as under different nonlinear effects, such as frequency compression and speech enhancement (i.e., noise suppression and dereverberation). Advantages and limitations of the investigated tools are presented and suggestions as to which metrics are to be used under different specific scenarios are given, thus serving as a useful guide for researchers and developers of assisted listening devices.

OBJECTIVE SPEECH QUALITY AND INTELLIGIBILITY PREDICTION

Over the last two decades, significant standardization efforts have been made by the International Telecommunications Union (ITU-T) to standardize both intrusive and nonintrusive algorithms for telephone speech using NH listeners [2]. On the other hand, only a handful of algorithms have been proposed that are specifically tuned to assistive listening devices. To overcome this limitation, recent studies have explored the use of NH-optimized tools, as well as proposed modifications to such tools to tailor them to assistive listening devices (e.g., [3]). In the following sections, several such measures, both intrusive and nonintrusive, are described. The choice of measures used in this study was guided not only by their applicability to the task at hand, but also by the availability of publicly available source code (or code that could be licensed at a reasonable cost).

INTRUSIVE METRICS

NORMALIZED COVARIANCE METRIC

The normalized covariance metric (NCM) measure estimates speech intelligibility based on the covariance between the envelopes of the time-aligned reference and processed speech signals [4]–[6]. Computation of NCM values depends on deriving speech temporal envelopes, via the Hilbert transform, from outputs of a gammatone filter bank used to emulate cochlear processing. The normalized correlation between the reference and processed speech envelopes produces an estimate of the so-called apparent signal-to-noise ratio (SNR) ($\mathrm{SNR}_{\mathrm{app}}$) given by

$$\mathrm{SNR}_{\mathrm{app}}(k) = \left[10\log_{10}\left(\frac{r_k^2}{1-r_k^2}\right)\right]_{[-15,15]}, \tag{1}$$

where r_k is the correlation coefficient between the reference and processed speech envelopes estimated in filter bank channel k (typically, 23 gammatone channels are used), and the [–15], [15] operator refers to the process of limiting and mapping $\mathrm{SNR}_{\mathrm{app}}$ into that range. The last step consists of linearly mapping the apparent SNR to the [0, 1] range using the following rule:

$$\mathrm{SNR}_{\mathrm{final}}^{\mathrm{NCM}}(k) = \frac{\max(\min(\mathrm{SNR}_{\mathrm{app}}(k), +15), -15) + 15}{30}. \qquad (2)$$

The $\mathrm{SNR}_{\mathrm{final}}^{\mathrm{NCM}}$ values are then weighted in each frequency channel according to the so-called articulation index (AI) weights $W(k)$ recommended in the American National Standards Institute (ANSI) S3.5 Standard [7]. The final NCM value is given by:

$$\mathrm{NCM} = \frac{\sum_{k=1}^{K=23} W(k)\, \mathrm{SNR}_{\mathrm{final}}^{\mathrm{NCM}}(k)}{\sum_{k=1}^{k=23} W(k)}. \qquad (3)$$

The NCM has been widely used to characterize the perceived intelligibility for CI users (e.g., [3] and [4]).

SHORT-TIME OBJECTIVE INTELLIGIBILITY

The short-time objective intelligibility (STOI) metric is based on a correlation coefficient between the temporal envelopes of the time-aligned reference and processed speech signal in short-time overlapped segments [8]. The signals are first decomposed by a 1/3-octave filter bank, segmented into short-time windows, normalized, clipped, and then compared by means of a correlation coefficient. The normalization step compensates for, e.g., different playback levels, which do not have a strong negative effect on intelligibility. Clipping, in turn, sets an upper bound on how severely degraded one speech time-frequency unit can be. According to [8], clipping is used to avoid changes in intelligibility prediction once speech has already been deemed "unintelligible." The resultant correlation coefficients correspond to short-time intermediate intelligibility measures for each of the segments, which are then averaged to one scalar value corresponding to the predicted speech intelligibility for the processed signal. The STOI was originally proposed to assess the intelligibility of time-frequency weighted noisy speech and enhanced speech for NH listeners. Nonetheless, a channel selection algorithm for CIs that employs STOI has been recently proposed [9].

PERCEPTUAL EVALUATION OF SPEECH QUALITY

The International Telecommunications Union ITU-T P.862 standard, also known as Perceptual Evaluation of Speech Quality (PESQ) [10], is a widely used objective quality measurement standard algorithm. As with most intrusive algorithms, the first step in PESQ processing is to time-align the reference and processed speech signals. Once the signals are time aligned, they are mapped to an auditory representation using a perceptual model based on power distributions over time-frequency and compressive loudness scaling, and then their differences are taken. Positive differences indicate that components such as noise are present, whereas negative differences indicate that components have been omitted. With PESQ, different scaling factors are applied to positive and negative disturbances to generate the so-called symmetrical and asymmetrical disturbances. The final PESQ quality score is obtained as a linear combination of the symmetrical and asymmetrical disturbances, with weights optimized using telephony data. While the original PESQ algorithm described in [10] was developed for narrow-band speech (8-kHz sampling rate), wideband (16 kHz) extensions were described in [11] and are used in the experiments described herein. It is important to emphasize that the P.862 standard was recently superseded by ITU-T Recommendation P.863 [also known as Perceptual Objective Listening Quality Assessment (POLQA); see [2] and references therein], thus covering a wider scope of distortions and speech bandwidths (e.g., superwideband). POLQA, however, is not used in this study as its source code is not publicly available and its license is very costly.

HEARING AID SPEECH QUALITY AND INTELLIGIBILITY INDICES

As originally described in [12], the HA speech quality index (HASQI) uses an auditory model to analyze the reference and processed signals from an HA. The auditory model was recently extended in [13] and now serves as the basis of a unified approach for predicting both intelligibility [14] and quality [15]. This HASQI Version 2 model is used in the experiments described herein. The auditory model includes the middle ear, an auditory filter bank, the dynamic-range compression mediated by the outer hair cells in the cochlea, two-tone suppression (where a tone at one frequency can reduce the cochlear output for a tone at a different frequency), and the onset enhancement inherent in the inner hair-cell neural firing behavior. Hearing impairment is incorporated in the model as a broadening of the auditory filters with increasing hearing loss, a reduction in the amount of dynamic-range compression, a reduction in the two-tone suppression, and a shift in the auditory threshold.

The HA speech intelligibility index (HASPI), in turn, combines two measures of signal fidelity. The first measure compares the evolution of the spectral shape over time for the processed signal with that of the reference signal. The second measure cross-correlates the high-level portions of the two signals in each frequency band. The envelope measure is sensitive to the dynamic signal behavior associated with consonants, while the cross-correlation measure is more responsive to preserving the harmonics in steady-state vowels. The HASQI quality model incorporates the effects of noise and nonlinear distortions, as well as linear spectral changes. The noise and nonlinear terms combine two measurements. The first measurement compares the time-frequency envelope modulation of the processed and reference signals and is similar to the envelope comparison used in HASPI. The second measurement is based on normalized signal cross-correlations in each frequency band. The linear term compares the long-term spectra and the spectral slopes. The final quality prediction is the product of the two terms. Both HASPI [14] and HASQI [15] have been evaluated for NH and HI listeners over a wide range of processing conditions, including additive stationary and modulated noise, nonlinear distortion, noise suppression, dynamic-range compression, frequency compression, feedback cancellation, and linear filtering.

PERCEPTION-MODEL-BASED QUALITY PREDICTION

In its original version, the perception-model-based quality prediction method, PEMO-Q, compares the auditory-inspired "internal representation" of the reference speech signal to that of its processed counterpart to objectively characterize the quality of the processed speech signal [16]. The auditory representation is obtained given the following signal processing chain. First, the signals are split into critical bands using a gammatone filter bank. Each subband is half-wave rectified and low-pass filtered at 1 kHz. Envelope signals are then thresholded to account for the absolute hearing threshold and passed through an adaptation chain consisting of five consecutive nonlinear feedback loops. Finally, the envelope signal is either lowpass filtered at 8-Hz modulation frequency (in PEMO-Q's optional "fast mode") or analyzed by a linear modulation filter bank comprising eight filters with center frequencies up to 129 Hz (i.e., in the default mode used here). When comparing the reference and processed signals, two quality measures are produced: the overall perceptual similarity measure (PSM) and a per-frame counterpart PSM_t.

PSM corresponds to the overall cross-correlation coefficient between the complete internal representations of the reference and processed speech signals. PSM_t, in turn, is a more refined measure and explicitly accounts for the temporal course of the instantaneous audio quality as derived from a temporal frame-by-frame correlation of internal representations. While PSM provides greater generalizability, PSM_t has been found to be more sensitive to small distortions [16]. Since the experiments described in this article will be dealing with a wider range of speech quality levels, the PSM measure will be used. PSM was also previously shown to reliably predict the quality of speech enhancement algorithms [17].

More recently, an extension to PEMO-Q was developed to account for hearing impairments (PEMO-Q-HI) for HA users [18]. In the modified version, sensorineural hearing losses are modeled by an instantaneous expansion and an attenuation stage applied before the adaptation stage. While the former accounts for the reduced dynamic compression caused by the loss of outer hair cells, the latter accounts for the loss of sensitivity due to loss of inner hair cells [19]. With PEMO-Q-HI, the amount of attenuation and expansion is quantified from the impaired listeners' audiograms, as detailed in [18].

NONINTRUSIVE METRICS

ITU-T RECOMMENDATION P.563

In 2004, the ITU-T standardized its first nonintrusive algorithm called ITU-T P.563 [20]. The P.563 algorithm extracts a number of signal parameters to detect one of six dominant distortion classes. The distortion classes are, in decreasing order of "annoyance": high level of background noise, signal interruptions, signal-correlated noise, speech robotization (voice with metallic sounds), unnatural male speech, and unnatural female speech. For each distortion class, a subset of the extracted parameters is used to compute an intermediate quality rating. Once a major distortion class is detected, the intermediate score is linearly combined with other parameters to derive a final quality estimate. Unnaturalness of the speech signal is characterized by vocal tract and linear prediction analysis of the speech signal. More specifically, the vocal tract is modeled as a series of tubes of different lengths and time-varying cross-sectional areas, which are then combined with higher-order statistics (skewness and kurtosis) of the linear prediction and cepstral coefficients and tested to see if they lie within the restricted range expected for natural speech. While P.563 was developed as an objective quality measure for NH listeners and telephony applications, a recent study has shown promising results with P.563 as a correlate of noise-excited vocoded speech intelligibility for NH listeners, thus simulating CI hearing [21]. Note that the ITU-T P.563 algorithm is only applicable to narrowband speech signals sampled at 8-kHz sampling rate.

ModA

The ModA [22] measure is based on the principle that the speech signal envelope is smeared by the late reflections in a reverberant room, thus affecting the modulation spectrum of the speech signal. To obtain the ModA metric, the signal is first decomposed into $N(=4)$ acoustic bands (lower cutoff frequencies of 300, 775, 1,375, and 3,676 Hz, as in [22]); the temporal envelopes for each acoustic band are then computed using the Hilbert transform, downsampled and grouped using a 1/3-octave filter bank with center frequencies ranging between 0.5 and 8 Hz. As in [22], 13 modulation filters are used to cover the 0.5–10 Hz modulation frequency range. For each acoustic frequency band, the so-called area under the modulation spectrum is computed (A_i) and finally averaged over all $N = 4$ acoustic bands to obtain the ModA measure, which has been used as an intelligibility correlate for CI users in reverberant and enhanced conditions [22].

SRMR

The SRMR was originally developed for reverberant and dereverberated speech and evaluated against subjective NH listener data [23]. The metric is computed as follows. First, the input speech signal is filtered by a gammatone filter bank with center frequencies ranging from 125 Hz to approximately half the sampling frequency, and with bandwidths characterized by the equivalent rectangular bandwidth. For 8-kHz and 16-kHz sampled speech signals, 23 and 32 filters are used, respectively. Temporal envelopes are then computed via the Hilbert transform for each of the filter bank outputs and used to extract modulation spectral energy for each critical band. To emulate frequency selectivity in the modulation domain [24], modulation frequency bins are grouped into eight overlapping modulation bands with center frequencies logarithmically spaced between 4 and 128 Hz. Finally, the SRMR value is computed as the ratio of the average modulation energy content available in the first four modulation bands (3–20 Hz, consistent with clean speech content) to the average modulation energy content available in the last four modulation bands (20–120 Hz), consistent with room acoustics information [25].

SRMR-CI AND SRMR-HA

To tailor the SRMR measure for CI, a few modifications were recently implemented [26], [27]. First, the gammatone filter bank was replaced by the filter bank used in the speech coding strategy of the CI devices used by the listeners in the subjective test. Second, speech content variability was reduced by means of a modulation spectrum thresholding scheme [27]. Finally, to model the reduced sensitivity of the HI listeners, the 4–128 Hz range of the eight modulation filter bank center frequencies of the original SRMR metric was reduced to 4–30 Hz. The SRMR-CI has been tested as a correlate of intelligibility for CI users under clean, noisy, reverberant, noise-plus-reverberation, and speech-enhanced conditions [26], [27].

Similar to the modified SRMR-CI metric previously described, an alternate modification to the original SRMR metric has been performed to tailor it to HA devices [28]. First, the gammatone filter bank in the original SRMR implementation was modified to take into account the listener's individual hearing loss thresholds obtained via an audiogram. More specifically, the Q-factors of each of the filters were adjusted to simulate the hearing loss due to outer hair cell damage. Hence, as hearing loss increased, so did the filter bandwidths (i.e., Q-factors decreased). Additionally, the temporal Hilbert envelopes were compressed using a nonlinear compression function, similar to that used in the HASQI metric, to further model outer hair cell losses. For HA devices, it was found that the original 4–128 Hz range of modulation filter bank center frequencies was optimal, thus no changes were implemented in the modulation filter bank. The SRMR-HA was tested as a correlate of subjective quality for HA users in noisy, reverberant, and speech-enhanced conditions [28].

EXPERIMENTAL SETUP

In this section, the data sets used in the experiments as well as the evaluation criteria that will be used to characterize the performance of the investigated metrics are described.

CI SPEECH INTELLIGIBILITY DATA SET

This database is described in full detail in [1] and in the references therein. The material comprises speech data presented to CI users within the framework of an intelligibility subjective test. The speech sentences presented to the CI users were taken from the well-known IEEE sentence corpus. Four recorded room impulse responses were convolved with the clean speech data to simulate reverberant speech with reverberation times (RT60) of 0.3, 0.6, 0.8, and 1 s. Speech-shaped noise was also added to the anechoic and the reverberant signals to generate noise-only and noise-plus-reverberation degradation conditions, respectively. Noise was added at SNRs of -5, 0, 5, and 10 dB for the anechoic samples and 5 and 10 dB for the reverberant samples. For the noise-plus-reverberation condition, the reverberant signals served as reference for SNR computation. Additionally, the database includes sentences enhanced using an ideal reverberant masking (IRM) strategy [29]. These sentences were under reverberant conditions with RT60 s of 0.6, 0.8, and 1 s, and all of the noise-plus-reverberation conditions previously described. The IRM algorithm was configured to use two to three different threshold values for each condition. Speech files were sampled at 16 kHz with 16-bit resolution.

Eleven adult CI users were recruited to participate in the subjective intelligibility experiments. The participants were all native speakers of American English with postlingual deafness and had an average age of 64 years. All participants had a minimum of one year experience using their device routinely, with some being bilaterally implanted for over six years. For consistency, all participants were temporarily fitted with a SPEAR3 research processor (22 filter bank channels with Mel-like spacing) with parameters matching the individual CI user's clinical settings. Participants were presented with 31 lists of 20 sentences randomly selected from the IEEE database, each list being corrupted by the aforementioned degradation conditions. Degraded stimuli were presented directly to the audio input of the research processor and the level was adjusted individually for comfort at the beginning of the experiment. Listeners were instructed to repeat aloud each sentence after its presentation. A tester then marked the words correctly identified by the subject according to the ground truth transcript. Finally, the number of words correctly recognized by the listener were divided by the total number of presented words to find the per-participant intelligibility scores. More details about the listening test can be found in [1].

HA SPEECH QUALITY DATA SETS

Two speech quality data sets collected with HA users were used in the experiments described herein. The first database explores the effects of frequency lowering, an amplification strategy for HI listeners with severe to profound high frequency sensorineural hearing loss that has gained renewed attention recently. Nonlinear frequency compression (NFC) is a particular type of frequency lowering algorithm, wherein the input spectral content beyond a cutoff frequency (CF) is compressed by a factor determined by the compression ratio (CR) before further processing by the HA. Thus, NFC moves high frequency energy to lower frequency regions (where there is better residual hearing acuity) increasing the chances of audibility and potential benefit. We refer the interested reader to [30] and the references therein for more details about the database and NFC processing.

The speech material presented to the listeners consisted of IEEE sentences, spoken by two males and two females and recorded through HAs with different NFC strategies; more specifically: 1) $\mathrm{CF} = 4\,\mathrm{kHz}$ and $\mathrm{CR} = 2{:}1$; 2) $\mathrm{CF} = 2\,\mathrm{kHz},\ \mathrm{CR} = 2{:}1$; 3) $\mathrm{CF} = 3\,\mathrm{kHz},\ \mathrm{CR} = 2{:}1$; 4) $\mathrm{CF} = 3\,\mathrm{kHz},\ \mathrm{CR} = 6{:}1$; and 5) $\mathrm{CF} = 3\,\mathrm{kHz},\ \mathrm{CR} = 10{:}1$. In addition, two "anchor" stimuli were created for each sentence: peak clipping at 25% of maximum signal amplitude and lowpass filtering at 2 kHz. In this study, the anchor conditions are not used during metric performance comparison to place emphasis solely on the effects of NFC. As such, of the available 32 stimuli [4 speakers × (5 NFC conditions + 2 anchors + 1 clean reference)], only 24 are used in the analysis presented in the section "Experimental Results."

Quality ratings of this database were obtained with 11 HI listeners with severe to profound hearing loss. Each participant was fitted with a Phonak Savia behind-the-ear (BTE) HA and seated in a double-walled sound booth in front of a speaker and a computer monitor. Ratings of speech quality were obtained using the [20–100]

Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) quality scale, with "20" referring to poor quality and "100" representing excellent quality. Participants selected and listened to the reference and test stimuli and then indicated their quality judgments by adjusting the corresponding sliders on the computer screen. Custom HA recordings were obtained for the purpose of objective speech quality prediction. To this end, the Phonak Savia BTE HA was programmed to match the amplification targets for each participant and was subsequently connected to a 2-cc coupler and placed inside a portable anechoic HA test box. The 32 stimuli within the database were then played back individually through the loudspeaker in the test box, and the resulting HA output was stored in a .wav file with 16-kHz sample rate and 16-bit resolution.

In the second database, the impact of HA speech enhancement on perceived speech quality was investigated in noise-only, reverberation-only, and noise-plus-reverberation listening conditions. Full details about the data set can be found in [28]. Twenty-two adult HA users (average age of 71 years) with moderate to severe sensorineural hearing loss profiles were recruited to participate in the subjective quality experiments. Each of the participants was fitted bilaterally with the Unitron experimental BTE HA and seated at the center of a loudspeaker array, first in a double-walled sound booth (RT60 $= 0.1$ s) and then in a reverberant chamber (RT60 $= 0.9$ s). In each of these rooms, sentences spoken by a male talker were played from a speaker at $0°$ azimuth and multitalker babble or speech-shaped noise at 0 or 5 dB SNR was played from speakers at 0, 90, 180, and 270° azimuth.

Participants listened to the degraded stimuli four times, each time with a different HA setting: omnidirectional microphone, adaptive directional microphone, partial strength signal enhancement (directionality, noise reduction, and speech enhancement algorithms operating below their maximum strengths), and full strength signal enhancement (all enhancement algorithms operating at maximum strength). Within each condition, subjects rated their perceived quality for each stimulus using the MUSHRA quality scale. Once again, a customized set of HA recordings was obtained to enable objective speech quality predictions. To this end, the bilateral HAs were programmed to match the amplification requirements for each HI participant and were then placed on a Bruel and Kjaer head and torso simulator (HATS). The HATS was then positioned in the center of the loudspeaker array in each of the two room environments. The same stimuli used in subjective speech quality experiments were played and the ensuing HA outputs were stored in .wav files with 16-kHz sample rate and 16-bit resolution. In the analysis described in the section "Experimental Results," the objective metrics were computed separately for the left and right channels (using the listeners' left and right audiograms, respectively) and then averaged into a final score that would be compared against the subjective ratings using the performance criteria described next. Moreover, all databases were also downsampled to 8 kHz, such that ITU-T P.563 could also be tested.

PERFORMANCE CRITERIA

To assess the performance of the tested algorithms, four performance criteria were used. As suggested in the literature, performance values are reported on a per-condition basis, where condition-averaged objective and subjective intelligibility/quality ratings are used to reduce intra- and intersubject variability [2]. First, linear relationships between predicted quality/intelligibility scores and subjective ratings are quantified via a Pearson correlation (ρ). Second, the ranking capability of the objective metrics is characterized by the Spearman rank correlation (ρ_{spear}), which is computed in a manner similar to ρ but with the original data values replaced by their ranks. These two measures together can provide insight into the need for a nonlinear monotonic mapping between the objective metric scale and the subjective rating scale. Here, a sigmoidal mapping function is used and once the objective values are mapped, a new Pearson correlation (termed ρ_{sig}) is computed and used as the third performance criteria. The sigmoid mapping is given by:

$$Y = \frac{1}{1+e^{-(\alpha_1 X - \alpha_2)}} \times 100\%, \tag{4}$$

where α_1 and α_2 are the fitting parameters, X represents the objective metric, and Y the mapped intelligibility/quality score.

Finally, the so-called epsilon insensitive root-mean-square estimation error (ε-RMSE) is used. This ε-RMSE measure differs from the conventional one as it considers only differences related to an epsilon-wide band around the target (subjective) quality/intelligibility value, thus taking the uncertainty of the subjective ratings into account. As proposed by ITU-T, epsilon can be defined as the 95% confidence interval (ci_{95}) of the subjective ratings and is given on a per-condition basis [31]. More specifically,

$$ci_{95}(c) = t(0.05,M)\frac{\sigma(c)}{\sqrt{M}}, \tag{5}$$

where c indexes a condition type, M corresponds to the total number of conditions, σ to the standard deviation of the per-condition subjective scores, and $t(0.05,M)$ to the t-value computed at a 0.05 significance level. As such, the per-condition ε-RMSE (c) is given by:

$$\varepsilon\text{-RMSE}(c) = \max(0, |Y(c) - S(c)| - ci_{95}(c)), \tag{6}$$

where $Y(c)$ corresponds to the average sigmoid-mapped intelligibility/quality score for a particular degradation condition c (out of a total of M conditions) and $S(c)$ is the corresponding average subjective score. The final ε-RMSE is then given by:

$$\varepsilon\text{-RMSE} = \sqrt{\frac{1}{M-d}\sum_{c=1}^{M} \varepsilon\text{-RMSE}(c)^2}, \tag{7}$$

where the degree of freedom d is set to "2" for the sigmoidal mapping function. An ideal objective metric will possess ρ_{sig} close to unity and an ε-RMSE close to zero.

When comparing the performance criteria of two or more metrics, it is important to characterize the statistical significance of the difference between them. For correlation-based criteria, a Fisher transformation z-test can be used; here, a significance level of 0.05 was used. For the ε-RMSE criterion, the following statistical significance test was used, as suggested by ITU-T [31]:

[TABLE 1] PER-CONDITION PERFORMANCE CRITERIA FOR THE CI INTELLIGIBILITY DATABASE. THE NUMBERS IN BOLD REPRESENT THE BEST ATTAINED PERFORMANCES (STATISTICALLY INDIFFERENT) AMONG ALL TESTED INTRUSIVE AND NONINTRUSIVE ALGORITHMS.

	ALL				NONENHANCED (NOISE/REVERB)				ENHANCED			
METRIC	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE
NCM	0.68	0.74	**0.87**	**9.03**	0.96	0.93	**0.93**	8.41	0.47	0.68	0.77	10.33
STOI	0.81	0.76	**0.89**	**7.05**	0.97	0.96	**0.97**	**0.6**	0.66	0.69	**0.92**	**3.82**
PESQ	–0.09	0.01	–0.02	26.85	–0.25	0.4	0.14	26.14	–0.09	0.21	–0.02	23.89
PEMO-Q	0.67	0.53	0.68	15.68	0.72	0.8	0.69	15.67	0.38	0.53	0.44	13.52
P.563	0.05	0.38	0.33	23.59	0.76	0.6	0.78	11.77	–0.79	0	–0.43	25.23
ModA	0.78	0.59	0.78	16.88	0.82	0.76	0.8	13.59	–0.13	–0.17	–0.07	18.42
SRMR	0.49	0.53	0.68	18.41	0.93	0.89	0.92	9.6	–0.35	–0.03	–0.37	23.16
SRMR-CI	0.86	0.77	**0.93**	**5.67**	0.98	0.98	**0.98**	**2.06**	0.65	0.5	**0.88**	**4.65**

$$T_{i,j} = \max\left(0, \frac{\varepsilon\text{-RMSE}_i^2}{\varepsilon\text{-RMSE}_j^2} - F(0.05, M, M)\right), \tag{8}$$

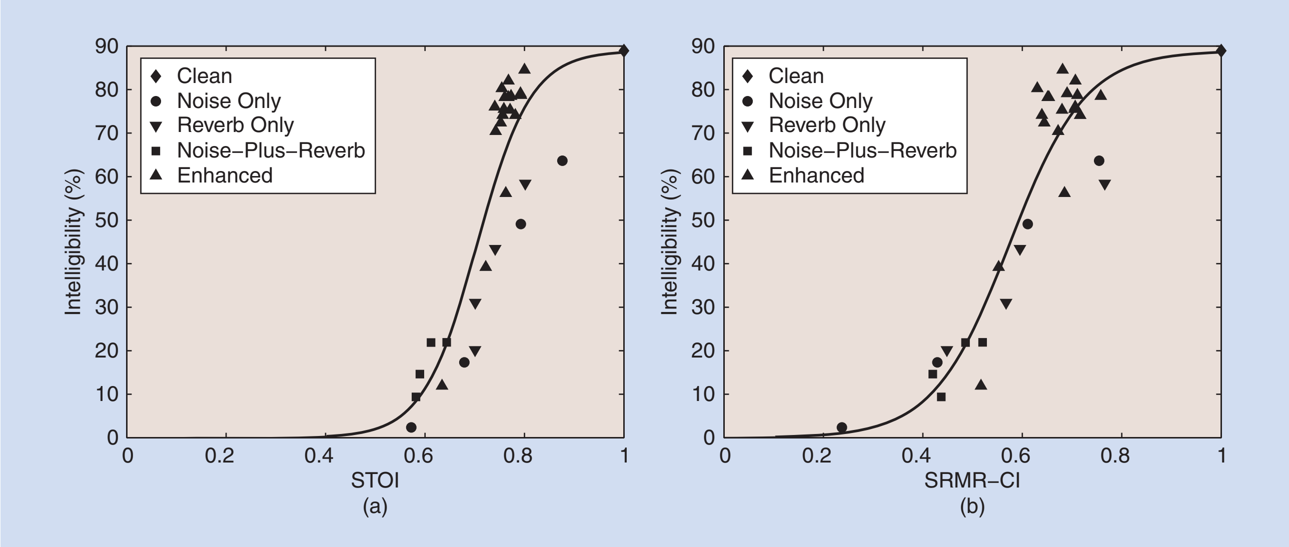
where $F(0.05, M, M)$ corresponds to the F-value computed at a 0.05 significance level. $T_{i,j} = 0$ indicates that metrics i and j achieved statistically equivalent ε-RMSEs, whereas a $T_{i,j} > 0$ indicates that metric i is statistically significant worse than j.

EXPERIMENTAL RESULTS

Table 1 presents the results obtained with four intrusive and four nonintrusive measures on the CI intelligibility database. Note that results for HASQI, HASPI, PEMO-Q-HI, and SRMR-HA have been omitted from the table, as they rely on the impaired listener's audiogram, which is not readily available from the CI participants. As can be seen from the table, the STOI and SRMR-CI measures achieved the highest ρ_{sig} and lowest ε-RMSE among the tested intrusive and nonintrusive metrics, respectively. The scatter plots in Figure 1(a) and (b) depict the subjective versus objective scores obtained for these two metrics, respectively, along with their fitted sigmoidal curves.

Table 2, in turn, presents the results obtained with seven intrusive and four nonintrusive measures on the HA nonlinear frequency compression quality database. Note that the results for SRMR-CI have been omitted from the table as they rely on filter bank information from CI devices. As observed, the PEMO-Q-HI metric achieved the best ρ_{sig} and ε-RMSE of the intrusive metrics, followed closely by the STOI metric (and the HASQI, in terms of ρ_{sig}). For the nonintrusive metrics, all tested measures performed poorly, with ModA achieving somewhat better performance. The scatter plots in Figure 2(a) and (b) depict the subjective versus objective scores obtained for the PEMO-Q-HI and ModA metrics, respectively, along with their fitted sigmoidal curves.

Finally, Table 3 presents the results obtained with seven intrusive and four nonintrusive metrics on the noisy, reverberant, and enhanced HA quality database. As in Table 2, SRMR-CI is omitted as it was developed for CI users and not HA. As can be seen, in the nonenhanced condition, all intrusive measures achieved similar ρ_{sig} values with PESQ achieving the lowest ε-RMSE, followed closely by STOI. For the enhanced condition, HASPI achieved the highest ρ_{sig}, but STOI, PESQ, and PEMO-Q-HI achieved lower ε-RMSE (over three times lower). For the nonintrusive metrics, ModA outperformed all others across both the enhanced and nonenhanced conditions. The scatter plots in Figure 3(a) and (b) depict the subjective versus objective scores obtained for the PESQ and ModA metrics, respectively, along with their fitted sigmoidal curves.



[FIG1] Scatterplots of subjective intelligibility versus objective scores for condition-averaged data points obtained from the (a) STOI and (b) SRMR-CI metrics for the CI intelligibility database.

DISCUSSION

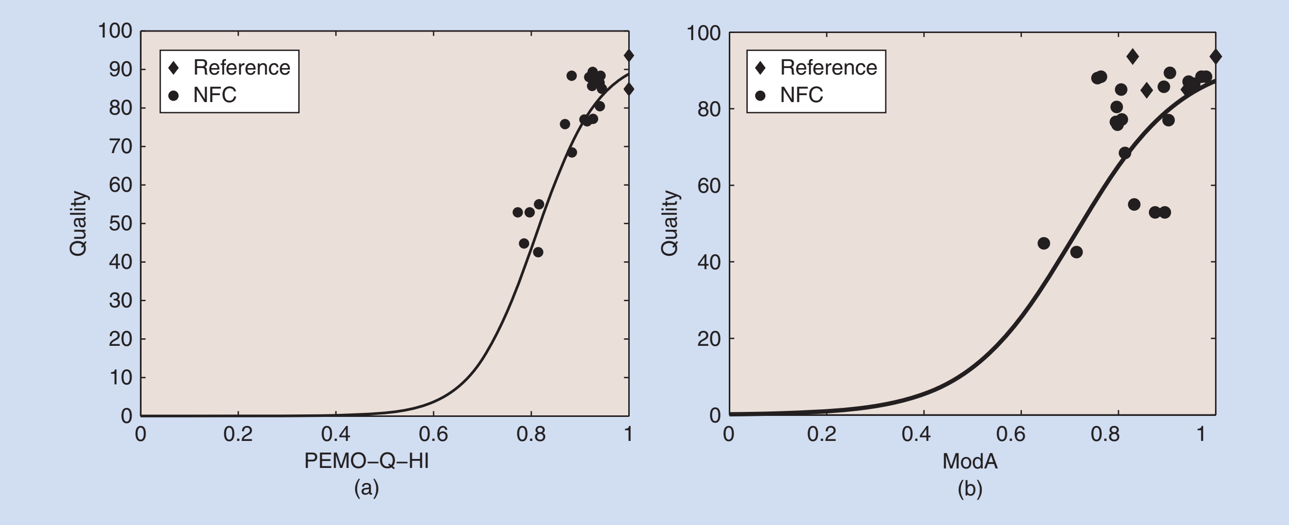
Table 4 summarizes the recommendations for metric usage based on distortion condition type (i.e., overall, nonenhanced, enhanced, NFC), assistive device (CI, HA), and the availability or unavailability of a reference signal (intrusive or nonintrusive). The recommended metrics include those that attained the highest ρ_{sig} and lowest ε-RMSE, shown in bold in the table, as well as all others which attained insignificantly different ρ_{sig} and ε-RMSE levels. A more detailed discussion is given next.

CI: NOISY AND ENHANCED CONDITIONS

For users of CI devices the STOI metric outperformed all other intrusive measures, thus corroborating the usefulness of the measure as a channel selection criteria for CI processing [9] (see Table 4). This was true for both nonenhanced and speech-enhanced conditions. The NCM metric, on the other hand, despite having similar processing stages with STOI and achieving

[TABLE 2] PER-CONDITION PERFORMANCE CRITERIA FOR THE HA NONLINEAR FREQUENCY COMPRESSION QUALITY DATABASE. THE NUMBERS IN BOLD REPRESENT THE BEST ATTAINED PERFORMANCES (STATISTICALLY INDIFFERENT) AMONG ALL TESTED INTRUSIVE AND NONINTRUSIVE ALGORITHMS.

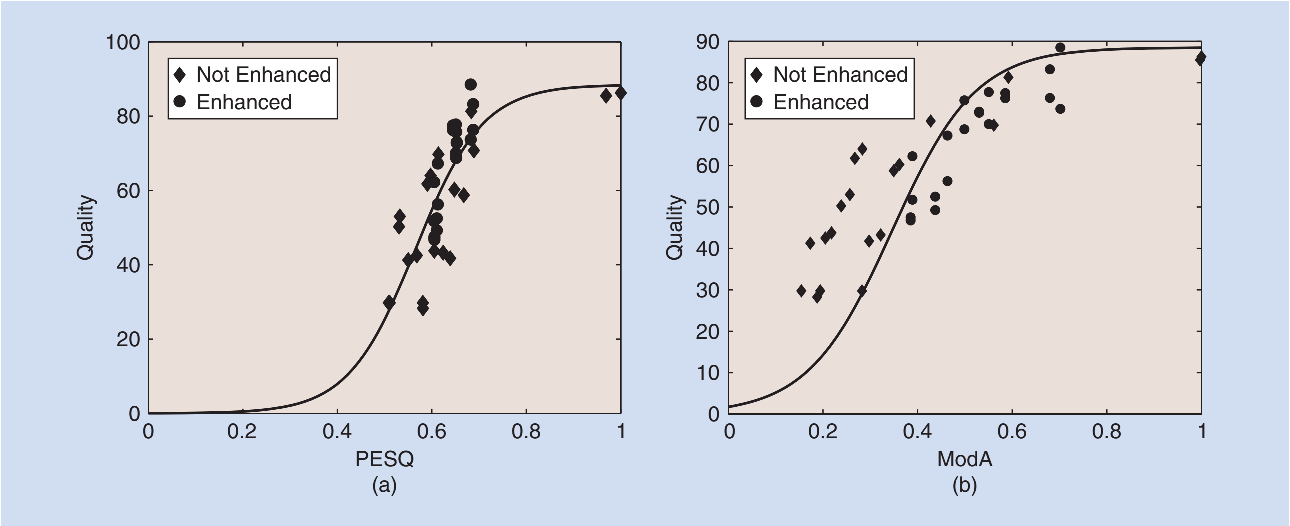
METRIC	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE
NCM	0.67	0.67	**0.89**	7.46
STOI	0.77	0.67	**0.92**	**2.24**
PESQ	0.62	0.56	0.79	**5.73**
HASQI	0.71	0.71	**0.93**	7.67
HASPI	0.83	0.72	0.81	9.9
PEMO-Q	0.67	0.6	0.79	**5.06**
PEMO-Q-HI	0.89	0.71	**0.92**	**1.83**
P.563	–0.27	–0.38	–0.33	23.25
ModA	0.52	0.48	**0.54**	**8.86**
SRMR	0.49	0.59	0.4	17.06
SRMR-HA	0.51	0.58	0.46	14.39



[FIG2] Scatterplots of subjective quality versus objective scores for condition-averaged data points obtained from the (a) PEMO-Q-HI and (b) ModA metrics for the HA nonlinear frequency compression quality database.

[TABLE 3] PER-CONDITION PERFORMANCE CRITERIA FOR THE HA REVERBERATION/ENHANCEMENT QUALITY DATABASE. THE NUMBERS IN BOLD REPRESENT THE BEST ATTAINED PERFORMANCES (STATISTICALLY INDIFFERENT) AMONG ALL TESTED INTRUSIVE AND NONINTRUSIVE ALGORITHMS.

	ALL				NON-ENHANCED				ENHANCED			
METRIC	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE
NCM	0.84	0.84	**0.83**	**6.61**	0.85	0.81	**0.81**	**8.54**	0.77	0.75	0.74	**7.67**
STOI	0.78	0.78	**0.77**	**6.21**	0.81	0.75	**0.78**	**6.26**	0.8	0.79	0.77	**4.11**
PESQ	0.76	0.8	**0.81**	**4.45**	0.76	0.74	**0.78**	**5.07**	0.7	0.68	0.72	**4.59**
HASQI	0.73	0.82	**0.81**	8.02	0.78	0.76	**0.77**	**10.8**	0.75	0.83	**0.86**	**5.6**
HASPI	0.71	0.86	**0.83**	12.95	0.8	0.83	**0.84**	**13.23**	0.71	0.87	**0.9**	15.57
PEMO-Q	0.81	0.88	**0.86**	8.11	0.85	0.8	**0.8**	**10.46**	0.77	0.83	**0.83**	**7.91**
PEMO-Q-HI	0.84	0.85	**0.83**	**6.23**	0.84	0.78	**0.77**	**9.01**	0.84	0.85	**0.84**	**4.18**
P.563	0.39	0.52	0.52	14.95	0.80	0.78	**0.8**	**7.5**	–0.22	–0.15	–0.23	22.38
ModA	0.86	0.9	**0.86**	**10.39**	0.83	0.84	**0.84**	**7.16**	0.82	0.91	**0.9**	**3.85**
SRMR	0.74	0.77	0.74	**8.19**	0.8	0.78	**0.75**	**9.45**	0.39	0.52	0.39	7.64
SRMR-HA	0.79	0.82	**0.77**	**9.9**	0.83	0.81	**0.75**	**10.99**	0.55	0.63	0.53	7.32



[FIG3] Scatterplots of subjective quality versus objective scores for condition-averaged data points obtained from the (a) PESQ and (b) ModA metrics for the HA reverberation/enhancement quality database.

insignificantly different ρ_{sig} values in the nonenhanced case, resulted in significantly higher ε-RMSE values. Such finding shows the importance of short-time processing for CI users. Interestingly, while PESQ and PEMO-Q have been shown to be highly correlated with subjective quality ratings of NH listeners in a number of telephony applications, poor performance was obtained for CI users, particularly under speech enhancement. For the nonintrusive measures, the SRMR-CI measure achieved the best results with performance levels in-line with those obtained with STOI, but with the advantage of not requiring a reference signal. In fact, when both noisy and enhanced conditions were considered overall, the SRMR-CI metric outperformed STOI across all four performance criteria. By incorporating CI processing percepts into the original SRMR measure, significant gains could be observed. Generally, the findings observed here resonate with those reported in the literature showing the importance of spectral envelopes for CI intelligibility.

[TABLE 4] A SUMMARY OF RECOMMENDED OBJECTIVE METRICS FOR DIFFERENT CONDITIONS. THE METRICS IN BOLD REPRESENT THOSE THAT ACHIEVED HIGHEST ρ_{sig} AND LOWEST ε-RMSE. THE METRIC SRMR-HA$_{\text{comp}}$ CORRESPONDS TO AN EXPLORATORY MEASURE DESCRIBED IN THE SECTION "HA: NFC CONDITIONS."

	CI		HA	
CONDITION	INTRUSIVE	NON-INTRUSIVE	INTRUSIVE	NONINTRUSIVE
COMBINED	**STOI,** NCM	**SRMR-CI**	**PESQ,** STOI, PEMO-Q-HI, NCM	**ModA,** SRMR-HA
NON-ENHANCED	**STOI**	**SRMR-CI**	ALL EXCEPT HASPI (**PEMO-Q**)	ALL (**ModA**)
ENHANCED	**STOI**	**SRMR-CI**	**PEMO-Q-HI,** HASQI, PEMO-Q	**ModA**
NFC	–	–	**PEMO-Q-HI,** STOI	SRMR-HA$_{\text{comp}}$

HA: NFC CONDITIONS

For users of HAs with frequency lowering strategies, PEMO-Q-HI and STOI attained insignificantly different ρ_{sig} and ε-RMSE results. The HASQI measure, in turn, resulted in the highest ρ_{sig}, but achieved a significantly higher ε-RMSE than the two aforementioned metrics. This higher error may be a result of the range of conditions used during training of the internal parameter (i.e., noise, linear, and nonlinear terms) mapping available in the HASQI. Notwithstanding, given the burgeoning popularity of such nonlinear frequency compression schemes for HI listeners with severe to profound high frequency sensorineural hearing loss, our results suggest that users have a few reliable intrusive metrics to choose from. On the other hand, the tested nonintrusive measures were not capable of correctly characterizing the perceptual artefacts caused by NFC in HA users. For example, none of the metrics surpassed the correlation threshold of 0.8 established by ITU-T during the competition that resulted in the P.563 Recommendation [20]. These findings motivate the need for more research on the development of innovative nonintrusive quality measures for frequency-lowering strategies. As an exploratory test, the modulation energy thresholding and modulation filter bank compression strategies implemented in the SRMR-CI metric (see the section "SRMR-CI and SRMR-HA") were tested on the original SRMR and SRMR-HA metrics and significant improvements $(p < 0.05)$ could be observed (e.g., $\rho_{\text{sig}} = 0.80$ and ε-RMSE $= 4.68$ with the so-called SRMR-HA$_{\text{comp}}$). In fact, these newly obtained results were in-line with some of the intrusive metrics, such as PEMO-Q, and suggest that further improvements may be obtained with nonintrusive measures.

HA: NOISY AND ENHANCED CONDITIONS

Finally, for HA users in complex listening environments comprising noise, reverberation, and noise-plus-reverberation, it was observed that all intrusive measures achieved insignificantly different ρ_{sig} and ε-RMSE values (with the exception of HASPI, in the latter case). In the scenario where nonlinear speech enhancement (noise suppression and dereverberation) was activated, three measures stood out: HASQI, PEMO-Q, and PEMO-Q-HI. Interestingly, for the nonenhanced and enhanced cases, HASPI, a metric tailored for intelligibility prediction, outperformed HASQI (its quality predictor counterpart) and all other metrics in terms of ρ_{sig}. Such findings resonate with what was mentioned in the section "HA: NFC Conditions" that alternate mappings of HASQI's internal parameters could be devised to reduce ε-RMSE. For nonintrusive measures, in turn, it was found that all tested metrics achieved insignificantly different ρ_{sig} values in the noisy condition, with ModA achieving the highest ρ_{sig} and lowest ε-RMSE. In the enhanced conditions, on the other hand, only ModA achieved levels above ITU-T's "acceptability threshold." Interestingly, in the nonenhanced conditions (i.e., noise-alone, reverberation-alone, and noise-plus-reverberation) ITU-T P.563 achieved reliable results in line with those obtained with SRMR-HA and ModA. With speech enhancement enabled, however, both P.563 and SRMR-HA performances decreased to unacceptable levels, thus suggesting that these two metrics are not capable of detecting and quantifying the effects of speech enhancement artefacts on perceived quality. These findings motivate the need for more research on the development of innovative nonintrusive quality measures for HA devices with nonlinear speech enhancement.

CONCLUSIONS

This article has provided a comprehensive review of 12 existing objective quality and intelligibility prediction algorithms that have been developed for NH and HI listeners who are users of assistive listening devices, such as HAs and CIs. The algorithms were tested on three common subjectively rated speech data sets: one with subjective ratings collected from CI users in noisy and reverberant environments, one from HA users in noisy and reverberant environments with and without speech enhancement, and one from HA users with NFC. The recommended metrics to be used under each condition (nonenhanced, enhanced, NFC) were tabulated for the two different assistive devices. In summary, for CI devices, two measures stood out: STOI (intrusive) and SRMR-CI (nonintrusive). For HA with NFC, several intrusive measures attained comparable results, including the recently proposed PEMO-Q-HI. None of the tested nonintrusive measures, on the other hand, achieved acceptable results, thus leading us to explore the development of a new metric called SRMR-HA$_{\text{comp}}$. Finally, for HA with speech enhancement enabled, the HASQI and PEMO-Q-HI intrusive measures stood out alongside ModA, a recently proposed nonintrusive measure. It is hoped that these insights will be useful not only for those in the assistive listening device research and development community but also clinicians, audiologists, and patients who wish to quickly gauge the performance of different devices across different practical environmental conditions.

ACKNOWLEDGMENTS

Tiago H. Falk and João F. Santos acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de Recherche du Quebec–Nature et Technologies. Vijay Parsa and Susan Scollie acknowledge funding from NSERC, the Oticon Foundation, and Phonak AG; James M. Kates and Kathryn Arehart received funding from GN ReSound and the National Institutes of Health R01 DC012289 (KA); Oldooz Harati received funding from the National Institute of Deafness and other Communication Disorders of the National Institutes of Health R01 DC 007527 (PI: Philipos C. Loizou); and Rainer Huber acknowledges funding from the German Research Foundation (DFG) FOR-1732.

AUTHORS

Tiago H. Falk (falk@emt.inrs.ca) received the Ph.D. degree from Queen's University, Kingston, Canada, in 2009. From 2009 to 2010, he was a postdoctoral fellow at the University of Toronto. In December 2010, he joined INRS-EMT (Montreal) as an assistant professor. His research interests include multimedia quality measurement and enhancement and human–machine interaction. He has published over 130 papers in top-tiered journals and conferences and has won four Best Paper Awards. He is a member of the IEEE Signal Processing Society's Speech and Language Technical Committee, the Sigma Xi Society, and the editorial board of *Journal of the Canadian Acoustical Association* and *Canadian Journal of Electrical and Computer Engineering*. He is a Senior Member of the IEEE.

Vijay Parsa (parsa@nca.uwo.ca) received the Ph.D. degree in biomedical engineering from the University of New Brunswick, Canada, in 1996. He then joined the Hearing Health Care Research Unit at the University of Western Ontario, where he worked on developing speech processing algorithms for audiology and speech language pathology applications. Between 2002 and 2007, he served as the Oticon Foundation chair in acoustic signal processing. He is currently an associate professor jointly appointed across the Faculties of Health Sciences and Engineering. His research interests are in speech signal processing with applications to hearing aids, assistive listening devices, and augmentative communication devices.

João F. Santos (joao.eel@gmail.com) received his B.S. degree in electrical engineering from the Federal University of Santa Catarina, Brazil, in 2011 and his M.Sc. degree in telecommunications from INRS in 2014, where he placed on the dean's list and was awarded the Best M.Sc Thesis Award. He is currently pursuing his Ph.D. degree in telecommunications at the same institute. His main research area is speech signal processing with an emphasis in speech quality assessment and enhancement for hearing aids and cochlear implants. He is also interested in applications of bioinspired algorithms and sparse representations to audio and speech processing.

Kathryn Arehart (kathryn.arehart@colorado.edu) is a professor in the Speech, Language, and Hearing Sciences Department at the University of Colorado at Boulder. Her laboratory's research focuses on understanding auditory perception and the impact hearing loss has on listening in complex auditory environments. Current projects include the study of individual factors (cognition, hearing loss, auditory processing) that affect the ability of older adults to successfully use advanced hearing-aid signal processing

strategies and the evaluation of signal processing algorithms with the goal of improving speech intelligibility and sound quality. She teaches courses in hearing science and audiology and is a certified clinical audiologist.

Oldooz Hazrati (oldooz.hazrati@gmail.com) received the B.S.E.E. and M.S.E.E. degrees from Amirkabir University of Technology (Tehran Polytechnic), in 2005 and 2008, respectively. She received her Ph.D. degree in electrical engineering from the University of Texas at Dallas (UTD) in 2012. Since January 2013, she has been a research associate with the Cochlear Implant and Speech Processing laboratories at UTD. Her primary research interests include signal processing for cochlear implants, speech dereverberation, and noise reduction. She has authored/coauthored 27 journal articles and conference papers in the field of signal processing for cochlear implants.

Rainer Huber (rainer.huber@hoertech.de) received the diploma and Ph.D. degrees in physics from the Universität Oldenburg, Germany, in 1998 and 2003, respectively. From 2001 to 2005, he was a research associate at the Medical Physics section at the Universität Oldenburg. Since 2005, he has been with HörTech (National Center of Competence for Hearing Aid System Technology) in Oldenburg, where he coleads the research and development section. His own research is concerned with development of objective sound quality models for normal hearing and hearing impaired listeners.

James M. Kates (James.Kates@colorado.edu) received B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT) in 1971 and the professional degree of electrical engineer from MIT in 1972. He retired in 2012 from hearing-aid manufacturer GN ReSound, where he held the position of research fellow. He is now a professor of hearing engineering research practice in the Department of Speech, Language, and Hearing Sciences at the University of Colorado at Boulder. His research interest is signal processing for hearing aids with a focus on predicting speech intelligibility and speech and music quality. He is a Senior Member of the IEEE, a fellow of the Acoustical Society of America, and a fellow of the Audio Engineering Society.

Susan Scollie (scollie@nca.uwo.ca) is an associate professor at the National Centre for Audiology, University of Western Ontario, in London, Canada. With colleagues, she developed version 5.0 of the DSL method for hearing aid fitting. Her current research focuses on the evaluation of DSL5, frequency compression signal processing, and outcomes for infants and children who use hearing aids.

REFERENCES

[1] O. Hazrati and P. C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *Int. J. Audiol.*, vol. 51, no. 6, pp. 437–443, June 2012.

[2] S. Moller, W. Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Mag.*, vol. 28, no. 6, pp. 18–28, 2011.

[3] J. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, "Performance comparison of intrusive objective speech intelligibility and quality metrics for cochlear implant users," in *Proc. InterSpeech*, 2012, vol. 1, pp. 1724–1727.

[4] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear Hear.*, vol. 32, no. 3, pp. 331–338, 2011.

[5] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[6] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1703–1716, 1996.

[7] ANSI S3.5-1997, "Methods for the calculation of the speech intelligibility index," ANSI, Tech. Rep., 1997.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] C. H. Taal, R. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in cochlear implants based on an intelligibility metric," in *Proc. 20th IEEE European Signal Processing Conf. (EUSIPCO)*, 2012, pp. 504–508.

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

[11] ITU-T Rec. P.862.2, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunications Union, Geneva, Switzerland, Nov. 2007.

[12] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381, 2010.

[13] J. Kates, "An auditory model for intelligibility and quality predictions," in *Proc. Meetings Acoustics*, vol. 19, no. 050184, 2013, pp. 1–9.

[14] J. Kates and K. Arehart, "The hearing aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov./Dec. 2014.

[15] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117, 2014.

[16] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, 2006.

[17] S. Goetze, E. Albertin, M. Kallinger, A. Mertins, and K.-D. Kammeyer, "Quality assessment for listening-room compensation algorithms," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, 2010, pp. 2450–2453.

[18] R. Huber, V. Parsa, and S. Scollie, "Predicting the perceived sound quality of frequency-compressed speech," *PlosOne*, to be published. DOI: 10.1371/journal.pone.0110260

[19] R. P. Derleth, T. Dau, and B. Kollmeier, "Modeling temporal and compressive properties of the normal and impaired auditory system," *Hear. Res.*, vol. 159, no. 1, pp. 132–149, 2001.

[20] ITU-T P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," International Telecommunications Union, Geneva, Switzerland, May 2004.

[21] S. Cosentino, T. Marquardt, D. McAlpine, and T. Falk, "Towards objective measures of speech intelligibility for cochlear implant users in reverberant environments," in *Proc. Int. Conf. Information Science, Signal Processing and Applications*, 2012, pp. 4710–4713.

[22] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 311–314, May 2013.

[23] T. 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*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[24] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1181–1196, 2000.

[25] T. Falk and W. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Measure.*, vol. 59, no. 4, pp. 978–989, 2010.

[26] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, "Objective speech intelligibility measurement for cochlear implant users in complex listening environments," *Speech Commun.*, vol. 55, no. 7–8, pp. 815 – 824, 2013.

[27] J. Santos and T. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE/ACM Trans. Audio, Speech, Lang. Processing*, vol. 22, no. 12, pp. 2197–2206, Dec. 2014.

[28] D. Suelzle, V. Parsa, and T. H. Falk, "On a reference-free speech quality estimator for hearing aids," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. EL412–EL418, 2013.

[29] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3221–3232, 2011.

[30] V. Parsa, S. Scollie, D. Glista, and A. Seelisch, "Nonlinear frequency compression effects on sound quality ratings of speech and music," *Trends Amplification*, vol. 17, no. 1, pp. 54–68, 2013.

[31] ITU-T, "Statistical evaluation procedure for P.OLQA," TD 12rev1 (WP 2/12), Mar. 2009.

[SP]

[Tuomas Virtanen, Jort F. Gemmeke, Bhiksha Raj, and Paris Smaragdis]

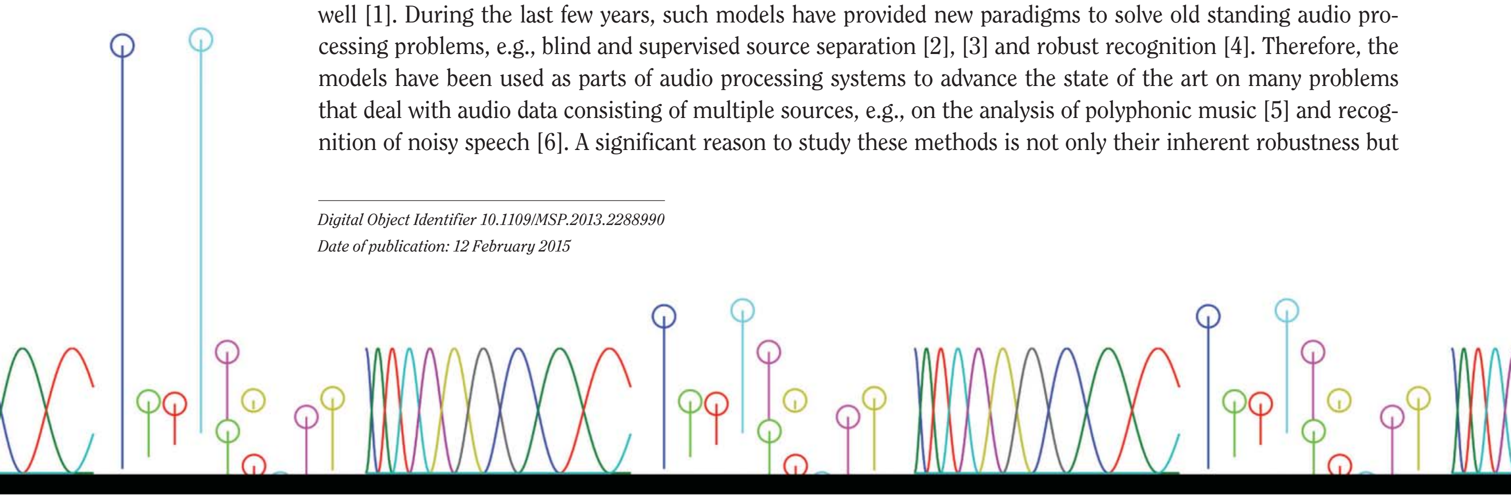
COMPOSITIONAL MODELS for Audio Processing

[Uncovering the structure of sound mixtures]

Many classes of data are composed as constructive combinations of parts. By constructive combination, we mean additive combination that does not result in subtraction or diminishment of any of the parts. We will refer to such data as *compositional data*. Typical examples include population or counts data, where the total count of a population is obtained as the sum of counts of subpopulations. To characterize such data, various mathematical models have been developed in the literature. These models, in conformance with the nature of the data, represent them as nonnegative linear combinations of parts, which themselves are also nonnegative to ensure that such a combination does not result in subtraction or diminishment. We will refer to such models as *compositional models*.

Although the notion of purely constructive composition most obviously applies to nonsignal data such as counts of populations, compositional models have frequently been employed to explain other forms of data as well [1]. During the last few years, such models have provided new paradigms to solve old standing audio processing problems, e.g., blind and supervised source separation [2], [3] and robust recognition [4]. Therefore, the models have been used as parts of audio processing systems to advance the state of the art on many problems that deal with audio data consisting of multiple sources, e.g., on the analysis of polyphonic music [5] and recognition of noisy speech [6]. A significant reason to study these methods is not only their inherent robustness but

Digital Object Identifier 10.1109/MSP.2013.2288990
Date of publication: 12 February 2015

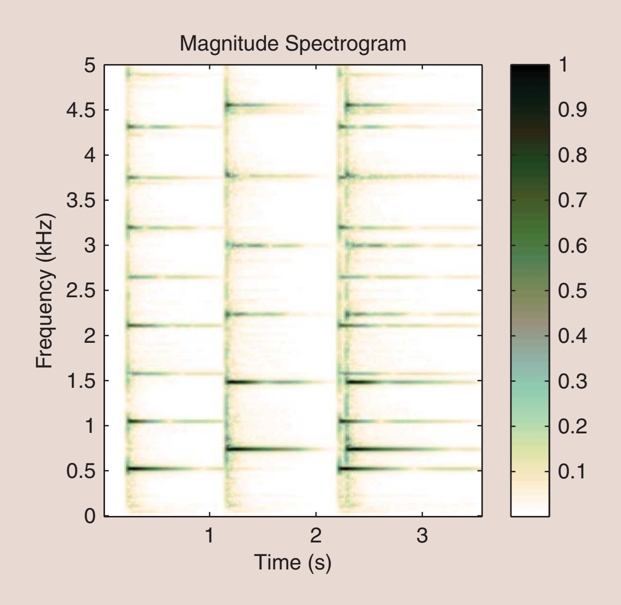


also the flexibility to use them in ways that are nonstandard in audio processing. In this article, we show how they can be powerful tools for processing audio data, providing highly interpretable audio representations and enabling diverse applications such as signal analysis and recognition [4], [7], [8], manipulation and enhancement [9], [10], and coding [11], [12].

THE BASIC PREMISE UNDERLYING THE APPLICATION OF COMPOSITIONAL MODELS TO AUDIO PROCESSING IS THAT SOUND CAN BE VIEWED AS BEING COMPOSITIONAL IN NATURE.

The basic premise underlying the application of compositional models to audio processing is that sound, too, can be viewed as being compositional in nature. The premise has intuitive appeal: sound, as we experience it, does indeed have compositional character. The sounds we hear are usually a medley of component sounds that are all concurrently present. Although a sound may mask others by its greater prominence, the sounds themselves do not generally cancel one another, except in a few cases when it is done intentionally, e.g., in adaptive noise cancellers. Even sounds produced by a single source are often compositions of component sounds from the source, e.g., the sound produced by a machine combines sounds from all of its parts, and music sounds are compositions of notes produced by various instruments.

The compositionality of sound is also evident in time–frequency characterizations of the signal, as illustrated by Figure 1. The figure shows a spectrogram—a visual representation of the magnitude of time–frequency components as a function of time and frequency—of a signal, which comprises two notes played individually at first and then played together. The spectral patterns characteristic of the individual notes are distinctly observable even when they are played together.



[FIG1] A magnitude spectrogram of a simple piano recording. Two notes are played in succession and then again in unison. We can visually identify these notes using their unique harmonic structure.

The compositional framework for sound analysis builds upon these impressions: it characterizes the sounds from any source as a constructive composition of atomic sounds that are characteristic of the source and postulates that the decomposition of the signal into its atomic parts may be achieved through the application of an appropriately constrained compositional model to an appropriate time–frequency representation of the signal. This, in turn, can be used to perform several of the tasks mentioned earlier.

The models themselves may take multiple forms. The nonnegative matrix factorization (NMF) models [3], [13] treat nonnegative time–frequency representations of the signal as matrices, which are decomposed into products of nonnegative component matrices. One of the matrices represents spectral patterns of the atomic parts, and the other represents their activation to the signal over time.

The probabilistic latent component analysis (PLCA) models treat the nonnegative time–frequency representations as histograms drawn from a mixture of multivariate multinomial random variables representing the atomic parts [14]. The two approaches can be shown to be equivalent as well as arithmetically identical under some circumstances [15].

The purpose of this article is to serve as an introduction to the application of compositional models to the analysis of sound. We first demonstrate the limitations of related algorithms that allow for the cancellation of parts and how compositional models can circumvent them, through an example. We then continue with a brief exposition on the type of time–frequency representations where compositional models may naturally be applied.

We subsequently explain the models themselves. Two of the most common formulations of compositional models are based on matrix factorization and PLCA. For brevity, we primarily present the matrix factorization perspective, although we also introduce the PLCA model briefly for completeness.

Within these frameworks, we address various issues, including how a given sound may be decomposed into the contributions of its atomic parts, how the parts themselves may be found, restrictions of the model vis-à-vis the number and nature of these parts and of the decomposition itself, and finally how the solutions to these problems make various applications possible.

WHY CONSTRUCTIVE COMPOSITION?

Before proceeding further, it may be useful to address a question that may already have struck you. Since the models themselves are effectively matrix decompositions, what makes the compositional model with its constraints on purely constructive composition different from other forms of matrix decompositions such as principal component analysis (PCA), independent component analysis (ICA), or other similar methods?

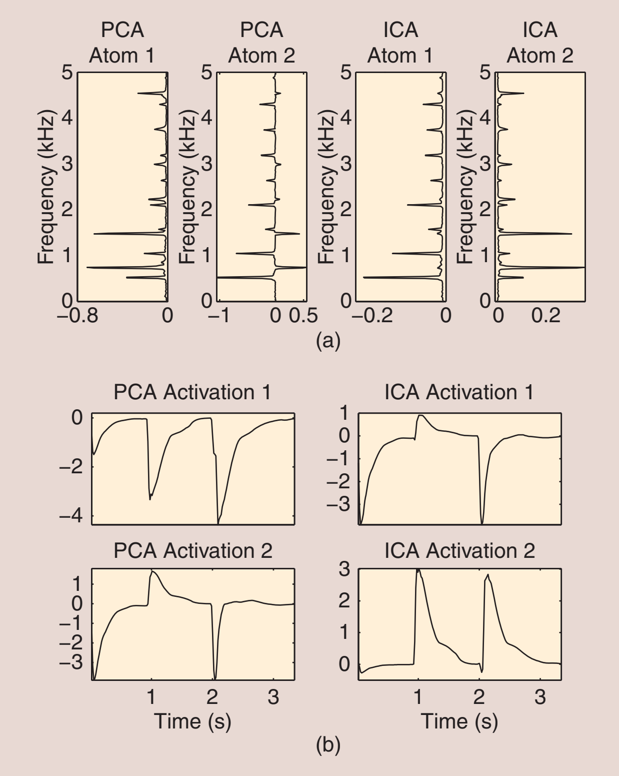
The answer is given as illustrated in Figure 2, which shows the outcome of PCA- and ICA-based decomposition of the spectrogram of Figure 1. Intuitively, the signal is entirely composed of only two notes, and an effective decomposition technique would discover these notes when they were played. PCA and ICA were employed to decompose the spectrogram into two bases and their activations. In both cases, a nearly perfect decomposition is achieved in the sense that the bases, when excited by their corresponding activations, combine to construct the original spectrogram nearly perfectly, reflecting the fact that the signal does indeed comprise only two basic elements (i.e., the two notes). However, an inspection of the actual bases discovered and their activations reveals a problem. PCA [see Figure 2(a)] discovers two bases that, although orthogonal to one another, are actually combinations of the two notes, and their corresponding activations provide no indication of the actual composition of the sound. In this particular example, ICA [see Figure 2(a)] discovers two bases whose activations track the actual activation of the notes in the signal. However, the discovered bases themselves have both negative and positive components, effectively characterizing the atomic units that compose the sound as having negative spectral magnitudes, which has no physical interpretation. More generally, even the degree of conformance to the underlying structure found in this particular example is usually not achieved. The intuitive dissonance is obvious—intuitively, the building blocks of this sound were the notes and both methods have failed to discover these effectively. Although we do not go into this further, the dissonance is more than intuitive; several of the solutions we develop later in the article through compositional models are simply not possible through normal matrix decomposition techniques such as PCA and ICA, which permit both constructive and destructive composition.

In contrast, Figure 3 shows the results obtained by decomposing the spectrogram of Figure 1 with NMF. The nonnegative factorization is observed to successfully uncover both the notes themselves (as defined by their spectra) and their activations. In practice, the discovered atoms will not always have as clearly associative semantics as in this example; for instance, in Figure 3, we have assumed that the correct number of atoms, two, is known a priori, and this is generally not the case. Nevertheless, the atoms that are discovered tend to be consistent spectral structures that compose the signal.

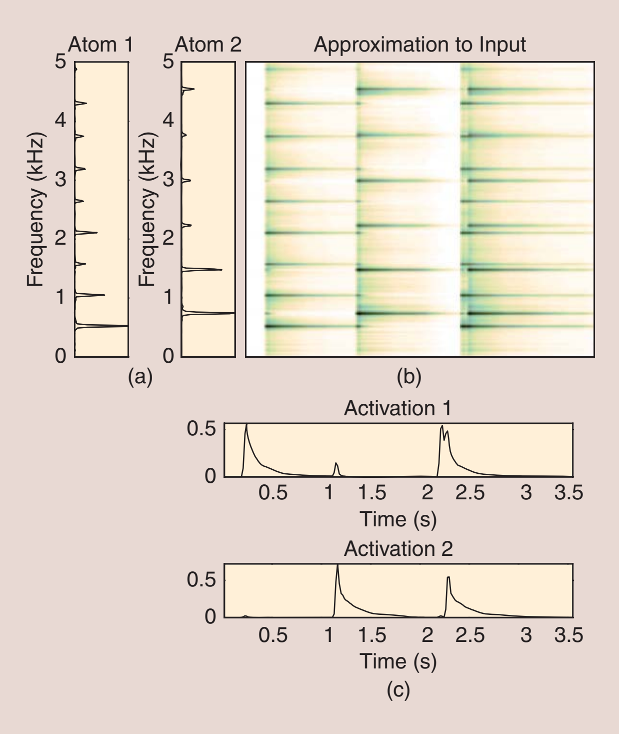
REPRESENTING AUDIO SIGNALS

As noted earlier, the constructive compositionality of sound is evidenced in the distribution of energy in time–frequency characterizations of the signal. This observation has a theoretical basis: the power in any frequency band of the sum of uncorrelated signals is the sum of the powers of the individual signals within that band. We will therefore employ time–frequency characterizations to represent audio signals.

The time–frequency characterizations of the signal are generally obtained through filter bank analysis. Thus, a signal $y[n]$, $n = 1 \cdots N$ is transformed into a matrix of values $Y[t, f]$, $t = 1 \cdots T$, $f = 1 \cdots F$, where T is the number of time frames, F is the number of filters in the filter bank, and $\tau = \lfloor N/T \rfloor$ is the period with which the output of the filter bank is sampled. It is also



[FIG2] The PCA and ICA analyses of the data in Figure 1: (a) the learned PCA and ICA atoms and (b) their corresponding activations. Compared to the learned parameters in Figure 3, we can see that these analyses are not resulting in an intuitive decomposition.



[FIG3] The NMF decomposition of the spectrogram of Figure 1: (a) the discovered atoms and (c) their corresponding activations and (b) is the approximation to the input.

known that the human auditory system effectively acts as a filter bank [16] and that the amplitude of a signal is encoded by the nonnegative number of the firings of neurons [17] (even though neurons encode amplitudes in a nonlinear manner). Thus, the signal representation used in compositional models has some similarities to the representation used in the human auditory system. For simplicity, the specific filter bank analysis we will use is the short-time Fourier transform (STFT), although other forms of time–frequency representations may also be used, some of which we will invoke later in the article. More specifically, we will work with the magnitude of these representations, i.e., with $|Y[t, f]|$.

[THE MAGNITUDE SPECTRA ARE MORE DETERMINISTIC AND CAN BE MODELED WITH A SIMPLE LINEAR MODEL.]

There are three main reasons for using compositional models on magnitudes of time–frequency representations. First, the purely constructive composition required by the compositional framework also necessitates the representations to be nonnegative. Second, the phase spectra (and therefore also the time–domain signals) of natural sounds are rather stochastic and therefore difficult to model, whereas the magnitude spectra are much more deterministic and can be modeled with a simple linear model. Third, the squared magnitude of time–frequency components of the signal represents the power in the various frequency bands. As mentioned earlier, theory dictates that the power in the sum of uncorrelated signals is the sum of the power in the component signals. Hence, the power in a signal composed from uncorrelated atomic units will be the sum of the power in the units. In practice, however, the time–frequency components of the signal are estimated from short-duration windows in which the above relationship does not hold exactly. Also, more than one component may be used to represent a single source, in which case the phases of the components are coherent. It has been empirically observed that the optimal magnitude exponent depends on the task at hand and how the performance is measured [18].

The original signal cannot be recovered directly from the magnitudes of the filter bank output alone; the phase is also required. This presents a problem since we often would like to reconstruct the signal from the output of the compositional analysis. For example, when a compositional model is used to separate out sources from a mixed signal, it is often desired to recover the time–domain signal from the separated time–frequency characterizations, which comprise only magnitude terms. The missing phase terms must be obtained through other means. As will be explained in the section "Source Separation," this can be accomplished, e.g., by using the phase of the mixed signal. Thus, compositional models do not, strictly speaking, perform signal separation but separation using a midlevel representation that allows separating latent parts of a mixture. Nevertheless, the separated midlevel representation, together with mixture phases, allows for reconstruction of signals that are close to source signals before mixing.

An important consideration in deriving time–frequency representations is that of time- and frequency-analysis resolution. Time–frequency representations have a fundamental limitation: the bandwidth, ΔF, of the filters, representing the minimum difference in frequencies that can be resolved is inversely proportional to the time resolution, ΔT, which represents the minimum distance in time between two segments of the signal that can be distinctly resolved. In the case of the STFT, in particular, ΔF is inversely proportional to the length in samples of the analysis window employed. Increasing the length of the analysis window increases the frequency resolution, but decreases the time resolution. Low time resolution analysis may result in the temporal blurring of rapidly changing events, such as those that occur in speech. On the other hand, low frequency resolution can result in the obscuring of frequency structures in signals such as music. Hence, the optimal time/frequency resolution will depend on the type of the signals we wish to analyze. For instance, music processing typically requires longer analysis frames (up to 100 ms), whereas speech processing typically applies shorter windows (tens of milliseconds).

COMPOSITIONAL MODELING OF AUDIO

In the following, we represent the magnitude spectrogram (which we will simply refer to as a *spectrogram* for brevity) as a matrix $\mathbf{Y} \in \mathcal{R}_+^{F \times T}$ comprising magnitudes of time–frequency components $|Y[f, t]|$. Here, $\mathcal{R}_+$ represents the set of nonnegative real numbers. Each column of the matrix $\mathbf{Y}$ is an F-component (magnitude) spectral vector $\mathbf{y}_t \in \mathcal{R}_+^{F \times 1}$, representing the magnitude spectrum of one slice or frame of the signal.

In alternate representation variants that attempt to explicitly capture the temporal dynamics of signals, a single column, $\mathbf{y}_t$, may represent multiple adjacent spectra concatenated into a single vector [4]. In such cases, $\mathbf{Y} \in \mathcal{R}_+^{LF \times T}$, where L is the number of frames that are concatenated together.

The compositional model represents the spectrogram as a nonnegative (purely constructive) linear combination of the contributions of atomic units (which we will simply refer to as *atoms* throughout). In its simplest form, the atomic units themselves are spectral vectors, representing steady-state sounds, and every spectral vector in the spectrogram can be decomposed into a nonnegative linear combination of these atoms. We describe two formalisms to achieve this decomposition.

COMPOSITIONAL MODELS AS MATRIX FACTORIZATION

The matrix factorization approach to compositional modeling treats the problem of decomposing a spectrogram into its atomic units as nonnegative matrix decomposition.

Let $\mathbf{a}_k$ represent any atom, representing spectral vectors in this context. In the matrix factorization approach, we will represent them as column vectors, i.e., $\mathbf{a}_k \in \mathcal{R}_+^{F \times 1}$. The atoms are indexed by $k = 1 \cdots K$, where K is the total number of atoms. Each spectral vector $\mathbf{y}_t$ is composed from all the atoms as $\mathbf{y}_t = \sum_{k=1}^{K} \mathbf{a}_k x_k[t]$, where $x_k[t]$ is the nonnegative activation of the kth atom in frame t. Thus, the spectrogram is modeled as the sum of factors having a fixed spectrum $\mathbf{a}_k$ and time-varying activation $x_k[t]$. Representing the activation of the kth atom to all of the spectral vectors in $\mathbf{Y}$ as a vector $\mathbf{x}_k = [x_k[1]\ x_k[2] \cdots x_k[T]]^\top$, we can represent the overall contribution of $\mathbf{a}_k$ to $\mathbf{Y}$ as $\mathbf{a}_k\mathbf{x}_k^\top$.

We can arrange all of the atoms $\mathbf{a}_k$, $k = 1 \cdots K$, as columns of a matrix $\mathbf{A} \in \mathcal{R}_+^{F\times K}$. We can similarly arrange the activation vectors of the atoms, $\mathbf{x}_k$, $k = 1 \cdots K$ as rows of the a matrix $\mathbf{X} \in \mathcal{R}_+^{K\times T}$. The composition of $\mathbf{Y}$ in terms of the atoms and their activations can now be written as

$$\mathbf{Y} \approx \mathbf{AX}, \tag{1}$$

where all entries are strictly nonnegative.

To decompose the signal into its atomic parts, we must determine the $\mathbf{A}$ and $\mathbf{X}$ that together satisfy (1) most closely. To do so, we define a scalar-valued divergence $D(\mathbf{Y}||\mathbf{AX})$ between the observed spectrogram $\mathbf{Y}$ and the decomposition $\mathbf{AX}$, which characterizes the error between the two. The minimum value of the divergence is zero, which is only reached if the error is zero, i.e., $\mathbf{Y} = \mathbf{AX}$. Typically, the divergence is calculated entry wise, i.e.,

$$D(\mathbf{Y}||\hat{\mathbf{Y}}) = \sum_{f,t} d(y_{f,t}, \hat{y}_{f,t}), \tag{2}$$

where $y_{f,t}$ and $\hat{y}_{f,t}$ are the (f,t)th entries of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, respectively, and $d()$ is the divergence between two scalars.

The optimal values $\mathbf{A}^*$ and $\mathbf{X}^*$ of $\mathbf{A}$ and $\mathbf{X}$ are obtained by minimizing this divergence.

$$\mathbf{A}^*, \mathbf{X}^* = \underset{\mathbf{A},\mathbf{X}}{\operatorname{argmin}}\, \mathrm{D}(\mathbf{Y}||\mathbf{AX}) \quad \mathbf{A} \succeq 0, \mathbf{X} \succeq 0. \tag{3}$$

Here, we assume that both the atoms $\mathbf{A}^*$ and their activations $\mathbf{X}^*$ must be obtained from the decomposition. However, if the atoms $\mathbf{A}$ are prespecified, then decomposition only requires estimation of the activations

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}}\, \mathrm{D}(\mathbf{Y}||\mathbf{AX}) \quad \mathbf{X} \succeq 0. \tag{4}$$

A similar solution may also be defined when $\mathbf{X}$ is specified, and $\mathbf{A}^*$ must be obtained.

The most commonly used divergence in matrix decomposition problems is the squared error: $\mathrm{D}(\mathbf{Y}||\mathbf{AX}) = ||\mathbf{Y} - \mathbf{AX}||_F^2$. However, in the context of audio modeling, other divergence measures have been found more appropriate [3], [19], [20]. Audio signals typically have a large dynamic range—i.e., the energy in high-frequency components can be tens of decibels lower than that in low-frequency components, even when both are perceptually equally important. The magnitude of errors in decomposition tends to be much larger in lower frequencies than in high ones. The squared error emphasizes the larger errors and, as a result, decompositions that minimize the squared error emphasize the accuracy in lower frequencies at the cost of perceptually important higher frequencies. Divergence measures that assign greater emphasis to low-energy components are required for audio.

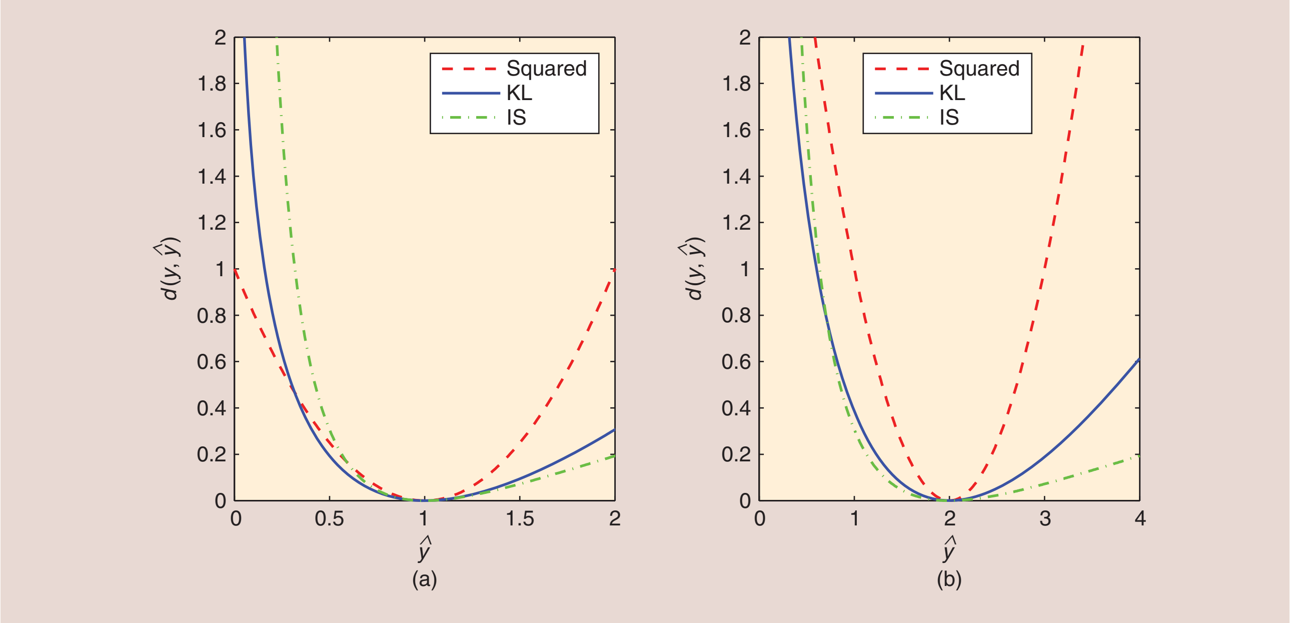
For representing audio, two commonly used divergences are the generalized Kullback–Leibler (KL) divergence

$$d_{\mathrm{KL}}(y, \hat{y}) = y \log(y/\hat{y}) - y + \hat{y} \tag{5}$$

and the Itakura–Saito (IS) divergence

$$d_{\mathrm{IS}}(y, \hat{y}) = y/\hat{y} - \log(y/\hat{y}) - 1. \tag{6}$$

The divergences in (5) and (6) and the squared error $d_{\mathrm{SQ}}(y, \hat{y}) = (y - \hat{y})^2$ are illustrated in Figure 4 for two values of y as the function of $\hat{y}$.



(a)

(b)

[FIG4] An illustration of the typical divergence functions used in NMF. The divergences are calculated for an observation (a) $y = 1$ and (b) $y = 2$ as the function of the model $\hat{y}$. The scale of the input affects the scale of the divergence.

The various divergences scale differently with their arguments. The squared error scales quadratically, meaning that $D_{\mathrm{SQ}}(\alpha \mathbf{Y} || \alpha \mathbf{AX}) = \alpha^2 D_{\mathrm{SQ}}(\mathbf{Y} || \mathbf{AX})$, the IS divergence is scale invariant, i.e., $D_{\mathrm{IS}}(\alpha \mathbf{Y} || \alpha \mathbf{AX}) = D_{\mathrm{IS}}(\mathbf{Y} || \mathbf{AX})$, while the KL divergence scales linearly: $D_{\mathrm{KL}}(\alpha \mathbf{Y} || \alpha \mathbf{AX}) = \alpha D_{\mathrm{KL}}(\mathbf{Y} || \mathbf{AX})$. The relative merits of the divergences may be inferred from this property: the squared error divergence puts undue emphasis on high-energy components, and the IS divergence fails to distinguish between the noise floor and higher-energy speech components. The KL divergence provides a good compromise between the two [3], [19], [20]. A generalization of the divergences in (5) and (6) is the beta divergence [21], which defines a set of divergences that are a function of a parameter β.

The divergences in (5) and (6) (KL, IS, or squared) can be obtained from maximum likelihood estimation of the parameters, when observed data is generated by a specific generative model (Poisson distribution, multiplicative Gamma noise, or additive Gaussian noise) independently at each time–frequency point [13]. Even though some of these models (e.g., the Poisson distribution) do not match well with the distribution of natural sounds, the statistical interpretation allows incorporating a prior distributions for the parameters.

The squared error and KL divergence are convex as the function of $\hat{y}$, and for these, the divergence $D(\mathbf{Y} || \hat{\mathbf{Y}})$ is also convex in $\hat{\mathbf{Y}}$. In this case, the optimization problem of (4) and its counterpart, where $\mathbf{X}$ is specified and $\mathbf{A}$ must be estimated, minimize a convex function, and can be solved by any convex optimization technique.

When $\hat{\mathbf{Y}}$ is itself a product of two matrices, e.g., $\hat{\mathbf{Y}} = \mathbf{AX}$, $D(\mathbf{Y} || \hat{\mathbf{Y}}) = D(\mathbf{Y} || \mathbf{AX})$ becomes biconvex in $\mathbf{A}$ and $\mathbf{X}$. This means that it is not jointly convex in both of these variables, but if either of them is fixed it is convex in the other. Therefore, (3) is biconvex and cannot directly be solved through convex optimization methods. Nevertheless, convex optimization methods may still be employed by alternately estimating one of $\mathbf{A}$ and $\mathbf{X}$, holding the other fixed to its current estimate.

A commonly used solution to estimating nonnegative decompositions is based on the so-called multiplicative updates. The parameters to be estimated are first initialized to random positive values and then iteratively updated by multiplying them with correction terms. The strength of the method stems from the ability of the updates to fulfill the nonnegativity constraints easily: provided that both the previous estimate and the correction term are nonnegative, and the updated term is guaranteed to be nonnegative as well. The multiplicative updates that decrease the KL divergence are given as

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{\frac{\mathbf{Y}}{\mathbf{AX}}\mathbf{X}^{\top}}{\mathbf{1}\mathbf{X}^{\top}} \tag{7}$$

and

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\top}\frac{\mathbf{Y}}{\mathbf{AX}}}{\mathbf{A}^{\top}\mathbf{1}}, \tag{8}$$

where $\mathbf{1}$ is an all-one matrix of the same size as $\mathbf{Y}$, $\otimes$ is an element-wise matrix product, and all the divisions are element wise. It can be easily seen that if $\mathbf{A}$ and $\mathbf{X}$ are nonnegative, the terms that are used to update them are also nonnegative. Thus, the updates obey the nonnegativity constraints. If both $\mathbf{A}$ and $\mathbf{X}$ must be estimated, (7) and (8) must be alternately computed. If one of the two is given and only the other must be estimated, then only the update rule for the appropriate variable need be iterated. For instance, if $\mathbf{A}$ is given, $\mathbf{X}$ can be estimated by iterating (8). In all cases, the KL divergence is guaranteed be nonincreasing under the updates. These multiplicative updates as well as rules for minimizing the squared error were proposed by Lee and Seung [22].

In addition to multiplicative updates, various alternative methods have been proposed, based on, e.g., second-order methods [23], projected gradient [1, pp. 267–268], etc. The methods can also be accelerated by active-set methods [24], [25]. Some divergences, such as the IS divergence, are not convex, and minimizing them requires more carefully designed optimization algorithms than the convex divergences [13].

There also exist divergences that aim at optimizing the perceptual quality of the representation [12], which are useful in audio coding applications. However, in most of the other applications of compositional models, such as source separation and signal analysis, the quality of the representation is affected more by its ability to isolate latent compositional units from a mixture signal, not by the ability to accurately represent the observations. Therefore, simple divergences such as the KL or IS are the most commonly used even in the applications where a mixture is separated into parts for listening purposes.

COMPOSITIONAL MODELS AS PLCA

The PLCA approach to compositional models treats the spectrogram of the signal as a histogram drawn from a mixture multinomial process, where the component multinomials in the mixture represent the atoms that compose the signal [14]. This model is an extension of probabilistic latent semantic indexing and probabilistic latent semantic analysis techniques that have been successfully used, e.g., for topic modeling of speech [26].

The generative model behind PLCA may be explained as follows. A stochastic process draws frequency indices randomly from a collection of multinomial distributions. In each draw, it first selects one of these component multinomials according to some probability distribution, $P(k)$, where k represents the selected multinomial. Subsequently, it draws the frequency, f, from the selected multinomial $P(f|k)$. Thus, the probability that a frequency, f, will be selected in any draw is given by $\sum_k P(k)P(f|k)$. To generate a spectral vector, the process draws frequencies several times. The histogram of the frequencies is the resulting spectral vector.

The mixture multinomial $\sum_k P(k)P(f|k)$ thus represents the distribution underlying a single spectral vector—the vector itself is obtained through several draws from this distribution. When we employ the model to generate an entire spectrogram comprising many spectral vectors, we make an additional assumption: that the component multinomials $P(f|k)$ are characteristic of the source that generates the sound, and represent the atomic units for the source. Hence, the set of component multinomials is the same for all vectors, and the only factor that changes from analysis frame to analysis frame is the probability distribution over k, which specifies how the component multinomials are chosen in any draw. The overall mixture multinomial distribution model for the spectrum of the tth analysis frame in the signal is given by

$$P_t(f) = \sum_{k=1}^{K} P_t(k) P(f \,|\, k), \tag{9}$$

where $P_t(k)$ represents the frame-specific a priori probability of k in the tth frame and $P(f \,|\, k)$ represents the multinomial distribution of f within the kth atom. Even though the formulation of the model is different from NMF, the models are conceptually similar: decomposition of a signal is equated to estimation of the atoms $P(f \,|\, k)$ and their activations $P_t(k)$ to each frame of the signal, given the spectrogram $Y[t, f]$.

The estimation can be performed using the expectation maximization algorithm [27]. The various components of the mixture multinomial distribution of (9) are initialized randomly and re-estimated through iterations of the following equations:

$$P_t(k \,|\, f) = \frac{P_t(k) P(f \,|\, k)}{\sum_{k'=1}^{k} P_t(k') P(f \,|\, k')}$$

$$P(f \,|\, k) = \frac{\sum_{t=1}^{T} P_t(k \,|\, f) Y[t, f]}{\sum_{t=1}^{T} \sum_{f'=1}^{F} P_t(k \,|\, f') Y[t, f]}, \tag{10}$$

$$P_t(k) = \frac{\sum_{f=1}^{F} P_t(k \,|\, f) Y[t, f]}{\sum_{k'=1}^{K} \sum_{f=1}^{F} P_t(k' \,|\, f) Y[t, f]}. \tag{11}$$

The contribution of the kth atom to the overall signal is the expected number of draws from the multinomial for the atom, given the observed spectrum, and is given by

$$Y_k[t, f] = Y[t, f] P_t(k \,|\, f) = Y[t, f] \frac{P_t(k) P(f \,|\, k)}{\sum_{k'=1}^{K} P_t(k') P(f \,|\, k')}.$$

This effectively distributes the intensity of $Y[t, f]$ using the posterior probability of the kth source in point $[t, f]$, and is equivalent to the Wiener-style reconstruction described in the next section.

The rest of this article is presented primarily through the matrix-factorization perspective for brevity. However, many of the NMF extensions described below are also possible within the PLCA framework, often in a manner that is more mathematically intuitive than the matrix-factorization framework. These include, e.g., tensor decompositions [27], convolutive representations, the imposition of temporal constraints [28], joint recognition of mixed signals, and the imputation of missing data [29]. We refer you to the studies mentioned previously for additional details of these models.

UNIQUENESS, REGULARIZATION, AND SPARSITY

The solutions to (3) and (4) are not always unique. We have noted that the divergence $D(\mathbf{Y} \,||\, \mathbf{AX})$ is biconvex in $\mathbf{A}$ and $\mathbf{X}$. As a result, when both $\mathbf{A}$ and $\mathbf{X}$ are to be estimated, multiple solutions may be obtained that result in the same divergence. Specifically, for any $\mathbf{Y} \in \mathcal{R}_+^{F \times T}$, if $(\mathbf{A} \in \mathcal{R}_+^{F \times K}, \mathbf{X} \in \mathcal{R}_+^{K \times T})$ is a solution that minimizes the divergence, then any matrix pair $(\tilde{\mathbf{A}} \in \mathcal{R}_+^{F \times K}, \tilde{\mathbf{X}} \in \mathcal{R}_+^{K \times T})$ such that $\tilde{\mathbf{A}}\tilde{\mathbf{X}} = \mathbf{AX}$ is also a solution. For $K \geq F$, in particular, trivial solutions also become possible. For $K = F$, $\mathbf{Y} = \mathbf{AX}$ can be made exact by simply setting $\mathbf{A} = \mathbf{I}$ and $\mathbf{X} = \mathbf{Y}$. For $K > F$, infinite exact decompositions may be found, for instance, simply by setting the first F columns of $\mathbf{A}$ to the identity matrix; the remaining dictionary atoms become irrelevant (and can be set to anything at all) as an exact decomposition can be obtained by setting their activations to zero.

Even if $\mathbf{A}$ is specified and only $\mathbf{X}$ must be estimated, the solution may not be unique although $D(\mathbf{Y} \,||\, \mathbf{AX})$ is convex in $\mathbf{X}$. This happens particularly when $\mathbf{A}$ is overcomplete, i.e., when $K \geq F$. Any F linearly independent columns of $\mathbf{A}$ can potentially be used to represent an F-dimensional vector with zero error. We can choose F linearly independent atoms from $\mathbf{A} \in \mathcal{R}_+^{F \times K}$ in up to $\binom{K}{F}$ ways, potentially giving us at least that many ways of decomposing any vector in $\mathbf{Y}$ in terms of the atoms in $\mathbf{A}$. If we permit combinations of more than F atoms, the number of minimum-divergence decompositions of a vector in terms of $\mathbf{A}$ can be much greater. The exact conditions for the uniqueness of the decompositions are studied in more detail in [30].

To reduce the ambiguity in the solution, it is customary to impose additional constraints on the decomposition, which is typically done through regularization terms that are added to the divergence to be minimized. Within the NMF framework, this modifies the optimization problem of (3) to

$$\mathbf{A}^*, \mathbf{X}^* = \underset{\mathbf{A},\mathbf{X}}{\operatorname{argmin}} D(\mathbf{Y} \,||\, \mathbf{AX}) + \lambda \boldsymbol{\Phi}(\mathbf{X}) \quad \mathbf{A} \geq 0, \mathbf{X} \geq 0, \tag{12}$$

where $\boldsymbol{\Phi}(\mathbf{X})$ is a differentiable, scalar function of $\mathbf{X}$ whose value decreases as the conformance of $\mathbf{X}$ to the desired constraint increases, and λ is a positive weight that is given for the regularization term.

The introduction of a regularization term as given in (12) can nevertheless still result in trivial solutions. Two solutions, $(\mathbf{A}, \mathbf{X})$ and $(\tilde{\mathbf{A}}, \tilde{\mathbf{X}})$, will result in identical divergence values if $\tilde{\mathbf{A}} = \epsilon^{-1}\mathbf{A}$ and $\tilde{\mathbf{X}} = \epsilon\mathbf{X}$, i.e., $D(\mathbf{Y} \,||\, \mathbf{AX}) = D(\mathbf{Y} \,||\, \tilde{\mathbf{A}}\tilde{\mathbf{X}})$. Structurally, the two solutions are identical since they are merely scaled versions of one another. On the other hand, the regularization terms for the two need not be identical: $\boldsymbol{\Phi}(\mathbf{X}) \neq \boldsymbol{\Phi}(\tilde{\mathbf{X}})$. As a result, the regularization term on the right-hand side of (12) can be minimized by simply scaling $\mathbf{X}$ by appropriate ϵ values and scaling $\mathbf{A}$ up by ϵ^{-1}, without actually modifying the decomposition obtained.

To avoid this problem, it becomes necessary to scale the atoms in $\mathbf{A}$ to have a constant ℓ_2 norm. Typically, this is done by normalizing every atom $\mathbf{a}_i$ in $\mathbf{A}$ such that $||\mathbf{a}_i||_2 = 1$ after every update.

Assuming that all atoms are normalized to unit ℓ_2 norm, for the KL divergence, the update rules from (8) are modified to

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^\top \dfrac{\mathbf{Y}}{\mathbf{AX}}}{\mathbf{A}^\top \mathbf{1} + \lambda \boldsymbol{\Phi}'(\mathbf{X})}, \tag{13}$$

where $\boldsymbol{\Phi}'(\mathbf{X})$ is the matrix derivative of $\boldsymbol{\Phi}(\mathbf{X})$ with respect to $\mathbf{X}$. The update rule for $\mathbf{A}$ remains unchanged, besides the additional requirement that atoms must be normalized after every iteration. There exists also ways to take the normalization into account in the update, which guarantee that the updates and normalization together decrease the value of the cost function [3], [31].

One of the most common constraints is that of sparsity, e.g., [4], [32], and [33]. A vector $\mathbf{x}$ is said to be sparse if the number of nonzero entries in it is fewer than the dimensionality of the vector itself, i.e., $\mathbf{x}_0 < F$. The fewer the nonzero elements, the sparser the vector is said to be. Sparsity is most commonly applied to the activations, i.e., the columns of the activation matrix $\mathbf{X}$. The sparsity constraint is most commonly included by employing the ℓ_1 norm of the activation matrix as a regularizer, i.e., $\boldsymbol{\Phi}(\mathbf{X}) = ||\mathbf{X}||_1 = \sum_k \sum_t |x_k[t]|$. This leads to the following update rule for the activations:

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^\top \dfrac{\mathbf{Y}}{\mathbf{AX}}}{\mathbf{A}^\top \mathbf{1} + \lambda}. \tag{14}$$

Other constraints may be similarly applied by modifying the regularization function $\boldsymbol{\Phi}(\mathbf{X})$ to favor the type of solutions desired. Similarly, regularization functions may be applied on dictionary $\mathbf{A}$, in which case the update rule of $\mathbf{A}$ should be modified. In the context of compositional models for audio, the types of regularizations applied on the dictionary include sparsity [32] and dissimilarity between learned atoms and generic speech templates [34].

It must be noted that despite the introduction of regularization terms, both (3) and (12) are still typically biconvex, and no algorithm is guaranteed to reach the global minimum in practice. Different algorithms and initializations lead to different solutions, and any solution obtained will, at best, be a local optimum. In practice, this can result in some degree of variation in the signal processing outcomes obtained through these decompositions.

The entire discussion in this section also applies to the PLCA decompositions, although the manner in which the regularization terms are applied within the PLCA framework is different. We refer the reader to [14], [27], and [35] for additional discussion of this topic.

SOURCE SEPARATION

Sound source separation refers to the problem of extracting a single or several signals of interest from a mixture containing multiple signals. This operation is central to many signal processing applications because the fundamental algorithms are typically built under the assumption that we operate on a clean target signal with minimal interference. Having the ability to remove unwanted components from a recording can allow us to perform subsequent operations that expect a clean input (e.g., speech recognition or pitch detection). We will predominantly focus on the case where we only observe a single-channel mixture and briefly discuss multichannel approaches later in the article.

The compositional model approach to the separation of signals from single-channel recordings addresses the problem in a rather simple manner. It assumes that any sound source can draw upon a characteristic set of atomic sounds to generate signals. Here, a source can refer to an actual sound source or to some other grouping of acoustic phenomena that should be jointly modeled, such as background noise or even a collection of sound classes that must be distinguished from a target class. A mixture of signals from distinct sources is composed of atoms from the individual sources. Hence, the separation of any particular component signal from a mixture only requires the segregation of the contribution of the atoms from that source from the mixture.

Mathematically, we can explain this as follows. We use the NMF formulation in our explanation. Let matrix $\mathbf{A}_s$ represent the set of atoms employed by the sth source. We will refer to it as a dictionary of atoms for that source. Any spectrogram $\mathbf{Y}_s$ from the sth source is composed from the atoms in the dictionary $\mathbf{A}_s$ as $\mathbf{Y}_s = \mathbf{A}_s\mathbf{X}_s$. A mixed signal $\mathbf{Y}_{\text{mix}}$ combining signals from several sources is given by

$$\mathbf{Y}_{\text{mix}} = \sum_s \mathbf{Y}_s = \sum_s \mathbf{A}_s\mathbf{X}_s. \tag{15}$$

Equation (15) can be written more compactly as follows. Let $\mathbf{A} = [\mathbf{A}_1\mathbf{A}_2\cdots]$ be a matrix composed by stacking the dictionaries for all the sources side by side. Let $\mathbf{X} = [\mathbf{X}_1^\top\mathbf{X}_2^\top\cdots]^\top$ be a matrix composed by stacking the activations for all the sources vertically. We can now express the mixed signal in compact form as

$$\mathbf{Y}_{\text{mix}} = \mathbf{AX}.$$

The contribution of the sth source to $\mathbf{Y}_{\text{mix}}$ is simply $\mathbf{Y}_s = \mathbf{A}_s\mathbf{X}_s$. In unsupervised source separation, both $\mathbf{A}$, $\mathbf{X}$ are estimated from the observation $\mathbf{Y}_{\text{mix}}$, followed by a process that identifies which source each atom is predominantly associated with. In a supervised scenario for separation, the dictionaries, $\mathbf{A}_s$, for each of the sources are known a priori. We address the problem of creating these dictionaries in the next section. Thus, $\mathbf{A}_s$ $\forall s$ are known, and thereby so is $\mathbf{A}$. $\mathbf{X}$ can now be estimated through iterations of (8).

The activations $\mathbf{X}_s^*$ of source s can be extracted from the estimated activation matrix $\mathbf{X}^*$ by selecting the rows corresponding to the atoms from the sth source. The estimated spectrogram for the sth source is then simply computed as

$$\hat{\mathbf{Y}}_s = \mathbf{A}_s\mathbf{X}_s^*. \tag{16}$$

An example of a source separation task using a dictionary representing isolated speech digits and a dictionary representing background noises is shown in Figure 5.

In practice, the decomposition will not be exact and we will only achieve approximate decomposition, i.e., $\mathbf{Y}_{\text{mix}} \approx \mathbf{AX}^*$, and as a consequence, $\mathbf{Y}_{\text{mix}}$ is not fully explained by the decomposition. Hence, the separated signal spectrograms given by (16) will not explain the mixed signal completely.

To be able to account for all the energy in the input signal, we can use an alternative method to extract the contributions of the individual sources. Although the separated signals do not completely explain the mixed signal, we assume that they do nevertheless successfully characterize the relative proportions of the individual signals in the mixture. This leads to the following estimate for the separated signals:

$$\mathbf{Y}_s^* = \mathbf{Y}_{\text{mix}} \otimes \frac{\mathbf{A}_s\mathbf{X}_s^*}{\mathbf{AX}},$$

where the last term is the ratio of the contribution of the sth source to all the sources in each time–frequency point. This filter response is used by the well-known Wiener filter, and the reconstruction is often referred to as the Wiener-style reconstruction.

If we wish to listen to these separated components, we need to convert them back to the time domain. At this point, we only have magnitude spectrogram representations $\mathbf{Y}_s^*$, so we need to find a way to create some phase values to be able to invert them back to a waveform. Although one can use magnitude inversion techniques [36], [37], a simple approach that leads to a reasonable quality is to use the phase of the original mixture. This leads to the following estimate for the separated complex spectrogram, which can be reverted to a time-domain signal:

$$\bar{\mathbf{Y}}_s^* = \bar{\mathbf{Y}}_{\text{mix}} \otimes \frac{\mathbf{A}_s\mathbf{X}_s^*}{\mathbf{AX}},$$

where $\bar{\mathbf{Y}}_s^*$ and $\bar{\mathbf{Y}}_{\text{mix}}$ represent complex spectrograms.

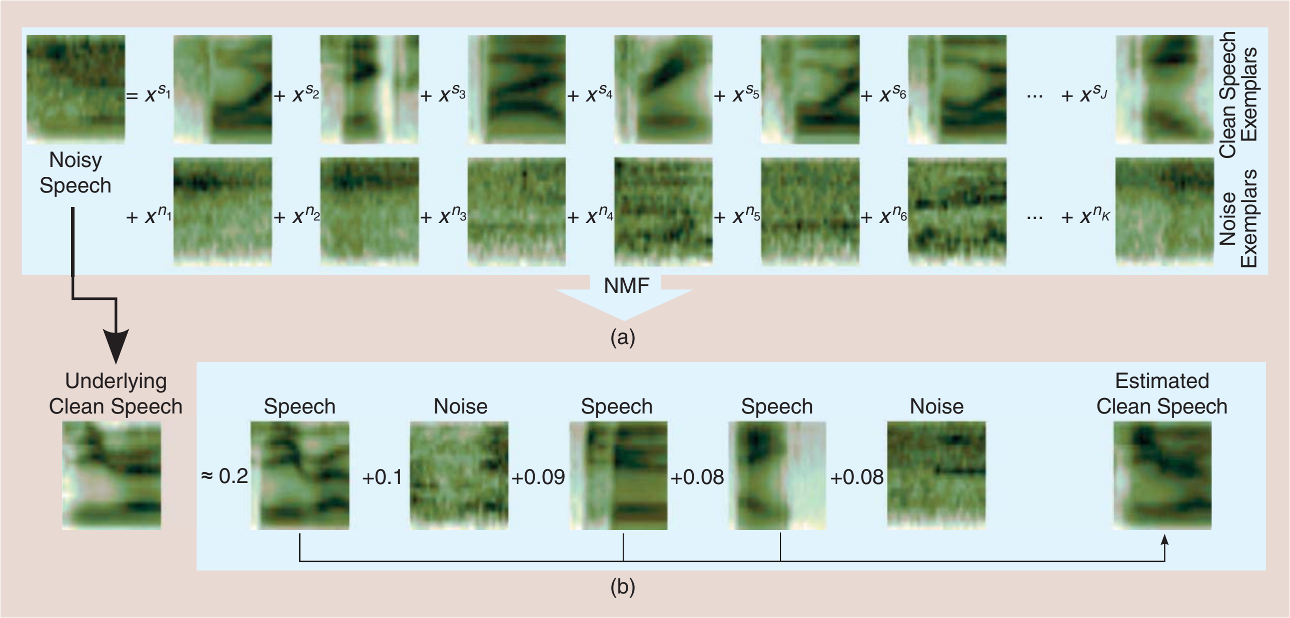
Although we have assumed in this section that the dictionaries for all sources are known, this is not essential. The technique may also be employed if the dictionary for one of the sources is not known. In this case, in addition to estimating the activation matrices, we must also estimate the unknown dictionary. This is done simply by using the same iterative updates as for NMF but with (7) only acting on the atoms reserved for modeling the unknown source.

DICTIONARY CREATION

The key to effective modeling and separation of sources is to have accurate dictionaries of atoms for each of the sources. The basic NMF (3) aims at estimating both the atoms and their activations from mixed data. Contrary to that, in supervised processing, source-specific dictionaries $\mathbf{A}_s$ are obtained in a training stage from a source-specific data set, and combined to form the whole dictionary. The dictionary is then kept fixed, and only the activations are estimated according to (4).

There are two main approaches for dictionary learning: the first attempts to learn dictionary atoms, which jointly describe the training data [38], [39], whereas the second approach uses samples from the training data itself as its dictionary atoms: a sampling-based approach [4], [35]. Good dictionaries have several properties. They should be capable of accurately describing the source and generalize well to unseen data. They should be kept relatively small to reduce computational complexity. They should be discriminative, meaning that sources cannot be well represented using a dictionary of another source. These requirements can be at odds with each other, e.g., because small, accurate dictionaries are often less discriminative. The various approaches for dictionary creation each have their strengths and weaknesses.

Let us denote the training data of source s as $\mathbf{D}_s$, a matrix with as its columns the training samples. The prevailing technique for dictionary learning is to use unsupervised NMF: For each data set, s we write $\mathbf{D}_s \approx \mathbf{A}_s\mathbf{X}_s$ and estimate the parameters using the optimization methods described in the previous sections. The activations $\mathbf{X}_s$ are discarded, and the dictionaries $\mathbf{A}_s$ of each source are concatenated as explained previously. To illustrate this, let us consider the piano and speech sounds described by the magnitude spectrograms in the left plots of Figure 6(a) and (b). We use unsupervised NMF on each individual sound to obtain a 16-atom dictionary, visualized in the plots on the right-hand sides of Figure 6(a) and (b). We can observe that the dictionaries capture



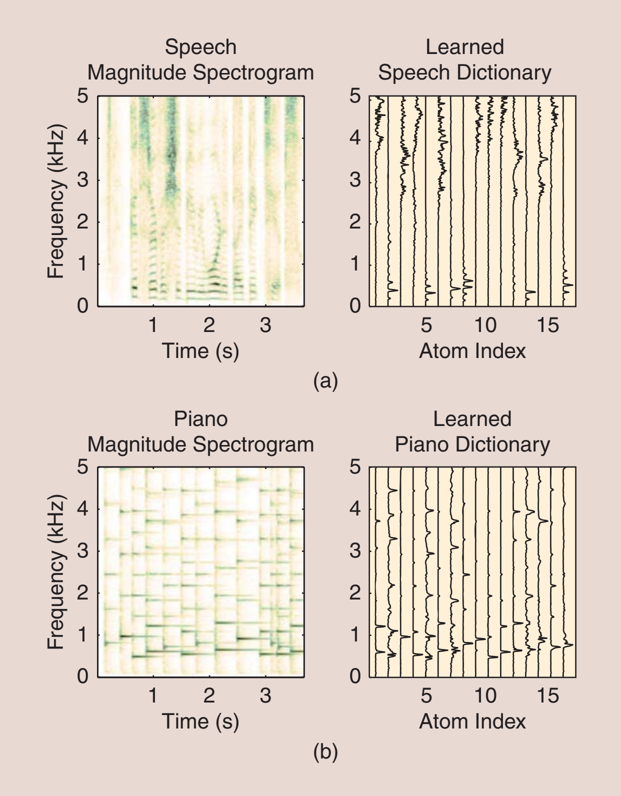
[FIG5] An example of supervised separation of noisy speech. In the top left corner, we display the noisy spectrogram of the isolated word zero corrupted with babble noise. In (a), we display parts of the speech and noise exemplar dictionaries. In (b) the five atoms with the highest weight are shown. The bottom left spectrogram illustrates the underlying clean speech, whereas the bottom right spectrogram shows the clean speech reconstruction.

the spectral character of each sound. The speech dictionary contains some atoms that have the harmonic structure of vowels, and others that contain a lot of broadband energy at the high frequencies, representing consonant sounds. For the piano dictionary, we obtain atoms that have a harmonic structure, each predominantly describing a different note.

An alternative dictionary-learning technique is based on clustering. In clustering, data samples are first grouped based on some measure of similarity, after which a representation of each group (cluster) becomes a dictionary atom. A popular technique is the k-means clustering approach [25]. Another alternative is given by dictionary-learning techniques employed in the field of sparse representations and compressed sensing [39], which aim at finding dictionaries that can sparsely represent a data set. Although most of these methods do not conform to the nonnegativity constraints of the compositional models we discuss in this article, at least one popular method, K-singular value decomposition (K-SVD), which has a nonnegative variant, has been used for the dictionary learning of audio signals [38].

The advantage of dictionary learning is that it typically yields dictionaries that generalize well to unobserved data. The NMF and sparsity-based methods both use the fact that atoms can linearly combine to model the training data, rather than having atoms that each individually need to model an observation as well as possible. This naturally leads to parts-based dictionaries, in which only parts of the spectra contain energy. This in turn leads to small dictionaries and very sparse representations, which may also be more interpretable for some phenomena. When the different sources are highly related, however, this may also be a disadvantage because a parts-based dictionary may no longer discriminative with respect to other dictionaries. The clustering approach typically yields dictionaries that are larger but more discriminative.

DICTIONARY ATOMS ARE ASSUMED TO REPRESENT BASIC ATOMIC SPECTRAL STRUCTURES THAT A SOUND SOURCE MAY PRODUCE.



[FIG6] Learning dictionaries from different sound classes. (a) An input magnitude spectrogram for a speech recording and the dictionary that was extracted from it. (b) A piano recording input and its corresponding dictionary. Note how both dictionaries capture salient spectral features from each sound.

While dictionary learning is a powerful method to create small dictionaries, it can be difficult to train overcomplete dictionaries, in which there are many more atoms than features. A large number of atoms would naturally increase the representation capability of the model, but learning overcomplete dictionaries from data then requires additional constraints such as sparsity and careful tuning, as will be discussed in the next section. As an alternative to learning the dictionaries representing training data, dictionary atoms can also be sampled from the data. Given a training data set $\mathbf{D}_s$, the dictionary $\mathbf{A}_s$ is constructed as a subset of the columns of $\mathbf{D}_s$.

By far, the simplest method is random sampling, where the dictionary is formed by a random subset of columns of $\mathbf{D}_s$. Interestingly, dictionaries obtained with this approach yield comparable and often superior results as more complex dictionary creation schemes [4]. The example in Figure 5 used randomly sampled atoms representing isolated speech digits and background noise.

The sampling methods typically require little tuning and allow for the creation of large, overcomplete dictionaries. A disadvantage is that they may not generalize as well to unseen data, and that smaller dictionaries are often incapable of accurately modeling a source because they disregard the fact that atoms can linearly combine to model an observation.

An alternative approach to dictionary creation, which avoids the need for training data, is to create dictionaries by using prior knowledge of the structure of the signals. For example, in music transcription, harmonic atoms that represent different fundamental frequencies have been successfully used [8]. In the excitation-filter model [5] described later in this article, atoms can describe filter bank responses and excitations. This approach is only used in a small number of specialized applications because, while it yields small dictionaries that generalize well, they are typically not very discriminative.

THE NUMBER OF ATOMS IN THE DICTIONARY

Let us now more carefully consider the issue of the number of atoms in the dictionary. Dictionary atoms are assumed to represent basic atomic spectral structures that a sound source may produce. A source may produce any number of distinct spectral structures. To accommodate all of them, the dictionary must ideally be large. When we attempt to learn large dictionaries, however, we run into a mathematical restriction: K becomes larger than F and, as a result, in the absence of other restrictions, trivial solutions for $\mathbf{A}$ can be obtained as explained earlier. Consequently, a learned dictionary with F or more atoms will generally be trivial and carry little information about the actual signal itself. Even if the dictionary is not learned through the decomposition but specified through other means such as through random draws from the training data, we run into difficulties when we attempt to explain any spectral vector in terms of this dictionary. In the absence of other restrictions, the decomposition of an $F \times 1$ spectral vector in terms of an $F \times K$ dictionary is not unique when $K \geq F$ as explained earlier.

THE SAMPLING METHODS TYPICALLY REQUIRE LITTLE TUNING AND ALLOW FOR THE CREATION OF LARGE, OVERCOMPLETE DICTIONARIES.

To overcome the nonuniqueness, additional constraints must be applied through appropriate regularization terms. The most common constraint that is applied is that of sparsity. Sparsity is most commonly applied to the activations, i.e., to the columns of the activation matrix $\mathbf{X}$. Intuitively, this is equivalent to the claim that although a source may draw from a large dictionary of atoms, any single spectral vector will only include a small number of these. Other commonly applied constraints are group sparsity, promotes sparsity over groups of atoms [40], and temporal continuity, which promotes smooth temporal variation of activations [3].

The number of atoms in the dictionary has a great impact on the decomposition, even when the number of atoms is fewer than F. Ideally, the number atoms should equal the number of latent compositional units within the signal. In certain cases, we might know exactly what this number might be (e.g., when learning a dictionary for a synthetic sound with a discrete number of states), but more often this information is not available and the number of atoms in the dictionary must be determined in other ways. A dictionary with too few elements will be unable to adequately explain all sounds from a given source, whereas one with too many elements may overgeneralize and explain unintended sounds that do not belong to that source as well, rendering it ineffective for most processing purposes. Although, in principle, the Bayesian information criterion can be employed to automatically obtain the optimal dictionary size, it is generally not as useful in this setting [41], and more sophisticated reasoning should be used. Sparsity can be used for automatic estimation of the number of atoms, e.g., by initializing the dictionary with a large number of atoms, enforcing sparsity on the activations, and reducing dictionary size by eliminating all atoms that exhibit consistently low activations [42]. Another approach is to make use of Bayesian formulations that allow for model selection in a natural way. For example, the Markov chain Monte Carlo methodology has been applied to estimate the size of a dictionary [41], [43].

In general, the trend is that larger dictionaries lead to better representations, and consequently superior signal processing, e.g., in terms of the separation quality [25], provided that they are appropriately acquired. The downside of larger dictionaries is, of course, increased computational complexity.

ANALYZING THE SEMANTICS OF SOUND

One of the fundamental goals in audio processing is the extraction of semantics from audio signals with ample applications such as music analysis, speech recognition, speaker identification, multimedia archive access, and audio event detection. The source separation applications described in previous sections are often used as a preprocessing step for conventional machine-learning techniques used in audio analysis, such as Gaussian mixture models (GMMs) and hidden Markov models. The compositional model itself, however, is also a powerful technique to extract meaning from audio signals and mixtures of audio signals.

As an example, let us consider a music transcription task. The goal is to transcribe the score of a music piece, i.e., the pitch and duration of the sounds (notes) that are played. Even when considering a recording in which only a single instrument, such as a piano, is playing, this is a challenging task since multiple notes can be played at once. Moreover, although each note is characterized by a single fundamental frequency, their energy may span the complete harmonic spectrum. These two aspects make music transcription difficult for conventional methods based on sinusoidal modeling and STFT spectrum analysis, in which notes are associated with a single frequency band, or machine-learning methods, which cannot model overlapping notes. An example using NMF is shown in Figure 7.

Information extraction using the compositional model works by associating each atom in the dictionary with metainformation, e.g., class labels indicating notes. With the observation described as a linear combination of atoms, the activation of these atoms then serves directly as evidence for the presence of (multiple) associated class labels. Formally, let us define a label-matrix $\mathbf{L}$, a binary matrix that associates each atom in $\mathbf{A}$ with one or multiple class labels. The dimensions of $\mathbf{L}$ are $Q \times K$, where Q is the total number of classes. A nonzero entry in the qth row of $\mathbf{L}$ indicates those atoms are associated with the label q. A straightforward method for classification is to calculate the label activations as

$$\mathbf{g}_t = \mathbf{L}\mathbf{x}[t], \qquad (17)$$

with $\mathbf{x}[t] = [x_1[t], x_2[t], \ldots, x_K[t]]^\top$ the atom activations of an observation $\mathbf{y}_t$. The entries of the Q-dimensional vector $\mathbf{g}$ are an unscaled score proportional to the presence of class labels in the observation. An example of this procedure is given in Figure 8, where the dictionary atoms of the source separation example of Figure 5 are now associated with word labels.

The formulation (17) is closely related to several other techniques such as k-nearest neighbor (k-NN) classification. When

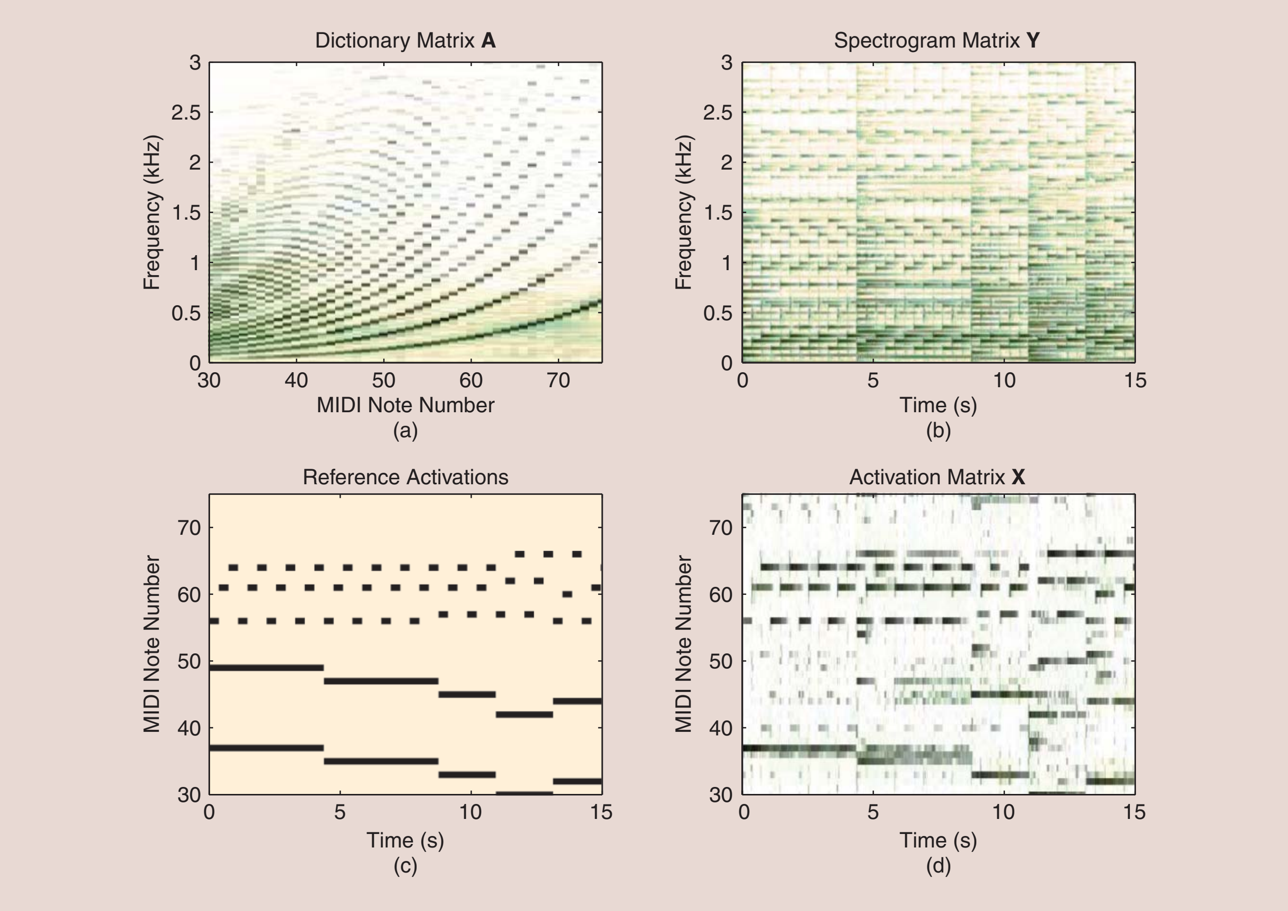
$\mathbf{x}[t]$ is maximally sparse (contains only one nonzero entry), (17) is in fact identical to nearest neighbor classification. For less sparse solutions, the difference is that the compositional model represents an observation as a combination of atoms, whereas k-NN represents an observation as a collection of k atoms that each individually are close to $\mathbf{y}_t$.

In literature, many different types of metainformation exist. In the music transcription example of Figure 7, dictionary atoms were associated with notes. Even in the previous application, source separation, we used metainformation by labeling atoms with a source identity. In speaker identification [44], atoms are associated with speaker identities. In simple speech processing tasks, such as phone classification [45] or word recognition [46], the associated labels are simply the phones or words themselves.

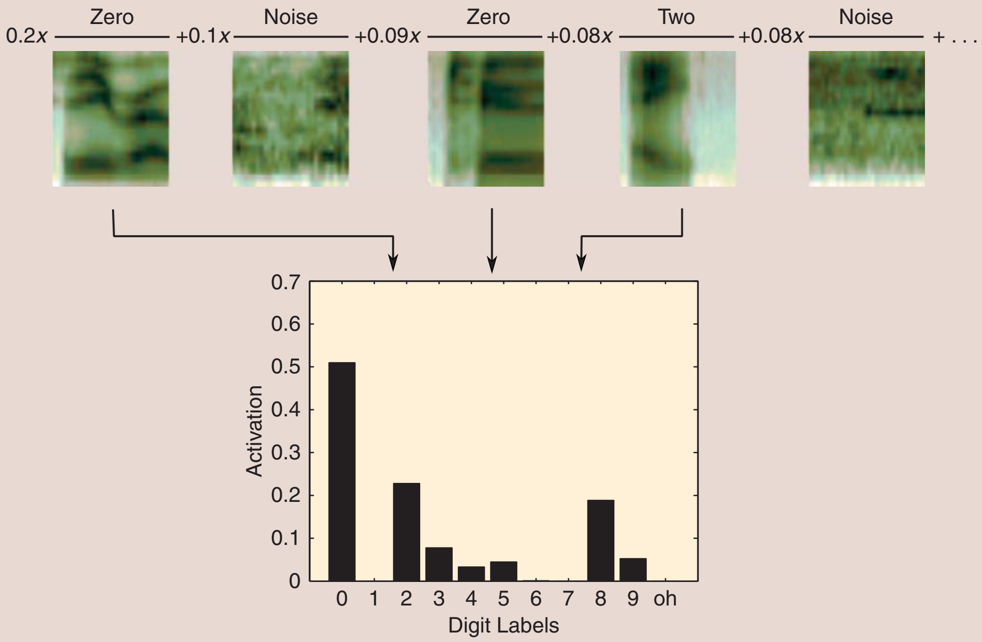
In these examples, the dictionary $\mathbf{A}$ is either constructed or sampled from training data, which makes it straightforward to associate labels to atoms. When the dictionary is learned from data, however, the appropriate mapping from atoms to labels is unclear. In this scenario, the mapping can be learned by first calculating atom activations on training data for which associated labels are known, followed by NMF or multiple regression. In [47], this approach was shown to improve the performance even with a sampled dictionary. Alternatively, we can treat either $\mathbf{g}_t$ or the activations $\mathbf{x}[t]$ as features for a conventional supervised machine-learning technique such as GMMs [48] or a neural network [49].

[INFORMATION EXTRACTION USING THE COMPOSITIONAL MODEL WORDS BY ASSOCIATING EACH ATOM IN THE DICTIONARY WITH METAINFORMATION.]

Another powerful aspect of the compositional model is that dictionary atoms can be as easily associated with other kinds of information, e.g., audio. Consider, for example, a bandwidth extension task [9], [50] where the goal is to estimate a full-spectrum audio signal given a bandwidth-limited audio signal. This is a useful operation to perform since, in many audio transmission cases, high-frequency information is removed to reduce the amount of



[FIG7] A music analysis example where a polyphonic mixture spectrogram (b) is decomposed into a set of note activations (d) using a dictionary consisting of spectra of piano notes (a). Each atom in the dictionary is associated with an MIDI note number. The reference note activations are given in (c). This example is an excerpt from Beethoven's *Moonlight Sonata*. Even though the activations are rather noisy and do not exactly match the reference, the structure of the music is much more clearly visible in the activation plot than in the spectrogram of the mixture signal.



[FIG8] By associating each dictionary atom from Figure 5 with a word label, the linear combination of speech atoms in Figure 5 serves directly as evidence for the underlying word classes. We observe that the word *zero*, underlying the noisy observation of Figure 5, does indeed obtain the highest score.

information to transmit, which negatively impacts intelligibility and the perception of quality. To use the compositional model approach for this task, two dictionaries are first constructed: a bandwidth-limited dictionary $\mathbf{A}$ and a full-bandwidth dictionary $\mathbf{L}$. The atoms in the dictionaries should be coupled, i.e., each atom in $\mathbf{A}$ should represent a band-limited version of the corresponding atom in $\mathbf{L}$. This can be done through training on parallel corpora of full-bandwidth and band-limited signals, or by calculating $\mathbf{L}$ from $\mathbf{A}$, if the details of the band-limitation process are known and can be modeled computationally. We then estimate the atom activations $\mathbf{x}[t]$ using the limited-bandwidth observation $\mathbf{y}_t$ and the limited-bandwidth dictionary $\mathbf{A}$. Finally, direct application of (17) serves as a replacement for the audio reconstruction $\mathbf{A}\mathbf{x}[t]$ and yields a full-bandwidth reconstruction. We illustrate this process in Figure 9. Very similar principles underlay voice conversion, in which the associated audio is another speaker [51], [52].

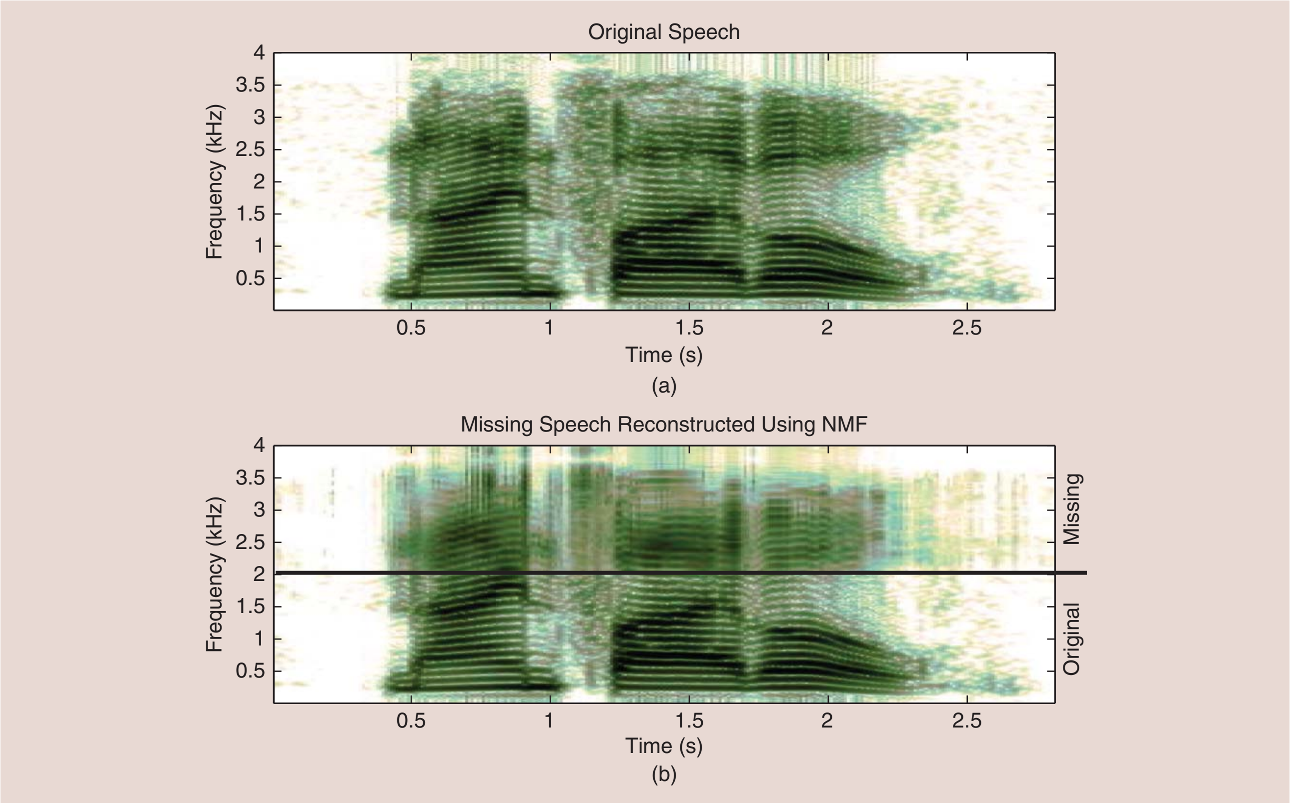
Missing data imputation [29], [53], [54] is closely related to bandwidth extension in that the goal is to estimate a full-spectrum audio signal, but with the difference that the missing data are not a set of predetermined frequency bands but rather arbitrary located time–frequency entries of the spectrogram. Algorithms for compositional models can be easily modified so that model parameters are estimated using only a part of the observed data (ignoring missing data) [29], [54], but the model output can be calculated also for entries corresponding to the missing data. Provided that there is a sufficient amount of observed (not missing) data, which will allow estimating the activations (and atoms in the case of unsupervised processing), reasonable estimates of missing values can be obtained because of dependencies between observed and missing values. In general, the quality of a model can be judged by its ability to make predictions, and the capability of compositional models to predict missing data also illustrates its effectiveness.

EXCITATION-FILTER MODEL AND CHANNEL COMPENSATION

Creating dictionaries from training data, as presented earlier, yields accurate representations as long as the data from which the dictionaries are learned match the observed data. In many practical situations, this is not the case, and there is a need to adapt the learned dictionaries. Moreover, we often have knowledge about the types of sources to be modeled, e.g., that they are musical instruments but do not have suitable training data to estimate the dictionaries in a supervised manner.

Natural sound sources can be modeled as an excitation signal being filtered by an instrument body filter or vocal tract filter. These kinds of excitation- or source-filter models have been very effective, e.g., in speech coding (several codecs use it). In addition to modeling the properties of a body filter, the filter can also model the response from a source to a microphone and, therefore, also do channel compensation.

In the context of compositional models, excitation-filter models have been found useful in, e.g., music processing [55], [56], where both the excitations and filters contain different type of information: excitations typically consists of harmonic spectra with different fundamental frequency values and are therefore useful in pitch estimation, whereas the filter carries instrument-dependent information that can be used for instrument recognition [5].



[FIG9] An example of bandwidth extension of the spoken sequence of digits "nine five oh." (a) The log-scaled spectrogram of the full-bandwidth signal. (b) The reconstruction of the top half obtained using only the 256 lowest frequency bands. For this reconstruction, an exemplar-based, speaker-dependent dictionary of 10,000 atoms was used, randomly extracted from a nonoverlapping data set. We can observe that although some fine detail is lost, the overall structure is captured very well.

Filtering, which corresponds to convolution in the time domain, can be expressed as a pointwise multiplication in the frequency domain. In the context of compositional models, the filtering can therefore be modeled in the magnitude spectral domain by pointwise multiplication of the magnitude spectrum of the excitation and the magnitude spectrum response of the filter. Assuming a fixed magnitude spectrum response of the filter that is denoted by the length-F column vector $\mathbf{h}$, the model for a filtered atom $\mathbf{a}_n$ is given as

$$\mathbf{a}_k = \mathbf{e}_k \otimes \mathbf{h}, \tag{18}$$

where $\mathbf{e}_k$ is the excitation of the kth atom. Here, all the atoms share the same filter, and the model for an input spectrum $\mathbf{y}_t$ in frame t is

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} (\mathbf{a}_k \otimes \mathbf{h}) x_k[t]. \tag{19}$$

When multiple sources are modeled, the atoms of each source can also have a separate filter [5]. The free parameters of an excitation-filter model can be estimated using the principles described in the previous sections—by applying iteratively update rules for each of the terms that decrease the divergence between an observed spectrogram and the model. Even for complex models like this, deriving update rules is rather straightforward using the principles presented in [3], [57], and [58].

Excitations can often be parameterized quite compactly: e.g., in music signal processing, it is known that many sources are harmonic and many sources have a distinct set of fundamental frequency values that they can produce, each corresponding to a harmonic spectrum with different fundamental f_0. Therefore, many excitation-filter models use a fixed set of harmonic excitations [5], [55], [58].

The filters, on the other hand, are specific to each instrument, recording environment, or microphone. To avoid the filter modeling harmonic structures when learned in an unsupervised manner, smooth filters over frequency can be obtained, e.g., by using constraints on two adjacent filter values [56], or by modeling filters a sum of smooth elementary filter atoms [55].

Figure 10 gives an example of an atom being modeled using the excitation-filter model. The filter is modeled as the sum of spectrally smooth filter atoms to make the filter also spectrally smooth. The excitation is a flat harmonic spectrum. The modeled atom can have a high frequency resolution, but it is parameterized only by the activations of few filter atoms and the pitch of the harmonic excitation. The model therefore offers an efficient way to adapt generic harmonic atoms to represent any harmonic signals.

The filter part of the excitation-filter model is able to compensate any linear channel effects. Therefore, the excitation-filter model can also be applied in a scenario in which the atoms in a dictionary that are acquired in specific conditions are viewed as excitations, and a filter is learned to accommodate the dictionary to a new condition.

AUDIO DEREVERBERATION

The excitation-filter model discussed in the previous section is only able to deal with filters whose length is smaller than one audio frame. Audio signals recorded in realistic indoor environments always contain some reverberation, which can have impulse response lengths much longer (typically hundreds of milliseconds to seconds) than frame lengths appropriate for audio compositional models (tens of milliseconds). Furthermore, reverberation is a commonly used effect in music production since a moderate amount of reverberation is found to be perceptually pleasant.

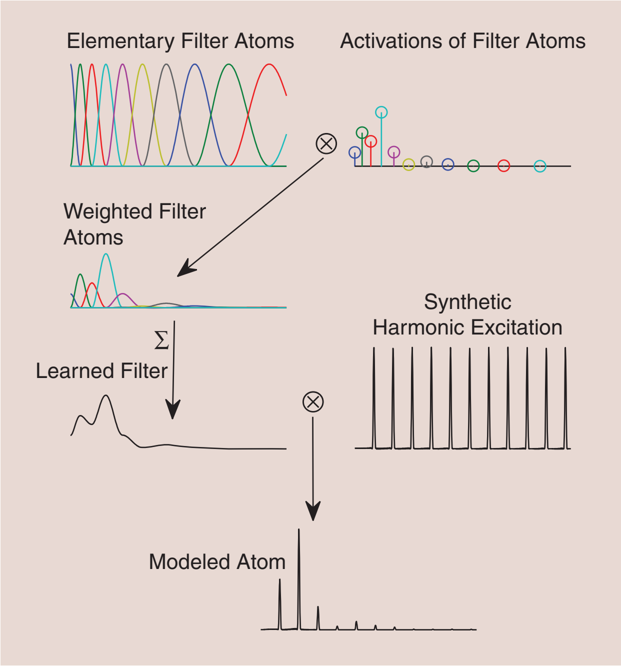
However, too much reverberation decreases the intelligibility of audio and interferes with many audio analysis algorithms. Therefore, there is a need for dereverberation methods and analysis methods that are robust to reverberations.

Reverberation can be formulated as a compositional process as a convolution between the magnitude spectrogram $|S[f,t]|$ of a dry, unreverberant signal, and the magnitude response $|H[f,t]|$ of a filter in the magnitude spectrogram domain [59], [60]

$$|Y[f,t]| \approx \sum_{\tau=0}^{M} |S[f,t-\tau]|\,|H[f,\tau]|, \tag{20}$$

$$\equiv |S[f,t]| * |H[f,t]|, \tag{21}$$

where M is the length of the filter (in frames). Blind estimation of dry signals and reverberation filters is not feasible since the model is ambiguous, and the roles of the source and the impulse response can end up swapped if other restrictions are not used. A suitable a priori information to regularize the model can be, e.g., sparseness [60] or a dictionary-based model [59]. Thus, in practice, we can model $|S(f,t)|$ using another compositional model. The model parameters can be estimated using the principles explained previously, i.e., by minimizing a divergence between an observed spectrogram and the model. Figure 11 gives an example of a reverberant speech spectrogram that is modeled as a convolution between a dry speech spectrogram and a spectrogram of filter.

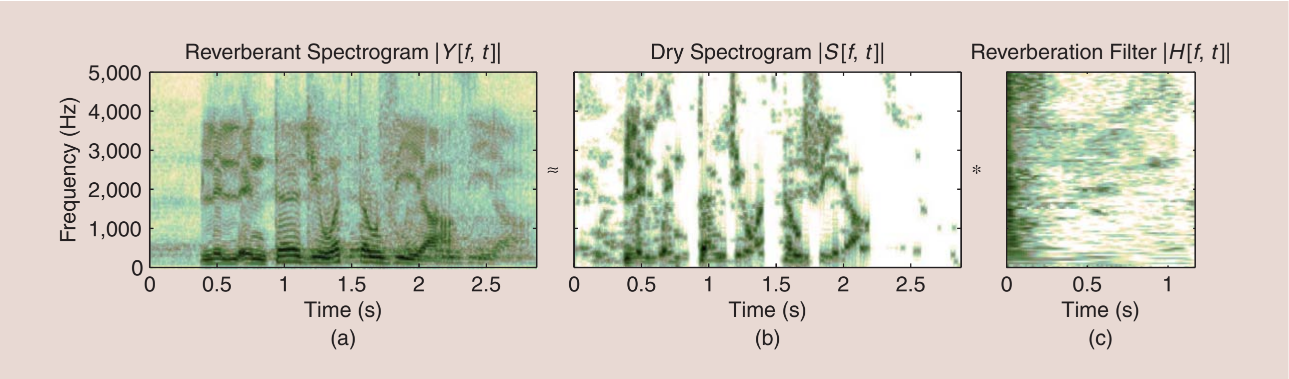


[FIG10] Modeling atoms with the excitation-filter model. The filter is modeled as the sum of elementary filter atoms (upper left), weighted by activations (upper right). The filter is pointwise multiplied by a synthetic harmonic excitation (right) to get an atom (bottom).

NONNEGATIVE MATRIX DECONVOLUTIONS

The basic unsupervised NMF model in (1) is limited in the sense that a random reordering of the frames and columns of the observation matrix $\mathbf{Y}$ does not affect the outcome of the result, i.e., the resulting $\mathbf{X}$ and $\mathbf{A}$ are just reordered versions of $\mathbf{X}$ and $\mathbf{A}$ that would have been obtained without reordering of $\mathbf{Y}$.

Let us illustrate the limitations of the model by the example in Figure 12(a), where few frames of the spectrogram are lost, e.g., because of packet loss. Even though the sounds in the example exhibit clear temporal structure that could be used to impute the missing values, the regular NMF cannot be used for this purpose since there is no data from which to estimate the



[FIG11] (a) The magnitude spectrogram of a reverberant signal can be approximated as (b) the convolution between the spectrograms of a dry signal and (c) the impulse response of the reverberation.

activations that correspond to the missing frames.

REVERBERATION CAN BE FORMULATED AS A COMPOSITIONAL PROCESS.

As in the previous example, sounds typically have strong temporal and spectral dependencies. Temporal context can be included in a compositional model by simply concatenating a number of adjacent observations to a long observation vector [4]. However, this increase of the dimensionality of the observations makes the inference of atoms more difficult—e.g., in the above example, we would need multiple atoms to represent all of the temporally shifted variants of the bird sounds.

The principles used to model reverberant spectrograms and estimate reverberation responses and dry signals can be extended to learn temporal and spectral patterns that span more than one frame or frequency bin, respectively. These nonnegative matrix deconvolution (NMD) [2], [33], [61] methods aim at modeling either temporal or spectral context.

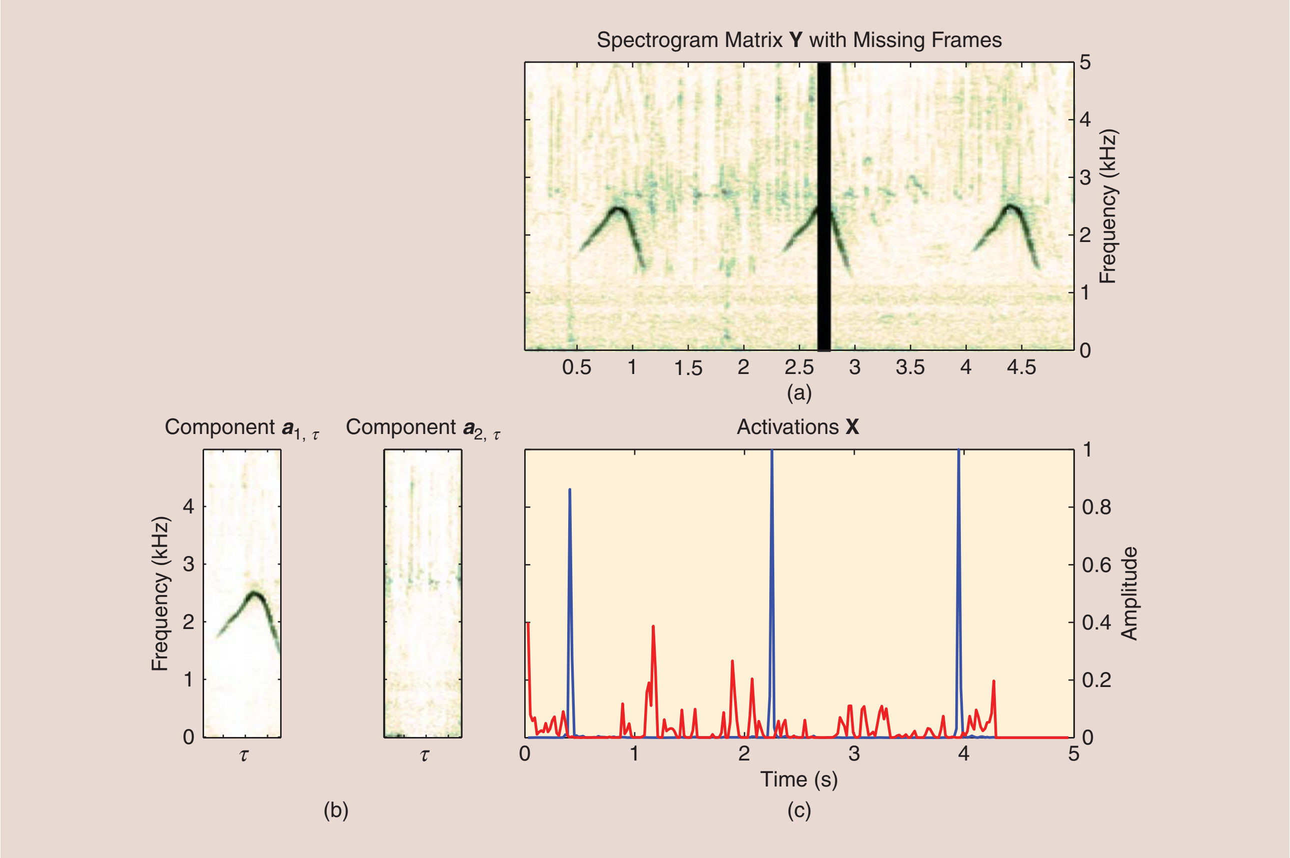
When the model is used in the time domain, it represents a spectrogram as a sum of temporally shifted and scaled versions of atomic spectrogram segments $\mathbf{a}_{n,\tau}$. As before, the atom vectors are indexed by n, but now also with τ, which is the frame index of the short-time spectrogram segment. An illustration of the model is given in Figure 12. Mathematically, the model for an individual mixture spectrogram frame $\mathbf{y}_t$ is given as

$$\mathbf{y}_t \approx \hat{\mathbf{y}}_t = \sum_{k=1}^{K} \sum_{\tau=0}^{L} \mathbf{a}_{k,\tau} x_k[t-\tau], \qquad (22)$$

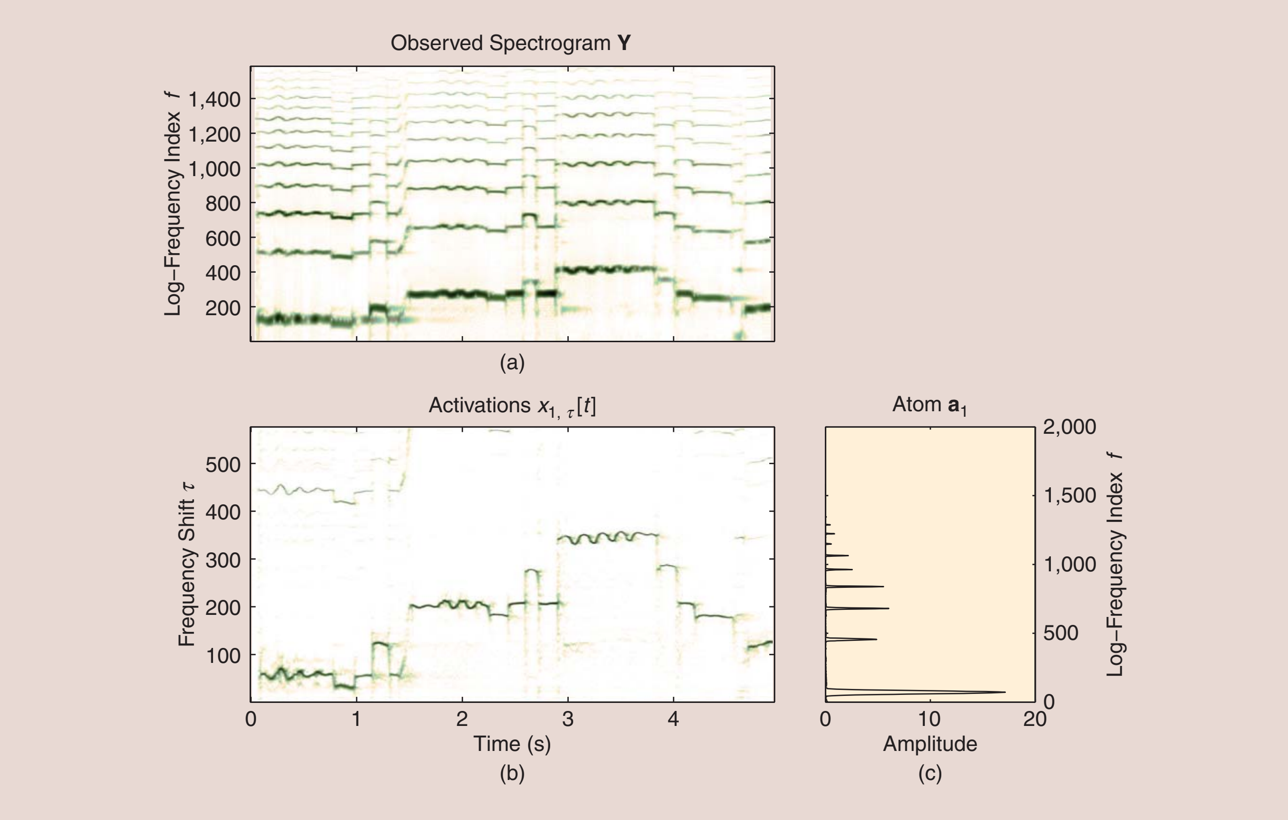
where L is the length of atomic spectrogram events. NMD gets its name from this formulation, as the contribution of a single atom is the convolution between the atom vectors and the activations.

Again, the parameters of the model can be obtained by minimizing a divergence between observations and the model while constraining the model parameters to nonnegative values. In an unsupervised scenario where both the atom vectors and their activations are estimated, care must be taken to limit the number of atoms and the length of events to avoid overfitting.

Convolution in frequency can be used to model pitch shifting of atoms. A limitation of the linear models, at least when a



[FIG12] An illustration of the NMD model. (a) The magnitude spectrogram of a signal consisting of three bird sounds (Friedmann's lark) and background noises. The spectrogram is modeled using NMD to decompose the signal into bird sounds (component 1) and background noises (component 2). (b) The compositional model represents the spectrogram as the weighted and delayed sum of two short event spectrogram segments. (c) The curves show the weights for each delay. The impulses in the curves correspond to the start times of bird sound events in the mixture. The events have been correctly found even though some of the frames in the mixture signal are missing (black vertical bar). Since NMD models the mixture as a sum of segments longer than the missing-frame segment, the model parameters can be used to predict the missing frames.



[FIG13] An illustration of NMD in frequency. (a) A spectrogram of a violin passage with a logarithmic frequency resolution has been decomposed into (c) a weighted sum of shifted versions of a single harmonic atom vector. (b) The activations for each pitch shift and each frame are illustrated. The model allows for representing notes of different pitches with a single harmonic spectrum that is shifted in frequency.

high-frequency resolution feature representation is used, is that a distinct atom is required for representing different pitches of a sound. However, both in speech and music signal processing, the sources to be modeled will be composed of spectra corresponding to different pitch values. If a logarithmic frequency resolution is used, a translation of a spectrum corresponds to a change in its fundamental frequency. Thus, by shifting the entries of a harmonic atom we can model different fundamental frequencies. In the framework of compositional models, we typically not constrain ourselves to a single pitch shift, but define a set of allowed shifts $\mathcal{L}$, and estimate activation $x_{k,\tau}[t]$ for each of the shifts in each frame. The model can be expressed as

$$\hat{y}_{f,t} = \sum_{k=1}^{K} \sum_{\tau \in \mathcal{L}} a_{f+\tau,k} x_{k,\tau}[t]. \qquad (23)$$

Above, $a_{f+\tau,k}$ is the spectrum of the kth atom at frequency f, shifted by τ frequency bins.

Figure 13 illustrates this model by using a single component to represent multiple pitches. The parameters of the model can again be estimated using the aforementioned principles. The plots illustrate that the resulting activations nicely represent the activity of different pitches, which can be useful in music and speech processing.

MULTICHANNEL TENSOR FACTORIZATION

When multichannel audio recordings are to be processed, tensor factorization of their spectrograms [62] has been found to be effective in taking advantage of the spatial properties of sources. In this framework, a spectrogram representation of each of the channels is calculated similarly to one-channel representations. The two-dimensional-spectrogram matrices Y_c of each channel c are concatenated to form a three-dimensional-tensor $\mathcal{Y}$, which entries are indexed as $\mathcal{Y}_{f,t,c}$, i.e., by frequency, time, and channel.

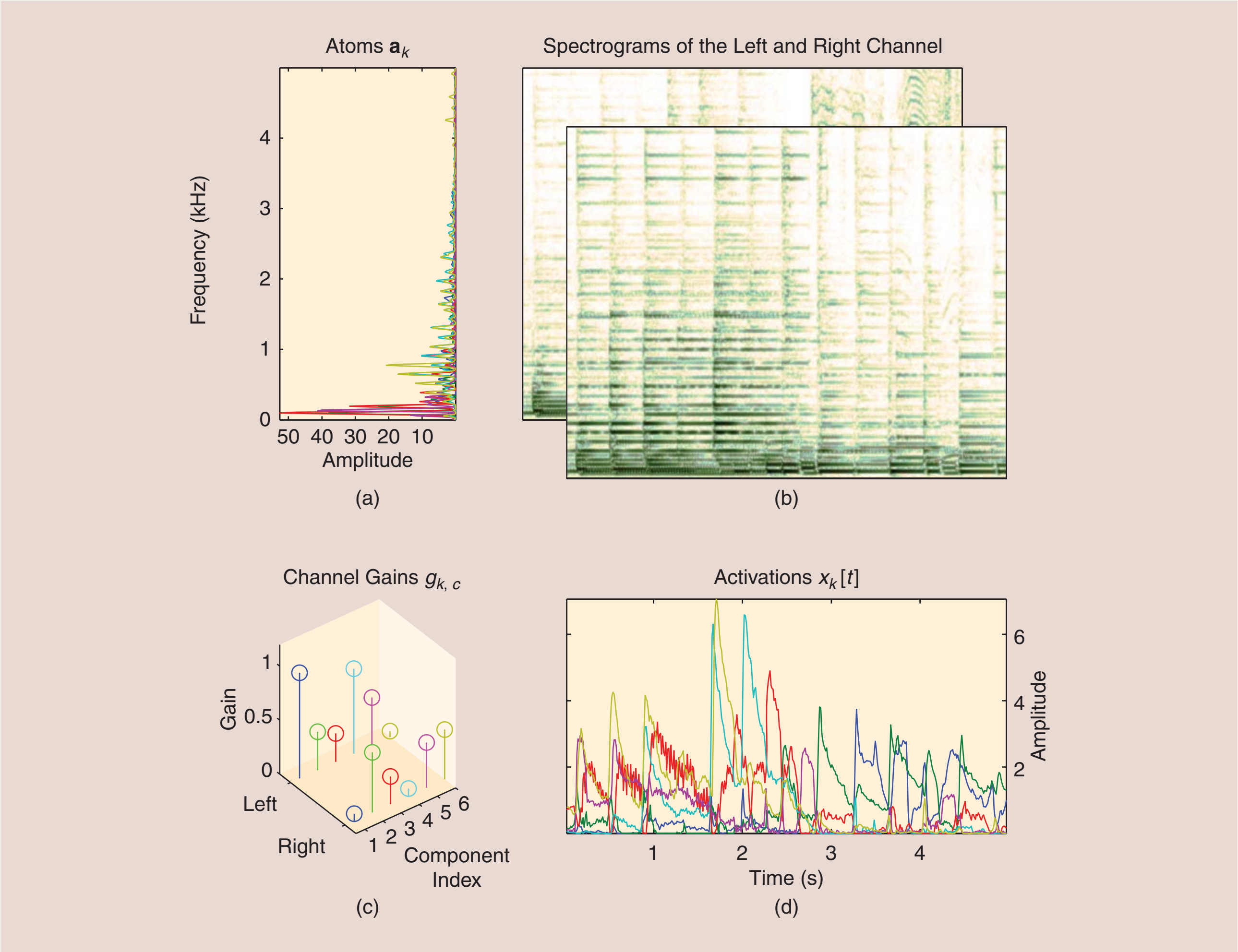
The basic tensor factorization model extends one-channel models by associating each atom with a channel gain $g_{k,c}$, which describes the amplitude of the kth atom in the cth channel.

The tensor factorization model is given as

$$\mathcal{Y}_{f,t,c} \approx \sum_{k=1}^{K} a_{f,k} g_{k,c} x_k[t]. \qquad (24)$$

The model is equivalent to parallel factor analysis (PARAFAC) or canonical polyadic decompositions [63], with the exception that all the parameters of the model are constrained to nonnegative values. Figure 14 illustrates the model.

In comparison to one-channel modeling, the tensor model is most effective in scenarios where the amplitudes of individual sources are different in each channel. The level differences depend on the way the signals are produced. For example, in commercially produced music, especially in music produced



[FIG14] The tensor factorization of multichannel audio. (a) Atom spectra, (b) an illustration of a stereo signal's left and right channel spectrograms factorized into an outer product of the atom spectra, (c) channel gains, and (d) activations in time. Each atom is represented with a different color.

between the 1960s and the 1980s, stereo panning was often used as a strong effect and sources may have significantly different amplitudes. Similarly, if a signal is captured by multiple microphones that are far away from each other, sourcewise amplitude differences between microphones are large. When a signal is captured by a microphone array where the microphones are close to each other, the amplitude differences between channels are typically small, but there are clear phase differences between the signals. In this scenario, techniques [10], [64] that model spectrogram magnitudes with the basic NMF model and phase differences between the channels with another model have shown potential.

DISCUSSION

Even though compositional models are a fairly new technique in the context of audio signal processing, as we have shown in this article, they are applicable to many fundamental audio processing tasks such as source separation, classification, and dereverberation. The compositional nature of the model, the modeling of a spectrogram as a nonnegative sum of atoms having a fixed spectrum and a time-varying gain, is intuitive and offers clear interpretations of the model parameters. This makes it easy to analyze representations obtained with the model, both algorithmically and manually, e.g., by visualizing the models.

The linear nature of the model also offers other advantages. Even when more complex models are used that combine multiple extensions described earlier, the linearity makes it straightforward to derive optimization algorithms for the estimation of the model parameters. Unlike some methods conventionally used for modeling multiple co-occurring sources (e.g., factorial hidden Markov models), the computational complexity of compositional model algorithms scales linearly as the function of the number of sources.

Compositional models have also some limitations. In the context of audio processing, they are mainly applied on magnitudes of time–frequency representations and require additional phase information for signal reconstruction. Therefore, the models have mainly applications in analyzing or processing existing signals, and their applicability in, e.g., sound synthesis is limited. Because of the linearity of the models, compositional models are also not

[THE ABILITY OF THE MODELS TO COUPLE ACOUSTIC AND OTHER TYPES OF INFORMATION ENABLES AUDIO ANALYSIS AND RECOGNITION DIRECTLY USING THE MODEL.]

well suited for modeling nonlinear phenomena. Compositional models use iterative algorithms for finding the model parameters, and their computational complexity is quite significant when large dictionaries are used. Thus, the accuracy of the models may need to be compromised in the case of real-time implementations. The optimization problems involved with compositional models are often nonconvex, and therefore, different algorithms and their initializations lead to different solutions, which needs to be taken into account when results obtained with the models are examined. Even though designing algorithms for new compositional models is in general rather straightforward, the sensitivity of the algorithms to get stuck into a local minimum far away from the global optimum increases as the structure of the model becomes more complex, and the model order increases. To get more accurate solutions with complex models, carefully designed initializations or regularizations may be needed.

Compositional models provide a single framework that enables modeling of several phenomena present in real-world audio: additive sources, sources consisting of multiple sound objects, convolutive noise, and reverberation. Frameworks that combine these in a systematic and flexible way have already been presented [57], [58]. Moreover, the ability of the models to couple acoustic and other types of information enables audio analysis and recognition directly using the model. To be able to handle all of this within a single framework is a great advantage in comparison to methods that tackle just a specific task since it offers the potential of jointly modeling multiple effects that affect each other, such as reverberation and source mixing.

ACKNOWLEDGMENTS

Tuomas Virtanen was financially supported by the Academy of Finland (grant 258708). The research of Jort F. Gemmeke was funded by the IWT-SBO project ALADIN contract 100049.

AUTHORS

Tuomas Virtanen (tuomas.virtanen@tut.fi) received the M.S. and doctor of science degrees in information technology from the Tampere University of Technology (TUT), Finland, in 2001 and 2006, respectively. He is an academy research fellow and an adjunct professor in the Department of Signal Processing, TUT. He is also a research associate in the Department of Engineering, Cambridge University, United Kingdom. He is known for his pioneering work on single-channel sound source separation using nonnegative matrix factorization-based techniques and their application to noise-robust speech recognition, music content analysis, and audio classification. His other research interests include the content analysis of audio signals and machine learning.

Jort F. Gemmeke (jgemmeke@amadana.nl) received the M.S. degree in physics from the Universiteit van Amsterdam in 2005. In 2011, he received the Ph.D. degree from the University of Nijmegen on the subject of noise robust automatic speech recognition (ASR) using missing data techniques. He is a postdoctoral researcher at KU Leuven, Belgium. He is known for pioneering the field of exemplar-based noise robust ASR. His research interests are automatic speech recognition, source separation, noise robustness, and acoustic modeling, in particular, exemplar-based methods and methods using sparse representations.

Bhiksha Raj (bhiksha@cs.cmu.edu) received the Ph.D. degree from Carnegie Mellon University (CMU) in 2000. From 2001 to 2008, he worked at Mitsubishi Electric Research Labs in Cambridge, Massachusetts, where he led the research effort on speech processing. He is an associate professor at the Language Technologies Institute of CMU with additional affiliations to the Machine Learning and Electrical and Computer Engineering Departments of the university. He has been at CMU since 2008. His research interests include speech and audio processing, automatic speech recognition, natural language processing, and machine learning.

Paris Smaragdis (paris@illinois.edu) received the Ph.D. degree from the Massachusetts Institute of Technology in 2003. He is an assistant professor in the Computer Science and Electrical Engineering Departments at the University of Illinois, Urbana Champaign, and a research scientist at Adobe. He is the inventor of frequency-domain ICA and several of the approaches that are now common in compositional model-based signal enhancement. His research interests are in computer audition, machine learning, and speech recognition.

REFERENCES

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

[2] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[4] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[5] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. Int. Conf. Music Information Retrieval*, Kobe, Japan, 2009, pp. 327–332.

[6] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL approach to the CHiME Challenge 2013: Multi-stream ASR exploiting BLSTM networks and sparse NMF," in *Proc. 2nd Int. Workshop on Machine Listening in Multisource Environments*, Vancouver, Canada, 2013, pp. 25–30.

[7] Y.-C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," *Pattern Recognit. Lett.*, vol. 26, no. 9, pp. 1327–1336, 2005.

[8] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian nonnegative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[9] D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in *Proc. EUROSPEECH*, Lisbon, Portugal, 2005, pp. 1505–1508.

[10] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[11] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio & music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, 2009.

[12] J. Nikunen and T. Virtanen, "Object-based audio coding using nonnegative matrix factorization for the spectrogram representation," in *Proc. 128th Audio Engineering Society Convention*, London, 2010.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computat.*, vol. 21, no. 3, pp. 793–830, 2009.

[14] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Proc. Neural Information Processing Systems*, Vancouver, Canada, 2007, pp. 1313–1320.

[15] C. Ding, T. Li, and W. Ping, "On the equivalence between nonnegative matrix factorization and probabilistic latent semantic indexing," *Computat. Stat. Data Anal.*, vol. 52, no. 8, pp. 3913–3927, 2008.

[16] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin: Springer-Verlag, 1990.

[17] B. C. J. Moore, Ed., *Hearing—Handbook of Perception and Cognition*, 2nd ed. San Diego, CA: Academic Press, 1995.

[18] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, Santander, Spain, 2012, pp. 1–6.

[19] J. Carabias-Orti, F. Rodriguez-Serrano, P. Vera-Candeas, F. Canadas-Quesada, and N. Ruiz-Reyes, "Constrained nonnegative sparse coding using learnt instrument templates for realtime music transcription," in *Proc. Engineering Applications of Artificial Intelligence*, 2013, pp. 1671–1680.

[20] F. Weninger and B. Schuller, "Optimization and parallelization of monaural source separation algorithms in the openBliSSART toolkit," *J. Signal Process. Syst.*, vol. 69, no. 3, pp. 267–277, 2012.

[21] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computat.*, vol. 23, no. 9, pp. 2421–2456, 2011.

[22] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Proc. Neural Information Processing Systems*, Denver, CO, 2000, pp. 556–562.

[23] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *Signal Process.*, vol. 87, no. 8, pp. 1904–1916, 2007.

[24] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM J. Sci. Comput.*, vol. 33, no. 6, pp. 3261–3281, 2011.

[25] T. Virtanen, J. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete nonnegative representations of audio," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 11, 2013.

[26] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1–2, pp. 177–196, 2001.

[27] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computat. Intell. Neurosci.*, vol. 2008, 2008.

[28] G. J. Mysore, P. Smaragdis, and B. Raj, "Nonnegative hidden Markov modeling of audio with application to source separation," in *Proc. 9th Int. Conf. Latent Variable Analysis and Signal Separation*, St. Malo, France, 2010, pp. 140–148.

[29] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for time-frequency representations of audio signals," *J. Signal Process. Syst.*, vol. 11, no. 3, pp. 361–370, 2011.

[30] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of NMF," *Computat. Intell. Neurosci.*, vol. 2008, 2008.

[31] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Budapest, Hungary, 2004, pp. 2529–2533.

[32] P. O. Hoyer, "Nonnegative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[33] P. D. O. Grady, "Sparse separation of underdetermined speech mixtures," Ph.D. dissertation, Natl. Univ. of Ireland, Maynooth, 2007.

[34] T. Virtanen, "Spectral covariance in prior distributions of nonnegative matrix factorization based speech separation," in *Proc. European Signal Processing Conf.*, Glasgow, Scotland, 2009, pp. 1933–1937.

[35] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Proc. Neural Information Processing Systems*, Vancouver, Canada, 2009, pp. 1705–1713.

[36] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 32, no. 2, pp. 236–242, 1984.

[37] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Processing Lett.*, vol. 20, no. 3, pp. 217–220, 2013.

[38] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD and its nonnegative variant for dictionary design," in *Proc. SPIE Conf. Wavelet Applications in Signal and Image Processing XI*, San Diego, CA, 2005, pp. 327–339.

[39] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Mag.*, vol. 24, no. 4, pp. 118–121, 2007.

[40] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Proc. IEEE Int. Conf. Audio, Speech and Signal Processing*, Prague, Czech Republic, 2011, pp. 21–24.

[41] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computat. Intell. Neurosci.*, vol. 2009, 2009.

[42] M. N. Schmidt and M. Mørup, "Infinite nonnegative matrix factorizations," in *Proc. European Signal Processing Conf.*, Aalborg, Denmark, 2010.

[43] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian nonnegative matrix factorization," in *Proc. 8th Int. Conf. Independent Component Analysis and Blind Signal Separation*, Paraty, Brazil, 2009, pp. 540–547.

[44] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proc. Interspeech 2012*, Portland, OR, Oregon.

[45] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. IEEE Int. Conf. Audio, Speech and Signal Processing*, Dallas, TX, 2010, pp. 4370–4373.

[46] J. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proc. European Signal Processing Conf.*, Glasgow, Scotland, 2009, pp. 24–28.

[47] K. Mahkonen, A. Hurmalainen, T. Virtanen, and J. F. Gemmeke, "Mapping sparse representation to state likelihoods in noise-robust automatic speech recognition," in *Proc. Interspeech 2011*, Florence, Italy, pp. 465–468.

[48] Y. Sun, B. Cranen, J. F. Gemmeke, L. Boves, L. ten Bosch, and M. M. Doss, "Using sparse classification outputs as feature observations for noise-robust ASR," in *Proc. Interspeech 2012*, Portland, OR.

[49] T. N. Sainath, D. Nahamoo, D. Kanevsky, and B. Ramabhadran, "Enhancing exemplar-based posteriors for speech recognition tasks," in *Proc. Interspeech 2012*, Portland, OR.

[50] B. Raj, R. Singh, M. Shashanka, and P. Smaragdis, "Bandwidth expansion with a Polya Urn model," in *Proc. IEEE Int. Conf. Audio, Speech and Signal Processing*, Honolulu, HI, 2007, pp. IV-597–IV-600.

[51] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Spoken Language Technology Workshop*, 2012, pp. 313–317.

[52] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using nonnegative spectrogram deconvolution," in *Proc. 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, 2013, pp. 201–206.

[53] J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE J. Sel. Top. Signal Processing*, vol. 4, no. 2, pp. 272–287, 2010.

[54] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Computational auditory induction as a missing-data model-fitting problem with Bregman divergence," *SIAM J. Sci. Comput.*, vol. 54, no. 5, pp. 658–676, 2011.

[55] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Top. Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[56] J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. Canadas-Quesada, "Musical instrument sound multi-excitation model for nonnegative spectrogram factorization," *IEEE J. Sel. Top. Signal Processing*, vol. 5, no. 6, pp. 1144–1158, 2011.

[57] Y. K. Yilmaz, A. T. Cemgil, and U. Simsekli, "Generalized coupled tensor factorization," in *Proc. Neural Information Processing Systems*, Granada, Spain, 2011, pp. 2151–2159.

[58] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[59] N. Yasuraoka, H. Kameoka, T. Yoshioka, and H. G. Okuno, "I-divergence-based dereverberation method with auxiliary function approach," in *Proc. IEEE Int. Conf. Audio, Speech and Signal Processing*, Prague, Czech Republic, 2011, pp. 369–372.

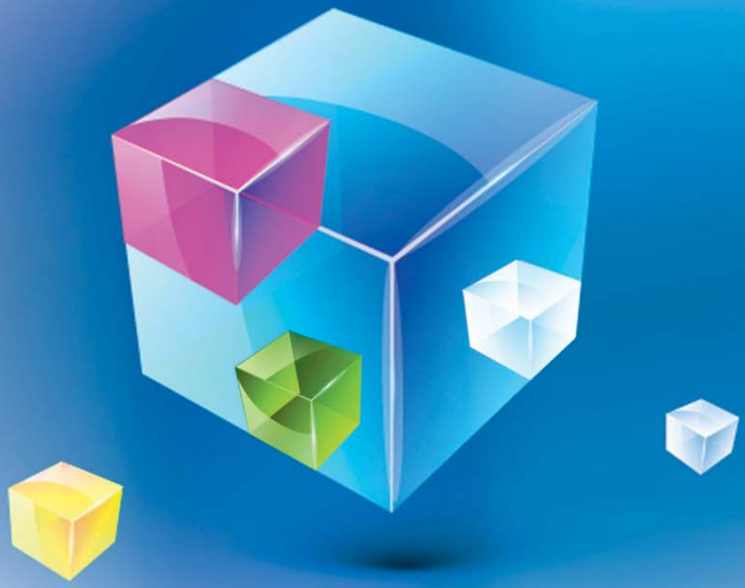
[60] R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in *Proc. IEEE Int. Conf. Audio, Speech and Signal Processing*, Dallas, TX, 2010, pp. 1914–1917.

[61] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proc. Int. Workshop on Machine Listening in Multisource Environments*, Florence, Italy, 2011, pp. 24–29.

[62] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical source separation," *Computat. Intell. Neurosci.*, vol. 2008, 2008.

[63] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," in *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[64] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Formulations and algorithms for multichannel complex NMF," in *Proc. IEEE Int. Conf. Audio, Speech and Signal Processing*, Prague, Czech Republic, 2011, pp. 229–232.



[Andrzej Cichocki, Danilo P. Mandic,
Anh Huy Phan, Cesar F. Caiafa,
Guoxu Zhou, Qibin Zhao, and
Lieven De Lathauwer]

TENSOR DECOMPOSITIONS for Signal Processing Applications

[From two-way to multiway component analysis]

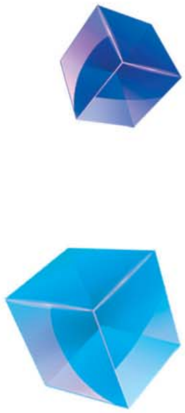


IMAGE LICENSED BY GRAPHIC STOCK

Digital Object Identifier 10.1109/MSP.2013.2297439
Date of publication: 12 February 2015

The widespread use of multisensor technology and the emergence of big data sets have highlighted the limitations of standard flat-view matrix models and the necessity to move toward more versatile data analysis tools. We show that higher-order tensors (i.e., multiway arrays) enable such a fundamental paradigm shift toward models that are essentially polynomial, the uniqueness of which, unlike the matrix methods, is guaranteed under very mild and natural conditions. Benefiting from the power of multilinear algebra as their mathematical backbone, data analysis techniques using tensor decompositions are shown to have great flexibility in the choice of constraints which match data properties and extract more general latent components in the data than matrix-based methods.

A comprehensive introduction to tensor decompositions is provided from a signal processing perspective, starting from the algebraic foundations, via basic canonical polyadic and Tucker models, to advanced cause-effect and multiview data analysis schemes. We show that tensor decompositions enable natural generalizations of some commonly used signal processing paradigms, such as canonical correlation and subspace techniques, signal separation, linear regression, feature extraction, and classification. We also cover computational aspects and point out how ideas from compressed sensing (CS) and scientific computing may be used for addressing the otherwise unmanageable storage and manipulation issues associated with big data sets. The

1053-5888/15©2015IEEE

concepts are supported by illustrative real-world case studies that highlight the benefits of the tensor framework as efficient and promising tools, inter alia, for modern signal processing, data analysis, and machine-learning applications; moreover, these benefits also extend to vector/matrix data through tensorization.

HISTORICAL NOTES

The roots of multiway analysis can be traced back to studies of homogeneous polynomials in the 19th century, with contributors including Gauss, Kronecker, Cayley, Weyl, and Hilbert. In the modern-day interpretation, these are fully symmetric tensors. Decompositions of nonsymmetric tensors have been studied since the early 20th century [1], whereas the benefits of using more than two matrices in factor analysis (FA) [2] have been apparent in several communities since the 1960s. The Tucker decomposition (TKD) for tensors was introduced in psychometrics [3], [4], while the canonical polyadic decomposition (CPD) was independently rediscovered and put into an application context under the names of canonical decomposition (CANDECOMP) in psychometrics [5] and parallel factor model (PARAFAC) in linguistics [6]. Tensors were subsequently adopted in diverse branches of data analysis such as chemometrics, the food industry, and social sciences [7], [8]. When it comes to signal processing, the early 1990s saw a considerable interest in higher-order statistics (HOS) [9], and it was soon realized that, for multivariate cases, HOS are effectively higher-order tensors; indeed, algebraic approaches to independent component analysis (ICA) using HOS [10]–[12] were inherently tensor based. Around 2000, it was realized that the TKD represents a multilinear singular value decomposition (MLSVD) [15]. Generalizing the matrix singular value decomposition (SVD), the workhorse of numerical linear algebra, the MLSVD spurred the interest in tensors in applied mathematics and scientific computing in very high dimensions [16]–[18]. In parallel, CPD was successfully adopted as a tool for sensor array processing and deterministic signal separation in wireless communication [19], [20]. Subsequently, tensors have been used in audio, image and video processing, machine learning, and biomedical applications, to name but a few areas. The significant interest in tensors and their quickly emerging applications is reflected in books [7], [8], [12], [21]–[23] and tutorial papers [24]–[31] covering various aspects of multiway analysis.

FROM A MATRIX TO A TENSOR

Approaches to two-way (matrix) component analysis are well established and include principal component analysis (PCA), ICA, nonnegative matrix factorization (NMF), and sparse component analysis (SCA) [12], [21], [32]. These techniques have become standard tools for, e.g., blind source separation (BSS), feature extraction, or classification. On the other hand, large classes of data arising from modern heterogeneous sensor modalities have a multiway character and are, therefore, naturally represented by multiway arrays or tensors (see the section "Tensorization—Blessing of Dimensionality").

Early multiway data analysis approaches reformatted the data tensor as a matrix and resorted to methods developed for classical two-way analysis. However, such a flattened view of the world and the rigid assumptions inherent in two-way analysis are not always a good match for multiway data. It is only through higher-order tensor decomposition that we have the opportunity to develop sophisticated models capturing multiple interactions and couplings instead of standard pairwise interactions. In other words, we can only discover hidden components within multiway data if the analysis tools account for the intrinsic multidimensional patterns present, motivating the development of multilinear techniques.

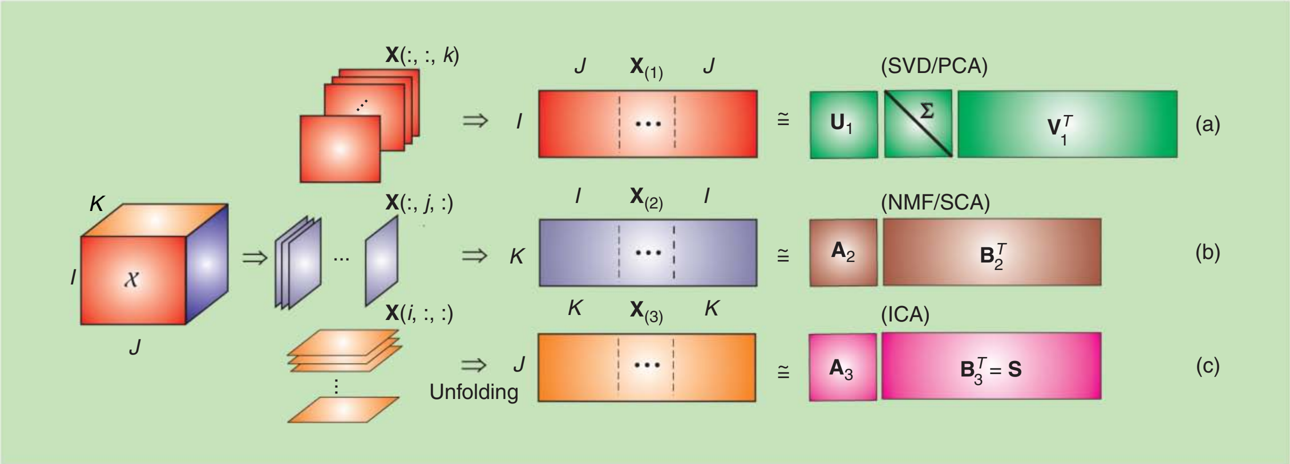
In this article, we emphasize that tensor decompositions are not just matrix factorizations with additional subscripts, multilinear algebra is much more structurally rich than linear algebra. For example, even basic notions such as rank have a more subtle meaning, the uniqueness conditions of higher-order tensor decompositions are more relaxed and accommodating than those for matrices [33], [34], while matrices and tensors also have completely different geometric properties [22]. This boils down to matrices representing linear transformations and quadratic forms, while tensors are connected with multilinear mappings and multivariate polynomials [31].

NOTATIONS AND CONVENTIONS

A tensor can be thought of as a multi-index numerical array, whereby the order of a tensor is the number of its modes or

[TABLE 1] BASIC NOTATION.

$\mathcal{A}, \mathbf{A}, \boldsymbol{a}, a$	TENSOR, MATRIX, VECTOR, SCALAR
$\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_R]$	MATRIX $\mathbf{A}$ WITH COLUMN VECTORS $\boldsymbol{a}_r$
$\boldsymbol{a}(:, i_2, i_3, \ldots, i_N)$	FIBER OF TENSOR $\mathcal{A}$ OBTAINED BY FIXING ALL BUT ONE INDEX
$\mathbf{A}(:, :, i_3, \ldots, i_N)$	MATRIX SLICE OF TENSOR $\mathcal{A}$ OBTAINED BY FIXING ALL BUT TWO INDICES
$\mathcal{A}(:, :, :, i_4, \ldots, i_N)$	TENSOR SLICE OF $\mathcal{A}$ OBTAINED BY FIXING SOME INDICES
$\mathcal{A}(\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N)$	SUBTENSOR OF $\mathcal{A}$ OBTAINED BY RESTRICTING INDICES TO BELONG TO SUBSETS $\mathcal{I}_n \subseteq \{1, 2, \ldots, I_n\}$
$\mathbf{A}_{(n)} \in \mathbb{R}^{I_n \times I_1 I_2 \cdots I_{n-1} I_{n+1} \cdots I_N}$	MODE-n MATRICIZATION OF TENSOR $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ WHOSE ENTRY AT ROW i_n AND COLUMN $(i_1 - 1) I_2 \cdots I_{n-1} I_{n+1} \cdots I_N + \cdots + (i_{N-1} - 1) I_N + i_N$ IS EQUAL TO $a_{i_1 i_2 \ldots i_N}$
$\mathrm{vec}(\mathcal{A}) \in \mathbb{R}^{I_N I_{N-1} \cdots I_1}$	VECTORIZATION OF TENSOR $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ WITH THE ENTRY AT POSITION $i_1 + \sum_{k=2}^{N} [(i_k - 1) I_1 I_2 \cdots I_{k-1}]$ EQUAL TO $a_{i_1 i_2 \ldots i_N}$
$\mathbf{D} = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_R)$	DIAGONAL MATRIX WITH $d_{rr} = \lambda_r$
$\mathcal{D} = \mathrm{diag}_N\,(\lambda_1, \lambda_2, \ldots, \lambda_R)$	DIAGONAL TENSOR OF ORDER N WITH $d_{rr \cdots r} = \lambda_r$
$\mathbf{A}^T,\ \mathbf{A}^{-1},\ \mathbf{A}^{\dagger}$	TRANSPOSE, INVERSE, AND MOORE–PENROSE PSEUDOINVERSE



[FIG1] **MWCA for a third-order tensor, assuming that the components are (a) principal and orthogonal in the first mode, (b) nonnegative and sparse in the second mode, and (c) statistically independent in the third mode.**

dimensions; these may include space, time, frequency, trials, classes, and dictionaries. A real-valued tensor of order N is denoted by $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and its entries by $a_{i_1, i_2, \ldots, i_N}$. Then, an $N \times 1$ vector $\boldsymbol{a}$ is considered a tensor of order one, and an $N \times M$ matrix $\mathbf{A}$ a tensor of order two. Subtensors are parts of the original data tensor, created when only a fixed subset of indices is used. Vector-valued subtensors are called *fibers*, defined by fixing every index but one, and matrix-valued subtensors are called *slices*, obtained by fixing all but two indices (see Table 1). The manipulation of tensors often requires their reformatting (*reshaping*); a particular case of reshaping tensors to matrices is termed *matrix unfolding* or *matricization* (see Figure 1). Note that a mode-n multiplication of a tensor $\mathcal{A}$ with a matrix $\mathbf{B}$ amounts to the multiplication of all mode-n vector fibers with $\mathbf{B}$, and that, in linear algebra, the tensor (or outer) product appears in the expression for a rank-1 matrix: $\boldsymbol{a}\boldsymbol{b}^T = \boldsymbol{a} \circ \boldsymbol{b}$. Basic tensor notations are summarized in Table 1, various product rules used in this article are given in Table 2, while Figure 2 shows two particular ways to construct a tensor.

INTERPRETABLE COMPONENTS IN TWO-WAY DATA ANALYSIS

The aim of BSS, FA, and latent variable analysis is to decompose a data matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ into the factor matrices $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_R] \in \mathbb{R}^{I \times R}$ and $\mathbf{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_R] \in \mathbb{R}^{J \times R}$ as

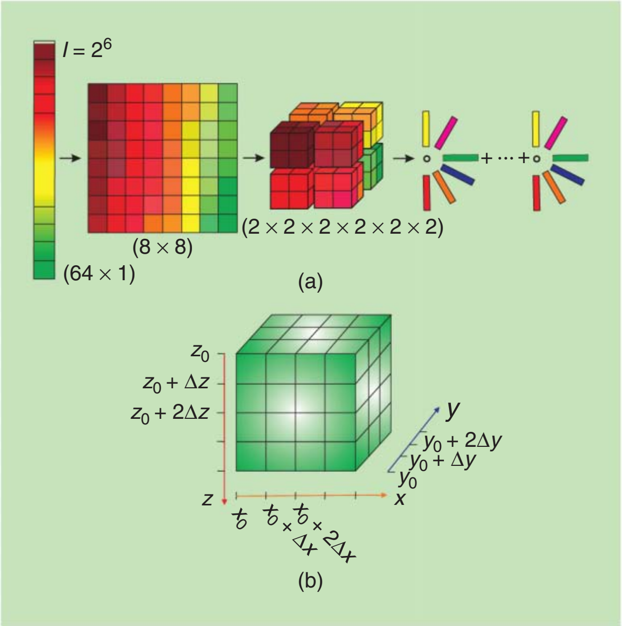
$$\mathbf{X} = \mathbf{A}\mathbf{D}\mathbf{B}^T + \mathbf{E} = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \boldsymbol{b}_r^T + \mathbf{E}$$
$$= \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \circ \boldsymbol{b}_r + \mathbf{E}, \qquad (1)$$

where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_R)$ is a scaling (normalizing) matrix, the columns of $\mathbf{B}$ represent the unknown source signals (factors or latent variables depending on the tasks in hand), the columns of $\mathbf{A}$ represent the associated mixing vectors (or factor loadings), while $\mathbf{E}$ is noise due to an unmodeled data part or model error. In other words, model (1) assumes that the data matrix $\mathbf{X}$ comprises hidden components $\boldsymbol{b}_r$ $(r = 1, 2, \ldots, R)$ that are mixed together in an unknown manner through coefficients $\mathbf{A}$, or, equivalently, that data contain factors that have an associated loading for every data channel. Figure 3(a) depicts the model (1) as a dyadic decomposition, whereby the terms $\boldsymbol{a}_r \circ \boldsymbol{b}_r = \boldsymbol{a}_r \boldsymbol{b}_r^T$ are rank-1 matrices.

The well-known indeterminacies intrinsic to this model are: 1) arbitrary scaling of components and 2) permutation of the rank-1 terms. Another indeterminacy is related to the physical meaning of the factors: if the model in (1) is unconstrained, it admits infinitely many combinations of $\mathbf{A}$ and $\mathbf{B}$. Standard matrix factorizations in linear algebra, such as QR-factorization, eigenvalue decomposition (EVD), and SVD, are only special

[TABLE 2] DEFINITION OF PRODUCTS.

$\mathcal{C} = \mathcal{A} \times_n \mathbf{B}$	MODE-n PRODUCT OF $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ AND $\mathbf{B} \in \mathbb{R}^{J_n \times I_n}$ YIELDS $\mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ WITH ENTRIES $c_{i_1 \cdots i_{n-1} j_n i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} a_{i_1 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} \, b_{j_n i_n}$ AND MATRIX REPRESENTATION $\mathbf{C}_{(n)} = \mathbf{B}\mathbf{A}_{(n)}$
$\mathcal{C} = [\![\mathcal{A}; \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \ldots, \mathbf{B}^{(N)}]\!]$	FULL MULTILINEAR PRODUCT, $\mathcal{C} = \mathcal{A} \times_1 \mathbf{B}^{(1)} \times_2 \mathbf{B}^{(2)} \cdots \times_N \mathbf{B}^{(N)}$
$\mathcal{C} = \mathcal{A} \circ \mathcal{B}$	TENSOR OR OUTER PRODUCT OF $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ AND $\mathcal{B} \in \mathbb{R}^{J_1 \times J_2 \times \cdots \times J_M}$ YIELDS $\mathcal{C} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N \times J_1 \times J_2 \times \cdots \times J_M}$ WITH ENTRIES $c_{i_1 i_2 \cdots i_N j_1 j_2 \cdots j_M} = a_{i_1 i_2 \cdots i_N} b_{j_1 j_2 \cdots j_M}$
$\mathcal{X} = \boldsymbol{a}^{(1)} \circ \boldsymbol{a}^{(2)} \circ \cdots \circ \boldsymbol{a}^{(N)}$	TENSOR OR OUTER PRODUCT OF VECTORS $\boldsymbol{a}^{(n)} \in \mathbb{R}^{I_n}$ $(n = 1, \ldots, N)$ YIELDS A RANK-1 TENSOR $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ WITH ENTRIES $x_{i_1 i_2 \ldots i_N} = a_{i_1}^{(1)} a_{i_2}^{(2)} \ldots a_{i_N}^{(N)}$
$\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$	KRONECKER PRODUCT OF $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ AND $\mathbf{B} \in \mathbb{R}^{J_1 \times J_2}$ YIELDS $\mathbf{C} \in \mathbb{R}^{I_1 J_1 \times I_2 J_2}$ WITH ENTRIES $c_{(i_1-1)J_1+j_1, (i_2-1)J_2+j_2} = a_{i_1 i_2} \, b_{j_1 j_2}$
$\mathbf{C} = \mathbf{A} \odot \mathbf{B}$	KHATRI–RAO PRODUCT OF $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_R] \in \mathbb{R}^{I \times R}$ AND $\mathbf{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R] \in \mathbb{R}^{J \times R}$ YIELDS $\mathbf{C} \in \mathbb{R}^{IJ \times R}$ WITH COLUMNS $\boldsymbol{c}_r = \boldsymbol{a}_r \otimes \boldsymbol{b}_r$



[FIG2] **Construction of tensors. (a) The tensorization of a vector or matrix into the so-called quantized format; in scientific computing, this facilitates supercompression of large-scale vectors or matrices. (b) The tensor is formed through the discretization of a trivariate function $f(x, y, z)$.**

cases of (1), and owe their uniqueness to hard and restrictive constraints such as triangularity and orthogonality. On the other hand, certain properties of the factors in (1) can be represented by appropriate constraints, making possible the unique estimation or extraction of such factors. These constraints include statistical independence, sparsity, nonnegativity, exponential structure, uncorrelatedness, constant modulus, finite alphabet, smoothness, and unimodality. Indeed, the first four properties form the basis of ICA [12]–[14], SCA [32], NMF [21], and harmonic retrieval [35].

TENSORIZATION—BLESSING OF DIMENSIONALITY

While one-way (vectors) and two-way (matrices) algebraic structures were, respectively, introduced as natural representations for segments of scalar measurements and measurements on a grid, tensors were initially used purely for the mathematical benefits they provide in data analysis; for instance, it seemed natural to stack together excitation–emission spectroscopy matrices in chemometrics into a third-order tensor [7].

The procedure of creating a data tensor from lower-dimensional original data is referred to as *tensorization*, and we propose the following taxonomy for tensor generation:

1) *Rearrangement of lower-dimensional data structures*: Large-scale vectors or matrices are readily tensorized to higher-order tensors and can be compressed through tensor decompositions if they admit a low-rank tensor approximation; this principle facilitates big data analysis [23], [29], [30] [see Figure 2(a)]. For instance, a one-way exponential signal $x(k) = az^k$ can be rearranged into a rank-1 Hankel matrix or a Hankel tensor [36]

$$\mathbf{H} = \begin{pmatrix} x(0) & x(1) & x(2) & \cdots \\ x(1) & x(2) & x(3) & \cdots \\ x(2) & x(3) & x(4) & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} = a\,\boldsymbol{b} \circ \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{b} = [1, z, z^2, \ldots]^T$. Also, in sensor array processing, tensor structures naturally emerge when combining snapshots from identical subarrays [19].

2) *Mathematical construction*: Among many such examples, the Nth-order moments (cumulants) of a vector-valued random variable form an Nth-order tensor [9], while in second-order ICA, snapshots of data statistics (covariance matrices) are effectively slices of a third-order tensor [12], [37]. Also, a (channel × time) data matrix can be transformed into a (channel × time × frequency) or (channel × time × scale) tensor via time-frequency or wavelet representations, a powerful procedure in multichannel electroencephalogram (EEG) analysis in brain science [21], [38].

3) *Experiment design*: Multifaceted data can be naturally stacked into a tensor; for instance, in wireless communications the so-called signal diversity (temporal, spatial, spectral, etc.) corresponds to the order of the tensor [20]. In the same spirit, the standard eigenfaces can be generalized to tensor faces by combining images with different illuminations, poses, and expressions [39], while the common modes in EEG recordings across subjects, trials, and conditions are best analyzed when combined together into a tensor [28].

4) *Natural tensor data*: Some data sources are readily generated as tensors [e.g., RGB color images, videos, three-dimensional (3-D) light field displays] [40]. Also, in scientific computing, we often need to evaluate a discretized multivariate function; this is a natural tensor, as illustrated in Figure 2(b) for a trivariate function $f(x, y, z)$ [23], [29], [30].

The high dimensionality of the tensor format is therefore associated with blessings, which include the possibilities to obtain compact representations, the uniqueness of decompositions, the flexibility in the choice of constraints, and the generality of components that can be identified.

CANONICAL POLYADIC DECOMPOSITION

DEFINITION

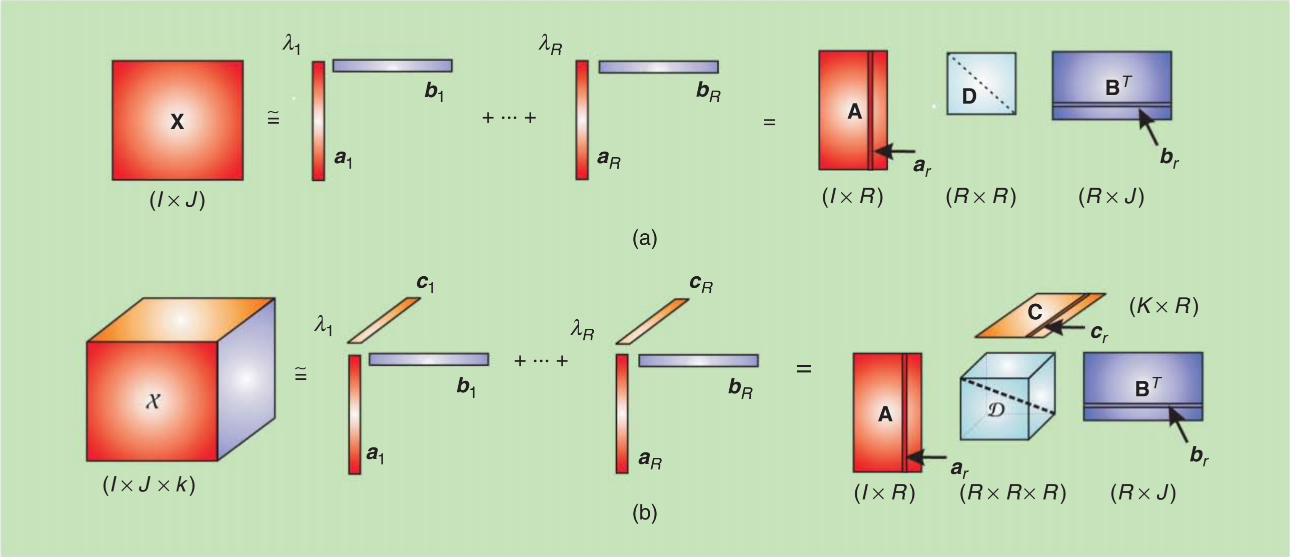
A polyadic decomposition (PD) represents an Nth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ as a linear combination of rank-1 tensors in the form

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \boldsymbol{b}_r^{(1)} \circ \boldsymbol{b}_r^{(2)} \circ \cdots \circ \boldsymbol{b}_r^{(N)}. \tag{3}$$

Equivalently, $\mathcal{X}$ is expressed as a multilinear product with a diagonal core

$$\begin{aligned} \mathcal{X} &= \mathcal{D} \times_1 \mathbf{B}^{(1)} \times_2 \mathbf{B}^{(2)} \cdots \times_N \mathbf{B}^{(N)} \\ &= [\![\mathcal{D}; \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \ldots, \mathbf{B}^{(N)}]\!], \end{aligned} \tag{4}$$

where $\mathcal{D} = \operatorname{diag}_N(\lambda_1, \lambda_2, \ldots, \lambda_R)$ [cf. the matrix case in (1)]. Figure 3 illustrates these two interpretations for a third-order



[FIG3] **The analogy between (a) dyadic decompositions and (b) PDs; the Tucker format has a diagonal core. The uniqueness of these decompositions is a prerequisite for BSS and latent variable analysis.**

tensor. The tensor rank is defined as the smallest value of R for which (3) holds exactly; the minimum rank PD is called *canonical PD* (*CPD*) and is desired in signal separation. The term *CPD* may also be considered as an abbreviation of CANDECOMP/PARAFAC decomposition, see the "Historical Notes" section. The matrix/vector form of CPD can be obtained via the Khatri–Rao products (see Table 2) as

$$\mathbf{X}_{(n)} = \mathbf{B}^{(n)}\mathbf{D}(\mathbf{B}^{(N)} \odot \cdots \odot \mathbf{B}^{(n+1)} \odot \mathbf{B}^{(n-1)} \odot \cdots \odot \mathbf{B}^{(1)})^T,$$
$$\mathrm{vec}(\mathcal{X}) = [\mathbf{B}^{(N)} \odot \mathbf{B}^{(N-1)} \odot \cdots \odot \mathbf{B}^{(1)}]\boldsymbol{d}, \qquad (5)$$

where $\boldsymbol{d} = [\lambda_1, \lambda_2, \ldots, \lambda_R]^T$.

RANK

As mentioned earlier, the rank-related properties are very different for matrices and tensors. For instance, the number of complex-valued rank-1 terms needed to represent a higher-order tensor can be strictly smaller than the number of real-valued rank-1 terms [22], while the determination of tensor rank is in general NP-hard [41]. Fortunately, in signal processing applications, rank estimation most often corresponds to determining the number of tensor components that can be retrieved with sufficient accuracy, and often there are only a few data components present. A pragmatic first assessment of the number of components may be through inspection of the multilinear singular value spectrum (see the "Tucker Decomposition" section), which indicates the size of the core tensor in the right-hand side of Figure 3(b). The existing techniques for rank estimation include the core consistency diagnostic (CORCONDIA) algorithm, which checks whether the core tensor is (approximately) diagonalizable [7], while a number of techniques operate by balancing the approximation error versus the number of degrees of freedom for a varying number of rank-1 terms [42]–[44].

UNIQUENESS

Uniqueness conditions give theoretical bounds for exact tensor decompositions. A classical uniqueness condition is due to Kruskal [33], which states that for third-order tensors, the CPD is unique up to unavoidable scaling and permutation ambiguities, provided that $k_{\mathrm{B}^{(1)}} + k_{\mathrm{B}^{(2)}} + k_{\mathrm{B}^{(3)}} \geq 2R + 2$, where the Kruskal rank k_{B} of a matrix $\mathbf{B}$ is the maximum value ensuring that any subset of k_{B} columns is linearly independent. In sparse modeling, the term $(k_{\mathrm{B}} + 1)$ is also known as the spark [32]. A generalization to Nth-order tensors is due to Sidiropoulos and Bro [45] and is given by

$$\sum_{n=1}^{N} k_{\mathrm{B}^{(n)}} \geq 2R + N - 1. \qquad (6)$$

More relaxed uniqueness conditions can be obtained when one factor matrix has full-column rank [46]–[48]; for a thorough study of the third-order case, we refer to [34]. This all shows that, compared to matrix decompositions, CPD is unique under more natural and relaxed conditions, which only require the components to be sufficiently different and their number not unreasonably large. These conditions do not have a matrix counterpart and are at the heart of tensor-based signal separation.

COMPUTATION

Certain conditions, including Kruskal's, enable explicit computation of the factor matrices in (3) using linear algebra [essentially, by solving sets of linear equations and computing (generalized) EVD] [6], [47], [49], [50]. The presence of noise in data means that CPD is rarely exact, and we need to fit a CPD model to the data by minimizing a suitable cost function. This is typically achieved by minimizing the Frobenius norm of the difference between the given data tensor and its CP approximation, or, alternatively, by least absolute error fitting when the noise is Laplacian [51]. The theoretical Cramér–Rao lower bound and

Cramér–Rao induced bound for the assessment of CPD performance were derived in [52] and [53].

Since the computation of CPD is intrinsically multilinear, we can arrive at the solution through a sequence of linear subproblems as in the alternating least squares (ALS) framework, whereby the least squares (LS) cost function is optimized for one component matrix at a time, while keeping the other component matrices fixed [6]. As seen from (5), such a conditional update scheme boils down to solving overdetermined sets of linear equations.

While the ALS is attractive for its simplicity and satisfactory performance for a few well-separated components and at sufficiently high signal-to-noise ratio (SNR), it also inherits the problems of alternating algorithms and is not guaranteed to converge to a stationary point. This can be rectified by only updating the factor matrix for which the cost function has most decreased at a given step [54], but this results in an N-times increase in computational cost per iteration. The convergence of ALS is not yet completely understood—it is quasilinear close to the stationary point [55], while it becomes rather slow for ill-conditioned cases; for more details, we refer to [56] and [57].

The conventional all-at-once algorithms for numerical optimization, such as nonlinear conjugate gradients, quasi-Newton, or nonlinear least squares (NLS) [58], [59], have been shown to often outperform ALS for ill-conditioned cases and to be typically more robust to overfactoring. However, these come at the cost of a much higher computational load per iteration. More sophisticated versions use the rank-1 structure of the terms within CPD to perform efficient computation and storage of the Jacobian and (approximate) Hessian; their complexity is on par with ALS while, for ill-conditioned cases, the performance is often superior [60], [61].

An important difference between matrices and tensors is that the existence of a best rank-R approximation of a tensor of rank greater than R is not guaranteed [22], [62] since the set of tensors whose rank is at most R is not closed. As a result, the cost functions for computing factor matrices may only have an infimum (instead of a minimum) so that their minimization will approach the boundary of that set without ever reaching the boundary point. This will cause two or more rank-1 terms go to infinity upon convergence of an algorithm; however, numerically, the diverging terms will almost completely cancel one another while the overall cost function will still decrease along the iterations [63]. These diverging terms indicate an inappropriate data model: the mismatch between the CPD and the original data tensor may arise because of an underestimated number of components, not all tensor components having a rank-1 structure, or data being too noisy.

CONSTRAINTS

As mentioned earlier, under quite mild conditions, the CPD is unique by itself, without requiring additional constraints. However, to enhance the accuracy and robustness with respect to noise, prior knowledge of data properties (e.g., statistical independence, sparsity) may be incorporated into the constraints on factors so as to facilitate their physical interpretation, relax the uniqueness conditions, and even simplify computation [64]–[66]. Moreover, the orthogonality and nonnegativity constraints ensure the existence of the minimum of the optimization criterion used [63], [64], [67].

APPLICATIONS

The CPD has already been established as an advanced tool for signal separation in vastly diverse branches of signal processing and data analysis, such as in audio and speech processing, biomedical engineering, chemometrics, and machine learning [7], [24], [25], [28]. Note that algebraic ICA algorithms are effectively based on the CPD of a tensor of the statistics of recordings; the statistical independence of the sources is reflected in the diagonality of the core tensor in Figure 3, i.e., in vanishing cross-statistics [11], [12]. The CPD is also heavily used in exploratory data analysis, where the rank-1 terms capture the essential properties of dynamically complex signals [8]. Another example is in wireless communication, where the signals transmitted by different users correspond to rank-1 terms in the case of line-of-sight propagation [19]. Also, in harmonic retrieval and direction of arrival type applications, real or complex exponentials have a rank-1 structure, for which the use of CPD is natural [36], [65].

EXAMPLE 1

Consider a sensor array consisting of K displaced but otherwise identical subarrays of I sensors, with $\tilde{I} = KI$ sensors in total. For R narrowband sources in the far field, the baseband equivalent model of the array output becomes $\mathbf{X} = \mathbf{A}\mathbf{S}^T + \mathbf{E}$, where $\mathbf{A} \in \mathbb{C}^{\tilde{I} \times R}$ is the global array response, $\mathbf{S} \in \mathbb{C}^{J \times R}$ contains J snapshots of the sources, and $\mathbf{E}$ is the noise. A single source $(R = 1)$ can be obtained from the best rank-1 approximation of the matrix $\mathbf{X}$; however, for $R > 1$, the decomposition of $\mathbf{X}$ is not unique, and, hence, the separation of sources is not possible without incorporating additional information. The constraints on the sources that may yield a unique solution are, for instance, constant modulus and statistical independence [12], [68].

Consider a row-selection matrix $\mathbf{J}_k \in \mathbb{C}^{I \times \tilde{I}}$ that extracts the rows of $\mathbf{X}$ corresponding to the kth subarray, $k = 1, \ldots, K$. For two identical subarrays, the generalized EVD of the matrices $\mathbf{J}_1\mathbf{X}$ and $\mathbf{J}_2\mathbf{X}$ corresponds to the well-known estimation of signal parameters via rotational invariance techniques (ESPRIT) [69]. For the case $K > 2$, we shall consider $\mathbf{J}_k\mathbf{X}$ as slices of the tensor $\mathcal{X} \in \mathbb{C}^{I \times J \times K}$ (see the section "Tensorization—Blessing of Dimensionality"). It can be shown that the signal part of $\mathcal{X}$ admits a CPD as in (3) and (4), with $\lambda_1 = \cdots = \lambda_R = 1$, $\mathbf{J}_k\mathbf{A} = \mathbf{B}^{(1)} \operatorname{diag}(b_{k1}^{(3)}, \ldots, b_{kR}^{(3)})$, and $\mathbf{B}^{(2)} = \mathbf{S}$ [19], and the consequent source separation under rather mild conditions—its uniqueness does not require constraints such as statistical independence or constant modulus. Moreover, the decomposition is unique even in cases when the number of sources, R, exceeds the number of subarray sensors, I, or even the total number of sensors, $\tilde{I}$. Note that particular array geometries, such as linearly and uniformly displaced subarrays, can be converted into a constraint on CPD, yielding a further relaxation of the uniqueness conditions, reduced sensitivity to noise, and often faster computation [65].

TUCKER DECOMPOSITION

Figure 4 illustrates the principle of TKD, which treats a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ as a multilinear transformation of a (typically dense but small) core tensor $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N}$ by the factor matrices $\mathbf{B}^{(n)} = [\boldsymbol{b}_1^{(n)}, \boldsymbol{b}_2^{(n)}, \ldots, \boldsymbol{b}_{R_n}^{(n)}] \in \mathbb{R}^{I_n \times R_n}$, $n = 1, 2, \ldots, N$ [3], [4], given by

$$\mathcal{X} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_N=1}^{R_N} g_{r_1 r_2 \cdots r_N} (\boldsymbol{b}_{r_1}^{(1)} \circ \boldsymbol{b}_{r_2}^{(2)} \circ \cdots \circ \boldsymbol{b}_{r_N}^{(N)}), \tag{7}$$

or equivalently

$$\begin{aligned} \mathcal{X} &= \mathcal{G} \times_1 \mathbf{B}^{(1)} \times_2 \mathbf{B}^{(2)} \cdots \times_N \mathbf{B}^{(N)} \\ &= [\![\mathcal{G}; \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \ldots, \mathbf{B}^{(N)}]\!]. \end{aligned} \tag{8}$$

Via the Kronecker products (see Table 2), TKD can be expressed in a matrix/vector form as

$$\begin{aligned} \mathbf{X}_{(n)} &= \mathbf{B}^{(n)} \mathbf{G}_{(n)} (\mathbf{B}^{(N)} \otimes \cdots \otimes \mathbf{B}^{(n+1)} \otimes \mathbf{B}^{(n-1)} \otimes \cdots \otimes \mathbf{B}^{(1)})^T \\ \mathrm{vec}(\mathcal{X}) &= [\mathbf{B}^{(N)} \otimes \mathbf{B}^{(N-1)} \otimes \cdots \otimes \mathbf{B}^{(1)}] \, \mathrm{vec}(\mathcal{G}). \end{aligned}$$

Although Tucker initially used the orthogonality and ordering constraints on the core tensor and factor matrices [3], [4], we can also employ other meaningful constraints.

MULTILINEAR RANK

For a core tensor of minimal size, R_1 is the column rank (the dimension of the subspace spanned by mode-1 fibers), R_2 is the row rank (the dimension of the subspace spanned by mode-2 fibers), and so on. A remarkable difference from matrices is that the values of $R_1, R_2, \ldots, R_N$ can be different for $N \geq 3$. The N-tuple $(R_1, R_2, \ldots, R_N)$ is consequently called the *multilinear rank* of the tensor $\mathcal{X}$.

LINKS BETWEEN CPD AND TUCKER DECOMPOSTION

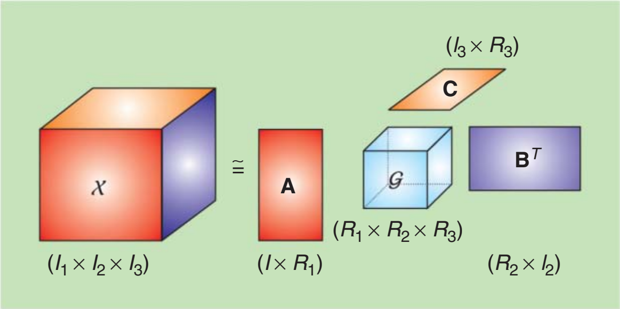
TKD can be considered an expansion in rank-1 terms (polyadic but not necessary canonical), as shown in (7), while (4) represents CPD as a multilinear product of a core tensor and factor matrices (but the core is not necessary minimal); Table 3 shows various other connections. However, despite the obvious interchangeability of notation, the CPD and TKD serve different purposes. In general, the Tucker core cannot be diagonalized, while the number of CPD terms may not be bounded by the multilinear rank. Consequently, in signal processing and data analysis, CPD is typically used for factorizing data into easy to interpret components (i.e., the rank-1 terms), while the goal of unconstrained TKD is most often to compress data into a tensor of smaller size (i.e., the core tensor) or to find the subspaces spanned by the fibers (i.e., the column spaces of the factor matrices).

UNIQUENESS

The unconstrained TKD is in general not unique, i.e., factor matrices $\mathbf{B}^{(n)}$ are rotation invariant. However, physically, the subspaces defined by the factor matrices in TKD are unique, while the bases in these subspaces may be chosen arbitrarily—their choice is compensated for within the core tensor. This becomes clear upon realizing that any factor matrix in (8) can be postmultiplied by any nonsingular (rotation) matrix; in turn, this multiplies the core tensor by its inverse, i.e.,

$$\begin{aligned} \mathcal{X} &= [\![\mathcal{G}; \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \ldots, \mathbf{B}^{(N)}]\!] \\ &= [\![\mathcal{H}; \mathbf{B}^{(1)}\mathbf{R}^{(1)}, \mathbf{B}^{(2)}\mathbf{R}^{(2)}, \ldots, \mathbf{B}^{(N)}\mathbf{R}^{(N)}]\!], \\ \mathcal{H} &= [\![\mathcal{G}; \mathbf{R}^{(1)^{-1}}, \mathbf{R}^{(2)^{-1}}, \ldots, \mathbf{R}^{(N)^{-1}}]\!], \end{aligned} \tag{9}$$

where the matrices $\mathbf{R}^{(n)}$ are invertible.



[FIG4] The Tucker decompostion of a third-order tensor. The column spaces of A, B, and C represent the signal subspaces for the three modes. The core tensor $\mathcal{G}$ is nondiagonal, accounting for the possibly complex interactions among tensor components.

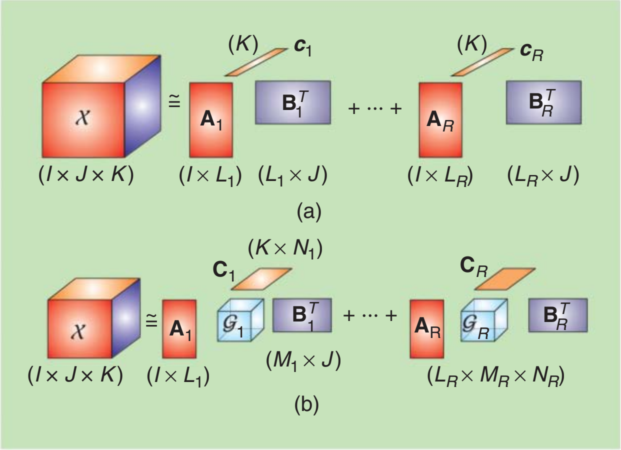
MULTILINEAR SVD

Orthonormal bases in a constrained Tucker representation can be obtained via the SVD of the mode-n matricized tensor $\mathbf{X}_{(n)} = \mathbf{U}_n \boldsymbol{\Sigma}_n \mathbf{V}_n^T$ (i.e., $\mathbf{B}^{(n)} = \mathbf{U}_n$, $n = 1, 2, \ldots, N$). Because of the orthonormality, the corresponding core tensor becomes

$$\mathcal{S} = \mathcal{X} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \cdots \times_N \mathbf{U}_N^T. \tag{10}$$

[TABLE 3] DIFFERENT FORMS OF CPD AND TUCKER REPRESENTATIONS OF A THIRD-ORDER TENSOR $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$.

CPD	TKD
TENSOR REPRESENTATION, OUTER PRODUCTS	
$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r$	$\mathcal{X} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} g_{r_1 r_2 r_3} \boldsymbol{a}_{r_1} \circ \boldsymbol{b}_{r_2} \circ \boldsymbol{c}_{r_3}$
TENSOR REPRESENTATION, MULTILINEAR PRODUCTS	
$\mathcal{X} = \mathcal{D} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$	$\mathcal{X} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$
MATRIX REPRESENTATIONS	
$\mathbf{X}_{(1)} = \mathbf{A}\,\mathbf{D}(\mathbf{C} \odot \mathbf{B})^T$	$\mathbf{X}_{(1)} = \mathbf{A}\,\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})^T$
$\mathbf{X}_{(2)} = \mathbf{B}\,\mathbf{D}(\mathbf{C} \odot \mathbf{A})^T$	$\mathbf{X}_{(2)} = \mathbf{B}\,\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})^T$
$\mathbf{X}_{(3)} = \mathbf{C}\,\mathbf{D}(\mathbf{B} \odot \mathbf{A})^T$	$\mathbf{X}_{(3)} = \mathbf{C}\,\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})^T$
VECTOR REPRESENTATION	
$\mathrm{vec}(\mathcal{X}) = (\mathbf{C} \odot \mathbf{B} \odot \mathbf{A})\boldsymbol{d}$	$\mathrm{vec}(\mathcal{X}) = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\,\mathrm{vec}(\mathcal{G})$
SCALAR REPRESENTATION	
$x_{ijk} = \sum_{r=1}^{R} \lambda_r \, a_{ir} \, b_{jr} \, c_{kr}$	$x_{ijk} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} g_{r_1 r_2 r_3} \, a_{ir_1} \, b_{jr_2} \, c_{kr_3}$
MATRIX SLICES $\mathbf{X}_k = \mathbf{X}(:,:,k)$	
$\mathbf{X}_k = \mathbf{A}\,\mathrm{diag}(c_{k1}, c_{k2}, \ldots, c_{kR})\,\mathbf{B}^T$	$\mathbf{X}_k = \mathbf{A} \sum_{r_3=1}^{R_3} c_{kr_3} \mathbf{G}(:,:,r_3)\mathbf{B}^T$



[FIG5] BTDs find data components that are structurally more complex than the rank-1 terms in CPD. (a) Decomposition into terms with multilinear rank $(L_r, L_r, 1)$. (b) Decomposition into terms with multilinear rank (L_r, M_r, N_r).

Then, the singular values of $\mathbf{X}_{(n)}$ are the Frobenius norms of the corresponding slices of the core tensor $\mathcal{S}$: $(\Sigma_n)_{r_n,r_n} = \|\mathcal{S}(:,:,\ldots,r_n,:,\ldots,:)\|_F$, with slices in the same mode being mutually orthogonal, i.e., their inner products are zero. The columns of $\mathbf{U}_n$ may thus be seen as multilinear singular vectors, while the norms of the slices of the core are multilinear singular values [15]. As in the matrix case, the multilinear singular values govern the multilinear rank, while the multilinear singular vectors allow, for each mode separately, an interpretation as in PCA [8].

LOW MULTILINEAR RANK APPROXIMATION

Analogous to PCA, a large-scale data tensor $\mathcal{X}$ can be approximated by discarding the multilinear singular vectors and slices of the core tensor that correspond to small multilinear singular values, i.e., through truncated matrix SVDs. Low multilinear rank approximation is always well posed; however, the truncation is not necessarily optimal in the LS sense, although a good estimate can often be made as the approximation error corresponds to the degree of truncation. When it comes to finding the best approximation, the ALS-type algorithms exhibit similar advantages and drawbacks to those used for CPD [8], [70]. Optimization-based algorithms exploiting second-order information have also been proposed [71], [72].

CONSTRAINTS AND TUCKER-BASED MULTIWAY COMPONENT ANALYSIS

Besides orthogonality, constraints that may help to find unique basis vectors in a Tucker representation include statistical independence, sparsity, smoothness, and nonnegativity [21], [73], [74]. Components of a data tensor seldom have the same properties in its modes, and for physically meaningful representation, different constraints may be required in different modes so as to match the properties of the data at hand. Figure 1 illustrates the concept of multiway component analysis (MWCA) and its flexibility in choosing the modewise constraints; a Tucker representation of MWCA naturally accommodates such diversities in different modes.

OTHER APPLICATIONS

We have shown that TKD may be considered a multilinear extension of PCA [8]; it therefore generalizes signal subspace techniques, with applications including classification, feature extraction, and subspace-based harmonic retrieval [27], [39], [75], [76]. For instance, a low multilinear rank approximation achieved through TKD may yield a higher SNR than the SNR in the original raw data tensor, making TKD a very natural tool for compression and signal enhancement [7], [8], [26].

BLOCK TERM DECOMPOSITIONS

We have already shown that CPD is unique under quite mild conditions. A further advantage of tensors over matrices is that it is even possible to relax the rank-1 constraint on the terms, thus opening completely new possibilities in, e.g., BSS. For clarity, we shall consider the third-order case, whereby, by replacing the rank-1 matrices $\boldsymbol{b}_r^{(1)} \circ \boldsymbol{b}_r^{(2)} = \boldsymbol{b}_r^{(1)} \boldsymbol{b}_r^{(2)T}$ in (3) by low-rank matrices $\mathbf{A}_r \mathbf{B}_r^T$, the tensor $\mathcal{X}$ can be represented as [Figure 5(a)]

$$\mathcal{X} = \sum_{r=1}^{R} (\mathbf{A}_r \mathbf{B}_r^T) \circ \boldsymbol{c}_r. \tag{11}$$

Figure 5(b) shows that we can even use terms that are only required to have a low multilinear rank (see the "Tucker Decomposition" section) to give

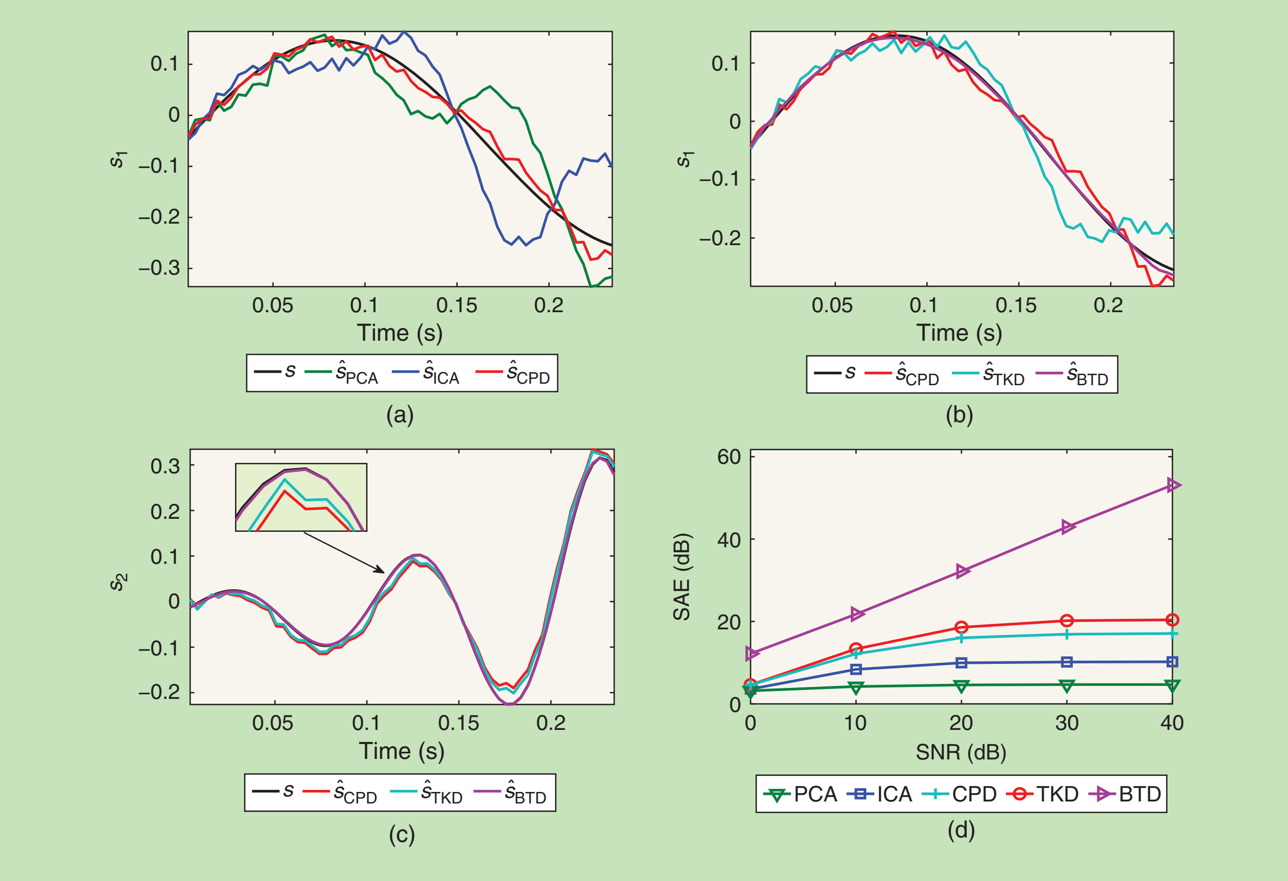
$$\mathcal{X} = \sum_{r=1}^{R} \mathcal{G}_r \times_1 \mathbf{A}_r \times_2 \mathbf{B}_r \times_3 \mathbf{C}_r. \tag{12}$$

These so-called block term decompositions (BTDs) in (11) and (12) admit the modeling of more complex signal components than CPD and are unique under more restrictive but still fairly natural conditions [77]–[79].

EXAMPLE 2

To compare some standard and tensor approaches for the separation of short duration correlated sources, BSS was performed on five linear mixtures of the sources $s_1(t) = \sin(6\pi t)$ and $s_2(t) = \exp(10t)\sin(20\pi t)$, which were contaminated by white Gaussian noise, to give the mixtures $\mathbf{X} = \mathbf{AS} + \mathbf{E} \in \mathbb{R}^{5\times 60}$, where $\mathbf{S}(t) = [\boldsymbol{s}_1(t), \boldsymbol{s}_2(t)]^T$ and $\mathbf{A} \in \mathbb{R}^{5\times 2}$ was a random matrix whose columns (mixing vectors) satisfy $\boldsymbol{a}_1^T \boldsymbol{a}_2 = 0.1$, $\|\boldsymbol{a}_1\|_2 = \|\boldsymbol{a}_2\|_2 = 1$. The 3-Hz sine wave did not complete a full period over the 60 samples so that the two sources had a correlation degree of $(|\boldsymbol{s}_1^T \boldsymbol{s}_2|)/(\|\boldsymbol{s}_1\|_2 \|\boldsymbol{s}_2\|_2) = 0.35$. The tensor approaches, CPD, TKD, and BTD employed a third-order tensor $\mathcal{X}$ of size 24 × 37 × 5 generated from five Hankel matrices whose elements obey $\mathcal{X}(i,j,k) = \mathbf{X}(k, i+j-1)$ (see the section "Tensorization—Blessing of Dimensionality"). The average squared angular error (SAE) was used as the performance measure. Figure 6 shows the simulation results, illustrating the following.

- PCA failed since the mixing vectors were not orthogonal and the source signals were correlated, both violating the assumptions for PCA.
- The ICA [using the joint approximate diagonalization of eigenmatrices (JADE) algorithm [10]] failed because the signals were not statistically independent, as assumed in ICA.



[FIG6] The blind separation of the mixture of a pure sine wave and an exponentially modulated sine wave using PCA, ICA, CPD, TKD, and BTD. The sources s_1 and s_2 are correlated and of short duration; the symbols $\hat{s}_1$ and $\hat{s}_2$ denote the estimated sources. (a)–(c) Sources $s_1(t)$ and $s_2(t)$ and their estimates using PCA, ICA, CPD, TKD, and BTD; (d) average squared angular errors (SAE) in estimation of the sources.

- Low-rank tensor approximation via a rank-2 CPD was used to estimate $\mathbf{A}$ as the third factor matrix, which was then inverted to yield the sources. The accuracy of CPD was compromised as the components of tensor $\mathcal{X}$ cannot be represented by rank-1 terms.
- Low multilinear rank approximation via TKD for the multilinear rank (4, 4, 2) was able to retrieve the column space of the mixing matrix but could not find the individual mixing vectors because of the nonuniqueness of TKD.
- BTD in multilinear rank-(2, 2, 1) terms matched the data structure [78]; it is remarkable that the sources were recovered using as few as six samples in the noise-free case.

HIGHER-ORDER COMPRESSED SENSING (HO-CS)

The aim of CS is to provide a faithful reconstruction of a signal of interest, even when the set of available measurements is (much) smaller than the size of the original signal [80]–[83]. Formally, we have available M (compressive) data samples $\boldsymbol{y} \in \mathbb{R}^M$, which are assumed to be linear transformations of the original signal $\boldsymbol{x} \in \mathbb{R}^I$ $(M < I)$. In other words, $\boldsymbol{y} = \boldsymbol{\Phi x}$, where the sensing matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times I}$ is usually random. Since the projections are of a lower dimension than the original data, the reconstruction is an ill-posed inverse problem whose solution requires knowledge of the physics of the problem converted into constraints. For example, a two-dimensional image $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2}$ can be vectorized as a long vector $\boldsymbol{x} = \text{vec}(\mathbf{X}) \in \mathbb{R}^I$ $(I = I_1 I_2)$ that admits sparse representation in a known dictionary $\mathbf{B} \in \mathbb{R}^{I \times I}$ so that $\boldsymbol{x} = \mathbf{B}\boldsymbol{g}$, where the matrix $\mathbf{B}$ may be a wavelet or discrete cosine transform dictionary. Then, faithful recovery of the original signal $\boldsymbol{x}$ requires finding the sparsest vector $\boldsymbol{g}$ such that

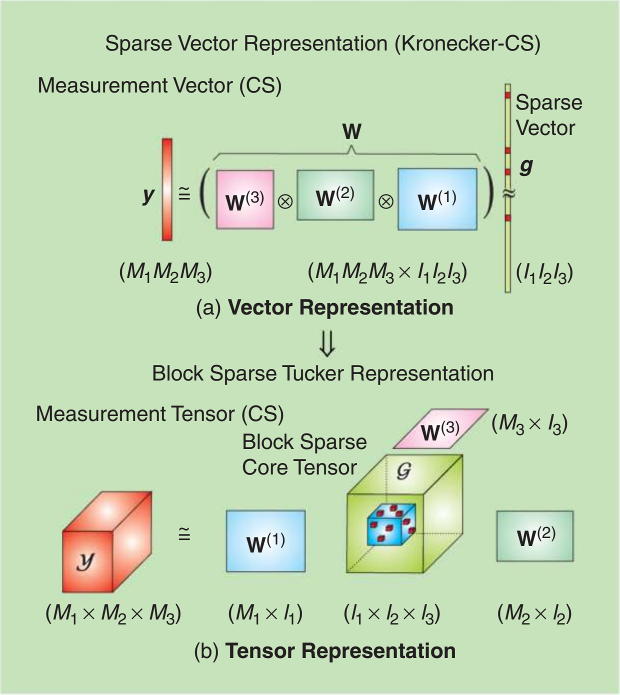
$$\boldsymbol{y} = \mathbf{W}\boldsymbol{g}, \text{ with } \|\boldsymbol{g}\|_0 \leq K, \qquad \mathbf{W} = \boldsymbol{\Phi}\mathbf{B}, \tag{13}$$

where $\|\cdot\|_0$ is the ℓ_0-norm (number of nonzero entries) and $K \ll I$.

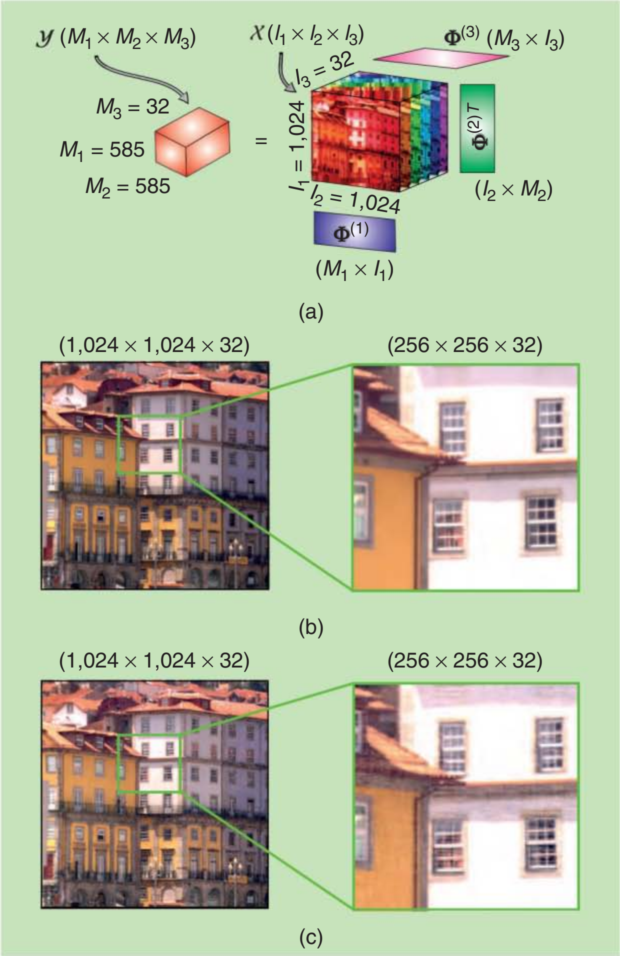
Since the ℓ_0-norm minimization is not practical, alternative solutions involve iterative refinements of the estimates of vector $\boldsymbol{g}$ using greedy algorithms such as the orthogonal matching pursuit (OMP) algorithm, or the ℓ_1-norm minimization algorithms $(\|\boldsymbol{g}\|_1 = \sum_{i=1}^{I} |g_i|)$ [83]. Low coherence of the composite dictionary matrix $\mathbf{W}$ is a prerequisite for a satisfactory recovery of $\boldsymbol{g}$ (and hence $\boldsymbol{x}$)—we need to choose $\boldsymbol{\Phi}$ and $\mathbf{B}$ so that the correlation between the columns of $\mathbf{W}$ is minimum [83].

When extending the CS framework to tensor data, we face two obstacles:

- loss of information, such as spatial and contextual relationships in data, when a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is vectorized.



[FIG7] **CS with a Kronecker-structured dictionary. OMP can perform faster if the sparse entries belong to a small subtensor, up to permutation of the columns of $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, and $\mathbf{W}^{(3)}$.**



[FIG8] **The multidimensional CS of a 3-D hyperspectral image using Tucker representation with a small sparse core in wavelet bases. (a) The Kronecker-CS of a 32-channel hyperspectral image. (b) The original hyperspectral image-RGB display. (c) The reconstruction (SP = 33%, PSNR = 35.51 dB)-RGB display.**

- data handling since the size of vectorized data and the associated dictionary $\mathbf{B} \in \mathbb{R}^{I \times I}$ easily becomes prohibitively large (see the section "Large-Scale Data and the Curse of Dimensionality"), especially for tensors of high order.

Fortunately, tensor data are typically highly structured, a perfect match for compressive sampling, so that the CS framework relaxes data acquisition requirements, enables compact storage, and facilitates data completion (i.e., inpainting of missing samples due to a faulty sensor or unreliable measurement).

KRONECKER-CS FOR FIXED DICTIONARIES

In many applications, the dictionary and the sensing matrix admit a Kronecker structure (Kronecker-CS model), as illustrated in Figure 7(a) [84]. In this way, the global composite dictionary matrix becomes $\mathbf{W} = \mathbf{W}^{(N)} \otimes \mathbf{W}^{(N-1)} \otimes \cdots \otimes \mathbf{W}^{(1)}$, where each term $\mathbf{W}^{(n)} = \boldsymbol{\Phi}^{(n)} \mathbf{B}^{(n)}$ has a reduced dimensionality since $\mathbf{B}^{(n)} \in \mathbb{R}^{I_n \times I_n}$ and $\boldsymbol{\Phi}^{(n)} \in \mathbb{R}^{M_n \times I_n}$. Denote $M = M_1 M_2 \cdots M_N$ and $I = I_1 I_2 \cdots I_N$, then, since $M_n \leq I_n$, $n = 1, 2, \ldots, N$, this reduces storage requirements by a factor of $(\Sigma_n I_n M_n)/(MI)$. The computation of $\mathbf{W}\boldsymbol{g}$ is affordable since $\boldsymbol{g}$ is sparse; however, computing $\mathbf{W}^T \boldsymbol{y}$ is expensive but can be efficiently implemented through a sequence of products involving much smaller matrices $\mathbf{W}^{(n)}$ [85]. We refer to [84] for links between the coherence of factor matrices $\mathbf{W}^{(n)}$ and the coherence of the global composite dictionary matrix $\mathbf{W}$.

Figure 7 and Table 3 illustrate that the Kronecker-CS model is effectively a vectorized TKD with a sparse core. The tensor equivalent of the CS paradigm in (13) is therefore to find the sparsest core tensor $\mathcal{G}$ such that

$$\mathcal{Y} \cong \mathcal{G} \times_1 \mathbf{W}^{(1)} \times_2 \mathbf{W}^{(2)} \cdots \times_N \mathbf{W}^{(N)}, \qquad (14)$$

with $\|\mathcal{G}\|_0 \leq K$, for a given set of modewise dictionaries $\mathbf{B}^{(n)}$ and sensing matrices $\boldsymbol{\Phi}^{(n)}$ $(n = 1, 2, \ldots, N)$. Working with several small dictionary matrices, appearing in a Tucker representation, instead of a large global dictionary matrix, is an example of the use of tensor structure for efficient representation; see also the section "Large-Scale Data and the Curse of Dimensionality."

A higher-order extension of the OMP algorithm, referred to as the *Kronecker-OMP algorithm* [85], requires K iterations to find the K nonzero entries of the core tensor $\mathcal{G}$. Additional computational advantages can be gained if it can be assumed that the K nonzero entries belong to a small subtensor of $\mathcal{G}$, as shown in Figure 7(b); such a structure is inherent to, e.g., hyperspectral imaging [85], [86] and 3-D astrophysical signals. More precisely, if the $K = L^N$ nonzero entries are located within a subtensor of size $(L \times L \times \cdots \times L)$, where $L \ll I_n$, then, by exploiting the block-tensor structure, the so-called N-way block OMP algorithm (N-BOMP) requires at most NL iterations, which is linear in N

[85]. The Kronecker-CS model has been applied in magnetic resonance imaging, hyperspectral imaging, and in the inpainting of multiway data [86], [84].

APPROACHES WITHOUT FIXED DICTIONARIES

In Kronecker-CS, the modewise dictionaries $\mathbf{B}^{(n)} \in \mathbb{R}^{I_n \times I_n}$ can be chosen so as best to represent the physical properties or prior knowledge about the data. They can also be learned from a large ensemble of data tensors, for instance, in an ALS-type fashion [86]. Instead of the total number of sparse entries in the core tensor, the size of the core (i.e., the multilinear rank) may be used as a measure for sparsity so as to obtain a low-complexity representation from compressively sampled data [87], [88]. Alternatively, a CPD representation can be used instead of a Tucker representation. Indeed, early work in chemometrics involved excitation–emission data for which part of the entries was unreliable because of scattering; the CPD of the data tensor is then computed by treating such entries as missing [7]. While CS variants of several CPD algorithms exist [59], [89], the oracle properties of tensor-based models are still not as well understood as for their standard models; a notable exception is CPD with sparse factors [90].

EXAMPLE 3

Figure 8 shows an original 3-D ($1{,}024 \times 1{,}024 \times 32$) hyperspectral image $\mathcal{X}$, which contains scene reflectance measured at 32 different frequency channels, acquired by a low-noise Peltier-cooled digital camera in the wavelength range of 400–720 nm [91]. Within the Kronecker-CS setting, the tensor of compressive measurements $\mathcal{Y}$ was obtained by multiplying the frontal slices by random Gaussian sensing matrices $\mathbf{\Phi}^{(1)} \in \mathbb{R}^{M_1 \times 1024}$ and $\mathbf{\Phi}^{(2)} \in \mathbb{R}^{M_2 \times 1024}$ $(M_1, M_2 < 1{,}024)$ in the first and second mode, respectively, while $\mathbf{\Phi}^{(3)} \in \mathbb{R}^{32 \times 32}$ was the identity matrix [see Figure 8(a)]. We used Daubechies wavelet factor matrices $\mathbf{B}^{(1)} = \mathbf{B}^{(2)} \in \mathbb{R}^{1024 \times 1024}$ and $\mathbf{B}^{(3)} \in \mathbb{R}^{32 \times 32}$, and employed the N-way block tensor N-BOMP to recover the small sparse core tensor and, subsequently, reconstruct the original 3-D image, as shown in Figure 8(b). For the sampling ratio SP=33% $(M_1 = M_2 = 585)$ this gave the peak SNR (PSNR) of 35.51 dB, while taking 71 min for $N_{\text{iter}} = 841$ iterations needed to detect the subtensor which contains the most significant entries. For the same quality of reconstruction (PSNR=35.51 dB), the more conventional Kronecker-OMP algorithm found 0.1% of the wavelet coefficients as significant, thus requiring $N_{\text{iter}} = K = 0.001 \times (1{,}024 \times 1{,}024 \times 32) = 33{,}555$ iterations and days of computation time.

LARGE-SCALE DATA AND THE CURSE OF DIMENSIONALITY

The sheer size of tensor data easily exceeds the memory or saturates the processing capability of standard computers; it is, therefore, natural to ask ourselves how tensor decompositions can be computed if the tensor dimensions in all or some modes are large or, worse still, if the tensor order is high. The term *curse of dimensionality*, in a general sense, was introduced by Bellman to refer to various computational bottlenecks when dealing with high-dimensional settings. In the context of tensors, the curse of dimensionality refers to the fact that the number of elements of an Nth-order $(I \times I \times \cdots \times I)$ tensor, I^N, scales exponentially with the tensor order N. For example, the number of values of a discretized function in Figure 2(b) quickly becomes unmanageable in terms of both computations and storing as N increases. In addition to their standard use (signal separation, enhancement, etc.), tensor decompositions may be elegantly employed in this context as efficient representation tools. The first question is, which type of tensor decomposition is appropriate?

[TABLE 4] STORAGE COST OF TENSOR MODELS FOR AN Nth-ORDER TENSOR $\mathcal{X} \in \mathbb{R}^{I \times I \times \cdots \times I}$ FOR WHICH THE STORAGE REQUIREMENT FOR RAW DATA IS $O(I^N)$.

1) CANONICAL POLYADIC DECOMPOSITION	$O(NIR)$
2) TUCKER	$O(NIR + R^N)$
3) TENSOR TRAIN	$O(NIR^2)$
4) QUANTIZED TENSOR TRAIN	$O(NR^2 \log_2(I))$

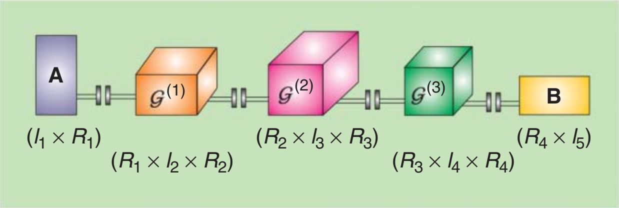
EFFICIENT DATA HANDLING

If all computations are performed on a CP representation and not on the raw data tensor itself, then, instead of the original I^N raw data entries, the number of parameters in a CP representation reduces to NIR, which scales linearly in N (see Table 4). This effectively bypasses the curse of dimensionality, while giving us the freedom to choose the rank, R, as a function of the desired accuracy [16]; on the other hand, the CP approximation may involve numerical problems (see the section "Canonical Polyadic Decomposition").

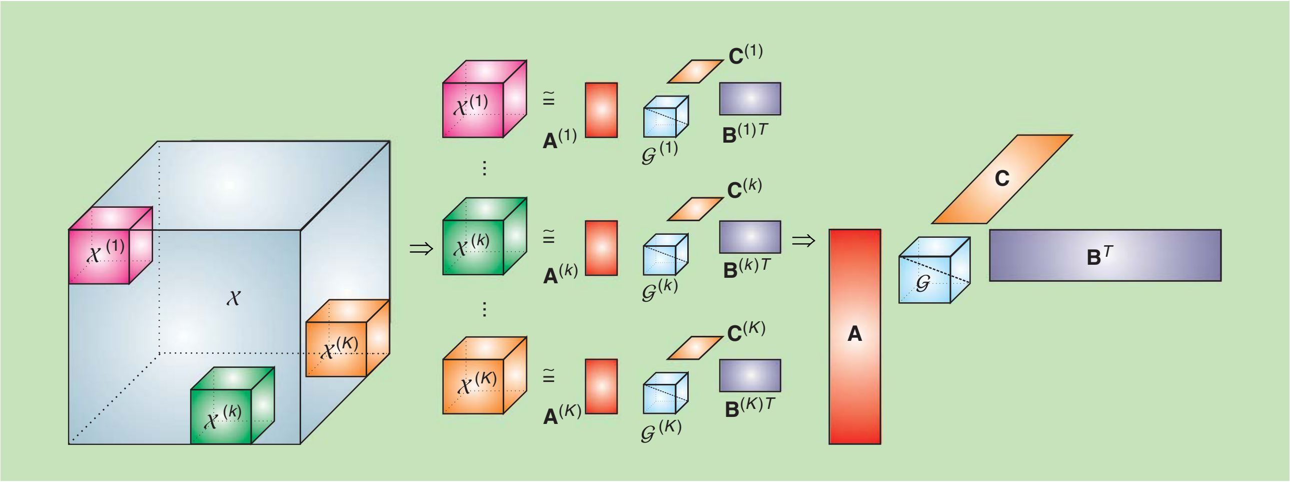
Compression is also inherent to TKD as it reduces the size of a given data tensor from the original I^N to $(NIR + R^N)$, thus exhibiting an approximate compression ratio of $(I/R)^N$. We can then benefit from the well understood and reliable approximation by means of matrix SVD; however, this is only useful for low N.

TENSOR NETWORKS

A numerically reliable way to tackle curse of dimensionality is through a concept from scientific computing and quantum information theory, termed *tensor networks*, which represents a tensor of a possibly very high order as a set of sparsely interconnected matrices and core tensors of low order (typically, order 3). These low-dimensional cores are interconnected via tensor contractions to provide a highly compressed representation of a data tensor. In addition, existing algorithms for the approximation of a given tensor by a tensor network have good numerical properties, making it



[FIG9] The TT decomposition of a fifth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_5}$, consisting of two matrix carriages and three third-order tensor carriages. The five carriages are connected through tensor contractions, which can be expressed in a scalar form as $x_{i_1,i_2,i_3,i_4,i_5} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_5=1}^{R_5} a_{i_1,r_1} g^{(1)}_{r_1,i_2,r_2} g^{(2)}_{r_2,i_3,r_3} g^{(3)}_{r_3,i_4,r_5} b_{r_4,i_5}$.



[FIG10] Efficient computation of CPD and TKD, whereby tensor decompositions are computed in parallel for sampled blocks. These are then merged to obtain the global components **A**, **B**, and **C**, and a core tensor $\mathcal{G}$.

possible to control the error and achieve any desired accuracy of approximation. For example, tensor networks allow for the representation of a wide class of discretized multivariate functions even in cases where the number of function values is larger than the number of atoms in the universe [23], [29], [30].

Examples of tensor networks are the hierarchical TKD and tensor trains (TTs) (see Figure 9) [17], [18]. The TTs are also known as matrix product states and have been used by physicists for more than two decades (see [92] and [93] and references therein). The PARATREE algorithm was developed in signal processing and follows a similar idea; it uses a polyadic representation of a data tensor (in a possibly nonminimal number of terms), whose computation then requires only the matrix SVD [94].

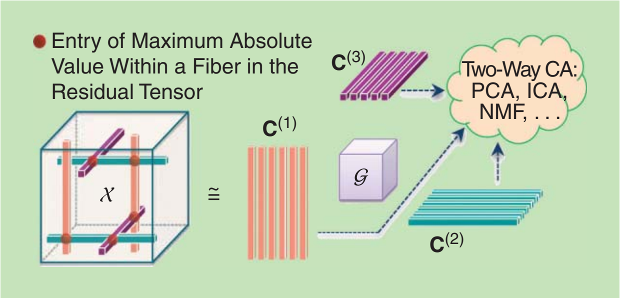
For very large-scale data that exhibit a well-defined structure, an even more radical approach to achieve a parsimonious representation may be through the concept of quantized or quantic tensor networks (QTNs) [29], [30]. For example, a huge vector $x \in \mathbb{R}^I$ with $I = 2^L$ elements can be quantized and tensorized into a $(2 \times 2 \times \cdots \times 2)$ tensor $\mathcal{X}$ of order L, as illustrated in Figure 2(a). If x is an exponential signal, $x(k) = az^k$, then $\mathcal{X}$ is a symmetric rank-1 tensor that can be represented by two parameters: the scaling factor a and the generator z (cf. (2) in the section "Tensorization—Blessing of Dimensionality"). Nonsymmetric terms provide further opportunities, beyond the sum-of-exponential representation by symmetric low-rank tensors. Huge matrices and tensors may be dealt with in the same manner. For instance, an Nth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, with $I_n = q^{L_n}$, can be quantized in all modes simultaneously to yield a $(q \times q \times \cdots \times q)$ quantized tensor of higher order. In QTN, q is small, typically $q = 2,3,4$, e.g., the binary encoding $(q = 2)$ reshapes an Nth -order tensor with $(2^{L_1} \times 2^{L_2} \times \cdots \times 2^{L_N})$ elements into a tensor of order $(L_1 + L_2 + \cdots + L_N)$ with the same number of elements. The TT decomposition applied to quantized tensors is referred to as the *quantized TT* (*QTT*); variants for other tensor representations have also been derived [29], [30]. In scientific computing, such formats provide the so-called supercompression—a logarithmic reduction of storage requirements: $O(I^N) \rightarrow O(N \log_q(I))$.

COMPUTATION OF THE DECOMPOSITION/REPRESENTATION

Now that we have addressed the possibilities for efficient tensor representation, the question that needs to be answered is how these representations can be computed from the data in an efficient manner. The first approach is to process the data in smaller blocks rather than in a batch manner [95]. In such a divide-and-conquer approach, different blocks may be processed in parallel, and their decompositions may be carefully recombined (see Figure 10) [95], [96]. In fact, we may even compute the decomposition through recursive updating as new data arrive [97]. Such recursive techniques may be used for efficient computation and for tracking decompositions in the case of nonstationary data.

The second approach would be to employ CS ideas (see the section "Higher-Order Compressed Sensing (HO-CS)") to fit an algebraic model with a limited number of parameters to possibly large data. In addition to enabling data completion (interpolation of missing data), this also provides a significant reduction of the cost of data acquisition, manipulation, and storage, breaking the curse of dimensionality being an extreme case.

While algorithms for this purpose are available both for low-rank and low multilinear rank representation [59], [87], an even more drastic approach would be to directly adopt sampled fibers as the bases in a tensor representation. In the TKD setting, we would choose the columns of the factor matrices $\mathbf{B}^{(n)}$ as mode-n fibers of the tensor, which requires us to address the following two problems: 1) how to find fibers that allow us to accurately represent the tensor and 2) how to compute the corresponding core tensor at a low cost (i.e., with minimal access to the data). The matrix counterpart of this problem (i.e., representation of a large matrix on the basis of a few columns and rows) is referred to as the *pseudoskeleton approximation* [98], where the optimal representation corresponds to the columns and rows that intersect in the submatrix of maximal volume (maximal absolute value of the determinant). Finding the optimal submatrix is computationally hard, but quasioptimal submatrices may be found by heuristic so-called cross-approximation methods that only require a limited, partial exploration of the data matrix. Tucker variants of this approach have been derived in [99]–[101] and are illustrated in Figure 11, while a cross-approximation for the TT format has been derived in [102]. Following a somewhat different idea, a tensor generalization of the CUR decomposition of matrices samples fibers on the basis of statistics derived from the data [103].


● Entry of Maximum Absolute Value Within a Fiber in the Residual Tensor
$\mathcal{X}$
≅
$\mathbf{C}^{(1)}$
$\mathbf{C}^{(2)}$
$\mathbf{C}^{(3)}$
$\mathcal{G}$
Two-Way CA: PCA, ICA, NMF, . . .

[FIG11] The Tucker representation through fiber sampling and cross-approximation: the columns of factor matrices are sampled from the fibers of the original data tensor $\mathcal{X}$. Within MWCA, the selected fibers may be further processed using BSS algorithms.

MULTIWAY REGRESSION—HIGHER-ORDER PARTIAL LS

MULTIVARIATE REGRESSION

Regression refers to the modeling of one or more dependent variables (responses), Y, by a set of independent data (predictors), X. In the simplest case of conditional mean square estimation (MSE), whereby $\hat{y} = E(y\,|\,x)$, the response y is a linear combination of the elements of the vector of predictors $\mathbf{x}$; for multivariate data, the multivariate linear regression (MLR) uses a matrix model, $\mathbf{Y} = \mathbf{XP} + \mathbf{E}$, where $\mathbf{P}$ is the matrix of coefficients (loadings) and $\mathbf{E}$ is the residual matrix. The MLR solution gives $\mathbf{P} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and involves inversion of the moment matrix $\mathbf{X}^T\mathbf{X}$. A common technique to stabilize the inverse of the moment matrix $\mathbf{X}^T\mathbf{X}$ is the principal component regression (PCR), which employs low-rank approximation of $\mathbf{X}$.

MODELING STRUCTURE IN DATA—THE PARTIAL LS

Note that in stabilizing multivariate regression, PCR uses only information in the X variables, with no feedback from the Y variables. The idea behind the partial LS (PLS) method is to account for structure in data by assuming that the underlying system is governed by a small number, R, of specifically constructed latent variables, called scores, that are shared between the X and Y variables; in estimating the number R, PLS compromises between fitting $\mathbf{X}$ and predicting $\mathbf{Y}$. Figure 12 illustrates that the PLS procedure: 1) uses eigenanalysis to perform contraction of the data matrix $\mathbf{X}$ to the principal eigenvector score matrix $\mathbf{T} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_R]$ of rank R and 2) ensures that the $\boldsymbol{t}_r$ components are maximally correlated with the $\boldsymbol{u}_r$ components in the approximation of the responses $\mathbf{Y}$, this is achieved when the $\mathbf{u}_r$'s are scaled versions of the $\boldsymbol{t}_r$'s. The Y-variables are then regressed on the matrix $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R]$. Therefore, PLS is a multivariate model with inferential ability that aims to find a representation of $\mathbf{X}$ (or a part of $\mathbf{X}$) that is relevant for predicting $\mathbf{Y}$, using the model

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \sum_{r=1}^{R} \boldsymbol{t}_r \boldsymbol{p}_r^T + \mathbf{E}, \tag{15}$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} = \sum_{r=1}^{R} \boldsymbol{u}_r \boldsymbol{q}_r^T + \mathbf{F}. \tag{16}$$

The score vectors $\boldsymbol{t}_r$ provide an LS fit of $\mathbf{X}$-data, while at the same time, the maximum correlation between $\boldsymbol{t}$ and $\boldsymbol{u}$ scores ensures a good predictive model for Y variables. The predicted responses $\mathbf{Y}_{\text{new}}$ are then obtained from new data $\mathbf{X}_{\text{new}}$ and the loadings $\mathbf{P}$ and $\mathbf{Q}$.

In practice, the score vectors $\boldsymbol{t}_r$, are extracted sequentially, by a series of orthogonal projections followed by the deflation of $\mathbf{X}$. Since the rank of $\mathbf{Y}$ is not necessarily decreased with each new $\boldsymbol{t}_r$, we may continue deflating until the rank of the $\mathbf{X}$-block is exhausted so as to balance between prediction accuracy and model order.

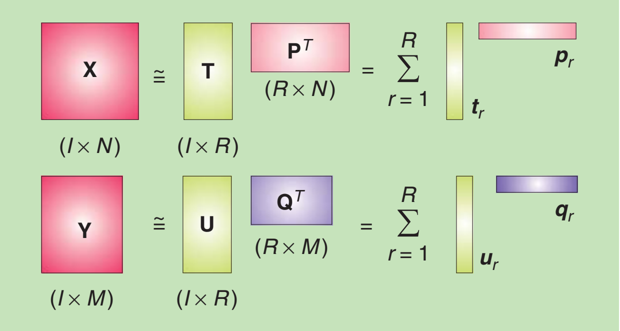
The PLS concept can be generalized to tensors in the following ways:

1) *Unfolding multiway data*. For example, tensors $\mathcal{X}(I \times J \times K)$ and $\mathcal{Y}(I \times M \times N)$ can be flattened into long matrices $\mathbf{X}(I \times JK)$ and $\mathbf{Y}(I \times MN)$ so as to admit matrix-PLS (see Figure 12). However, such flattening prior to standard bilinear PLS obscures the structure in multiway data and compromises the interpretation of latent components.

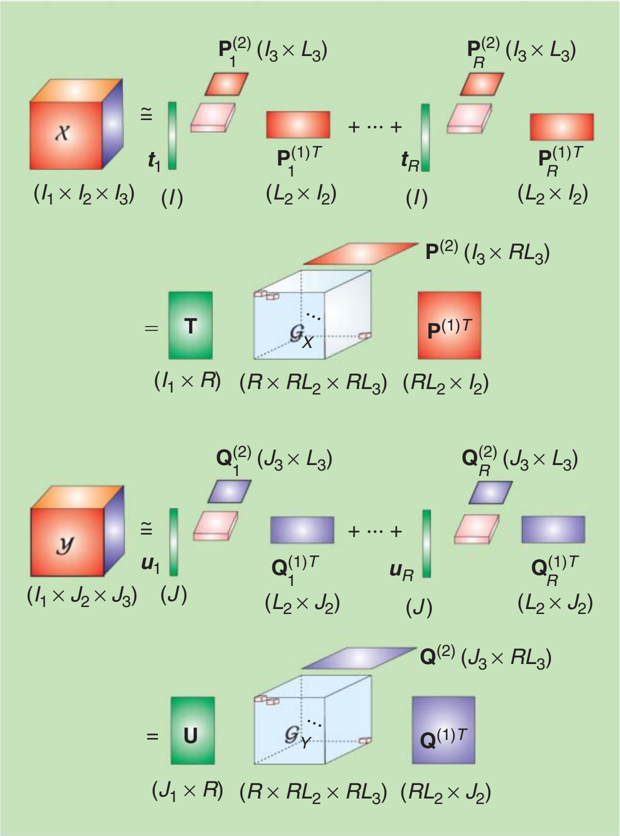
2) *Low-rank tensor approximation*. The so-called N-PLS attempts to find score vectors having maximal covariance with response variables, under the constraints that tensors $\mathcal{X}$ and $\mathcal{Y}$ are decomposed as a sum of rank-1 tensors [104].

3) *A BTD-type approximation*. As in the higher-order PLS (HOPLS) model shown in Figure 13 [105], the use of block terms within HOPLS equips it with additional flexibility, together with a more physically meaningful analysis than unfolding-PLS and N-PLS.

The principle of HOPLS can be formalized as a set of sequential approximate decompositions of the independent tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and the dependent tensor $\mathcal{Y} \in \mathbb{R}^{J_1 \times J_2 \times \cdots \times J_M}$ (with $I_1 = J_1$) so as to ensure maximum similarity (correlation) between the scores $\boldsymbol{t}_r$ and $\boldsymbol{u}_r$ within the matrices $\mathbf{T}$ and $\mathbf{U}$, based on



[FIG12] The basic PLS model performs joint sequential low-rank approximation of the matrix of predictors $\mathbf{X}$ and the matrix of responses $\mathbf{Y}$ so as to share (up to the scaling ambiguity) the latent components—columns of the score matrices $\mathbf{T}$ and $\mathbf{U}$. The matrices $\mathbf{P}$ and $\mathbf{Q}$ are the loading matrices for predictors and responses, and $\mathbf{E}$ and $\mathbf{F}$ are the corresponding residual matrices.



[FIG13] The principle of HOPLS for third-order tensors. The core tensors $\mathcal{G}_{\mathrm{X}}$ and $\mathcal{G}_{\mathrm{Y}}$ are block-diagonal. The BTD-type structure allows for the modeling of general components that are highly correlated in the first mode.

$$\mathcal{X} \cong \sum_{r=1}^{R} \mathcal{G}_{\mathrm{X}}^{(r)} \times_1 \boldsymbol{t}_r \times_2 \mathbf{P}_r^{(1)} \cdots \times_N \mathbf{P}_r^{(N-1)} \tag{17}$$

$$\mathcal{Y} \cong \sum_{r=1}^{R} \mathcal{G}_{\mathrm{Y}}^{(r)} \times_1 \boldsymbol{u}_r \times_2 \mathbf{Q}_r^{(1)} \cdots \times_N \mathbf{Q}_r^{(M-1)}. \tag{18}$$

A number of data-analytic problems can be reformulated as either regression or similarity analysis [analysis of variance (ANOVA), autoregressive moving average modeling (ARMA), linear discriminant analysis (LDA), and canonical correlation analysis (CCA)], so that both the matrix and tensor PLS solutions can be generalized across exploratory data analysis.

EXAMPLE 4

The predictive power of tensor-based PLS is illustrated on a real-world example of the prediction of arm movement trajectory from the electrocorticogram (ECoG). Figure 14(a) illustrates the experimental setup, whereby the 3-D arm movement of a monkey was captured by an optical motion capture system with reflective markers affixed to the left shoulder, elbow, wrist, and hand; for full details, see http://neurotycho.org. The predictors (32 ECoG channels) naturally build a fourth-order tensor $\mathcal{X}$ (time × channel_no × epoch_length × frequency) while the movement trajectories for the four markers (response) can be represented as a third-order tensor $\mathcal{Y}$ (time × 3D_marker_position × marker_no). The goal of the training stage is to identify the HOPLS parameters: $\mathcal{G}_{\mathrm{X}}^{(r)}, \mathcal{G}_{\mathrm{Y}}^{(r)}, \mathbf{P}_r^{(n)}, \mathbf{Q}_r^{(n)}$ (see Figure 13). In the test stage, the movement trajectories, $\mathcal{Y}^*$, for the new ECoG data, $\mathcal{X}^*$, are predicted through multilinear projections: 1) the new scores, $\boldsymbol{t}_r^*$, are found from new data, $\mathcal{X}^*$, and the existing model parameters: $\mathcal{G}_X^{(r)}, \mathbf{P}_r^{(1)}, \mathbf{P}_r^{(2)}, \mathbf{P}_r^{(3)}$, and 2) the predicted trajectory is calculated as $\mathcal{Y}^* \approx \sum_{r=1}^{R} \mathcal{G}_{\mathrm{Y}}^{(r)} \times_1 \boldsymbol{t}_r^* \times_2 \mathbf{Q}_r^{(1)} \times_3 \mathbf{Q}_r^{(2)} \times_4 \mathbf{Q}_r^{(3)}$. In the simulations, standard PLS was applied in the same way to the unfolded tensors.

Figure 14(c) shows that although the standard PLS was able to predict the movement corresponding to each marker individually, such a prediction is quite crude as the two-way PLS does not adequately account for mutual information among the four markers. The enhanced predictive performance of the BTD-based HOPLS [the red line in Figure 14(c)] is therefore attributed to its ability to model interactions between complex latent components of both predictors and responses.

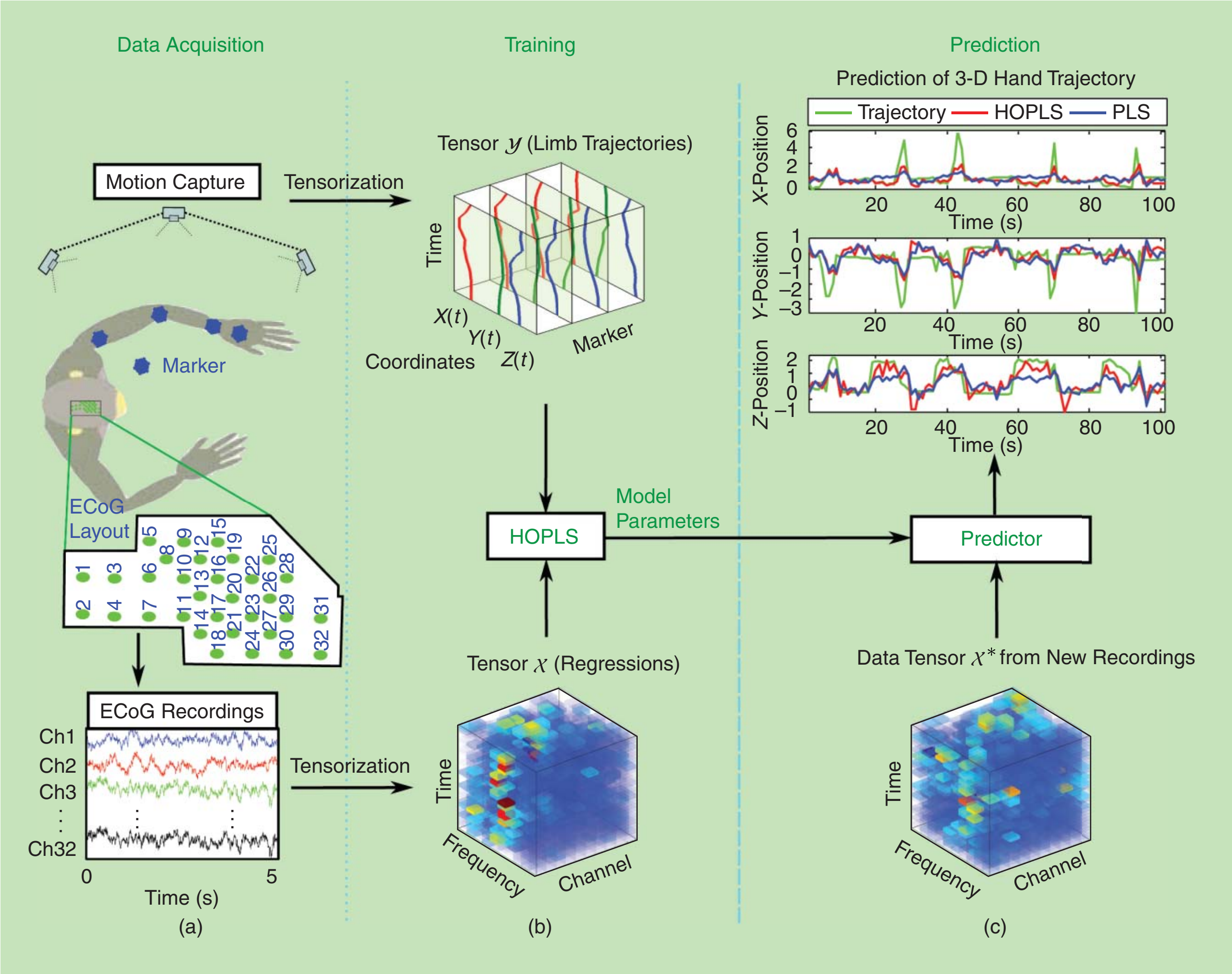
LINKED MULTIWAY COMPONENT ANALYSIS AND TENSOR DATA FUSION

Data fusion concerns the joint analysis of an ensemble of data sets, such as multiple views of a particular phenomenon, where some parts of the scene may be visible in only one or a few data sets. Examples include the fusion of visual and thermal images in low-visibility conditions and the analysis of human electrophysiological signals in response to a certain stimulus but from different subjects and trials; these are naturally analyzed together by means of matrix/tensor factorizations. The coupled nature of the analysis of such multiple data sets ensures that we are able to account for the common factors across the data sets and, at the same time, to guarantee that the individual components are not shared (e.g., processes that are independent of excitations or stimuli/tasks).

The linked multiway component analysis (LMWCA) [106], shown in Figure 15, performs such a decomposition into shared and individual factors and is formulated as a set of approximate joint TKD of a set of data tensors $\mathcal{X}^{(k)} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $(k = 1, 2, \ldots, K)$

$$\mathcal{X}^{(k)} \cong \mathcal{G}^{(k)} \times_1 \mathbf{B}^{(1,k)} \times_2 \mathbf{B}^{(2,k)} \cdots \times_N \mathbf{B}^{(N,k)}, \tag{19}$$

where each factor matrix $\mathbf{B}^{(n,k)} = [\mathbf{B}_C^{(n)}, \mathbf{B}_I^{(n,k)}] \in \mathbb{R}^{I_n \times R_n}$ has 1) components $\mathbf{B}_C^{(n)} \in \mathbb{R}^{I_n \times C_n}$ (with $0 \le C_n \le R_n$) that are common (i.e., maximally correlated) to all tensors and 2) components $\mathbf{B}_I^{(n,k)} \in \mathbb{R}^{I_n \times (R_n - C_n)}$ that are tensor specific. The objective is to estimate the common components $\mathbf{B}_C^{(n)}$, the individual components $\mathbf{B}_I^{(n,k)}$, and, via the core tensors $\mathcal{G}^{(k)}$, their mutual interactions. As in MWCA (see the section "Tucker Decomposition"), constraints may be imposed to match data properties [73], [76]. This enables a more general and flexible framework than group ICA and independent vector analysis, which also performs linked analysis of multiple data sets but assume that 1) there exist only common components and 2) the corresponding latent variables are statistically independent [107], [108]. Both are quite stringent and limiting assumptions. As an alternative to TKD, coupled tensor decompositions may be of a polyadic or even block term type [89], [109].



[FIG14] **The prediction of arm movement from brain electrical responses. (a) The experiment setup. (b) The construction of the data and response tensors and training. (c) The new data tensor (bottom) and the predicted 3-D arm movement trajectories (*X, Y, Z* coordinates) obtained by tensor-based HOPLS and standard matrix-based PLS (top).**

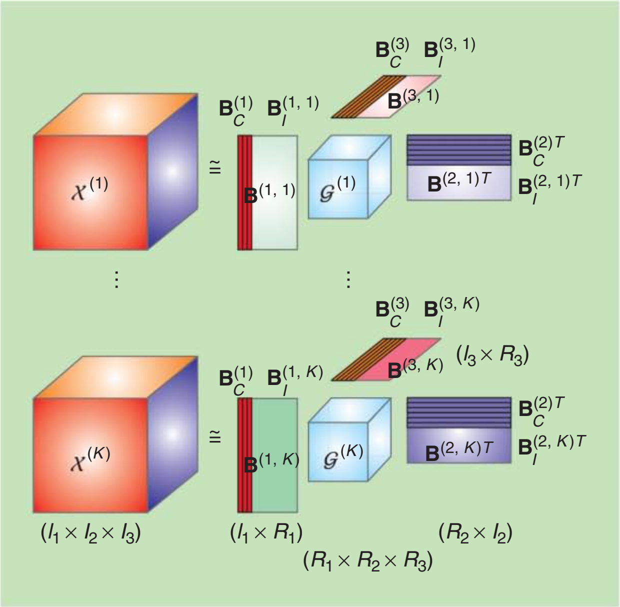
EXAMPLE 5

We employed LWCA for classification based on common and distinct features of natural objects from the ETH-80 database (http://www.d2.mpi-inf.mpg.de/Data sets/ETH80) whereby the discrimination among objects was performed using only the common features. This data set consists of 3,280 images in eight categories, each containing ten objects with 41 views per object. For each category, the training data were organized in two distinct fourth-order $(128 \times 128 \times 3 \times I_4)$ tensors, where $I_4 = 10 \times 41 \times 0.5p$, where p denotes the fraction of training data. LMWCA was applied to these two tensors to find the common and individual features, with the number of common features set to 80% of I_4. In this way, eight sets of common features were obtained for each category. The test sample label was assigned to the category whose common features matched the new sample best (evaluated by canonical correlations) [110]. Figure 16 compares LMWCA with the standard K-nearest neighbors (K-NNs) and LDA classifiers (using 50 principal components as features), all averaged over 50 Monte Carlo runs. The enhanced classification results for LMWCA are attributed to the fact that the classification makes use of only the common components and is not hindered by components that are not shared across objects or views.

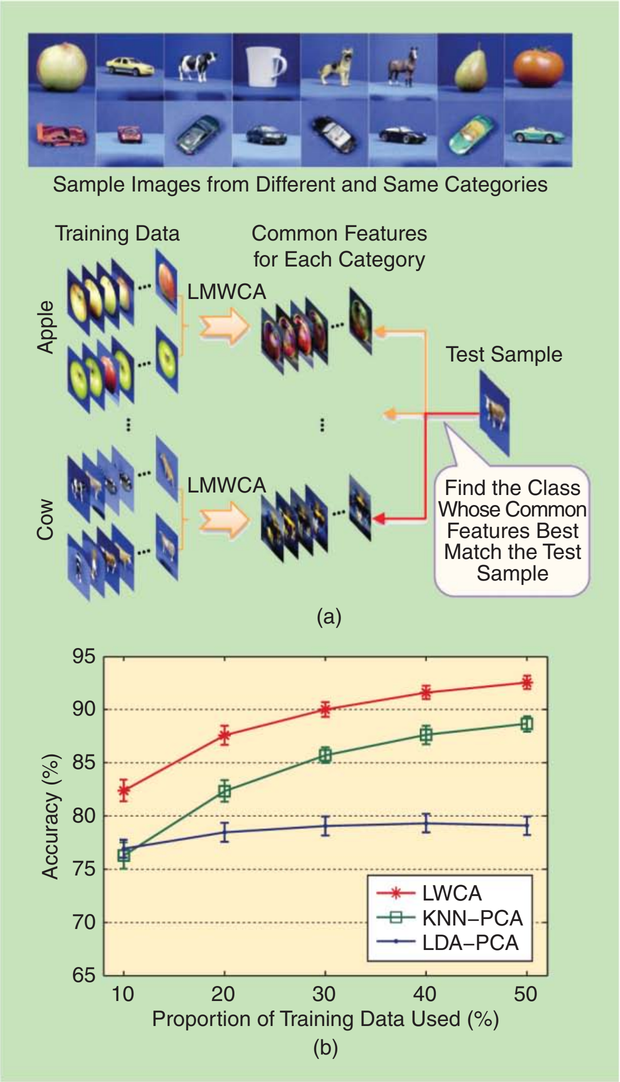
SOFTWARE

The currently available software resources for tensor decompositions include:

- The tensor toolbox, a versatile framework for basic operations on sparse and dense tensors, including CPD and Tucker formats [111].
- The TDALAB and TENSORBOX, which provide a user-friendly interface and advanced algorithms for CPD, nonnegative TKD, and MWCA [112], [113].
- The Tensorlab toolbox builds upon the complex optimization framework and offers numerical algorithms for computing the CPD, BTD, and TKD; the toolbox includes a library of constraints (e.g., nonnegativity and orthogonality) and the possibility to combine and jointly factorize dense, sparse, and incomplete tensors [89].



[FIG15] **Coupled TKD for LMWCA. The data tensors have both shared and individual components. Constraints such as orthogonality, statistical independence, sparsity, and nonnegativity may be imposed where appropriate.**



[FIG16] **The classification of color objects belonging to different categories. By using only common features, LMWCA achieves a high classification rate, even when the training set is small. (a) Classification based on LMWCA. (b) Performance comparison.**

- The *N*-way toolbox, which includes (constrained) CPD, TKD, and PLS in the context of chemometrics applications [114]; many of these methods can handle constraints (e.g., nonnegativity and orthogonality) and missing elements.
- The TT toolbox, the Hierarchical Tucker toolbox, and the Tensor Calculus library provide tensor tools for scientific computing [115]–[117].
- Code developed for multiway analysis is also available from the Three-Mode Company [118].

CONCLUSIONS AND FUTURE DIRECTIONS

We live in a world overwhelmed by data, from multiple pictures of Big Ben on various social Web links to terabytes of data in multiview medical imaging, while we may also need to repeat the scientific experiments many times to obtain the ground truth. Each snapshot gives us a somewhat incomplete view of the same object and involves different angles, illumination, lighting conditions, facial expressions, and noise.

We have shown that tensor decompositions are a perfect match for exploratory analysis of such multifaceted data sets and have illustrated their applications in multisensor and multimodal signal processing. Our emphasis has been to show that tensor decompositions and multilinear algebra open up completely new possibilities for component analysis, as compared with the flat view of standard two-way methods.

Unlike matrices, tensors are multiway arrays of data samples whose representations are typically overdetermined (fewer parameters in the decomposition than the number of data entries). This gives us an enormous flexibility in finding hidden components in data and the ability to enhance both robustness to noise and tolerance to missing data samples and faulty sensors. We have also discussed multilinear variants of several standard signal processing tools such as multilinear SVD, ICA, NMF, and PLS and have shown that tensor methods can operate in a deterministic way on signals of very short duration.

At present, the uniqueness conditions of standard tensor models are relatively well understood and efficient computation algorithms do exist. However, for future applications, several challenging problems remain to be addressed in more depth.

- A whole new area emerges when several decompositions that operate on different data sets are coupled, as in multiview data where some details of interest are visible in, e.g., only one mode. Such techniques need theoretical support in terms of existence, uniqueness, and numerical properties.
- As the complexity of advanced models increases, their computation requires efficient iterative algorithms, extending beyond the ALS class.

- The estimation of the number of components in data and the assessment of their dimensionality would benefit from automation, especially in the presence of noise and outliers.
- Both new theory and algorithms are needed to further extend the flexibility of tensor models, e.g., for the constraints to be combined in many ways and tailored to the particular signal properties in different modes.
- Work on efficient techniques for saving and/or fast processing of ultra-large-scale tensors is urgent; these now routinely occupy terabytes, and will soon require petabytes of memory.
- Tools for rigorous performance analysis and rule of thumb performance bounds need to be further developed across tensor decomposition models.
- Our discussion has been limited to tensor models in which all entries take values independently of one another. Probabilistic versions of tensor decompositions incorporate prior knowledge about complex variable interaction, various data alphabets, or noise distributions, and so promise to model data more accurately and efficiently [119], [120].
- The future computational, visualization, and interpretation tools will be important next steps in supporting the different communities working on large-scale and big data analysis problems.

It is fitting to conclude with a quote from the French novelist Marcel Proust: “The voyage of discovery is not in seeking new landscapes but in having new eyes.” We hope to have helped to bring to the eyes of the signal processing community the multidisciplinary developments in tensor decompositions and to have shared our enthusiasm about tensors as powerful tools to discover new landscapes.

AUTHORS

Andrzej Cichocki (cia@brain.riken.jp) received the Ph.D. and Dr.Sc. (habilitation) degrees all in electrical engineering from the Warsaw University of Technology, Poland. He is currently a senior team leader of the Laboratory for Advanced Brain Signal Processing at RIKEN Brain Science Institute, Japan, and a professor at the Systems Research Institute, Polish Academy of Science, Poland. He has authored more than 400 publications and four monographs in the areas of signal processing and computational neuroscience. He is an associate editor of *IEEE Transactions on Signal Processing* and *Journal of Neuroscience Methods*.

Danilo P. Mandic (d.mandic@imperial.ac.uk) is a professor of signal processing at Imperial College London, United Kingdom, and has been working in the area of nonlinear and multidimensional adaptive signal processing and time-frequency analysis. His publication record includes two research monographs, *Recurrent Neural Networks for Prediction* and *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*, an edited book, *Signal Processing for Information Fusion*, and more than 200 publications on signal and image processing. He has been a guest professor at KU Leuven, Belgium, and a frontier researcher at RIKEN Brain Science Institute, Tokyo, Japan.

Anh Huy Phan (phan@brain.riken.jp) received the Ph.D. degree from the Kita Kyushu Institute of Technology, Japan, in 2011. He worked as a deputy head of the Research and Development Department, Broadcast Research and Application Center, Vietnam Television, and is currently a research scientist at the Laboratory for Advanced Brain Signal Processing and a visiting research scientist at the Toyota Collaboration Center, RIKEN Brain Science Institute, Japan. He has served on the editorial board of *International Journal of Computational Mathematics*. His research interests include multilinear algebra, tensor computation, blind source separation, and brain–computer interfaces.

Cesar F. Caiafa (ccaiafa@gmail.com) received the Ph.D. degree in engineering from the Faculty of Engineering, University of Buenos Aires, in 2007. He is currently an adjunct researcher with the Argentinean Radioastronomy Institute (IAR)—CONICET and an assistant professor with Faculty of Engineering, the University of Buenos Aires. He is also a visiting scientist at the Laboratory for Advanced Brain Signal Processing, BSI—RIKEN, Japan.

Guoxu Zhou (zhouguoxu@brain.riken.jp) received the Ph.D. degree in intelligent signal and information processing from the South China University of Technology, Guangzhou, in 2010. He is currently a research scientist at the Laboratory for Advanced Brain Signal Processing at RIKEN Brain Science Institute, Japan. His research interests include statistical signal processing, tensor analysis, intelligent information processing, and machine learning.

Qibin Zhao (qbzhao@brain.riken.jp) received the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China, in 2009. He is currently a research scientist at the Laboratory for Advanced Brain Signal Processing in RIKEN Brain Science Institute, Japan, and a visiting research scientist in the BSI Toyota Collaboration Center, RIKEN-BSI. His research interests include multiway data analysis, brain–computer interface, and machine learning.

Lieven De Lathauwer (Lieven.DeLathauwer@kuleuven-kulak.be) received the Ph.D. degree from the Faculty of Engineering, KU Leuven, Belgium, in 1997. From 2000 to 2007, he was a research associate with the Centre National de la Recherche Scientifique, France. He is currently a professor with KU Leuven. He is affiliated with the group Science, Engineering, and Technology of Kulak, the Stadius Center for Dynamical Systems, Signal Processing, and Data Analytics of the Electrical Engineering Department (ESAT), and iMinds Future Health Department. He is an associate editor of *SIAM Journal on Matrix Analysis and Applications* and was an associate editor of *IEEE Transactions on Signal Processing*. His research focuses on the development of tensor tools for engineering applications.

REFERENCES

[1] F. L. Hitchcock, “Multiple invariants and generalized rank of a p-way matrix or tensor,” *J. Math. Phys.*, vol. 7, no. 1, pp. 39–79, 1927.

[2] R. Cattell, “Parallel proportional profiles and other principles for determining the choice of factors by rotation,” *Psychometrika*, vol. 9, pp. 267–283, 1944.

[3] L. R. Tucker, “The extension of factor analysis to three-dimensional matrices,” in *Contributions to Mathematical Psychology*, H. Gulliksen and N. Frederiksen, Eds. New York: Holt, Rinehart and Winston, 1964, pp. 110–127.

[4] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sept. 1966.

[5] J. Carroll and J.-J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of ‘Eckart-Young’ decomposition,” *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sept. 1970.

[6] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis," *UCLA Working Pap. Phonet.*, vol. 16, pp. 1–84, 1970.

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

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

[9] C. Nikias and A. Petropulu, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[10] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," in *IEE Proc. F (Radar and Signal Processing)*, vol. 140, no. 6, IET, 1993, pp. 362–370.

[11] P. Comon, "Independent component analysis: A new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[12] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. New York, Academic, 2010.

[13] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. Hoboken, NJ: Wiley, 2003.

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[15] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[16] G. Beylkin and M. Mohlenkamp, "Algorithms for numerical analysis in high dimensions," *SIAM J. Sci. Comput.*, vol. 26, no. 6, pp. 2133–2159, 2005.

[17] J. Ballani, L. Grasedyck, and M. Kluge, "Black box approximation of tensors in hierarchical Tucker format," *Linear Algebr. Appl.*, vol. 433, no. 2, pp. 639–657, 2011.

[18] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.

[19] N. Sidiropoulos, R. Bro, and G. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[20] N. Sidiropoulos, G. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Processing*, vol. 48, no. 3, pp. 810–823, 2000.

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

[22] J. Landsberg, *Tensors: Geometry and Applications*. AMS, 2012.

[23] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus* (ser. Springer series in computational mathematics). Heidelberg: Springer, 2012, vol. 42.

[24] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Trans. Knowledge Data Eng.*, vol. 21, no. 1, pp. 6–20, 2009.

[25] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Sept. 2009.

[26] P. Comon, X. Luciani, and A. L. F. de Almeida, "Tensor decompositions, alternating least squares and other tales," *J. Chemomet.*, vol. 23, no. 7–8, pp. 393–405, 2009.

[27] H. Lu, K. Plataniotis, and A. Venetsanopoulos, "A survey of multilinear subspace learning for tensor data," *Pattern Recognit.*, vol. 44, no. 7, pp. 1540–1551, 2011.

[28] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdisc. Rew.: Data Mining Knowled. Discov.*, vol. 1, no. 1, pp. 24–40, 2011.

[29] B. Khoromskij, "Tensors-structured numerical methods in scientific computing: Survey on recent advances," *Chemomet. Intell. Lab. Syst.*, vol. 110, no. 1, pp. 1–19, 2011.

[30] L. Grasedyck, D. Kressner, and C. Tobler, "A literature survey of low-rank tensor approximation techniques," *CGAMM-Mitteilungen*, vol. 36, no. 1, pp. 53–78, 2013.

[31] P. Comon, "Tensors: A brief introduction," *IEEE Signal Processing Mag.*, vol. 31, no. 3, pp. 44–53, May 2014.

[32] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[33] J. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebr. Appl.*, vol. 18, no. 2, pp. 95–138, 1977.

[34] I. Domanov and L. De Lathauwer, "On the uniqueness of the canonical polyadic decomposition of third-order tensors—Part I: Basic results and uniqueness of one factor matrix and part II: Uniqueness of the overall decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 855–903, 2013.

[35] M. Elad, P. Milanfar, and G. H. Golub, "Shape from moments—An estimation theory perspective," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1814–1829, 2004.

[36] N. Sidiropoulos, "Generalizing Caratheodory's uniqueness of harmonic parameterization to N dimensions," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1687–1690, 2001.

[37] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and É. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[38] F. Miwakeichi, E. Martnez-Montes, P. Valds-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, "Decomposing EEG data into space–time–frequency components using parallel factor analysis," *NeuroImage*, vol. 22, no. 3, pp. 1035–1045, 2004.

[39] M. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. European Conf. on Computer Vision (ECCV)*, Copenhagen, Denmark, May 2002, vol. 2350, pp. 447–460.

[40] M. Hirsch, D. Lanman, G.Wetzstein, and R. Raskar, "Tensor displays," in *Proc. Int. Conf. Computer Graphics and Interactive Techniques, SIGGRAPH 2012*, Los Angeles, CA, USA, Aug. 5-9, 2012, *Emerging Technologies Proc.*, 2012, pp. 24–42.

[41] J. Håstad, "Tensor rank is NP-complete," *J. Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[42] M. Timmerman and H. Kiers, "Three mode principal components analysis: Choosing the numbers of components and sensitivity to local optima," *Br. J. Math. Stat. Psychol.*, vol. 53, no. 1, pp. 1–16, 2000.

[43] E. Ceulemans and H. Kiers, "Selecting among three-mode principal component models of different types and complexities: A numerical convex-hull based method," *Br. J. Math Stat Psychol.*, vol. 59, no. 1, pp. 133–150, May 2006.

[44] M. Mørup and L. K. Hansen, "Automatic relevance determination for multiway models," *J. Chemomet., Special Issue: In Honor of Professor Richard A. Harshman*, vol. 23, no. 7–8, pp. 352–363, 2009.

[45] N. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemomet.*, vol. 14, no. 3, pp. 229–239, 2000.

[46] T. Jiang and N. D. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models," *IEEE Trans. Signal Processing*, vol. 52, no. 9, pp. 2625–2636, 2004.

[47] L. De Lathauwer, "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 28, no. 3, pp. 642–666, 2006.

[48] A. Stegeman, "On uniqueness conditions for CANDECOMP/PARAFAC and INDSCAL with full column rank in one mode," *Linear Algebr. Appl.*, vol. 431, no. 1–2, pp. 211–227, 2009.

[49] E. Sanchez and B. Kowalski, "Tensorial resolution: A direct trilinear decomposition," *J. Chemomet.*, vol. 4, no. 1, pp. 29–45, 1990.

[50] I. Domanov and L. De Lathauwer, "Canonical polyadic decomposition of third-order tensors: Reduction to generalized eigenvalue decomposition," *SIAM Anal. Appl.*, vol. 35, no. 2, pp. 636–660, 2014.

[51] S. Vorobyov, Y. Rong, N. Sidiropoulos, and A. Gershman, "Robust iterative fitting of multilinear models," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2678–2689, 2005.

[52] X. Liu and N. Sidiropoulos, "Cramér-Rao lower bounds for low-rank decomposition of multidimensional arrays," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 2074–2086, Sept. 2001.

[53] P. Tichavský, A.-H. Phan, and Z. Koldovský, "Cramér-Rao-induced bounds for CANDECOMP/PARAFAC tensor decomposition," *IEEE Trans. Signal Processing*, vol. 61, no. 8, pp. 1986–1997, 2013.

[54] B. Chen, S. He, Z. Li, and S. Zhang, "Maximum block improvement and polynomial optimization," *SIAM J. Optim.*, vol. 22, no. 1, pp. 87–107, 2012.

[55] A. Uschmajew, "Local convergence of the alternating least squares algorithm for canonical tensor approximation," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 2, pp. 639–652, 2012.

[56] M. J. Mohlenkamp, "Musings on multilinear fitting," *Linear Algebr. Appl.*, vol. 438, no. 2, pp. 834–852, 2013.

[57] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[58] P. Paatero, "The multilinear engine: A table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model," *J. Computat. Graph. Stat.*, vol. 8, no. 4, pp. 854–888, Dec. 1999.

[59] E. Acar, D. Dunlavy, T. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemomet. Intell. Lab. Syst.*, vol. 106, no. 1, pp. 41–56, 2011.

[60] A.-H. Phan, P. Tichavský, and A. Cichocki, "Low complexity damped Gauss-Newton algorithms for CANDECOMP/PARAFAC," *SIAM J. Matrix Anal. Appl. (SIMAX)*, vol. 34, no. 1, pp. 126–147, 2013.

[61] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical Polyadic Decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms and a new generalization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 695–720, 2013.

[62] V. de Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 1084–1127, Sept. 2008.

[63] W. Krijnen, T. Dijkstra, and A. Stegeman, "On the non-existence of optimal solutions and the occurrence of "degeneracy" in the CANDECOMP/PARAFAC model," *Psychometrika*, vol. 73, no. 3, pp. 431–439, 2008.

[64] M. Sørensen, L. De Lathauwer, P. Comon, S. Icart, and L. Deneire, "Canonical Polyadic Decomposition with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 4, pp. 1190–1213, 2012.

[65] M. Sørensen and L. De Lathauwer, "Blind signal separation via tensor decomposition with Vandermonde factor: Canonical polyadic decomposition," *IEEE Trans. Signal Processing*, vol. 61, no. 22, pp. 5507–5519, Nov. 2013.

[66] G. Zhou and A. Cichocki, "Canonical Polyadic Decomposition based on a single mode blind source separation," *IEEE Signal Processing Lett.*, vol. 19, no. 8, pp. 523–526, 2012.

[67] L.-H. Lim and P. Comon, "Nonnegative approximations of nonnegative tensors," *J. Chemomet.*, vol. 23, nos. 7–8, pp. 432–441, 2009.

[68] A. van der Veen and A. Paulraj, "An analytical constant modulus algorithm," *IEEE Trans. Signal Processing*, vol. 44, no. 5, pp. 1136–1155, 1996.

[69] 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, 1989.

[70] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-$R_1, R_2, \ldots, R_N$ approximation of higher-order tensors," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[71] B. Savas and L.-H. Lim, "Quasi-Newton methods on Grassmannians and multilinear approximations of tensors," *SIAM J. Sci. Comput.*, vol. 32, no. 6, pp. 3352–3393, 2010.

[72] M. Ishteva, P.-A. Absil, S. Van Huffel, and L. De Lathauwer, "Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme," *SIAM J. Matrix Anal. Appl.*, vol. 32, no. 1, pp. 115–135, 2011.

[73] G. Zhou and A. Cichocki, "Fast and unique Tucker decompositions via multiway blind source separation," *Bull. Polish Acad. Sci.*, vol. 60, no. 3, pp. 389–407, 2012.

[74] A. Cichocki, "Generalized component analysis and blind source separation methods for analyzing mulitchannel brain signals," in *Statistical and Process Models for Cognitive Neuroscience and Aging*. Lawrence Erlbaum Associates, 2007, pp. 201–272.

[75] M. Haardt, F. Roemer, and G. D. Galdo, "Higher-order SVD based subspace estimation to improve the parameter estimation accuracy in multi-dimensional harmonic retrieval problems," *IEEE Trans. Signal Processing*, vol. 56, no. 7, pp. 3198–3213, July 2008.

[76] A.-H. Phan and A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional data sets," *Nonlinear Theory Appl., IEICE*, vol. 1, no. 1, pp. 37–68, 2010.

[77] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms—Part I and II," *SIAM J. Matrix Anal. Appl. (SIMAX) Special Issue on Tensor Decompositions and Applications*, vol. 30, no. 3, pp. 1022–1066, 2008.

[78] L. De Lathauwer, "Blind separation of exponential polynomials and the decomposition of a tensor in rank-$(L_r, L_r, 1)$ terms," *SIAM J. Matrix Anal. Appl.*, vol. 32, no. 4, pp. 1451–1474, 2011.

[79] L. De Lathauwer, "Block component analysis: A new concept for blind source separation," in *Proc. 10th Int. Conf. LVA/ICA, Tel Aviv*, Israel, Mar. 12–15, 2012, pp. 1–8.

[80] E. 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, 2006.

[81] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[82] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[83] Y. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, vol. 20. New York: Cambridge Univ. Press, 2012, p. 12.

[84] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Processing*, vol. 21, no. 2, pp. 494–504, 2012.

[85] C. Caiafa and A. Cichocki, "Computing sparse representations of multidimensional signals using Kronecker bases," *Neural Computat.*, vol. 25, no. 1, pp. 186–220, 2013.

[86] C. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *WIREs Data Mining Knowled. Discov.*, vol. 3, no. 6, pp. 355–380, 2013.

[87] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Prob.*, vol. 27, no. 2, pp. 1–19, 2011.

[88] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. K. Suykens, "Learning with tensors: A framework based on convex optimization and spectral regularization," *Mach. Learn.*, vol. 94, no. 3, pp. 303–351, Mar. 2014.

[89] L. Sorber, M. Van Barel, and L. De Lathauwer. (2014, Jan.). Tensorlab v2.0. [Online]. Available: www.tensorlab.net

[90] N. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Processing Lett.*, vol. 19, no. 11, pp. 757–760, 2012.

[91] D. Foster, K. Amano, S. Nascimento, and M. Foster, "Frequency of metamerism in natural scenes," *J. Opt. Soc. Amer. A*, vol. 23, no. 10, pp. 2359–2372, 2006.

[92] A. Cichocki, "Era of big data processing: A new approach via tensor networks and tensor decompositions (invited talk)," in *Proc. 2013 Int. Workshop on Smart Info-Media Systems in Asia, SISA-2013*, Nagoya, Japan, Oct. 1, 2013, 2013, 30 pages. [Online]. Available: http://arxiv.org/pdf/1403.2048.pdf

[93] R. Orus, "A practical introduction to tensor networks: Matrix product states and projected entangled pair states," *J. Chem. Phys.*, 2013.

[94] J. Salmi, A. Richter, and V. Koivunen, "Sequential unfolding SVD for tensors with applications in array signal processing," *IEEE Trans. Signal Processing*, vol. 57, no. 12, pp. 4719–4733, 2009.

[95] A.-H. Phan and A. Cichocki, "PARAFAC algorithms for large-scale problems," *Neurocomputing*, vol. 74, no. 11, pp. 1970–1984, 2011.

[96] S. K. Suter, M. Makhynia, and R. Pajarola, "TAMRESH: Tensor approximation multiresolution hierarchy for interactive volume visualization," *Comput. Graph. Forum*, vol. 32, no. 3, pp. 151–160, 2013.

[97] D. Nion and N. Sidiropoulos, "Adaptive algorithms to track the PARAFAC decomposition of a third-order tensor," *IEEE Trans. Signal Processing*, vol. 57, no. 6, pp. 2299–2310, June 2009.

[98] S. A. Goreinov, N. L. Zamarashkin, and E. E. Tyrtyshnikov, "Pseudo-skeleton approximations by matrices of maximum volume," *Math. Notes*, vol. 62, no. 4, pp. 515–519, 1997.

[99] C. Caiafa and A. Cichocki, "Generalizing the column-row matrix decomposition to multi-way arrays," *Linear Algebr. Appl.*, vol. 433, no. 3, pp. 557–573, 2010.

[100] S. A. Goreinov, "On cross approximation of multi-index array," *Doklady Math.*, vol. 420, no. 4, pp. 404–406, 2008.

[101] I. Oseledets, D. V. Savostyanov, and E. Tyrtyshnikov, "Tucker dimensionality reduction of three-dimensional arrays in linear time," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 939–956, 2008.

[102] I. Oseledets and E. Tyrtyshnikov, "TT-cross approximation for multidimensional arrays," *Linear Algebr. Appl.*, vol. 432, no. 1, pp. 70–88, 2010.

[103] M. W. Mahoney, M. Maggioni, and P. Drineas, "Tensor-CUR decompositions for tensor-based data," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 957–987, 2008.

[104] R. Bro, "Multiway calibration. Multilinear PLS," *J. Chemomet.*, vol. 10, no. 1, pp. 47–61, 1996.

[105] Q. Zhao, C. Caiafa, D. Mandic, Z. Chao, Y. Nagasaka, N. Fujii, L. Zhang, and A. Cichocki, "Higher-order partial least squares (HOPLS): A generalized multilinear regression method," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 35, no. 7, pp. 1660–1673, 2013.

[106] A. Cichocki, "Tensors decompositions: New concepts for brain data analysis?" *J. Control, Measure*., Syst. Integr. (SICE), vol. 47, no. 7, pp. 507–517, 2011.

[107] V. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *Neuroimage*, vol. 45, pp. 163–172, 2009.

[108] Y.-O. Li, T. Adali, W. Wang, and V. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Trans. Signal Processing*, vol. 57, no. 10, pp. 3918–3929, Oct. 2009.

[109] E. Acar, T. Kolda, and D. Dunlavy, "All-at-once optimization for coupled matrix and tensor factorizations," in *Proc. Mining and Learning with Graphs, (MLG'11)*, San Diego, CA, August 2011.

[110] G. Zhou, A. Cichocki, S. Xie, and D. Mandic. (2013). Beyond canonical correlation analysis: Common and individual features analysis. *IEEE Trans. Pattern Anal. Mach. Intell*. [Online]. Available: http://arxiv.org/abs/1212.3913

[111] B. Bader, T. G. Kolda et al. (2012, Feb.). MATLAB tensor toolbox version 2.5. [Online]. Available: http://www.sandia.gov/ tgkolda/TensorToolbox/

[112] G. Zhou and A. Cichocki. (2013). TDALAB: Tensor decomposition laboratory, LABSP, Wako-shi, Japan. [Online]. Available: http://bsp.brain.riken.jp/TDALAB/

[113] A.-H. Phan, P. Tichavský, and A. Cichocki. (2012). TENSORBOX: A MATLAB package for tensor decomposition, LABSP, RIKEN, Japan. [Online]. Available: http://www.bsp.brain.riken.jp/ phan/tensorbox.php

[114] C. Andersson and R. Bro. (2000). The N-way toolbox for MATLAB. [Online]. *Chemomet. Intell. Lab. Syst.*, 52(1), pp. 1–4, 2000. Available: http://www.models.life.ku.dk/nwaytoolbox

[115] I. Oseledets. (2012). TT-toolbox 2.2. [Online]. Available: https://github.com/oseledets/TT-Toolbox

[116] D. Kressner and C. Tobler. (2012). htucker—A MATLAB toolbox for tensors in hierarchical Tucker format. MATHICSE, EPF Lausanne. [Online]. Available: http://anchp.epfl.ch/htucker

[117] M. Espig, M. Schuster, A. Killaitis, N. Waldren, P. Wähnert, S. Handschuh, and H. Auer. (2012). Tensor calculus library. [Online]. Available: http://gitorious.org/tensorcalculus

[118] P. Kroonenberg. The three-mode company: A company devoted to creating three-mode software and promoting three-mode data analysis. [Online]. Available: http://three-mode.leidenuniv.nl/.

[119] Z. Xu, F. Yan, and A. Qi, "Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis," in *Proc. 29th Int. Conf. Machine Learning (ICML-12), ser. ICML'12*. Omnipress, July 2012, pp. 1023–1030.

[120] K. Yilmaz and A. T. Cemgil, "Probabilistic latent tensor factorisation," in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation, cPCI-S*, 2010, vol. 6365, pp. 346–353.

[SP]

Dave Zachariah
and Petre Stoica

Cramér–Rao Bound Analog of Bayes' Rule

The estimation of multiple parameters is a common task in signal processing. The Cramér–Rao bound (CRB) sets a statistical lower limit on the resulting errors when estimating parameters from a set of random observations. It can be understood as a fundamental measure of parameter uncertainty [1], [2]. As a general example, suppose $\boldsymbol{\theta}$ denotes the vector of sought parameters and that the random observation model can be written as

$$\mathbf{y} = \mathbf{x}_\theta + \mathbf{w}, \tag{1}$$

where $\mathbf{x}_\theta$ is a function or signal parameterized by $\boldsymbol{\theta}$ and $\mathbf{w}$ is a zero-mean Gaussian noise vector. Then the CRB for $\boldsymbol{\theta}$ has the following notable properties:

1) For a fixed $\boldsymbol{\theta}$, the CRB for $\boldsymbol{\theta}$ decreases as the dimension of $\mathbf{y}$ increases.

2) For a fixed $\mathbf{y}$, if additional parameters $\tilde{\boldsymbol{\theta}}$ are estimated, then the CRB for $\boldsymbol{\theta}$ increases as the dimension of $\tilde{\boldsymbol{\theta}}$ increases.

3) If adding a set of observations $\tilde{\mathbf{y}}$ requires estimating additional parameters $\tilde{\boldsymbol{\theta}}$, then the CRB for $\boldsymbol{\theta}$ decreases as the dimension of $\tilde{\mathbf{y}}$ increases, provided the dimension of $\tilde{\boldsymbol{\theta}}$ does not exceed that of $\tilde{\mathbf{y}}$ [3]. This property implies both 1) and 2) above.

4) Among all possible distributions of $\mathbf{w}$ with a fixed covariance matrix, the CRB for $\boldsymbol{\theta}$ attains its maximum when $\mathbf{w}$ is Gaussian, i.e., the Gaussian scenario is the "worst case" for estimating $\boldsymbol{\theta}$ [4]–[6].

In this lecture note, we show a general property of the CRB that quantifies the interdependencies between the parameters in $\boldsymbol{\theta}$. The presented result is valid for more general models than (1) and also generalizes the result in [7] to vector parameters. It will be illustrated via two examples.

Digital Object Identifier 10.1109/MSP.2014.2365593
Date of publication: 12 February 2015

RELEVANCE

In probability theory, the chain rule and Bayes' rule are useful tools to analyze the statistical interdependence between multiple random variables and to derive tractable expressions for their distributions. In this lecture note, we provide analogs of the chain rule and Bayes' rule for the CRB associated with multiple parameters. The results are particularly useful when estimating parameters of interest in the presence of nuisance parameters.

PREREQUISITIES

The reader needs basic knowledge about linear algebra, elementary probability theory, and statistical signal processing.

PRELIMINARIES

We will consider a general scenario in which we observe an $n \times 1$ random vector $\mathbf{y}$. Its probability density function (pdf) $p(\mathbf{y};\boldsymbol{\theta})$ is parameterized by a $k \times 1$ deterministic vector $\boldsymbol{\theta}$. The goal is to estimate $\boldsymbol{\theta}$, or subvectors of $\boldsymbol{\theta}$, given $\mathbf{y}$.

Let $l(\boldsymbol{\theta}) \triangleq ln\, p(\mathbf{y};\boldsymbol{\theta})$ denote the log-likelihood function, and let $\hat{\boldsymbol{\theta}}$ be any unbiased estimator. Then the mean square error (MSE) matrix $P_{\hat{\theta}} \triangleq \mathrm{E}[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^*]$ is bounded from below by the inverse of the Fisher information matrix $\mathbf{J}_\theta \triangleq -\mathrm{E}[\partial_\theta^2 l(\boldsymbol{\theta})]$, where ∂_θ^2 denotes the second-order differential or Laplacian operator with respect to $\boldsymbol{\theta}$. That is, $\mathbf{P}_{\hat{\theta}} \succeq \mathbf{J}_\theta^{-1}$, assuming from hereon that $\mathbf{J}_\theta$ is nonsingular. This is the Cramér–Rao inequality [2], [8], [9].

The determinant of the MSE matrix, $|\mathbf{P}_{\hat{\theta}}|$, is a scalar measure of the error magnitude. For unbiased estimators, $|\mathbf{P}_{\hat{\theta}}|$ equals the "generalized variance" of errors [10]. By defining $\mathrm{CRB}(\boldsymbol{\theta}) \triangleq |\mathbf{J}_\theta^{-1}|$, the generalized error variance is bounded by

$$|\mathbf{P}_{\hat{\theta}}| \geq \mathrm{CRB}(\boldsymbol{\theta}).$$

In the following, we are interested in subvectors or elements of $\boldsymbol{\theta}$. Letting $\boldsymbol{\theta} = [\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top]^\top$, we can write the Fisher information matrix in block form,

$$\begin{aligned}\mathbf{J}_\theta &= -\mathrm{E}\begin{bmatrix} \partial_\alpha^2 l(\boldsymbol{\alpha},\boldsymbol{\beta}) & \partial_\alpha \partial_\beta l(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ \partial_\beta \partial_\alpha l(\boldsymbol{\alpha},\boldsymbol{\beta}) & \partial_\beta^2 l(\boldsymbol{\alpha},\boldsymbol{\beta}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{J}_\alpha & \mathbf{J}_{\alpha\beta} \\ \mathbf{J}_{\beta\alpha} & \mathbf{J}_\beta \end{bmatrix}. \end{aligned} \tag{2}$$

MAIN RESULT

Let $\mathbf{a}$ and $\mathbf{b}$ be two random vectors. Two useful rules in probability theory are the chain rule

$$p(\mathbf{a},\mathbf{b}) = p(\mathbf{a}\,|\,\mathbf{b})\,p(\mathbf{b}) \tag{3}$$

and Bayes' rule

$$p(\mathbf{a}) = \frac{p(\mathbf{b})}{p(\mathbf{b}\,|\,\mathbf{a})}\,p(\mathbf{a}\,|\,\mathbf{b}). \tag{4}$$

Now consider two parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. When both are unknown, their joint CRB bound is given by

$$\mathrm{CRB}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \left| \begin{bmatrix} \mathbf{J}_\alpha & \mathbf{J}_{\alpha\beta} \\ \mathbf{J}_{\beta\alpha} & \mathbf{J}_\beta \end{bmatrix}^{-1} \right|. \tag{5}$$

The bound for α with known β is simply

$$\mathrm{CRB}(\boldsymbol{\alpha}\,|\,\boldsymbol{\beta}) = |\mathbf{J}_\alpha^{-1}|, \tag{6}$$

and the bound for $\boldsymbol{\alpha}$ with unknown $\boldsymbol{\beta}$ is

$$\mathrm{CRB}(\boldsymbol{\alpha}) = |(\mathbf{J}_\alpha - \mathbf{J}_{\alpha\beta}\mathbf{J}_\beta^{-1}\mathbf{J}_{\beta\alpha})^{-1}|. \tag{7}$$

[Equation (7) follows by evaluating the inverse in (5) and extracting the upper-left block corresponding to $\boldsymbol{\alpha}$.] Equations (6) and (7) are the respective CRB analogs of conditional and marginal distributions for random variables.

IEEE TRANSACTIONS ON

SIGNAL AND INFORMATION PROCESSING OVER NETWORKS

Now accepting paper submissions

The new *IEEE Transactions on Signal and Information Processing over Networks* publishes high-quality papers that extend the classical notions of processing of signals defined over vector spaces (e.g. time and space) to processing of signals and information (data) defined over networks, potentially dynamically varying. In signal processing over networks, the topology of the network may define structural relationships in the data, or may constrain processing of the data. Topics of interest include, but are not limited to the following:

Adaptation, Detection, Estimation, and Learning

- Distributed detection and estimation
- Distributed adaptation over networks
- Distributed learning over networks
- Distributed target tracking
- Bayesian learning; Bayesian signal processing
- Sequential learning over networks
- Decision making over networks
- Distributed dictionary learning
- Distributed game theoretic strategies
- Distributed information processing
- Graphical and kernel methods
- Consensus over network systems
- Optimization over network systems

Communications, Networking, and Sensing

- Distributed monitoring and sensing
- Signal processing for distributed communications and networking
- Signal processing for cooperative networking
- Signal processing for network security
- Optimal network signal processing and resource allocation

Modeling and Analysis

- Performance and bounds of methods
- Robustness and vulnerability
- Network modeling and identification

Modeling and Analysis (cont.)

- Simulations of networked information processing systems
- Social learning
- Bio-inspired network signal processing
- Epidemics and diffusion in populations

Imaging and Media Applications

- Image and video processing over networks
- Media cloud computing and communication
- Multimedia streaming and transport
- Social media computing and networking
- Signal processing for cyber-physical systems
- Wireless/mobile multimedia

Data Analysis

- Processing, analysis, and visualization of big data
- Signal and information processing for crowd computing
- Signal and information processing for the Internet of Things
- Emergence of behavior

Emerging topics and applications

- Emerging topics
- Applications in life sciences, ecology, energy, social networks, economic networks, finance, social sciences, smart grids, wireless health, robotics, transportation, and other areas of science and engineering

Editor-in-Chief: Petar M. Djurić, Stony Brook University (USA)

To submit a paper, go to: https://mc.manuscriptcentral.com/tsipn-ieee

IEEE COMMUNICATIONS SOCIETY®   IEEE Signal Processing Society®   IEEE COMPUTER SOCIETY®

Digital Object Identifier 10.1109/MSP.2015.2396047

By applying the Schur determinant formula [8], [11]

$$\left|\begin{bmatrix}\mathbf{J}_{\alpha} & \mathbf{J}_{\alpha\beta}\\ \mathbf{J}_{\beta\alpha} & \mathbf{J}_{\beta}\end{bmatrix}\right| = |\mathbf{J}_{\alpha}||\mathbf{J}_{\beta}-\mathbf{J}_{\beta\alpha}\mathbf{J}_{\alpha}^{-1}\mathbf{J}_{\alpha\beta}| = |\mathbf{J}_{\beta}||\mathbf{J}_{\alpha}-\mathbf{J}_{\alpha\beta}\mathbf{J}_{\beta}^{-1}\mathbf{J}_{\beta\alpha}|,$$

along with $|\mathbf{J}^{-1}|=|\mathbf{J}|^{-1}$, to (5)–(7), we can now state the CRB analogs of the chain rule (3),

$$\mathrm{CRB}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathrm{CRB}(\boldsymbol{\alpha}\,|\,\boldsymbol{\beta})\mathrm{CRB}(\boldsymbol{\beta}) \quad (8)$$

and of Bayes' rule (4),

$$\mathrm{CRB}(\boldsymbol{\alpha}) = \frac{\mathrm{CRB}(\boldsymbol{\beta})}{\mathrm{CRB}(\boldsymbol{\beta}\,|\,\boldsymbol{\alpha})}\mathrm{CRB}(\boldsymbol{\alpha}\,|\,\boldsymbol{\beta}). \quad (9)$$

The results are, of course, symmetric, i.e., one can interchange $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

From (8) we see that the joint error bound for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ equals the error bound for $\boldsymbol{\alpha}$, when $\boldsymbol{\beta}$ is known, multiplied by the error bound for $\boldsymbol{\beta}$. More interestingly, (9) tells us that the error bound for $\boldsymbol{\alpha}$ is equal to the bound for $\boldsymbol{\alpha}$ when $\boldsymbol{\beta}$ is known, multiplied by a factor, viz. $\mathrm{CRB}(\boldsymbol{\beta})/\mathrm{CRB}(\boldsymbol{\beta}\,|\,\boldsymbol{\alpha}) \geq 1$, that quantifies the influence of $\boldsymbol{\beta}$ on one's ability to estimate $\boldsymbol{\alpha}$.

REMARK 1

The rules can be applied to cases with any number of additional parameters, besides $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Consider, for instance, the case of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is an unknown nuisance parameter. Then applying the chain rule twice yields

$$\begin{aligned}\mathrm{CRB}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) &= \mathrm{CRB}(\boldsymbol{\gamma}\,|\,\boldsymbol{\alpha},\boldsymbol{\beta})\\ &\quad\ \mathrm{CRB}(\boldsymbol{\alpha}\,|\,\boldsymbol{\beta})\mathrm{CRB}(\boldsymbol{\beta})\\ &= \mathrm{CRB}(\boldsymbol{\gamma}\,|\,\boldsymbol{\alpha},\boldsymbol{\beta})\\ &\quad\ \mathrm{CRB}(\boldsymbol{\beta}\,|\,\boldsymbol{\alpha})\mathrm{CRB}(\boldsymbol{\alpha}),\end{aligned} \quad (10)$$

where the factors without $\boldsymbol{\gamma}$ signify that the nuisance parameter is unknown. Combining the two expressions in (10) yields the analog of Bayes' rule (9) for any number of additional parameters.

The joint error bound for a set of parameters $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \ldots$ can be similarly decomposed by a recursive application of the chain rule to analyze their interdependency and its impact on estimation.

REMARK 2

The CRB analog of Bayes' rule (9) generalizes the result in [7], which concerns only scalar parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ amid a vector of nuisance parameters $\boldsymbol{\gamma}$. Our proof of (9) is also more direct than in [7].

REMARK 3

These results are also applicable to the posterior, or Bayesian, CRB (PCRB), in which $\boldsymbol{\theta}$ is modeled as a random variable with a prior distribution. The PCRB is valid for the entire class of estimators $\hat{\boldsymbol{\theta}}$, whether biased or not [2]. The posterior Cramér–Rao inequality is then $\mathbf{P}_{\hat{\theta}} \succeq \mathbf{J}_{\theta}^{-1}$, where $\mathbf{J}_{\theta} \triangleq -\mathrm{E}[\partial_{\theta}^{2} ln\, p(\mathbf{y},\boldsymbol{\theta})]$ is the Bayesian Fisher information matrix, $p(\mathbf{y},\boldsymbol{\theta})$ is the joint pdf and the expectation is with respect to this pdf. Letting $\boldsymbol{\theta} = [\boldsymbol{\alpha}^{\top}\boldsymbol{\beta}^{\top}]^{\top}$, the matrix can be partitioned correspondingly,

$$\mathbf{J}_{\theta} = \begin{bmatrix}\mathbf{J}_{\alpha} & \mathbf{J}_{\alpha\beta}\\ \mathbf{J}_{\beta\alpha} & \mathbf{J}_{\beta}\end{bmatrix},$$

and thereby the results (8)–(10) can be applied to the PCRB as well.

EXAMPLES

Next, we illustrate via two examples how a decomposition like (9) can be used for analysis. The examples show that, by quantifying the impact of nuisance parameters, it is possible to study the tradeoff between the gain of obtaining them through independent side information versus estimating them jointly with the parameters of interest.

LINEAR MIXED MODEL

Consider a linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} + \mathbf{w} \in \mathbb{R}^{n},$$

where $\mathbf{w}$ is Gaussian noise with covariance matrix $v\mathbf{I}$, and $\mathbf{x} \in \mathbb{R}^{k_x}$ and $\mathbf{z} \in \mathbb{R}^{k_z}$ are unknown parameters. The matrices are known and $\mathrm{rank}([\mathbf{A}\ \mathbf{B}]) = k_x + k_z < n$, which implies that the parameters $\mathbf{x}$ and $\mathbf{z}$ are embedded into two distinct range spaces, $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{B})$, respectively. Here $\mathcal{R}(\mathbf{A})$ denotes the linear subspace spanned by the columns of $\mathbf{A}$. Under these conditions the joint Fisher information matrix equals [9]

$$\begin{bmatrix}\mathbf{J}_{x} & \mathbf{J}_{xz} & \mathbf{J}_{xv}\\ \mathbf{J}_{zx} & \mathbf{J}_{z} & \mathbf{J}_{zv}\\ \mathbf{J}_{vx} & \mathbf{J}_{vz} & \mathbf{J}_{v}\end{bmatrix} = \frac{1}{v}\begin{bmatrix}\mathbf{A}^{\top}\mathbf{A} & \mathbf{A}^{\top}\mathbf{B} & 0\\ \mathbf{B}^{\top}\mathbf{A} & \mathbf{B}^{\top}\mathbf{B} & 0\\ 0 & 0 & \frac{n}{2v}\end{bmatrix}.$$

From this expression, we see that the bound for v is independent of that for $\mathbf{x}$ and $\mathbf{z}$. That is, $\mathrm{CRB}(\mathbf{x},\mathbf{z},v) = \mathrm{CRB}(\mathbf{x},\mathbf{z})\mathrm{CRB}(v)$. This is a CRB analog of the independence for random variables. Furthermore, we obtain $\mathrm{CRB}(\mathbf{z}\,|\,\mathbf{x}) = |\mathbf{J}_{z}^{-1}| = |v(\mathbf{B}^{\top}\mathbf{B})^{-1}| = v^{k_z}|\mathbf{B}^{\top}\mathbf{B}|^{-1}$ and $\mathrm{CRB}(\mathbf{z}) = |(\mathbf{J}_{z}-\mathbf{J}_{zx}\mathbf{J}_{x}^{-1}\mathbf{J}_{xz})^{-1}| = |v(\mathbf{B}^{\top}\mathbf{B}-\mathbf{B}^{\top}\mathbf{A}(\mathbf{A}^{\top}\mathbf{A})^{-1}\mathbf{A}^{\top}\mathbf{B})^{-1}| = v^{k_z}|\mathbf{B}^{\top}\Pi_{\mathrm{A}}^{\perp}\mathbf{B}|^{-1}$, where $\Pi_{\mathrm{A}}^{\perp}$ is the projector onto the orthogonal complement of $\mathcal{R}(\mathbf{A})$.

The increase in the error bound for $\mathbf{x}$ due to the lack of information about $\mathbf{z}$ can now be quantified using (9)

$$\mathrm{CRB}(\mathbf{x}) = \frac{|\mathbf{B}^{\top}\mathbf{B}|}{|\mathbf{B}^{\top}\Pi_{\mathrm{A}}^{\perp}\mathbf{B}|}\mathrm{CRB}(\mathbf{x}\,|\,\mathbf{z}), \quad (11)$$

where the factor $|\mathbf{B}^{\top}\Pi_{\mathrm{A}}^{\perp}\mathbf{B}|$ measures the alignment of $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{B})$. When the range spaces are orthogonal we have that $|\mathbf{B}^{\top}\Pi_{\mathrm{A}}^{\perp}\mathbf{B}| = |\mathbf{B}^{\top}\mathbf{B}|$, and by (11) the bound for $\mathbf{x}$ is unaffected by one's ignorance about $\mathbf{z}$. In scenarios where it is possible to obtain $\mathbf{z}$ through additional side-information or calibration instead of estimation, the cost can be weighed against the reduction of the error bound for $\mathbf{x}$ by the given factor $|\mathbf{B}^{\top}\Pi_{\mathrm{A}}^{\perp}\mathbf{B}|/|\mathbf{B}^{\top}\mathbf{B}|$.

This example has illustrated the interdependencies between the unknown parameters $\mathbf{x}$, $\mathbf{z}$, and v. Next we consider an example where the unknown parameters become asympotically independent as the number of samples n grows large.

SINE-WAVE FITTING

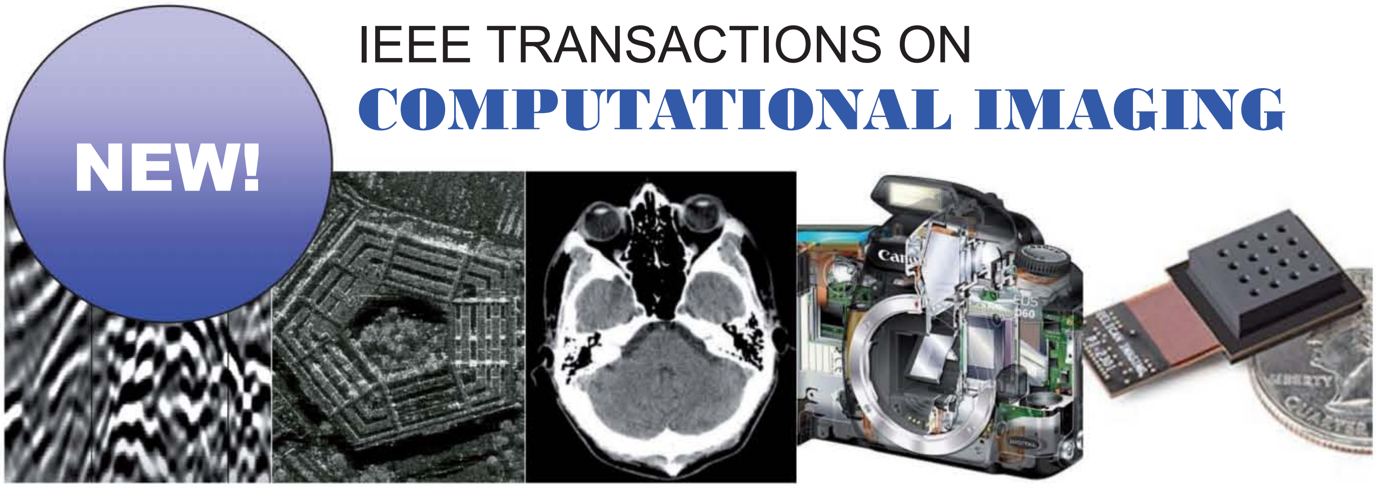
Sine-wave fitting is a problem that arises in system testing, e.g., of waveform recorders, and IEEE Standard 1057 formalizes procedures to do so (see [12] and references therein).

Consider n uniform samples of a sinusoid in noise

$$y(k) = \alpha\sin(\omega k + \phi) + C + w(k),$$

where $w(k)$ is a Gaussian white noise process with variance v and $k = 0, \ldots, n-1$. The amplitude α and phase ϕ of the sinusoidal signal, along with the offset C, are of interest. In certain cases, the frequency ω of the test signal may be obtained separately from the estimation of α, ϕ and C. For simplicity, we first consider an alternative parameterization of the sinusoid: $\alpha\sin(\omega k + \phi) = A\cos(\omega k) + B\sin(\omega k)$, where $A = \alpha\sin(\phi)$ and $B = \alpha\cos(\phi)$. The parameters are $\boldsymbol{\theta} = [A\ B\ C\ \omega\ v]^{\top}$.

IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING



The new IEEE Transactions on Computational Imaging seeks original manuscripts for publication. This new journal will publish research results where computation plays an integral role in the image formation process. All areas of computational imaging are appropriate, ranging from the principles and theory of computational imaging, to modeling paradigms for computational imaging, to image formation methods, to the latest innovative computational imaging system designs. Topics of interest include, but are not limited to the following:

Imaging Models and Representation

- Statistical-model based methods
- System and image prior models
- Noise models
- Graphical and tree-based models
- Perceptual models

Computational Sensing

- Coded source methods
- Structured light
- Coded aperture methods
- Compressed sensing
- Light-field sensing
- Plenoptic imaging
- Hardware and software systems

Computational Image Creation

- Sparsity-based methods
- Statistically-based inversion methods, Bayesian regularization
- Super-resolution, multi-image fusion
- Learning-based methods, Dictionary-based methods
- Optimization-based methods; proximal iterative methods, ADMM

Computational Photography

- Non-classical image capture, Generalized illumination
- Time-of-flight imaging
- High dynamic range imaging
- Focal stacks

Computational Consumer Imaging

- Cell phone imaging
- Camera-array systems
- Depth cameras

Computational Acoustic Imaging

- Multi-static ultrasound imaging
- Photo-acoustic imaging
- Acoustic tomography

Computational Microscopic Imaging

- Holographic microscopy
- Quantitative phase imaging
- Multi-illumination microscopy
- Lensless microscopy

Tomographic Imaging

- X-ray CT
- PET
- SPECT

Magnetic Resonance Imaging

- Diffusion tensor imaging
- Fast acquisition

Radar Imaging

- Synthetic aperture imaging
- Inverse synthetic imaging
- Terahertz imaging

Geophysical Imaging

- Multi-spectral imaging
- Ground penetrating radar
- Seismic tomography

Multi-spectral Imaging

- Multi-spectral imaging
- Hyper-spectral imaging
- Spectroscopic imaging

Editor-in-Chief: W. Clem Karl, Boston University.
To submit a paper go to: https://mc.manuscriptcentral.com/tci-ieee

Digital Object Identifier 10.1109/MSP.2015.2396046

As shown in [12], the Fisher information matrix can be decomposed into $\mathbf{J}_\theta = \bar{\mathbf{J}}_\theta + \tilde{\mathbf{J}}_\theta$, where $\bar{\mathbf{J}}_\theta$, shown in the box at the bottom of the page, contains the dominant terms and $\tilde{\mathbf{J}}_\theta$ contains the remainder, so that $\mathbf{J}_\theta^{-1} \simeq \bar{\mathbf{J}}_\theta^{-1}$ for large n. Using this approximation we now analyze the bounds for A, B, and C by application of (9).

First, let $\boldsymbol{\theta}' = [A\ B\ C\ v]^\top$ be the parameter vector without ω. Then

$$\begin{aligned} \mathrm{CRB}(\omega) &= |J_\omega - \mathbf{J}_{\omega\theta'}\mathbf{J}_{\theta'}^{-1}\mathbf{J}_{\theta'\omega}|^{-1} \\ &\simeq 2v\bigg(\frac{(A^2+B^2)n^3}{3} \\ &\quad -\frac{(A^2+B^2)n^3}{4}\bigg)^{-1} \\ &= \frac{2v}{n^3}\frac{12}{(A^2+B^2)}. \end{aligned}$$

Second, let $\boldsymbol{\theta}'' = [B\ C\ v]^\top$ be the parameter vector without ω and A. Then

$$\begin{aligned} \mathrm{CRB}(\omega\,|\,A) &= |J_\omega - \mathbf{J}_{\omega\theta''}\mathbf{J}_{\theta''}^{-1}\mathbf{J}_{\theta''\omega}|^{-1} \\ &\simeq 2v\bigg(\frac{(A^2+B^2)n^3}{3} \\ &\quad -\frac{A^2n^3}{4}\bigg)^{-1} \\ &= \frac{2v}{n^3}\frac{12}{(A^2+B^2)+3B^2}. \end{aligned}$$

Thus $\mathrm{CRB}(\omega)/\mathrm{CRB}(\omega\,|\,A) = 1 + 3B^2/(A^2+B^2) \in [1,4]$. Note that the dominant terms of $\mathbf{J}_{\theta'}$ and $\mathbf{J}_{\theta''}$ are diagonal, making their inverses particularly easy to compute. Applying (9), we obtain

$$\begin{aligned} \mathrm{CRB}(A) &\simeq \left(1+\frac{3B^2}{A^2+B^2}\right)\mathrm{CRB}(A\,|\,\omega) \\ \mathrm{CRB}(B) &\simeq \left(1+\frac{3A^2}{A^2+B^2}\right)\mathrm{CRB}(B\,|\,\omega) \\ \mathrm{CRB}(C) &\simeq \mathrm{CRB}(C\,|\,\omega), \end{aligned}$$

where the bounds for B and C are derived in a similar manner as for A. This shows that the bound for the offset C becomes independent of the knowledge of the frequency ω as n increases, while the bounds for A and B are inflated by factors ranging between one and four due to one's ignorance about ω.

When considering the original parameterization $\boldsymbol{\vartheta} = [\alpha\ \phi\ C\ \omega\ v]^\top$ there exists an invertible relation, $\boldsymbol{\vartheta} = \mathbf{g}(\boldsymbol{\theta}) = [\sqrt{A^2+B^2}\ \arctan(A/B)\ C\ \omega\ v]^\top$. Therefore we have that $\mathbf{J}_\vartheta^{-1} = \partial_\theta \mathbf{g}(\boldsymbol{\theta})\mathbf{J}_\theta^{-1}\partial_\theta \mathbf{g}(\boldsymbol{\theta})^\top$ [2], where ∂_θ denotes the first-order differential or gradient with respect to $\boldsymbol{\theta}$ and

$$\partial_\theta \mathbf{g}(\theta) = \begin{bmatrix} \sin\phi & \cos\phi & \mathbf{0} \\ \frac{\cos\phi}{\alpha} & -\frac{\sin\phi}{\alpha} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Exploiting the approximation $\mathbf{J}_\theta^{-1} \simeq \bar{\mathbf{J}}_\theta^{-1}$ once again, one obtains [12]

$$\begin{aligned} \mathrm{CRB}(\alpha) &\simeq \mathrm{CRB}(\alpha\,|\,\omega) \\ \mathrm{CRB}(\phi) &\simeq 4\mathrm{CRB}(\phi\,|\,\omega). \end{aligned}$$

This shows that, in large samples, the error bound for the amplitude α also becomes independent of knowledge about the frequency ω, whereas not knowing ω inflates the bound for the phase ϕ by a factor of four.

For large data records, the cost of precalibrating the frequency can be weighed against a reduction of the error bound for the phase, while the error bounds for the amplitude and offset will not be improved.

WHAT WE HAVE LEARNED

An analog of Bayes' rule for the CRB has been derived. This analogous rule enables a formalized decomposition and quantification of the mutual dependencies between multiple unknown parameters. The use of the rule was illustrated in two estimation problems.

$$\bar{\mathbf{J}}_\theta = \begin{bmatrix} \bar{J}_A & \bar{J}_{AB} & \bar{J}_{AC} & \bar{J}_{A\omega} & \bar{J}_{Av} \\ \bar{J}_{BA} & \bar{J}_B & \bar{J}_{BC} & \bar{J}_{B\omega} & \bar{J}_{Bv} \\ \bar{J}_{CA} & \bar{J}_{CB} & \bar{J}_C & \bar{J}_{C\omega} & \bar{J}_{Cv} \\ \bar{J}_{\omega A} & \bar{J}_{\omega B} & \bar{J}_{\omega C} & \bar{J}_\omega & \bar{J}_{\omega v} \\ \bar{J}_{vA} & \bar{J}_{vB} & \bar{J}_{vC} & \bar{J}_{v\omega} & \bar{J}_v \end{bmatrix} = \frac{1}{2v}\begin{bmatrix} n & 0 & 0 & -\frac{Bn^2}{2} & 0 \\ 0 & n & 0 & \frac{An^2}{2} & 0 \\ 0 & 0 & 2n & 0 & 0 \\ \frac{-Bn^2}{2} & \frac{An^2}{2} & 0 & \frac{(A^2+B^2)n^3}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{n}{v} \end{bmatrix}.$$

AUTHORS

Dave Zachariah (dave.zachariah@it.uu.se) received the M.S. degree in electrical engineering from Royal Institute of Technology (KTH), Stockholm, Sweden, in 2007. He received the Tech. Lic. and Ph.D. degrees in signal processing from KTH in 2011 and 2013, respectively. He is currently a postdoctoral researcher at Uppsala University in Sweden.

Petre Stoica (ps@it.uu.se) is a researcher and educator in the field of signal processing and its applications to radar/sonar, communications and biomedicine. He is a professor of signal and system modeling at Uppsala University in Sweden and a member of the Royal Swedish Academy of Engineering Sciences, the Romanian Academy (honorary), the European Academy of Sciences, and the Royal Society of Sciences in Uppsala.

REFERENCES

[1] H. Cramér, "A contribution to the theory of statistical estimation," *Scand. Actuarial J.*, vol. 1946, no. 1, pp. 85–94, 1946.

[2] H. Van Trees and K. Bell, *Detection Estimation and Modulation Theory, Part I. Detection Estimation and Modulation Theory*, 2nd ed. Hoboken, NJ: Wiley, 2013.

[3] P. Stoica and J. Li, "Study of the Cramér-Rao bound as the numbers of observations and unknown parameters increase," *IEEE Signal Process. Lett.*, vol. 3, no. 11, pp. 299–300, 1996.

[4] P. Stoica and P. Babu, "The Gaussian data assumption leads to the largest Cramér-Rao bound," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 132–133, 2011.

[5] S. Park, E. Serpedin, and K. Qaraqe, "Gaussian assumption: The least favorable but the most useful," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 183–186, 2013.

[6] M. Stein, A. Mezghani, and J. Nossek, "A lower bound for the Fisher information measure," *IEEE Signal Process. Lett.*, vol. 21, pp. 796–799, July 2014.

[7] A. D'Amico, "A "reciprocity" property of the unbiased Cramér-Rao bound for vector parameter estimation," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 615–619, 2014.

[8] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice Hall, 1988.

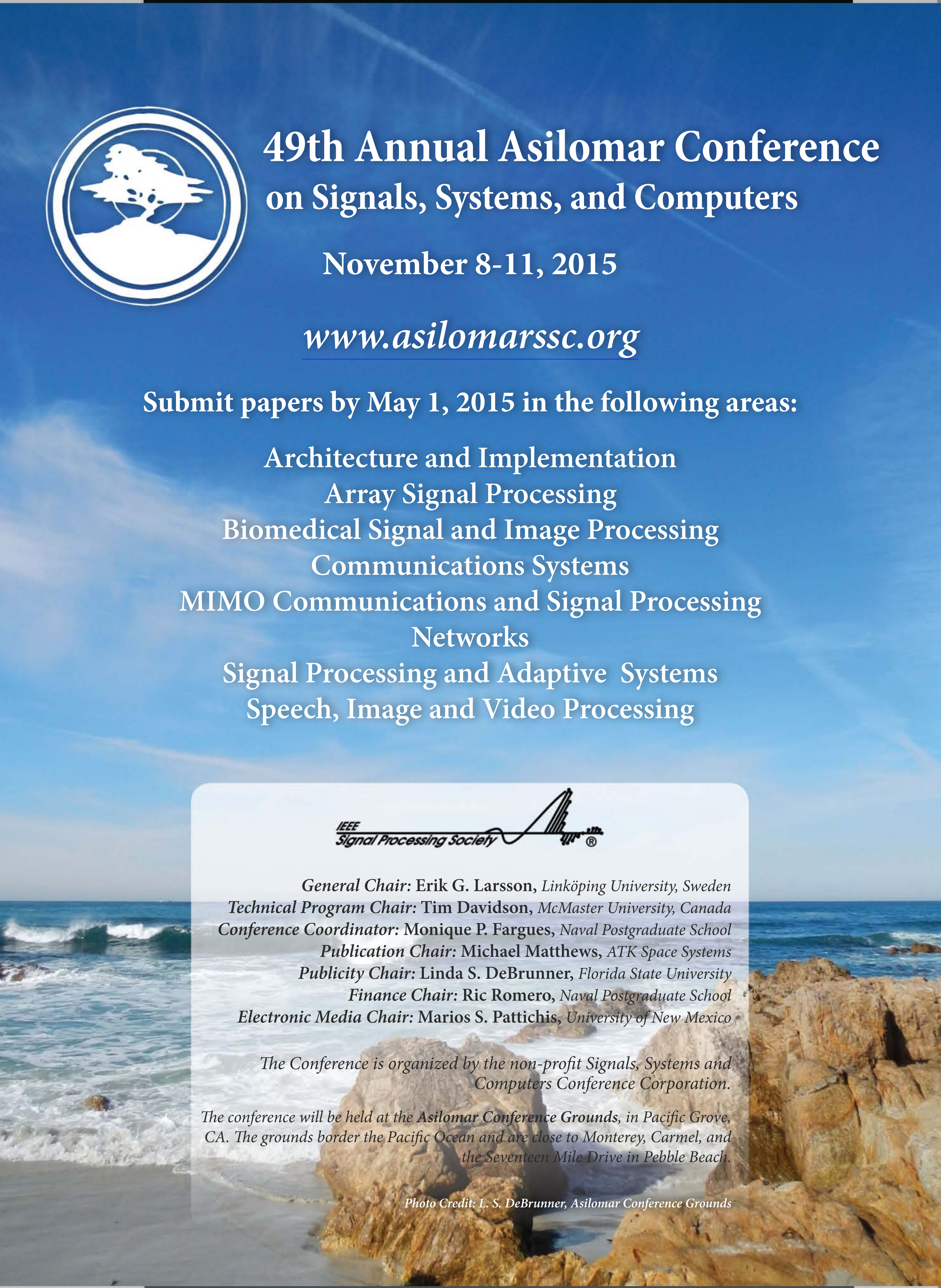
[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[10] T. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. (Wiley Series in Probability and Statistics). Hoboken, NJ: Wiley, 2003.

[11] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[12] T. Andersson and P. Händel, "IEEE standard 1057, Cramér–Rao bound and the parsimony principle," *IEEE Trans. Instrum. Measure.*, vol. 55, no. 1, pp. 44–53, 2006.

[SP]



49th Annual Asilomar Conference
on Signals, Systems, and Computers

November 8-11, 2015

www.asilomarssc.org

Submit papers by May 1, 2015 in the following areas:

Architecture and Implementation
Array Signal Processing
Biomedical Signal and Image Processing
Communications Systems
MIMO Communications and Signal Processing
Networks
Signal Processing and Adaptive Systems
Speech, Image and Video Processing

IEEE Signal Processing Society®

General Chair: Erik G. Larsson, *Linköping University, Sweden*
Technical Program Chair: Tim Davidson, *McMaster University, Canada*
Conference Coordinator: Monique P. Fargues, *Naval Postgraduate School*
Publication Chair: Michael Matthews, *ATK Space Systems*
Publicity Chair: Linda S. DeBrunner, *Florida State University*
Finance Chair: Ric Romero, *Naval Postgraduate School*
Electronic Media Chair: Marios S. Pattichis, *University of New Mexico*

The Conference is organized by the non-profit Signals, Systems and Computers Conference Corporation.

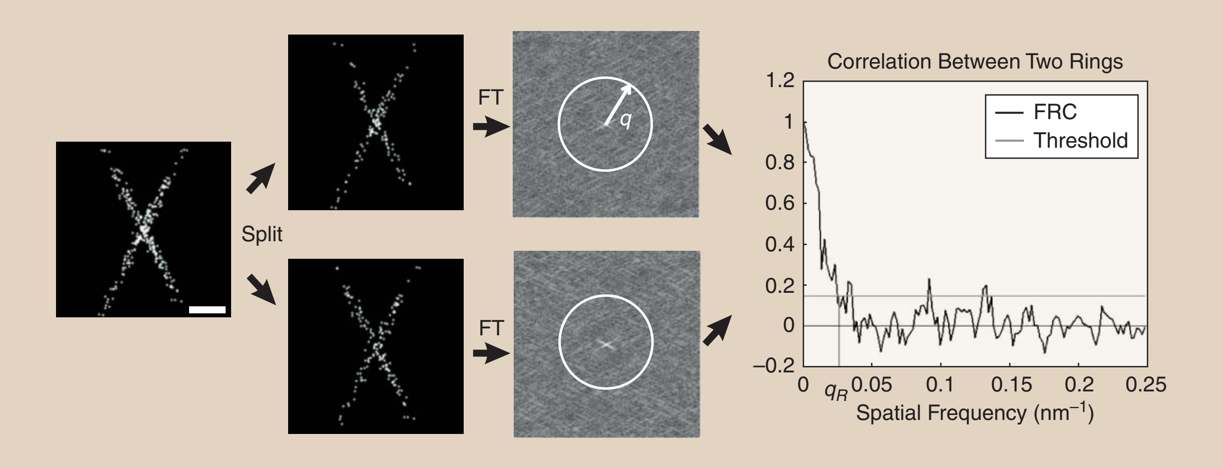
*The conference will be held at the **Asilomar Conference Grounds**, in Pacific Grove, CA. The grounds border the Pacific Ocean and are close to Monterey, Carmel, and the Seventeen Mile Drive in Pebble Beach.*

Photo Credit: L. S. DeBrunner, Asilomar Conference Grounds

ERRATA

In the article "Image Processing and Analysis for Single-Molecule Localization Microscopy" by B. Rieger et al. in the January 2015 issue of *IEEE Signal Processing Magazine* [1], the two white circles in the gray boxes in Figure 4 were displaced due to a production error.

The correct FIgure 4 appears below. We apologize for the errors and any confusion they may have caused.



[FIG4] A schematic illustration of FRC resolution computation. The localizations are divided into two halves, and their Fourier transforms are correlated over the perimeters of circles in Fourier space of radius q. The resulting FRC curve decays with spatial frequency, and the image resolution is taken to be the inverse of the spatial frequency q_R where the FRC curve drops below the threshold 1/7.

Reference

[1] B. Rieger, R. P. J. Nieuwenhuizen, and S. Stallinga, "Image processing and analysis for single-molecule localization microscopy," *IEEE Signal Processing Mag.*, vol. 32, no. 1 pp. 49–57, Jan. 2015.

Digital Object Identifier 10.1109/MSP.2015.2389292

Digital Object Identifier 10.1109/MSP.2015.2396713

Siwei Ma, Tiejun Huang, Cliff Reader, and Wen Gao

AVS2—Making Video Coding Smarter

AVS2 is a new generation of video coding standard developed by the IEEE 1857 Working Group under project 1857.4. AVS2 is also the second-generation video coding standard established by the Audio and Video Coding Standard (AVS) Working Group of China; the first-generation AVS1 was developed by the AVS Working Group and issued as Chinese national standard GB/T 20090.2-2006 in 2006. The AVS Working Group was founded in 2002 and is dedicated to providing the digital audio-video industry with highly efficient and economical coding/decoding technologies. So far, the AVS1 video coding standard is widely implemented in regional broadcasting, communication, and digital video entertainment systems. As the successor of AVS1, AVS2 is designed to achieve significant coding efficiency improvements relative to the preceding H.264/MPEG4-AVC and AVS1 standards. The basic coding framework of AVS2 is similar to the conterminous HEVC/H.265, but AVS2 can provide more efficient compression for certain video applications such as surveillance as well as low-delay communication such as videoconferencing. AVS2 is making video coding smarter by adopting intelligent coding tools that not only improve coding efficiency but also help with computer vision tasks such as object detection and tracking.

BACKGROUND

The AVS Working Group was established in March 2002 in China. The mandate of the group is to establish generic technical standards for the compression, decoding, processing, and representation of digital audio-video content, thereby enabling digital audio-video equipment and systems with highly efficient and economical coding/decoding technologies. After more than a decade, the working group has published a series of standards, including AVS1, which is the culmination of the first stage of work.

[AVS2 IS MAKING VIDEO CODING SMARTER BY ADOPTING INTELLIGENT CODING TOOLS THAT NOT ONLY IMPROVE CODING EFFICIENCY, BUT ALSO HELP WITH COMPUTER VISION TASKS SUCH AS OBJECT DETECTION AND TRACKING.]

Table 1 shows the time line of the AVS1 video coding standard (for short, AVS1). In AVS1, six profiles were defined to meet the requirements of various applications. The Main Profile focuses on digital video applications like commercial broadcasting and storage media, including high-definition video applications. It was approved as a national standard in China: GB/T 20090.2-2006. It was followed by the Enhanced Profile, an extension of the Main Profile with higher coding efficiency, targeting the needs of multimedia entertainment, such as movie compression for high-density storage. The Surveillance Baseline and Surveillance Profiles focus on video surveillance applications, considering in particular the characteristics of surveillance videos, i.e., high noise levels, relatively low encoding complexity, and requirements for easy event detection and search. The Portable Profile targets mobile video applications with lower resolution, low computational complexity, and robust error resiliency to meet the wireless environment. The latest Broadcasting Profile is also an improvement of the Main Profile and targets high-quality, high-definition TV (HDTV) broadcasting. It was approved and published as an industry standard by the State of China Broadcasting Film and Television Administration in July 2012.

AVS standards are also being recognized internationally. In 2007, the Main

[TABLE 1] TIME LINE OF AVS1 VIDEO CODING STANDARD.

TIME	PROFILE	TARGET APPLICATION(s)	MAJOR CODING TOOLS
DECEMBER 2003	MAIN	TV BROADCASTING	8 × 8 BLOCK-BASED INTRAPREDICTION, TRANSFORM AND DEBLOCKING FILTER; VARIABLE BLOCK SIZE MOTION COMPENSATION (16 × 16 ~ 8 × 8)
JUNE 2008	SURVEILLANCE BASELINE	VIDEO SURVEILLANCE	BACKGROUND-PREDICTIVE PICTURE FOR VIDEO CODING, ADAPTIVE WEIGHTING QUANTIZATION (AWQ), CORE FRAME CODING
SEPTEMBER 2008	ENHANCED	DIGITAL CINEMA	CONTEXT BINARY ARITHMETIC CODING (CBAC), AWQ
JULY 2009	PORTABLE	MOBILE VIDEO COMMUNICATION	8 × 8/4 × 4 BLOCK TRANSFORM
JULY 2011	SURVEILLANCE	VIDEO SURVEILLANCE	BACKGROUND MODELING BASED CODING
MAY 2012	BROADCASTING	HDTV	AWQ, ENHANCED FIELD CODING

Digital Object Identifier 10.1109/MSP.2014.2371951
Date of publication: 12 February 2015

1053-5888/15©2015IEEE

IEEE GlobalSIP'15—Call for Papers

2015 IEEE Global Conference on Signal and Information Processing – Orlando, Florida

General Chairs: Jose Moura and Dapeng Oliver Wu
Technical Program Chairs: Mihaela van der Schaar, Xiaodong Wang, and Hsiao-Chun Wu

The IEEE Global Conference on Signal and Information Processing (GlobalSIP) is a recently launched flagship conference of the *IEEE Signal Processing Society*. GlobalSIP' 15 will be held in Orlando, Florida, USA, December 14–16, 2015. The conference will focus broadly on signal and information processing with an emphasis on up-and-coming signal processing themes. The conference will feature world-class speakers, tutorials, exhibits, and technical sessions consisting of poster or oral presentations. GlobalSIP' 15 technical program will be comprised of a main program and several co-located symposia on special topics. Technical paper submissions are solicited in the interest topics which may include, but are not limited to:

- Signal processing in communications and networks, including green communication and Signal processing in optical communication
- Image and video processing
- Selective topics in speech and language processing
- Signal processing in security applications
- Signal processing in finance
- Signal processing in energy and power systems
- Signal processing in genomics and bioengineering (physiological, pharmacological and behavioral)
- Signal processing for social media networks
- Neural signal processing
- Seismic signal processing
- Selective topics in statistical signal processing
- Graph-theoretic signal processing
- Machine learning
- Compressed sensing, sparsity analysis, and applications
- Big data processing, heterogeneous information processing and informatics
- Human machine interfaces
- Multimedia transmission, indexing and retrieval, and playback challenges
- Hardware and real-time implementations
- Other novel and significant Applications of selected areas of signal processing

Submission of Papers: Prospective authors are invited to submit full-length papers, with up to four pages for technical content including figures and possible references, and with one additional optional 5th page containing only references. Manuscripts should be original (not submitted/published anywhere else) and written in accordance with the standard IEEE double-column paper template. All paper submissions should be carried out through EDAS system (http://edas.info). A selection of best papers and best student papers will be made by the GlobalSIP 2015 best paper award committee upon recommendations from Technical Committees.

Notice: The IEEE Signal Processing Society enforces a "no-show" policy. Any accepted paper included in the final program is expected to have at least one author or qualified proxy attend and present the paper at the conference. Authors of the accepted papers included in the final program who do not attend the conference will be subscribed to a "No-Show List", compiled by the Society. The "no-show" papers will not be published by the IEEE on IEEE *Xplore* or other public access forums, but these papers will be distributed as part of the on-site electronic proceedings and the copyright of these papers will belong to the IEEE.

Timeline for paper submission:

May 15, 2015:	Paper submission deadline
June 30, 2015:	Review results announced
September 5, 2015:	Camera-ready papers due

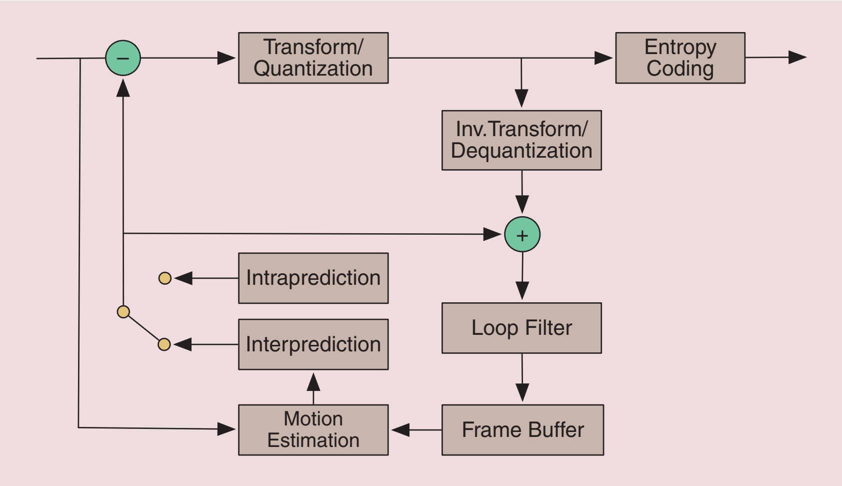
Digital Object Identifier 10.1109/MSP.2015.2396712

Profile was accepted as an option of video codecs for Internet Protocol Television (IPTV) applications by the International Telecommunication Union–Telecommunication Standardization Sector (ITU-T) Focus Group on IPTV standardization [1]. The IEEE 1857 Working Group was established in 2012 to work on IEEE standards for advanced audio and video coding, based on individual members of the IEEE Standards Association from the AVS Working Group. The IEEE 1857 Working Group meets three to four times annually to discuss the standard technologies, syntax, and so on. Until now, the IEEE 1857 Working Group has finished three parts of IEEE 1857 standards, including IEEE 1857-2013 for video, IEEE 1857.2-2013 for audio, and IEEE 1857.3-2013 for system [2].

AVS standards have been developed in compliance with the AVS intellectual property rights (IPR) policy. This policy includes up-front commitment by participants to license essential patents with declaration of default licensing terms—royalty-free without compensation [(RAND-RF) and otherwise under reasonable and nondiscriminatory terms], or participation in the AVS patent pool, or RAND. The disclosure of published patent applications and granted patents is required, and the existence of unpublished applications is also required if the RAND option is taken. The licensing terms are also considered in the adoption of proposals for AVS standards when all technical factors are equal. Reciprocity in licensing is required. The protection of participants's IPR is provided to guard against the situation in which the IPR of a participant are disclosed by another party. AVS has encouraged the establishment of a Patent Pool Administration (PPA) that is independent from the AVS Working Group, which only focuses on the standards. The AVS standards are also fully compliant with the IPR policy of IEEE standards.

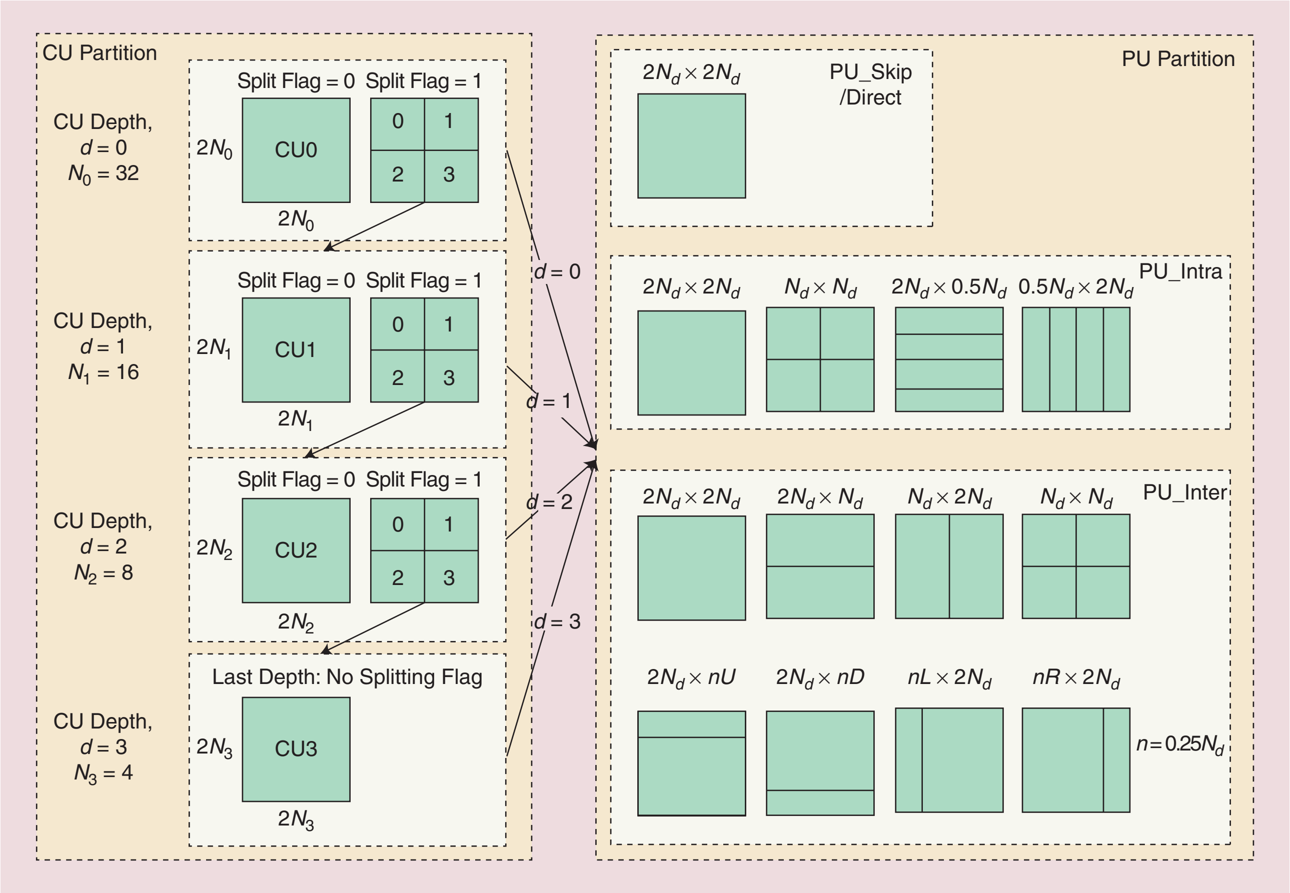
[AS A SUCCESSOR OF AVS1, AVS2 IS DESIGNED TO IMPROVE CODING EFFICIENCY FOR HIGHER RESOLUTION VIDEOS AND PROVIDE EFFICIENT COMPRESSION SOLUTIONS FOR VARIOUS KINDS OF VIDEO APPLICATIONS.]

Based on the success of AVS1 and the recent research and standardization works, AVS has been working on a new generation of video coding technologies called AVS2 (or more specifically, Part 2 in the AVS2 series standards). In fact, since 2005 and before the AVS2 project officially started, AVS has been continuously working on an AVS-X project to explore more efficient coding techniques. AVS2 was started formally by issuing a call for platforms in March 2012. By October 2012, a reference platform (RD 1.0) based on the AVS1 reference software was developed for AVS2 [3]. After that, AVS2 continued to improve its coding efficiency, and the standard in committee draft 2.0 was finalized in June 2014. It has been approved as a project of IEEE standard, IEEE 1857.4, and a project of Chinese national standard, both of which are expected to be finished by the end of 2014 at the time of this writing.



[FIG1] The coding framework of an AVS2 encoder.

As a successor of AVS1, AVS2 is designed to improve coding efficiency for higher-resolution videos and provide efficient compression solutions for various kinds of video applications. Compared to the preceding coding standards, AVS2 adopts smarter coding tools that are adapted to satisfy the new requirements identified from emerging applications. First, more flexible prediction block partitions are used to further improve prediction accuracy, e.g., square and non-square partitions, which are more adaptive to the image content especially in edge areas. Related to the prediction structure, transform block size is more flexible and can be up to 64 × 64 pixels. After transformation, context adaptive arithmetic coding is used for the entropy coding of the transformed coefficients. A two-level coefficient scan and coding method can encode the coefficients of large blocks more efficiently. Moreover, for low-delay communication applications, e.g., video surveillance, video conference, etc., where the background usually does not often change, a background picture model-based coding method is developed in AVS2. The background picture constructed from original pictures or decoded pictures is used as a reference picture to improve prediction efficiency. Test results show that this background picture-based prediction coding can improve coding efficiency significantly. Furthermore, the background picture can also be used for object detection and tracking for intelligent surveillance. In addition, to support object tracking among multiple cameras in surveillance applications, navigation information such as those from the global positioning system and BeiDou Navigation Satellite System of China is also defined, which mainly includes timing, location, and movement information. Finally, aiming at more intelligent surveillance video coding, AVS2 also started a



[FIG2] **(a) The maximum possible recursive CU structure in AVS2. (LCU size = 64, maximum hierarchical depth = 4). (b) Possible PU splitting for skip, intramodes, and intermodes in AVS2, including symmetric and asymmetric prediction (*d*=1, 2 for intraprediction, and *d*= 0,1,2 for interprediction).**

digital media content description project in which visual objects in the images or videos are described with multilevel features for facilitating visual object based storage, retrieval, and interactive applications, etc.

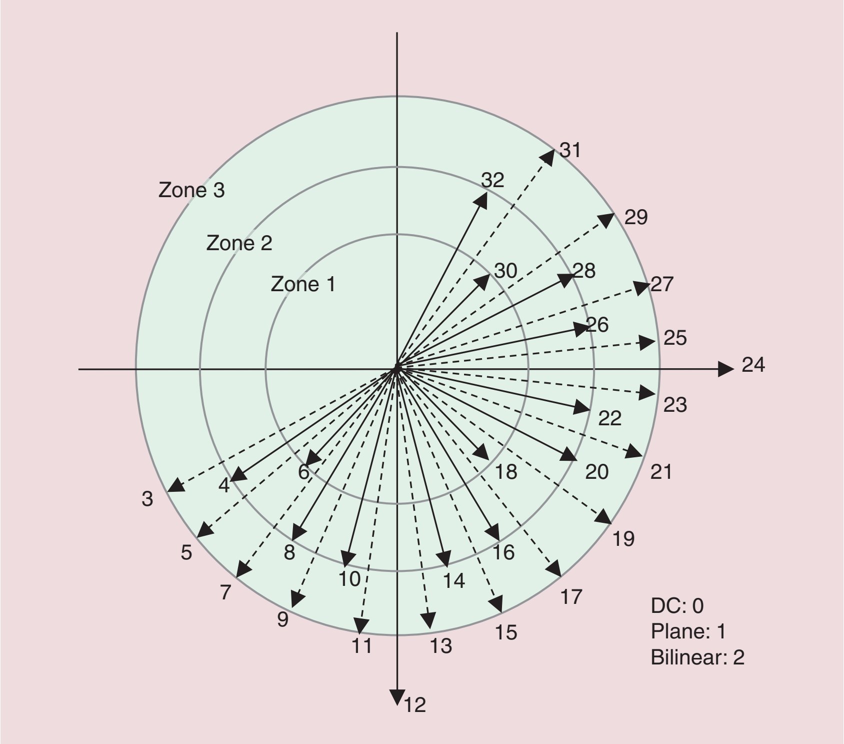
This column will provide a short overview of AVS2 video coding technology and a performance comparison with other video coding standards.

TECHNOLOGY AND KEY FEATURES

Similar to previous coding standards, AVS2 adopts the traditional prediction/transform hybrid coding framework, as shown in Figure 1. Within the framework, a more flexible coding structure is adopted for efficient high-resolution video coding, and more efficient coding tools are developed to make full use of the textual information and temporal redundancies. These tools can be classified into four categories: 1) prediction coding (including intraprediction and interprediction), 2) transform, 3) entropy coding, and 4) in-loop filtering. We will give a brief introduction to the coding framework and coding tools.

CODING FRAMEWORK

In AVS2, a coding unit (CU)-, prediction unit (PU)-, and transform unit (TU)-based coding/prediction/transform structure is adopted to represent and organize the encoded data [3]. First, pictures are split into largest coding units (LCUs), which consist of $2N \times 2N$ samples of a luminance component and associated chrominance samples with $N = 8, 16$, or 32. One LCU can be a single CU or can be split into four smaller CUs with a quad-tree partition structure; a CU can be recursively split until it reaches the smallest CU size limit, as shown in Figure 2(a). Once the splitting of the CU hierarchical tree is finished, the leaf node CUs can be further split into PUs. PU is the basic unit for intra- and interprediction and allows multiple different shapes to encode irregular image patterns, as shown in Figure 2(b). The size of a PU is limited to that of a CU with various square or rectangular shapes. More specifically, both intra- and interprediction partitions can be symmetric or asymmetric. Intraprediction partitions vary in the set $\{2N \times 2N, N \times N, 2N \times 0.5N, 0.5N \times 2N\}$, while inter-prediction partitions vary in the set $\{2N \times 2N, 2N \times N, N \times 2N, 2N \times nU, 2N \times nD, nL \times 2N, nR \times 2N\}$, where U, D, L,and R are the abbreviations of "Up," "Down," "Left," and "Right," respectively. Besides CU and PU, TU is also defined to represent the basic unit for transform coding and quantization. The size of a TU cannot exceed that of a CU, but it is independent of the PU size.



[FIG3] An illustration of directional prediction modes.

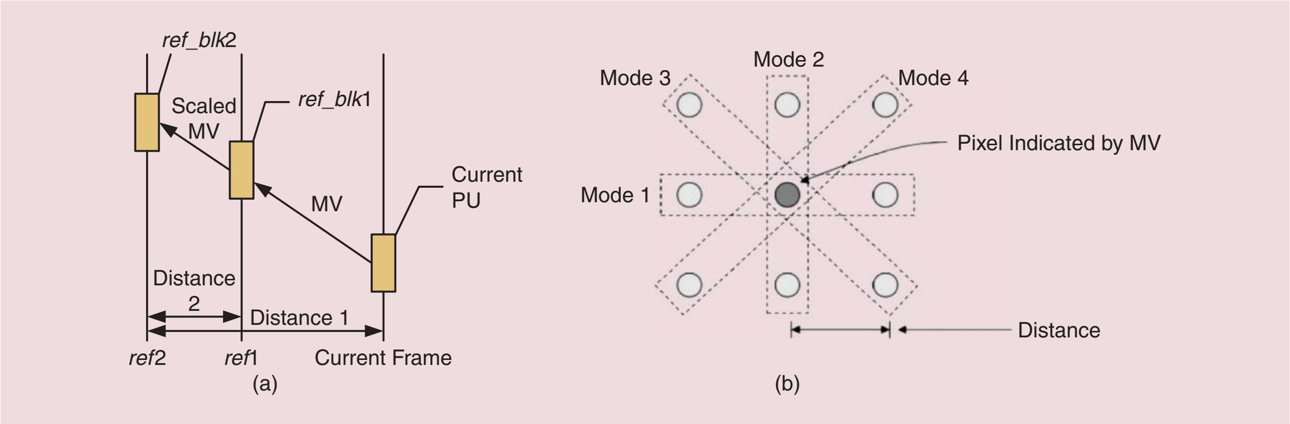
INTRAPREDICTION

Intraprediction is used to reduce the redundancy existing in the spatial domain of the picture. Block partition-based directional prediction is used for AVS2 [5]. As shown in Figure 2, besides the square PU partitions, nonsquare partitions, called *short distance intra prediction* (*SDIP*), are adopted by AVS2 for more efficient intraluminance prediction [4], where the nearest reconstructed boundary pixels are used as the reference sample in intraprediction. For SDIP, a $2N \times 2N$ PU is horizontally/vertically partitioned into four prediction blocks. SDIP is more adaptive to the image content, especially in edge area. But for the complexity reduction, SDIP is used in all CU sizes except a 64 × 64 CU. For each prediction block in the partition modes, a total of 33 prediction modes are supported for luminance, including 30 angular modes [5], a plane mode, a bilinear mode, and a DC mode. Figure 3 shows the distribution of the prediction directions associated with the 30 angular modes. Each sample in a PU is predicted by projecting its location to the reference pixels applying the selected prediction direction. To improve the intraprediction accuracy, the subpixel precision reference samples must be interpolated if the projected reference samples locate on a noninteger position. The noninteger position is bounded to 1/32 sample precision to avoid floating point operation, and a four-tap linear interpolation filter is used to get the subpixel.

For the chrominance component, the PU size is always $N \times N$, and five prediction modes are supported, including vertical prediction, horizontal prediction, bilinear prediction, DC prediction, and the prediction mode derived from the corresponding luminance prediction mode [6].

INTERPREDICTION

Compared to the spatial intraprediction, interprediction focuses on exploiting the temporal correlation between the consecutive pictures to reduce the temporal redundancy. Multireference prediction has been used since the H.264/AVC standard, including both short-term and long-term reference pictures. In AVS2, long-term reference picture usage is extended further, which can be constructed from a sequence of long-term decoded pictures, e.g., background picture used in surveillance coding, which will be discussed separately later. For short-term reference prediction in AVS2, F frames are defined as a special P frame [7], in addition to the traditional P and B frames. More specifically, a P frame is a forward-predicted frame using a single reference picture, while a B frame is a bipredicted frame that consists of forward,



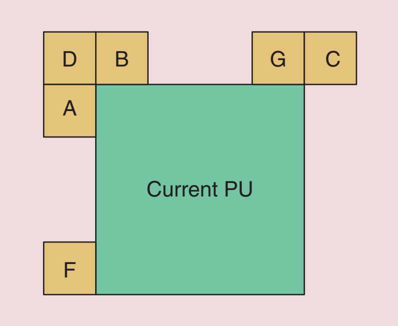
[FIG4] (a) Temporal multihypothesis mode. (b) Spatial multihypothesis mode.

backward, biprediction, and symmetric prediction, using two reference frames.

In a B frame, in addition to the conventional forward, backward, bidirectional, and skip/direct prediction modes, symmetric prediction is defined as a special biprediction mode, wherein only one forward motion vector (MV) is coded and the backward MV is derived from the forward MV. For an F frame, besides the conventional single hypothesis prediction mode in a P frame, multihypothesis techniques are added for more efficient prediction, including the advanced skip/direct mode [8], temporal multihypothesis prediction mode [9], and spatial directional multihypothesis (DMH) prediction mode [10].

In an F frame, an advanced skip/direct mode is defined using a competitive motion derivation mechanism. Two derivation methods are used: one is temporal and the other is spatial. Temporal multihypothesis mode combines two predictors along the predefined temporal direction, while spatial multihypothesis mode combines two predictors along the predefined spatial direction. For temporal derivation, the prediction block is obtained by an average of the prediction blocks indicated by the MV prediction (MVP) and the scaled MV in a second reference. The second reference is specified by the reference index transmitted in the bit stream. For temporal multihypothesis prediction, as shown in Figure 4, one predictor *ref_blk*1 is generated with the best MV *MV* and a reference frame *ref*1 is searched by motion estimation, and then this MV is linearly scaled to a second reference to generate another predictor *ref_blk*2. The second reference *ref*2 is specified by the reference index transmitted in the bit stream. In DMH mode, as specified in Figure 4, the seed predictors are located on the line crossing the initial predictor obtained from motion estimation. The number of seed predictors is restricted to eight. If one seed predictor is selected for combined prediction, for example "Mode 1," then the index of the seed predictor "1" will be signaled in the bit stream.

For spatial derivation, the prediction block may be obtained from one or two prediction blocks specified by the motion copied from its spatial neighboring blocks. The neighboring blocks are illustrated in Figure 5. They are searched in a predefined order F, G, C, A, B, D, and the selected neighboring block is signaled in the bit stream.



[FIG5] An illustration of neighboring blocks A, B, C, D, F, and G for MVP.

MOTION VECTOR PREDICTION AND CODING

MVP plays an important role in interprediction, which can reduce the redundancy among MVs of neighboring blocks and thus save large numbers of coding bits for MVs. In AVS2, four different prediction methods are adopted, as tabulated in Table 2. Each of them has its unique usage. Spatial MVP is used for the spatial derivation of Skip/Direct mode in F frames and B frames. Temporal MVP is used for temporal derivation of Skip/Direct mode in P frames and F frames. Spatial-temporal-combined MVP is used for the joint temporal and spatial derivation of Skip/Direct mode in B frames. For other cases, median prediction is used.

In AVS2, the MV is in quarter-pixel precision for the luminance component, and the subpixel is interpolated with an eight-tap DCT interpolation filter (DCT-IF) [11]. For the chrominance component, the MV derived from luminance with 1/8 pixel precision and a four-tap DCT-IF is used for subpixel interpolation [12]. After the MVP, the MV difference (MVD) is coded in the bit stream. However, redundancy may still exist in MVD, and to further save coding bits of MVs, a progressive MV resolution adaptation method is adopted in AVS2 [13]. In this scheme, the MVP is firstly rounded to the nearest integer sample position, and then the MV is rounded to a half-pixel precision if its distance from MVP is larger than a by a threshold. Finally, the resolution of the MVD is decreased to half-pixel precision if it is larger than a threshold.

TRANSFORM

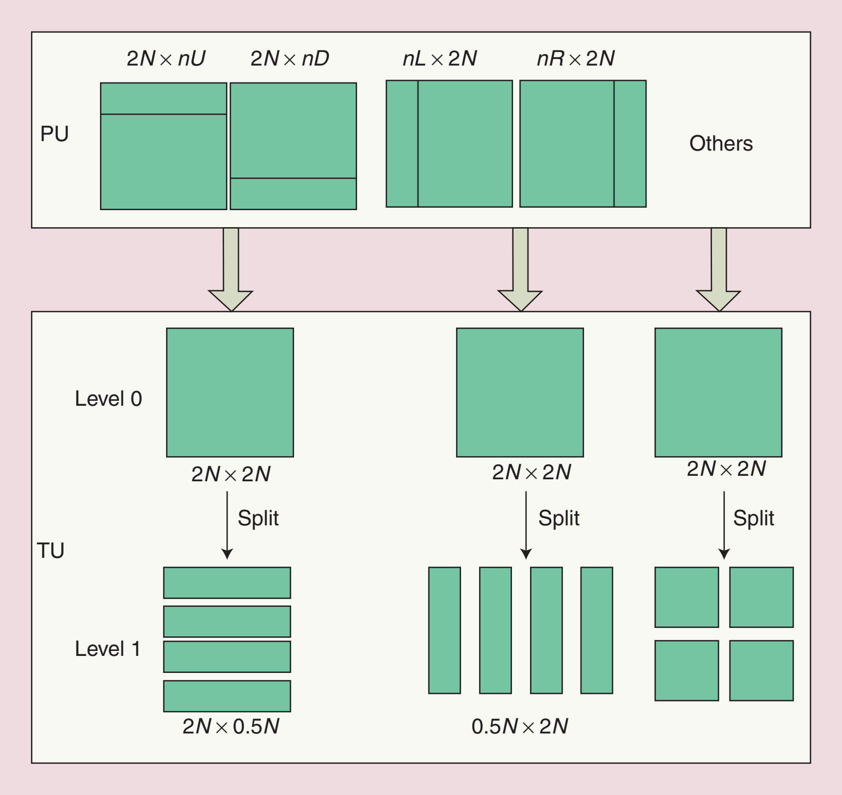
Two-level transform coding is utilized to further compress the predicted residual. For a CU with symmetric prediction unit partition, the TU size can be $2N \times 2N$ or $N \times N$ signaled by a transform split flag. Thus, the maximum transform size is 64 × 64, and the minimum transform size is 4 × 4. For the TU size 4 × 4 to 32 × 32, an integer transform (IT) that closely approximates the performance of the discrete cosine transform (DCT) is used; while for the 64 × 64 transform, a logical transform (LOT) [14] is applied to the residual. A five-three-tap integer wavelet transform is first performed on a 64 × 64 block discarding the low-high (LH), high-low (HL), and (high-high) HH-bands, and then a normal 32 × 32 IT is applied to the low-low (LL)-band. For a CU that has an asymmetric PU partition, a $2N \times 2N$ IT is used in the first level and a nonsquare transform [15] is used in the second level, as shown in Figure 6. Moreover, in the latest AVS2 standard, a secondary transform was adopted for intraprediction residual (for more details see the latest AVS specification document N2120 on the AVS FTP Web site [21]).

ENTROPY CODING

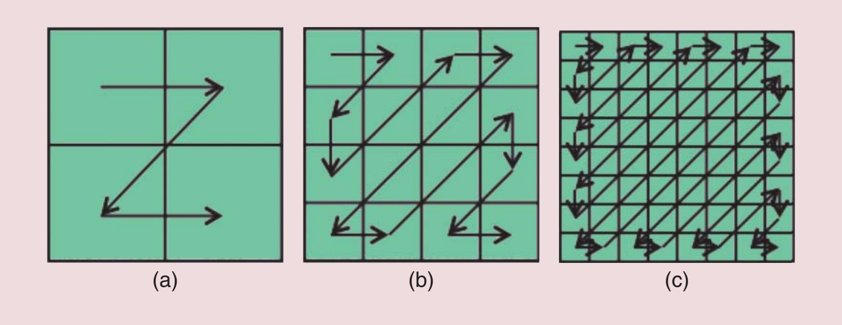
After transform and quantization, a two-level coding scheme is applied to the

[TABLE 2] MV PREDICTION METHODS IN AVS2.

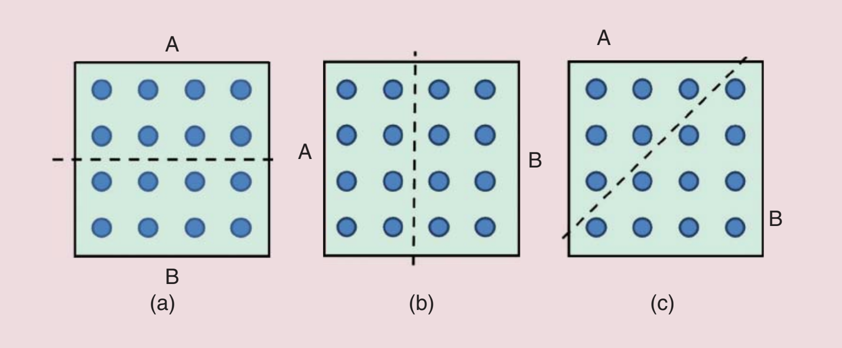
METHOD	DETAILS
MEDIAN	USING THE MEDIAN MV VALUES OF THE NEIGHBORING BLOCKS.
SPATIAL	USING THE MVs OF SPATIAL NEIGHBORING BLOCKS.
TEMPORAL	USING THE MVs OF TEMPORAL COLLOCATED BLOCKS.
SPATIAL-TEMPORAL COMBINED	USING THE TEMPORAL MVP FIRST IF IT IS AVAILABLE, AND SPATIAL MVP IS USED INSTEAD IF THE TEMPORAL MVP IS NOT AVAILABLE.



[FIG6] A PU partition and two-level transform coding.



[FIG7] A subblock scan for transform blocks of size (a) 8 × 8, (b) 16 × 16, and (c) 32 × 32 transform blocks; each subblock represents a 4 × 4 CG.



[FIG8] A subblock region partitions of 4 × 4 CG in an intraprediction block.

transform coefficient blocks [16]. A coefficient block is partitioned into 4 × 4 coefficient groups (CGs), as shown in Figure 7. Then zig-zag scanning and context-adaptive binary arithmetic coding (CABAC) is performed at both the CG level and coefficient level. At the CG level for a TU, the CGs are scanned in zig-zag order, and the CG position indicating the position of the last nonzero CG is coded first, followed by a bin string of significant CG flags indicating whether the CG scanned in zig-zag order contains nonzero coefficients. At the coefficient level, for each nonzero CG, the coefficients are further scanned into the form of (*run*, *level*) pair in zig-zag order. Level and run refer to the magnitude of a nonzero coefficient and the number of zero coefficients between two nonzero coefficients, respectively. For the last CG, the coefficient position that denotes the position of the last nonzero coefficient in scan order is coded first. For a nonlast CG, a last run is coded that denotes number of zero coefficients after the last nonzero coefficient in zig-zag scan order. And then the (*level*, *run*) pairs in a CG are coded in reverse zig-zag scan order.

For the context modeling used in the CABAC, AVS2 employs a mode-dependent context selection design for intraprediction blocks [17]. In this context design, 34 intraprediction modes are classified into three prediction mode sets: vertical, horizontal, and diagonal. Depending on the prediction mode set, each CG is divided to two regions, as shown in Figure 8. The intraprediction modes and CG regions are applied in the context coding of syntax elements including the last CG position, last coefficient position, and run value.

IN-LOOP FILTERING

Artifacts such as blocking artifacts, ringing artifacts, color biases, and blurring artifacts are quite common in compressed video, especially at medium and low bit rate. To suppress those artifacts, deblocking filtering, sample adaptive offset (SAO) filtering [18], and adaptive loop filter (ALF) [19] are applied to the reconstructed pictures sequentially.

Deblocking filtering aims to remove the blocking artifacts caused by block transform and quantization. The basic unit for the deblocking filter is an 8 × 8 block. For each 8 × 8 block, the deblocking filter is used only if the boundary belongs to either of the CU, PU, or TU boundaries.

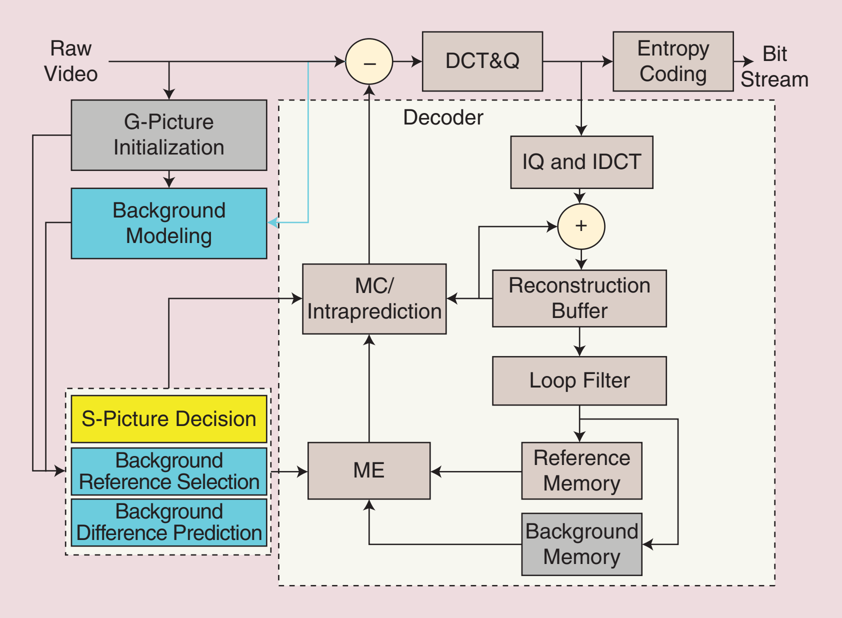
After the deblocking filter, an SAO filter is applied to reduce the mean sample distortion of a region, where an offset is added to the reconstructed sample to reduce ringing artifacts and contouring artifacts. There are two kinds of offset: edge offset (EO) and band offset (BO) mode. For the EO mode, the encoder can select and signal a vertical, horizontal, downward-diagonal, or upward-diagonal filtering direction. For BO mode, an offset value that directly depends on the amplitudes of the reconstructed samples is added to the reconstructed samples.

ALF is the last stage of in-loop filtering. There are two stages in this process. The first stage is filter coefficient derivation. To train the filter coefficients, the encoder classifies reconstructed pixels of the luminance component into 16 categories, and one set of filter coefficients is trained for each category using Wiener–Hopf equations to minimize the mean squared error between the original frame and the reconstructed frame. To reduce the redundancy between these 16 sets of filter coefficients, the encoder will adaptively merge them based on the rate-distortion performance. At its maximum, 16 different filter sets can be assigned for the luminance component and only one for the chrominance components. The second stage is a filter decision, which includes both the frame level and LCU level. First, the encoder decides whether frame-level adaptive loop filtering is performed. If frame level ALF is on, then the encoder further decides whether the LCU level ALF is performed.

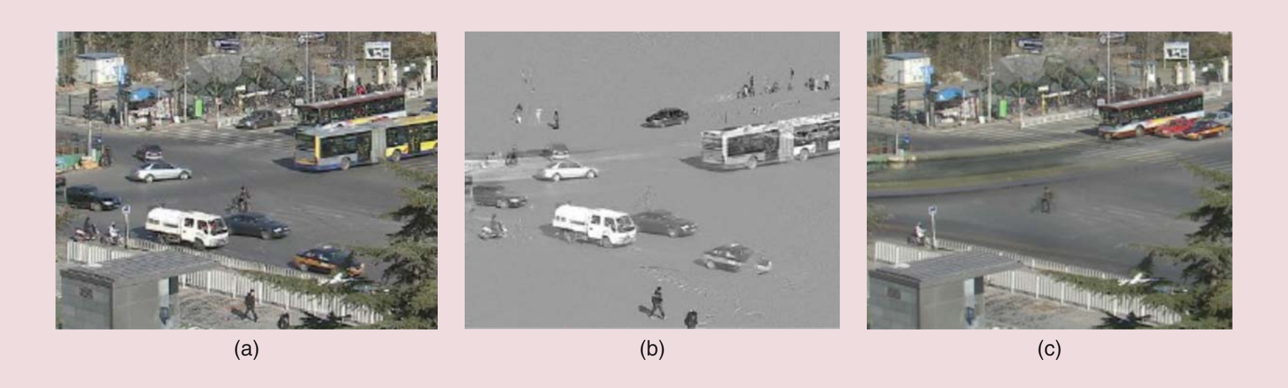
SMART SCENE VIDEO CODING

More and more videos being captured in specific scenes (such as surveillance video and videos from the classroom, home, courthouse, etc.) are characterized by a temporally stable background. The redundancy originating from the background could be further reduced. AVS2 developed a background picture model-based coding method [20], which is illustrated in Figure 9. G-pictures and S-pictures are defined to further exploit the temporal redundancy and facilitate video event generation such as object segmentation and motion detection. The G-picture is a special I-picture, which is stored in a separate background memory. The S-picture is a special P-picture, which can be only predicted from a reconstructed G-picture or a virtual G-picture, which does not exist in the actual input sequence but is modeled from input pictures and encoded into the stream to act as a reference picture.

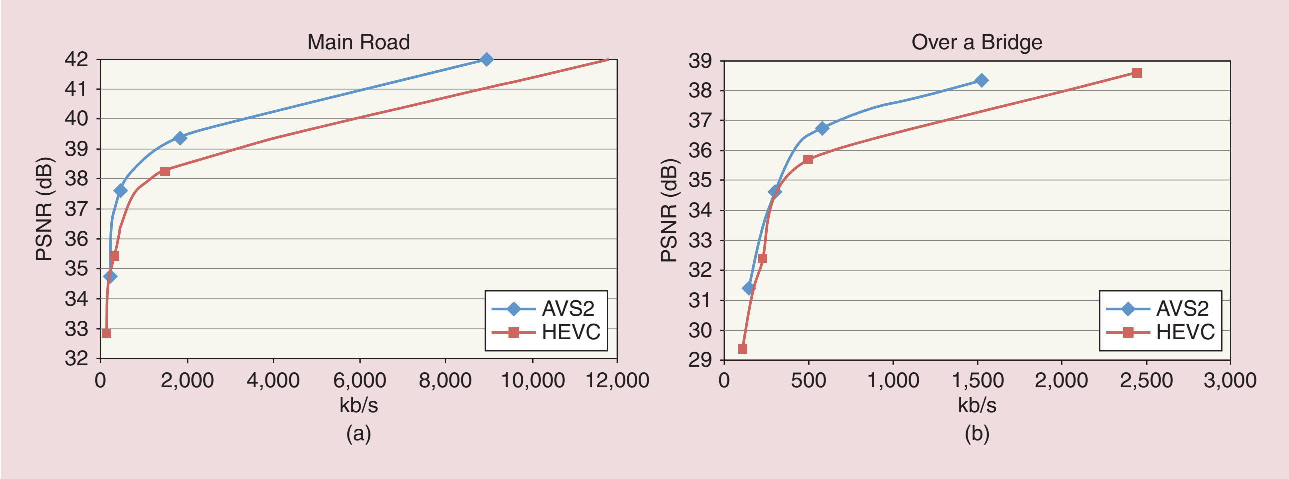
The G-picture is initialized by background initialization and updated by background modeling with methods such as median filtering, fast implementation



[FIG9] A background picture-based scene coding in AVS2.



[FIG10] Examples of the background picture and the difference frame between the original picture and the background picture: (a) original picture, (b) difference frame, and (c) background picture.



[FIG11] A performance comparison between AVS2 and HEVC for surveillance videos: (a) main road and (b) over a bridge.

of a Gaussian mixture model, etc. In this way, the selected or generated G-picture can well represent the background of a scene with rare occluding foreground objects and noise. Once a G-picture is obtained, it is encoded and the reconstructed picture is stored into the background memory in the encoder/decoder and updated only if a new G-picture is selected or generated. After that, S-pictures can be involved in the encoding process by an S-picture decision. Except that it uses a G-picture as a reference, the S-picture owns similar properties as the traditional I-picture such as error resilience and random access (RA). Therefore, the pictures that should be coded as traditional I-pictures can be candidate S-pictures, such as the first picture of one group of pictures, or scene change, etc. Besides bringing about more prediction opportunity for those background blocks that normally dominate a picture, an additional benefit from the background picture is a new prediction mode called *background difference prediction*, as shown in Figure 10, which can improve foreground prediction performance by excluding the background influence. It can be seen that, after background difference prediction, the background redundancy is effectively removed. Furthermore, according to the predication modes in the AVS2 compression bit stream, the blocks of an AVS2 picture could be classified as background blocks, foreground blocks, or blocks on the edge area. Obviously, this information is very helpful for possible subsequent vision tasks such as object detection and tracking. Object-based coding has already been proposed in MPEG-4; however, object segmentation remains a challenging problem, which constrains the application of object-based coding. Therefore AVS2 uses simple background modeling instead of accurate object segmentation, which is easier and provides a good tradeoff between coding efficiency and complexity.

[AVS2 HAS BEEN DEVELOPED IN ACCORDANCE WITH AVS AND IEEE IPR POLICIES TO ENSURE RAPID LICENSING OF ESSENTIAL PATENTS AT COMPETITIVE ROYALTY RATES.]

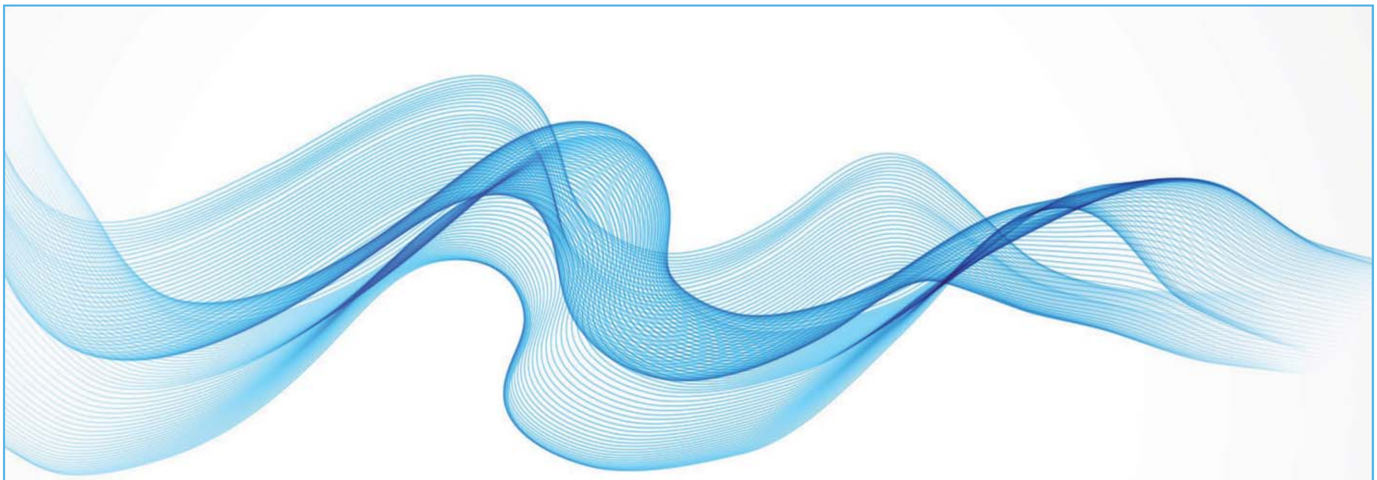
To provide convenience for applications like event detection and searching, AVS2 added some novel high-level syntax to describe the region of interest (ROI). In the region extension, the region number, event ID, and coordinates for top left and bottom right corners are included to show what number the ROI is, what event happened, and where it lies.

PERFORMANCE COMPARISON

The major target applications of AVS2 are high-quality TV broadcasting and scene videos. For high-quality broadcasting, RA is necessary and may be achieved by inserting intraframes at a fixed interval, e.g, 0.5 s. And for high-quality video capture and editing, all intracoding (AI) is required. For scene video applications, e.g., video surveillance or videoconference, low delay (LD) needs to be guaranteed. According to the applications, we tested

[TABLE 3] BIT RATE SAVING OF AVS2 PERFORMANCE COMPARISON WITH AVS1 AND HEVC.

SEQUENCES	AI CONFIGURATION		RA CONFIGURATION		LD CONFIGURATION
	AVS2 VERSUS AVS1	AVS2 VERSUS HEVC	AVS2 VERSUS AVS1	AVS2 VERSUS HEVC	AVS2 VERSUS HEVC
UHD	31.2%	2.4%	50.3%	–0.4%	
1080P	33%	0.8%	50.3%	0.3%	
1200P					37.9%
SD					26.2%
OVERALL	32.1%	1.6%	50.3%	–0.1%	32.1%



FREE SPS STUDENT MEMBERSHIP FOR 2015

You're in the beginning stages of your career. Membership in the IEEE Signal Processing Society can help you lay the groundwork for many years of success. You can have it all in 2015 - and for free! Membership includes:

- Discounts on conference registration fees;
- Eligibility to apply for travel grants to attend SPS flagship conferences including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and IEEE International Conference on Image Processing (ICIP);
- Networking and job opportunities at the ICASSP Student Career Luncheon;
- Eligibility to enter our student competition, the Signal Processing Cup, for a US$5,000 grand prize;
- Involvement opportunities through SPS's local Chapters - more than 130 worldwide;
- Free electronic and digital subscriptions to *IEEE Signal Processing Magazine*, *Inside Signal Processing e-Newsletter*, and the *IEEE Signal Processing Society Content Gazette*;
- Access to cutting-edge educational resources, including SigView, SPS's online video tutorial portal.

See everything Signal Processing Society membership can do for you:
http://signalprocessingsociety.org

Already an IEEE member? Join SPS for free now!

(You must have already renewed your IEEE membership for 2015 to use this offer)

- Visit http://ieee.org/join
- On the left side, click "Societies and Special Interest Groups"
- Click "IEEE Signal Processing Society," then "Join the IEEE Signal Processing Society"
- When you reach the catalog page, click "Add Item(s)" and sign in with your IEEE account *Note: Free offer applies only to basic membership. For US$8.00, enhance your membership for more great benefits!*
- Once logged in, click "Proceed to Checkout"
- When you reach the shopping cart, enter the promotion code SP15STUAD and click "Apply"
- Complete check out and congratulations! Welcome to SPS!

Not yet an IEEE Student Member?

Get a free SPS membership with the purchase of an IEEE Student membership!

- Visit http://ieee.org/join
- Click "Join as a student" on the bottom right to create your new IEEE Student member account
- After your IEEE account is created, complete the membership application and proceed to "Do you want to add any memberships and subscriptions?"
- Select "Signal Processing Society membership" and click "add selected item"
- Click "Proceed to Checkout"
- When you reach the shopping cart, enter the promotion code SP15STUNW and click "Apply"
- Complete check out and congratulations! Welcome to SPS!

Note: *Must be an active IEEE Student or Graduate Student member. This offer does not apply to SPS Students or Graduate Students renewing for 2015.*

IEEE

Digital Object Identifier 10.1109/MSP.2015.2396413

the performance of AVS2 with three different coding configurations AI, RA, and LD, similar to the high-efficiency video coding (HEVC) common test conditions and Bjøntegaard delta bit rate is used for bit rate saving evaluation. The ultrahigh-definition (UHD) and 1080p test sequences are the common test sequences used in AVS, including partial test sequences used in HEVC, such as Traffic (UHD) and Kimono1(1080P), etc. All of these sequences and the surveillance/videoconference sequences used for LD testing are available on the AVS Web site [21].

Table 3 summarizes the rate distortion performance of AVS2 for three test cases. As shown in the table, for RA and AI configurations, AVS2 shows comparable performance as HEVC and outperforms AVS1 with significant bits saving, up to 50% for RA. For surveillance and videoconference video coding, AVS2 outperforms HEVC by 32.1%, and the curves in Figure 11 show the results on two surveillance video sequences. For the coding configurations more reasonable for scene video coding, AVS2's gain is more significant. It should be pointed out that the results are tested with the current AVS2 reference software RD9.2, which is still under optimization, and the performance of AVS2 may be improved further.

CONCLUSIONS

This column gives an overview of the upcoming AVS2 standard. AVS2 is an application-oriented coding standard, and different coding tools have been developed according to various application characteristics and requirements. For high-quality broadcasting, flexible prediction and transform coding tools have been incorporated. For surveillance video and videoconferencing applications, AVS2 bridges video compression with machine vision by incorporating smart coding tools, e.g., background picture modeling and location/time information etc., thereby making video coding smarter and more efficient. Compared to the previous AVS1 coding standards, AVS2 achieves significant improvement in coding efficiency and flexibility. AVS2 has been developed in accordance with AVS and IEEE IPR policies to ensure rapid licensing of essential patents at competitive royalty rates. In the development of AVS2, the favorability of licensing terms was also considered in the adoption of proposals for AVS standards, and the formation of a patent pool is expected in the near future.

Several directions are currently being explored for future extensions of AVS2, including three-dimensional video coding and media description for smarter coding. Related standardization work has started in the AVS Working Group.

RESOURCES

AVS documents and reference software can be found in [21]. AVS products information can be found in [22].

ACKNOWLEDGMENT

This research was sponsored by the National Science Foundation of China under award 61322106.

AUTHORS

Siwei Ma (swma@pku.edu.cn) is a professor at the National Engineering Lab of Video Technology, Peking University, China, and a cochair of the AVS Video Subgroup.

Tiejun Huang (tjhuang@pku.edu.cn) is a professor at the National Engineering Lab of Video Technology, Peking University, China, and the secretary-general of the AVS Working Group.

Cliff Reader (cliff@reader.com) is an adjunct professor at the National Engineering Lab of Video Technology, and the chair of the AVS Intellectual Property Rights Subgroup.

Wen Gao (wgao@pku.edu.cn) is a professor at the National Engineering Lab of Video Technology, Peking University, China, and the chair of the AVS Working Group.

REFERENCES

[1] ITU-T, "HSTP-MCTB Media coding toolbox for IPTV: Audio and video codecs," technical paper, ITU-T Study Group 16 Working Party 3 meeting, Geneva, Switzerland, 10 July 2009.

[2] S. Ma, S. Wang, and W. Gao, "Overview of IEEE 1857 video coding standard," in *Proc. IEEE Int. Conf. Image Processing,* Melbourne, Australia, Sept. 2013, pp. 1500–1504.

[3] Q. Yu, S. Ma, Z. He, Y. Ling, Z. Shao, L. Yu, W. Li, X. Wang, Y. He, M. Gao, X. Zheng, J. Zheng, I.-K. Kim, S. Lee, and J. Park, "Suggested video platform for AVS2," in *Proc. 42nd AVS Meeting*, Guilin, China, Sept. 2012, AVS_M2972.

[4] Q. Yu, X. Cao, W. Li, Y. Rong, Y. He, X. Zheng, and J. Zheng, "Short distance intra prediction," in *Proc. 46th AVS Meeting*, Shenyang, China, Sept. 2013, AVS_M3171.

[5] Y. Piao, S. Lee, and C. Kim, "Modified intra mode coding and angle adjustment," in *Proc. 48th AVS Meeting*, Beijing, China, Apr. 2014, AVS_M3304.

[6] Y. Piao, S. Lee, I.-K. Kim, and C. Kim, "Derived mode (DM) for chroma intra prediction," in *Proc. 44th AVS Meeting,* Luoyang, China, Mar. 2013, AVS_M3042.

[7] Y. Lin and L. Yu, "F frame CE: Multi forward hypothesis prediction," in *Proc. 48th AVS Meeting*, Beijing, China, Apr. 2014, AVS_M3326.

[8] Z. Shao and L. Yu, "Multi-hypothesis skip/direct mode in P frame," in *Proc. 47th AVS Meeting*, Shenzhen, China, Dec. 2013, AVS_M3256.

[9] Y. Ling, X. Zhu, L. Yu, J. Chen, S. Lee, Y. Piao, and C. Kim, "Multi-hypothesis mode for AVS2," in *Proc. 47th AVS Meeting*, Shenzhen, China, Dec. 2013, AVS_M3271.

[10] I.-K. Kim, S. Lee, Y. Piao, and C. Kim, "Directional multi-hypothesis prediction (DMH) for AVS2," in *Proc. 45th AVS Meeting*, Taicang, China, June 2013, AVS_M3094.

[11] H. Lv, R. Wang, Z. Wang, S. Dong, X. Xie, S. Ma, and T. Huang, "Sequence level adaptive interpolation filter for motion compensation," in *Proc. 47th AVS Meeting*, Shenzhen, China, Dec. 2013, AVS_M3253.

[12] Z. Wang, H. Lv, X. Li, R. Wang, S. Dong, S. Ma, T. Huang, and W. Gao, "Interpolation improvement for chroma motion compensation," in *Proc. 48th AVS Meeting*, Beijing, China, Apr., 2014, AVS_M3348.

[13] J. Ma, S. Ma, J. An, K. Zhang, and S. Lei, "Progressive motion vector precision," in *Proc. 44th AVS Meeting*, Luoyang, China, Mar. 2013, AVS_M3049.

[14] S. Lee, I.-K. Kim, M.-S. Cheon, N. Shlyakhov, and Y. Piao, "Proposal for AVS2.0 reference software," in *Proc. 42nd AVS Meeting*, Guilin, China, Sept. 2012, AVS_M2973.

[15] W. Li, Y. Yuan, X. Cao, Y. He, X. Zheng, and J. Zhen, "Non-square quad-tree transform," in *Proc. 45th AVS Meeting*, Taicang, China, June 2013, AVS_M3153.

[16] J. Wang, X. Wang, T. Ji, and D. He, "Two-level transform coefficient coding," in *Proc. 43rd AVS Meeting*, Beijing, China, Dec. 2012, AVS_M3035.

[17] X. Wang, J. Wang, T. Ji, and D. He, "Intra prediction mode based context design," in *Proc. 45th AVS Meeting*, Taicang, China, June 2013, AVS_M3103.

[18] J. Chen, S. Lee, C. Kim, C.-M. Fu, Y.-W. Huang, and S. Lei, "Sample adaptive offset for AVS2," in *Proc. 46th AVS Meeting*, Shenyang, China, Sept. 2013, AVS_M3197.

[19] X. Zhang, J. Si, S. Wang, S. Ma, J. Cai, Q. Chen, Y.-W. Huang, and S. Lei, "Adaptive loop filter for AVS2," in *Proc. 48th AVS Meeting*, Beijing, China, Apr. 2014, AVS_M3292.

[20] S. Dong, L. Zhao, P. Xing, and X. Zhang, "Surveillance video coding platform for AVS2," in *Proc. 47th AVS Meeting*, Shenzhen, China, Dec. 2013, AVS_M3221.

[21] AVS Working Group Web Site. [Online]. Available: http://www.avs.org.cn

[22] AVS Industry Alliance Web Site. [Online]. Available: http://www.avsa.org.cn

[SP]

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
Asilomar Conference	169	www.asilomarssc.org	
IEEE MDL/Marketing	3	www.ieee.org/go/trymdl	
Mathworks	CVR 4	www.mathworks.com/ltc	+1 508 647 7040
Mini Circuits	CVR 2, 5, CVR 3	www.minicircuits.com	+1 718 934 4500

advertisers SALES OFFICES

James A. Vick
Sr. 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

Mindy Belfer
Advertising Sales Coordinator
Phone: +1 732 562 3937
Fax: +1 732 981 1855
m.belfer@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

Digital Object Identifier 10.1109/MSP.2014.2373676

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 ASIA/FAR EAST/PACIFIC RIM
Louise Smith
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
les.ieeemedia@ieee.org
Europe, Africa, Middle East
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: +224 616 3034;
Fax: +1 847 729 4269
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
Louise Smith
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
les.ieeemedia@ieee.org
Europe, Africa, Middle East

[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.)

2015

[APRIL]

Data Compression Conference (DCC)
7–9 April, Snowbird, Utah, United States.
URL: http://www.cs.brandeis.edu/~dcc/index.html

14th IEEE International Conference on Information Processing in Sensor Networks (IPSN)
13–17 April, Seattle, Washington, United States.
General Chair: Suman Nath
URL: http://ipsn.acm.org/2015

First IEEE Conference on Network Softwarization (NetSoft)
13–17 April, London, United Kingdom.
General Cochairs: Prosper Chemouil and George Pavlou
URL: http://sites.ieee.org/netsoft/

12th IEEE International Symposium on Biomedical Imaging (ISBI)
16–19 April, Brookyln, New York, United States.
General Chairs: Elsa Angelini and Jelena Kovacevic
URL: http://biomedicalimaging.org/2015/

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
19–24 April, Brisbane, Australia.
General Cochairs: Vaughan Clarkson and Jonathan Manton
URL: http://icassp2015.org/

[MAY]

31st Picture Coding Symposium (PCS)
31 May–3 June, Cairns, Australia.
General Chairs: David Taubman and Mark Pickering
URL: http://www.pcs2015.org

[JUNE]

Third IEEE International Workshop on Compressed Sensing Theory and Its Applications to Radar, Sonar, and Remote Sensing (CoSeRa)
22–24 June, Pisa, Italy.
General Chairs: Fulvio Gini and Joachim Ender
URL: http://www.cosera2015.iet.unipi.it/

16th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)
28 June–1 July, Stockholm, Sweden.
General Chairs: Joakim Jaldén and Björn Ottersten
URL: http://www.spawc2015.org/

IEEE International Conference on Multimedia and Expo (ICME)
29 June–3 July, Turin, Italy.
General Chairs: Enrico Magli, Stefano Tubaro, and Anthony Vetro
URL: http://www.icme2015.ieee-icme.org/index.php

[JULY]

Third IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)
12–15 July, Chengdu, China.
General Chairs: Yingbo Hua and Dezhong Yao
URL: http://www.chinasip2015.org/

[AUGUST]

IEEE Signal Processing and SP Education Workshop (SPW)
9–12 August, Salt Lake City, Utah, United States.
General Chair: Todd Moon
URL: http://spw2015.coe.utah.edu

12th IEEE International Conference on Advanced Video- and Signal-Based Surveillance (AVSS)
25–28 August, Karlsruhe, Germany.
General Chairs: Jürgen Beyerer and Rainer Stiefelhagen
URL: http://avss2015.org

[SEPTEMBER]

Sensor Signal Processing for Defence (SSPD)
9–10 September, Edinburgh, United Kingdom.
http://www.see.ed.ac.uk/drupal/udrc/sspd/

IEEE International Conference on Image Processing (ICIP)
28 September–1 October, Quebec City, Quebec, Canada.
URL: http://www.icip2015.org/

[OCTOBER]

IEEE International Workshop on Multimedia Signal Processing (MMSP)
19–21 October, Xiamen, China.
General Chairs: Xiao-Ping Zhang, Oscar C. Au, and Jonathan Li
URL: http://www.mmsp2015.org/

[DECEMBER]

IEEE 6th International Workshop on Computational Advances in Multisensor Adaptive Processing (CAMSAP)
13–16 December, Cancun, Mexico.

IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)
13–17 December, Scottsdale, Arizona, United States.
URL: http://www.asru2015.org/

IEEE Global Conference on Signal and Information Processing (GlobalSIP)
14–16 December, Orlando, Florida, United States.
General Chairs: José M.F. Moura and Dapeng Oliver Wu

2016

[MARCH]

41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
21–25 March, Shanghai, China.
URL: http://dmlab.sjtu.edu.cn/icassp/icassp2016.html

[SP]

Digital Object Identifier 10.1109/MSP.2014.2372551
Date of publication: 12 February 2015

USB & ETHERNET RF SWITCH MATRIX

DC to 18 GHz from $385 ea.

We're adding more models and more functionality to our line of RF switch matrices. All models now feature switch cycle counting with automatic calibration interval alerts based on actual usage, an industry first! This function improves test reliability and saves you money. Our new RC-series models feature both USB and Ethernet control, so you can run your test setup from anywhere in the world! Rugged aluminum cases on all models house our patented mechanical switches with extra-long life of 10 years/100 million cycles of guaranteed performance!*

Our easy-to-install, easy-to-use GUI will have you up and running in minutes for step-by-step control, full automation, or remote operation. They're fully compatible with most third-party lab software,† adding capabilities and efficiency to existing setups with ease! Visit minicircuits.com today for technical specifications, performance data, quantity pricing, and real time availability – or call us to discuss your custom programming needs – and think how much time and money you can save!

USB Control Switch Matrices

Model	# Switches (SPDT)	IL (dB)	VSWR (:1)	Isolation (dB)	RF P_{MAX} (W)	Price $ (Qty. 1-9)
NEW USB-1SP4T-A18	1 (SP4T)	0.25	1.2	85	2	795.00
USB-1SPDT-A18	1	0.25	1.2	85	10	385.00
USB-2SPDT-A18	2	0.25	1.2	85	10	685.00
USB-3SPDT-A18	3	0.25	1.2	85	10	980.00
USB-4SPDT-A18	4	0.25	1.2	85	10	1180.00
USB-8SPDT-A18	8	0.25	1.2	85	10	2495.00

NEW USB and Ethernet Control Switch Matrices

Model	# Switches (SPDT)	IL (dB)	VSWR (:1)	Isolation (dB)	RF P_{MAX} (W)	Price $ (Qty. 1-9)
RC-1SP4T-A18	1 (SP4T)	0.25	1.2	85	2	895.00
RC-2SP4T-A18	2 (SP4T)	0.25	1.2	85	2	2195.00
RC-1SPDT-A18	1	0.25	1.2	85	10	485.00
RC-2SPDT-A18	2	0.25	1.2	85	10	785.00
RC-3SPDT-A18	3	0.25	1.2	85	10	1080.00
RC-4SPDT-A18	4	0.25	1.2	85	10	1280.00
RC-8SPDT-A18	8	0.25	1.2	85	10	2595.00

*The mechanical switches within each model are offered with an optional 10 year extended warranty. Agreement required. See data sheets on our website for terms and conditions. Switches protected by US patents 5,272,458; 6,650,210; 6,414,577; 7,633,361; 7,843,289; and additional patents pending.

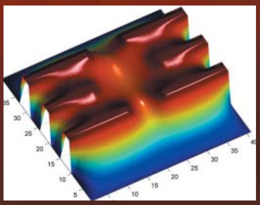
†See data sheet for a full list of compatible software.

NEW FEATURE
SWITCH CYCLE COUNTING

Mini-Circuits®

www.minicircuits.com P.O. Box 350166, Brooklyn, NY 11235-0003 (718) 934-4500 sales@minicircuits.com

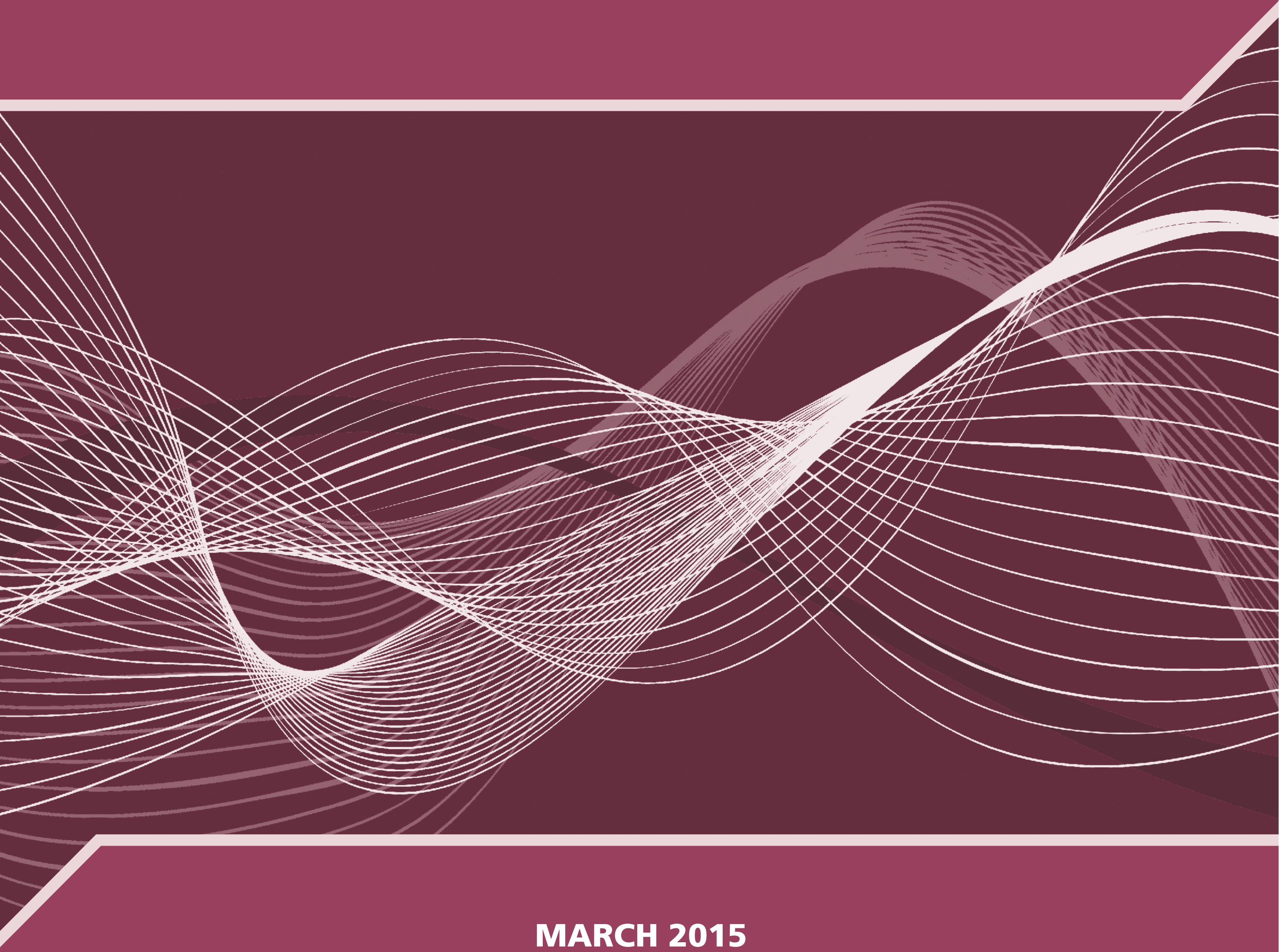
521 rev G



IEEE SIGNAL PROCESSING SOCIETY

CONTENT GAZETTE

[ISSN 2167-5023]



MARCH 2015

IEEE Signal Processing Society®

IEEE

ASRU 2015

IEEE Automatic Speech Recognition and Understanding Workshop

December 13-17, 2015 Scottsdale, Arizona, USA

http://asru2015.org

Call for Papers

The fourteenth biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU) will be held on December 13-17, 2015 in Scottsdale, Arizona - USA. The ASRU workshop meets every two years and has a tradition of bringing together researchers from academia and industry in an intimate and collegial setting to discuss problems of common interest in **automatic speech recognition, understanding, and related fields of research**.

Topics and focus

Authors are encouraged to submit contributions in all areas of spoken language processing, with emphasis placed on the following topics:

- Automatic speech recognition
- Spoken language understanding
- Speech-to-text systems
- Spoken dialog systems
- Multilingual language processing
- Robustness in automatic speech recognition
- Spoken document retrieval
- Speech-to-speech translation
- Text-to-speech systems
- Spontaneous speech processing
- Speech summarization
- New applications of automatic speech recognition

Format

The workshop features one keynote and one or two invited talks a day. Regular papers are presented as posters. See http://asru2015.org for formatting guidelines. ASRU 2015 will also include challenge tasks, panel discussions and demo sessions.

Paper Submission

Prospective authors are invited to submit full-length, 4-6 page papers, including figures, plus 1-2 additional pages for references only. All papers will be handled and reviewed electronically.

Schedule

Paper due date: **Friday July 10, 2015**
Paper Notification: Friday Sept 11, 2015
Registration opens: Friday Sept 11, 2015
Demo/toolkit deadline: Friday Sept 25, 2015
Paper Camera ready version due: Friday Oct 2, 2015
Demo/toolkit notification date: Friday Oct 9, 2015
Author and early registration end: Friday Oct 23, 2015
Demo/toolkit camera ready version due: Monday Oct 26, 2015
Workshop: Dec 13-17, 2015

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed

MEDLINE U.S. National Library of Medicine

JANUARY 15, 2015 VOLUME 63 NUMBER 2 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

IEEE TRANSACTIONS ON SIGNAL PROCESSING (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $602.50 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

FEBRUARY 1, 2015 VOLUME 63 NUMBER 3 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

IEEE TRANSACTIONS ON SIGNAL PROCESSING (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $602.50 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

Association for Computing Machinery

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed

MEDLINE U.S. National Library of Medicine

JANUARY 2015 VOLUME 23 NUMBER 1 ITASFA (ISSN 2329-9290)

EDITORIAL

REGULAR PAPERS

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (ISSN 2329-9290) is published bimonthly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $339.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE/ACM TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society ® Association for Computing Machinery www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

FEBRUARY 2015 VOLUME 23 NUMBER 2 ITASFA (ISSN 2329-9290)

REGULAR PAPERS

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (ISSN 2329-9290) is published bimonthly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $339.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE SAM 2016
Rio de Janeiro
Brazil

The Ninth IEEE Sensor Array and Multichannel Signal Processing Workshop

IEEE

10th-13th July 2016,Rio de Janeiro, Brazil

IEEE Signal Processing Society

Call for Papers

General Chairs
Rodrigo C. de Lamare,
PUC-Rio, Brazil and University of York, United Kingdom

Martin Haardt,
TU Ilmenau, Germany

Technical Chairs
Aleksandar Dogandzic,
Iowa State University, USA

Vítor Nascimento,
University of São Paulo, Brazil

Special Sessions Chair
Cédric Richard,
University of Nice, France

Publicity Chair
Maria Sabrina Greco,
Universityof Pisa, Italy

Important Dates

Special Session Proposals
29thJanuary , 2016

Submission of Papers
26thFebruary, 2016

Notification of Acceptance
29thApril , 2016

Final Manuscript Submission
16th May, 2016

Advance Registration
16th May, 2016

Technical Program
The SAM Workshop is an important IEEE Signal Processing Society event dedicated to sensor array and multichannel signal processing. The organizing committee invites the international community to contribute with state-of-the-art developments in the field. SAM 2016 will feature plenary talks by leading researchers in the field as well as poster and oral sessions with presentations by the participants.

Welcome to Rio de Janeiro! – The workshop will be held at the Pontifical Catholic University of Rio de Janeiro, located in Gávea, in a superb area surrounded by beaches, mountains and the Tijuca National Forest, the world's largest urban forest. Rio de Janeiro is a world renowned city for its culture, beautiful landscapes, numerous tourist attractions and international cuisine. The workshop will take place during the first half of July about a month before the 2016 Summer Olympic Games when Rio will offer plenty of cultural activities and festivities, which will make SAM 2016 a memorable experience.

Research Areas
Authors are invited to submit contributions in the following areas:
□ Adaptive beamforming
□ Array processing for biomedical applications
□ Array processing for communications
□ Blind source separation and channel identification
□ Computational and optimization techniques
□ Compressive sensing and sparsity-based signal processing
□ Detection and estimation
□ Direction-of-arrival estimation
□ Distributed and adaptive signal processing
□ Intelligentsystems and knowledge-based signal processing
□ Microphone and loudspeaker array applications
□ MIMO radar
□ Multi-antenna systems: multiuser MMO, massive MIMO and space-time coding
□ Multi-channel imaging and hyperspectral processing
□ Multi-sensor processing for smart grid and energy
□ Non-Gaussian, nonlinear, and non-stationary models
□ Performance evaluations with experimental data
□ Radar and sonar array processing
□ Sensor networks
□ Source Localization, Classification and Tracking
□ Synthetic aperture techniques
□ Space-time adaptive processing
□ Statistical modelling for sensor arrays
□ Waveform diverse sensors and systems

Submission of papers – Full-length four-page papers will be accepted only electronically.

Special session proposals – They should be submitted by e-mail to the Technical Program Chairs and the Special Sessions Chair and include a topical title, rationale, session outline, contact information, and list of invited speakers.

IEEE TRANSACTIONS ON IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society® www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

JANUARY 2015 VOLUME 24 NUMBER 1 IIPRE4 (ISSN 1057-7149)

PAPERS

IEEE

IEEE TRANSACTIONS ON IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

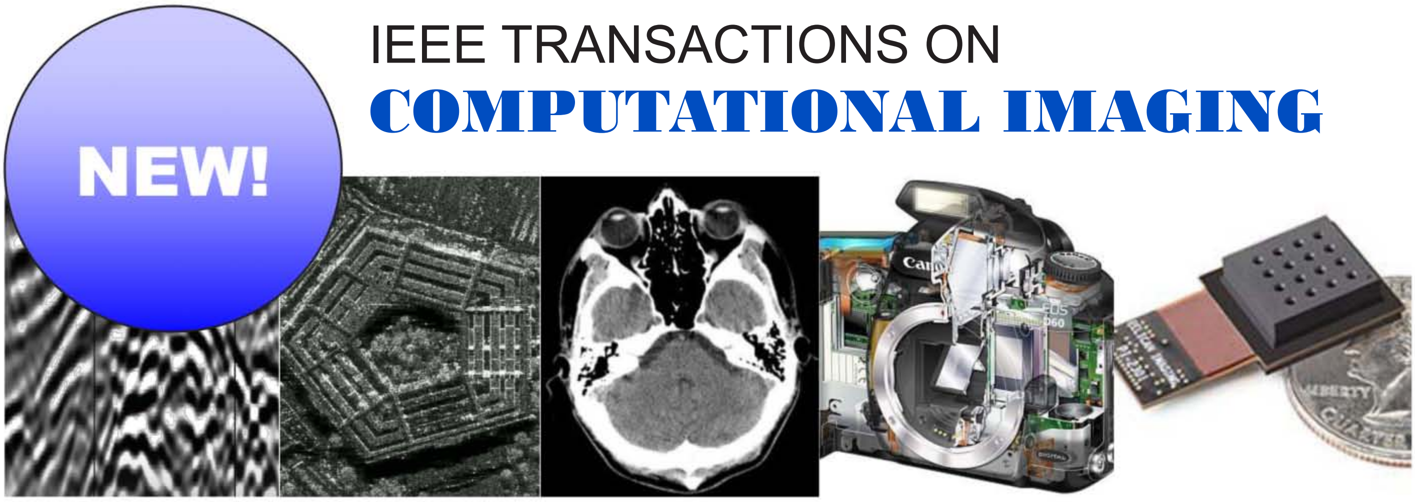
www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

FEBRUARY 2015 VOLUME 24 NUMBER 2 IIPRE4 (ISSN 1057-7149)

PAPERS

IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING



The new IEEE Transactions on Computational Imaging seeks original manuscripts for publication. This new journal will publish research results where computation plays an integral role in the image formation process. All areas of computational imaging are appropriate, ranging from the principles and theory of computational imaging, to modeling paradigms for computational imaging, to image formation methods, to the latest innovative computational imaging system designs. Topics of interest include, but are not limited to the following:

Imaging Models and Representation

- Statistical-model based methods
- System and image prior models
- Noise models
- Graphical and tree-based models
- Perceptual models

Computational Sensing

- Coded source methods
- Structured light
- Coded aperture methods
- Compressed sensing
- Light-field sensing
- Plenoptic imaging
- Hardware and software systems

Computational Image Creation

- Sparsity-based methods
- Statistically-based inversion methods, Bayesian regularization
- Super-resolution, multi-image fusion
- Learning-based methods, Dictionary-based methods
- Optimization-based methods; proximal iterative methods, ADMM

Computational Photography

- Non-classical image capture, Generalized illumination
- Time-of-flight imaging
- High dynamic range imaging
- Focal stacks

Computational Consumer Imaging

- Cell phone imaging
- Camera-array systems
- Depth cameras

Computational Acoustic Imaging

- Multi-static ultrasound imaging
- Photo-acoustic imaging
- Acoustic tomography

Computational Microscopic Imaging

- Holographic microscopy
- Quantitative phase imaging
- Multi-illumination microscopy
- Lensless microscopy

Tomographic Imaging

- X-ray CT
- PET
- SPECT

Magnetic Resonance Imaging

- Diffusion tensor imaging
- Fast acquisition

Radar Imaging

- Synthetic aperture imaging
- Inverse synthetic imaging
- Terahertz imaging

Geophysical Imaging

- Multi-spectral imaging
- Ground penetrating radar
- Seismic tomography

Multi-spectral Imaging

- Multi-spectral imaging
- Hyper-spectral imaging
- Spectroscopic imaging

Editor-in-Chief: W. Clem Karl, Boston University.
To submit a paper go to: https://mc.manuscriptcentral.com/tci-ieee

IEEE Signal Processing Society ® EMB GRSS

IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

JANUARY 2015 VOLUME 10 NUMBER 1 ITIFA6 (ISSN 1556-6013)

EDITORIAL

PAPERS

IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

www.signalprocessingsociety.org

FEBRUARY 2015 VOLUME 10 NUMBER 2 ITIFA6 (ISSN 1556-6013)

PAPERS

ANNOUNCEMENTS

IEEE GlobalSIP'15–Call for Papers

2015 IEEE Global Conference on Signal and Information Processing – Orlando Florida
General Chairs: Jose Moura and Dapeng Oliver Wu
Technical Program Chairs: Mihaela van der Schaar, Xiaodong Wang, and Hsiao-Chun Wu

The IEEE Global Conference on Signal and Information Processing (GlobalSIP) is a recently launched flagship conference of the *IEEE Signal Processing Society*. GlobalSIP' 15 will be held in Orlando, Florida, USA, December 14-16, 2015. The conference will focus broadly on signal and information processing with an emphasis on up-and-coming signal processing themes. The conference will feature world-class speakers, tutorials, exhibits, and technical sessions consisting of poster or oral presentations. GlobalSIP' 15 technical program will be comprised of a main program and several co-located symposia on special topics. Technical paper submissions are solicited in the interest topics which may include, but are not limited to:

- Signal processing in communications and networks, including green communication and Signal processing in optical communication
- Image and video processing
- Selective topics in speech and language processing
- Signal processing in security applications
- Signal processing in finance
- Signal processing in energy and power systems
- Signal processing in genomics and bioengineering (physiological, pharmacological and behavioral)
- Signal processing for social media networks
- Neural signal processing
- Seismic signal processing
- Selective topics in statistical signal processing
- Graph-theoretic signal processing
- Machine learning
- Compressed sensing, sparsity analysis, and applications
- Big data processing, heterogeneous information processing and informatics
- Human machine interfaces
- Multimedia transmission, indexing and retrieval, and playback challenges
- Hardware and real-time implementations
- Other novel and significant Applications of selected areas of signal processing

Submission of Papers: Prospective authors are invited to submit full-length papers, with up to four pages for technical content including figures and possible references, and with one additional optional 5th page containing only references. Manuscripts should be original (not submitted/published anywhere else) and written in accordance with the standard IEEE double-column paper template. All paper submissions should be carried out through EDAS system (http://edas.info). A selection of best papers and best student papers will be made by the GlobalSIP 2015 best paper award committee upon recommendations from Technical Committees.

Notice: The IEEE Signal Processing Society enforces a "no-show" policy. Any accepted paper included in the final program is expected to have at least one author or qualified proxy attend and present the paper at the conference. Authors of the accepted papers included in the final program who do not attend the conference will be subscribed to a "No-Show List", compiled by the Society. The "no-show" papers will not be published by IEEE on IEEEXplore or other public access forums, but these papers will be distributed as part of the on-site electronic proceedings and the copyright of these papers will belong to the IEEE.

Timeline for paper submission:

May 15, 2015:	Paper submission deadline
June 30, 2015:	Review results announced
September 5, 2015:	Camera-ready papers due

IEEE TRANSACTIONS ON MULTIMEDIA

A PUBLICATION OF
THE IEEE CIRCUITS AND SYSTEMS SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY

http://www.signalprocessingsociety.org/tmm/

JANUARY 2015 VOLUME 17 NUMBER 1 ITMUF8 (ISSN 1520-9210)

EDITORIAL

PAPERS

3-D Audio/Video Processing

Sparse Signal Processing

Multimodal Human-Machine Interfaces and Interaction

Quality Assessment and User Experience

Content Description and Annotation

Multimedia Search and Retrieval

Transport and Automotive

Immersive Communication and Networking

Wireless/Mobile Multimedia

CORRESPONDENCES

3-D Audio/Video Processing

Quality Assessment and User Experience

ANNOUNCEMENTS

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society®

www.ieee.org/sp/index.html

FEBRUARY 2015 VOLUME 9 NUMBER 1 IJSTGY (ISSN 1932-4553)

ISSUE ON VISUAL SIGNAL PROCESSING FOR WIRELESS NETWORKS

EDITORIAL

PAPERS

IEEE Journal of Selected Topics in Signal Processing (ISSN 1932-4553) is published eight times a year in February, March, April, June, August, September, October, and December by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $371.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE Journal of Selected Topics in Signal Processing, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: PitneyBowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ONM5W3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.ieee.org/sp/index.html

FEBRUARY 2015 VOLUME 22 NUMBER 2 ISPLEM (ISSN 1070-9908)

LETTERS

IEEE SIGNAL PROCESSING LETTERS (ISSN 1070–9908) is published quarterly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $309.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE SIGNAL PROCESSING LETTERS, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.ieee.org/sp/index.html

MARCH 2015 VOLUME 22 NUMBER 3 ISPLEM (ISSN 1070-9908)

LETTERS

IEEE SIGNAL PROCESSING LETTERS (ISSN 1070–9908) is published quarterly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $309.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE SIGNAL PROCESSING LETTERS, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

CORRESPONDENCE

Image and Multidimensional Signal Processing

2015 IEEE International Workshop on Multimedia Signal Processing

Xiamen, China, October 19 – October 21, 2015

http://www.mmsp2015.org

General Chairs
Xiao-Ping Zhang – *Ryerson U, Canada*
Oscar C. Au – *HKUST, Hong Kong*
Jonathan Li – *Xiamen U, China*
Technical Chairs
Tao Mei – *Microsoft Research Asia*
Gene Cheung – *NII, Japan*
Special Session Chairs
John Paisley – *Columbia U, USA*
Yap-Peng Tan – *NTU, Singapore*
Overview Chairs
Homer Chen–*NTU, Taiwan*
Anthony Vetro – *MERL, USA*
Local Arrangement Chair
Xinghao Ding –*Xiamen U, China*
Rongrong Ji – *Xiamen U, China*
Finance Chairs
Chia-Wen Lin – *NTHU, Taiwan*
Yue Huang – *Xiamen U, China*
Publications Chairs
Vicky Zhao – *U. Alberta, Canada*
Delu Zeng – *Xiamen U, China*
Publicity Chairs
Lina Stankovic – *U. Strathclyde, UK*
Ivan Bajic – *Simon Fraser U,. Canada*
Registration Chair
Liujuan Cao – *Xiamen U, China*
Demo Chair
Wenxin Hong – *Xiamen U, China*
Industry Liaison
Alexander Loui – *Kodak, USA*
North America Liaison
Antonio Ortega, *USC, USA*
Asia Liason
Feng Wu – *USTC, China*
Europe Liaison
Fernando Pereira – *IST-IT, Portugal*

Tentative Call for Papers

MMSP 2015 is the 17th International Workshop on Multimedia Signal Processing. The workshop is organized by the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. This year's event has a ***Heterogeneous Big Data Analytics in Multimedia*** theme. The workshop will bring together researchers and developers in multimedia signal processing and applications to share their latest achievements and explore future directions and synergies in these exciting areas.

Papers are solicited in (but not limited to) the following topics, covering this year's theme and the general scope of multimedia signal processing:

- Theories and applications for heterogeneous big media data analytics
- Semantic extraction and knowledge mining from heterogeneous big media data
- Massive-scale media detection and recognition
- Content-based analysis, retrieval and annotation for big media data
- Feature learning for heterogeneous big media data representation
- Multimedia security, forensic, privacy for big data
- Multimedia quality assessment and enhancement
- Affective computing and cross-media sentiment analysis
- Media algorithm optimization and complexity analysis
- Multimedia in economics, finance, business analytics
- Multimedia signals in geomatics
- Image/video coding and processing
- Speech/audio recognition and processing
- Multimedia communications and interactions

Top 10% Paper Award

This award is granted to as many as 10% of the total paper submissions, and is open to all accepted papers. Papers will be evaluated based on originality, technical contribution, and presentation quality during the workshop.

Paper Submission

Prospective authors should submit full-length papers of 6 pages in two-column IEEE format, including author affiliation and address, figures, tables and references, to the submission website. Only electronic submissions are accepted. Paper submission implies the intent of at least one of the authors to register and present the paper, if accepted.

Important Dates

Proposals for Special Sessions:	March 20, 2015
Submission of Paper:	May 28, 2015
Notification of acceptance:	July 6, 2015

IEEE SignalProcessing MAGAZINE

[VOLUME 32 NUMBER 2 MARCH 2015]

ASSISTED LISTENING

SIGNAL PROCESSING TECHNIQUES

COMPOSITIONAL MODELS FOR AUDIO PROCESSING

TENSOR DECOMPOSITIONS FOR SIGNAL PROCESSING APPLICATIONS

A MEDICAL SENSOR REVOLUTION

CRAMÉR–RAO BOUND ANALOG OF BAYES' RULE

IEEE Signal Processing Society®

IEEE

[CONTENTS]

[VOLUME 32 NUMBER 2]

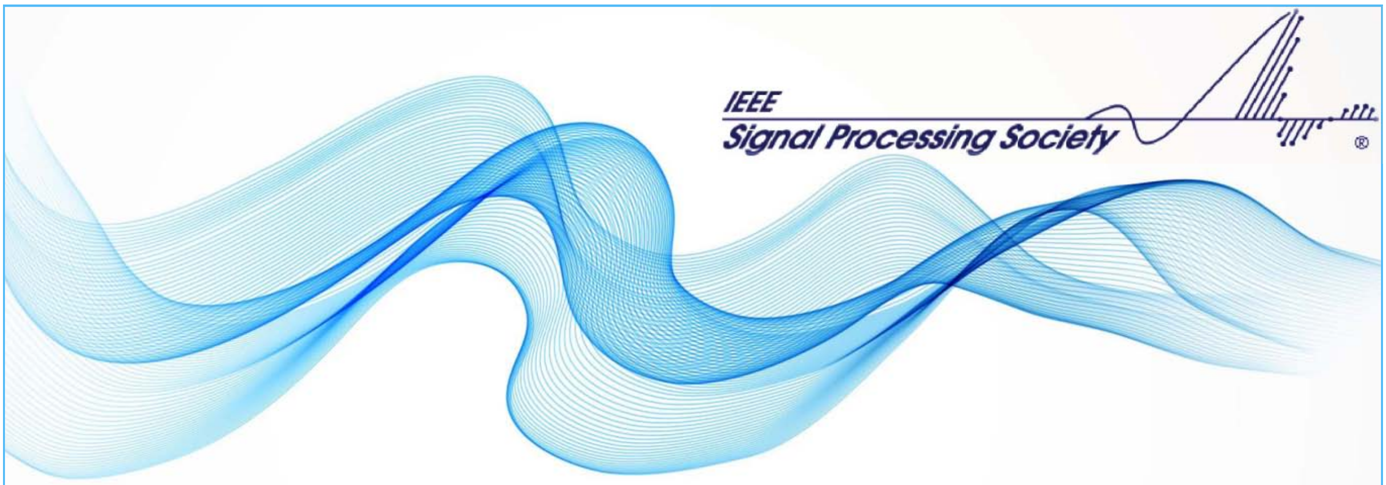
[SPECIAL SECTION—SIGNAL PROCESSING TECHNIQUES FOR ASSISTED LISTENING]

[FEATURES]

[COLUMNS]

[DEPARTMENTS]

Digital Object Identifier 10.1109/MSP.2014.2372532



FREE SPS STUDENT MEMBERSHIP FOR 2015

You're in the beginning stages of your career. Membership in the IEEE Signal Processing Society can help you lay the groundwork for many years of success. You can have it all in 2015 - and for free! Membership includes:

- Discounts on conference registration fees;
- Eligibility to apply for travel grants to attend SPS flagship conferences including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and IEEE International Conference on Image Processing (ICIP);
- Networking and job opportunities at the ICASSP Student Career Luncheon;
- Eligibility to enter our student competition, the Signal Processing Cup, for a US$5,000 grand prize;
- Involvement opportunities through SPS's local Chapters - more than 130 worldwide;
- Free electronic and digital subscriptions to *IEEE Signal Processing Magazine*, *Inside Signal Processing eNewsletter*, and the *IEEE Signal Processing Society Content Gazette*;
- Access to cutting-edge educational resources, including SigView, SPS's online video tutorial portal.

See everything Signal Processing Society membership can do for you:
http://signalprocessingsociety.org

Already an IEEE member? Join SPS for free now!

(You must have already renewed your IEEE membership for 2015 to use this offer)

- Visit http://ieee.org/join
- On the left side, click "Societies and Special Interest Groups"
- Click "IEEE Signal Processing Society," then "Join the IEEE Signal Processing Society"
- When you reach the catalog page, click "Add Item(s)" and sign in with your IEEE account *Note: Free offer applies only to basic membership. For US$8.00, enhance your membership for more great benefits!*
- Once logged in, click "Proceed to Checkout"
- When you reach the shopping cart, enter the promotion code SP15STUAD and click "Apply"
- Complete check out and congratulations! Welcome to SPS!

Not yet an IEEE Student Member?

Get a free SPS membership with the purchase of an IEEE Student membership!

- Visit http://ieee.org/join
- Click "Join as a student" on the bottom right to create your new IEEE Student member account
- After your IEEE account is created, complete the membership application and proceed to "Do you want to add any memberships and subscriptions?"
- Select "Signal Processing Society membership" and click "add selected item"
- Click "Proceed to Checkout"
- When you reach the shopping cart, enter the promotion code SP15STUNW and click "Apply"
- Complete check out and congratulations! Welcome to SPS!

***Note:** Must be an active IEEE Student or Graduate Student member. This offer does not apply to SPS Students or Graduate Students renewing for 2015.*

IEEE

CAMSAP 2015

Call for Papers

The Sixth IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing

Cancun, Mexico
December 13 – 16, 2015
http://inspire.rutgers.edu/CAMSAP2015

Following the success of the first five editions of the IEEE workshop on Computational Advances in Multi-Sensor Adaptive Processing, we are pleased to announce the sixth workshop in this series. IEEE CAMSAP 2015 will be held in historic Cancun, Mexico, and will feature a number of plenary talks from the world's leading researchers in the area, special focus sessions, and contributed papers. All papers will undergo peer review in order to provide feedback to the authors and ensure a high-quality program.

Topics and applications of interest for the workshop include, but are not limited to, the following.

TOPICS OF INTEREST

- Array processing, waveform diversity, space-time adaptive processing
- Convex optimization and relaxation
- Computational linear & multi-linear algebra
- Computer-intensive methods in signal processing (bootstrap, MCMC, EM, particle filtering, etc.)
- Signal and information processing over networks
- Sparse signal processing

APPLICATIONS

- Big data
- Biomedical signal processing
- Communication systems
- Computational imaging
- Radar
- Sensor networks
- Smart grids
- Sonar

Submission of Papers: Prospective authors are invited to submit original full-length papers, with up to four pages for technical content including figures and references, using the formatting guidelines on the website for reviewing purposes. All accepted papers must be presented at the workshop to appear in the proceedings. Best student paper awards, selected by a CAMSAP committee, will also be presented at the workshop.

Special Session Proposals: In addition to contributed sessions, the workshop will also have a number of special sessions. Prospective organizers of special sessions are invited to submit a proposal form, available on the workshop website, by e-mail to the Special Sessions Chair.

IMPORTANT DEADLINES

Submission of proposals for special sessions	February 15, 2015
Notification of special session acceptance	March 15, 2015
Submission of papers	June 15, 2015
Notification of paper acceptance	September 11, 2015
Final paper submission	October 11, 2015

IEEE Signal Processing Society

General Chairs

Petar M. Djurić
petar.djuric@stonybrook.edu
Stony Brook University,
USA

Jean-Yves Tourneret
jean-yves.tourneret@enseeiht.fr
University of Toulouse,
France

Technical Program Chairs

Fulvio Gini
f.gini@ing.unipi.it
University of Pisa,
Italy

Cédric Richard
cedric.richard@unice.fr
Nice Sophia-Antipolis University,
France

Finance Chair

Marius Pesavento
mpesa@nt.tu-darmstadt.de
University of Darmstadt,
Germany

Special Sessions Chair

Maria Sabrina Greco
m.greco@iet.unipi.it
University of Pisa,
Italy

Publicity and Publications Chair

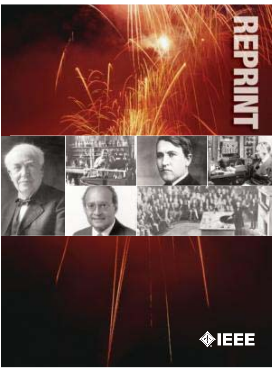
Waheed U. Bajwa
waheed.bajwa@rutgers.edu
Rutgers University,
USA

Local Arrangements Chair

Mónica F. Bugallo
monica.bugallo@stonybrook.edu
Stony Brook University,
USA

Join Twitter Conversation
#CAMSAP2015

IEEE



IEEE ORDER FORM FOR REPRINTS

Purchasing IEEE Papers in Print is easy, cost-effective and quick.

Complete this form, send via our secure fax (24 hours a day) to 732-981-8062 or mail it back to us.

PLEASE FILL OUT THE FOLLOWING

Author: ______________________________

Publication Title: ______________________________

Paper Title: ______________________________

____

RETURN THIS FORM TO:
IEEE Publishing Services
445 Hoes Lane
Piscataway, NJ 08855-1331

Email the Reprint Department at reprints@ieee.org for questions regarding this form

PLEASE SEND ME

■ 50 ■ 100 ■ 200 ■ 300 ■ 400 ■ 500 or ________(in multiples of 50) reprints.

■ YES ■ NO Self-covering/title page required. COVER PRICE: $74 per 100, $39 per 50.

■ $58.00 Air Freight must be added for all orders being shipped outside the U.S.

■ $21.50 must be added for all USA shipments to cover the cost of UPS shipping and handling.

PAYMENT

■ Check enclosed. Payable on a bank in the USA.

■ Charge my: ■ Visa ■ Mastercard ■ Amex ■ Diners Club

Account # ______________________________ Exp. date ______________

Cardholder's Name (please print): ______________________________

■ Bill me (you must attach a purchase order) Purchase Order Number ______________

Send Reprints to: Bill to address, if different:

______________________ ______________________

______________________ ______________________

______________________ ______________________

______________________ ______________________

Because information and papers are gathered from various sources, there may be a delay in receiving your reprint request. This is especially true with postconference publications. Please provide us with contact information if you would like notification of a delay of more than 12 weeks.

Telephone: ______________Fax: ______________ Email Address: ______________

2012 REPRINT PRICES (without covers)

Number of Text Pages

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40	41-44	45-48
50	$129	$213	$245	$248	$288	$340	$371	$408	$440	$477	$510	$543
100	$245	$425	$479	$495	$573	$680	$742	$817	$885	$953	$1021	$1088

Larger quantities can be ordered. Email reprints@ieee.org with specific details.

Tax Applies on shipments of regular reprints to CA, DC, FL, MI, NJ, NY, OH and Canada (GST Registration no. 12534188).

Prices are based on black & white printing. Please call us for full color price quote, if applicable.

IEEE
Advancing Technology for Humanity

2015 IEEE MEMBERSHIP APPLICATION

(students and graduate students must apply online)

Start your membership immediately: Join online www.ieee.org/join

Please complete both sides of this form, typing or **printing in capital letters**. Use only English characters and abbreviate only if more than 40 characters and spaces per line. We regret that incomplete applications cannot be processed.

1 Name & Contact Information

Please PRINT your name as you want it to appear on your membership card and IEEE correspondence. As a key identifier for the IEEE database, circle your last/surname.

□ Male □ Female Date of birth (Day/Month/Year) ______/______/______

____
Title First/Given Name Middle Last/Family Surname

▼ Primary Address □ Home □ Business (All IEEE mail sent here)

____
Street Address

____
City State/Province

____
Postal Code Country

____
Primary Phone

____
Primary E-mail

▼ Secondary Address □ Home □ Business

____
Company Name Department/Division

____
Street Address City State/Province

____
Postal Code Country

____
Secondary Phone

____
Secondary E-mail

To better serve our members and supplement member dues, your postal mailing address is made available to carefully selected organizations to provide you with information on technical services, continuing education, and conferences. Your e-mail address is not rented by IEEE. Please check box only if you do not want to receive these postal mailings to the selected address. □

2 Attestation

I have graduated from a three- to five-year academic program with a university-level degree.
□ Yes □ No

This program is in one of the following fields of study:
□ Engineering
□ Computer Sciences and Information Technologies
□ Physical Sciences
□ Biological and Medical Sciences
□ Mathematics
□ Technical Communications, Education, Management, Law and Policy
□ Other (please specify): ________________

This academic institution or program is accredited in the country where the institution is located. □ Yes □ No □ Do not know

I have ______ years of professional experience in teaching, creating, developing, practicing, or managing within the following field:

□ Engineering
□ Computer Sciences and Information Technologies
□ Physical Sciences
□ Biological and Medical Sciences
□ Mathematics
□ Technical Communications, Education, Management, Law and Policy
□ Other (please specify): ________________

3 Please Tell Us About Yourself

Select the numbered option that best describes yourself. This information is used by IEEE magazines to verify their annual circulation. Please enter numbered selections in the boxes provided.

A. Primary line of business → □

1. Computers
2. Computer peripheral equipment
3. Software
4. Office and business machines
5. Test, measurement and instrumentation equipment
6. Communications systems and equipment
7. Navigation and guidance systems and equipment
8. Consumer electronics/appliances
9. Industrial equipment, controls and systems
10. ICs and microprocessors
11. Semiconductors, components, sub-assemblies, materials and supplies
12. Aircraft, missiles, space and ground support equipment
13. Oceanography and support equipment
14. Medical electronic equipment
15. OEM incorporating electronics in their end product (not elsewhere classified)
16. Independent and university research, test and design laboratories and consultants (not connected with a mfg. co.)
17. Government agencies and armed forces
18. Companies using and/or incorporating any electronic products in their manufacturing, processing, research or development activities
19. Telecommunications services, telephone (including cellular)
20. Broadcast services (TV, cable, radio)
21. Transportation services (airline, railroad, etc.)
22. Computer and communications and data processing services
23. Power production, generation, transmission and distribution
24. Other commercial users of electrical, electronic equipment and services (not elsewhere classified)
25. Distributor (reseller, wholesaler, retailer)
26. University, college/other educational institutions, libraries
27. Retired
28. Other________________

B. Principal job function → □

1. General and corporate management
2. Engineering management
3. Project engineering management
4. Research and development management
5. Design engineering management —analog
6. Design engineering management —digital
7. Research and development engineering
8. Design/development engineering —analog
9. Design/development engineering—digital
10. Hardware engineering
11. Software design/development
12. Computer science
13. Science/physics/mathematics
14. Engineering (not elsewhere specified)
15. Marketing/sales/purchasing
16. Consulting
17. Education/teaching
18. Retired
19. Other________________

C. Principal responsibility → □

1. Engineering and scientific management
2. Management other than engineering
3. Engineering design
4. Engineering
5. Software: science/mngmnt/engineering
6. Education/teaching
7. Consulting
8. Retired
9. Other ________________

D. Title → □

1. Chairman of the Board/President/CEO
2. Owner/Partner
3. General Manager
4. VP Operations
5. VP Engineering/Dir. Engineering
6. Chief Engineer/Chief Scientist
7. Engineering Management
8. Scientific Management
9. Member of Technical Staff
10. Design Engineering Manager
11. Design Engineer
12. Hardware Engineer
13. Software Engineer
14. Computer Scientist
15. Dean/Professor/Instructor
16. Consultant
17. Retired
18. Other________________

Are you now or were you ever a member of IEEE?
□ Yes □ No If yes, provide, if known:

____
Membership Number Grade Year Expired

4 Please Sign Your Application

I hereby apply for IEEE membership and agree to be governed by the IEEE Constitution, Bylaws, and Code of Ethics. I understand that IEEE will communicate with me regarding my individual membership and all related benefits. **Application must be signed.**

____
Signature Date

Over Please

Information for Authors

(Updated/Effective September 17, 2014)

The IEEE TRANSACTIONS ON SIGNAL PROCESSING is published online twice per month (semimonthly) covering advances in the theory and application of signal processing. The scope is reflected in the EDICS: the Editor's Information and Classification Scheme. Please consider the journal with the most appropriate scope for your submission.

Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Comment Correspondences (brief items that provide comment on a paper previously published in the TRANSACTIONS). Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.

Every manuscript must (a) provide a clearly defined statement of the problem being addressed, (b) state why it is important to solve the problem, and (c) give an indication as to how the current solution fits into the history of the problem, including bibliographic references to related work rather than restating established algorithms and scientific principles.

In order to be considered for review, a paper must be within the scope of the journal and represent a novel contribution. A paper is a candidate for an Immediate Rejection if it is of limited novelty, e.g. a straightforward combination of theories and algorithms that are well established and are repeated on a known scenario, no new experimental data or new application. Experimental contributions will be rejected without review if there is insufficient experimental data. The TRANSACTIONS are published in English. Papers that have a large number of typographical and/or grammatical errors will also be rejected without review.

By submission/resubmission of your manuscript to this TRANSACTIONS, you are acknowledging that you accept the rules established for publication of manuscripts, including agreement to pay all overlength page charges, color charges, and any other charges and fees associated with publication of the manuscript. Such charges are not negotiable and cannot be suspended.

New and revised manuscripts should be prepared following the "Manuscript Submission" guidelines below, and submitted to the online manuscript system ScholarOne Manuscripts. After acceptance, finalized manuscripts should be prepared following the "Final Manuscript Submission Guidelines" below. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will only access your manuscript electronically via the ScholarOne Manuscripts system.

Manuscript Submission. Please follow the next steps.

1. *Account in ScholarOne Manuscripts*. If necessary, create an account in the on-line submission system ScholarOne Manuscripts. Please check first if you already have an existing account which is based on your e-mail address and may have been created for you when you reviewed or authored a previous paper.
2. *Electronic Manuscript*. Prepare a PDF file containing your manuscript in double-column, single-spaced format using a font size of 10 points or larger, having a margin of at least 1 inch on all sides. For a regular paper, the manuscript may not exceed 13 double-column pages, including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references.

   Upload this version of the manuscript as a PDF file "double.pdf" to the ScholarOneManuscripts site. You are encouraged to also submit a single-column, double-spaced version (11 point font or larger), but page length restrictions will be determined by the double-column version.

   For regular papers, the *revised* manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references.

   Proofread your submission, confirming that all figures and equations are visible in your document before you "SUBMIT" your manuscript. Proofreading is critical; once you submit your manuscript, the manuscript cannot be changed in any way. You may also submit your manuscript as a PostScript or MS Word file. The system has the capability of converting your files to PDF, however it is your responsibility to confirm that the conversion is correct and there are no font or graphics issues prior to completing the submission process.
3. *Additional Documents for Review*. Please upload pdf versions of all items in the reference list which are not publicly available, such as unpublished (submitted) papers. Other materials for review such as supplementary tables and figures may be uploaded as well. Reviewers will be able to view these files only if they have the appropriate software on their computers. Use short filenames without spaces or special characters. When the upload of each file is completed, you will be asked to provide a description of that file.
4. *Multimedia Materials*. IEEE Xplore can publish multimedia files (audio, images, video) and Matlab code along with your paper. Alternatively, you can provide the links to such files in a README file that appears on Xplore along with your paper. For details, please see http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect6 under "Multimedia." To make your work reproducible by others, the TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper. Files that are to be included with the final paper must be uploaded for consideration in the review process.
5. *Submission.* After uploading all files and proofreading them, submit your manuscript by clicking "Submit." A confirmation of the successful submission will open on screen containing the manuscript tracking number and will be followed with an e-mail confirmation to the corresponding and all contributing authors. Once you click "Submit," your manuscript cannot be changed in any way.
6. *Copyright Form and Consent Form.* By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with the IEEE copyright policies, authors are required to sign and submit a completed "IEEE Copyright and Consent Form" prior to publication by the IEEE.

   The IEEE recommends authors to use an effective electronic copyright form (eCF) tool within the ScholarOne Manuscripts system. You will be redirected to the "IEEE Electronic Copyright Form" wizard at the end of your original submission; please simply sign the eCF by typing your name at the proper location and click on the "Submit" button.

Comment Correspondence. Comment Correspondences provide brief comments on material previously published in the TRANSACTIONS. A comment correspondence may not exceed 2 pages in double-column, single double-spaced format, using 9 point type, with margins of 1 inch minimum on all sides, and including: title, names and contact information for authors, abstract, text, references, and an appropriate number of illustrations and/or tables. Comment Correspondences are submitted in the same way as regular manuscripts (see "Manuscript Submission" above for instructions).

Manuscript Length. Papers published on or after 1 January 2007 can now be up to 10 pages, and any paper in excess of 10 pages will be subject to over length page charges. The IEEE Signal Processing Society has determined that the standard manuscript length shall be no more than 10 published pages (double-column format, 10 point type) for a regular submission. Manuscripts that exceed these limits will incur mandatory over length page charges, as discussed below. Since changes recommended as a result of peer review may require additions to the manuscript, it is strongly recommended that you practice economy in preparing original submissions.

Exceptions to manuscript length requirements may, under extraordinary circumstances, be granted by the Editor-in-Chief. However, such exception does not obviate your requirement to pay any and all over length or additional charges that attach to the manuscript.

Resubmission of Previously Rejected Manuscripts. Authors of manuscripts rejected from any journal are allowed to resubmit their manuscripts only once. At the time of submission, you will be asked whether your manuscript is a new submission or a resubmission of an earlier rejected manuscript. If it is related to a manuscript previously rejected by any journal, you are expected to submit supporting documents identifying the previous submission and detailing how issues raised in the previous reviews have been addressed. Papers that do not disclose connection to a previously rejected paper or that do not provide documentation as to changes made may be immediately rejected.

Full details of the resubmission process can be found in the Signal Processing Society "Policy and Procedures Manual" at http://www.signalprocessingsociety.org/about/governance/policy-procedure/.

Author Misconduct.

Author Misconduct Policy: Plagiarism includes copying someone else's work without appropriate credit, using someone else's work without clear delineation of citation, and the uncited reuse of an authors previously published work that also involves other authors. Plagiarism is unacceptable.

Self-plagiarism involves the verbatim copying or reuse of an authors own prior work without appropriate citation; it is also unacceptable. Self-plagiarism includes duplicate submission of a single journal manuscript to two different journals, and submission of two different journal manuscripts which overlap substantially in language or technical contribution.

Authors may only submit original work that has not appeared elsewhere in a journal publication, nor is under review for another journal publication. Limited overlap with prior journal publications with a common author is allowed only if it is necessary for the readability of the paper. If authors have used their own previously published work as a basis for a new submission, they are required to cite the previous work and very briefly indicate how the new submission offers substantively novel contributions beyond those of the previously published work.

It is acceptable for conference papers to be used as the basis for a more fully developed journal submission. Still, authors are required to cite related prior work; the papers cannot be identical; and the journal publication must include novel aspects.

Digital Object Identifier 10.1109/TSP.2014.2358875

Author Misconduct Procedures: The procedures that will be used by the Signal Processing Society in the investigation of author misconduct allegations are described in the IEEE SPS Policies and Procedures Manual.

Author Misconduct Sanctions: The IEEE Signal Processing Society will apply the following sanctions in any case of plagiarism, or in cases of self-plagiarism that involve an overlap of more than 25% with another journal manuscript:

1) immediate rejection of the manuscript in question;
2) immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors; immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors;
3) prohibition against each of the authors for any new submissions, either individually, in combination with the authors of the plagiarizing manuscript, or in combination with new co-authors, to all of the Society's publications (journals, conferences, workshops). The prohibition shall continue for one year from notice of suspension.

Further, plagiarism and self-plagiarism may also be actionable by the IEEE under the rules of Member Conduct.

Submission Format.

Authors are encouraged to prepare manuscripts employing the on-line style files developed by IEEE. All manuscripts accepted for publication will require the authors to make final submission employing these style files. The style files are available on the web at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2 under "Template for all TRANSACTIONS." (LaTeX and MS Word).

Authors using LaTeX: the two PDF versions of the manuscript needed for submission can both be produced by the IEEEtran.cls style file. A double-spaced document is generated by including \documentclass[11pt,draftcls,onecolumn]{IEEEtran} as the first line of the manuscript source file, and a single-spaced double-column document for estimating the publication page charges via \documentclass[10pt,twocolumn,twoside]{IEEEtran} for a regular submission, or \documentclass[9pt,twocolumn,twoside]{IEEEtran} for a Correspondence item.

- *Title page and abstract:* The first page of the manuscript shall contain the title, names and contact information for all authors (full mailing address, institutional affiliations, phone, fax, and e-mail), the abstract, and the EDICS. An asterisk * should be placed next to the name of the Corresponding Author who will serve as the main point of contact for the manuscript during the review and publication processes.

  An abstract must be a well-written stand-alone paragraph 150-250 words long, with no displayed equations, footnotes, references or tabular material. The abstract should indicate the scope of the paper and summarize the author's conclusions, making it a useful tool for information retrieval. Visit http://www.signalprocessingsociety.org/publications/periodicals/tsp/tsp-author-info/ for specifications and description.
- *EDICS:* All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at http://www.signalprocessingsociety.org/publications/periodicals/tsp/TSP-EDICS/
- NOTE: EDICS are necessary to begin the peer review process. Upon submission of a new manuscript, please choose the EDICS categories that best suit your manuscript. Failure to do so will likely result in a delay of the peer review process.
- The EDICS category should appear on the first page—i.e., the title and abstract page—of the manuscript.
- *Illustrations and tables:* Each figure and table should have a caption that is intelligible without requiring reference to the text. Illustrations/tables may be worked into the text of a newly-submitted manuscript, or placed at the end of the manuscript. (However, for the final submission, illustrations/tables must be submitted separately and not interwoven with the text.)

  Illustrations in color may be used but, unless the final publishing will be in color, the author is responsible that the corresponding grayscale figure is understandable.

  In preparing your illustrations, note that in the printing process, most illustrations are reduced to single-column width to conserve space. This may result in as much as a 4:1 reduction from the original. Therefore, make sure that all words are in a type size that will reduce to a minimum of 9 points or 3/16 inch high in the printed version. Only the major grid lines on graphs should be indicated.
- *Abbreviations:* This TRANSACTIONS follows the practices of the IEEE on units and abbreviations, as outlined in the Institute's published standards. See http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/authors/transjnl/auinfo07.pdf for details.
- *Mathematics:* All mathematical expressions must be legible. Do not give derivations that are easily found in the literature; merely cite the reference.

Final Manuscript Submission Guidelines.

Upon formal acceptance of a manuscript for publication, instructions for providing the final materials required for publication will be sent to the Corresponding Author. Finalized manuscripts should be prepared in LaTeX or MS Word, and are required to use the style files established by IEEE, available at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2.

Instructions for preparing files for electronic submission are as follows:

- For regular papers, the final manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. Without expressed approval from the Editor-in-Chief, papers that exceed 16 pages in length will not publish.
- Files must be self-contained; that is, there can be no pointers to your system setup.
- Include a header to identify the name of the TRANSACTIONS, the name of the author, and the software used to format the manuscript.
- Do not import graphics files into the text file of your finalized manuscript (although this is acceptable for your initial submission). If submitting on disk, use a separate disk for graphics files.
- Do not create special macros.
- Do not send PostScript files of the text.
- File names should be lower case.
- Graphics files should be separate from the text, and not contain the caption text, but include callouts like "(a)," "(b)."
- Graphics file names should be lower case and named fig1.eps, fig2.tif, etc.
- Supported graphics types are EPS, PS, TIFF, or graphics created using Word, Powerpoint, Excel or PDF. Not acceptable is GIF, JPEG, WMF, PNG, BMP or any other format (JPEG is accepted for author photographs only). The provided resolution needs to be at least 600 dpi (400 dpi for color).
- Please indicate explicitly if certain illustrations should be printed in color; note that this will be at the expense of the author. Without other indications, color graphics will appear in color in the online version, but will be converted to grayscale in the print version.

IEEE supports the publication of author names in the native language alongside the English versions of the names in the author list of an article. For more information, please visit the IEEE Author Digital Tool Box at the following URL: http://www.ieee.org/publications_standards/publications/authors/auth_names_native_lang.pdf

Additional instructions for preparing, verifying the quality, and submitting graphics and multimedia files are available via http://www.ieee.org/publications_standards/publications/authors/authors_journals.html.

Open Access.

This publication is a hybrid journal, allowing either Traditional manuscript submission or Open Access (author-pays OA) manuscript submission. Upon submission, if you choose to have your manuscript be an Open Access article, you commit to pay the discounted $1,750 OA fee if your manuscript is accepted for publication in order to enable unrestricted public access. Any other application charges (such as over-length page charge and/or charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. If you would like your manuscript to be a Traditional submission, your article will be available to qualified subscribers and purchasers via IEEE Xplore. No OA payment is required for Traditional submission.

Page Charges.

Voluntary Page Charges. Upon acceptance of a manuscript for publication, the author(s) or his/her/their company or institution will be asked to pay a charge of $110 per page to cover part of the cost of publication of the first ten pages that comprise the standard length (six pages, in the case of Technical Correspondences until their publication will be discontinued).

Mandatory Page Charges. The author(s) or his/her/their company or institution will be billed $220 per each page in excess of the first ten published pages for regular papers and six published pages for technical correspondence until their publication will be discontinued. These are mandatory page charges and the author(s) will be held responsible for them. They are not negotiable or voluntary. The author(s) signifies his willingness to pay these charges simply by submitting his/her/their manuscript to the TRANSACTIONS. The Publisher holds the right to withhold publication under any circumstance, as well as publication of the current or future submissions of authors who have outstanding mandatory page charge debt.

Color Charges. Color figures which appear in color only in the electronic (Xplore) version can be used free of charge. In this case, the figure will be printed in the hardcopy version in grayscale, and the author is responsible that the corresponding grayscale figure is intelligible. Color reproduction in print is expensive, and all charges for color are the responsibility of the author. The estimated costs are as follows. There will be a charge of $62.50 for each figure; this charge may be subject to change without notification. In addition, there are printing preparation charges which may be estimated as follows: color reproductions on four or fewer pages of the manuscript: a total of approximately $1045; color reproductions on five pages through eight pages: a total of approximately $2090; color reproductions on nine through 12 pages: a total of approximately $3135, and so on. Payment of fees on color reproduction is not negotiable or voluntary, and the author's agreement to publish the manuscript in the TRANSACTIONS is considered acceptance of this requirement.

IEEE TRANSACTIONS ON

SIGNAL AND INFORMATION PROCESSING OVER NETWORKS

Now accepting paper submissions

The new *IEEE Transactions on Signal and Information Processing over Networks* publishes high-quality papers that extend the classical notions of processing of signals defined over vector spaces (e.g. time and space) to processing of signals and information (data) defined over networks, potentially dynamically varying. In signal processing over networks, the topology of the network may define structural relationships in the data, or may constrain processing of the data. Topics of interest include, but are not limited to the following:

Adaptation, Detection, Estimation, and Learning
- Distributed detection and estimation
- Distributed adaptation over networks
- Distributed learning over networks
- Distributed target tracking
- Bayesian learning; Bayesian signal processing
- Sequential learning over networks
- Decision making over networks
- Distributed dictionary learning
- Distributed game theoretic strategies
- Distributed information processing
- Graphical and kernel methods
- Consensus over network systems
- Optimization over network systems

Communications, Networking, and Sensing
- Distributed monitoring and sensing
- Signal processing for distributed communications and networking
- Signal processing for cooperative networking
- Signal processing for network security
- Optimal network signal processing and resource allocation

Modeling and Analysis
- Performance and bounds of methods
- Robustness and vulnerability
- Network modeling and identification

Modeling and Analysis (cont.)
- Simulations of networked information processing systems
- Social learning
- Bio-inspired network signal processing
- Epidemics and diffusion in populations

Imaging and Media Applications
- Image and video processing over networks
- Media cloud computing and communication
- Multimedia streaming and transport
- Social media computing and networking
- Signal processing for cyber-physical systems
- Wireless/mobile multimedia

Data Analysis
- Processing, analysis, and visualization of big data
- Signal and information processing for crowd computing
- Signal and information processing for the Internet of Things
- Emergence of behavior

Emerging topics and applications
- Emerging topics
- Applications in life sciences, ecology, energy, social networks, economic networks, finance, social sciences, smart grids, wireless health, robotics, transportation, and other areas of science and engineering

Editor-in-Chief: Petar M. Djurić, Stony Brook University (USA)

To submit a paper, go to: https://mc.manuscriptcentral.com/tsipn-ieee

IEEE COMMUNICATIONS SOCIETY® IEEE Signal Processing Society® IEEE COMPUTER SOCIETY

2015 IEEE SIGNAL PROCESSING SOCIETY MEMBERSHIP APPLICATION

Mail to: IEEE OPERATIONS CENTER, ATTN: Louis Curcio, Member and Geographic Activities, 445 Hoes Lane, Piscataway, New Jersey 08854 USA
or Fax to (732) 981-0225 (credit card payments only.)
For info call (732) 981-0060 or 1 (800) 678-IEEE or E-mail: new.membership@ieee.org

IEEE

1. PERSONAL INFORMATION

NAME AS IT SHOULD APPEAR ON IEEE MAILINGS: SEND MAIL TO: □ **Home Address** OR □ **Business/School Address**
If not indicated, mail will be sent to home address. Note: Enter your name as you wish it to appear on membership card and all correspondence.
PLEASE PRINT Do not exceed 40 characters or spaces per line. Abbreviate as needed. Please circle your last/surname as a key identifier for the IEEE database.

TITLE FIRST OR GIVEN NAME MIDDLE NAME SURNAME/LAST NAME

HOME ADDRESS

CITY STATE/PROVINCE POSTAL CODE COUNTRY

2.

Are you now or were you ever a member of IEEE? □ **Yes** □ **No**
If yes, please provide, if known:

MEMBERSHIP NUMBER_______________

Grade____________**Year Membership Expired:**______________

3. BUSINESS/PROFESSIONAL INFORMATION

Company Name

Department/Division

Title/Position Years in Current Position

Years in the Profession Since Graduation □ PE State/Province

Street Address

City State/Province Postal Code Country

4. EDUCATION

A baccalaureate degree from an IEEE recognized educational program assures assignment of "Member" grade. For others, additional information and references may be necessary for grade assignment.

A.
Baccalaureate Degree Received Program/Course of Study

College/University Campus

State/Province Country Mo./Yr. Degree Received

B.
Highest Technical Degree Received Program/Course of Study

College/University Campus

State/Province Country Mo./Yr. Degree Received

5.

Full signature of applicant

6. DEMOGRAPHIC INFORMATION – ALL APPLICANTS -

Date Of Birth ______________ □ Male □ Female
Day Month Year

7. CONTACT INFORMATION

Office Phone/Office Fax Home Phone/Home Fax

Office E-Mail Home E-Mail

8. 2015 IEEE MEMBER RATES

IEEE DUES Residence	16 Aug 14-28 Feb 15 Pay Full Year	1 Mar -15 Aug 15 Pay Half Year**
United States	$193.00 □	$96.50 □
Canada (incl. GST)	$171.25 □	$85.63 □
Canada (incl. HST for PEI)	$184.30 □	$92.15 □
Canada (incl. HST for Nova Scotia)	$185.75 □	$92.88 □
Canada (incl. HST for NB, NF and ON)	$182.85 □	$91.43 □
Canada (incl. GST and QST Quebec)	$185.71 □	$92.86 □
Africa, Europe, Middle East	$158.00 □	$79.00 □
Latin America	$149.00 □	$74.50 □
Asia, Pacific	$150.00 □	$75.00 □

Canadian Taxes (GST/HST): All supplies, which include dues, Society membership fees, online products and publications (except CD-ROM and DVD media), shipped to locations within Canada are subject to the GST of 5% or the HST of 13%, 14% or 15%, depending on the Province to which the materials are shipped. GST and HST do not apply to Regional Assessments. (IEEE Canadian Business Number 12563 4188 RT0001)
Value Added Tax (VAT) in the European Union: In accordance with the European Union Council Directives 2002/38/EC and 77/388/EEC amended by Council Regulation (EC)792/2002, IEEE is required to charge and collect VAT on electronic/digitized products sold to private consumers that reside in the European Union. The VAT rate applied is the EU member country standard rate where the consumer is resident. (IEEE's VAT registration number is EU826000081)
U.S. Sales Taxes: Please add applicable state and local sales and use tax on orders shipped to **Alabama, Arizona, California, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Massachusetts, Maryland, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, West Virginia, Wisconsin.** Customers claiming a tax exemption must include an appropriate and properly completed tax-exemption certificate with their first order.

2015 SPS MEMBER RATES

		16 Aug-28 Feb Pay Full Year	1 Mar-15 Aug Pay Half Year
Signal Processing Society Membership Fee*		$ 20.00 □	$ 10.00 □

Fee includes: **IEEE Signal Processing Magazine** (electronic and digital), **Inside Signal Proc. eNewsletter** (electronic) and **IEEE Signal Processing Society Content Gazette** (electronic).

Add $15 to enhance SPS Membership and also receive:		$15.00 □	$ 7.50 □

IEEE Signal Processing Magazine (print) and **SPS Digital Library**: online access to Signal Processing Magazine, Signal Processing Letters, Journal of Selected Topics in Signal Processing, Trans. on Audio, Speech, and Language Processing, Trans. on Image Processing, Trans. on Information Forensics and Security and Trans. on Signal Processing.

Publications available only with SPS membership:

Signal Processing, IEEE Transactions on:	Print	$190.00 □	$ 95.00 □
Audio, Speech, and Lang. Proc., IEEE/ACM Trans. on:	Print	$145.00 □	$ 72.50 □
Image Processing, IEEE Transactions on:	Print	$188.00 □	$ 94.00 □
Information Forensics and Security, IEEE Trans. on:	Print	$163.00 □	$ 81.50 □
IEEE Journal of Selected Topics in Signal Processing:	Print	$160.00 □	$ 80.00 □
Affective Computing, IEEE Transactions on:	Electronic	$ 35.00 □	$ 17.50 □
Biomedical and Health Informatics, IEEE Journal of:	Print	$ 55.00 □	$ 27.50 □
	Electronic	$ 40.00 □	$ 20.00 □
	Print & Electronic	$ 65.00 □	$ 32.50 □
IEEE Cloud Computing	Electronic and Digital	$ 39.00 □	$ 19.50 □
New! **IEEE Trans. on Cognitive Comm. &Networking**	Electronic	$ 26.00 □	$ 13.00 □
New! **IEEE Trans. on Computational Imaging**	Electronic	$ 28.00 □	$ 14.00 □
New! **IEEE Trans. on Big Data**	Electronic	$ 25.00 □	$ 12.50 □
New! **IEEE Trans. on Molecular, Biological, & Multi-scale Communications**	Electronic	$ 24.00 □	$ 12.00 □
IEEE Internet of Things Journal	Electronic	$ 26.00 □	$ 13.00 □
IEEE Trans. on Cloud Computing	Electronic	$ 42.00 □	$ 21.00 □
IEEE Trans. on Computational Social Systems	Electronic	$ 30.00 □	$ 15.00 □
New! **IEEE Trans. on Signal & Info Proc. Over Networks**	Electronic	$ 28.00 □	$ 14.00 □
IEEE Biometrics Compendium:	Online	$ 30.00 □	$ 15.00 □
Computing in Science & Engrg. Mag.:	Electronic and Digital	$ 39.00 □	$ 19.50 □
	Print	$ 149.00 □	$ 74.50 □
Medical Imaging, IEEE Transactions on:	Print	$ 74.00 □	$ 37.00 □
	Electronic	$ 53.00 □	$ 26.50 □
	Print & Electronic	$ 89.00 □	$ 44.50 □
Mobile Computing, IEEE Transactions on:	ELE/Print Abstract/CD-ROM	$ 40.00 □	$ 20.00 □
Multimedia, IEEE Transactions on:	Electronic	$ 42.00 □	$ 21.00 □
IEEE MultiMedia Magazine:	Electronic and Digital	$ 39.00 □	$ 19.50 □
	Print	$149.00 □	$ 74.50 □
Network Science and Engrg., IEEE Trans. on:	Electronic	$ 33.00 □	$ 16.50 □
IEEE Reviews in Biomedical Engineering:	Print	$ 25.00 □	$ 12.50 □
	Electronic	$ 25.00 □	$ 12.50 □
	Print & Electronic	$ 40.00 □	$ 20.00 □
IEEE Security and Privacy Magazine:	Electronic and Digital	$ 39.00 □	$ 19.50 □
	Print	$149.00 □	$ 74.50 □
IEEE Sensors Journal:	Print	$150.00 □	$ 75.00 □
	Electronic	$ 50.00 □	$ 25.00 □
Smart Grid, IEEE Transactions on:	Print	$100.00 □	$ 50.00 □
	Electronic	$ 40.00 □	$ 20.00 □
	Print & Electronic	$120.00 □	$ 60.00 □
Wireless Communications, IEEE Transactions on:	Print	$120.00 □	$ 60.00 □
	Electronic	$ 48.00 □	$ 24.00 □
	Print & Electronic	$120.00 □	$ 60.00 □
IEEE Wireless Communications Letters:	Print	$ 80.00 □	$ 40.00 □
	Electronic	$ 18.00 □	$ 9.00 □
	Print & Electronic	$ 95.00 □	$ 47.50 □
New! **IEEE Life Sciences Letters** (Open Access Pub)	Electronic	No Fee	

*IEEE membership required or requested
Affiliate application to join SP Society only. Amount Paid $______________

9.

IEEE Membership Affiliate Fee (See pricing in Section 8**)** $______________

Signal Processing Society Fees $______________

Canadian residents pay 5% GST or 13% HST
Reg. No. 125634188 on Society payment(s) & pubs only Tax $______________

AMOUNT PAID WITH APPLICATION TOTAL $______________
Prices subject to change without notice.
□ **Check or money order enclosed Payable to IEEE on a U.S. Bank**
□ **American Express** □ **VISA** □ **MasterCard**
□ **Diners Club**

Exp. Date/ Mo./Yr.

Cardholder Zip Code Billing Statement Address/USA Only

Full signature of applicant using credit card Date

10. WERE YOU REFERRED?

□ Yes □ No If yes, please provide the follow information:
Member Recruiter Name:______________
IEEE Recruiter's Member Number (Required): ______________

IEEE Signal Processing Society

2015 IEEE SIGNAL PROCESSING SOCIETY STUDENT MEMBERSHIP APPLICATION

(Current and reinstating IEEE members joining SPS complete areas 1, 2, 8, 9.)
Mail to: **IEEE OPERATIONS CENTER, ATTN: Louis Curcio, Member and Geographic Activities, 445 Hoes Lane, Piscataway, New Jersey 08854 USA**
or Fax to (732) 981-0225 (credit card payments only.)
For info call (732) 981-0060 or 1 (800) 678-IEEE or E-mail: new.membership@ieee.org

IEEE

1. PERSONAL INFORMATION

NAME AS IT SHOULD APPEAR ON IEEE MAILINGS: SEND MAIL TO: □ **Home Address** OR □ **Business/School Address**
If not indicated, mail will be sent to home address. Note: Enter your name as you wish it to appear on membership card and all correspondence.
PLEASE PRINT Do not exceed 40 characters or spaces per line. Abbreviate as needed. Please circle your last/surname as a key identifier for the IEEE database.

TITLE FIRST OR GIVEN NAME MIDDLE NAME SURNAME/LAST NAME

HOME ADDRESS

CITY STATE/PROVINCE POSTAL CODE COUNTRY

2.

Are you now or were you ever a member of IEEE? □ **Yes** □ **No**
If yes, please provide, if known:

MEMBERSHIP NUMBER__|_|_|_|_|_|_|_|_|_

Grade____________**Year Membership Expired:**____________

3. BUSINESS/PROFESSIONAL INFORMATION

Company Name

Department/Division

Title/Position Years in Current Position

Years in the Profession Since Graduation □ PE State/Province

Street Address

City State/Province Postal Code Country

4. EDUCATION

A baccalaureate degree from an IEEE recognized educational program assures assignment of "Member" grade. For others, additional information and references may be necessary for grade assignment.

A.
Baccalaureate Degree Received Program/Course of Study

College/University Campus

State/Province Country Mo./Yr. Degree Received

B.
Highest Technical Degree Received Program/Course of Study

College/University Campus

State/Province Country Mo./Yr. Degree Received

5.

Full signature of applicant

6. DEMOGRAPHIC INFORMATION – ALL APPLICANTS -

Date Of Birth ________________ □ Male □ Female
Day Month Year

7. CONTACT INFORMATION

Office Phone/Office Fax Home Phone/Home Fax

Office E-Mail Home E-Mail

8. 2015 IEEE STUDENT MEMBER RATES

IEEE DUES Residence	16 Aug 14-28 Feb 15 Pay Full Year	1 Mar -15 Aug 15 Pay Half Year**
United States	$32.00 □	$16.00 □
Canada (incl. GST)	$33.60 □	$16.80 □
Canada (incl. HST for NB, NF, and ON)	$36.16 □	$18.08 □
Canada (incl. HST for Nova Scotia)	$36.80 □	$18.40 □
Canada (incl. HST for PEI)	$36.48 □	$18.24 □
Canada (incl. GST and QST Quebec)	$36.79 □	$18.40 □
Africa, Europe, Middle East, Latin America, Asia, Pacific	$27.00 □	$13.50 □

Canadian Taxes (GST/HST): All supplies, which include dues, Society membership fees, online products and publications (except CD-ROM and DVD media), shipped to locations within Canada are subject to the GST of 5% or the HST of 13%,14% or 15%, depending on the Province to which the materials are shipped. GST and HST do not apply to Regional Assessments. (IEEE Canadian Business Number 12563 4188 RT0001)
Value Added Tax (VAT) in the European Union: In accordance with the European Union Council Directives 2002/38/EC and 77/388/EEC amended by Council Regulation (EC)792/2002, IEEE is required to charge and collect VAT on electronic/digitized products sold to private consumers that reside in the European Union. The VAT rate applied is the EU member country standard rate where the consumer is resident. (IEEE's VAT registration number is EU826000081)
U.S. Sales Taxes: Please add applicable state and local sales and use tax on orders shipped to **Alabama, Arizona, California, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Massachusetts, Maryland, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, West Virginia, Wisconsin.** Customers claiming a tax exemption must include an appropriate and properly completed tax-exemption certificate with their first order.

IEEE Signal Processing Society

2015 SPS STUDENT MEMBER RATES

		16 Aug-28 Feb Pay Full Year	1 Mar-15 Aug Pay Half Year
Signal Processing Society Membership Fee*		$10.00 □	$ 5.00 □

Fee includes: **IEEE Signal Processing Magazine** (electronic and digital), **Inside Signal Processing eNewsletter** (electronic) and **IEEE Signal Processing Society Content Gazette** (electronic).

Add $8 to enhance SPS Membership and also receive:		$ 8.00 □	$ 4.00 □

IEEE Signal Processing Society Magazine (print) and **SPS Digital Library**: online access to Signal Processing Magazine, Signal Processing Letters, Journal of Selected Topics in Signal Processing, Trans. on Audio, Speech, and Language Processing, Trans. on Image Processing, Trans. on Information Forensics and Security and Trans. on Signal Processing.

Publications available only with SPS membership:

Signal Processing, IEEE Transactions on:	Print	$ 95.00 □	$ 47.50 □
Audio, Speech, and Lang. Proc., IEEE/ACM Trans. on:	Print	$ 73.00 □	$ 36.50 □
Image Processing, IEEE Transactions on:	Print	$ 94.00 □	$ 47.00 □
Information Forensics and Security, IEEE Trans. on:	Print	$ 82.00 □	$ 41.00 □
IEEE Journal of Selected Topics in Signal Processing:	Print	$ 80.00 □	$ 40.00 □
Affective Computing, IEEE Transactions on:	Electronic	$ 18.00 □	$ 9.00 □
Biomedical and Health Informatics, IEEE Journal of:	Print	$ 28.00 □	$ 14.00 □
	Electronic	$ 20.00 □	$ 10.00 □
	Print & Electronic	$ 65.00 □	$ 32.50 □
IEEE Cloud Computing	Electronic and Digital	$ 20.00 □	$ 10.00 □
New! **IEEE Trans. on Cognitive Comm. &Networking**	Electronic	$ 13.00 □	$ 6.50 □
New! **IEEE Trans. on Computational Imaging**	Electronic	$ 14.00 □	$ 7.00 □
New! **IEEE Trans. on Big Data**	Electronic	$ 13.00 □	$ 6.50 □
New! **IEEE Trans. on Molecular, Biological, & Multi-Scale Communications**	Electronic	$ 12.00 □	$ 6.00 □
IEEE Internet of Things Journal	Electronic	$ 13.00 □	$ 6.50 □
IEEE Trans. on Cloud Computing	Electronic	$ 21.00 □	$ 10.50 □
IEEE Trans. on Computational Social Systems	Electronic	$ 15.00 □	$ 7.50 □
New! **IEEE Trans. on Signal & Info Proc. Over Networks**	Electronic	$ 14.00 □	$ 7.00 □
IEEE Biometrics Compendium:	Online	$ 15.00 □	$ 7.50 □
Computing in Science & Engrg. Mag.:	Electronic and Digital	$ 20.00 □	$ 10.00 □
	Print	$ 75.00 □	$ 37.50 □
Medical Imaging, IEEE Transactions on:	Print	$ 37.00 □	$ 18.50 □
	Electronic	$ 27.00 □	$ 13.50 □
	Print & Electronic	$ 45.00 □	$ 22.50 □
Mobile Computing, IEEE Transactions on:			
	ELE/Print Abstract/CD-ROM	$ 20.00 □	$ 10.00 □
Multimedia, IEEE Transactions on:	Electronic	$ 21.00 □	$ 10.50 □
IEEE MultiMedia Magazine:	Electronic and Digital	$ 20.00 □	$ 10.00 □
	Print	$ 75.00 □	$ 37.50 □
Network Science and Engrg., IEEE Trans. on:	Electronic	$ 17.00 □	$ 8.50 □
IEEE Reviews in Biomedical Engineering:	Print	$ 13.00 □	$ 6.50 □
	Electronic	$ 13.00 □	$ 6.50 □
	Print & Electronic	$ 20.00 □	$ 10.00 □
IEEE Security and Privacy Magazine:	Electronic and Digital	$ 20.00 □	$ 10.00 □
	Print	$ 75.00 □	$ 37.50 □
IEEE Sensors Journal:	Print	$150.00 □	$ 75.00 □
	Electronic	$ 28.00 □	$ 14.00 □
Smart Grid, IEEE Transactions on:	Print	$ 50.00 □	$ 25.00 □
	Electronic	$ 20.00 □	$ 10.00 □
	Print & Electronic	$ 60.00 □	$ 30.00 □
Wireless Communications, IEEE Transactions on:	Print	$ 60.00 □	$ 30.00 □
	Electronic	$ 24.00 □	$ 12.00 □
	Print & Electronic	$ 60.00 □	$ 30.00 □
IEEE Wireless Communications Letters:	Print	$ 40.00 □	$ 20.00 □
	Electronic	$ 9.00 □	$ 4.50 □
	Print & Electronic	$ 48.00 □	$ 24.00 □
New! **IEEE Life Sciences Letters** (Open Access Pub)	Electronic	No Fee	

*IEEE membership required or requested
Affiliate application to join SP Society only. Amount Paid $__________

9.

IEEE Membership Fee (See pricing in Section 8) $__________
Signal Processing Society Fees $__________
Canadian residents pay 5% GST or 13% HST
Reg. No. 125634188 on Society payment(s) & pubs only Tax $__________

AMOUNT PAID WITH APPLICATION TOTAL $__________
Prices subject to change without notice.
□ **Check or money order enclosed Payable to IEEE on a U.S. Bank**
□ **American Express** □ **VISA** □ **MasterCard** □ **Diners Club**

Exp. Date/ Mo./Yr.

Cardholder Zip Code Billing Statement Address/USA Only

Full signature of applicant using credit card Date

10. WERE YOU REFERRED?

□ Yes □ No If yes, please provide the follow information:
Member Recruiter Name:____________________
IEEE Recruiter's Member Number (Required): ________________

IEEE Signal Processing Society®

IEEE®