

# IEEE Signal Processing MAGAZINE

Volume 40 | Number 4 | June 2023

## Signal Processing Society 75 YEARS OF SERVICE

Celebrating Past Breakthroughs  
and Navigating the Future With Care

SPS: The Social Aspects  
of the Organization

Science Can Change a Teen's  
Life: A Testimony

Embracing the Ethical Challenges  
of Future Breakthroughs

SENSOR ARRAY SP

BIOIMAGING

CONFERENCES AND WORKSHOPS

WOMEN IN SP

SIGNAL PROCESSING SOCIETY

GRAPH SIGNAL PROCESSING

BEAMFORMING

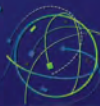
MULTI-ANTENNA COMMUNICATIONS

MULTICHANNEL SP

MULTIMEDIA SP

PART 1

IEEE  
Signal  
Processing  
Society



CELEBRATING 75 YEARS

IEEE



**IEEE**  
**Signal**  
**Processing**  
Society™

CELEBRATING 75 YEARS

## Honoring IEEE's first society

On 2 June 1948, the Professional Group on Audio of the IRE was formed, establishing what would become the IEEE society structure we know today.

75 years later, this group — now the IEEE Signal Processing Society — is the technical home to nearly 20,000 passionate, dedicated professionals and a bastion of innovation, collaboration, and leadership.

*Thank you!*

**Celebrate with us:**

<https://linktr.ee/ieeesps>

# Contents

Volume 40 | Number 4 | June 2023

## SPECIAL SECTION

### 75TH ANNIVERSARY OF SIGNAL PROCESSING SOCIETY SPECIAL ISSUE

- 3 FROM THE GUEST EDITORS**  
Rodrigo Capobianco Guido,  
Tulay Adali, Emil Björnson,  
Laure Blanc-Féraud,  
Ulisses Braga-Neto, Behnaz  
Ghoraani, Christian Jutten,  
Alle-Jan Van Der Veen,  
Hong Vicky Zhao, and Xiaoxing Zhu
- 14 EMPOWERING THE GROWTH OF  
SIGNAL PROCESSING**  
Athina Petropulu, José M.F. Moura,  
Rabab Kreidieh Ward, and  
Theresa Argiropoulos
- 23 THE EVOLUTION OF WOMEN  
IN SIGNAL PROCESSING AND  
SCIENCE, TECHNOLOGY,  
ENGINEERING, AND MATHEMATICS**  
Rabab Kreidieh Ward
- 36 IEEE SIGNAL PROCESSING  
SOCIETY FLAGSHIP CONFERENCES  
OVER THE PAST 10 YEARS**  
Ana I. Perez-Neira,  
Fernando Pereira, Carlo Regazzoni,  
and Caroline Johnson
- 46 HOW THE 1969 IEEE CONVENTION  
AND EXHIBITION CHANGED MY  
LIFE FOREVER**  
John Edwards
- 49 GRAPH SIGNAL PROCESSING**  
Geert Leus, Antonio G. Marques,  
José M.F. Moura, Antonio Ortega,  
and David I Shuman



### ON THE COVER

This issue celebrates the SP society's 75th anniversary. We look back on past breakthroughs and ponder the ethical challenges associated with future advances

COVER IMAGE: ©SHUTTERSTOCK.COM/G.DOLPHINNY

- 61 FROM NANO TO MACRO**  
Selin Aiyente, Alejandro F. Frangi,  
Erik Meijering, Arrate Muñoz-  
Barrutia, Michael Liebling, Dimitri  
Van De Ville, Jean-Christophe Olivo-  
Marin, Jelena Kovačević,  
and Michael Unser



- 72 MULTIMEDIA SIGNAL PROCESSING**  
Ivan V. Bajic<sup>†</sup>, Marta Mrak,  
Frédéric Dufaux, Enrico Magli,  
and Tsuhan Chen
- 80 TWENTY-FIVE YEARS OF SENSOR  
ARRAY AND MULTICHANNEL  
SIGNAL PROCESSING**  
Wei Liu, Martin Haardt, Maria S. Greco,  
Christoph F. Mecklenbräuker,  
and Peter Willett
- 92 THREE MORE DECADES IN ARRAY  
SIGNAL PROCESSING RESEARCH**  
Marius Pesavento, Minh Trinh-  
Hoang, and Mats Viberg
- 107 TWENTY-FIVE YEARS OF SIGNAL  
PROCESSING ADVANCES FOR  
MULTIANTENNA COMMUNICATIONS**  
Emil Björnson, Yonina C. Eldar,  
Erik G. Larsson, Angel Lozano,  
and H. Vincent Poor
- 118 TWENTY-FIVE YEARS OF  
ADVANCES IN BEAMFORMING**  
Ahmet M. Elbir, Kumar Vijay Mishra,  
Sergiy A. Vorobyov,  
and Robert W. Heath Jr.



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\$248 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 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 © 2023 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.**

Digital Object Identifier 10.1109/MSP.2023.3262435

## 8 From the Editor

Celebrating Technological Breakthroughs and Navigating the Future With Care  
*Christian Jutten and Athina Petropulu*

## 132 Dates Ahead



132

©SHUTTERSTOCK.COM/PHILIPPOS PHILIPPOU

The IEEE Signal Processing Society will celebrate its 75th Anniversary during the International Conference on Acoustic, Speech and Signal Processing, to be held in Rhodes Island, Greece, 4–10 June 2023.

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

## EDITOR-IN-CHIEF

Christian Jutten—Université Grenoble Alpes, France

## AREA EDITORS

## Feature Articles

Laure Blanc-Féraud—Université Côte d'Azur, France

## Special Issues

Xiaoxiang Zhu—German Aerospace Center, Germany

## Columns and Forum

Rodrigo Capobianco Guido—São Paulo State University (UNESP), Brazil

H. Vicky Zhao—Tsinghua University, R.P. China

## e-Newsletter

Hamid Palangi—Microsoft Research Lab (AI), USA

## Social Media and Outreach

Emil Björnson—KTH Royal Institute of Technology, Sweden

## EDITORIAL BOARD

Massoud Babaie-Zadeh—Sharif University of Technology, Iran

Waheed U. Bajwa—Rutgers University, USA

Caroline Chau—French Center of National Research, France

Mark Coates—McGill University, Canada

Laura Cottatellucci—Friedrich-Alexander

University of Erlangen-Nuremberg, Germany

Davide Dardari—University of Bologna, Italy

Mario Figueiredo—Instituto Superior Técnico, University of Lisbon, Portugal

Sharon Gannot—Bar-Ilan University, Israel

Yifan Gong—Microsoft Corporation, USA

Rémi Gribonval—Inria Lyon, France

Joseph Guerci—Information Systems

Laboratories, Inc., USA

Ian Jermyn—Durham University, U.K.

Ulugbek S. Kamilov—Washington University, USA

Patrick Le Callet—University of Nantes, France

Sanghoon Lee—Yonsei University, Korea

Danilo Mandic—Imperial College London, U.K.

Michalis Matthaiou—Queen's University Belfast, U.K.

Phillip A. Regalia—U.S. National Science Foundation, USA

Gaël Richard—Télécom Paris, Institut Polytechnique de Paris, France

Reza Sameni—Emory University, USA

Ervin Sejdic—University of Pittsburgh, USA

Dimitri Van De Ville—Ecole Polytechnique

Fédérale de Lausanne, Switzerland

Henk Wymeersch—Chalmers University of Technology, Sweden

## ASSOCIATE EDITORS—COLUMNS AND FORUM

Ulisses Braga-Neto—Texas A&M University, USA

Cagatay Candan—Middle East Technical University, Turkey

Wei Hu—Peking University, China

Andres Kwasinski—Rochester Institute of Technology, USA

Xingyu Li—University of Alberta, Edmonton, Alberta, Canada

Xin Liao—Hunan University, China

Piya Pal—University of California San Diego, USA

Hemant Patil—Dhirubhai Ambani Institute of Information and Communication Technology, India

Christian Ritz—University of Wollongong, Australia

## ASSOCIATE EDITORS—e-NEWSLETTER

Abhishek Appaji—College of Engineering, India

Subhro Das—MIT/IBM Watson AI Lab, IBM Research, USA

Behnaz Ghorani—Florida Atlantic University, USA

Panagiotis Markopoulos—The University of Texas at San Antonio, USA

## IEEE SIGNAL PROCESSING SOCIETY

Athina Petropulu—President

Min Wu—President-Elect

Ana Isabel Pérez-Neira—Vice President, Conferences

Roxana Saint-Nom—VP Education

Kenneth K.M. Lam—Vice President, Membership

Marc Moonen—Vice President, Publications

Alle-Jan van der Veen—Vice President, Technical Directions

## IEEE SIGNAL PROCESSING SOCIETY STAFF

Richard J. Baseil—Society Executive Director

William Colacchio—Senior Manager, Publications and Education Strategy and Services

Rebecca Wollman—Publications Administrator

## IEEE PERIODICALS MAGAZINES DEPARTMENT

Sharon Turk, *Journals Production Manager*

Katie Sullivan, *Senior Manager, Journals Production*

Janet Dudar, *Senior Art Director*

Gail A. Schnitzer, *Associate Art Director*

Theresa L. Smith, *Production Coordinator*

Mark David, *Director, Business Development - Media & Advertising*

Felicia Spagnoli, *Advertising Production Manager*

Peter M. Tuohy, *Production Director*

Kevin Lisankie, *Editorial Services Director*

Dawn M. Melley, *Senior Director, Publishing Operations*

Digital Object Identifier 10.1109/MSP.2023.3262437

**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.

Rodrigo Capobianco Guido<sup>1</sup>, Tulay Adali<sup>2</sup>, Emil Björnson<sup>3</sup>,  
 Laure Blanc-Féraud<sup>4</sup>, Ulisses Braga-Neto<sup>5</sup>, Behnaz Ghoraani<sup>6</sup>, Christian Jutten<sup>7</sup>,  
 Alle-Jan Van Der Veen<sup>8</sup>, Hong Vicky Zhao<sup>9</sup>, and Xiaoxing Zhu<sup>10</sup>

## IEEE Signal Processing Society: Celebrating 75 Years of Remarkable Achievements

It is our great pleasure to introduce the first part of this special issue to you! The IEEE Signal Processing Society (SPS) has completed 75 years of remarkable service to the signal processing community. When the Society was founded in 1948, we couldn't imagine, for instance, how wireless networks of smartphones would be able to connect us easily at all times, or that an image processing algorithm would be able to detect cancer in a few seconds. Those are just simple examples of the immense technological progress over the past 75 years, which became possible thanks in great part to the dedicated work of professional members of the SPS.

### Celebrating 75 years

A special issue of *IEEE Signal Processing Magazine* was published 25 years ago to celebrate the 50th anniversary of the SPS. To celebrate the 75th anniversary, we have focused on what has happened during the previous 25 years in the field of signal processing, in addition to the main perspectives considering both societal and technical aspects in different domains covered by our Society. In response to an open call for papers, we received 41 white paper submissions. Among those, 18 were selected and invited to be considered for publication upon submission of a full version. Finally, 11 were accepted for

inclusion in this first part of the special issue, while the remaining ones will appear in the upcoming second part.

The first three articles in this first part of the special issue focus on the history of the SPS. The article by Petropulu (SPS president) et al. [A1] describes the extraordinary growth we have witnessed in the field of digital signal processing (DSP) since 1998, where the SPS played a fundamental role in promoting cross-disciplinary collaboration and knowledge sharing. Then, the article by Ward (former SPS President) [A2] focuses on women researchers and volunteers and their active role within the SPS.

Finally, Pérez-Neira (SPS vice president, conferences) et al. [A3] present an article that comments on the most prominent SPS conferences and their evolution. These articles also discuss the main challenges and opportunities for the SPS.

Next, we have a powerful testimony by Edwards [A4], who has contributed significantly to our magazine and Society over the years. He begins by recalling a very special occasion: the day he was 14 years old and visited the 1969 IEEE International Convention & Exhibition and decided on his future career. Then, using his unique journalistic skills, he narrates lots of interesting events with significant value to our DSP community.

As signal processing can be classified along techniques and methods such as sampling, transforms, statistical techniques including machine learning, and so on, it can also be partitioned into major application areas, such as speech and audio, image processing and multimedia, communication and sensor array processing. Our technical committees (TCs) and unified Editors Information Classification Scheme (EDICS) reflect these dual partitionings. The selected feature articles included in this special issue provide a cross section of those fields. Particularly, in this first issue, we present seven of these

feature articles. The first one, authored by Leus et al. [A5] describes the role of graph signal processing for signal analysis over the recent decades in a variety of applications, including image and video processing; social, transportation, communication, and brain networks; recommender systems; financial engineering; distributed control; and learning. The second feature article is by Aviyente et al. [A6]. In it, the authors offer a brief history of the IEEE Bioimaging and Signal Processing TC, providing an overview of the main technological and methodological contributions and highlight promising new directions. Then, Bajić et al. [A7] review both the history of multimedia signal processing as well



as the IEEE Multimedia Signal Processing TC, with a focus on the last three decades.

The fourth feature article we present in this special issue is authored by Liu et al. [A8], where an overview of the IEEE Sensor Array and Multichannel TC and its activities are introduced, followed by the main technological advances and new developments in the area along with promising future research directions. The fifth feature article, authored by Pesavento et al. [A9], presents an overview and advances in multiple-input, multiple-output systems, including details on direction of arrival, direction of departure, time delay of arrival, and Doppler mechanisms. The sixth feature article is authored by Björnson et al. [A10] and presents the story of wireless communication technologies over the past 25 years, including the advances in air interface, channel coding, source compression, connection protocols, and related areas, covering from 2G to 5G technologies. Finally, the seventh feature article, authored by Elbir et al. [A11], describes relevant details on the development of beamformers, emphasizing minimum-variance distortionless response strategies and the corresponding major breakthroughs over the past decades.

This concludes the first part of this special issue. In the second part, to be published in the magazine's July issue, another set of relevant articles will appear, concluding our efforts to group together the most significant contributions received to celebrate the 75th anniversary of the SPS. We would like to specially express our gratitude to all our contributing authors and reviewers, in addition to Rebecca Wollman, who efficiently helped us with all the administrative details, and the entire team, led by Sharon Turk, who brilliantly promoted and supervised the editorial process.

We sincerely hope that you enjoy reading the first part of this special issue.

## Acknowledgment

Rodrigo Capobianco Guido is the lead guest editor of this special issue.

## Guest Editors



**Rodrigo Capobianco Guido** (guido@ieee.org) received his Ph.D. degree in computational applied physics from the University of São Paulo (USP), Brazil, in 2003. Following two postdoctoral programs in signal processing at USP, he obtained the title of associate professor in signal processing, also from USP, in 2008. Currently, he is an associate professor at São Paulo State University, São José do Rio Preto, São Paulo, 15054-000, Brazil. He has been an area editor of *IEEE Signal Processing Magazine* and was recently included in Stanford University's rankings of the world's top 2% scientists. His research interests include signal and speech processing based on wavelets and machine learning. He is a Senior Member of IEEE.



**Tulay Adali** (adali@umbc.edu) received her Ph.D. degree in electrical engineering from North Carolina State University. She is a distinguished university professor at the University of Maryland, Baltimore County, Baltimore, MD 21250 USA. She is chair of IEEE Brain and past vice president of technical directions for the IEEE Signal Processing Society (SPS). She is a Fulbright Scholar and an SPS Distinguished Lecturer. She received a Humboldt Research Award, an IEEE SPS Best Paper Award, the University System of Maryland Regents' Award for Research, and a National Science Foundation CAREER Award. Her research interests include statistical signal processing and machine learning and their applications, with an emphasis on applications in medical image analysis and fusion. She is a Fellow of IEEE and a fellow of the American Institute for Medical and Biological Engineering.



**Emil Björnson** (emilbjo@kth.se) is a full (tenured) professor of wireless communication at the KTH Royal Institute of Technology,

Stockholm, 100 44, Sweden. He received the 2018 and 2022 IEEE Marconi Prize Paper Awards in Wireless Communications, the 2019 EURASIP Early Career Award, the 2019 IEEE Communications Society Fred W. Ellersick Prize, the 2019 IEEE Signal Processing Magazine Best Column Award, the 2020 Pierre-Simon Laplace Early Career Technical Achievement Award, the 2020 Communication Theory Technical Committee Early Achievement Award, the 2021 IEEE Communications Society Radio Communications Committee Early Achievement Award, and the 2023 IEEE Communications Society Outstanding Paper Award. His work has also received six Best Paper Awards at conferences. He is a Fellow of IEEE, and a Digital Futures and Wallenberg Academy fellow.

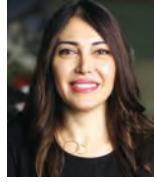


**Laure Blanc-Féraud** (laure.blanc-feraud@univ-cotedazur.fr) received her Ph.D. degree and habilitation to conduct research in inverse problems in image processing from University Côte d'Azur in 1989 and 2000, respectively. She is a researcher with Informatique Signaux et Systèmes at Sophia Antipolis (I3S) Lab, the University Côte d'Azur, Centre national de la recherche scientifique (CNRS), Sophia Antipolis, 06900 France. She served/serves on the IEEE Biomedical Image and Signal Processing Technical Committee (2007–2015; 2019–) and has been general technical chair (2014) and general chair (2021) of the IEEE International Symposium on Biomedical Imaging. She has been an associate editor of *SIAM Imaging Science* (2013–2018) and is currently an area editor of *IEEE Signal Processing Magazine*. She headed the French national research group GDR Groupement de recherche–Information, Signal, Image et ViSion (ISIS) of CNRS on Information, Signal Image and Vision (2021–2018). Her research interests include inverse problems in image processing using partial differential equation and optimization. She is a Fellow of IEEE.



**Ulisses Braga-Neto** (ulisses@tamu.edu) received his Ph.D. degree in electrical and computer engineering from Johns Hopkins University in 2002. He is a professor in the Electrical and Computer Engineering Department, Texas A&M University, College Station TX 77843 USA. He is founding director of the Scientific Machine Learning Lab at the Texas A&M Institute of Data Science. He is an associate editor of *IEEE Signal Processing Magazine* and a former elected member of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee and the IEEE Biomedical Imaging and Signal Processing Technical Committee. He has published two textbooks and more than 150 peer-reviewed journal articles and conference papers. He received the 2009 National Science Foundation CAREER Award. His research focuses on machine

learning and statistical signal processing. He is a Senior Member of IEEE.



**Behnaz Ghoraani** (bghoraani@fau.edu) received her Ph.D. from the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada, followed by a Postdoctoral Fellow period with the Faculty of Medicine, University of Toronto, Toronto, Canada. She is an associate professor of electrical engineering and computer science at Florida Atlantic University, Boca Raton FL 33431 USA, with a specialization in biomedical signal analysis, machine learning, wearable and assistive devices for rehabilitation, and remote home monitoring. She is an associate editor of *IEEE Journal of Biomedical and Health Informatics* and *BioMedical Engineering OnLine Journal*. Her research has received recognition through multiple

best paper awards and the Gordon K. Moe Young Investigator Award. Her research has been funded by grants from the National Institutes of Health, the National Science Foundation (including a CAREER Award), and the Florida Department of Health. She is an esteemed member of the Board of Scientific Counselors of National Library of Medicine, as well as the IEEE SPS Biomedical Signal and Image Professional Technical Committee. She has also taken on the role of the IEEE Women in Signal Processing Committee Chair and an Area Editor for the IEEE SPM eNewsletter.



**Christian Jutten** (christian.jutten@grenoble-inp.fr) received his Ph.D. and Doctor es Sciences degrees from Grenoble Polytechnic Institute, France, in 1981 and 1987, respectively. He was an associate



Google is looking forward to connecting with the community at ICASSP 2023! We welcome you to stop by the Google table during the conference to network with members of our research teams. You can also scan the QR code below to view our list of accepted papers at ICASSP this year.



professor (1982–1989) and a professor (1989–2019), and has been a professor emeritus since September 2019 at University Grenoble Alpes, Saint-Martin-d’Hères 38400. He was an organizer or program chair of many international conferences, including the first Independent Component Analysis Conference in 1999 (ICA’99) and the 2009 IEEE International Workshop on Machine Learning for Signal Processing. He was the technical program cochair of ICASSP 2020. Since 2021, he has been editor-in-chief of *IEEE Signal Processing Magazine*. Since the 1980s, his research interests have been in machine learning and source separation, including theory and applications (brain and hyperspectral imaging, chemical sensing, and speech). He is a Fellow of IEEE and a fellow of the European Association for Signal Processing.



**Alle-Jan Van Der Veen** (a.j.vanderveen@tudelft.nl) received his Ph.D. in system theory at the Circuits and Systems Group,

Department of Electrical Engineering, TU Delft, The Netherlands, with a postdoctoral research position at Stanford University, USA. He is a professor and chair of the Signal Processing Systems group at Delft University of Technology, Delft, 2628, The Netherlands. He was editor-in-chief of *IEEE Transactions on Signal Processing* and *IEEE Signal Processing Letters*. He was an elected member of the IEEE Signal Processing Society (SPS) Board of Governors. He was chair of the IEEE SPS Fellow Reference Committee, chair of the IEEE SPS Signal Processing for Communications Technical Committee, and technical cochair of ICASSP 2011 (Prague). He is currently the IEEE SPS vice president of technical directions (2022–2024). His research interests are in the areas of array signal processing and signal processing for communication, with applications to radio astronomy and sensor network local-

ization. He is a Fellow of IEEE and a fellow of the European Association for Signal Processing.



**Hong Vicky Zhao** (vzhao@tsinghua.edu.cn) received her Ph.D. degree in electrical engineering from the University of Maryland,

College Park, in 2004. Since May 2016, she has been an associate professor with the Department of Automation, Tsinghua University, Beijing, 100084, China. She received the IEEE Signal Processing Society 2008 Young Author Best Paper Award. She is the coauthor of “Multimedia Fingerprinting Forensics for Traitor Tracing” (Hindawi, 2005), “Behavior Dynamics in Media-Sharing Social Networks” (Cambridge University Press, 2011), and “Behavior and Evolutionary Dynamics in Crowd Networks” (Springer, 2020). She was a member of the IEEE Signal Processing Society Information Forensics and Security Technical Committee and the Multimedia Signal Processing Technical Committee. She is the senior area editor, area editor, and associate editor of *IEEE Signal Processing Letters*, *IEEE Signal Processing Magazine*, *IEEE Transactions on Information Forensics and Security*, and *IEEE Open Journal of Signal Processing*. Her research interests include media-sharing social networks, information security and forensics, digital communications, and signal processing.



**Xiaoxing Zhu** (xiaoxiang.zhu@tum.de) received her Dr.-Ing. degree and her “Habilitation” in signal processing from the

Technical University of Munich (TUM), in 2011 and 2013, respectively. She is the chair professor for data science in Earth observation at TUM, Munich, 80333, Germany. She was founding head of the “EO Data Science” Department at the Remote Sensing Technology Institute, German Aerospace Center. Since October 2020, she has served as a director of the TUM Munich Data Science Institute. She is

currently a visiting artificial intelligence professor at the European Space Agency’s Phi Lab. Her research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications to tackling societal grand challenges, e.g., global urbanization, the United Nations’ sustainable development goals, and climate change. She is a Fellow of IEEE.

## Appendix: Related Articles

- [A1] A. Petropulu, J. M. F. Moura, R. K. Ward, and T. Argiropoulos, “Empowering the growth of signal processing,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 14–22, Jul. 2023, doi: 10.1109/MSP.2023.3262905.
- [A2] R. K. Ward, “The evolution of women in signal processing and science, technology, engineering, and mathematics,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 23–35, Jul. 2023, doi: 10.1109/MSP.2023.3236475.
- [A3] A. I. Perez-Neira, F. Pereira, C. Regazzoni, and C. Johnson, “IEEE signal processing society flagship conferences over the past 10 years,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 36–45, Jul. 2023, doi: 10.1109/MSP.2023.3240852.
- [A4] J. Edwards, “How the 1969 IEEE convention and exhibition changed my life forever,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 46–48, Jul. 2023, doi: 10.1109/MSP.2023.3253254.
- [A5] G. Leus, A. G. Marques, J. M. F. Moura, A. Ortega, and D. I. Shuman, “Graph signal processing,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 49–60, Jul. 2023, doi: 10.1109/MSP.2023.3262906.
- [A6] S. Aviyente et al., “From nano to macro,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 61–71, Jul. 2023, doi: 10.1109/MSP.2023.3242833.
- [A7] I. V. Bajić, M. Mrak, F. Dufaux, E. Magli, and T. Chen, “Multimedia signal processing,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 72–79, Jul. 2023, doi: 10.1109/MSP.2023.3260989.
- [A8] W. Liu, M. Haardt, M. S. Greco, C. F. Mecklenbräuker, and P. Willett, “Twenty-five years of sensor array and multichannel signal processing,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 80–91, Jul. 2023, doi: 10.1109/MSP.2023.3258060.
- [A9] M. Pesavento, M. Trinh-Hoang, and M. Viberg, “Three more decades in array signal processing research,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 92–106, Jul. 2023, doi: 10.1109/MSP.2023.3255558.
- [A10] E. Björnson, Y. C. Eldar, E. G. Larsson, A. Lozano, and H. V. Poor, “Twenty-five years of signal processing advances for multiantenna communications,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 107–117, Jul. 2023, doi: 10.1109/MSP.2023.3261505.
- [A11] A. M. Elbir, K. V. Mishra, S. A. Vorobyov, and R. W. Heath Jr., “Twenty-five years of advances in beamforming,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 118–131, Jul. 2023, doi: 10.1109/MSP.2023.3262366.





## Get Published in the New *IEEE Open Journal of Signal Processing*

Submit a paper today to the premier new open access journal in signal processing.



In keeping with IEEE's continued commitment to providing options supporting the needs of all authors, in 2020, IEEE introduced the high-quality publication, the *IEEE Open Journal of Signal Processing*.

**In recognition of author funding difficulties during this unprecedented time, the IEEE Signal Processing Society is offering a reduced APC of USD\$995 with no page limits for regular papers. (This offer cannot be combined with any other discounts.)**

We invite you to have your article peer-reviewed and published in the new journal. This is an exciting opportunity for your research to benefit from the high visibility and interest the journal will generate.

Your research will also be exposed to 5+ million unique monthly users of the IEEE Xplore® Digital Library.

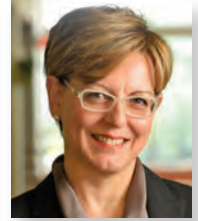
*The high-quality IEEE Open Journal of Signal Processing* will draw on IEEE's expert technical community's continued commitment to publishing the most highly cited content. The editor-in-chief is the distinguished Prof. Brendt Wohlberg, who specializes in signal and image processing, inverse problems, and computational imaging.

**The rapid peer-reviewed process targets a publication time frame within 10-15 weeks for most accepted papers. This journal is fully open and compliant with funder mandates, including Plan S.**

### Submit your paper today!

The high-quality IEEE Open Journal of Signal Processing launched in IEEE Xplore® in January 2020 and welcomes submissions of novel technical contributions.

[Click here to learn more](#)



## Celebrating Technological Breakthroughs and Navigating the Future With Care

The 75th anniversary of the IEEE Signal Processing Society (SPS) is an ideal time to look at the rapid advances in our field and the many ways that these increasingly powerful technologies have transformed our professions and the world. This is not just a time to celebrate past achievements and pat ourselves on the back, but also to educate young students and innovators about the history of our profession, the challenges we have overcome, and the breakthroughs that have led to the incredible growth of Signal Processing (SP). More importantly, this reflection will help shape our path forward, by inspiring new innovations, and also bringing awareness of the ethical issues associated with evolving and emerging technologies. This awareness will help us to develop meaningful safeguards and ensure responsible use of these technologies.

The 75th anniversary of the SPS coincides with another important 75th anniversary, that of a tiny yet mighty device: the transistor. This is not a mere coincidence. From their birth, signal and image processing (SIP) has been strongly associated with technological advances, especially in electronics and computers. In fact, SIP requires both

sensors, for recording signals and images and computers, for implementing smart and efficient processing.

### Fantastic growth during the last decades

For the sake of our younger SPS scientists, let's start with a few milestones in the common history of SIP and hardware. The first computer was very big. ENIAC, built in 1943, was about 170 m<sup>2</sup> and 27 tons, with a power consumption of 150 kW. It had a very limited computational capacity: approximately 0.2 ms for addition or subtraction, 2 ms for multiplication, and up to 65 ms for performing a division or a square root! ENIAC was a decimal machine, but, a few years later, in 1946, in the framework of the Electronic Discrete Variable

Automatic Computer (EDVAC) project, the concept of the von Neumann machine appeared, using binary coding and computing. The basic components of these machines were electronic tubes.

By the 1950s and up to end of the 1960s, a few computers were built with transistors as discrete components. In 1958, Kilby (a Nobel Prize winner in 2000) invented the first integrated circuit, which was patented in 1964 by Texas Instruments. This discovery had an incredible impact on the development of the digital world.

The first microprocessor appeared in 1971: the Intel 4004, a 4-bit microprocessor with 2,300 transistors and a clock frequency of about 100 kHz. This 16-pin integrated circuit (about 3.8 × 2.8 cm) had a computational power similar to that of ENIAC! Of course, advances in microelectronics provided increasingly powerful integrated circuits and microprocessors. Here are just a few milestones, to show this impressive growth:

- 1972: Intel 8080: 8 bits; 3,500 transistors; and clock of 200 kHz
- 1979: Intel 8088: 16 bits; 29,000 transistors; and clock of 5 MHz
- 1989: Intel 80486: 32 bits; 1,200,000 transistors; and clock of 16–100 MHz.

Today, microprocessors are 64 bits and multicore, with more than 2 million transistors and a clock of about 5 GHz!

The microprocessor Intel 8088 was the basic component of the first IBM personal computer built in 1981. With 16 kB of random-access memory (RAM), extensible to 256 kB, and a floppy disk of 160 kB, its price was quite high, and it was primarily used by companies and, later, by some laboratories.

Until the 1980s, images were recorded using a Vidicon camera, based on a cathodic ray tube, which provides an image by the scanning of an electron beam. At that time, it was impossible to store such images in computer memory because the time access of the memory was not compatible with the speed of the scanning (30 frames/s

**The 75th anniversary of the SPS coincides with another important 75th anniversary, that of a tiny yet mighty device: the transistor.**

and about 500 lines), and the capacity of the memory (even dynamic RAM) was too small—fewer than 256 kB [1]. In 1970, Boyle and Smith (Nobel Prize winners in 2009) published a paper on charge-coupled semiconductor devices (CCDs) [2], which could be used as image sensors. The first commercial image CCD sensors were proposed by Fairchild in 1974, with  $100 \times 100$  pixels. Then, in 1983, Sony developed the first mass-produced consumer video camera based on a CCD sensor (CCD-G5) with  $384 \times 491$  pixels. Now, the size of CCD or CMOS image sensors in a camera is about  $8,000 \times 6,000$  pixels or better!

Advances in technology have had a strong impact in many domains for the development of other electronic devices and sensors, especially in medicine, remote sensing, transportation, and telecommunications.

Christian's experiences as a researcher in his university lab in France provide some important perspectives on the impact of increasingly powerful technologies. "In 1980, about 20 researchers in three labs shared access to two 16-bit computers: HP 1000 and T1600 (from the French company Télémécanique). On average, we could use one machine for about 1 h per day, with a personal partition of 24 kB of memory—for both the program and the data! Programs were written in Fortran, and there was no graphical output: we had to manually draw curves from the numerical results." One of Christian's friends designed methods for doing

handwritten character recognition: despite small images of  $128 \times 32$  pixels, computations had to be done using integers since coding and computing in

floating point were impossible using 24-kB memory.

Later, in 1985, Christian's lab got its first PC, and it was possible to use other languages, like Pascal and Basic. "But the performance was still very limited,"

he notes. "A very simple program of source separation required about one hour to converge. In the lab, we did some simulations on computers, but, typically, Ph.D. students also built dedicated machines." To overcome the computer's slowness, Christian implemented

**Young people wonder how it was possible to get work done, locate journal articles, do comprehensive research, share ideas, and communicate with each other without e-mail and the Internet.**

**75 years... That's big.  
Almost as big as some of our datasets.**



**DATAOCEAN AI**  
YOUR GLOBAL DATA PARTNER

**ASR**

**TTS**

**CV**

**NLP**

**Lexicon**

You used to know us as Speechocean, but we've changed our name because we do so much more. We still have 1,000+ off-the-shelf datasets ready to license, and extensive data collection, annotation and algorithm training capabilities. Drop by Booth 18 and talk to us about how we can help with your global data needs, visit us at [www.dataoceanai.com](http://www.dataoceanai.com), or email us at [meetme@dataoceanai.com](mailto:meetme@dataoceanai.com).

the source separation algorithm with operational amplifiers, field effect transistors, and other discrete components (Figure 1). The convergence of this analog implementation required only a few milliseconds, and he added a low-pass RC circuit to slow down the convergence speed so that it became observable!

Christian's team worked on artificial neural networks (ANNs), and, by the end of the 1980s, a few Ph.D. students had designed new systolic and parallel architectures with the related software for overcoming the limitations of classic computers for simulating ANNs.

Currently, e-mail and Internet access are essential tools in our lives, both personal and professional. We've become so used to fast, reliable, 24-h connectivity that we see it as a crisis if our web server is down for more than a few minutes. Young people wonder how it was possible to get work done, locate journal articles, do comprehensive research, share ideas, and communicate with each other without e-mail and the Internet. Christian remembers that

**Christian remembers that one of the first e-mails that he sent in 1985 came back three weeks later, with an error message and the list of servers through which it had passed!**

one of the first e-mails that he sent in 1985 came back three weeks later, with an error message and the list of servers through which it had passed! Before reliable Internet and e-mail connectivity became available at the beginning of the 1990s, we had access to some printed journals in the lab or university library, and,

when we had to share documents with collaborators outside our workplace, we did so by fax.

In that era, writing articles and papers for journals and conferences was also much more tricky. "We used an electric typewriter," says Christian. "If we needed to change the font, such as when typing equations or Greek characters, we had to replace the typeball." When Christian's lab acquired its first LaserJet in 1989, it became so easy to print a text with different fonts in one step using PostScript. It was also possible to design figures on a computer and to add them to the text, and later it became easy to include photos and images, too.

Now, tablets, laptops, PCs, and even smartphones are so powerful and fast, with huge memories, tens of gigabytes,

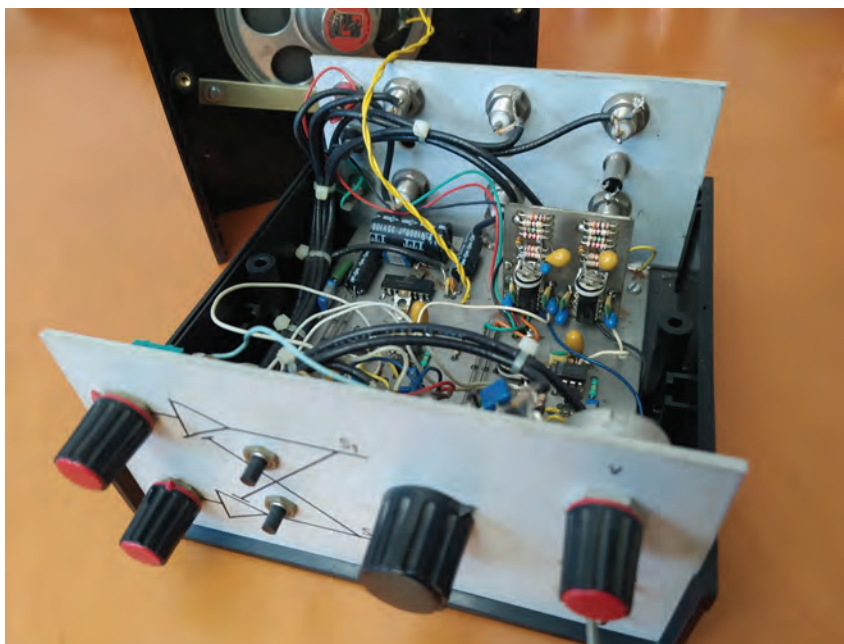
and hard disks of a few terabytes. For very complex simulations and computations, researchers can share university-based and national computing centers with incredibly powerful machines. All of these means of high-performance computations seem commonplace today, but it's important to be mindful that the growth of these tools has been extraordinarily fast over the last decades. The growth in SIP followed a similar trajectory, and it is easy to understand why image processing, computational imaging, wireless communications, and forensics, to name a few, didn't appear until the 1990s since they required devices, sensors, and computers that didn't exist or were not powerful enough.

### **Challenges for the future: Growth versus ethics and ecology**

In the 2020s, the developments in integrated circuits have led to GPUs whose highly parallel architectures are well-suited for efficiently performing a large number of operations. Multicore computers and GPUs provide researchers with the tools to train deep neural networks more quickly and efficiently than was previously possible. These tools enabled the development of large-scale deep learning frameworks, such as TensorFlow and PyTorch, making it easier for researchers and engineers to experiment with artificial intelligence (AI) models. These tools also supported the development of large-scale language models, such as the ChatGPT, developed by OpenAI, enabling language translation, chatbots, and content generation.

With all of the computational power available today, AI can analyze large amounts of data and identify patterns and insights that might be difficult for humans to detect. AI is now capable of performing a wide range of tasks, including image recognition, natural language processing, decision making, and even creative tasks, such as music composition and art generation.

While AI can accelerate the pace of scientific discovery, it also poses several concerns. AI systems may perpetuate and amplify biases that exist in society,



**FIGURE 1.** This analog electronic implementation of a source separation algorithm was about 1 million times faster than the simulation on a PC available in 1985.

such as racial or gender bias. This can happen when the AI system is trained on biased data, or if the algorithm itself is designed in a way that perpetuates bias. Another problem with AI methods is that they operate as “black boxes,” meaning that their inner workings are not transparent or easily understandable by humans. This can make it difficult to explain how the AI system arrived at a particular decision or prediction and can also make it challenging to identify and correct errors or biases in the system. When it comes to AI language models, such as the ChatGPT, there are serious concerns stemming from the kind of information it is accessed and potential violations of data privacy and intellectual rights. ChatGPT designs its answers by utilizing various resources available on the web and other servers, but the accuracy of these sources cannot always be guaranteed. Additionally, it is important to consider whether such AI tools respect academic integrity. When scientists or students write a paper or report, they must carefully cite all sources used; otherwise, the work may be considered plagiarism. Unfortunately, in ChatGPT’s answers, sources are not always accurately referenced.

More research is needed to make AI systems more explainable and trustworthy by using interpretable or explainable machine learning algorithms. Such algorithms should produce results that can be easily understood by humans and provide insights into how the AI system arrived at its predictions or decisions. Additionally, it is crucial for such systems to provide a measure of uncertainty in their answers, such as confidence intervals or standard deviations, similar to scientific practices, where results are often reported with a range of values that account for possible variations in the data or measurement errors.

Another significant concern with AI methods, particularly deep learning algorithms, is that they consume a lot of

power. This is because these algorithms require large amounts of computational resources to train and run. They use huge servers and high-performance GPUs, which require high power, large amounts of memory, and also communications between servers and GPUs. The energy consumption is only going to increase as the use of AI continues to grow, and more powerful AI systems are developed. In addition to the environmental impact of energy consumption, high levels of power consumption can also result in higher operating costs and can limit the scalability and accessibility of AI systems.

There is increasing research and development focused on developing more energy-efficient AI systems. This involves a range of techniques and strategies, including the use of specialized hardware, such as tensor processing units; the development of more efficient algorithms and architectures; and the use of techniques, such as model compression and pruning, to reduce the computational requirements of AI systems. In addition to these technical approaches, there is also a need for broader policy and regulatory measures to encourage the development and adoption of energy-efficient AI systems. This could include incentives for energy-efficient design, regulations on the energy consumption of AI systems, and the development of standards and benchmarks to encourage the use of more energy-efficient AI technologies [4], [5], [6]. It is also important to consider whether AI is necessary to solve the problem at hand or whether simpler and less costly solutions exist. Furthermore, it’s crucial to evaluate the impact of any proposed AI solution on both humans and the environment. In evaluating and comparing AI systems, one should use metrics that take into account both performance

and complexity or power consumption, such as the Akaike criterion [7] or similar ones.

The rapid evolution of technology has opened the doors to many extraordinary breakthroughs that have had an incred-

**With all of the computational power available today, AI can analyze large amounts of data and identify patterns and insights that might be difficult for humans to detect.**

ible impact on our field and will continue to transform the world. Let’s celebrate these achievements, but let’s also be mindful that, with these promising technologies, there can also be significant peril—to scientific progress,

to society and human well-being, and to the ecological environment. Innovation comes with great responsibility. Let us all do our best to be smart and thoughtful as we navigate the future.

In this *IEEE Signal Processing Magazine* special issue celebrating the 75th anniversary of the SPS, you will find additional insights into the history of SPS during the last decades and more technical articles about the evolution, breakthroughs, and discoveries in different domains in SIP.

## References

- [1] S. Matsue et al., “A 256 K dynamic RAM,” in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, USA, 1980, pp. 232–233, doi: 10.1109/ISSCC.1980.1156048.
- [2] W. S. Boyle and G. E. Smith, “Charge coupled semiconductor devices,” *Bell Syst. Tech. J.*, vol. 49, no. 4, pp. 587–593, Apr. 1970, doi: 10.1002/j.1538-7305.1970.tb01790.x.
- [3] C. Jutten and J. Héroult, “Analog implementation of a permanent unsupervised learning algorithm,” in *Proc. NATO Workshop Neurocomputing*, Les Arcs, France, 1989, pp. 145–152, doi: 10.1007/978-3-642-76153-9\_18.
- [4] E. Azarkhish, D. Rossi, I. Loi, and L. Benini, “Neurostream: Scalable and energy efficient deep learning with smart memory cubes,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 2, pp. 420–434, Feb. 2018, doi: 10.1109/TPDS.2017.2752706.
- [5] W. J. McKibbin and A. C. Morris, *Policy Challenges for the Global Transition to a Low-Carbon Economy*. Washington, DC, USA: Brookings Institution, 2019.
- [6] V. Galaz et al., “Artificial intelligence, systemic risks, and sustainability,” *Technol. Soc.*, vol. 67, Nov. 2021, Art. no. 101741, doi: 10.1016/j.tech-soc.2021.101741.
- [7] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974, doi: 10.1109/TAC.1974.1100705.





2023.ieeeicip.org

# International Conference on Image Processing

OCTOBER 8 -11, 2023



IEEE

The 30th IEEE International Conference on Image Processing (ICIP 2023) will be held in Kuala Lumpur, Malaysia, on October 8-11, 2023. ICIP is the world's largest and most comprehensive technical conference focused on image and video processing and computer vision.



## Topics of interest included but not limited to:

- Sensing, Representation, Modeling, and Registration
- Motion Estimation, Registration, and Fusion
- Synthesis, Rendering, and Visualization
- Deep Learning for Images and Videos
- Computational Imaging
- Learning with Limited Labels
- Restoration and Enhancement
- Image & Video Interpretation and Understanding
- Compression, Coding, and Transmission
- Detection, Recognition, Retrieval, and Classification
- Color, Multi-spectral, and Hyper-spectral Imaging
- Biometrics, Forensics, and Security
- Stereoscopic, Multi-view, and 3D Processing
- Biomedical and Biological Image Processing
- Image & Video Quality Models
- Emerging Applications and Systems

## Regular ICIP Paper Submission

Authors are invited to submit full-length papers (up to 4 pages for technical content including figures and references, and one optional 5th page containing only references). Submission instructions, templates for paper format, and the “no show” policy are available at the conference website. All accepted papers will be published in IEEE Xplore.

## Open Journal of Signal Processing (OJSP) Paper Submission

For the first time at ICIP, authors are invited to submit their papers, aligned with the conference scope and intended for presentation at ICIP, to the IEEE Open Journal of Signal Processing (OJSP). These papers may be up to 8 pages + 1 page for references, and their review will be managed by the editorial board of OJSP, expedited to ensure that a decision is made prior to finalization of the ICIP technical program. Accepted papers will be published in OJSP and will also be scheduled for presentation at ICIP. Submission implies a commitment that an author will attend the conference to present the paper at ICIP if it is accepted.

Acceptance of submissions to OJSP under this “expedited review for ICIP” program may close early if demand is unexpectedly heavy.

## Author Rebuttal

At the discretion of the Area Chairs, some authors may be asked for a rebuttal of the reviews received.

## Journal Paper Presentation

Authors of papers published in all IEEE Signal Processing Society fully owned journals as well as in IEEE TCI, IEEE Signal Processing Letters, IEEE TIP, IEEE TMI, IEEE TM, IEEE TIFS, IEEE TSP, IEEE Journal on STSP, and IEEE SPM will be given the opportunity to present their works at ICIP 2023, subject to space availability and approval by the Technical Program Committee. Visit the conference website for more details.

## Tutorials, Special Sessions, and Challenge Sessions Proposals

Tutorials will be held on October 8, 2023. Tutorial proposals must include a title, outline, description of the materials to be covered/distributed, and the contact information, biography and selected publications of the presenter(s). Special sessions and challenge sessions proposals must include a topical title, rationale, session outline, contact information, and a list of invited papers/participants. For detailed submission guidelines, please visit the conference website.

## Open Preview

Open Preview allows conference proceedings to be available in the IEEE Xplore Digital Library, free of charge, to all customers, 30 days prior to the conference start date, through the conference end date.



## ORGANIZING COMMITTEE

### General Co-Chairs

Norliza Mohd Noor (UTM)  
 Gaurav Sharma (U. Rochester)  
 Mohan Kankanhalli (NUS)

### Technical Program Co-Chairs

Eduardo AB da Silva (UFRJ)  
 Stefan Winkler (NUS)  
 Jing-Ming Guo (NTUST)

### Finance Chair

Mohammad Faizal Ahmad Fauzi (MMU)  
 Aly Farag (Louisville U.)

### Plenary Co-Chairs

Shri Narayanan (USC)  
 Syed Abdul Rahman Syed Abu Bakar (UTM)

### Special Session Co-Chairs

Chin Tat-Jun (Adelaide U.)  
 Wong Lai Kuan (MMU)

### Tutorials/Co-Located Chair

Rajasvaran Logeswaran (City U.)

### Publication Co-Chairs

Chong-Wah Ngo (SMU)  
 John See (HW)

### Publicity Co-Chairs

Syed Khaleel Ahmed (Consultant)  
 Susanto Rahardja (SIT)

### Industry/Exhibit Chair

Liu JianQuan (NEC)  
 Hezerul Abdul Karim (MMU)

### Innovation Program/ Competition Co-Chairs

Wong Kok Sheik (Monash)  
 Jean Luc Dugelay (EUROCOM)

### Local Arrangement Co-Chairs

Vijanth Asirvadam (UTP)  
 Hoo Wai Lam (UM)

### Registration Co-Chairs

Nor'aini Abdul Jalil (Wavesmiles)  
 Haidawati Mohamad Nasir (UniKL)

### Awards Co-Chairs

Jocelyn Chanussot (Grenoble U.)  
 Ma Kai Kuang (NTU)

### Student Activity Co-Chairs (Post Doctoral Consortium)

Kushsairy Abdul Kadir (UniKL)  
 Fabrice Meriaudeau (Bourgogne U.)

### Social Media Chair

Mohd Norzali Haji Mohd (UTHM)  
 Loh Yuen Peng (MMU)

### Keynote Chair

Cheng Wen-Huang (NYCU)  
 Ba- Ngu Vo (Curtin U.)

## IMPORTANT DATES

Challenge Proposal Deadline:  
**November 23, 2022**

Challenge Proposal Acceptance Notification:  
**December 14, 2022**

Special Session Proposal Deadline:  
**December 7, 2022**

Special Session Proposal Acceptance Notification:  
**January 4, 2023**

Tutorial Proposal Deadline:  
**January 11, 2023**

Tutorial Proposal Acceptance Notification:  
**February 1, 2023**

Regular & Special Session Paper Submission Deadline:  
**February 15, 2023**

Challenge Paper Submission Deadline:  
**March 1, 2023**

Journal Presentation Request Deadline:  
**May 31, 2023**

Paper Acceptance Notification:  
**June 21, 2023**

Journal Presentation Acceptance Notification:  
**June 21, 2023**

Final Paper Submission Deadline:  
**July 5, 2023**

Author Registration Deadline:  
**July 12, 2023**

Supported by:



Connect with us!



# Empowering the Growth of Signal Processing

*The evolution of the IEEE Signal Processing Society*



©SHUTTERSTOCK.COM/TRIFF

**S**ignal processing (SP) is a “hidden” technology that has transformed the digital world and changed our lives in so many ways. The field of digital SP (DSP) took off in the mid-1960s, aided by the integrated circuit and increasing availability of digital computers. Since then, the field of DSP has grown tremendously and fueled groundbreaking advances in technology across a wide range of fields with profound impact on society. The IEEE Signal Processing Society (SPS) is the world’s premier professional society for SP scientists and professionals. Through its high-quality publications, conferences, and technical and educational activities, the SPS has played a pivotal role in advancing the theory and applications of SP. It has been instrumental in promoting cross-disciplinary collaboration and knowledge sharing among researchers, practitioners, and students in the field. This article highlights the SP advances between 1998 and mid-2023 and the evolution of the SPS to empower the growth of SP.

## Introduction

Without hyperbole, SP is behind much of the digital world we live in today. The field of DSP took off in the mid-1960s, aided by the integrated circuit of Kilby and Noyce in the 1950s, the microprocessors of Texas Instruments and Intel in the 1960s, and the increasing availability of digital computers. A big push into the field can be attributed to the fast Fourier transform (FFT), by James Cooley and John Tukey, which reduced from  $O(N^2)$  to  $O(N \log N)$  the computation time of the FT. This allowed many SP algorithms that were already available to be implementable in close to real time. Around the same time, the first book on DSP, by Ben Gold and Charles Rader, appeared [1]. Since then, the field of DSP has grown tremendously and fueled groundbreaking advances in technology across many fields with profound impact on society. For example, DSP has revolutionized the way we create, store, and transmit audio and video content. DSP has enabled digital audio processing, high-quality audio recordings, and streaming services. Similarly, digital video processing techniques, such as compression, have made it possible to transmit high-quality video content over various networks. DSP has played a crucial role in the development of wireless



communication systems and the smartphone, which has become so ubiquitous in all aspects of our daily life that it is difficult for most people to imagine life without it. Techniques such as channel coding, and equalization have made it possible to achieve high data rates and reliable wireless communication over long distances. These techniques led to the widespread adoption of wireless technologies, such as Wi-Fi, Bluetooth, and cellular networks. DSP techniques, such as array processing, have played determining roles in geophysics exploration, radar, sonar, and other related applications. DSP has been instrumental in the development of medical imaging techniques, such as magnetic resonance imaging, computed tomography scans, and ultrasound. These technologies rely on DSP algorithms to process raw data and create high-resolution images of the human body, enabling doctors to diagnose and treat a wide variety of medical conditions with greater accuracy and precision. DSP has also enabled significant advancements in speech and audio recognition. Techniques such as voice recognition, speech to text, and music recognition rely on DSP algorithms to analyze and classify audio signals. This has led to the development of many popular applications, including virtual assistants, transcription services, and music streaming platforms. DSP has enabled the development of advanced control systems for a variety of applications, including robotics, aerospace, and automotive industries. DSP algorithms are used to analyze sensor data and control the behavior of complex systems, with high accuracy and precision.

The SPS is the world's premier professional society for SP scientists and professionals. It has nearly 20,000 members across 120+ countries. Through high-quality publications, conferences, technical, and educational activities, the SPS advances and disseminates state-of-the-art scientific information and resources, educates the SP community, and, by bringing people together, catalyzes advances in the field of SP.

The SPS has had many names since it was established, on 2 June 1948, as the first Professional Group on Audio of the Institute of Radio Engineers (IRE). In 1963, the IRE merged with the American Institute of Electrical Engineers to form IEEE, and the Professional Group on Audio became the IEEE Audio Group, in 1964. In 1976, the IEEE Audio Group was renamed the IEEE Acoustics, Speech, and Signal Processing (ASSP) Society, reflecting the Society's expanding scope beyond audio processing to include SP in a broader sense. In 1989, the ASSP Society changed its name to the SPS, due to the growing field of image processing.

The SPS currently has 12 technical committees (TCs), 3 technical working groups (TWGs) and 2 megatrend initiatives that support a broad selection of SP-related activities associated with specific areas of study within the SP field. The TCs are actively involved in awards, conferences, publications, and educational activities. The Society's leadership leans heavily on TC members for their advice on specific areas within SP. The TCs are

- 1) Applied Signal Processing Systems TC
- 2) Audio and Acoustic Signal Processing TC
- 3) Bio Imaging and Signal Processing TC
- 4) Computational Imaging TC
- 5) Image, Video, and Multidimensional Signal Processing TC



- 6) Information Forensics and Security TC
- 7) Machine Learning for Signal Processing TC
- 8) Multimedia Signal Processing TC
- 9) Sensor Array and Multichannel TC
- 10) Signal Processing for Communications and Networking TC
- 11) Signal Processing Theory and Methods TC
- 12) Speech and Language Processing TC.

The TWGs include

- 1) Industry TWG
- 2) Integrated Sensing and Communication TWG
- 3) Synthetic Aperture TWG.

The megatrend initiatives are:

- 1) Autonomous Systems Initiative
- 2) Data Science Initiative.

The SPS currently publishes several high-impact periodicals, including *IEEE Signal Processing Magazine*; *IEEE Open*

*Journal of Signal Processing (OJ-SP); IEEE Journal of Selected Topics in Signal Processing; IEEE Signal Processing Letters; IEEE/ACM Transactions on Audio, Speech, and Language Processing; IEEE Transactions on Information Forensics and Security; IEEE Transactions on Image Processing; IEEE Transactions on Signal Processing; IEEE Signal Processing Society Content Gazette; and Inside Signal Processing Newsletter.* The SPS publishes about 3,000 journal papers annually.

Reflecting the highly interdisciplinary nature of SP, the SPS publishes jointly with other IEEE Societies a growing list of journals, including *IEEE Transactions on Computational Imaging, IEEE Transactions on Signal and Information Processing Over Networks, IEEE Transactions on Multimedia, IEEE Transactions on Big Data, IEEE Journal on Biomedical and Health Informatics, IEEE Transactions on Cognitive Communications and Networking, IEEE Transactions on Medical Imaging, IEEE Transactions on Mobile Computing, IEEE Transactions on Wireless Communications, and IEEE Wireless Communications Letters.*

The SPS is also involved with a large number of IEEE-level publications, including *IEEE Sensors Journal, IEEE Control Systems Letters, IEEE Transactions on Affective Computing, IEEE Computing in Science and Engineering Magazine, IEEE Internet of Things Journal, IEEE Internet of Things Magazine, IEEE Transactions on Computational Social Systems, IEEE Life Science Letters, IEEE MultiMedia Magazine, IEEE Transactions on Network Science and Engineering, IEEE Reviews in Biomedical Engineering, IEEE Transactions on Smart Grid, IEEE Security & Privacy Magazine, IEEE Transactions on Artificial Intelligence, IEEE Transactions on Green Communications and Networking, IEEE Transactions on Quantum Engineering, IEEE Transactions on Computational Social Systems, IEEE Transactions on Machine Learning in Communications and Networking, IEEE Journal of Indoor and Seamless Positioning and Navigation, and IEEE Transactions on Radar.*

The Society also organizes several conferences and workshops each year as a sole sponsor [7]. The two flagship conferences are the International Conference on Acoustics Speech and Signal Processing (ICASSP) and the International Conference on Image Processing (ICIP). ICASSP was first held in 1976 and is, in a sense, the continuation of the Arden House workshop, which was first held in 1968 with a focus on the FFT. The first ICIP was held in 1994. Other SPS solely sponsored workshops, mainly focused on the areas covered by the SPS TCs and on new areas the SPS is exploring, include the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU); IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP); IEEE Data Science and Learning Workshop (DSLW); IEEE Workshop on Image, Video, and Multimedia Signal Processing (IVMSP); IEEE Workshop on Machine Learning for Signal Processing (MLSP); IEEE Workshop on Multimedia Signal Processing (MMSP); IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM); IEEE Workshop on Spoken Language Technology (SLT); IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC); IEEE Workshop on Statistical Signal Processing (SSP);

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA); and IEEE International Workshop on Information Forensics and Security (WIFS). Several thousand people attend our conferences annually, and conference recordings are kept in our SPS Resource Center for later access.

The SPS also cosponsors a growing list of conferences and workshops, including the IEEE International Conference on Multimedia and Expo (ICME), IEEE International Symposium on Biomedical Imaging (ISBI), IEEE Conference on Advanced Video and Signal-Based Surveillance (AVSS), ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE Workshop on Signal Processing Systems (SiPS) and IEEE Conference on Artificial Intelligence (IEEE CAI).

Over the years, the SPS has played a pivotal role in advancing the theory and applications of SP. It has been instrumental in promoting cross-disciplinary collaboration and knowledge sharing among researchers, practitioners, and students in the field.

## **Key developments in SP and the SPS**

The developments in SP and the evolution of the SPS up to 1998 are described in [2], which was published on the 50th anniversary of the Society. In this article, we summarize some of that history and expand on developments after 1998 until mid-2023.

### *The 1940s and 1950s: The advent of DSP*

The Wiener filter, in the 1940s and 1950s, and Kalman and Kalman-Bucy filtering, in 1960, addressed the processing of noisy signals with many applications, from radar to communications and guidance and control. In the 1960s, SP techniques were developed in geophysics exploration for oil discovery, with Burg developing his linear prediction algorithm that also found wide application in speech processing. Array processing techniques, such as Capon's, detect seismic events and track underwater targets. The analysis of time series motivated by detection of underground nuclear explosions led Cooley and Tukey, at IBM, to propose, in 1965, a new fast implementation for the FT, the now ubiquitous FFT. The FFT reduced the computation time of the FT by orders of magnitude. This allowed many available SP algorithms to be implementable in close to real time. Around the same time, the first book on DSP appeared, by Gold and Rader [1]. Toward the end of the decade, the statistical theory of SP was finding more and more applications, with detection and estimation as major areas of activity. In 1968, Harry L. Van Trees penned the seminal book on detection, estimation, and modulation theory, summarizing the main tenets of the theory with applications to radar, sonar, and communications [3]. But these SP developments were happening in parallel to the core of DSP, as DSP was then emerging from speech and audio and radar applications through work, for example, at Lincoln Laboratories and Bell Labs.

### *The 1970s: DSP receives public attention and the rise of personal computers*

The invention of the integrated circuit, by Jack Kilby, of Texas Instruments, in 1958 [8], and also independently by

Robert Noyce, of Fairchild, in 1959, significantly accelerated the development of digital computers. In 1971, Texas Instruments introduced the TMS 1802NC, and Intel created the Intel 4004, the first microprocessors with a 4-bit chip and a clock speed of 108 kHz [9]. In 1981, IBM introduced the first personal computer with a built-in hard disk, the IBM PC 5150 [10]. It had a 5.25-in floppy disk drive, 16 KB of random-access memory (RAM), and a 4.77-MHz Intel 8088 processor. In the 1970s and early 1980s, typical RAM sizes were in the range of a few hundred bytes to a few kilobytes, while the read-only memory (ROM) sizes were of the order of kilobytes. Hard disks were not yet widely available for personal computers, so data were typically stored on floppy disks with capacities of a few hundred kilobytes to a few megabytes. As computer technology advanced throughout the 1980s, clock speeds and memory capacities increased rapidly. By the end of the decade, personal computers were running at speeds of several tens of megahertz and had RAM capacities of several megabytes. These developments paved the way for the emergence of DSP and image processing, which rely heavily on fast processing speeds and large amounts of memory.

In the 1970s, DSP started receiving increased attention from the general public. During that time, in Britain, the BBC began using eight-track digital audio recorders with error correction [2], [4]. Thomas Stockham showed how DSP could restore old recordings of Enrico Caruso. In 1978, Texas Instruments designed a popular toy called Speak & Spell, which taught spelling by pronouncing a word and providing input on whether a spelling attempt was correct. Key DSP technologies were crucial to those advances, such as speech compression and the availability of the first integrated circuits for SP. During that decade, landmark books on DSP by Alan V. Oppenheim and Ronald W. Schaffer (1975) and Lawrence Rabiner and Ben Gold (1975) as well as the first book on digital speech processing, by Rabiner and Schaffer (1978), appeared [11], [12], [13]. Toward the end of the decade, Ralph O. Schmidt, with his MUSIC algorithm (published in the open literature in 1979), [14], and Georges Bienvenu and Laurent Kopp (1979), [15] introduced high-resolution subspace-based techniques to detect and localize nearby sources.

During that decade, the ASSP Society membership grew from 5,299 to 8,619 members. SPS publications included *IEEE Transactions on Audio and Electroacoustics*; *IEEE Transactions on Acoustics, Speech, and Signal Processing (T-ASSP)*; *IEEE Newsletter on Audio and Electroacoustics*; and *IEEE Acoustics, Speech, and Signal Processing Newsletter*.

### *The 1980s: DSP a key player in data storage, image and video processing, and medical imaging*

With the emergence of personal computers and cellular phones, array SP and digital communications became major areas of activity. Increasing levels of recording density required sophisticated new detection algorithms to read back accurately the recorded bits. Wavelets also appeared on the scene along with the first CDs, offering a new digital format for storing and playing music. The CD quickly replaced vinyl records and

cassette tapes as the dominant music format. Biotechnology emerged as a significant field of study, with the development of new techniques for genetic engineering, gene sequencing, and biopharmaceutical production. This decade also witnessed much interest in digital image processing that laid the ground for video processing growth and led to important advances in a wide variety of applications, including multimedia, computer vision, medical imaging, image and video compression, virtual reality, and biometrics and facial recognition, to mention a few. The Society's journals and conferences were the publication of choice for much of the work on wavelets.

DSP was a key player in these technologies, and the ASSP Society membership grew from 8,619 to 15,925. SPS publications included *T-ASSP*; *IEEE Acoustics, Speech, and Signal Processing Newsletter*; and *IEEE ASSP Magazine*.

The growth of the Society led to the development of the Publications Board, the Conference Board, and the Awards Committee. In 1981, the ASSP Society joined the IEEE Engineering in Medicine and Biology Society (EMBS), the IEEE Nuclear and Plasma Sciences Society, and the IEEE Sonics and Ultrasonics Society to establish a new quarterly journal, *IEEE Transactions on Medical Imaging*. To address the increasing need for more content, in 1984, ASSP newsletter became *IEEE ASSP Magazine*.

### *The 1990s: Distributed web and new organizational structure for SPS' rapid growth*

The World Wide Web rapidly grew in popularity, revolutionizing the way people accessed and shared information online. Personal computers became more affordable and widespread, and mobile phones became smaller, more affordable, and more popular. CD-ROMs became a popular storage medium for computer software, music, and video, replacing floppy disks and cassette tapes. Through the decade, disk drives recording densities grew at faster rates than Moore's law allowing for storing ever increasing amounts of data and requiring new signal processing algorithms to retrieve the data. JPEG was standardized in 1992, followed by the H.261 and MPEG conference and video standards. Digital cameras began to replace film cameras. E-mail became a widely used form of communication, voice over Internet Protocol technology was introduced, and e-commerce emerged. GPS became available for civilian use, allowing for accurate location tracking and navigation. Advances in wireless communications and Wi-Fi allowed for communication and computing anytime, anyplace, anywhere. Again, DSP played a big role in those technologies. On the research front, among many other areas, compressed sensing techniques experienced significant activity, with many papers appearing in SPS journals and conferences.

A new journal, *IEEE Transactions on Image Processing*, was introduced, in 1992, as a quarterly publication but quickly became monthly. Also in 1991, *T-ASSP* was renamed *IEEE Transactions on Signal Processing (TSP)*. This was the year that the JPEG standard was established. *IEEE Transactions on Speech and Audio Processing* was introduced in 1993, *IEEE Signal Processing Letters* in 1994, and *IEEE*

*Transactions on Multimedia* in 1999. The SPS also cosponsored four other journals, including *IEEE Transactions on Evolutionary Computing*, *IEEE Transactions on Fuzzy Systems*, *IEEE Transactions on Medical Imaging*, and *IEEE Transactions on Neural Networks*. *IEEE ASSP Magazine* became *IEEE Signal Processing Magazine* in 1991. ICIP was established to address the rapidly growing field of image processing, and it was held for the first time in 1994.

During this decade, the ASSP Society became the SPS, and the membership grew to 19,835. The growth necessitated major revisions of the Society bylaws. A new administrative structure was approved, in 1993, so that the Society would be headed by a Board of Governors (BoG) consisting of the Society officers and 12 members at large. A smaller Executive Committee was created that would act on Society matters between the biannual BoG meetings.

In 1993, Mercy Kowalczyk became the first executive director of the Society. She hired Theresa Argiropoulos, in 1993, to assist with operational support. At that time, the three SPS publications were managed externally, by Peirce and Barbara Wheeler. In 1996, Nancy DeBlasi was hired to transition the publications operations in-house, and by the end of the year, Deborah Blazek was also hired to support the publications business. In 1998, additional staff was hired to address the growing workload in operations (Linda Skeahan) and publications (Kathy Jackson and Jo-Ellen Snyder). By 1998, the submission and peer review of papers of the *Transitions on Signal Processing* moved to the online management system Manuscript Control (MC).

In 1996, the Society selected its first logo, which was designed by Gabriel Thomas, at the time a graduate student at the University of Texas at Austin. This SPS logo was used between 1996 and 2019.



The Society took its first steps in electronic publishing, making the proceedings of the 1993 ICASSP available on CD-ROM—a first for IEEE conferences. In 1997, *IEEE Signal Processing Letters* was one of the first IEEE publications to be made available online. The following year, all Society transactions as well as letters were available online.

### *The 2000s: Distributed information processing and new SPS technical activities structure*

With the emergence and proliferation of sensor networks, smartphones, and social media platforms, cloud computing became available, making it possible to store and access data over the Internet, revolutionizing the way businesses and individuals store and access information. Streaming services, such as Spotify, were introduced, and Netflix expanded, changing the way we consume entertainment. The use of artificial intelligence (AI) became more prevalent in the 2000s, with

advancements in signal processing and machine learning, and with applications such as speech recognition and image recognition. AI has since become increasingly important in many industries, including health care, finance, and transportation. Sensor networks entered the scene, leading to a burst of research in distributed and decentralized information processing and optimization that became significant new areas to utilize the way data are collected, stored, and processed.

In 2002, the first ISBI was held; it was cosponsored and run by the EMBS and SPS. To strengthen the Society's coverage of biomedical topics, the Bio Imaging and Signal Processing TC was established, in 2004. The Society's interest in security issues with emerging technologies led to the 2006 creation of the Information Forensics and Security Technical Community. In 2008, the Image and Multidimensional Signal Processing TC, which was established in 1991, changed its name to the Image, Video, and Multidimensional Signal Processing TC. The Neural Networks for Signal Processing TC, which was founded in 1990, became the Machine Learning for Signal Processing TC, in 2003.

In 2006, the scope of *IEEE Transactions on Speech and Audio Processing* was expanded, with the journal becoming *IEEE Transactions on Speech, Audio, and Language Processing*. Continued progress was made with the Society's efforts to publish high-quality and relevant periodicals. In 2004, *IEEE Signal Processing Society Magazine* was ranked number 1 among IEEE journals in the Journal Citation Report, and it has since remained among the top journals in the field of electrical and electronic engineering as well as computer science. The magazine is widely recognized for its high-quality articles and practical tutorials that cover a wide range of topics in SP.

In 2006, *IEEE Transactions on Information Forensics and Security* was launched, and the following year, *IEEE Journal of Selected Topics in Signal Processing* was introduced. In 2009, the SPS became a technical cosponsor of two new IEEE publications, *IEEE Biometrics Compendium* and *IEEE Transactions on Affective Computing*. That year, IEEE International Workshop on Information Forensics and Security (WIFS) was also introduced.

In 1999, the Society approved a plan to digitize all its content, including journals, workshops and conference proceedings, newsletters, and other publications sponsored by the SPS. This led, in 2002, to the Society's Signal Processing Electronic Library (SPeL), containing all material published by the SPS from its 1948 beginnings through 2005. SPeL was released on two DVD-ROMs and enthusiastically received by SPS and other IEEE Members. It was intended to be a useful travel companion for SPS members. But events and technology dictated otherwise, and the Society donated the SPeL digital content to IEEE to form the emergent IEEE digital library, now *IEEE Xplore*. In 2001, SPS became the first IEEE Society with submission and peer review of all the journal papers handled online through MC.

During this decade, IEEE witnessed a drop in its membership, and the SPS membership changed from 19,835 to 14,897, coinciding with a shift of journal subscriptions toward

institutional customers instead of individual subscribers. The launch of IEEE *Xplore* gave broad access to SPS digital content through universities' and companies' *Xplore* subscriptions.

Other Society changes included the 2007 formation of a TC review committee charged with conducting formal reviews of the TCs, chaired by the president-elect. From this experience, the Society reformulated its Executive Committee and created the vice president, technical directions position, and the Technical Directions Committee was elevated to the Technical Directions Board, in 2007.

In 2008, effective 2010, the Society decided that all fully sponsored periodical publications of the SPS would become available in electronic format for free to all Society members as a member benefit. To help authors search for papers of interest, the Society introduced the monthly *IEEE Signal Processing Society Content Gazette*, in 2010, which contained the table of contents pages of all the Society's periodical publications. The digital versions of *IEEE Signal Processing Magazine* and the *Gazette* were also provided free to members.

Over the years, staffing grew to support the Society's expanding operations and new initiatives. Staff worked closely with the volunteer leadership on all areas of the Society's activities. The conference business was also growing, and the first conference staff was hired to provide support to that activity.

### *The 2010s: Higher-speed communications and emphasizing membership services*

The popularity of smartphones dramatically increased in the 2010s. The introduction of 4G networks made it possible to access high-speed Internet on mobile devices. Internet of Things (IoT) technology became prevalent with the increasing number of connected devices, such as smart homes, wearables, and the industrial IoT. Graph-based data proliferated, initiating a new area in SP, graph SP. AI technology continued to progress with advancements in signal processing, deep learning and machine learning. In commercial SP technologies, AI is now used in a wide range of industries, including health care, finance, and manufacturing. Cloud computing continued to advance in the 2010s, making it possible to scale information technology infrastructure efficiently, making it easier for businesses to grow. The development of self-driving cars emerged with the potential to revolutionize transportation.

With such dramatic growth in commercial SP technologies, the Society focused its efforts on enhancing member services. In 2011, the SPS Membership Board was formed, and the position of regional director at large was created to bring regional perspective to the BoG and Membership Board. As a result of the Membership Board's formation, the vice president, awards and membership position was separated into the vice president, membership and Awards Board chair positions, and only the vice president, membership was a member of the Executive Committee. In 2014, the Executive Committee was further amended, with the president-elect also taking on the responsibilities of vice president, finance. With the increased emphasis on member services, SPS membership grew from 14,897 to 18,730 members in this decade.

Student membership was also cultivated. The first annual Signal Processing Cup (SP Cup) was established, in 2014. The SP Cup is a student competition in which graduate and undergraduate students work in groups to solve real-world problems by using SP methods and techniques. The program was expanded, in 2017, to include the Video and Image Processing Cup and, in 2020, the Five-Minute Video Clip Contest. The Student Career Luncheon at ICASSP was launched to help students explore job opportunities by connecting them with industry representatives. Since 2015, the Women in Signal Processing Luncheon has become an SPS-sponsored event at all major SPS conferences, and similarly, the Young Professionals (YPs) luncheon event launched in 2016, emphasizing the role of women and YPs in the Society. The Young Professionals Development Workshop was introduced at ICASSP 2019.

Other SP initiatives include the 2013 launch of the IEEE Global Conference on Signal and Information Processing (GlobalSIP) along with the IEEE China Summit on Signal and Information Processing (ChinaSIP). In 2018, the BoG approved discontinuing GlobalSIP after 2019, and beginning in 2016, ChinaSIP continued for a few years as the SPS Signal/Data Science Forum.

In 2013, Richard Baseil became the SPS executive director. That year, two special interest groups (SIGs) were established to address technical areas in big data and the IoT. In 2015, a third SIG was approved, on computational imaging, which was later elevated to a TC, in 2018.

In 2014, the SPS teamed up with the Association for Computing Machinery to jointly publish *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. That year, two member benefits were introduced: SigView, an online portal of video tutorials with valuable educational content, and SigPort, an online archive of manuscripts, reports, theses, and supporting materials providing early exposure and peer feedback on work that is in progress. SigView videos were later relocated to the SPS Resource Center. Recognizing that data are a key element of SP, SigPort was duplicated and expanded to become IEEE DataPort, now an IEEE-wide product. Two new journals were added in 2015, *IEEE Transactions on Signal and Information Processing Over Networks* and *IEEE Transactions on Computational Imaging*.

In 2015, the Society also looked to popularize and promote SP and its applications to the general public. Target audiences included high-school and college students and other non-SP professionals. Several visibility videos were posted on the SPS YouTube channel: "What Is Signal Processing," "Signal Processing and Machine Learning," "Signal Processing in Free Viewpoint Television," "The Benefits of Spoken Language Technology," "Multimedia Forensics," and "Under the Radar." To enhance global reach, some videos were translated into Arabic, Spanish, and Mandarin. Videos about careers in SP were also created as well as other SP technology-related videos.

In 2016, the SPS created the IEEE Foundation Student and Young Professionals Fund, which is aimed toward enabling SP student and YPs programs and activities. In 2019, to promote student engagement, the student/graduate student Society

membership fee was set to US\$1 annually. The Society also focused on diversity, equity, and inclusion (DEI), and its first diversity statement was created along with a diversity pledge and a DEI webpage [6].

Ethical use of SP techniques has always been an emphasis of the Society. Privacy is an ongoing concern among many SP-related fields, such as biometric data and datasets, so ethical guidelines continue to evolve for our publications and conferences.

SPS conferences started strengthening industry links. ICIP 2016 included an innovation program that featured state-of-the-art vision technologies, innovation challenges, talks by innovation leaders and entrepreneurs, tutorials, and networking. ICIP 2016 also launched a Visual Technology Innovation Award for industry leaders. ICIP 2017 also featured industry-focused keynote talks, panels, and programs related to several existing and emerging technologies. These industry initiatives have since been adapted by some ICIPs.

Another SPS change occurred in 2019. Historically, the BoG was responsible for electing Society presidents, but that year, it opened the vote to all members. A petition process was also instituted to give members additional opportunities to be heard. The Society also updated its logo to reflect the expanding world of DSP.



### *The 2020s: Autonomous systems, AI, data sciences, and new outlooks for SP*

The first few years of the 2020s have been turbulent, to say the least. The COVID-19 pandemic impacted all aspects of life. Lockdowns and social distancing measures increased the demand for video conferencing, e-commerce, virtual events, online learning platforms, and digital health technologies, such as telemedicine and virtual care. Meanwhile, the rollout of 5G networks is providing faster and more reliable Internet connections, and the development of 6G is aiming to generate even higher rates to support autonomous vehicles and smart cities. AI applications, such as deep learning, natural language processing, and predictive analytics, are being used in a wide range of industries, including health care, finance, and transportation. Quantum computing technology promises to solve problems that are currently unsolvable with traditional computers, in areas such as cryptography, drug discovery, and weather forecasting.

The SPS has two megatrend initiatives: the SPS Autonomous Systems Initiative (ASI) and the SPS Data Sciences Initiative (DSI). ASI aims at highlighting the central role of SP in the design and development of autonomous systems, a multidisciplinary area cutting across AI, robotics, and the IoT. DSI coordinates the activities of the various TCs on data science, another area at the heart of SP. SP takes a broad view of signals and data since, once digitized, signals are data. To address the new set of applications, SP journals and conferences capture

much of the research activity in distributed and decentralized peer-to-peer and networked environments and, in graph SP (GSP), the new theories and applications of graph-based data. Our publications and conferences continue exploiting a priori information about structure in problems/data, connections to physical applications (such as 3D audio, radar, and ultrasound), social and other emerging applications, and connection to computational platforms and scenarios, e.g., distributed computing, edge computing, and processing at the device, through peer-to-peer communications, possibly with no cloud and edge connectivity. DSI launched very successful webinars on brain research and GSP and has established a working group that works with the Education Board on incorporating topics around data science in academic and postacademic education curricula. The SPS Education Webinar program also grew significantly in 2022, when we offered a total of 55 webinars on cutting-edge topics. Some webinars are author solicitations—invitations based on *Xplore* article analytics—and some are arranged by the various TC and SPS initiatives.

In addition to all its journals being hybrid open access, the SPS now has a fully open access journal, *OJ-SP*. *OJ-SP* recently introduced new paper categories; in addition to regular papers, it now accepts short papers (eight + one pages long), overview papers, and dataset/competition/challenges papers.

To date, the SPS has 20 financially and technically cosponsored journals. In 2022, the SPS added *IEEE Journal on Indoor and Seamless Positioning and Navigation* (open access), *IEEE Transactions on Radar Systems* (hybrid), and *IEEE Transactions on Machine Learning in Communications and Networking* (open access).

In 2019, TWGs were established, and three have been created to address the areas of industry, integrated sensing and communication, and synthetic aperture. With the creation of the Synthetic Aperture TWG, the SPS is now involved in the development of standards. The SPS Synthetic Aperture Standards Committee continues to experience steady growth and increasing interest from the research community, which is a testament to the need for market-driven standards in this technology space.

To alleviate global pandemic restrictions, in 2020, ICASSP provided free remote access for nonmembers and nonauthors. Over 16,000 attendees joined the ICASSP virtual platform, most of whom were not SPS members, confirming that ICASSP topics, trends, and technologies are increasingly popular and growing at a very fast pace. The global mainstream interest in SP highlights the strength, dynamism, and diversity of our community. Indeed, ICASSP is the home of cutting-edge research in many areas, including speech and language processing, audio and acoustic SP, machine learning for SP and communications, distributed optimization and information processing, graph SP, and image, video, and multidimensional SP. For the first time, ICASSP 2023 will host satellite workshops, which will foster cross-discipline exchanges of ideas and promote focused events in topics at the cutting edge of our field. We expect that these will become permanent features in future ICASSPs. ICIP is the premier forum for presenting technological advances and research results in the fields of theoretical, experimental,

and applied image and video processing, and it continues to attract high-quality research in these areas. It is increasingly becoming the conference venue of interest in research related to image and video deep learning methods.

With continued efforts to increase and strengthen industry participation at our conferences, industry involvement has progressively increased. At ICASSP 2022, an industry program was incorporated in the technical program. It included a full parallel industry track with a corresponding open call for participation, high-profile industry keynote speakers, industry expert sessions, industry workshops, and the traditional show-and-tell demonstrations. The SPS is now participating in IEEE DiscoveryPoint for Communications, which is a platform to meet the technical information needs of practicing product design engineers working in communications.

In 2020, the SPS established the Education Board to serve members' continuing education needs and promote SP education broadly. It also set strategic goals to boost educational and training offerings. ICASSP 2022 offered education-oriented 10-h courses, providing in-depth and multisided understanding of a topic and a final quiz to cap each course. Upon completion of each course, attendees were provided professional development certificates for training hours. These courses are now offered on demand for SPS members in the SPS Resource Center. The SPS plans to continue offering such courses in future ICASSPs and ICIPs.

For SPS students, a new member benefit is the SPS Scholarship Program, launched in 2023. The SPS is now awarding multiple scholarships up to a total of US\$7,000 for up to three years of consecutive support to students who have expressed interest and commitment to pursuing SP education and real-world career experiences. Students and graduate students from all 10 IEEE Regions are eligible for this program.

The Society has also provided initiatives to stimulate and grow entrepreneurship to enable SP-related discoveries to impact applications. The SPS is now offering an Entrepreneurship Forum in conjunction with ICASSP to promote entrepreneurship in the SP community by sharing entrepreneurship journeys, discussing challenges and opportunities in translating SP research into commercial applications, providing a forum for pitching, and, ultimately, training a new generation of SP entrepreneurs. The first Entrepreneurship Forum was held at ICASSP 2022 with great success; it included a pitching competition, where the SPS offered US\$10,000 in prizes.

Diversity, equity, and inclusion have remained a key focus of SPS efforts. The SPS recognizes the importance of diversity and inclusion and has several ongoing efforts to widen the pipeline of women and underrepresented minorities interested in signal processing. Initiatives include outreach programs geared to pre-college students and, in particular, female and underrepresented students. The SPS understands that the key to increasing diversity in SP is a more diverse faculty providing opportunities, mentors, and role models that inspire students for excellence. In that spirit, in 2020, the SPS established the Promoting Diversity in Signal Processing (PROGRESS) Workshop to help women and underrepresented minorities pursue academic positions in

SP. PROGRESS is offered to both SPS members and nonmembers every year, in conjunction with ICASSP and ICIP. In 2021, the SPS started offering Mentoring Experiences for Underrepresented Young Researchers, connecting them with established researchers in the field [5], [6]. Also in 2021, the SPS started planning K-12 outreach initiatives to increase the visibility of the Society and the SP discipline to young students worldwide by developing exciting impactful educational programs.

The SPS strives to create an environment in which women and underrepresented minorities members feel included and appreciated. It is encouraging to see that our efforts have been paying off; today, the BoG includes members from nine of the 10 IEEE Regions, and over half of the voting members are female. To enhance our commitment to diversity, the SPS has revised its governance documents, using gender-neutral language. ICASSP 2023 will provide a lactation room and nongender-specific bathrooms, and we plan to add those as permanent features to all our conferences and workshops.

On the ethics front, the SPS has formed a team of volunteers representing various TCs to develop recommendations for responsible research and the ethical use of technology. The team is focusing on guidelines for authors, encouraging them to consider not only the potential benefits of their research but also the potential negative societal impacts and to adopt measures to mitigate risk. It is also developing guidelines for promoting explainable machine learning and solutions with low computational and memory cost and ensuring that SP-enabled developments are compatible with human well-being.

So far in this decade, SPS membership has grown from 18,730 to 19,164 members and is expected to surpass 20,000 during the Society's 75th anniversary year.

The SPS has engaged IEEE at large, with its leaders assuming leadership positions within IEEE-level boards and committees. In the past 20 years, SPS volunteers have served almost uninterruptedly on the IEEE Board of Directors as directors of Division IX (at least eight directors), vice presidents of technical activities (four), vice presidents of educational activities (two), vice presidents of the IEEE Publications Services and Products Board (two), and presidents of IEEE (three). In these positions, they steered IEEE into financially sound operations, promoting a more diverse, equitable, and inclusive organization, adopting open access and open science, exploring new membership models, adopting an IEEE-wide mobility policy, and fostering new services for professionals in IEEE's areas of interest.

## The beyond

The Society envisions that open access publishing will continue to grow, conferences will continue to expand both physically and virtually, membership activities will increase, educational opportunities will expand, and diversity and ethics will remain pillars in all aspects of our future commitments to our members and the general public. SP is a key ingredient in many new technologies and products, and its strengths continue to be enhanced by evolving computer and communications capabilities and novel algorithms. From digital and statistical to distributed and graph SP, from radar and communications to speech, images,

and language technologies, from the physical world to the social networks to space technologies, SP professionals are data scientists, AI developers and practitioners, and machine learning specialists, and SPS is stepping up in these areas to meet their needs. Opportunities abound!

## The SPS staff

The SPS staff has always played an important role in maintaining the continuity of the Society. Its members are highly capable professionals who work harmoniously with SPS volunteers and have the knowledge and skills to turn ideas into reality.

The current composition of the SPS staff is as follows:

- **Richard Baseil:** executive director
- **Administration:**
  - **Theresa Argiropoulos:** director, operations
  - **Deborah Blazek:** administrator, committees and governance
  - **George Olekson:** chapter and operations associate
  - **Jessica Perry:** membership communications and experience specialist
  - **Jaqueline Rash:** administrator, membership program and events
- **Conferences:**
  - **Caroline Johnson:** senior manager, conference strategy and services
  - **Nicole Allen:** senior conference administrator
  - **Samantha Esposito:** conference administrator
- **Publications:**
  - **William Colacchio:** senior manager, publication and education strategy and services
  - **Rebecca Wollman:** publications administrator
  - **Michelle Demydenko:** society peer-review and education program administrator
  - **Nanette Januszkiewicz:** society peer-review and education program administrator
  - **Mikaela Langdon:** society peer-review and education program administrator
  - **Rupal Bhatt:** web administrator.

## Acknowledgment

The authors would like to thank Richard Baseil for his valuable input on this article.

## Authors

**Athina Petropulu** (athinap@rutgers.edu) received her Ph.D. degree in engineering. She is Distinguished professor in the Department of Electrical and Computer Engineering, Rutgers University, Rutgers, NJ 08854 USA. She is the 2022-2023 president of the IEEE Signal Processing Society and was the editor-in-chief of *IEEE Transactions on Signal Processing*. She was a recipient of the 2005 IEEE Signal Processing Magazine Best Paper Award, the 2021 Barry Carlton Award by the IEEE Aerospace and Electronic Systems Society, and the 2023 Stephen O. Rice Prize by the IEEE Communications Society. Her research interests include signal processing, communications, networking, radar signal processing, security, and spectrum sharing. She is a Fellow of IEEE.

**José M.F. Moura** (moura@ece.cmu.edu) received his D.Sc. degree in electrical engineering and computer science. He is the Philip L. and Marsha Dowd University Professor at Carnegie Mellon University, Pittsburgh, PA 15913 USA. He was the editor-in-chief of *IEEE Transactions on Signal Processing*, president of the IEEE Signal Processing Society (SPS), and 2019 IEEE president and chief executive officer. He received the SPS Claude Shannon-Harry Nyquist Technical Achievement Award and the SPS Norbert Wiener Society Award as well as the 2023 IEEE Jack S. Kilby Signal Processing Medal. His research interests include statistical, algebraic, distributed, and graph signal processing. He is a Fellow of IEEE, a member of the Academy of Sciences of Portugal, and a member of the U.S. National Academy of Engineering.

**Rabab Kreidieh Ward** (rababw@ece.ubc.ca) received her Ph.D. degree in electrical engineering. She is the IEEE vice president, education and a professor emeritus in the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada. She was the president of the IEEE Signal Processing Society (SPS) and an IEEE director. She is the recipient of the 2023 IEEE Fourier Award for Signal Processing and IEEE SPS Norbert Wiener Society Award. Her research interests include signal and image processing and their applications to cable TV, multimedia, medical imaging, infant cry signals, and brain computer interfaces. She is a Fellow of IEEE.

**Theresa Argiropoulos** (t.argiropoulos@ieee.org) is the director of operations for the IEEE Signal Processing Society.

## References

- [1] B. Gold and C. Rader, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ, USA: Lincoln Laboratory report, 1969 and Prentice-Hall, 1972.
- [2] F. Nebeker, *The IEEE Signal Processing Society: Fifty Years of Service 1948 to 1998*. New Brunswick, NJ, USA: IEEE History Center, 1998.
- [3] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. Hoboken, NJ, USA: Wiley, 1967.
- [4] G. M. McNally, "Digital audio in broadcasting," *IEEE ASSP Mag.*, vol. 2, no. 4, pp. 26–44, Oct. 1985, doi: 10.1109/MASSP.1985.1163754.
- [5] IEEE PROGRESS. [Online]. Available: <https://ieeeprogess.org/>
- [6] "Diversity, equity and inclusion," IEEE Signal Process. Soc., Piscataway, NJ, USA. [Online]. Available: <https://signalprocessingsociety.org/our-story/diversity-equity-and-inclusion>
- [7] A. I. Perez-Neira, F. Pereira, C. Regazzoni, and C. Johnson, "IEEE signal processing society flagship conferences over the past 10 years," *IEEE Signal Process. Mag.*, to be published.
- [8] J. S. Kilby, "Invention of the integrated circuit," *IEEE Trans. Electron Devices*, vol. 23, no. 7, pp. 648–654, Jul. 1976, doi: 10.1109/T-ED.1976.18467.
- [9] "Intel 4004 microprocessor," Intel Corp., Santa Clara, CA, USA, 2018. [Online]. Available: <https://www.intel.com/content/www/us/en/history/museum-story-of-intel-4004.html>
- [10] P. Norton, *Inside the IBM PC: Access to Advanced Features and Programming*. Bowie, MD, USA: R.J. Brady, 2014.
- [11] A. V. Oppenheim and R. W. Schaffer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.
- [12] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.
- [13] L. R. Rabiner and R. W. Schaffer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.
- [14] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," presented at the RADC Spectr. Estimation Workshop, Rome Air Development Center, Griffiss Air Force Base, New York, NY, USA, Oct. 1979, pp. 243–258.
- [15] G. Bienvenu and L. Kopp, "Principe de la goniometrie passive adaptative," in *Proc. 7<sup>eme</sup> Colloque sur le Traitement du Signal et ses Appl. (GRETSI)*, Nice, France, May 1979, pp. 106/1–106/10.





Rabab Kreidieh Ward 

# The Evolution of Women in Signal Processing and Science, Technology, Engineering, and Mathematics



©SHUTTERSTOCK.COM/TRIFF

When I began writing this 75th anniversary article celebrating women in signal processing (SP), I reread the 1998 editorial titled “Fifty Years of Signal Processing: 1948–1998” [1]. At that time, IEEE had more than 300,000 members in 150 nations, the world’s largest professional technical Society. Within the IEEE umbrella, there were 37 IEEE Societies and technical groups, and the IEEE Signal Processing Society (SPS) was the oldest among its many Societies.

The 50th anniversary piece was a celebration of the major players in SP and the historic growth of the SPS. It featured many important scientists in SP and the SPS, including numerous blurbs, quotes, and personal recollections from male leaders in the SPS. The first nod to a woman inventor doesn’t occur until the 1970s, mentioning Susan A. Webber in the field of subband coding breakthroughs. Readers have to wait until the 1980s section to see the face and profile of an SPS woman member: Delores Etter, the Society’s first woman president, in 1988. Beneath her smiling photo and blurb, the piece acknowledges that, “As in most areas of science and engineering, there were relatively few women in SP until the last one or two decades, when their numbers increased markedly.” It mentions Marie Dolan and Carol McGonegal (who served on the Digital Signal Processing (DSP) Technical Committee (TC) in the 1970s), and various other women members of the SPS Board of Governors (BOG), starting with Edith L.R. Corliss (1973–1975). Leah Jamieson joined the BOG in 1981, Maureen Quirk in 1986, and Fay Boudreaux-Bartels in 1989, followed in the 1990s by Marcia A. Bush, Candice Kamm, Quirk, Sarah Rajala, Sally Wood, and Jamieson, who began her two-year term as Society president in 1998 (see Figure 1).

Jamieson is the second and final woman given a profile entry in that 55-page anniversary celebration piece, followed by a tip of the hat to Quirk and SPS Executive Director Mercy Kowalczyk, who were involved with revising the Society Constitution and Bylaws in 1993.

I know that during those 50 years, and in the 25 years since, great strides have been made by women to close the gender gap in society, at the SPS, and in science, technology, engineering,

Digital Object Identifier 10.1109/MSP.2023.3236475  
Date of current version: 1 June 2023

and mathematics (STEM). I know from personal experience that we have accomplished many firsts, despite many roadblocks. As a young woman living in Beirut in the 1960s, I had the highest grades in the country, but I couldn't study engineering at the American University of Beirut, so I had to go to Egypt for my engineering education, where approximately 18% of students were women. Soon after, I became the first women member of the Lebanese Professional Engineering Society. Later I completed my Ph.D. in electrical engineering at the University of California at Berkeley, and I was only the second woman to earn a Ph.D. there, in 1972; the first was an Egyptian named Kawthar Zaki.

In 1970, women accounted for 38% of the U.S. workforce; 8% were in STEM fields and only 3% were in engineering [2]. Like so many women, I couldn't find a job in academia that acknowledged my expertise. I was a sessional lecturer for two years at the University of British Columbia (UBC), and then I went abroad, had children, and eventually became the first woman in the engineering faculty at University of Zimbabwe. Later, in the early 1980s, I became the first woman engineering professor in BC, which made me Canada's first woman holding a Ph.D. to become professor of electrical engineering, and later, in 1998, to become Fellow of the Royal Society of Canada. At that time, most women I knew in science or engineering in Canada were appointed on short terms as sessional lecturers, so the majority of my colleagues and students were male.

Gradually, over time, more young women chose engineering, and some became established leaders in their fields, including Lina Karam, who was appointed in 2020 as the dean of engineering at the Lebanese American University in Lebanon. I was appointed as director of the Institute for Computing, Information and Cognitive Systems at the UBC, and later as its Natural Sciences and Engineering Research Coordinator and Advisor at UBC's Vice President (VP) Research Office. Some of my work has been licensed to U.S. and Canadian industries and has resulted in many accolades. Most notable are the IEEE Signal Processing Society Norbert Wiener Society Award, in 2008, and the R.A. MacLachlan Award, the highest award of the Association of Professional Engineers in BC, emphasizing significant technical contributions and leadership to engineering "that characterize the profession at its best."

In 2020, I became an international member of the National Academy of Engineering. This year, I am the recipient of the 2023 IEEE Fourier Award for Signal Processing. Among my various awards, the dearest to my heart and the one that I feel I deserve most, is the highly competitive Killam Senior Award for Excellence in Mentoring, which I received in 2013.

But sadly, today many women still face many of the challenges that I encountered decades ago. According to the IEEE-USA's 2022 Annual Salary Survey, the gap for IEEE women members grew in 2021, by almost US\$6,000, with the proportion of IEEE women engineers remaining at under 10%, the same number for the past decade [3]. The news from other data collection sources is similarly distressing. "As the demand for STEM talent increases, women's share of those jobs remain relatively flat," according to the 2020 Women in Stem Workforce Index, which found that in the United States, women hold only one in four STEM jobs [4]. Other troubling aspects of the Index include that the largest STEM occupation, computers and math, a field that has exploded in growth in the past decades, women's share of jobs actually decreased from 44% in 1990 to 27% in 2018, and women made up only 15% of the engineering and surveying workforce, the lowest representation among STEM workers. The STEM pay gap actually increased by 3% between 2010 and 2015 [5] and has flatlined since, at 27% in computers and math, 16% in engineering, and 26% in management positions [4], with women consistently underrepresented at the executive, high-level leadership level [2].

The situation is even more grim for U.S. women of color (WOC) in STEM [6]: 13% of STEM bachelor's degrees, 12% of master's degrees, 7% of doctorate degrees, and they represent only 4.8% of the workforce. Among science and engineering jobs, the numbers are even worse: 2.3% for Hispanic/Latina women, 2.5% for Black women, and 0.07% for indigent women.

I will provide more big-picture numbers later and also gender-specific statistics from the IEEE and the SPS, but first, on this 75th anniversary of the SPS, I want to celebrate and feature some of the many women SPS members who have worked so very hard to grow our Society, our research fields, and our world. I want you to hear their personal anecdotes, struggles, and victories. I want you to learn about the positive work



**FIGURE 1.** Women Presidents of SPS, from left: Delores Etter (1988–1989), Leah Jamieson (1998–1999), Rabab K. Ward (2016–2017), Athina Petropulu (2022–2023), and President-Elect Min Wu (2022–2023).

they're doing to encourage and support the next generation of brilliant women so that girls and young women from all walks of life, ethnicities, and cultural backgrounds have role models and heroes whose footprints they can follow, whose strides will encourage the next generations of women in SP and STEM to take great leaps and blaze their own trails in this world.

### Women leaders and innovators at the SPS

Since I became a member of the Society in 1988, I've had the pleasure of meeting many fantastic women in STEM, including many of the women mentioned in the 50th anniversary publication. These women broke down gender barriers at the SPS level in academia and industry in all corners of the globe.

My involvement with the IEEE and the SPS has been crucial to my career success. To me, it was more than a professional home. I was exposed to new technical topics, and I have learned so much from my colleagues about strategic planning, creating common goals, embracing change, forging effective leadership and management, and the importance of rewards. Many of these colleagues are incredible women leaders.

Let's start with Etter, the first woman president of the SPS. She received a Ph.D. in electrical engineering from the University of New Mexico in 1979 and became a faculty member in the Department of Electrical and Computer Engineering (ECE) with a focus on speech recognition, software engineering, and adaptive SP [7]. She also worked at Sandia National Laboratories, working in seismic SP. In 1998, she became the Deputy Under Secretary of Defense for Science and Technology, overseeing the American Defense Science and Technology Program. She also ran the Defense Modeling and Simulation Office, the Department of Defense's high-energy laser research program, and was the principle U.S. representative at the North Atlantic Treaty Organization's Research and Technology Board.

In the 2000s, she joined the faculty of the U.S. Naval Academy, becoming the first Office of Naval Research Distinguished Chair in Science and Technology. She was also elected member of the National Academy of Engineering and was Assistant Secretary of the Navy for Research Development and Acquisitions, overseeing the purchases of military machinery and IT. The prestigious Dr. Delores M. Etter Top Scientists and Engineers Award is named for her.

I reached out to Etter to talk about the history of women in SP and her memories of those early years. She presented her first paper at an IEEE Asilomar Conference in 1978, and in 1979, she presented another paper at ICASSP. "I remember standing in the hall with the conference guide, trying to decide which of the parallel sessions to attend," she recalls. "I was wearing a brown linen suit with a white blouse with lace on the collar. I had my name tag clearly visible on the collar of my jacket to show my name and university affiliation. While I was standing there, another attendee walked up to me and handed me his coat. I took it, and then looked around to see why he handed it to me. Down the hall was a sign for coat check. I am sure that I was frowning as I handed him back his coat and pointed down the hall!"

Etter says that in those days there were few women attendees at SP conferences and SPS governance. "I wanted to help

provide more visibility to the other women," she says. Etter began volunteering in conference activities and "quickly realized that the SPS decisions were made by the Administrative Committee," which included no women, and "no members west of the Mississippi." That would change in 1983, thanks to Etter, who campaigned for a position on the BOG. "I was able to get on the ballot and get addresses for SPS members in California," she recalls. "I sent them a letter asking for their vote so that there would be broader representation geographically." Etter was elected to the BOG, and she began a decade of significant involvement with the SPS, including chairing many key committees, and as editor-in-chief (EIC) of *IEEE Signal Processing Society Magazine* (1986–1987) and *IEEE Transactions on Signal Processing (TSP)* (1993–1995.)

Jamieson is another trailblazer in SP and the SPS. After receiving her Ph.D. in electrical engineering and computer science (CS) at Princeton, she became a distinguished professor at Purdue and later dean of engineering, specializing in speech processing and parallel SP. In 2007, she became president of IEEE, and chair of both the Purdue and the National Global Women in Tech organizations.

Jamieson got her start at the SPS in the 1980s, volunteering. "There is no question that my experiences in the SPS contributed to many of my future successes," she acknowledges. "Several of my fondest memories as a member of the SPS are the people: new friendships, new colleagues, opportunities to work with some truly amazing people through my years on Acoustic Speech and Signal Processing (ASSP)/SPS committees and boards, and the truly wonderful SPS staff. My memorable experiences on the Board of Governors and as president included shoe shopping with SPS Executive Director Mercy Kowalczyk, something she said she didn't get to do with her other presidents. Colleagues in the Society offered me encouragement over many, many years."

Some of these colleagues were men, including Al Oppenheim and SPS Presidents Tariq Durrani (1994–1995) and Don Johnson (1996–1997). Jamieson went on to have numerous IEEE posts, including 2003 IEEE VP Technical Activities, 2005 IEEE VP Publications, 2007 IEEE president and CEO, and 2012–2016 president of the IEEE Foundation.

"I had my first experience with strategic planning when I was on the Board of Governors," she says. "As president of the IEEE Foundation, we developed a five-year 'Strategy for the Future.' Fostering collaboration became a central theme of much of my work at IEEE, and as dean of engineering."

As SPS president during the 50th anniversary of the SPS, Jamieson says, "I think we would have been hard-pressed to do an article about women in 1998." At that time, the climate for women in academia was described as "chilly" according to a 1996 book, *The Chilly Classroom Climate: A Guide to Improve the Education of Women*, coauthored by Bernice Resnick Sandler [8]. Known as the *godmother* of the 1972 U.S. educational amendment Title IX legislation prohibiting discrimination based on race, color, religion, gender, and national origin [9], her research on gender bias in academia documents women students' many hurdles, from hostility and

denigration, to the various ways that professors overlooked, ignored, and dismissed women students, from lack of eye contact and dialogue, to patronizing, simplistic responses to their questions or comments [10].

As the years passed, and research became increasingly collaborative, interdisciplinary and global, so did education research, with increased emphasis on teamwork, ethics training, and community outreach, including the IEEE program Engineering Projects in Community Service (EPICS), which Jamieson cofounded and directed. EPICS increased diversity, including that 33% of CS EPIC students were women, compared to only 11.5% nationally. These programs underline the need to connect girls and women in STEM to real-world, community-based issues and needs that will benefit the world. As IEEE grew, so too did its publishing output and global readership, strategic planning, global offices, and key messaging, including its 2010 core purpose of “Advancing Technology for Humanity.”

Yet despite many efforts, women engineers continue to experience a higher attrition rate in the workforce, lower pay scale, and many tensions between work and personal life responsibilities. Campus life is also still chilly. “Engineering students still tell the same ‘boys club’ stories,” Jamieson noted in her keynote talk at the 2021 ICASSP conference [11]: “Male lab partners who assume the woman will take notes while he does the experiments; women leading design teams whose members won’t pay attention to their leadership; unwanted sexual advances; faculty who shrug off concerns of women who come to them for help in dealing with these issues.”

Long-time SPS Member Quirk didn’t have such negative experiences at the professional level. “I found that male engineers were very supportive of women,” she says. “They never belittled researchers because they were women. Engineers are much more interested in people getting the answer rather than any attribute a person might have.” Quirk provided me with an amusing anecdote from 1984, the year she joined the ASSP Conference Board Committee, the first woman on that committee. At that time, she was working at the Jet Propulsion Laboratory in Pasadena, and attended a DSP workshop. “Tom Quatieri gave a talk about sinusoidal representation for speech,” she recalls of the event that included few women researchers. “He mentioned that it worked far better on women. I started clapping, then there was silence. After a moment, everyone started to laugh and clap. Later when I read his paper, he mentioned that the signal speech reconstruction method was ‘pronounced for low-pitched speakers.’” Two years later, Quirk was appointed SPS secretary until 1991; from 1993 until 1996, she was treasurer, then conference VP from 1997 until 2000.

Wood, another key woman member of the SPS starting in the 1980s, underlines the importance of networking and mentoring opportunities. In those days, “there were not many women in SPS, and we all knew each other,” she says. “When I got my B.S. degree, I was told that, in the United States, only 2% of practicing engineers were women. At the SPS, I benefited from informal mentoring from a number of more senior SPS members. As a Society, I think SPS serves its members well by having a broad range of professional activities and

venues for engagement. SPS is an intellectually vibrant and collegial community, which attracts so many women.”

As a professor of ECE, and current associate dean for graduate studies at the Santa Clara University School of Engineering, Wood became an IEEE Distinguished Lecturer (DL) in 2003. She says that her proudest moments include serving as SPS VP of Awards, and becoming an IEEE Fellow.

Another important factor in the growing number of women in our field is that the SPS and the IEEE grew its membership at the global level, attracting many new members from around the world, including Asia, the Middle East, and Europe. I have met many incredible SPS women colleagues from all parts of the world, and I can only mention some of them here whom I have served with on different SPS committees: Urbashi Mitra, Sheila Hemami, Yan Sun, Tulay Adali, Bhuvana Ramabhadran and Behnaz Ghoraani from the United States, Deepa Kundur, Z. Jane Wang, Octavia Dobre, and Mahsa Pourazad from Canada, Roxana Saint-Nom from Argentina, Hong (Vicky) Zhao from China, Helen Meng and Pascal Fung from Hong Kong, Anubha Gupta from India, and Maria Sabrina Greco, Christine Guillemot, Isabel Trancoso, and Josiane Zerubia from Europe.

Zerubia is the first woman from outside North America whom I have served with on the SPS BOG. She has been active member of the SPS for more than 25 years. As director of research at Center INRIA since 1989, she has headed many labs and groups, including Scene Analysis and Symbolic Image Processing, Variational and Stochastic Models for Image Processing, Models of spatio-temporal structure for high-resolution image processing, and AI and Remote Sensing on board for the New Space. A Fellow since 2003, Zerubia acknowledges, “It is not always easy to be a successful woman in SP and scientific fields. Male and female colleagues could try to push you down. The only way to survive is to work harder and to always get better results.” Zerubia also credits the work of her male counterparts who “strongly supported” women members, including former SPS Presidents Jose Moura and Ali Sayed. “My vision for the future for women in SP is that we need to encourage young ladies to choose to learn math and physics at a young age [and give them opportunities] to learn SP at university. We also need role models in SP. Mine are Rabab Ward and Jelena Kovačević.”

Kovačević is a specialist in wavelet theory and biomedical imaging and a long-time advocate of women in STEM [12]. She grew up in the former Yugoslavia and credits her parents for putting her on the path to a career in math, providing her “infinite confidence” that she could do anything she wanted in life. She attended Columbia University and was one of only a handful of women Ph.D. students in the electrical engineering department, from where she graduated in 1991. That decade, she worked at Bell Labs in New Jersey and cofounded xWaveforms. In both academia and industry, “I did hear an occasional, ‘She got the job because she is a woman,’ comment and ignored it,” she says. But once she became the department head of ECE at Carnegie Mellon in 2014, she learned that only 21% of undergrads in her department were women. She listened to “heart-wrenching” stories from women students about the

hardships they faced simply because of their gender. “I educated myself. I attended a leadership academy for women at Carnegie Mellon. I read articles. I discovered that gender bias in STEM fields abounds. Even though we know that diversity, in gender and race, makes us smarter, better people.” She took action at the departmental level. “We completely revamped our faculty-hiring process, educated faculty on unconscious bias, had broad and inclusive search committees, and published our search procedures,” she says. “We also hosted prominent career-building workshops and events like Rising Stars in EECS (electrical engineering computer science), and Judith Resnik Year of Women in ECE.” Within three years, the number of women undergraduate students grew to 27%. The department included five women junior faculty members, growing the number of women staff members to 18%. In 2018, Kovačević became dean of New York University’s Tandon School of Engineering, the first woman to head the school since it was founded in 1854.

Kovačević has been an active member of the SPS for more than 30 years, former EIC of *IEEE Transactions in Image Processing*, former member-at-large of the BOG, and winner of an IEEE SPS Technical Achievement Award, which she counts among her proudest career moments. “We still have a lot of work to do on campus and after graduation,” she says. “Many women go to Silicon Valley, which isn’t welcoming to women. We need to make this a wider conversation: that gender equality in STEM is also a social issue that everyone needs to change, so that all parents; educators; and employers; all the elders in our culture, advocate for equality so that all children can follow their passions to have opportunities to fail and learn and succeed.”

As a long-time professor, I appreciate the importance of growing the number of professional women faculty members in academia. And many other women SPS members have the same goal, including our current SPS President Athina Petropulu.

“I have to admit that in the first few years I felt isolated at SPS conferences,” says the 1991 Ph.D. graduate of ECE from Northeastern University. “After I got involved as a volunteer (through a TC membership first), I started having a network, which made a big difference. Women in SP (WISP) is a great opportunity to feel part of a community. I am very proud to have been EIC of *IEEE TSP* and SPS VP Conferences. While there has been progress, women still do not get as many nominations for awards and recognition [(DL and distinguished industry speakers (DISs)]. Women represent untapped capital. If we make them feel included and comfortable, they will unfold their talents and the field of SP will be so much richer.”

Petropulu is a distinguished professor at Rutgers ECE, and she’s active on many levels at IEEE and the SPS, including as IEEE Technical Activities Board member, and a former BOG member-at-large at the SPS. She has won numerous awards, including the Barry Carlton Best Paper Award at the IEEE Aerospace and Electronic Systems Society, where she served as a DL in 2019. While president-elect at the SPS, Petropulu received approval for a new faculty diversity-building workshop she conceived and named *Promoting Diversity in Signal Processing (PROGRESS)* [13]. This workshop was inspired

by iRedefine, a program she spearheaded, as president of the Electrical and Computer Engineering Department Head Association. “The idea is to motivate and prepare women and underrepresented minorities to consider academia,” she says. Between 2017 and 2018, iRedefine helped 36.6% of the 66 student participants get academic jobs. “We all see that there are very few women faculty,” says Petropulu. “China has over 50% female students, but still very few faculty. Who is going to inspire those women to become leaders when they hit the job market? Companies recognize the value of diversity and have the means (high salaries) to lure women. But academia does not offer high salaries. How can it compete with industry for the best? At PROGRESS, we provide information on how to put together application materials, CVs (curriculum vitae), give mock interviews, and also professional training on how to negotiate. Since PROGRESS is for all the world, we have panels focusing on different countries.”

PROGRESS attracted 202 students at its start in conjunction with ICIP 2020. It’s now an ongoing SPS workshop at ICASSP and ICIP conferences and some mentoring teleconferences. Panel members represented a diverse group of global academic leaders from Beirut to Bangalore, to Buenos Aires and Hong Kong. The exit surveys for their first workshop showed that interest in pursuing professional academia more than doubled [14].

Piya Pal, one of our younger senior members of the SPS, agrees that mentorship is a key factor for women in STEM. “SPS has a lot of activities planned during conferences, which are very encouraging for young people,” she says. “But I think the real work happens behind the scenes through forming personal relationships between a mentor and a mentee.” Born in Calcutta, Pal did her Ph.D. at the California Institute of Technology (Caltech) in 2013 and is now an assistant professor at the University of California, San Diego’s Jacobs School of Engineering. “Encouragement from the [SPS] community and the visibility of my work at an early stage played an important role for my career development,” she acknowledges. Student paper awards were essential for her early career; her doctoral thesis was awarded the 2014 Charles and Ellen Wilts Prize for Outstanding Thesis in Electrical Engineering at Caltech. “I was also honored to receive the Early Career Technical Achievement Award from the SPS and the U.S. PECASE Award for my works on sparse sampling techniques,” she says.

On the challenges for women in STEM, Pal acknowledges that “people (both men and women) can jump to quick conclusions (which are often wrong) about another person’s work, and this is usually due to lack of proper technical understanding, or sometimes even due to deep-rooted biases. When faced with these situations, I have always tried my best to fight back purely on a technical basis and not let my personal emotions get in the way toward establishing the scientific truth.”

## **Women in the SPS: Challenges and opportunities**

Many key women and men at the SPS have spent decades working very hard to open doors for women in SP. Yet many recent stats show that progress for women in STEM has plateaued over the past decade, particularly in leadership in academia

and industry, and for WOC. I would like to now turn the spotlight on IEEE and the SPS to check our progress in the past 25 years and discuss new hurdles and opportunities.

When I became president-elect of the SPS in 2014, my priorities included growing the number of women involved in SP, inspiring our young members to get involved with the SPS, and ultimately seek out fruitful careers in SP and all STEM fields. Since the start of my involvement in the SPS in 1998, I have found it to be very supportive of diversity, and its presidents take special care and consideration of the various methods for cultivating and advancing women's participation in our Society. It has been an honor to continue these endeavors, both as president and a senior member of the SPS. In the past decade, I have had many lively discussions about this topic with Past Presidents Mostafa Kaveh, Jose Moura, Ray Liu, Alex Acero, Ali Sayed, and Ahmed Tewfik. As the SPS shared my goal to increase its women members, we looked at ways to enhance women members' experiences at the SPS. Senior women members know from personal experience that diversity, networking, and mentorships are crucial to both personal success

and the vitality of any organization, and we wanted to grow these opportunities. The WISP Subcommittee was approved by the SPS BOG in May 2014, with the leadership of Kostas Plataniotis, the SPS VP Membership at that time. That year, we started holding the WISP luncheon at all major SPS conferences, including ICASSP, ICIP, and GlobalSip. These luncheons, which also welcomed male SPS members, featured women speakers discussing ways to build and advance women's careers so that everyone can benefit and thrive. These luncheons were well attended, especially by newcomers. I loved attending them, seeing colleagues, meeting new people, and participating in discussions with speakers and panelists.

When I was president in 2017, and under the support and direction of Nikos Sidiropoulos, the WISP Subcommittee was elevated to the committee level, directly reporting to the SPS Membership Board. (The chairs of WISP thus far include Antonia Papandreou-Suppappola, Namrata Vaswani, and currently, Celia Shahnaz.) Initially, the WISP Committee focused on the luncheons, and they have since become active in hosting other events, including International Women's Day.

## IEEE Signal Processing Society Member Quotes

It was not always easy to break into the field and to make friends. In the beginning, it could be lonely as the only woman on a committee or in the room. Luckily, that problem does not exist anymore as more women are joining the IEEE Signal Processing Society (SPS). It brings diversity into the Society and generally, into the field. I have always felt supported during the Women in Signal Processing (WISP) meetings at the major conferences. I met fantastic colleagues, shared experiences, and made some of my best friends. Being able to mentor younger colleagues was always an inspiration. But there are still not enough awards given to women who would richly deserve them. New awards can be created. Similarly for keynotes, there should be a push to have more women speaking. My vision for the future for women in signal processing? Equity. —*Sabine Süssstrunk*, IEEE Fellow; head of the Image and Visual Representation Lab, and director of the Digital Humanities Institute (2015–2020) at Ecole polytechnique fédérale de Lausanne; member of the Executive Committee of Swiss National Science Foundation.

Some of my proudest moments are my paper awards, the successful conferences I organized, and the pride of having improved the *IEEE Signal Processing Letters* performance. I can see a new generation of thought leaders and pioneers. —*Anna Scaglione*, IEEE Fellow; Cornell University, member of the SPS Board of Governors BOG (2011–2013), editor-in-chief of *IEEE Signal Processing Letters* (2012–2013).

Just around the time I got my Ph.D., my advisor decided to leave his tenured faculty position and left academia. I

was an academic orphan. My Ph.D. was in adaptive filtering and I was thinking that neural networks with a statistical connection could be a fruitful research direction. Hence, I decided to make it to the NNSP workshop, which in 1993 was held in a small town in Greece. That trip indeed helped shape my career. I found a vibrant and friendly community with NNSP (now MLSP) TC, which I also chaired (2003–2005 and 2021–2013). The interactions within this and then the broader SPS community provided important support. That is why, when serving as Vice President (VP) Technical Directions, I worked on multiple initiatives to reach out to young professionals to help them connect easily with our technical activities within the SPS, and find a community that nurtures them. —*Tülay Adali*, IEEE Fellow; distinguished university professor, University of Maryland Baltimore County; SPS VP Technical Directions (2019–2021); chair of the IEEE Brain Initiative.

Luckily, I have had great mentorship through the years that supported my growth. SPS has dedicated women in signal processing committees and arranges dedicated events at major conferences to help female students, which are commendable. However, it is often hard to measure if such events have led to sustained impacts on women's careers. It is wonderful to see many of our senior women are taking up leadership positions in SPS, which for sure will inspire more to follow, leading to broader and more diverse participation in our community across all races, genders, and demographic areas. —*Yuejie Chi*, Carnegie Mellon University; IEEE SPS Distinguished Lecturer (2022–2023); IEEE Information Theory Society Goldsmith Lecturer (2021).

A healthy portion of the time is devoted to networking, allowing me to meet many new women members from all parts of the world, and learn about their country-specific challenges and situations. Many women wanted to know how they could start volunteering at the SPS. A woman from China was surprised that in North America, we had so many initiatives to encourage women to go into STEM, saying that in her country there are now many women engineering students and professional engineers, and that women are encouraged to do what they like to do. Women in some Arab countries said that approximately half of engineering students are women, although they face challenges getting employment in some fields, and so, some women start their own businesses or work in a related profession. A Canadian resident from Mexico said that women students in Mexico are very optimistic about getting into engineering programs at university, but they have fewer career prospects as male engineers tend to prefer hiring male graduates. I learned that there were many variations from one region to the next, even among neighboring countries.

The SPS has also spearheaded an informal event for senior women faculty members that began in 2018 at ICASSP, thanks to Yonina Eldar. “It is important to have a more intimate forum than the WISP luncheons, where senior colleagues can network and also discuss issues having to do with more advanced stages of our careers,” says Eldar. “It’s an informal women’s gathering with the goal of celebrating each other’s success and enjoying each other’s company, and of course creating a supportive network.” The event has since become an ICASSP tradition.

The SPS still holds the networking luncheons at ICASSP/ICIP, although during the peak of the COVID-19 pandemic, all events were held virtually. In general, the pandemic has caused many setbacks in academia, when women in STEM identified loss of mentoring and networking as a significant issue, with a years-long impact on their educations and careers [15]. It shows just how important these networking opportunities are for women in our field.

Recent SPS statistics provide some evidence that our various programs have benefited women in SP, particularly at the student level. But there’s certainly much room for improvement.

I have attended the Women in Signal Processing events at ICASSP, whenever possible, both as a student and later as a professor. I can meet with my friends and professors from all over the world, make new contacts within the community, and discuss relevant and timely issues that women in STEM (science, technology, engineering, and math) fields may face. However, it is not usually clear how to get involved into SPS activities other than attending those events. SPS is a large community, and it is easy to feel lost. —*Tanaya Guha*, University of Glasgow; honorary associate professor, University of Warwick; chair, IEEE Women in Engineering (WIE) Vancouver Section, IEEE Multimedia Systems and Applications TC (2021–2024); presently, Editorial Board member, *Nature Scientific Reports*.

As a daughter and a wife, I always face the expectations to take care of my aging parents and my own family. I moved back and forth to balance my research career and family duties. Now I am self-employed to do independent research in my field while taking care of my families. I hope I can have free access to IEEE e-library from home without having to physically visit a university in the future, which is important for a self-employed woman researcher. Women in signal processing can advance technologies to improve the quality of human life if they are inventive and persevering. —*Huiqun Deng*, IEEE Senior Member; self-employed.

The lack of women in most forums pushes you toward a male mentality. To overcome that, I try to be present in different representative bodies so I can influence and invite women on board, or help them to gain visibility. WISP is

the best tool and must keep on growing in members and visibility in many different activities, which should be organized for women and men. —*Ana Perez-Neira*, IEEE Fellow; Universitat Politècnica de Catalunya; SPS VP Conferences (2021–2023), general chair ICASSP 2020 (with more than 15,000 virtual attendees).

You always need to do much more than a man to be recognized. I did more. Women still need to fight to break the glass ceiling, but their competencies are better recognized and this also encourages young women to pursue a career in SP. The SPS is an international organization, so the recognition is more objective than in local and small professional committees. The international recognition helps in supporting our local recognition and professional career promotion. My fondest experience at SPS is chairing ICIP 2014 in Paris. —*Beatrice Pesquet*, IEEE Fellow; Télécom ParisTech; SPS BOG member (2017–2019), chair of SPS Image, Video, and Multidimensional Signal Processing (IVMSP) technical committee, and of SPS International Conference on Image Processing (ICIP) IDSP TCs.

I had to overcome shyness early in my career due to few women at conferences. Women’s meetings at conferences are motivating. More advertising and events with successful women in the field would be interesting as well as financial support for young and promising researchers, and help from more experienced colleagues. —*Mariane Petraglia*, IEEE Senior Member; Federal University of Rio de Janeiro, Professional Trajectory Award Recipient; IEEE WIE Unicamp, Brazil (2015).

An ad hoc committee chaired by Mari Ostendorf was tasked with collecting statistics and information about women IEEE Members in the field of SP and how they fare in awards and in leadership roles. Besides Ostendorf, the ad hoc committee also included Petropulu, Beatrice Pesquet-Popescu, and Eve Riskin.

In September 2016, I received their report. It was an illuminating read, highlighting that women made up only 9.4% of SPS members and 10.6% of IEEE Members, and although 10% of SPS fellows were women, reflecting their substantial technical achievements, since 1990, only 2.2% of SPS major (nonservice) awards were earned by women.

“Our primary findings are that women in the SPS are grossly underrepresented in technical achievement-related awards (Society, technical achievement, education) relative to their percentage representation in the Society, which is itself low relative to representation in IEEE overall ... the trends are consistent with those for the major IEEE awards, where the numbers are significant. The representation of women among plenary speakers at the SPS flagship conferences also appears to be unreasonably low ...” [16]

The committee found that the single biggest issue was related to the nomination process. Women members were nominated at much lower level than their male counterparts. This

## Women in Science, Technology, Engineering, and Mathematics by the Numbers

In 1970, women accounted for 38% of the U.S. workforce, but only 8% of science, technology, engineering, and mathematics (STEM) occupations, and 3% of engineering jobs. By 2019, the proportion of women had reached 48% of the U.S. workforce and 27% of the STEM workforce. Yet in the computer and engineering fields, the largest among STEM occupations at 80%, women represented only about a quarter of the computer workforce and 15% of engineering occupations [2].

A 2020 global snapshot of women in engineering (WIE) jobs found that women’s representation ranged from 11% in Brazil, to 14% in the United Kingdom, to 20% in India [21]. In the European Union, women account for only 32% of the high-tech workforce [22].

### *Gender pay gap*

Women typically earn less than their male counterparts in all fields, including STEM. Among the 70 STEM occupations in the U.S. Census Bureau, women earned more than men in only one STEM field (computer network architects) [2].

According to 2020 data on workers aged 35–44, women in the United States earned 30% less than men, and that pay gap increased with age. In STEM occupations, in 2019, women earned US\$0.816 for every dollar that men earned [6]. In the United Kingdom, the pay gap for women engineers is 11%, and by the age of 35, 57% drop out of the profession despite the fact that the country has a shortage in the field [23].

### *The leaky pipeline*

The pipeline starts leaking during childhood [6]. A 2019 U.S. study asked school kids to draw a scientist. Only 28% depicted a woman scientist. The majority of boys drew male characters, and girls did the same twice as often as the girls who depicted a woman scientist. Another 2019 U.S. meta-analysis of gender stereotypes in science [24] found that although 70% of girls aged six drew a woman, only 25% of girls aged 16 chose to depict a woman. When students reach middle school, boys are more than twice as likely as girls to

choose science or engineering careers, according to 2019 research.

Almost 50% of U.S. women in science and engineering majors switch to non-STEM faculties, compared to 33% of men. Fifty-seven percent of Bachelor of Arts (BA) recipients are women, but only 39% are STEM degrees, with the lion’s share in biological sciences, math and statistics, and physical sciences. Only 19% of BAs given to women were in computer sciences, and 21% in engineering [National Science Foundation (NSF) 2017–2019].

Additionally, U.S. women in STEM receive only 44.3% of master’s degrees and 41% of doctorate degrees, and 36% are postdoctoral fellows. Yet only 29% are employed in STEM fields. In engineering, only 13% of working engineers are women, earning 10% less than male counterparts [25], and as many as one in four of them will quit this profession after the age of 30.

The situation is much more grim for undergraduate U.S. women of color (WOC) in STEM, with 5% Asian, 5% Hispanic/Latina, 3% Black, and 0.16% identifying as American Indigenous. WOC represent roughly 17% of undergraduates, but only 9% are in STEM. WOC also receive only 12% of master’s degrees and 7% of doctorate degrees and make up only 5% of the STEM workforce [6].

### *Academia*

As of 2019, women are only 34.5% of faculty at academic institutions, and fewer than 3.5% are Hispanic, Black, or Indigenous. Twenty-eight percent of tenured STEM faculty are women, and less than 3% are Hispanic, Black, or Indigenous.

### *Career crunch*

According to 2019 U.S. NSF statistics, women represent 52% of the college-educated workforce, but only 29% of workers in science and engineering. In computer sciences, it’s 25%, and in engineering, we were only 16% of the workforce. In general, the disparity in income for STEM occupations is 16%, with the highest gender wage gaps among health care, physical scientists, and computer occupations [26]. As mentioned previously, the situation with



is a crucial aspect of career success: when women were nominated, the success rate almost doubled.

This research suggested that women have to make much bigger strides than men in the same field, which is consistent with other gender-specific literature in this field. Unconscious bias is a culprit, which has repeatedly been linked to the gender divide. Research in academia has found that a CV with a male's name gets a higher rating [17], and in academic fellowship applications, women with competence matched to their male counterparts, such as publication volume and impact, were given significantly lower scores [18]. Even recommendation letters for women medical faculty members resulted in lower reviews [19].

WOC is even worse. In the science and engineering workforce, Hispanic/Latinas, Black women, and Indigenous women count for only 2.3, 2.5, and 0.07%, respectively.

#### *Leadership*

Among U.S. government labs and research centers, 86% of the directors are white men while only 5% are women, and no WOC are represented at the director level. Only 26% of STEM-related leadership positions are held by women, including 3% of WOC. Between 2013 and 2019, women counted for only 8% of CEOs at biotechnology and initial public offering companies [6].

#### *Science academies*

Globally, women represent 33% of researchers, but only 12% of members of national science academies [27]. A 2021 Gender Insight report found that women memberships in National Academies included highs of 25% in Mexico and Canada, 19% in Malaysia and the United States, 15% in Brazil, 11% in Singapore, 10% in the United Kingdom, and 9% in India [21].

#### *R&D*

According to the United Nations Educational, Scientific and Cultural Organization, Central Asia has the highest number of women in R&D at 48.5%, followed by 45.8% in Latin America and the Caribbean, and 40.9% in Arab States [21].

#### *And the prize goes to*

Between 1901 and 2019 there were 616 Nobel Laureates in Physics, Science and Medicine. Only 19 of these prize winners were women. According to one study of National Institutes of Health funding between 2006 and 2017, women as first-time principal investigators received US\$40,000 less than male counterparts [6].

#### *Blatant discrimination*

According to a 2018 National Academies of Sciences, Engineering and Medicine survey, 50% of women in STEM academia experience sexual harassment. Another 2018 study found that half of women in STEM jobs experienced discrimination, 9% higher than their non-STEM

The ad hoc committee report acknowledges that, "People tend to hire others who have similar backgrounds to theirs, and the same trends seem to hold in SPS nominations for awards and invited talks. Recognizing unconscious biases is a critical step to reducing their impact on judgments. In addition, the reality of these biases means that boards need to be proactive about building a diverse candidate pool, in nominations for awards but also for lecturers, TCs, and board members."

Although IEEE and the SPS had various policies in place to provide gender balance among editors and TCs, the committee found no similar policies for nominees. They underlined the need for specific methods to bridge the gender gap: in leadership

counterparts. A whopping 70% of women in STEM report that they are routinely the target of biases and microaggressions related to their merits and competence. Even more chilling, 90% of STEM workers that do report sexual misconduct experience some form of retaliation [6].

#### *COVID-19: A disturbing new normal*

A 2021 report by the U.S. National Academies of Science, Engineering and Medicine found that women in STEM "face a myriad of systemic inequities" and "disproportionate hardships," suggesting "that the disruptions caused by the COVID-19 pandemic endangered the engagement, experience, and retention of women in academic STEM, and may roll back some of the achievement gains made by women in the academy to date" [28]. These hardships include loss of work-life boundaries; reduced productivity; isolation from networks, communities, and mentorships; increased issues with setting work-life boundaries, due in part to home childcare responsibilities; and psychological issues, ranging from burnout and sleep problems, to anxiety and depression. The report found that these various pandemic-related issues have been more pronounced for WOC [29].

#### *The benefits of closing the gender gap*

The European Institute for Gender Equality found that decreasing the gender gap in STEM fields could result in more than one million jobs, grow gross domestic product of the European Union by up to €820 billion by 2050, and potentially close the gender wage gap [30].

In other research about healthy workplace dynamics, research has consistently found that workers, and their organizations, thrive in environments that provide workers with three basic needs: autonomy, competence, and interconnectedness [31]. Psychological safety is another key factor that breeds inclusiveness, trust, and mutual respect, particularly when provided by leaders and executives. Google's Project Aristotle crunched the numbers among its teams, looking at numerous factors, and finding that psychological safety was the one key factor for successful teams [32].



Women SPS EICs	4	2	3	4	0	1	0	1	0	0	0	0	0	0
Total SPS EICs in time frame	16	17	11	12	4	7	3	2	1	2	1	3	1	2
Percentage of women SPS EICs	25%	11.76%	27.27%	33.33%	0%	14.29%	0%	50%	0%	0%	0%	0%	0%	0%
Women SPS TC chairs	4	5	4	5	1	1	0	0	0	0	—	—	—	—
Total SPS TC chairs in time frame	34	31	31	26	22	19	16	14	9	5	—	—	—	—
Percentage of women SPS TC chairs	11.76%	16.13%	12.9%	19.23%	4.55%	5.26%	0%	0%	0%	0%	—	—	—	—
Women SPS DLs	5	6	2	2	0	1	1	0	—	—	—	—	—	—
Total SPS DLs	26	26	25	28	28	18	14	3	—	—	—	—	—	—
Percentage of women SPS DLs	19.23%	23.08%	8%	7.14%	0%	5.56%	7.14%	0%	—	—	—	—	—	—
Women SPS DISs *	1	—	—	—	—	—	—	—	—	—	—	—	—	—
Total SPS DISs	20	—	—	—	—	—	—	—	—	—	—	—	—	—
Percentage of women SPS DISs	5%	—	—	—	—	—	—	—	—	—	—	—	—	—

training activities, by annually tracking the percentage of women in all aspects of Society membership activities, including major awards nominees and winners, paper awards and DLs, and by making meaningful policy changes to address any gender gaps.

The committee also recommended training programs to address unconscious biases in leadership training, methods for reducing the number of “all-male nomination slates,” and including more male members among the WISP Committee to increase and diversify the pool of nominators. A silver lining among all this data was that as of 2016, SPS women student membership had grown to 21.8% of the total student members. But the number of graduate women students was lower (16.5%), and women also represented less than 10% of nonstudent members. “As expected, the membership statistics show a leaky pipeline,” the report acknowledges, citing a “particularly big drop from graduate student members to members.” Unfortunately, this trend has not changed since. In 2021, although women student undergraduate memberships had risen to 31% of the total student numbers, for the graduate students it was 16.3%, and the number of nonstudent members was 9%.

Whatever the specific causes, this “leaky pipeline” dilemma is ubiquitous in STEM gender research, which has often found that after postsecondary graduation, we tend to lose far too many talented, bright women to other fields. In STEM and at the IEEE and SPS levels, there’s a pressing need to retain women members in IEEE, the SPS, and in academia, research, and industry.

Recent IEEE statistics from 2020 found that since 1993, the percentage of IEEE (and also SPS) women members increased from 6 to 13%. And since 2009, the proportion of women IEEE Fellows has doubled, from 3 to 6%, and SPS membership has also grown from 5 to 9%.

We recently gathered new statistics on women senior membership numbers, award recipients, women BOG members, and other important statistics on women in leadership roles at IEEE and the SPS. The SPS-compiled data included only the Society Awards and not the Paper Awards. We found that the number of women SPS fellows more than doubled since 2016, to 208. The number of women SPS award recipients during the last five-year interval (2017–2021) remained the same as the previous five-year period at five, representing 12% of total SPS awardees. But the number of awards increased in the last five-year period to nine awards, up from four between 2012 and 2016, when 18.5% of recipients were women. The original four awards were

- 1) the Carl Friedrich Gauss Education Award (formerly the Education Award)
- 2) the Claude Shannon–Harry Nyquist Technical Achievement Award (formerly the Technical Achievement Award)
- 3) the IEEE Signal Processing Society Norbert Wiener Award (formerly the Society Award)
- 4) the Leo L. Beranek Meritorious Service Award (formerly the Meritorious Service Award).

The new awards are

- 1) the Industrial Innovation Award, established in 2015
- 2) the Amar G. Bose Industrial Leader Award, also established in 2015

- 3) the Meritorious Regional/Chapter Service Award was introduced in 2017
- 4) the Pierre-Simon Laplace Early Career Technical Achievement Award began in 2019
- 5) the Meritorious Regional Distinguished Teacher Award was introduced in 2020.

For the 2012–2016 time frame, two women received the Claude Shannon–Harry Nyquist Technical Achievement Award: Eldar (2013) and Kovačević (2016). Three women received the Leo L. Beranek Meritorious Service Award: Petropulu (2012), yours truly (2013), and Min Wu in 2015. Between 2017 and 2021, two women received the Pierre-Simon Laplace Early Career Technical Achievement Award: Yuejie Chi (2019) and Piya Pal (2020). Tara Sainath was given the IEEE SPS Industrial Innovation Award in 2021. The Leo L. Beranek Meritorious Service Award was given to Ostendorf (2017), Helen Meng (2019) and in 2022 to Tulay Adali.

Among the IEEE-level SPS-related awards given to women, in 2011, Ingrid Daubechies received the IEEE Jack S. Kilby Signal Processing Medal, and Julia Hirschberg was given the James L. Flanagan Speech and Audio Processing Technical Field Award. In 2012, this award was given to Janet Baker (and her husband James Baker), and in 2018, to Ostendorf. I received the 2023 IEEE Fourier Award for Signal Processing.

On the IEEE leadership level, IEEE has had four women presidents: Jamieson, Martha Sloan, and since 2017, Karen Bartleson and Kathy Land. Karen Panetta served as an IEEE Women in Engineering (WIE) chair (2007–2009), and Shahnaz is presently the IEEE WIE chair-elect. Hemami was IEEE VP Publications Services and Products (2012–2016), and Evangelia Micheli-Tzanakou was IEEE VP Education (2007–2011). In 2022, I was elected as the 2023 IEEE VP Education.

In the SPS, the number of women in the SPS BOG increased to 16 in the last five years, and women make up four of the 34 SPS TC chairs and four of the 16 SPS EICs. Among SPS DLs, five of 26 are women, but only one woman (Dilek Hakkani-Tur) is among the 20 SPS DISs. The latter award was established in 2018.

Some of these numbers are certainly a dramatic improvement from earlier decades; there were no women EICs until the 1980s, no women TC chairs before the 1990s, only two women fellows until the mid-1980s, and no women IEEE-level SPS awards recipients before 2007 (see Table 1).

Women SPS members and women in STEM fields have made many great strides, despite the inequities that they continue to face. Many of my women colleagues share a cause for optimism and celebration, in large part thanks to the participation of women members and leaders in the field. We have more than 900 IEEE WIE groups worldwide and both an IEEE and a Technical Activities Board Diversity and Inclusion Committee, while almost all IEEE Societies and Councils have women or equity, diversity, and inclusion committees or subcommittees, and there are many other women-focused committees. We also have an IEEE conferences Code of Conduct, with a zero “tolerance for discrimination, harassment, or bullying in any form at IEEE-related events.”

Diversity is a key aspect of any healthy ecosystem, in nature and the cultural institutions we nurture. Wu, our current SPS president-elect, is an ideal leader for continuing to grow our diversity. “Through many volunteer roles, I have gained experiences, broadened my horizon, developed leadership skills, and made friends and formed comradeships around the world,” says the specialist in information forensics and security and multimedia SP. “Being an SPS member for about 25 years (starting as a student member in graduate school), I couldn’t have foreseen that two decades later, I would contribute directly to blazing a trail to diversify the leadership of SPS,” she says, acknowledging that she has “overcome the twists and turns” in her career including “many forms of implicit bias and double-standard treatment.” Currently the associate dean for graduate programs at the University of Maryland’s A. James Clark School of Engineering, Wu was born and raised in China, did her Bachelor of Science at Tsinghua University in Beijing, and her Ph.D. in electrical engineering at Princeton University. She offered some sage advice that has helped her overcome institutional and cultural gender biases. “Quietly biting our lips won’t help in the long run, nor lead to the greater good,” she said. “Get support and sounding boards from mentors and supportive colleagues. I have received broad support from members around the world, including many whom I have worked with over the years in various capacities. Take strides in doing good work, technical work and serving the community. Continue inclusive excellence, for more women as well as other underrepresented groups. Nurture a big heart and fair mind for the greater good and work with male colleagues to build a strong, vibrant community full of positive energy.”

As a girl, I was fascinated and gripped by Marie Curie and her incredible achievements. She said, “Life is not easy for any of us, but what of that? We must have perseverance, and above all, confidence in ourselves” [20].

She was my idol and I used to dream that one day, I would also do great things. As a young girl, I did not know exactly what my own great thing would be, but I had much confidence that I would be able to do it. That’s thanks in large part to my mother, who taught me and my siblings that with hard work, we could do anything we wanted and reach any goals we set for ourselves, no matter how high we set those goals. This self-confidence was my savior as I certainly faced many obstacles. But whenever I came up against a closed door, I looked for another door to open.

We need to nurture and cultivate that confidence. We need to start with children, encouraging young girls and young women to follow their passions. I believe that more of us women leaders should visit elementary and secondary schools and talk to students, girls and also boys about the beauty of invention, the mysteries of the universe, the fact that engineering can solve so many of our social and environmental problems and advance technology for all of humanity. It is interesting that there has been a healthy increase in women students studying health-related engineering, and many biomedical strides and inventions have been made by experts

in SP to help foster a better life for humanity, to literally help save lives.

In every country, and all cultures, we still have a lot to do, and a long way to go, to bridge the gender divide, fix the leaky pipeline, and rise above the recent plateaus in the gender wage gap, the number of women in engineering and CS professions, and among women in leadership positions in academia and industry. Women, and men, need to continue to actively participate in closing these gender gaps; by mentoring girls and young women and providing them with opportunities to succeed at all levels of society; by giving them the chance to make mistakes, and rewarding them when they do succeed; by pushing beyond our own cultural limitations and internal biases to recognize their talents, no matter their gender, color, and ethnicity; by helping women open their own doors—in academia, industry, and all aspects of life.

## Acknowledgment

I wish to thank a number of people for their assistance in writing this 75th anniversary article. For the invitation to write this article, many thanks to the guest editor team of the SPM Special Issue for the SPS 75th anniversary, and its Chair, Christian Jutten. Special thanks also go to VP Membership, K.V.S. Hari, and the SPS Membership Board for their support. Thanks also to Jessica Perry for creating and disseminating the questionnaire to many SPS women members, to the women respondents that provided input, to Danielle Egan for her great help in editing and researching the article, and to George Olekson for gathering the data on SPS women members.

This article was submitted for publication on 15 October 2022 and reviewed and accepted on 26 December 2022.

## References

[1] "Fifty years of signal processing: The IEEE signal processing society and its technologies, 1948–1998." *IEEE Signal Process. Soc.*, Piscataway, NJ, USA, 1998. [Online]. Available: <https://signalprocessingsociety.org/uploads/history/history.pdf>

[2] A. Martinez and C. Christnacht, "Women making gains in STEM occupations but still underrepresented." U.S. Census Bureau, Washington, DC, USA, Jan. 26, 2021. [Online]. Available: <https://www.census.gov/library/stories/2021/01/women-making-gains-in-stem-occupations-but-still-underrepresented.html>

[3] T. S. Perry, "Inflation-adjusted income for U.S. Engineers drops: Insights from IEEE-USA's annual salary survey in six charts," *IEEE Spectr.*, Sep. 15, 2022. [Online]. Available: <https://spectrum.ieee.org/electrical-engineer-salary>

[4] "Women in stem workforce index," UC San Diego Extension, La Jolla, CA, USA, 2020. [Online]. Available: [https://extendedstudies.ucsd.edu/getattachment/community-and-research/center-for-research-and-evaluation/Accordion/Research-Reports-and-Publications/Women-in-STEM-Workforce-Index-FINAL-for-CRE-7\\_22\\_20.pdf.aspx?lang=en-US](https://extendedstudies.ucsd.edu/getattachment/community-and-research/center-for-research-and-evaluation/Accordion/Research-Reports-and-Publications/Women-in-STEM-Workforce-Index-FINAL-for-CRE-7_22_20.pdf.aspx?lang=en-US)

[5] "Women in STEM USA statistics." Stem Women, Liverpool, U.K., May 21, 2021. [Online]. Available: <https://www.stemwomen.com/women-in-stem-usa-statistics>

[6] I. Singh, "By the numbers: Women in STEM: What do the statistics reveal about ongoing gender disparities?" *Yale Scientific*, Nov. 2020. [Online]. Available: <https://www.yalescientific.org/2020/11/by-the-numbers-women-in-stem-what-do-the-statistics-reveal-about-ongoing-gender-disparities/>

[7] Library of Congress, Washington, DC, USA. [Online]. Available: <https://www.smu.edu/Lyle/Academics/Departments/CS/People/Emeritus/EtterDelores>

[8] "Papers of Bernice Resnick Sandler, 1963-2008." Harvard Univ., Cambridge, MA, USA, 1963-2008. [Online]. Available: <https://web.archive.org/web/20180703162227/http://oasis.lib.harvard.edu/oasis/deliver/~sch01194>

[9] "Bernice Resnick Sandler," National Women's Hall of Fame, Seneca Falls, NY, USA. [Online]. Available: <https://www.womenofthehall.org/women-of-the-hall/voices-great-women/bernice-resnick-sandler/>

[10] B. R. Sandler, "The chilly climate," Illinois Wesleyan Univ., Bloomington, IL, USA, 2005. [Online]. Available: <https://sun.iwu.edu/~mgardner/Articles/chillyclimate.pdf>

[11] L. H. Jamieson, "Reflections on signal processing, engineering, education, leadership, IEEE, women, and change," presented at the ICASSP Women Signal Process., Toronto, ON, Canada, Jun. 2021.

[12] J. Kovačević, NYU Tandon School of Engineering, New York, NY, USA, 2018. [Online]. Available: <https://engineering.nyu.edu/faculty/jelena-kovacevic>

[13] "Progress initiative," *IEEE Signal Process. Soc.* [Online]. Available: <https://signalprocessingsociety.org/professional-development/progress-initiative>

[14] A. Petropulu, "IEEE signal processing society PROGRESS: Support for underrepresented talent in the field of signal processing." *IEEE Signal Process. Mag.*, vol. 38, no. 3, May 2021, doi: 10.1109/MSP.2021.3067588.

[15] "Report on potential impact of COVID-19 on academic women in STEM." National Academy of Sciences, Washington, DC, USA, 2021. [Online]. Available: <https://nap.nationalacademies.org/catalog/26061/the-impact-of-covid-19-on-the-careers-of-women-in-academic-sciences-engineering-and-medicine>

[16] M. Ostendorf, A. Petropulu, B. Pesquet-Popescu, and E. Riskin, "Promoting women in signal processing," Ad Hoc Committee Rep., 2016.

[17] R. E. Steinpreis et al., "Science faculty's subtle gender Biases Favor male students," *Proc. Nat. Acad. Sci. USA Amer.*, vol. 109, no. 7/8, pp. 16,474–16,479, Oct. 2012. [Online]. Available: <https://genderedinnovations.stanford.edu/institutions/bias.html>, doi: 10.1023/A:1018839203698.

[18] F. Trix and C. Psenka, "Exploring the color of glass: Letters of recommendation for female and male medical faculty," *Discourse Soc.*, vol. 14, no. 2, pp. 191–220, 2003, doi: 10.1177/0957926503014002277.

[19] J. M. Madera, M. R. Hebl, and R. C. Martin, "Gender and letters of recommendation for academia: Agentic and communal differences." *J. Appl. Psychol.*, vol. 94, no. 6, pp. 1591–1599, Nov. 2009, doi: 10.1037/a0016539.

[20] "Marie Curie." Goodreads. [Online]. Available: [https://www.goodreads.com/author/quotes/126903.Marie\\_Curie](https://www.goodreads.com/author/quotes/126903.Marie_Curie)

[21] "Global stem workforce," Society of Women Engineers, Chicago, IL, USA, 2022. [Online]. Available: <https://swe.org/research/2022/global-stem-workforce/>

[22] "Women in science, technology, engineering, and mathematics (STEM) quick take," Catalyst, New York, NY, USA, Aug. 23, 2022. [Online]. Available: <https://www.catalyst.org/research/women-in-science-technology-engineering-and-mathematics-stem/>

[23] "Gender pay gap report reveals women underrepresented at high level engineering jobs in the UK," Stem Women, Liverpool, U.K., Feb. 2020. [Online]. Available: <https://www.stemwomen.com/gender-pay-gap-report-reveals-women-underrepresented-at-high-level-engineering-jobs-in-the-uk>

[24] D. I. Miller, K. M. Nolla, A. H. Eagly, and D. H. Uttal, "The development of children's gender-science stereotypes: A meta-analysis of 5 decades of U.S.: Draw-A-Scientist studies," *Child Develop.*, vol. 89, no. 6, pp. 1943–1955, Nov./Dec. 2018, doi: 10.1111/cdev.13039.

[25] R. Rincon, "SWE research update: Women in engineering by the numbers," All Together, Society of Women Engineers, Chicago, IL, USA, Sep. 11, 2018. [Online]. Available: <https://alltogether.swe.org/2018/09/swe-research-update-women-in-engineering-by-the-numbers/>

[26] M. Kantrowitz, "Women achieve gains in STEM fields," *Forbes*, Apr. 7, 2022. [Online]. Available: <https://www.forbes.com/sites/markkantrowitz/2022/04/07/women-achieve-gains-in-stem-fields/?sh=cfee9b15ac57>

[27] "Gender barriers in education," STEM for Alleurasia (UNDP/UNICEF), New York, NY, USA, 2022. [Online]. Available: <https://stem4alleurasia.org/gender-in-stem/gender-barriers-in-education>

[28] E. Higginbotham and M. Lund Dahlberg, Eds. *The Impact of COVID-19 on the Careers of Women in Academic Sciences, Engineering, and Medicine*. Washington, DC, USA: National Academies of Sciences, Engineering, and Medicine, 2021. [Online]. Available: <https://www.nationalacademies.org/our-work/investigating-the-potential-impact-of-covid-19-on-the-careers-of-women-in-academic-science-engineering-and-medicine>

[29] E. Higginbotham and M. Lund Dahlberg, Eds. *Front Matter: The Impact of COVID-19 on the Careers of Women in Academic Sciences, Engineering, and Medicine*. Washington, DC, USA: National Academies of Sciences, Engineering, and Medicine, 2021, pp. 3–12. [Online]. Available: <https://nap.nationalacademies.org/read/26061/chapter/2>

[30] "How gender equality in STEM education leads to economic growth," European Institute for Gender Equality, Vilnius, Lithuania, 2022. [Online]. Available: <https://eige.europa.eu/gender-mainstreaming/policy-areas/economic-and-financial-affairs/economic-benefits-gender-equality/STEM>

[31] W. Unanue, M. E. Gómez, D. Cortez, J. C. Oyanedel, and A. Mendiburo-Seguel, "Revisiting the link between job satisfaction and life satisfaction: The role of basic psychological needs," *Frontiers Psychol.*, vol. 8, May 2017, Art. no. 680, doi: 10.3389/fpsyg.2017.00680.

[32] C. Duhigg, "What Google learned from its quest to build the perfect team," *The New York Times*, Feb. 25, 2016. [Online]. Available: <https://www.nytimes.com/2016/02/28/magazine/what-google-learned-from-its-quest-to-build-the-perfect-team.html?smprod=nytcore-iphone&smid=nytcore-iphone-share>

# IEEE Signal Processing Society Flagship Conferences Over the Past 10 Years



©SHUTTERSTOCK.COM/TRIFF

**T**hroughout the IEEE Signal Processing Society's (SPS's) history, conferences have functioned as a main way to connect within the Society, bringing together the signal processing research community to discuss and debate, establish research collaborations, and have a good time. These immersive conference experiences, to which attendees travel from all over the world to be together for a set period of time, have certainly been challenged over these past years, but SPS leadership was able to guide and steer through the constant unanticipated changes, with continued financial stability and growing momentum for a more inclusive future. This article gives an overview of the evolution of SPS conferences in the past decade and presents the challenges ahead.

## Introduction

SPS conferences are overseen by the vice president (VP) for conferences, who chairs the SPS Conferences Board and serves on the IEEE Board of Governors for a three-year term. The VP for conferences has the direct overall responsibilities for the development, design, operation, and improvement of the Society's conferences and workshops and their proceedings. Under the direction and guidance of the VP for conferences, throughout the years, all the SPS conferences and workshops have remained adaptive, flexible, and aimed at consistent improvement of the member and customer experience.

The two flagship conferences of the SPS are ICASSP and ICIP. An open call for proposals is held for each of these conferences annually, and members of the Society formulate a complete plan for the location, organizing committee, timeline, and innovations for the conference for review and selection by the SPS Conferences Board. For many volunteer conference organizers for ICASSP and ICIP, this is a five-plus year commitment. The SPS conference organizers have consistently shown their ability to be innovative in terms of new engaging programming and event formats, being versatile and proactive and able to lead an immense operation to

execute these conferences flawlessly throughout the years. The relationship between the conference organizers and the VP for conferences and SPS Conferences Board is a critical element to the success of the Society throughout the years, and none of the innovations highlighted in the following would have been possible without the effort of the conference organizers.

Due to the joint efforts of the SPS Conferences Board and the conference organizers, ICASSP and ICIP have been trending toward growth in terms of the number of papers submitted and attendees and the amount of content being offered throughout this period.

Throughout time, with the onboarding of each new VP for conferences, new ideas and priorities are infused into the SPS flagship conferences, but there are many consistent goals and efforts: 1) focusing on creating an equitable and inclusive Society by removing financial and access barriers and providing additional opportunities to engage as well as by adhering to a statement on diversity and inclusion, 2) maintaining the high quality of the technical content and presentations, 3) providing excellent opportunities for networking and idea exchanges, 4) streamlining operational tasks and offering transparent and supportive guidance to the conference organizers, and 5) leveraging metrics and data for decision making by the Society leadership.

## Evolution in the past decade

### *From 2015 to 2017: Growing competition from other conference models*

In the period from 2015 to 2017, the major goals to be pursued were related to the search for a dynamic equilibrium in a changing world. On one hand, the tradition of the SPS with respect to quality and scope in flagship and Technical Committee (TC) workshops and conferences was consolidated. On the other hand, it was perceived that it was necessary to explore new scientific communication forms and ways to eliminate possible barriers to make possible the involvement of larger communities, such as students and industrial people. The Society, in that period, felt and tried to react to the growing competition from other conference models over the previous several years. At the same time, the SPS was confident that its members had strong scientific content that was used to communicate according to certain well-assessed modalities. For example, research works that were more consolidated typically targeted journal publications, while initial and limited works with promising results and interesting ideas were typically targeted for the conference format. All papers followed a careful and serious review process.

However, there may have been a gap in successfully reaching larger audiences, including the corporate and industry audience, that was avoided by other conferences

using alternative models. Examples that were often cited at that time were other conferences in computer vision and deep learning, which took advantage of the explosion of machine learning with deep learning and convolutional networks to grow considerably. In that case, submitted conference papers were subject to more severe selection and review, resulting in conference papers that were often more cited and more impactful than journal papers. Such conferences achieved large industrial participation, appearing to be more selective despite parallel workshops being created to collect papers in a less selective modality and keeping

the number of active people participating in the conferences high. The SPS made the decision to remain more open and inclusive for its members, with a lower rejection rate of around 55%. With this in mind, the SPS focused on being a top choice, offering highly attractive conference modalities for the community to show its research work. Many discussions and some experiments took place to see how new models of conferences could be

explored and which modifications to flagship ones could be defined. Some directions were continued from previous years, some lines within existing conferences were potentiated, and some experiments were evaluated and discontinued, such as the IEEE Global Conference on Signal and Information Processing (GlobalSip), a flagship conference originally introduced to increase industrial attraction within the SPS. Summing up, the attempt to modernize the conference offerings of the Society followed two main lines: 1) proposals of new services within existing conferences and 2) attempts to introduce new conferences.

### New services in existing conferences

Strong cooperation in this direction was in place with the SPS Membership Board, as many of the ideas that were explored came from the objective to introduce new services for members. In addition to membership, the citation index became, in this period, a central issue for SPS conferences, as a matter related to the attractiveness and impact of research works and, consequently, to the careers of researchers in the field presenting their work.

During this period, many new services were introduced. Among them was the SPS Signal Processing Repository (SigPort) platform, which made available the ability for authors to upload and share complementary materials to their papers and share with members and attendees. This was first tried at the GlobalSip conference, where slides and posters were made available to attendees. SigPort is still available today as a repository for SPS members and for supplemental conference materials.

Open Preview was also trialed at ICIP 2016 as a first-time offering within IEEE. This new service allowed for the possibility to speed up the citation of conference papers by publishing the preconference final papers in

**With the onboarding of each new VP for conferences, new ideas and priorities are infused into the SPS flagship conferences, but there are many consistent goals and efforts.**

IEEE *Xplore* as open access for the one-month period before the conference for anyone who visited IEEE *Xplore*. Following the results of the ICIP 2016 trial, this program is now a standard for ICIP, ICASSP, and other SPS conferences and workshops as well as other conferences within IEEE (see Figure 1).

Within this period, new criteria were designed to allow papers recently published in SPS journals to be presented during flagship conferences and TC workshop sessions, increasing the visibility of these papers and allowing authors to collect in-person feedback at conferences. At the same time, the consolidation of procedures allowing the management of events, such as the IEEE Signal Processing Cup, was begun to favor the involvement of younger active members, including students, in conferences.

The necessity of finding ways to actively attract and involve more industrial members was also a central point. A subcommittee was created to discuss which forms for industrial participation could be put in place to promote initiatives and activity, such as dedicated sessions and simplified review modalities for papers not to be published on IEEE *Xplore* but using SigPort. At that time, social media was not yet at full development despite having had a considerable advance in the previous 10 years. For example, the difference between the tools to manage the conference program for ICIP, in Genova, Italy, in 2005 and the one used in this period was a big jump. Review processes as well as tools for assessing conference programs were made

different, allowing organizers to concentrate on more specific aspects by relying on a more robust and spread digitalization. However, in the period under description, connectivity problems were not unusual at conferences, and while the progressive elimination of many hard materials (only the elders remember the weight of ICASSP and ICIP proceedings to be carried back to the lab on behalf of supervisors ...) in favor of digital counterparts progressively made easier the life of attendees, some transition issues were to be considered, implying the coexistence of hard and soft materials. This required much work in the SPS Conferences Board to determine and discuss initial guidelines that were updated progressively to help conference organizers in their job. Through the improvement of a continuous relationship with conference organizers along with the presentation of proposals, the selection, realization, and postconference closure procedures were continuously performed to keep the process under control. IEEE staff had a central role in that, especially in all administrative aspects. Also, communication related to technical aspects was considered to be improved. The SPS Conferences Board introduced, at that time, the liaison member for each flagship conference to try to guarantee better communication among the central boards and the conference committees. The participation in joint meetings with the TCs Board took advantage of liaison members to better synchronize the essential role of TCs in review processes with technical chairs of conference committees.

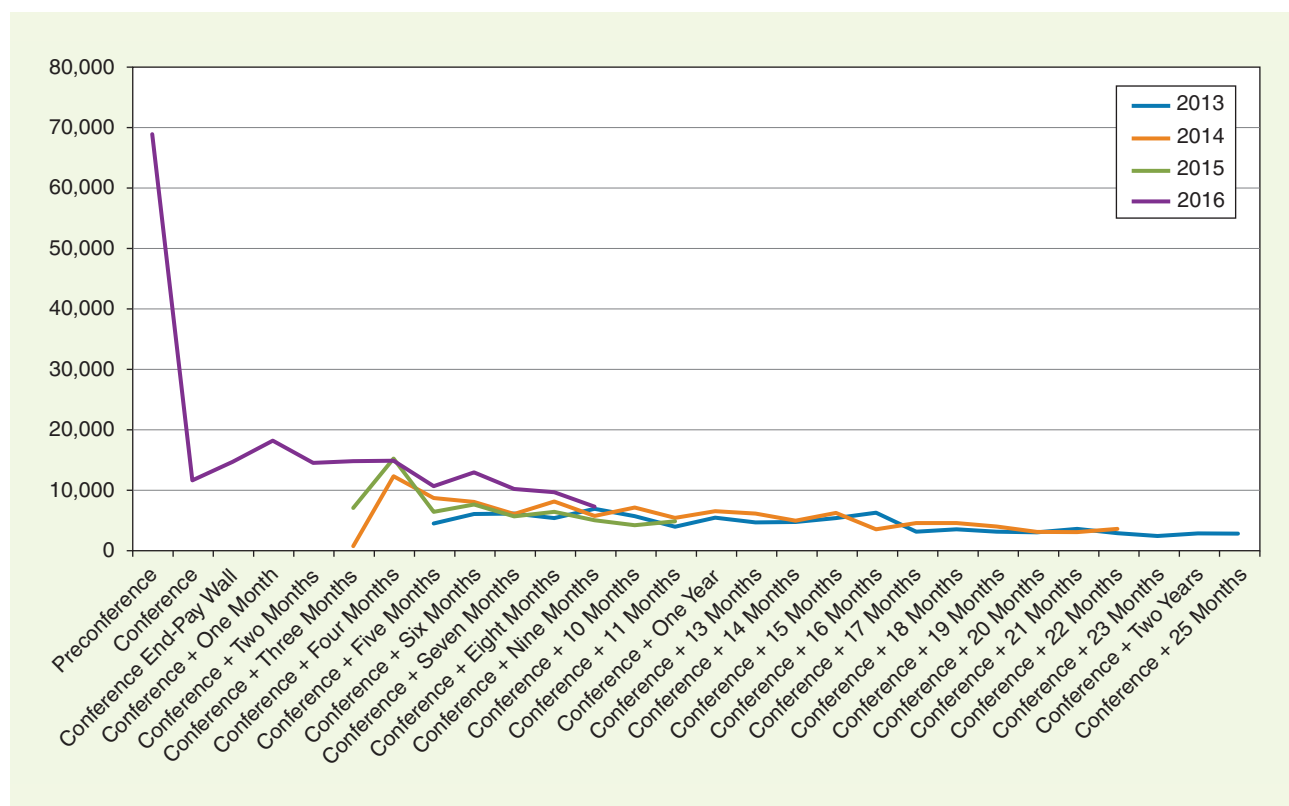


FIGURE 1. The total downloads from IEEE *Xplore* through time, using the conference itself as reference moment, including both PDF and HTML downloads.



### New conferences

A second major point was related to defining a policy about new conferences. Together with the issue of maintaining high quality, budget issues also had to be considered. On the one hand, the income streams from conferences had to be kept limited so as not to increase too much the cost for organizers and attendees. On the other hand, conferences were becoming an important source for guaranteeing the capability of the Society to put new services in place. A good tradeoff could be reached only if the Society could identify new highly attractive research lines that captured the interest of its old and new members (including industrial ones) while providing new formats. So, in this period, attention toward making the Society able to monitor the growth of new research fields proactively was supported by the SPS Conferences Board and, in general, by the other governing boards. This activity led to results related to making plans to start new TCs and special interest groups that were coupled with conference events following a dedicated conference strategy. On the other side, a critical evaluation of experiments started in previous years was carried on, trying to efficiently learn from actions and, eventually, errors. The GlobalSIP conference is an example. The details here are not relevant to different positions and opinions in evaluating GlobalSIP and the evaluation of whether the conference was significant or should have been discontinued, but the end of the process is known. While GlobalSIP had merits in trying to address an enlarged set of members and was the place to do experiments (for example, SigPort usage was started there), it was perceived that the results obtained in terms of attendee numbers and the enlargement of Society membership as well as added value to the Society's conference offerings did not allow its continuation. Moreover, lessons were learned about the difficulty to set up an efficient promotion and review process for a new conference to allow results to be collected in the short time that characterizes our days. Some solutions were attempted to make the review process integrated with TC activities, but the high review workload perception in this domain was learned as a difficult obstacle. Reviews carried on by TCs for ICASSP and ICIP in addition to TC workshops were already a very huge load, and so having a third flagship conference was a major overload. Despite the efforts of GlobalSIP, the conference organizers were impressive, and the experiments carried on were useful lessons for activities moved later to other SPS conferences and workshops.

### Human relationships

In this period, nothing of what happened in the following years was predictable, and the conference was a concept that could not be separated from traveling and meeting in person. So, the conference experience was mostly central despite some experiments on streaming plenary lectures and relevant moments for a virtual audience. For instance,

ICIP 2017, in Beijing, China, was very important for Society networking/social and global outreach (Figure 2 is a snapshot of the banquet at ICIP 2017). Fully virtual meetings were discussed but were theoretical, and no one could have forecast the speed at which the pandemic made them concrete and changed our behaviors. Meeting in person had some advantages for establishing personal contacts but also some issues. The duration of the SPS Conferences Board meeting in previous years was somehow legendary. During this period, there was a kind of competition between the Conferences Board and the SPS Publication Board, led by Thrastos Pappas, to arrive at reasonable durations while keeping the necessary discussion.

Also with conference organizers, the exchange of experiences was very rich. Just one of them is recalled here regarding Andre' Morin, who contributed to organizing a very successful ICIP meeting in Quebec in 2015. When we were informed that he passed some years later, his contribution was not forgotten in suggesting and realizing improvements to SPS conferences. In-person conferences concern meeting people and sometimes people quite different from typical attendees: someone may recall that at the Quebec conference, we had the opportunity to meet, in person, an iconic figure in our community. The unchanging woman from the Lena picture became a kind advanced-age Swedish mother who told her history. Someone said it was inappropriate and so on and so on. From the point of view



FIGURE 2. The banquet at ICIP 2017, in Beijing.

of the Conferences Board chair at that time, it was a human moment in our conference, reminding us that behind our papers and professional activities, there is life. The lesson that was learned is that the way social events based on the history of a community like the SPS are chosen can generate different reactions at a conference. However, this cannot limit the choices of organizers in highlighting different moments of our past. In this way, they can take care of the way events are presented and different sensibilities are considered. A good conference organizer has to balance the program among consolidated parts and social events to enrich the human experience at a conference. Morin, in that case, was able to do that. Who among older SPS guys does not remember Magdy Bayoumi dressed as a pharaoh on a camel under a Pyramid in Egypt or guiding a jazz band in New Orleans? Behind the curtain, too, all the incredible and well-prepared work done by the SPS staff that, in this period, faced some changes, was enriched by the possibility to meet in person, for example, in the morning executive meetings of the Conferences Board as well as in the extended ones. More in general, each in-person meeting with all the volunteers who were involved deserves a final big thank you to all conference organizers, volunteers, and Conferences Board members and the staff that contributed to the success of SPS conferences in this period. So, human relationships were a crucial aspect at that point. Nevertheless, the Society would have to face different times....

*From 2018 to 2020: Really changing times, one way or another ...*

The three strategic cornerstones for this 2018–2020 period were certainly 1) to increase the efficiency and effectiveness of the conference organization processes, 2) to upgrade the conference experiences for all attendees, and 3) to effectively cherish inclusion and diversity beyond just words.

One of the surprising features of this period was the need to be flexible and adaptable to move many of the Society’s flagship conference dates and/or locations throughout this period, starting with the last-minute move of ICASSP 2018 from Seoul, South Korea, to Calgary, Canada, due to the growing tension in the Korean peninsula but then quickly continuing with rescheduling the dates of all following confirmed ICASSPs so they would not be held during Ramadan. Then, the COVID-19 “earthquake” took place, which moved all 2020 SPS events to virtual. These were real changes and challenges, during which the Society focused on keeping the interests and values of its members and communities at the forefront of all decisions made, with a focus on opportunities for creating a more equitable, diverse, and inclusive conference ecosystem.

**While much attention had always been given to the before- and during-conference periods, the same had not been happening with the postconference period.**

Conference organization

To be able to effectively organize a successful conference, it was found critical to clearly define first what a “successful conference” is. After much discussion, it was agreed by the SPS Conferences Board members that the key factors for a successful conference would be

- participants’ satisfaction, measured with surveys and feedback
- the involvement of students, young professionals, women, and industry
- the number of participants and number of submissions
- the impact of papers and number of IEEE *Xplore* downloads
- innovative initiatives and new forms of interaction
- financial health.

The financial factor was strategically included as the last one to signal to SPS members that healthy finances are critical but that certainly, attendees’ satisfaction and fulfillment are above everything.

A key development in this period was the completion and continuous refinement of the SPS “Conference Organizer Guidelines” [1], built to play a key role in the relationship between the SPS and conference organizers for years to come. From the document, the SPS has created this set of guidelines for all SPS financially sponsored and cosponsored technical meetings, with the main purpose

to help organizers create coherent conference and workshop experiences over the years for the attendees while also accommodating innovations, creativity, and diversity. Since its first edition, this document has been consistently improved and completed to address all SPS conference and workshop procedures and customs and has contributed to significantly reducing conference organizers’ entrance barriers, especially for newcomers, with a welcoming reference document to ensure transparency of expectations. Since the organizers initially sign their agreement with these guidelines, this is the closest thing to a contract between the SPS and conference organizers.

Naturally, all conferences start with a proposal by the organizers, and preparing a proposal is a process that may be long and complex. Since this process must be also transparent and fair, the conference proposal submission instructions and the proposal review procedures have been reviewed and improved to offer a clearer pipeline. The proposal outline was updated to include key SPS values, such as engaging students and the local community and encouraging inclusive activities and events for all attendees. To effectively see beyond proposal submissions, it was approved, in this period, that all candidate sites would be visited at some appropriate stage by an SPS senior staff member and an SPS Conferences Board member to cocreate a detailed report to be given to the SPS Conferences Board before any site selection. This process has allowed having a much clearer idea of candidate sites’

strengths and weaknesses and even the strength and dedication of an organizing committee.

To help the conference finances through greater sponsorship, a patron and exhibitor prospectus template has been created and a sponsorship sales support company contracted to help build longer-term relationships with sponsors in a more proactive and time-stable way. This was an important move from the previous stage, where each organizing committee would restart the sponsorship gathering process, without much coherence over the years and thus without much stability and sustainability.

While much attention had always been given to the before- and during-conference periods, the same had not been happening with the postconference period, notably regarding automated and consistent conference data collection to better inform decisions about future conferences and workshops. In fact, it often happened at many SPS management meetings that attendees asked questions about the statistics from previous conferences, and the reply was commonly “not available” or “not consistently reported.” Many decisions and choices really depend on past statistics. How many students? How many young engineers? How many low-income countries’ participants? How many local participants? How many students are at banquets? And the list is endless . . . . For this reason, it has been decided to start a more effective conference data collection process and adopt an appropriate management and query system to be able to answer the coming questions, benchmark performance, and set targets.

### Conference experience

It was often commented that the SPS flagship conference format was too static and not evolving as much as it should, maybe excluding the technological paraphernalia. In this period, the vision was again to upgrade the overall conference experience, starting from the submission and reviewing process and continuing to the interactions and connections among all participants at an event. At the core of this vision are the people, i.e., the paper submitters, the reviewers, the authors, the conference participants, the session chairs, the local people, and so on. And all these people should communicate, connect, interact, discuss, enjoy, and have fun in old and novel ways.

Again, the before-conference period deserved much attention, notably the reviewing process, which is always very critical, especially for the authors whose papers are rejected. The feedback from authors and attendees was that the SPS paper review process was largely considered “rigid” and nonerror resilient; moreover, the conference technical program chairs struggled to find enough reviewers as well as increase the quality of the reviews. At this stage, the key objectives were to increase the efficiency of the reviewing process (e.g., immediate rejects), the quality of the reviews,

the involvement of the authors (e.g., rebuttals and reviewer discussions), the transparency of the entire reviewing process, and the reviewers pool (notably, with younger people) and create a pipeline for new reviewers and more potential SPS members.

At conference time, a critical issue to surpassing “old style” experiences was the exploitation of the increasing amount of technological paraphernalia available beyond the individual usage of computers and mobile phones. This has taken two key directions, one related to hardware and the other to software. The best exploitation example of new hardware were the e-poster sessions at ICIP 2019, in Taipei, Taiwan, where the usual paper posters were substituted by large screens and the poster presenters could exploit many

new forms of presentation and communication (see Figure 3). The key lesson was that it takes time for people to learn to use these new capabilities since most e-posters were still “old style” static images, but there were also many creative surprises. On the software front, a single SPS Events app was created for all SPS flagship conferences, which really makes

attendees’ life much easier and is ecologically friendly since much paper program printing is avoided.

Ideally, SPS conferences should leave a positive mark by compensating for their carbon footprint. This led to the idea of organizing events with the local community, not only researchers, academics, and companies but also high-school students and teachers. This could involve bringing them to the conference venue but also taking top scientists attending the conference to the local schools. This initiative should have had its first blast at ICASSP 2020, in Barcelona, Spain, but COVID-19 changed those plans. Naturally, in 2020, the COVID-19 pandemic totally changed, in weeks, the critical conference organization issues and the attendees’ experience, but this is a topic for a following section.

### Diversity and inclusion

Diversity and inclusion were major goals in this period, and these goals assumed multiple facets. The first one to

**In 2020, the COVID-19 pandemic totally changed, in weeks, the critical conference organization issues and the attendees’ experience.**



**FIGURE 3.** The e-poster session at ICIP 2019, in Taipei.

mention could be geographical diversity. During this period, the first SPS flagship conference happened in the Persian Gulf area, i.e., ICIP 2020, in Abu Dhabi, United Arab Emirates (unfortunately moved to online due to COVID-19 but returning in 2024). Moreover, the first SPS flagship conferences in India and Malaysia have been approved for ICASSP 2025 and ICIP 2023, respectively. The second goal to mention could be restrictions for SPS flagship conferences now that all major religious holidays have been identified to avoid any time clash with these conferences out of respect for all faiths; several ICASSPs had to be rescheduled to follow this new rule.

Another important inclusion direction was toward low-income countries. To include more people from these areas of the world, discounts and travel grants have been defined; for example, authors from these countries now have very discounted registration fees (e.g., US\$210 at ICIP 2022), and all people from these countries have symbolic tutorial fees (e.g., US\$20 at ICIP 2022) and totally free virtual attendance. While the attendance numbers from these countries are still low, there has been growing hope that these actions will facilitate increased participation in the future.

Finally, student participation opportunities have required major attention to make students feel more included in all conference functions and avoid a two-tier conference sensation. Students are the future of the Society. Although maybe not the most important, a very symbolic initiative was the inclusion in the SPS “Conference Organizer Guidelines” of the sentences “Although not mandatory, it is strongly recommended that the banquet is included in the registration fees for all types of attendees at no additional cost... Banquets do not have to be formal and expensive, and emphasis could be placed on creating an inclusive environment for all attendees.” Only a very low percentage of students usually participate in the conference banquets, thus creating a segregation impression. However, more initiatives have been taken to include students, notably, organizing student luncheons with top researchers, defining very low tutorial fees (e.g., US\$20 at ICIP 2022), allocating travel funding for students and young professionals (also for workshops), and creating student cups, 5-min video clip contests, job fairs, and student research networking events.

### The pandemic

While much brainstorming had been dedicated to virtual events and virtual participation, only small developments happened before March 2020, e.g., streaming plenaries and keynote talks on Facebook Live, and nobody was really ready for the landslide that happened with the emergence of the COVID-19 pandemic. The migration to fully virtual events was meteoric and total in a couple of months, and the impacts may last for a long time; it remains to be seen. What did not happen for decades happened in a few weeks,

and, surprisingly, with the intense efforts of all involved and the amazing readiness of technology, the SPS flagship conferences entered a new era, with rather limited suffering. Naturally, the move to virtual raised many difficult issues, some more legal, e.g., what to do with the contracts related to the physical venues already contracted (ICASSP 2020 was in May, just a couple of months after the pandemic began), others about the attendees’ experience, e.g., what application, presentation, and communication functionalities to use for the virtual conference, and, finally, others more financial, e.g., what registration fees model to use for the virtual conferences. For the first SPS flagship conference during COVID times, i.e., ICASSP 2020, in Barcelona, the decision was for fully recorded presentations, including the keynotes, although with real-time questions at the end; this has changed over time, with more and more presentations happening in real time and recorded for later viewing, notably, in different time zones. For all of 2020, it was decided that nonauthor registrations would be totally free, which led to record registration numbers, notably, around 16,000 for ICASSP 2020 (rising from around 3,500 at ICASSP 2019, in Brighton, United Kingdom). While it was great to have access to these new members of the community, it was soon clear that free registrations do not necessarily imply participation, and thus, the decision for 2021 and beyond was to have minimum registration fees for virtual participation.

After the shocking forced experience, the SPS started thinking about the post-COVID future, notably, how conferences should be in the future after what was learned in 2020. Clearly, virtual conferences allow remote participation and increase inclusion but also prevent the warm connections that only physical conferences allow. What about hybrid conferences, allowing people to choose to be present or not, including authors? The impression was that many people would simply stay at home because they could and that the “amazing” atmosphere of past SPS flagship conferences would never return. Are conferences really essential beyond journals if they do not produce physical human contact? And what about the technical discussions around a dinner table, with opinions from all around the world, with real laughs and real beer? These are some of the questions that the years to come will help to answer... Overall, 2020 was a year of forced change, but it allowed the SPS to show its strength, resilience, and commitment to an inclusive future. And that was great to see.

Against all odds, 2021 was still restricted by the pandemic. However, the SPS tried to use that to its advantage. ICASSP 2021 and ICIP 2021 were finally held virtually, in Toronto, Canada, and Anchorage, Alaska, respectively, and this helped the Society to keep on learning about the

**The migration to fully virtual events was meteoric and total in a couple of months, and the impacts may last for a long time; it remains to be seen.**

*From 2021 to 2023: Toward a marketplace model*

Against all odds, 2021 was still restricted by the pandemic. However, the SPS tried to use that to its advantage. ICASSP 2021 and ICIP 2021 were finally held virtually, in Toronto, Canada, and Anchorage, Alaska, respectively, and this helped the Society to keep on learning about the

best virtual platforms and how to organize virtual technical programs. Not only must the virtual lecture and poster sessions offer the best experience to the attendees but the conference networking and social events should not be lost. The survey that the SPS is now regularly doing after each of its two flagship conferences helps to steer the organization of future events. Clearly, people are longing to meet in person at the conferences, and 2022 seemed to slowly bring things back to normal in terms of COVID-19. In any case, the organizing committees are still working with plans A, B, and even C to adapt to the possibly changing situation. This was the case of ICASSP 2022, in Singapore, planned to be held mainly in person until a new pandemic outbreak appeared in China just two months before the conference. Once more, the SPS thanks its volunteers for the great job that they do when facing these complicated situations.

Despite the roller coaster that the organization of events has become, building on the achievements of the past years helps to continue offering attractive conferences of high quality. Figure 4 describes the evolution of ICASSP in the past 10 years and how the number of submitted papers and the ratio of the number of attendees versus submitted papers in the

past three editions are the largest ever. The planned strategic cornerstones for this 2021–2023 period are 1) to stabilize the hybrid model, 2) to make the industry program and participation grow, and 3) to incorporate other kinds of submissions, which can enrich the conference papers. Since our conferences should be a powerful tool to help SPS membership grow, we will continue fostering initiatives that promote diversity and new services and liaise with other sister Societies and events. The vision for this period is that our flagship conferences should be a bubbling marketplace with heterogeneity and a wide range of products to offer.

**We will continue fostering initiatives that promote diversity and new services and liaise with other sister Societies and events.**

#### Hybrid conferences

The discussions that have been carried out within the SPS Conferences Board and some ad hoc meetings stated the positive

aspects that the virtual component brings to our conferences: increasing the range of attendees that can participate (e.g., students and members of industry), promoting the green and sustainable aspect of our events, and allowing joint nonco-located events, among others. For this reason, the SPS has intensively helped the organizing committees of our flagship conferences to include this component, even in an in-person event. That has led to a central track, around which

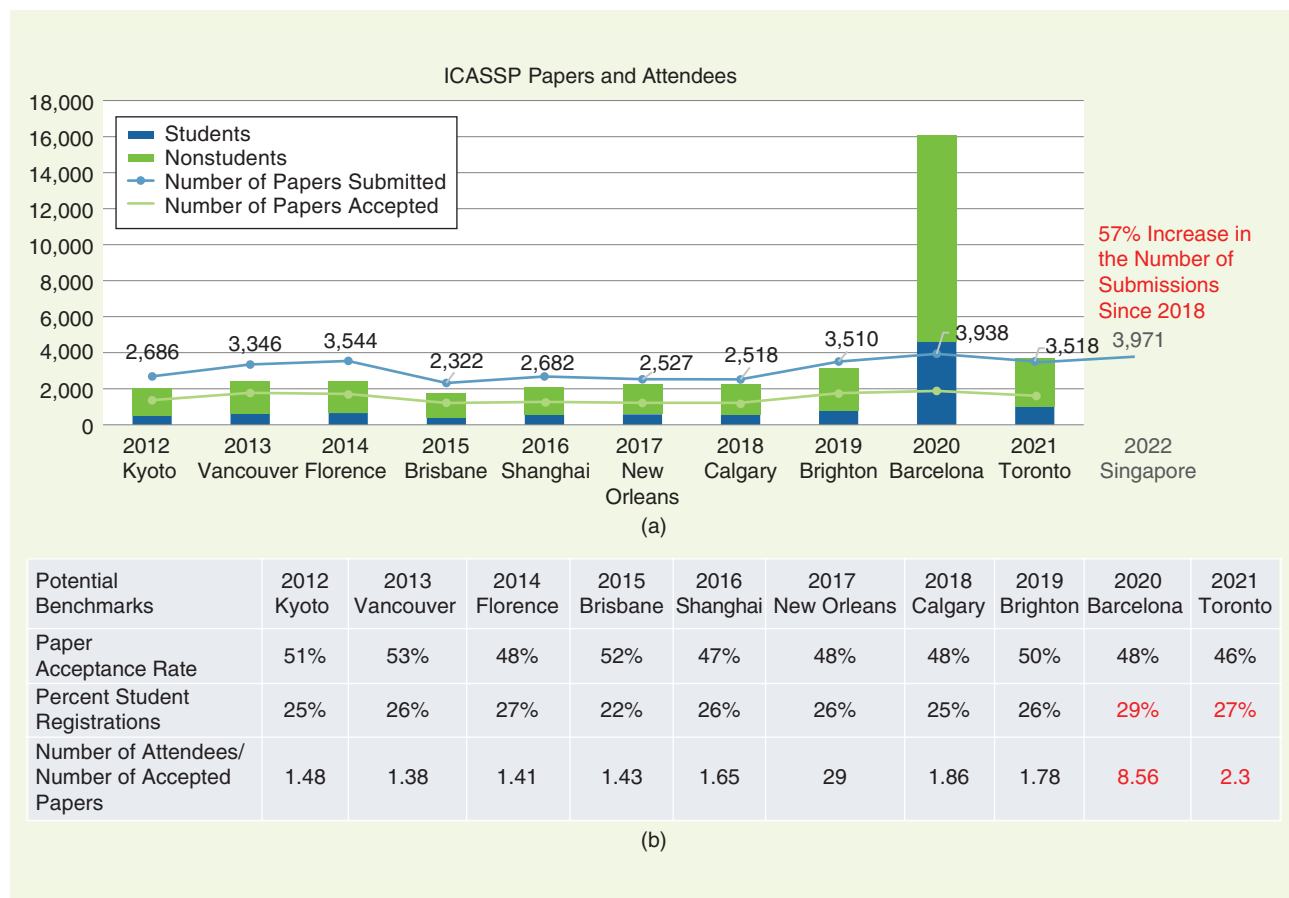


FIGURE 4. (a) ICASSP paper and attendee counts and (b) potential benchmarks (2012–2021).

a conference is structured, that is held in an auditorium with real-time streamed activities all day long and during all the days of the conference (e.g., keynote talks and panels). The regular and special oral and poster sessions are still looking for their best format and depend very much on the organizing committee's decisions and the conference venue's capabilities. The imaginative and good solutions that are devised then remain for the next editions. Importantly, the virtual platforms that can be used are well-known now, as is how to link them with the regular paper submission platforms, which helps very much in the whole process.

Authors' participation is still an open topic for discussion regarding whether they must attend in person or not. However, those that have already experienced going back to the in-person conferences feel happy with the experience and realize that it is indeed a very important aspect of our job as researchers. The spontaneity of in-person encounters cannot be replaced by any virtual tool so far—maybe when the metaverse comes, who knows, but let us live the present and midterm future, which is already difficult to predict. Also, among the authors, there is a feeling that the virtual component has brought in much work with all the material that has to be submitted: papers, videos, slides, and posters (all of it in quite a short period of time). The SPS is still looking for ways to combine different kinds of participation so that it ends up with a win-win situation for all attendees.

### Industry program and participation

The flavor of SPS conferences is mainly academic, and the events stand out because of the academic quality of their technical programs. However, the presence of industry lags a little bit behind that in other Societies. We assume that one of the main reasons is that there is no main supporting signal processing industry, as, for instance, is the case in communications and power electronics, but several industries. In any case, signal processing has become the “silent” core of many different industries, and we can play this in our favor, together with the deep connection between signal processing and artificial intelligence. To help increase the participation of industry in our flagship conferences, we have defined some key performance indicators (KPIs) and set some thresholds to be met: 1) the number of sponsorships and exhibitors, 2) the size of the industry program, and 3) the number of attendees and papers.

Taking the baton from the past VP for conferences, who put in place a patron and exhibitor prospectus template and promoted contracting with a sponsorship sales support company for ICASSP and ICIP, the SPS recently signed a multiyear contract with the company to build longer-term relationships. The KPIs were part of the contract. Also, there are several industry members of the Society that should regularly attend ICASSP and ICIP but do not. In addition, to

share some of the motivations of the academic attendees, the Society should further work into its conferences other kinds of motivation that are more specific to industry members, such as recruiting (i.e., it is good to have direct contact with potential candidates) and stay up to date with latest trends, which is seen as an essential part of their job. They strongly appreciate that a conference deploys a specific industrial program with specific workshops, panels, special sessions, and keynotes. For this reason, in 2021 and 2022, the Society consolidated and enhanced the industrial program, especially regarding industry-driven keynotes, talks, tutorials, and workshops. Also, industrial forums and panels are good to showcase the industrial trends in the near, medium, and long terms. The more the sponsors get closer involvement in the elaboration of the conference program the better, and it is important that this be reflected in the patron prospectus and planned from the very beginning. Closer involvement of industry in Society activities must be achieved: the Signal Processing Cup, SPS Video and Imaging Processing Cup, and SPS Grand Challenges should not be exclusively organized by academic TCs. The Show and Tell Demonstration, for instance, is a good forum for quick presentations of new products and technologies. These and other new activities should be envisaged by the Society and the Conferences Board, which has already

incorporated in the conference guidelines the new structure that the conference organizing committees must have to be able to cope with the increasing number of activities and promote growth. One very good example of the new industry-oriented activities is the Entrepreneurship Forum, which was successfully organized for the first time at ICASSP 2022, in Singapore.

Finally, keeping the participation hybrid (both virtual and in person) is necessary to including our industry members, who may have more travel restrictions than in academia. Also, planning for alternative participation formats in a conference that do not necessarily require a full paper can help. This leads to the next topic about incorporating other kinds of submissions, which can enrich the conference papers.

Enrich the types of conference submissions

The goal of our flagship conferences is to offer a vibrant marketplace that gives the possibility to interact and network around the signal processing topic. To open the floor to admitting different types of conference paper submissions and planning for alternative participation formats that do not necessarily require a full paper can help to advance in this direction. For example, workshops and sessions with short papers and extended abstracts targeting participants from industry have already begun to be implemented, at ICASSP 2022. Papers/abstracts may be part of the proceedings but do not really need to be published in *IEEE Xplore* (mitigating the <50% acceptance ratio).

**The spontaneity of in-person encounters cannot be replaced by any virtual tool so far—maybe when the metaverse comes, who knows, but let us live the present and midterm future.**

Another avenue is that authors of papers published and accepted in SPS journals may present their work at ICASSP and ICIP in appropriate tracks. These papers will neither be reviewed nor included in the proceedings. It is high time to promote this alternative, which was a little bit hidden in the call for papers. In addition, at ICASSP 2023, for the first time, *IEEE Open Journal of Signal Processing (OJSP)* will provide a special track for longer submissions, with the same processing timeline as ICASSP. Accepted papers will be published in *OJSP* and presented at the conference but will not be included in the conference proceedings. With this, the Society begins its journey toward open access for conferences. This is a topic that has already taken many brainstorming sessions within the Board of Governors and that will continue to do so. In the absence of a disruptive and clear winning solution, the strategy is to try different alternatives that have previously been well discussed and build on them.

The discussions about how to further increase the quality of our conferences, at least within some specific tracks, are still open. In our opinion, this must be so and reflects the continuous aim for improvement and adaptation in our Society.

And much more ...

Importantly, the SPS Conferences Board keeps track of different trials and discussions in the past so as not to repeat past actions that failed. However, it also keeps an eye on how the world and circumstances evolve. A brilliant idea in the past may not have succeeded because it did not come at the right time. The Conferences Board agreed that ICIP, much smaller in size than ICASSP, can be a good testbed for new ideas. This can also help the conference to revive in front of others that are currently doing extremely well and create a very strong competition.

Other original activities that are enriching our conferences are the series of educational courses that were initiated at ICASSP 2022 and the Promoting Diversity in Signal Processing workshop (already in its fifth edition). The Society keeps having many open questions to debate around the key factors for a successful conference, which is very good, as it means that the SPS is continuously observing its competitors and searching for improvement. The increasing capabilities that have been gained in data analytics will help very much in shaping a good strategy. Also, the better advertisement tools that the Society has acquired in different social media reinforces the beacon role of our flagship conferences.

## Conclusions

In these few pages, we have tried to show how SPS flagship conferences have become a platform to share the latest and innovative ideas with peers all around the world. As we have commented, the conferences are in a continuous adaptation process to create the best ecosystem in every moment to help science grow with the exchange of the top ideas. Diversity, inclusion, and quality are the ultimate goals, and new changes will keep being introduced so that they become more and more a reality.

## Authors

**Ana I. Perez-Neira** (aperez@cttc.es) received her Ph.D. degree in telecommunications from the Universitat Politècnica de Catalunya. She is a full professor in the Department of Signal Theory and Communication, Universitat Politècnica de Catalunya, Spain, and the director of Centre Tecnològic de Telecomunicacions de Catalunya, Castelldefels, 08860 Barcelona, Spain. Pérez-Neira is a member of the Board of Governors of the IEEE Signal Processing Society (SPS) and the SPS vice president for conferences (2021–2023). Her research interests include signal processing for communications, focused on satellite communications. Pérez-Neira is a Fellow of IEEE and the European Association for Signal Processing as well as a member of the Real Academy of Science and Arts of Barcelona.

**Fernando Pereira** (fp@lx.it.pt) received his Ph.D. degree in electrical and computer engineering from the Instituto Superior Técnico (IST), Universidade de Lisboa. He is a full professor at IST, University of Lisbon, Portugal, and a senior researcher at Instituto de Telecomunicações, 1049-001 Lisboa, Portugal. Pereira is an area editor of *Signal Processing: Image Communication Journal* and an associate editor of *EURASIP Journal on Image and Video Processing*, and he has been elected to serve on the IEEE Signal Processing Society Board of Governors and the European Signal Processing Society Board of Directors. His research interests include video analysis, representation, coding, description and adaptation, and advanced multimedia services. Pereira is a Fellow of IEEE, the European Association for Signal Processing, and the Institution of Engineering and Technology.

**Carlo Regazzoni** (carlo.regazzoni@unige.it) received his Ph.D. degree in telecommunications and signal processing from the University of Genova. He is a full professor of cognitive signal processing and telecommunications in the Department of Naval, Electrical, Electronic, and Telecommunications Engineering, University of Genova, I-16145 Genova, Italy. He is a coordinator of a joint Ph.D. course on interactive and cognitive environments between the University of Genova and other European universities. His research interests include data fusion, emergent self-awareness in autonomous systems, cognitive multimedia signal processing, Bayesian machine learning, and cognitive radio. He is Senior member of IEEE.

**Caroline Johnson** (c.j.johnson@ieee.org) received her B.S. degree in marketing from Rowan University and her Certified Meeting Professional certification from the Events Industry Council. She is the senior manager of conference strategy and services for the IEEE Signal Processing Society and a Certified Meeting Professional in New Brunswick, NJ 08854 USA. She is a member of the Professional Convention Management Association, International Congress and Convention Association, and the American Society of Association Executives.

## Reference

[1] SPS Conference Organizer Guidelines, "Conference resources." [Online]. Available: <https://signalprocessingsociety.org/events/conference-resources>, 2023, pp. 1–71.



# How the 1969 IEEE Convention and Exhibition Changed My Life Forever

*The story of how a single IEEE event attracted over 60,000 people and gave one young man a lifelong career*



©SHUTTERSTOCK.COM/TRIFF

Let me begin by telling you that 1969 was a great year—a really, really, really great year.

It was the year of the *Apollo 11* Moon landing, Woodstock, and the New York Mets winning the World Series. It was the year I took my mother and sister to see *Hair* on Broadway.

But all of that happened shortly after what made 1969 such a really great year. That was the day I visited the IEEE '69 International Convention & Exhibition and, in the process, found my future career.

## An important call

The phone rang. It was my best friend, Jonathan Bird, a junior at Brooklyn Technical High School.

“So, what’s up?” I asked.

“I’m going to the IEEE show tomorrow,” he said.

“The *what* show?” I replied.

“It’s a show for electrical engineers,” he explained. “All the big electronics companies will be showing their stuff there.”

My interest was piqued. Ever since visiting the 1964–1965 World’s Fair, all things electronic had fascinated me. I was an avid *Popular Electronics* reader, eagerly devouring the latest news about developments in integrated circuits, computers, telecommunications, lasers, and so on. All in all, it was a great time to be a 14-year-old tech-crazed kid.

The idea of attending an event packed with news and presentations about the latest electronic breakthroughs sounded far more appealing than spending another dreary day at Junior High School 119 in Glendale, Queens.

“I want to go with you,” I told Jonathan.

“You’re too young,” he shot back. “They’re only admitting high school and college students.”

“I don’t care,” I answered. “I’m going with you.”

He sighed. It was something he did quite frequently when in my company.

## Welcome to Wonderland

Bright and early the next morning, I met Jonathan outside my Queens, New York apartment. A bus and three subway rides



later we emerged into the light at Manhattan's Columbus Circle. The New York Coliseum was directly in front of us. The International-style structure, opened in 1956, featured both a low building with an exhibition space and a 26-story office block. A sign outside declared: "Welcome to the IEEE Convention and Exhibition" (Figure 1).

Jonathan and I entered the exhibition space's foyer. My friend presented his school ID to one of a series of women positioned behind glass windows. As a student at Brooklyn Tech, New York City's elite engineering school, he was immediately granted a pass.

Then, it was my turn. I approached the window looking completely flustered.

"Oh, no," I groaned while rapidly patting my shirt, trouser, and jacket pockets. "I can't find my ID! This is horrible!"

If I was on Broadway instead of Columbus Circle, I probably would have been nominated for a 1969 Tony Award.

Jonathan sprang into action "I'll vouch for him," he told the lady. "I don't want my friend to get in trouble. We're supposed to write a report about our visit."

The lady smiled, nodded, and handed me a pass. Jonathan looked at me and rolled his eyes, ever so slightly.

Hooray!!! I was in!!!

## A different world

Walking onto any of the event's four convention floors was like stepping through a portal into a future world. Virtually every major electronics manufacturer was represented—as well as many minor players—all hoping to gain attention for their innovations. The event attracted approximately 600 exhibitors.

The exhibition had its own unique miniecosystem, quite separate from the world surrounding it. It was, in several ways, something of a throwback to earlier times. While the world outside

of the Coliseum was enveloped in turmoil, protest, and revolution, IEEE '69 was a calm and regimented oasis. As the exhibits and technical sessions (there were 48 of those) looked toward the future, the attendees and corporate representatives seemed blissfully oblivious to what was happening in the outside world. While the world was marching to the Beatles' *Helter Skelter*, the engineers were listening to Henry Mancini.

Of course, in an ironic twist of fate, it was the engineers who were the true revolutionaries, responsible for numerous technologies that changed the world forever.

Virtually all of the IEEE '69's 60,000 attendees were men. Nearly all wore the uniform of the day: a dark suit

with thin lapels, a white shirt, and a dark, not-too-wide necktie. Smoking was permitted just about everywhere; many engineers at the time smoked pipes. Women were generally relegated to support roles, dispensing admission passes, checking coats, and demonstrating various devices, batteries, cables, and other products.

Certain exhibits resonate with me to this day. For some reason, I was fascinated by the Nixie readout tubes at the Texas Instruments booth. They seemed so futuristic and cool. I figured that we would one day see these softly glowing little gems all over the place. What I didn't know was that four years earlier engineers at RCA Laboratories had developed something called a *liquid crystal display*.

Hewlett-Packard has a strong presence at IEEE '69. I recall that the company was promoting its brand-new Model 5360 A computing counter, which could measure the distance from the Earth to the Moon within one foot of accuracy. William Hewlett himself was there to meet and greet both current and potential customers.

**Of course, in an ironic twist of fate, it was the engineers who were the true revolutionaries, responsible for numerous technologies that changed the world forever.**

*Unlocking the Future*

Electrical  
Electronics  
Engineering

Monday through Thursday

4 FLOORS OF EXHIBITS  
NEW YORK COLISEUM

48 TECHNICAL SESSIONS  
NEW YORK HILTON

- 48 TECHNICAL SESSIONS at the New York Hilton. Hours: 10:00-12:30; 2:00-4:30.
- FOUR FLOORS OF EXHIBITS at the N. Y. Coliseum including over 600 firms. Hours: 10 a.m.-8 p.m. 4 Days.
- GALA ANNUAL BANQUET—Wednesday 7:15 p.m. N. Y. Hilton Grand Ballroom — \$18.00.
- FREE SHUTTLE BUSES between the Hilton and the Coliseum — every few minutes.
- REGISTRATION — Good all days — Technical Sessions and exhibits. In and out privileges. — IEEE Members \$3.00. Non-members \$6.00. Ladies \$1.00. High School Students \$3.00 if accompanied by an adult — One student per adult. Thursday only — limit of 3 students per adult.
- REG-IDENT CARD speeds request for exhibitors' literature. Ask for one when registering.
- ESCALATORS/EXPRESS ELEVATORS to the Fourth Floor.

**IEEE '69** INTERNATIONAL  
CONVENTION & EXHIBITION  
MARCH 24-27, 1969

FIGURE 1.

The Motorola booth was a special treat. Since Jonathan and I were both avid radio experimenters, we questioned a friendly Motorola representative about his company's latest gear. He then asked us to wait a minute. He soon came back with a real *Apollo* space helmet, which he quickly clamped onto Jonathan's head. When Jonathan spoke, a nearby speaker amplified his voice and he sounded like a genuine astronaut! The representative then removed the helmet from Jonathan's head and placed it on me. Four months later, as I was watching the *Apollo 11* Moon landing, I thought to myself: "A few months ago I wore a helmet just like Neil Armstrong's!"

The one thing Jonathan and I wanted to see most—an actual functioning computer—we were never able to find. We searched all four exhibit floors to no avail. If there were any on-site computers, they were hidden to us. That's actually not surprising, given the fact that even the era's early mini-computers, such as Digital Equipment Corporation's PDP-8 models, weighed somewhere between 80 and 250 pounds—far from easily transportable.

The closest we ever came to encountering a real computer was a Model 33 teletype that was positioned at one of the exhibit booths (which one, I can no longer remember). The primitive workstation was linked to a remote computer via a telephone line through an acoustic coupler.

This particular teletype allowed users to play a game that required entering numeric values in order to land a vehicle on the Moon. The actual gameplay was a question-and-answer session, in which the game asked its user for the rocket fuel burn rate at each turn, which the user would then enter as a number from 0 to 200. Once the vehicle contacted the lunar surface, the player was given a report on the vehicle's landing speed and remaining fuel. My vehicle consistently crashed into the Moon. Even Motorola's helmet wouldn't have saved me.

After several hours, it was time to head home. Jonathan and I were burdened with so much product literature that we decided to hop on the courtesy shuttle bus to the New York Hilton, which brought us closer to the subway line leading back to Queens. Unfortunately, the bus was standing room only. Worse yet was the suffocating pipe and cigarette smoke. Still, we made it back to Queens safe and sound.

## Life changing

As I previously noted, the IEEE '69 International Convention & Exhibition changed my life forever. In the years that followed, I continued to read every electronics magazine I could lay my hands on. I homebrewed projects, built kits,

experimented with radio propagation, hooked an early fax machine to my ham radio to send dirty pictures across the United States, and even restored a Model 19 teletype (Baudot, not ASCII) to communicate with other experimenters. I wanted to become an electrical engineer but, alas, my math skills were just too poor to even consider the possibility. (I could, however, send and receive Morse code at 20 words per minute.)

So, I chose to do the next best thing. I decided to write about technology! Over the decades, I've written countless articles on just about every technology imaginable. In the

pre-Internet days, I was a daily columnist for both the CompuServe and Prodigy services. I've also written for the *New York Times*, *Washington Post*, the Massachusetts Institution of Technology, *CIO Magazine*, *Electronic Design*, *PC Week*, *MacWeek*, and, of course, *IEEE Signal Processing Magazine*. It has been my honor to write about the engineers

and scientists whose research has led to the interconnected world that now allows instant communication, has led to an endless number of scientific and medical innovations, and places the world's collected knowledge at our fingertips. I am blessed.

Jonathan also never became an electrical engineer, although he certainly possessed the necessary skills. Like me, he decided to seek a media career. I'm saddened to say, however, that he died tragically young in 1994, a victim of the AIDS epidemic. I think about him every day.

The IEEE '69 International Convention & Exhibition's theme was "Unlocking the Future." Well, it certainly helped to unlock mine.

## Fast forward

Every so often, perhaps once or twice a year, I dream that I'm back at IEEE '69. In this recurring dream, I'm surrounded by engineers who are shouting, saying things like: "It's impossible!" "Let me see that!" "What's that kid up to?"

As I hold the iPhone 14 Pro in my hand for everyone to see, a team of serious-looking men, headed by someone who bears a striking resemblance Efrem Zimbalist Jr., moves toward me, and...

Then I wake up.

## Author

**John Edwards** (jedwards@johnedwardsmedia.com) is a technology writer based in Gilbert, AZ 85234 USA. Follow him on Twitter @TechJohnEdwards.



**What I didn't know was that four years earlier engineers at RCA Laboratories had developed something called a liquid crystal display.**

Geert Leus<sup>1</sup>, Antonio G. Marques<sup>2</sup>, José M.F. Moura<sup>3</sup>,  
Antonio Ortega<sup>4</sup>, and David I Shuman<sup>5</sup>

# Graph Signal Processing

*History, development, impact, and outlook*



©SHUTTERSTOCK.COM/TRIFF

**S**ignal processing (SP) excels at analyzing, processing, and inferring information defined over regular (first continuous, later discrete) domains such as time or space. Indeed, the last 75 years have shown how SP has made an impact in areas such as communications, acoustics, sensing, image processing, and control, to name a few. With the digitalization of the modern world and the increasing pervasiveness of data-collection mechanisms, information of interest in current applications oftentimes arises in non-Euclidean, irregular domains. Graph SP (GSP) generalizes SP tasks to signals living on non-Euclidean domains whose structure can be captured by a weighted graph. Graphs are versatile, able to model irregular interactions, easy to interpret, and endowed with a corpus of mathematical results, rendering them natural candidates to serve as the basis for a theory of processing signals in more irregular domains.

The term *graph signal processing* was coined a decade ago in the seminal works of [1], [2], [3], and [4]. Since these papers were published, GSP-related problems have drawn significant attention, not only within the SP community [5] but also in machine learning (ML) venues, where research in graph-based learning has increased significantly [6]. Graph signals are well-suited to model measurements/information/data associated with (indexed by) a set where 1) the elements of the set belong to the same class (regions of the cerebral cortex, members of a social network, weather stations across a continent); 2) there exists a relation (physical or functional) of proximity, influence, or association among the different elements of that set; and 3) the strength of such a relation among the pairs of elements is not homogeneous. In some scenarios, the supporting graph is a physical, technological, social, information, or biological network where the links can be explicitly observed. In many other cases, the graph is implicit, capturing some notion of dependence or similarity across nodes, and the links must be inferred from the data themselves. As a result, GSP is a broad framework that encompasses and extends classical SP methods, tools, and algorithms to application domains of the modern technological world, including social, transportation, communication,

Digital Object Identifier 10.1109/MSP.2023.3262906  
Date of current version: 1 June 2023

and brain networks; recommender systems; financial engineering; distributed control; and learning. Although the theory and application domains of GSP continue to expand, GSP has become a technology with wide use. It is a research domain pursued by a broad community, the subject of not only many journal and conference articles, but also of textbooks [5], special issues of different journals, symposia, workshops, and special sessions at ICASSP and other SP conferences.

In this article, we provide an overview of the evolution of GSP, from its origins to the challenges ahead. The first half is devoted to reviewing the history of GSP and explaining how it gave rise to an encompassing framework that shares multiple similarities with SP, and especially digital SP (DSP). A key message is that GSP has been critical to develop novel and technically sound tools, theory, and algorithms that, by leveraging analogies with and the insights of DSP, provide new ways to analyze, process, and learn from graph signals. In the second half, we shift focus to review the impact of GSP on other disciplines. First, we look at the use of GSP in data science problems, including graph learning and graph-based deep learning. Second, we discuss the impact of GSP on applications, including neuroscience and image and video processing. We finally conclude with a brief discussion of the emerging and future directions of GSP.

## The early roots

The roots of GSP can be traced to algebraic and spectral graph theory, harmonic analysis, numerical linear algebra, and specific applications of these ideas to areas such as data representations for high-dimensional data, pattern recognition, (fast) transforms, image processing, computer graphics, statistical physics, partial differential equations, semisupervised learning (SSL), and neuroscience. Algebraic graph theory [7] dates back to the 1700s, and spectral graph theory [8] dates back to the mid-1900s. They study mathematical properties of graphs and link the graph structure to the spectrum (eigenvalues and eigenvectors) of matrices related to the graph. However, they generally did not consider potential signals that could be living on the graph.

In the late 1990s and early 2000s, graph-based methods for analyzing and processing data became more popular, independently, in a number of disciplines, including computer graphics [9], image processing [10], graphical models in Bayesian statistics [11], [12], dimensionality reduction [13], SSL [14], and neuroscience (e.g., the detailed history included in [15]). For example, in computer graphics, Taubin utilized graph Laplacian eigenvectors to perform surface smoothing by applying a low-pass graph filter to functions defined on polyhedral surfaces [9], and later used similar ideas to compress polygonal meshes. In image processing, weighted graphs can be defined with edges being a function of pixel distance and intensity differences. Such semilocal and nonlocal graphs were exploited for denoising (bilateral filtering), image smoothing, and image segmentation (e.g., in [10] and [16]). Graphical models [12]—in particular, undirected graphical models, also referred to as *Markov random fields*—model data as a family of random variables (the vertices), with

the graph edges capturing their probabilistic dependencies. Through the graph, these models sparsely encode complex probability distributions in high-dimensional spaces. Graphical models have been widely used in Bayesian statistics and Bayesian probabilistic approaches, kernel regression methods, statistical learning, and statistical mechanics [17]. We return to SSL and neuroscience and their connections with GSP in the “SSL” and “Applications to Neuroscience” sections, respectively.

Also in the late 1990s, two new models were introduced for random networks (graphs) to model the structure of complex engineered systems, going well beyond the classical Erdős–Rényi random graphs: real-world large networked systems exhibit small-world characteristics (the Watts–Strogatz model) and scale-free degree distributions (the Barabási–Albert model). This led to a flurry of activity, usually referred to as *network science*, concerned with analyzing and designing complex systems like telecommunication, power grid, and large-scale infrastructure networks [18]. Although the central focus of network science was on properties of the network and its nodes (e.g., centralities, shortest paths, and clustering coefficients), network science researchers also leveraged graphs to explore the dynamics of processes such as percolation, traffic flows, synchronization, and epidemic spread [18, Part 5], often adopting mean field approximations. For example, in the investigation of the susceptible-infected-susceptible epidemiological model in scale-free graphs in [19], each vertex can be seen as having a 0/1 (susceptible/infected) signal residing on it. Advancements in network science have certainly informed the subsequent development of GSP.

In parallel, a stream of new methods for analyzing data on graphs were investigated. These methods tried specifically to combine 1) intuition and dictionary constructions for performing computational harmonic analysis on data on Euclidean domains with 2) generalizable ways to incorporate the structure of the underlying graph into the data transforms. For example, one of the first general wavelet constructions for signals on graphs was the spatial wavelet transform of [20], which was defined directly in the vertex domain. In the seminal work of Crovella and Kolaczyk [21], diffusion wavelets were constructed by 1) creating a multiresolution of approximation spaces, each spanned by graph signals generated by diffusing a unit of energy outwards from each vertex for a fixed amount of time, and 2) computing orthogonal diffusion wavelets to serve as basis functions for the detail spaces that are the sequential orthogonal complements of the approximation spaces. Spectral graph wavelets [1] traded off the orthogonality of diffusion wavelets for a simpler generative method for each wavelet atom: define a pattern in the graph spectral domain and localize that pattern to be centered at each vertex of the graph. Meanwhile, the algebraic SP approach [22], [23] showed that classical SP can be captured by a triplet defined by a shift operator. Different shifts lead to different SP models and different Fourier transforms. In particular, it showed that a shift based on Chebyshev polynomials, appropriate for lattice models like in images, leads to standard block transforms such as the discrete cosine transform (DCT) and Karhunen–Loève transform (KLT), which can be

understood as Fourier transforms on certain graphs. Numerous other types of multiresolution transforms and dictionaries for data residing on graphs, trees, and compact manifolds were investigated in the subsequent few years. These included lifting and pyramid transforms, graph filter banks, tight spectral frames, vertex-frequency transforms that generalized the classical short-time Fourier transform, and learned dictionaries (see [24] and [25] for a more complete literature review and list of references). GSP arose from these different fields, coalescing multiple perspectives into a common framework and set of ideas. In the last decade, this unifying framework has evolved into a full-fledged theory and technology.

## The theoretical underpinnings

Ten years ago, [1], [2], [3], and [4] introduced the field of GSP and established many of its foundations. Remarkably, these works approached the problem from two different perspectives. Inspired by graph theory and harmonic analysis, the authors of [1] and [2] use the graph Laplacian as the core of their theory, naturally generalizing concepts such as frequencies and filter banks to the graph domain. Differently, the authors of [3] and [4] follow an algebraic approach, under which the multiplication of a graph signal by the adjacency matrix of the supporting graph yields the most basic operation of shift for a graph signal. Based on this simple operation, more advanced tools such as filtering, graph Fourier transforms (GFTs), graph frequency, or total variation can be generalized to the vertex and spectral graph domains. Rather than being considered competing approaches, these works brought complementary views and tools and, jointly, contributed to increasing the attention on the field. After introducing some common notations, this section reviews these two approaches and then explains how they were merged into an integrated framework that facilitated drawing links with classical SP and propelled the growth of GSP.

### Basic definitions and notational conventions

The goal in GSP is to leverage SP and graph theory tools to analyze and process signals defined over a network domain, with notable examples including technological, social, gene, brain, knowledge, financial, marketing, and blog networks. In these setups, graphs are used to both index the data and represent relations/similarities/dependencies among the locations of the data. We denote the underlying weighted graph by  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega)$ , where  $\mathcal{V} := \{1, \dots, N\}$  denotes the set of  $N$  graph vertices;  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$  denotes the set of graph edges; and  $\omega : \mathcal{E} \rightarrow \mathbb{R}$  is a weight function that assigns a real-valued weight to each edge, with a higher edge weight representing a stronger similarity or dependency between the two vertices connected by that edge. A graph with edge weights all equal to one is called *unweighted*. A graph signal contains information associated with each vertex of the graph. For simplicity, we focus our discussion on scalar, real-valued graph signals (each signal is a mapping from  $\mathcal{V}$  to  $\mathbb{R}$ ), but the values associated with each node could be discrete, complex, or even vectors (e.g., when multiple features per node are observed). Each scalar, real-valued graph signal can equivalently be represented as an

$N$ -dimensional vector  $\mathbf{x} := [x_1, \dots, x_N]^T$ , with  $x_i$  (also written sometimes as  $[\mathbf{x}]_i$ ) representing the value of the signal at vertex  $i$ . An example of a graph signal is shown in Figure 1.

To gain some insight, consider the problem of studying Twitter patterns. Assume that we have  $N$  Twitter users: each vertex  $i \in \mathcal{V}$  represents a user  $i$ , and each edge  $e = (i, j) \in \mathcal{E}$  captures that two users  $i$  and  $j$  follow each other. The data,  $x_i$ , indexed by node  $i$  could, e.g., be the number of tweets that user  $i$  tweeted in a given time interval. In a second application, to understand traffic flow in cities, we can examine the number of pickups of for-hire vehicles (e.g., taxis, Uber or Lyft cars, and so on) over a given time period. The graph  $\mathcal{G}$  can be the city road map, with the vertices  $i \in \mathcal{V}$  representing intersections, and the edges  $e \in \mathcal{E}$  representing road segments between intersections. The data  $x_i$  at each vertex  $i$  might, e.g., be the number of pickups close to that intersection over the time period of interest. The graphs  $\mathcal{G}$  in such real-world applications can be modeled as undirected (if  $(i, j) \in \mathcal{E}$ , then  $(j, i) \in \mathcal{E}$ ), or directed (e.g., to capture one-way streets).

Classical SP signals such as audio and image signals that reside on Euclidean domains can also be viewed as graph signals. Consider for instance, finite-length discrete-time 1D signals, e.g., the  $N$  vertices of the graph are the time instances  $i = 0, \dots, N - 1$ , with  $N$  being the window length. As the signal value  $x_{i+1}$  at time  $i + 1$  is usually closely related to the signal value  $x_i$  at the preceding time, there is a directed edge from vertex  $i$  to vertex  $i + 1$ . At  $i = N - 1$ , there are different options for the boundary conditions; here, we consider the periodic boundary condition, which means that the time instant “next” to the terminal instant  $N - 1$  is  $i = 0$ . The resulting “time graph” is then a directed cycle  $\mathcal{G}_{dc}$  (see Figure 2). By similar reasoning, vertices in the image graph represent the pixels, and because the image brightness or color  $x_{i,j}$  at pixel  $(i, j)$  is usually highly related to the brightness or colors of its four neighboring pixels, there are undirected edges from  $(i, j)$  to its neighboring pixels. The corresponding graph is then an undirected 2D lattice.

At the core of GSP are  $N \times N$  matrices that encode the graph’s topology. The most prominent are 1) the weighted adjacency matrix  $\mathbf{A}$ , whose  $(i, j)$ -entry is the edge weight  $\omega((i, j))$  if  $(i, j) \in \mathcal{E}$  and zero otherwise; 2) the combinatorial (or nonnormalized) graph Laplacian  $\mathbf{L} := \mathbf{D} - \mathbf{A}$ , where  $\mathbf{D} = \text{diag}(\mathbf{A}\mathbf{1})$  is the diagonal matrix of vertex degrees (sums of the weights of the edges adjacent to each vertex) and  $\mathbf{1}$  is an  $N \times 1$  vector of all ones; and 3) the normalized graph Laplacian  $\mathbf{L}_{\text{norm}} := \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ . We elaborate on the role of these matrices in the next section.

### The spectral approach for GSP

Classical Fourier analysis of a 1D signal decomposes the signal into a linear combination of complex exponential functions (continuous or discrete) at different frequencies, with increasing frequencies corresponding to higher rates of oscillation and basis functions that are less smooth. The spectral approach to GSP [1], [2] generalizes this classical Fourier analysis by writing graph signals as linear combinations of a basis of graph signals with the property that the basis vectors can be (roughly) ordered

according to how fast they oscillate across the graph, or, related, how smooth they are with respect to the underlying graph structure. By “smooth” in this context, we mean that the values of the graph signal at each pair of neighboring vertices are similar.

The operator that captures this notion of smoothness with respect to the underlying (undirected) graph is the graph Laplacian  $\mathbf{L}$ . It is a discrete difference operator as we have

$$[\mathbf{L}\mathbf{x}]_i = \sum_{j=1}^N A_{i,j}(x_i - x_j) = \sum_{j \in \mathcal{N}_i} A_{i,j}(x_i - x_j)$$

where  $\mathcal{N}_i$  is the neighborhood of node  $i$  and  $A_{i,j}$  is the  $(i, j)$ -entry of the adjacency matrix  $\mathbf{A}$ . Because  $\mathbf{L}$  is a real symmetric matrix, it has a set of orthonormal eigenvectors  $\{\mathbf{v}_\ell\}_{\ell=0}^{N-1}$  and a set of real nonnegative eigenvalues  $\{\lambda_\ell\}_{\ell=0}^{N-1}$ . Assuming a connected graph, it can further be shown that there is only one eigenvalue zero, e.g.,  $\lambda_0 = 0$ , with corresponding eigenvector  $\mathbf{v}_0 = \mathbf{1}/\sqrt{N}$ . In matrix form, we obtain  $\mathbf{L} = \mathbf{V}\text{diag}(\boldsymbol{\lambda})\mathbf{V}^\top$ , with  $\mathbf{V} = [\mathbf{v}_0, \dots, \mathbf{v}_{N-1}]$  and  $\boldsymbol{\lambda} = [\lambda_0, \dots, \lambda_{N-1}]^\top$ .

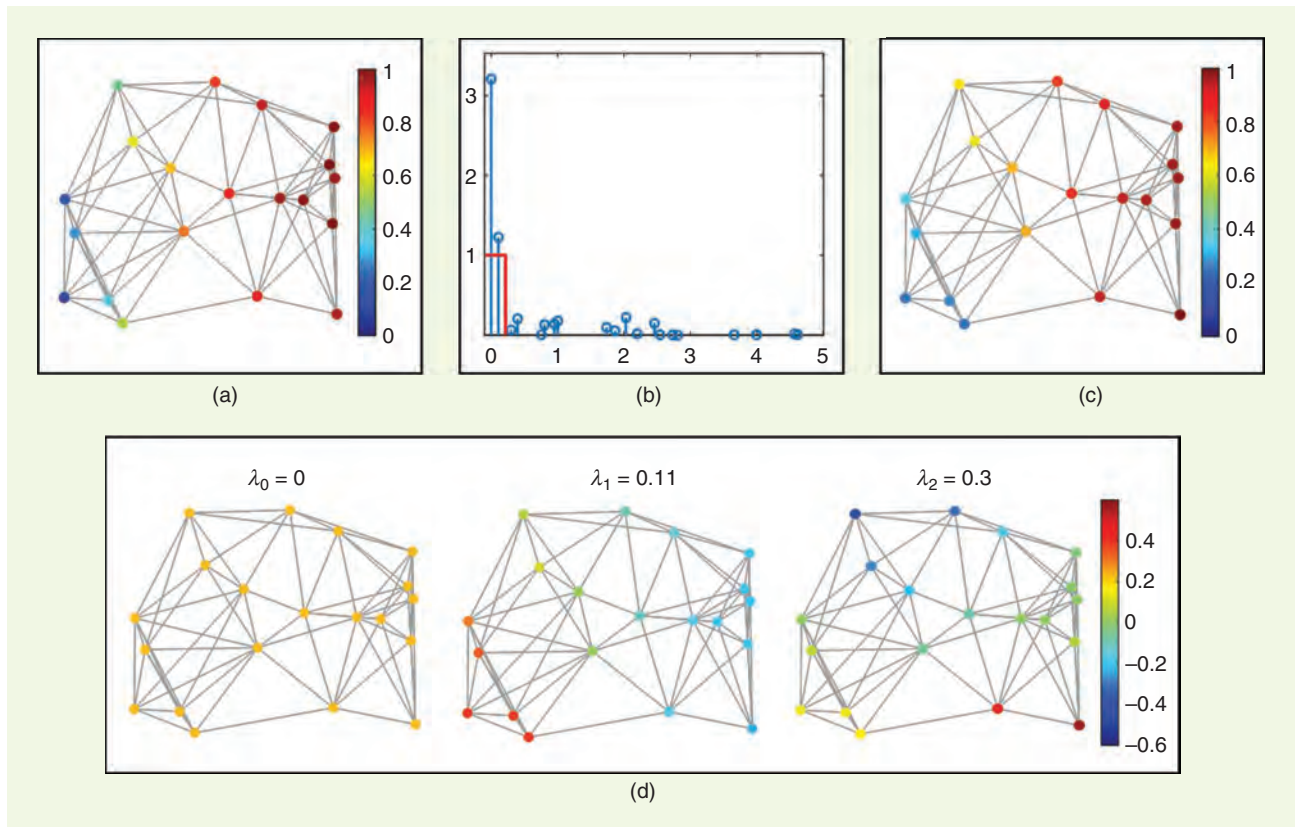
Importantly, the graph Laplacian can also be viewed as a graph extension of the time-domain Laplacian operator  $\partial^2/\partial t^2$ . Just as the 1D complex exponentials—the eigenfunctions of the time-domain Laplacian operator—capture a notion of frequency, we can interpret the graph Laplacian eigenvectors as graph frequency vectors, with the associated graph Laplacian eigenvalues capturing a notion of the rate of oscillation [2].

The Laplacian operator introduces a measure of smoothness for a graph signal  $\mathbf{x}$ , through the graph Laplacian quadratic form

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \sum_{(i,j) \in \mathcal{E}} A_{i,j}(x_i - x_j)^2 \quad (1)$$

which penalizes large differences between signal values at strongly connected vertices. Because  $\mathbf{v}_\ell^\top \mathbf{L}\mathbf{v}_\ell = \lambda_\ell$ , it is then clear from (1) that the larger the graph frequency  $\lambda_\ell$ , the less smooth (or more variable) the graph Laplacian eigenvector  $\mathbf{v}_\ell$ . So, with the indexing convention  $0 = \lambda_0 < \lambda_1 \leq \dots \leq \lambda_{N-1}$ , the graph frequency vectors  $\{\mathbf{v}_\ell\}_{\ell=0}^{N-1}$  are ordered according to increasing variability (see Figure 1). Using the Laplacian eigenvectors as the basis, we can now define a GFT as  $\mathbf{V}^\top$ . It transforms a graph signal  $\mathbf{x}$  into its frequency components as  $\hat{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$ .

Graph filters can then be interpreted as operators that modify the different frequency components of a signal  $\mathbf{x}$  individually. That is, the graph filter operation can be represented in the graph Fourier domain by  $\mathcal{H} : \mathbb{R} \rightarrow \mathbb{R}$  so that  $[\hat{\mathbf{y}}]_\ell = \mathcal{H}(\lambda_\ell)[\hat{\mathbf{x}}]_\ell$ . In most cases, the spectral function  $\mathcal{H}$  (oftentimes referred to as a *kernel*) is set to a prespecified analytical form (typically parametric) that promotes certain properties in the output signals [e.g., rectangular kernels promote smoothness and remove noise (see Figure 1)]. However, nonparametric approaches can also be used. Equally as important, Shuman et al. [2] also illustrate how graph filters can be used to interpolate missing



**FIGURE 1.** (a) An example of a graph with a color-coded graph signal on top. (b) The signal in the graph frequency domain and in red the frequency response of a potential low-pass graph filter. (c) The filtered graph signal. (d) The first three eigenvectors of the graph Laplacian ordered with decreasing smoothness (increasing eigenvalue).

values, and to design signal dictionaries whose atoms concentrate their energy around a few frequencies or vertices, highlighting their relevance in a number of applications.

### The algebraic approach for GSP

In classical SP, convolution is a key building block present in many algorithms, including filtering, sampling, and interpolation. In defining convolution and filtering, the time shift, that is, the unit delay that transforms a signal into a delayed version of itself, plays a critical role. The output of a linear time-invariant filter is a weighted linear combination of delayed versions of the input. Similarly, the discrete Fourier transform (DFT) can be understood as the transformation that diagonalizes every linear time-invariant filter and provides an alternative description for signals and filters.

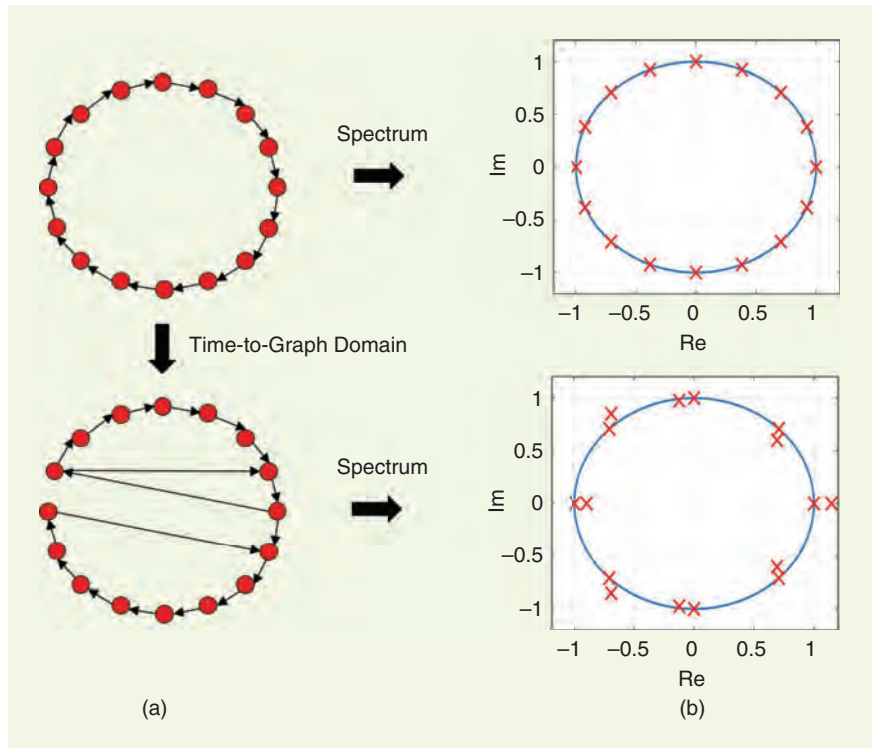
In extending these ideas to GSP, the two key contributions of [3] and [4] are 1) highlighting the relevance of defining a “graph-aware” operator that plays the role of the “most basic operation” to be performed on a signal  $\mathbf{x}$  defined over a graph  $\mathcal{G}$ ; and 2) setting this operation as  $\mathbf{A}\mathbf{x}$ , i.e., the multiplication of the graph signal  $\mathbf{x}$  by the adjacency matrix  $\mathbf{A}$  of  $\mathcal{G}$ . The motivation for the latter choice is twofold. First,  $\mathbf{A}$  is a simple (parsimonious and linear) operator that combines the values of  $\mathbf{x}$  in a manner that accounts for the local connectivity of  $\mathcal{G}$ . Second, when particularized to time-varying signals defined over the directed cycle  $\mathcal{G}_{dc}$ , using  $\mathbf{A}_{dc}\mathbf{x}$  is equivalent to the classical unit delay, i.e.,  $[\mathbf{A}_{dc}\mathbf{x}]_{i+1} = [\mathbf{x}]_i$ .

How can this basic, graph-aware operator be leveraged to design 1) linear graph filters that are applied to a graph signal to generate another graph signal and 2) linear transforms that provide an alternative representation for a graph signal? In classical SP, the basic, nontrivial operation applied to a signal is the unit delay (time shift); in other words, the simplest filter is the time-shift filter  $z^{-1}$ . Because graphs are finite, we consider DSP with finite signals, and, for simplicity, with periodic signal extensions. Generic linear filters are then polynomials of this basic operator of the form  $p(z) = p_0 + p_1z^{-1} + \dots + p_{N-1}z^{-(N-1)}$ , with  $z^{-1}$  being the consecutive application of the operator  $z^{-1}$  to a time signal  $l$  times. DSP polynomial filters are shift invariant in the sense that  $z^{-1} \cdot p(z) = p(z) \cdot z^{-1}$ .

Hence, to address the first question, [3] sets the simplest signal operation in GSP as multiplication by the adjacency matrix  $\mathbf{A}$  and, subsequently, defines graph filters as (matrix) polynomials of the form  $p(\mathbf{A}) = p_0\mathbf{I}_N + p_1\mathbf{A} + \dots + p_{N-1}\mathbf{A}^{(N-1)}$ . It is easy to see that polynomial filters are  $\mathbf{A}$  invariant, in the sense that

$\mathbf{A} \cdot p(\mathbf{A}) = p(\mathbf{A}) \cdot \mathbf{A}$ . Apart from the theoretical motivation, the polynomial definition exhibits a number of advantages. When applied to a graph signal  $\mathbf{x}$ , the operation  $\mathbf{A}\mathbf{x}$  can be understood as a local linear combination of the signal values at one-hop neighbors. Similarly,  $\mathbf{A}^2\mathbf{x}$  is a local linear combination of  $\mathbf{A}\mathbf{x}$ , reaching values that are in the two-hop neighborhood. From this point of view, a graph filter  $p(\mathbf{A})$  represented by a polynomial of order  $L$  is mixing values that are at most  $L$  hops away, with the polynomial coefficients  $\{p_l\}_{l=0}^L$  representing the strength given to each of the neighborhoods. Another advantage is that if  $\mathbf{A}$  is set to  $\mathbf{A}_{dc}$  (the graph representing the support of classical time signals), the graph polynomial definition  $p(\mathbf{A}_{dc})$  reduces to the classical time-shift definition  $p(z^{-1})$  so that graph filters become linear time-invariant filters.

To address the second question, [3] defines the GFT as the linear transform that diagonalizes these graph filters of the form  $p(\mathbf{A})$ . Letting  $\mathbf{A} = \mathbf{V}\text{diag}(\boldsymbol{\lambda})\mathbf{V}^{-1}$  be the eigendecomposition of the (possibly directed) adjacency matrix  $\mathbf{A}$ , then  $p(\mathbf{A}) = p(\mathbf{V}\text{diag}(\boldsymbol{\lambda})\mathbf{V}^{-1}) = \mathbf{V}(p(\text{diag}(\boldsymbol{\lambda})))\mathbf{V}^{-1}$  (note that we use  $\mathbf{V}^{-1}$  now instead of  $\mathbf{V}^T$  because the eigenvectors are not necessarily orthonormal as for the Laplacian). In other words, matrix polynomials can be understood as operators that transform the input by 1) multiplying it by the matrix  $\mathbf{V}^{-1}$ , 2) applying a diagonal operator  $p(\text{diag}(\boldsymbol{\lambda}))$ , and 3) transforming the result back to the vertex domain with a multiplication by  $\mathbf{V}$ . The GFT of a graph signal and the signal spectral representation is then set as the multiplication by  $\mathbf{V}^{-1}$ , and the frequency response of a filter is found by calculating  $p(\text{diag}(\boldsymbol{\lambda}))$  (similar to the spectral approach description in the previous



**FIGURE 2.** (a) From the directed cycle representing the time domain to a general graph. (b) Eigenvalues (spectrum) of the related adjacency matrices.

section). From the GFT of the signal, common SP concepts can now be defined in GSP [4], including ordering graph frequencies from low and high graph frequencies, or designing low- and high-pass graph filters. Figure 2 shows the generalization of the time domain to a more general graph domain. The applications in [3] to data prediction, graph signal compression, data classification, and customer behavior prediction for service providers, and in [4] to filter design and malfunction detection in sensor networks show the breadth of application domains.

### The benefits of a joint framework

Although having different origins, the approaches in [1] and [2], and in [3] and [4] bring complementary perspectives. The work in [1] and [2] relies on the graph Laplacian to capture the structure of  $\mathcal{G}$ , uses its eigendecomposition to characterize graph signals and define filtering operations, and draws clear links with existing graph-based techniques in a number of applications. In [3] and [4], the focus is on the shift operation in the vertex domain, postulating the use of the adjacency matrix as the building block to design GSP algorithms, and unveiling a number of similarities with classical SP. Although some early works mixed the features of [1] and [2], and of [3] and [4] (e.g., the use of polynomials based on the Laplacian matrix), the publication of these four papers and related works led to the emergence of works that combine both approaches under a common framework. One way to do so is to define a generic “graph-shift operator” (GSO) that plays a dual role: 1) it can be viewed as the most basic operation applied to a graph signal, and 2) it codifies the structure of the graph in a more generic way than  $\mathbf{L}$  or  $\mathbf{A}$  so that it can be used to tackle a broader range of setups. Under this framework, the linear GSO  $\mathbf{S} \in \mathbb{R}^{N \times N}$  has been set to different adjacency matrices (e.g., one and two hops), different graph Laplacians (e.g., combinatorial, normalized, and random walk), the precision matrix of a Gaussian–Markov random field, or even combinations of those. Based on the eigendecomposition of this operator, given by  $\mathbf{S} = \mathbf{V} \text{diag}(\boldsymbol{\lambda}) \mathbf{V}^{-1}$ , linear graph filtering can be equivalently understood as an operator that is linear and orthogonal (diagonal) in the frequency domain defined by  $\mathbf{V}^{-1}$ , or as the multiplication by a matrix that is a linear combination of successive applications (powers) of the GSO  $\mathbf{S}$ :

$$\mathbf{H}(\mathbf{S}) = \mathbf{V} \text{diag}(\hat{\mathbf{h}}) \mathbf{V}^{-1} \quad \text{or} \quad \mathbf{H}(\mathbf{S}) = \sum_{l=0}^{N-1} h_l \mathbf{S}^l \quad (2)$$

where the  $\mathbf{H}(\mathbf{S})$  notation is used to emphasize the dependence on the GSO  $\mathbf{S}$ . The first definition in (2) focuses on the frequency domain, with the filter parameters being the  $N$ -dimensional frequency response  $\hat{\mathbf{h}} = [\hat{h}_0, \dots, \hat{h}_{N-1}]^\top$ . The second definition in (2) focuses on the vertex domain, with the parameters of the filter being the  $N$  filter taps  $\mathbf{h} = [h_0, \dots, h_{N-1}]^\top$ . Although we focus on degree  $N - 1$  polynomials, thanks to the Cayley–Hamilton theorem, the definition in (2) can represent a matrix polynomial of any degree [3]. With these models at hand, the literature promptly addressed tasks such as prediction, classification, compression, filter identification, and filter design in

graph/network contexts [3], [26], [27]. The particular solution obtained for any of these tasks depends on the GSO at hand as well as the assumptions on the graph filter. For example, if the goal is to estimate the graph-based linear mapping from a set of input–output pairs collected in matrices  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M]$  and  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_M]$ , one requires  $M = N$  input–output pairs if no structure is assumed for  $\mathbf{H}$ , and a single  $M = 1$  pair if one assumes that  $\mathbf{H}$  is a graph filter. Furthermore, defining the counterparts of classical finite-impulse response (FIR) and infinite-impulse response (IIR) filters as  $\mathbf{H}_{\text{FIR}}(\mathbf{S}) = \sum_{l=0}^{L-1} b_l \mathbf{S}^l$  and  $\mathbf{H}_{\text{IIR}}(\mathbf{S}) = (\sum_{l=0}^{L-1} a_l \mathbf{S}^l)^{-1}$ , respectively, identifying such filters from input–output observations is feasible, even if only a subset (with cardinality larger than  $2L$ ) of the signal values is observed [27]. Additionally, using the definitions in (2), it is not difficult to show that any cascade/parallel/feedback connection of graph filters can also be written as a graph filter, opening the door to make and exploit connections between graph-network processes and classical tools in control.

A natural next step is to use (2) to model certain properties of classes of graph signals of interest. To be more specific, consider that we model a graph signal  $\mathbf{x} \in \mathbb{R}^N$  from a class of interest as  $\mathbf{x} = \mathbf{H}(\mathbf{S})\mathbf{z}$ , with  $\mathbf{z}$  being a hidden seed signal and  $\mathbf{H}(\mathbf{S})$  a generative graph filter that “transfers” some of the properties of  $\mathbf{S}$  to  $\mathbf{x}$ . Although mathematically simple, modeling graph signals as  $\mathbf{x} = \mathbf{H}(\mathbf{S})\mathbf{z}$  has proven to be fruitful. A typical approach is to assume some parsimonious structure on either  $\mathbf{z}$ , the filter  $\mathbf{H}(\mathbf{S})$ , or both, and then analyze the impact of those assumptions on the properties of  $\mathbf{x}$ . Standard assumptions have included  $\mathbf{H}(\mathbf{S})$  being a band-limited graph filter so that  $\mathbf{x}$  is graph-band limited [28],  $\mathbf{H}(\mathbf{S})$  being low pass so that  $\mathbf{x}$  is smooth [2], [29], [30],  $\mathbf{z}$  being a white signal so that  $\mathbf{x}$  is graph stationary [31], [32], or  $\mathbf{z}$  being sparse so that  $\mathbf{x}$  is a diffused graph signal [33], as well as combinations of those. More importantly, the combination of the generative model  $\mathbf{x} = \mathbf{H}(\mathbf{S})\mathbf{z}$  and one or more of the previous structural assumptions have been leveraged to successfully generalize a number of estimation and learning tasks to the graph domain. Early examples investigated in the literature included signal denoising, sampling and interpolation, input identification, blind deconvolution, dictionary design, SSL, classification, and the generalization of stationarity to graph domains (see, e.g., [24] for a detailed review). Although covering all of these tasks goes beyond the scope of this article, we next discuss three illustrative milestones: 1) sampling and interpolation, 2) source identification and blind deconvolution, and 3) statistical descriptions of random graph signals.

We start with a simple sampling and interpolation setup that, due to its practical relevance, received early attention from multiple research groups [34]. Consider the sampling set  $\mathcal{M} \subseteq \mathcal{V}$  with cardinality  $M \leq N$ , and define the selection matrix  $\Phi_{\mathcal{M}} \in \{0, 1\}^{M \times N}$  as the  $M$  rows of the  $N \times N$  identity matrix indexed by the set  $\mathcal{M}$ . The sampled signal  $\mathbf{x}_{\mathcal{M}} := \Phi_{\mathcal{M}} \mathbf{x}$  collects the values of the graph signal  $\mathbf{x}$  at the vertex set  $\mathcal{M}$ . The goal is to use  $\mathbf{x}_{\mathcal{M}}$ , along with  $\mathbf{S}$ , to recover  $\mathbf{x}$ , leveraging the structure of the graph. As the problem is ill-posed, we need to assume and enforce some structure on  $\mathbf{x}$ . Two widely adopted approaches



are to 1) assume that  $\mathbf{x}$  is  $K$ -band-limited, i.e., it is in the span of the first  $K$  eigenvectors of  $\mathbf{S}$ , for some  $K < N$ , or 2) assume that the signal  $\mathbf{x}$  is smooth with respect to the underlying graph, which can be generically modeled as the norm of  $\|\mathbf{x} - \mathbf{H}(\mathbf{S})\mathbf{x}\|$  being small, where  $\mathbf{H}(\mathbf{S})$  is a low-pass filter tuned to promote a particular notion of smoothness. We denote the subspace of  $K$ -band-limited signals by  $\mathcal{X}(\mathbf{V}_K) := \{\mathbf{V}_K\boldsymbol{\beta} \text{ for all } \boldsymbol{\beta} \in \mathbb{R}^K\}$ . The statement that  $\mathbf{x} \in \mathcal{X}(\mathbf{V}_K)$  is equivalent to saying that  $\mathbf{x}$  is generated via a graph filter with  $\hat{\mathbf{h}} = [\mathbf{1}_K, \mathbf{0}_{N-K}]^\top$ . These two alternative assumptions lead to the following optimization problems for interpolation, respectively:

$$\begin{aligned} \mathbf{x}^* &= \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}(\mathbf{V}_K)} \|\mathbf{x}_M - \Phi_M \mathbf{x}\|_2^2 \text{ or} \\ \mathbf{x}^* &= \operatorname{argmin}_{\mathbf{x}} \|\mathbf{x}_M - \Phi_M \mathbf{x}\|_2^2 + \alpha \|\mathbf{I} - \mathbf{H}(\mathbf{S})\|_2^2 \mathbf{x} \end{aligned} \quad (3)$$

with the weight  $\alpha$  controlling the trade-off between minimizing the smoothness of  $\mathbf{x}^*$  and how similar  $\mathbf{x}^*$  and  $\mathbf{x}$  are for the nodes in  $\mathcal{M}$ . For band-limited signals, if  $M \geq K$  and  $(\Phi_M \mathbf{V}_K)$  is full rank, the signal  $\mathbf{x}$  can be identified from its samples  $\mathbf{x}_M$  via  $\mathbf{x} = \mathbf{V}_K (\Phi_M \mathbf{V}_K)^\dagger \mathbf{x}_M$  [28]. Although this is also true for time signals, other popular results in classical SP, such as ideal low-pass filters being the optimal interpolators or regularly spaced sampling being optimal, do not hold true for the graph domain due to the lack of regularity in  $\mathcal{G}$ . Regarding the second optimization problem in (3), the solution is  $\mathbf{x}^* = (\Phi_M^\top \Phi_M + \alpha (\mathbf{I} - \mathbf{H}(\mathbf{S})\mathbf{H}^\top(\mathbf{S}))^{-1} \Phi_M^\top) \mathbf{x}_M$ . In this case, we can interpret  $\Phi_M^\top \mathbf{x}_M$  as a zero-padded graph signal that is smoothly diffused through the graph by  $(\Phi_M^\top \Phi_M + \alpha (\mathbf{I} - \mathbf{H}(\mathbf{S})\mathbf{H}^\top(\mathbf{S}))^{-1})^{-1}$ .

Using the model  $\mathbf{x} = \mathbf{H}(\mathbf{S})\mathbf{z}$ , source identification and blind deconvolution have also been generalized to the graph setting [33]. In both, the signal  $\mathbf{z}$  is assumed to be sparse. For source identification, given a sampled version of  $\mathbf{x}$ , the goal is to identify the locations and nonzero values of  $\mathbf{z}$ , which can be viewed as source nodes whose inputs are diffused throughout the network represented by  $\mathbf{S}$ . For blind deconvolution, the goal is to use  $\mathbf{x}$  to identify both the sparse input  $\mathbf{z}$  and the generating filter  $\mathbf{H}(\mathbf{S})$ , with a classical assumption being that the coefficients  $\mathbf{h}$  are sparse, or that the filter has a parsimonious FIR/IIR structure. Inspired by those works, generalizations were also investigated for demixing setups where the aggregation of multiple signals is observed (e.g., the sum of several network processes, each with different sources and dynamics).

Our last example to illustrate the benefits of a common GSP framework is the statistical description of random graph signals. Characterizing random processes is a challenging task even for regular time-varying signals, with stationarity models excelling at finding a sweet spot between practical relevance and analytical tractability. With this in mind, multiple efforts were carried out to generalize the definition of stationarity to graph signals [31], [32]. The key step was to say that a zero-mean random graph signal  $\mathbf{x}$  is stationary in a normal GSO  $\mathbf{S}$  if it can be modeled as  $\mathbf{x} = \mathbf{H}(\mathbf{S})\mathbf{z}$ , with  $\mathbf{z}$  being white. This is equivalent to saying that the covariance matrix  $\mathbf{C}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$  can be written as a polynomial of the GSO  $\mathbf{S}$ , illustrating the relationship between the underlying graph and the statisti-

cal properties of the graph signal, and establishing meaningful links with Gaussian–Markov random fields that assume  $\mathbf{S} = \mathbf{C}_x^{-1}$ . With this definition, counterparts to concepts and tools such as the power spectral density, periodogram, Wiener filter, and autoregressive moving average models were developed [31]. These developments provide new ways to design graph-based covariance estimators and denoise graph signals as well as a rigorous framework to better model, understand, and control random processes residing on a graph.

We close this section by highlighting that, although some instances of the problems discussed had been investigated well before the GSP framework was put forth (e.g., denoising based on smooth priors given by powers of the Laplacian, or source identification based on graph-diffusion processes), those early works were mostly disconnected and focused on particular setups. The advent of GSP and use of a common language and theoretical framework served a number of purposes: 1) facilitating the identification of connections between and differences among existing works, 2) bringing different research communities together, 3) enabling the design of more complex processing architectures that use early works as building blocks, 4) providing a new set of tools for graph signals based on the generalization of classical SP schemes to the graph domain, and 5) aiding the development of novel, theoretically grounded solutions to graph-based problems that had been solved in a heuristic manner.

## The impact of GSP on data science

GSP has transformed how the SP community deals with irregular geometric data; however, it has also contributed to areas that go beyond SP, having a significant impact on data science-related disciplines. To illustrate this, we next review several of the data science problems where GSP-based approaches have made significant contributions.

### Graph learning

The field of GSP was originally conceived with a given graph ( $\mathcal{G}$  or  $\mathbf{S}$ ) in mind. Such a graph could originate from a physical network, such as transportation, communication, social, or structural brain networks. However, in many applications, the graph is an implicit object that describes relationships or levels of association among the variables. In some cases, the links of those graphs can be based on expert domain knowledge (e.g., activation properties in protein-to-protein networks), but in many other cases, the graph must be inferred from the data themselves. Examples include graphs for image processing where the edges are defined based on both pixel distance and intensity differences, a  $k$ -nearest neighbor graph for SSL where edges connect data points with similar sets of features, or correlation graphs for functional brain networks. In those cases, the problem to solve can be formulated as “given a collection of  $M$  graph signals  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M] \in \mathbb{R}^{N \times M}$ , find an  $N \times N$  sparse graph matrix  $\mathbf{S}$  describing the relations among the nodes of the graph.” Clearly, such a problem is severely ill-posed, and models used to relate the properties of the graph and the signals are key to address it in a meaningful way.

Learning a graph from data is a topic on its own, with roots in statistics, network science, and ML (see [11] and references therein). Initial approaches focused on the information associated with each node separately, so that the existence of the link  $(i, j)$  in the graph was decided based only on the  $i$ th and  $j$ th row of  $\mathbf{X}$ . Contemporary (more advanced) approaches look at the problem as finding a mapping from  $\mathbf{X}$  to  $\mathbf{S}$ , with graphical lasso (GL) being the most prominent example. GL is tailored for Gaussian–Markov random fields and sets the graph to a sparsified version of the precision matrix so that  $\mathbf{S} \approx ((1/M)\mathbf{X}\mathbf{X}^T)^{-1}$  [11]. The contribution of GSP to the problem of graph learning [30], [35] falls into this second class of approaches, where the more sophisticated (spectral and/or polynomial) relationships between the signals and the graph can be fully leveraged. One cluster of early GSP works focused on learning a graph  $\mathbf{S}$  that made the signals in  $\mathbf{X}$  smooth with respect to the learned graph [29]. If smoothness is promoted using a Laplacian-based total-variation regularizer  $\sum_{m=1}^M \mathbf{x}_m^T \mathbf{L} \mathbf{x}_m$ , the formulation leads to a kernel-ridge regression problem with the pseudoinverse of  $\mathbf{L}$  as the kernel, and meaningful links with GL can be established [35]. A second set of GSP-based topology inference methods model the data  $\mathbf{X}$  as resulting from a diffusion process over the sought graph  $\mathbf{S}$  through a graph filter. The key questions when modeling the observations as  $\mathbf{x}_m = \mathbf{H}(\mathbf{S})\mathbf{z}_m$  are then the assumptions (if any) about the diffusing filter  $\mathbf{H}(\mathbf{S})$  and the input signals  $\mathbf{z}_m$ . Assuming the inputs  $\mathbf{z}_m$  to be white, which is tantamount to assuming that the signals  $\mathbf{x}_m$  are stationary in  $\mathbf{S}$ , leads to a model where the covariance (precision) matrix of the observations is a polynomial of the sought GSO  $\mathbf{S}$ , all having the same eigenvectors. This not only provides a common umbrella to several existing graph-learning methods but also a new (spectral and/or polynomial) way to address graph estimation [36], [37]. Indeed, the fact that GSP offers a well-understood framework for modeling graph signals has propelled the investigation of multiple generalizations of the aforementioned methods, tackling, e.g., directed graphs, causal structure identification, presence of hidden nodes whose signals are never observed, dynamic networks, multilayer graphs, and nonlinear models of interaction. The interested reader is referred to [30] and the references therein for more details.

### Network science

As discussed in the previous section, advancements in network science informed subsequent developments in GSP. It is now also the case that GSP techniques have been used to address network science problems such as clustering and community mining. We mention three examples here. First, in [38], spectral graph wavelets are utilized to develop a new, fast, multiscale community mining protocol. Second, by graph-spectral filtering random graph signals, feature vectors can be efficiently constructed for each vertex in a manner such that the distances between vertices based on these feature vectors resemble the distances based on standard spectral clustering feature vectors. In [39], a detailed account is provided of how that approach and other new sampling and interpolation methods

developed for GSP can be used to accelerate spectral clustering by avoiding  $k$ -means. Third, [40] uses spectral graph wavelets to learn structural embeddings that help identify vertices that have similar structural roles in the network, even though they may be distant in the graph.

### SSL

The goal of SSL is to utilize a combination of labeled and unlabeled data to predict the labels of the unlabeled data points. The labels may be discrete (semisupervised classification) or continuous (semisupervised regression). Many of the graph-based SSL methods (e.g., [14]) investigated by the ML community in the early 2000s constructed an undirected, weighted-similarity graph, with each vertex representing one data point (either labeled or unlabeled), and then diffused the known labels across the graph to infer the labels at the unlabeled vertices. This approach can also be thought of as compelling the vector of labels to be smooth with respect to the underlying graph. Mathematically, this results in optimization problems with at least two terms: a fitting term that ensures that the vector of labels exactly or approximately matches the known labels on the vertices corresponding to the labeled data points, and a regularization term of the form  $\mathbf{x}^T \mathbf{H}(\mathbf{S})\mathbf{x}$  for some (symmetric) GSO  $\mathbf{S}$  and (low-pass) graph filter  $\mathbf{H}(\mathbf{S})$  that enforces global smoothness of the signal [41] (as discussed in the “Graph Neural Networks” section).

Rather than enforcing global smoothness of the labels with respect to the underlying graph, another GSP approach to SSL is to encourage the labels to be piecewise smooth with respect to the graph by modeling them as a sparse linear combination of graph wavelet atoms [42]. Regularization problems resulting from this approach feature the same fitting term as mentioned previously, but the additional term in the objective function captures the sparsity prior through the norm (or mixed norm) of the coefficients used to synthesize the labels as a linear combination of the graph wavelets. Finally, in GSP parlance, SSL is intimately related to graph signal interpolation so that most of the results regarding the sampling and reconstruction of (band-limited) graph signals, can be (and have been) applied to SSL.

### Graph neural networks

Neural networks (NNs) are nonlinear data processing architectures composed of multiple layers, each of which combines (mixes) the inputs linearly via matrix multiplication and then applies a scalar nonlinear function to each of the entries of the output. The values of the mixing matrices  $\{\Theta_\ell\}_{\ell=1}^L$  are considered the parameters of the architecture. To avoid an excess of parameters, a standard approach is to impose some parsimonious structure on the mixing matrices (e.g., Toeplitz, low-rank, and sparse), giving rise to different families of NNs. Given the success of NNs—and convolutional NNs in particular—in processing regular data such as speech and images, a natural question is how best to generalize these architectures to data defined over irregular graph domains. In this context, the ML learning community investigated graph NNs that incorporate the graph ( $\mathcal{G}$  or  $\mathbf{S}$ ) into NN architectures in different ways [6],

[43]. GSP offers a principled way to address this question, postulating that the matrices  $\{\Theta_\ell\}_{\ell=1}^L$  have the form of a graph filter  $\{\mathbf{H}_\ell(\mathbf{S})\}_{\ell=1}^L$ . This offers both a flexible way to incorporate the graph (with the selection of the GSO  $\mathbf{S}$  being application dependent) and also provides a range of options for parameterizing the graph filter (e.g., polynomial, rational, and diffusion filters). Similarly, a number of generalizations and novel architectures that leverage GSP have been proposed, including pooling schemes based on sampling over graphs, graph-recurrent NNs, architectures defined over product graphs, and NNs based on graphon filters [44]. GSP has not only provided a common framework to better understand the contributions of and links between many of the existing works but has also facilitated novel contributions on subjects such as transferability, robustness, or sensitivity with respect to the graph [45].

### Graph-time processing

In many applications, a time series, as opposed to a scalar value, is observed at each node of the graph  $\mathcal{G}$ . If the length of each time series is  $T$ , the data at hand can be arranged in the form of a matrix  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$ , which can be viewed as a collection of  $N$  time series (one per node of the graph), a collection of  $T$  graph signals, or as a single signal  $\text{vec}(\mathbf{X}) \in \mathbb{R}^{NT}$  that varies across both time and the nodes of the graph  $\mathcal{G}$ . The first approaches to handle time-varying graph signals were based on product graphs that combine a graph of the vertices with a graph for the time domain (e.g., a directed cycle graph  $\mathcal{G}_{\text{dc}}$ ) to obtain a single larger graph  $\mathcal{G} \times \mathcal{G}_{\text{dc}}$  with  $NT$  nodes [46], [47]. This interpretation allows for the use of standard GSP tools such as the GFT transform and graph filters, with the joint GFT being the Kronecker product of the original GFT  $\mathbf{V}^{-1}$  and the DFT matrix  $\mathbf{F}^H$ , and the joint GSO some chosen product (e.g., Kronecker, Cartesian, and strong) of the respective GSOs. Indeed, the joint spectrum of the time-varying graph signal  $\text{vec}(\mathbf{X})$  can be analyzed this way, and joint, graph-time filters can be adopted for their denoising or interpolation. In their most general form, those filters need not be separable over the graph and time domains, thereby increasing their modeling and processing potential.

Later, vector autoregressive (VAR) processes were considered for graph-time processing. A VAR models a vector process by expressing the current vector as a matrix-weighted version of past vectors plus some innovation, i.e.,  $\mathbf{x}_t = \sum_{p=1}^P \mathbf{A}_p \mathbf{x}_{t-p} + \mathbf{e}_t$ . Considering that the vectors we are handling are graph signals, the underlying graph structure can be incorporated in such VAR models in different ways, leading to different GSP extensions. One direction is to replace the matrix weights by graph filters, i.e.,  $\mathbf{A}_p = \mathbf{H}_p(\mathbf{S})$ , leading to graph VAR processes [48]. In such models, the graph filter can be implemented in the graph frequency domain or as a polynomial of the GSO in the vertex domain. Furthermore, causal models have been assumed where the polynomial order of the graph filter cannot be larger than the time delay on which the filter operates [49]. Another extension of VAR models also considers the interaction between the different nodes of the current vector, i.e.,  $\mathbf{x}_t = \mathbf{A}_0 \mathbf{x}_t + \sum_{p=1}^P \mathbf{A}_p \mathbf{x}_{t-p} + \mathbf{e}_t$ , where  $\mathbf{A}_0$  has a zero diagonal.

It further enforces sparsity on all of the matrix weights. In such structural VAR processes [50], the matrix weights can be viewed as graph-adjacency matrices that link the current data on a node with past data on the same node as well as with current and past data on neighboring nodes. Extensions to non-linear versions have also been considered.

### The value of GSP in science and engineering applications

Not surprisingly, GSP methods have been applied to engineering networks where a clear definition of the graph follows from a physical network. These include communication networks (e.g., developing distributed schemes to estimate the channels), smart grids, power networks, (e.g., designing distributed resource allocation algorithms for power flow), water networks, and transportation networks (e.g., developing graph-based architectures to predict traffic delay). Similarly, GSP has also contributed to applications where the network is not explicitly observable but can be inferred from additional information, such as social networks, meteorological prediction, genetics, and financial engineering. Although all of the previous examples are meaningful and relevant, here we briefly highlight the two areas with the largest and most consistent GSP activity over the past decade: neuroscience and image and video processing.

#### Applications to neuroscience

Graphs have a long history in neuroscience because they can be used to represent different relationships and pairwise connections between regions of the brain, taking each region to be a vertex [15]. An anatomical brain graph captures structural connections between the regions, as measured, e.g., via fiber tracts in white matter captured through diffusion magnetic resonance imaging (MRI). A functional brain graph, on the other hand, aims to capture pairwise interdependencies between activity that is measured in the different brain regions. Identifying the functional brain graph has been studied extensively for different reasons and with different modalities, the most common of which is functional MRI (fMRI). Often, such studies also involve the estimation of dynamic graphs [51], [52]. During a sequence of task and rest periods, it has, for instance, been shown that on- and off-task functional brain graphs differ substantially [51]. Recent work also demonstrates that dynamics in the functional brain graph even exist during resting-state fMRI, with meaningful correlations with electroencephalograph, demographic, and behavioral data [52].

Interestingly, most of the graph-based approaches in neuroscience consist of first identifying a brain graph and then using graph-theoretical and network science tools to analyze its properties. From this point of view, GSP tools can be (and have been) leveraged for learning brain graphs [53]. However, GSP really shines when it comes to analyzing how the measured activity pattern—the brain signal—behaves in relation to a brain graph (either anatomical or functional, related to one or multiple subjects). In other words, GSP provides a technology to merge the brain function, contained in the

brain signal, with the brain graph (see [53] and references therein). Specifically, the GFT has been used to analyze cognitive behavior. For example, [54] shows that there is a relationship between the energy of the high-frequency content of an fMRI signal and the attention-switching ability of an individual. There is further research from the same group that states that, when learning a task, the correlations between the learning rate and the energies of the low-/high-frequency content of an fMRI signal change with the exposure time, i.e., they depend on how familiar we are with the task. In addition to the GFT, graph wavelets and Slepian's have been used to reveal localized frequency content in the brain [53], and graph filters have been used as diffusion operators to model disease progression in dementia. Although these results demonstrate the potential GSP has for neuroscience, we believe this pairing is still in its infancy, and that there is plenty of room for exploration.

### *Applications to image and video processing*

As noted earlier, widely used techniques in image and video processing, including transforms such as the DCT and the KLT, segmentation methods, and image filtering can be interpreted from a GSP perspective [55]. In recent years, the emergence of a broader understanding of GSP has led to a further evolution of how graph-based approaches are used for image processing. As an example, although the DCT or asymmetric discrete sine transform are formed by the eigenvectors of path graphs with equal edge weights, extensions have been proposed where graph edges with lower weights can be introduced in between pixels corresponding to image contours [56]. In these approaches, as in input-dependent image filtering [57], the image structure is first analyzed (e.g., contours detected), and then transforms adapted to the image characteristics are selected, with the choice of transform sent as side information.

A particularly promising application of GSP methods is to point cloud processing and compression. Each point in a point cloud is defined by its coordinates in 3D space and has associated with it an attribute (e.g., color or reflectance). Although points are in a Euclidean domain, their positions, on the surfaces of the objects in the scene, are irregular and make it natural to develop a graph-based processing approach. Transforms have been proposed that leverage or are closely related to the GFT of a point cloud graph [58]. These methods are fundamental algorithms for geometry-based point cloud compression. Additionally, point cloud processing has become a major application domain for graph ML, with applications in areas such as denoising [59].

### **The future ahead**

The focus of this article has been on reviewing the early results and growth of GSP, with an eye not only on the SP community but also the applications and data science problems that have benefited from GSP. We close by discussing some of the emerging directions and open problems that we believe will shape the future of the field.

One emerging area in the field of GSP is dynamic graphs; more specially, how to estimate them, and how to process time-varying graph signals residing on them. Graphs are rarely static; think, for instance, about social networks with new users or changing connections, or functional brain networks determined by a specific task that is carried out at a particular time instant. As a result, GSP tools, theory, and algorithms need to be extended to such scenarios. There is already quite some work on graph topology identification for dynamic graphs, but most of these methods link consecutive graphs in the cost function, making the problems computationally challenging [30], [50]. Adaptive methods (of the correction-only or prediction-correction type) try to tackle this issue, but tracking rates are still low. Processing signals residing on time-varying graphs have not been studied in depth, and this is clearly an area where many opportunities arise.

Extending GSP to higher-order graphs is another important future direction. Some applications are characterized by a graph domain where more than two nodes can interact; think, for instance, about a coauthorship network where groups of coauthors who collaborated on a paper are linked together, or about movie graphs in recommender systems, where movies starring the same actor form a group. Such graphs where an edge can join more than two nodes are called *higher-order graphs*. Popular abstractions of higher-order graphs are simplicial complexes and cell complexes. A simplicial/cell complex is a collection of subsets of the set of nodes satisfying certain properties. Whereas in a simplicial complex, the subsets satisfy the subset inclusion property (e.g., there needs to be links among each pair of the three coauthors of a paper), in a cell complex, they do not. However, both types of complexes share a similar recursive relationship between the higher-order Laplacians, leading to a hierarchical processing architecture that can process node signals over edges, edge signals over triangles/polygons (for a simplicial/cell complex), and so on. A less restrictive representation of a higher-order graph is a hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \omega)$ , where  $\omega$  is a function that assigns a weight to each hyperedge in  $\mathcal{E}$ . Hyperedges can connect more than two vertices in  $\mathcal{V}$ . Some recent overviews on higher-order networks, with focuses on GSP and network science, respectively, can be found in [60] and [61]. There are still many open issues in higher-order GSP, including the exploration of connections to adjacent fields such as topological data analysis and computational geometry.

Many other open problems—extending GSP to include uncertainty in the signals and graphs, design of exact and approximate Bayesian (recursive) estimators able to track variations across nodes and time, developing GSP models for categorical data, generalizing GSP results to continuous manifold (geometric) data, incorporating GSP tools into reinforcement learning and spatiotemporal control, and so on—are also expected to play important roles in the future of the discipline. If the first years of GSP combined theoretical developments with practical applications by placing a stronger focus on the former, we expect that the coming years will see an increased emphasis on applications, along with important efforts on learning and statistical schemes.

## Acknowledgments

An extended version of this article (including additional references) is available at <http://arxiv.org/abs/2303.12211>. G. Leus is partially supported by the TTW-OTP project GraSPA (project number 19497) financed by the Dutch Research Council (NWO). A. G. Marques acknowledges the support of the Spanish NSF grant PID2019-105032GB-I00/AEI/10.13039/501100011033. The work of J. M. F. Moura was partially supported by NSF Grant CCN 1513936.

## Authors

**Geert Leus** (g.j.t.leus@tudelft.nl) received his Ph.D. degree in electrical engineering from KU Leuven in 2000. He is a professor at the Delft University of Technology, 2628CD Delft, The Netherlands. He serves as chair of the EURASIP Signal Processing for Multisensor Systems Technical Area Committee and editor-in-chief of *EURASIP Signal Processing*. He received the 2021 EURASIP Individual Technical Achievement Award, a 2005 IEEE SPS Best Paper Award, and a 2002 IEEE SPS Young Author Best Paper Award. He served as a member-at-large on the Board of Governors of the IEEE Signal Processing Society, chair of the IEEE International Conference on Signal Processing and Communications Technical Committee, and editor-in-chief of *EURASIP Journal on Advances in Signal Processing*. He is a Fellow of IEEE and the European Association for Signal Processing.

**Antonio G. Marques** (antonio.garcia.marques@urjc.es) received his doctorate degree in telecommunications engineering (with highest honors) from Carlos III University of Madrid in 2007. He is a professor with the Department of Signal Theory and Communications, King Juan Carlos University, 28942 Madrid, Spain. He has received multiple paper awards and was also a recipient of the 2020 EURASIP Early Career Award. His research interests lie in the areas of signal processing, machine learning, and network science. He is a Member of IEEE, the European Association for Signal Processing, and the European Laboratory for Learning and Intelligent Systems Society.

**José M.F. Moura** (moura@andrew.cmu.edu) received his D.Sc. degree in electrical engineering and computer science from the Massachusetts Institute of Technology. He is the Philip Marsha Dowd University Professor, the Department of Electrical and Computer Engineering, Carnegie Mellon University (CMU), Pittsburgh, PA 15213 USA. His patented detector (co-inventor Alek Kavcic) is in more than 60% of computers sold in the last 18 years (4 billion). CMU settled with Marvell its infringement for US\$750 million. He was the 2019 IEEE president and CEO. He holds honorary doctorate degrees from the University of Strathclyde and Universidade de Lisboa and has received the Great Cross and Order of Infante D. Henrique. He received the 2023 IEEE Kilby Signal Processing Medal. His research interests include statistical, distributed, and graph signal processing. He is a Fellow of IEEE, the American Association for the Advancement of Science, and the National Academy of Inventors, and a mem-

ber of the Portugal Academy of Sciences and the U.S. National Academy of Engineering.

**Antonio Ortega** (antonio.ortega@sipi.usc.edu) received his Ph.D. degree in electrical engineering from Columbia University. He is Dean's Professor of Electrical and Computer Engineering, at the University of Southern California, Los Angeles, CA 90089 USA. He has received several paper awards, including the 2016 Signal Processing Magazine award. He is the author of the book, "Introduction to Graph Signal Processing," published by Cambridge University Press in 2022. He was editor-in-chief of *IEEE Transactions of Signal and Information Processing over Networks* and served on the Board of Governors of the IEEE Signal Processing Society. His recent research work focuses on graph signal processing, machine learning, and multimedia compression. He is a Fellow of IEEE and the European Association for Signal Processing.

**David I Shuman** (dshuman@olin.edu) received his Ph.D. degree in electrical engineering systems from the University of Michigan in 2010. He is a professor of data science and applied mathematics at Franklin W. Olin College of Engineering, Needham, MA 02492 USA. He is an associate editor of *IEEE Transactions on Signal Processing* and has received multiple IEEE best paper awards. His research interests include signal processing on graphs, computational harmonic analysis, and stochastic scheduling problems.

## References

- [1] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 129–150, Mar. 2011, doi: 10.1016/j.acha.2010.04.005.
- [2] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013, doi: 10.1109/MSP.2012.2235192.
- [3] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013, doi: 10.1109/TSP.2013.2238935.
- [4] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014, doi: 10.1109/TSP.2014.2321121.
- [5] L. Stanković et al., *Data Analytics on Graphs*, Boston, MA, USA: Now Publishers, 2021.
- [6] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017, doi: 10.1109/MSP.2017.2693418.
- [7] C. Godsil and G. Royle, *Algebraic Graph Theory*. Berlin, Germany: Springer-Verlag, 2001.
- [8] D. M. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs: Theory and Application*. New York, NY, USA: Academic, 1980.
- [9] G. Taubin, "A signal processing approach to fair surface design," in *Proc. 22nd Annu. Conf. Comp. Graph. Interactive Techn. (SIGGRAPH)*, 1995, pp. 351–358.
- [10] A. Elmoataz, O. Lezoray, and S. Bougleux, "Nonlocal discrete regularization on weighted graphs: A framework for image and manifold processing," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1047–1060, Jul. 2008, doi: 10.1109/TIP.2008.924284.
- [11] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York, NY, USA: Springer-Verlag, 2009.
- [12] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends® Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, Nov. 2008, doi: 10.1561/2200000001.
- [13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003, doi: 10.1162/089976603321780317.

- [14] A. Smola and R. Kondor, "Kernels and regularization on graphs," in *Proc. Conf. Learn. Theory (COLT)*, Aug. 2003, pp. 144–158, doi: 10.1007/978-3-540-45167-9\_12.
- [15] O. Sporns, *Networks of the Brain*. Cambridge, MA, USA: MIT Press, 2010.
- [16] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vision (ICCV)*, Jan. 1998, pp. 839–846, doi: 10.1109/ICCV.1998.710815.
- [17] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.
- [18] M. Newman, *Networks*. Oxford, U.K.: Oxford Univ. Press, 2018.
- [19] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, Apr. 2001, Art. no. 3200, doi: 10.1103/PhysRevLett.86.3200.
- [20] M. Crovella and E. Kolaczyk, "Graph wavelets for spatial traffic analysis," in *Proc. IEEE 22nd Annu. Joint Conf. IEEE Comput. Commun. Soc.*, 2003, pp. 1848–1857, doi: 10.1109/INFCOM.2003.1209207.
- [21] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 53–94, Jun. 2006, doi: 10.1016/j.acha.2006.04.004.
- [22] M. Püschel and J. M. F. Moura, "Algebraic signal processing theory: Foundation and 1-D time," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3572–3585, Aug. 2008, doi: 10.1109/TSP.2008.925261.
- [23] M. Püschel and J. M. F. Moura, "Algebraic signal processing theory: 1-D space," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3586–3599, Aug. 2008, doi: 10.1109/TSP.2008.925259.
- [24] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018, doi: 10.1109/JPROC.2018.2820126.
- [25] D. I. Shuman, "Localized spectral graph filter frames: A unifying framework, survey of design considerations, and numerical comparison," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 43–63, Nov. 2020, doi: 10.1109/MSP.2020.3015024.
- [26] J. S. Segarra, A. G. Marques, and A. Ribeiro, "Optimal graph-filter design and applications to distributed linear network operators," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4117–4131, Apr. 2017, doi: 10.1109/TSP.2017.2703660.
- [27] E. Isufi, A. Loukas, A. Simonetto, and G. Leus, "Autoregressive moving average graph filtering," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 274–288, Jan. 2017, doi: 10.1109/TSP.2016.2614793.
- [28] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Dec. 2015, doi: 10.1109/TSP.2015.2469645.
- [29] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, Dec. 2016, doi: 10.1109/TSP.2016.2602809.
- [30] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019, doi: 10.1109/MSP.2018.2890143.
- [31] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 5911–5926, Nov. 2017, doi: 10.1109/TSP.2017.2739099.
- [32] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3462–3477, Jul. 2017, doi: 10.1109/TSP.2017.2690388.
- [33] S. Segarra, G. Mateos, A. G. Marques, and A. Ribeiro, "Blind identification of graph filters," *IEEE Trans. Signal Process.*, vol. 65, no. 5, pp. 1146–1159, Mar. 2017, doi: 10.1109/TSP.2016.2628343.
- [34] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, Nov. 2020, doi: 10.1109/MSP.2020.3016908.
- [35] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 44–63, May 2019, doi: 10.1109/MSP.2018.2887284.
- [36] J. S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, Sep. 2017, doi: 10.1109/TSIPN.2017.2731051.
- [37] B. Pasdoloup, V. Gripon, G. Mercier, D. Pastor, and M. G. Rabbat, "Characterization and inference of graph diffusion processes from observations of stationary signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 3, pp. 481–496, Sep. 2018, doi: 10.1109/TSIPN.2017.2742940.
- [38] N. Tremblay and P. Borgnat, "Graph wavelets for multiscale community mining," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5227–5239, Oct. 2014, doi: 10.1109/TSP.2014.2345355.
- [39] N. Tremblay and A. Loukas, "Approximating spectral clustering via sampling: A review," in *Sampling Techniques for Supervised or Unsupervised Tasks*, F. Ros and S. Guillaume, Eds. Cham, Switzerland: Springer-Verlag, 2020, pp. 129–183.
- [40] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec, "Learning structural node embeddings via diffusion wavelets," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2018, pp. 1320–1329, doi: 10.1145/3219819.3220025.
- [41] D. I. Shuman, P. Vandergheynst, D. Kressner, and P. Frossard, "Distributed signal processing via Chebyshev polynomial approximation," *IEEE Trans. Signal Process. Netw.*, vol. 4, no. 4, pp. 736–751, Dec. 2018, doi: 10.1109/TSIPN.2018.2824239.
- [42] D. I. Shuman, M. Faraji, and P. Vandergheynst, "Semi-supervised learning with spectral graph wavelets," in *Proc. Int. Conf. Sampling Theory Appl. (SampTA)*, 2011.
- [43] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 3844–3852.
- [44] A. G. Marques, N. Kiyavash, J. M. F. Moura, D. V. D. Ville, and R. Willett, "Graph signal processing: Foundations and emerging directions," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 11–13, Nov. 2020, doi: 10.1109/MSP.2020.3020715.
- [45] L. Ruiz, F. Gama, and A. Ribeiro, "Graph neural networks: Architectures, stability, and transferability," *Proc. IEEE*, vol. 109, no. 5, pp. 660–682, May 2021, doi: 10.1109/JPROC.2021.3055400.
- [46] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sep. 2014, doi: 10.1109/MSP.2014.2329213.
- [47] F. Grassi, A. Loukas, N. Perraudin, and B. Ricaud, "A time-vertex signal processing framework: Scalable processing and meaningful representations for time-series on graphs," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 817–829, Feb. 2018, doi: 10.1109/TSP.2017.2775589.
- [48] E. Isufi, A. Loukas, N. Perraudin, and G. Leus, "Forecasting time series with VARMA recursions on graphs," *IEEE Trans. Signal Process.*, vol. 67, no. 18, pp. 4870–4885, Sep. 2019, doi: 10.1109/TSP.2019.2929930.
- [49] J. Mei and J. M. F. Moura, "Signal processing on graphs: Causal modeling of unstructured data," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 2077–2092, Apr. 2017, doi: 10.1109/TSP.2016.2634543.
- [50] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. IEEE*, vol. 106, no. 5, pp. 787–807, May 2018, doi: 10.1109/JPROC.2018.2804318.
- [51] R. P. Monti, P. Hellyer, D. Sharp, R. Leech, C. Anagnostopoulos, and G. Montana, "Estimating time-varying brain connectivity networks from functional MRI time series," *NeuroImage*, vol. 103, pp. 427–443, Dec. 2014, doi: 10.1016/j.neuroimage.2014.07.033.
- [52] M. G. Preti, T. A. W. Bolton, and D. Van De Ville, "The dynamic functional connectome: State-of-the-art and perspectives," *NeuroImage*, vol. 160, pp. 41–54, Oct. 2017, doi: 10.1016/j.neuroimage.2016.12.061.
- [53] W. Huang, T. A. W. Bolton, J. D. Medaglia, D. S. Bassett, A. Ribeiro, and D. Van De Ville, "A graph signal processing perspective on functional brain imaging," *Proc. IEEE*, vol. 106, no. 5, pp. 868–885, May 2018, doi: 10.1109/JPROC.2018.2798928.
- [54] J. D. Medaglia et al., "Functional alignment with anatomical networks is associated with cognitive flexibility," *Nature Human Behav.*, vol. 2, no. 2, pp. 156–164, Feb. 2018, doi: 10.1038/s41562-017-0260-9.
- [55] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph spectral image processing," *Proc. IEEE*, vol. 106, no. 5, pp. 907–930, May 2018, doi: 10.1109/JPROC.2018.2799702.
- [56] W. Hu, G. Cheung, A. Ortega, and O. C. Au, "Multiresolution graph Fourier transform for compression of piecewise smooth images," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 419–433, Jan. 2015, doi: 10.1109/TIP.2014.2378055.
- [57] P. Milanfar, "A tour of modern image filtering: New insights and methods, both practical and theoretical," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 106–128, Jan. 2013, doi: 10.1109/MSP.2011.2179329.
- [58] R. L. De Queiroz and P. A. Chou, "Compression of 3D point clouds using a region-adaptive hierarchical transform," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3947–3956, Aug. 2016, doi: 10.1109/TIP.2016.2575005.
- [59] D. Valsesia, G. Fracastoro, and E. Magli, "Deep graph-convolutional image denoising," *IEEE Trans. Image Process.*, vol. 29, pp. 8226–8237, Aug. 2020, doi: 10.1109/TIP.2020.3013166.
- [60] M. T. Schaub, Y. Zhu, J.-B. Seby, T. M. Roddenberry, and S. Segarra, "Signal processing on higher-order networks: Livin' on the edge... and beyond," *Signal Process.*, vol. 187, Oct. 2021, Art. no. 108149, doi: 10.1016/j.sigpro.2021.108149.
- [61] F. Battiston et al., "Networks beyond pairwise interactions: Structure and dynamics," *Phys. Rep.*, vol. 874, pp. 1–92, Aug. 2020, doi: 10.1016/j.physrep.2020.05.004.

Selin Aviyente , Alejandro F. Frangi , Erik Meijering , Arrate Muñoz-Barrutia ,  
Michael Liebling , Dimitri Van De Ville , Jean-Christophe Olivo-Marin ,  
Jelena Kovačević , and Michael Unser 

# From Nano to Macro

*An overview of the IEEE Bio Image and Signal Processing Technical Committee*



©SHUTTERSTOCK.COM/TRIFF

The Bio Image and Signal Processing (BISP) Technical Committee (TC) of the IEEE Signal Processing Society (SPS) promotes activities within the broad technical field of biomedical image and signal processing. Areas of interest include medical and biological imaging, digital pathology, molecular imaging, microscopy, and associated computational imaging, image analysis, and image-guided treatment, alongside physiological signal processing, computational biology, and bioinformatics.

## Introduction

BISP has 40 members and covers a wide range of Editors Information Classification Scheme, including CIS-MI: medical imaging; BIO-MIA: medical image analysis; BIO-BI: biological imaging; BIO: biomedical signal processing; BIO-BCI: brain/human-computer interfaces; and BIO-INFR: bioinformatics. BISP plays a central role in the organization of the IEEE International Symposium on Biomedical Imaging (ISBI) and contributes to the technical sessions at the ICASSP and the ICIP. In this article, we provide a brief history of the TC, review the technological and methodological contributions its community delivered, and highlight promising new directions we anticipate.

## Historical context

Until 2002, the signal processing activities related to biomedical imaging were overseen by the Image and Multidimensional Digital Signal Processing Committee of the SPS and typically presented in topical sessions at ICIP and ICASSP. The SPS also cosponsored *IEEE Transactions on Medical Imaging*. Yet, at the turn of the century, the importance of imaging in medicine and biology was becoming increasingly apparent. At the same time, advanced signal processing played an ever increasing role in the reconstruction and analysis of the vast volume of images produced. This realization was reinforced by the creation of the National Institute of Bioimaging and Bioengineering (NIBIB) by the U.S. National Institutes of Health (NIH) and U.S. Congress in December 2000

as an agency solely dedicated to the advancement of imaging technology and bioengineering. The latter was an official recognition of the crucial role of engineering in biomedical research and of the necessity to fund such research activities. This motivated the SPS and IEEE Engineering in Medicine and Biology Society (EMBS) to join forces and demonstrate leadership in biomedical imaging research.

Accordingly, it was decided to launch a new regular meeting on biomedical imaging: the ISBI (Figure 1), in close collaboration with NIBIB. The task of organizing this conference was given to Michael Unser (SPS representative) and Zhi-Pei Liang (EMBS representative). The fact that Prof. Unser had spent the larger part of his career at the NIH facilitated the interaction with NIBIB, which committed to supporting the first edition of ISBI that took place in Washington, DC, USA, in July 2002. The unique aspect of ISBI was to cover the whole spectrum and range of imaging, from nano (electron and optical microscopy) to macro (medical imaging modalities) [1].

### Creation of a dedicated TC

With the creation of ISBI and its establishment as the IEEE flagship conference in biomedical imaging, the next step was to put in place a structure to promote the conference and ensure its scientific quality. Since Prof. Unser with his team had formulated the vision for ISBI, he was instructed to form the SPS BISP TC and to make suggestions for its initial membership. In addition to its strategic role in bioimaging, BISP was given the mission to oversee the SPS activities in biomedical signal processing (e.g., the analysis of physiological signals) and bioinformatics—in short, to be responsible for all signal processing activities in medicine and biology and to maintain a liaison with its sister TC in EMBS, the Technical Community on Biomedical Imaging and Image Processing. Since the inception of the TC, BISP members have also actively participated in cross-Society activities,

such as the IEEE Life Science Technical Community and the IEEE Brain Technical Community.

### Workshops and conferences

The inaugural ISBI was held between 7 and 10 July 2002, at the Ritz-Carlton Hotel, Washington, DC. The meeting was organized jointly by the SPS and the EMBS. Significant support was provided by both the NIH and NIBIB (US\$40,000 in grants and approximately 50 paid registrants). The conference was a huge success, providing a venue for interdisciplinary exchange with researchers from both medical and biological imaging areas. It was also well attended by NIH representatives. Dr. Elias Zerhouni gave the opening address as the then newly appointed NIH director. This attracted many observers, including members of the press. Dr. Roderic Pettigrew also delivered a speech—the very first in his new function as the director of NIBIB. Both directors expressed a strong interest in the conference and commented on the need to strengthen the links between the engineering and biomedical research communities.

ISBI 2002 had two parts to the scientific program: 1) the contributed papers reviewed by the technical program committee and 2) the invited papers. Out of 355 submitted papers, 73 were accepted for oral presentation and 142 for poster presentation. The invited program consisted of 10 special sessions that were organized by leading researchers in the field.

Despite a very active and growing community, BISP membership has increased only moderately, yet the TC has always strived for a well-balanced representation across the broad range of subfields it covers. A key task for BISP members was to ensure that papers submitted to ISBI (or to dedicated tracks at ICASSP and ICIP) would benefit from the availability of a highly qualified pool of reviewers and editors. BISP members also participated in the many activities related to ISBI as members of the organizing committee. Since 2006, ISBI has been held regularly as an annual four-day conference. Figure 2 summarizes with a word cloud the keywords from the keynote titles since ISBI's inception. Outstanding clinical and technical speakers delivered their visions for the field, relevant trends, or challenges ahead. Among our distinguished speakers, there were Nobel Prize winners and top NIH officers.

In 2022, ISBI was held for the first time as a fully hybrid conference in Kolkata, India, with every session having both physical and online speakers and audiences. Out of 785 submitted papers, 309 were accepted. In addition to the regular paper sessions, there were five special sessions, five plenary talks, six challenges, and six tutorials. In addition to ISBI, BISP has been an active contributor to ICASSP since 2006, with the number of submitted papers increasing from 100 in 2006 to 222 in 2020.

### Biomedical image and signal acquisition across scales

Biomedical data come in many shapes and forms. BISP focuses on digital images and signals, which can be automatically processed and analyzed by advanced computational methods.



**FIGURE 1.** The original ISBI logo designed by Annette Unser (graphic artist and sister of the founding chair), with the eye projecting a distinctive vision for ISBI. Observers have suggested that the central motif illustrates the Fourier slice theorem, or for the more pessimistic ones, the typical artifacts of the filtered back-projection reconstruction algorithm.



To study biological processes in health and disease, many images and signal acquisition techniques have been developed in the past century, reflecting the fact that biological phenomena occur at different spatial and temporal scales (Figure 3). Before discussing methodological advances, we survey some of the most prominent modalities for molecular and cellular imaging, tissue imaging, anatomical and functional medical imaging, neuroimaging, physiological signal recording, and several data types in bioinformatics.

### Molecular and cellular imaging

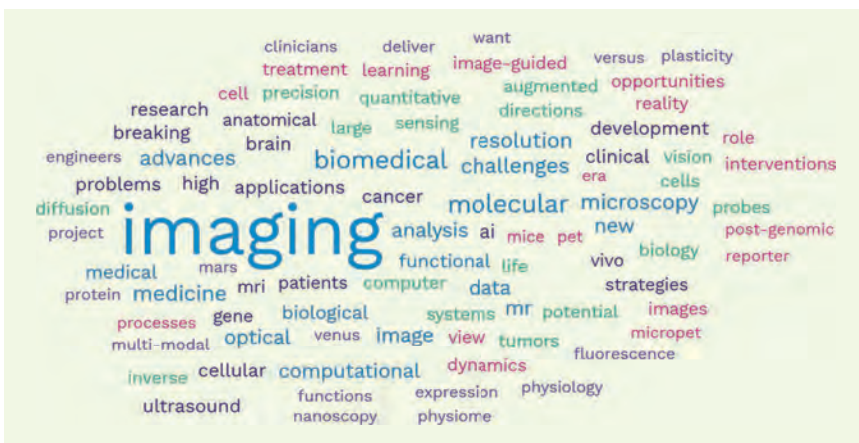
Molecular and cellular imaging has undergone multiple revolutions in the past three decades, moving from a mainly qualitative to a mostly quantitative field thanks to advances in molecular probes as well as imaging modalities [2], [3], [4]. With the advent of the green fluorescent protein, pioneered by Osamu Shimomura, Martin Chalfie, and Roger Y. Tsien (Nobel Prize in Chemistry, 2008), microscopy has become one of the key tools in biological research [3]. Fluorescence microscopy became a fast-growing field to study (quantitatively and often within a high-throughput content setup) processes and organelles within living cells and organisms. More recently, a vast leap has been made with the invention of superresolution techniques, based on seminal work by Eric Betzig, Stefan W. Hell, and William E. Moerner (Nobel Prize in Chemistry, 2014).

Another recent development is selective plane illumination (light sheet) microscopy, which allows long-term biological studies of living organisms with rapid acquisition, high resolution, and minimal phototoxicity. Classical image and signal processing methods

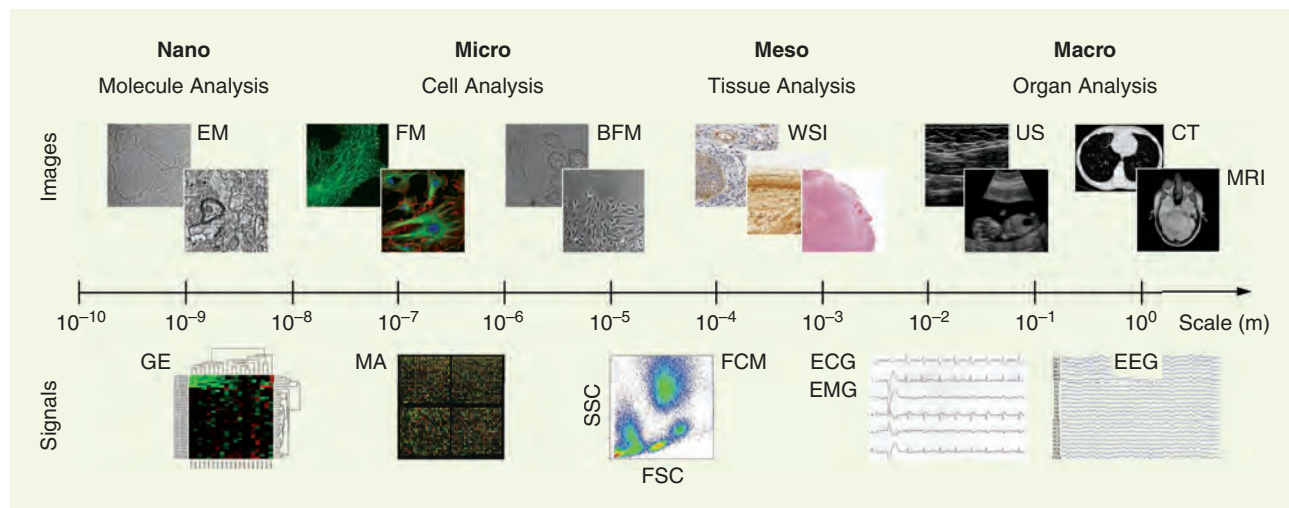
as well as modern deep learning-based methods are increasingly used not only for reconstruction and deconvolution of the data produced by advanced microscopy imaging modalities but also for enabling downstream tasks such as segmentation, classification, tracing, and tracking [4], [5], [6]. Fluorescence microscopy has enabled the study of dynamic processes within cells and complements structural and static imaging modalities, such as scanning probe microscopy, electron microscopy (Nobel Prize in Physics, 1986), and cryo-electron microscopy (Nobel Prize in Chemistry, 2017), which have become part of the vast arsenal of tools for the life sciences. In recent years, several new journals or sections in established publications, e.g., *Biological Imaging*, *Frontiers in Bioinformatics*, and *Cell Reports Methods*, have been launched to host the increasing number of publications in this domain.

### Tissue imaging

Microscopy can also be employed to study biological phenomena at the tissue level. Rather than imaging individual cells



**FIGURE 2.** A word cloud with the most frequent keywords from the titles of all the ISBI keynotes in the past 20 years. Clinicians, Nobel Prize winners, and NIH officials have given these talks, among other contributors.



**FIGURE 3.** Examples of the many data acquisition modalities in biomedical image and signal processing operating at various spatial scales. BFM, bright-field microscopy; CT, computed tomography; ECG, electrocardiography; EEG, electroencephalography; EM: electromagnetic; EMG, electromyography; FCM, flow cytometry; FM, fluorescence microscopy; GE, gene expression; MA, microarray; MRI, magnetic resonance imaging; US, ultrasound; WSI, whole-slide imaging.

(molecular/cellular imaging) for basic research or anatomical structures and entire organs (medical imaging) for clinical diagnostics, the imaging of tissue slides prepared from biopsies enables characterizing and grading disease processes *ex vivo* as revealed by abnormal cell arrangements and tissue architectures. This is especially important in researching and diagnosing pathologies, notably the many types of cancer (the field of oncology), known to manifest themselves first at the cell and tissue level (histopathology). Recent advances in digital whole-slide imaging (WSI) systems (sometimes referred to as *virtual microscopy*) have created unprecedented opportunities for computer-aided diagnosis in histopathology. Both image and signal processing play a prominent role in histopathological image analysis, especially for breast cancer, prostate cancer, lung cancer, tumor pathology in many other forms of cancer, and cancer prognosis.

Review papers have summarized and commented on the challenges and opportunities in this domain [7], [8]. Typical tasks include the detection and segmentation of cell nuclei, glands, and lymphocytes and computing various quantitative morphological features for classification. This, in turn, requires effective techniques for image normalization as well as feature selection and dimensionality reduction. Analysis of the spatial arrangements of tissues is often facilitated by graph-based representation and topological modeling. The challenges in histopathological image analysis are not only due to the high complexities of the image structures but also to the typically large image sizes, on the order of tens of thousands by tens of thousands of pixels, at multiple magnifications.

Traditionally, tissue classification has been performed using handcrafted features and machine learning methods, such as support vector machines and random forests. Still, there is now growing evidence that deep artificial neural networks (NNs) provide fast and reliable image analysis on a par with seasoned pathologists and can serve as a synergistic tool for the latter to improve accuracy and throughput. However, the full adoption of deep learning methods in pathology is hindered by the lack of large and reliably annotated image cohorts documenting the large diversity of diseases and the high variability of disease traits, calling for efficient automated annotation methods.

### *Medical imaging*

*Medical imaging* refers to the imaging techniques and processes to gain insights into the interior of a body for clinical diagnosis or medical intervention as well as visual representations of the function of organs or tissues. Medical imaging can be divided into structural or anatomical imaging and functional or physiological imaging. Many medical imaging techniques have been invented since the discovery of X-rays by Wilhelm Conrad Röntgen in 1895, which form the basis of projection X-rays and computed tomography (CT). In 1946, Bloch and Purcell (Nobel Prize in Physics, 1952) independently discovered nuclear magnetic resonance, which formed the basis of MRI (Paul Lauterbur and Sir Peter Man-

sfield in the 1970s, Nobel Prize in Physiology or Medicine, 2003). MRI developed into a platform technology with many specialized techniques, providing insights into anatomy, perfusion, diffusion, and deformation. Ultrasound (US) was used in medicine since World War II, but it was not until the late 1970s that US imaging was popularized as a clinical imaging modality.

In 1963, David Kuhl and Roy Edwards introduced emission reconstruction tomography, a method that later became single-photon emission CT (SPECT). Sir Godfrey Hounsfield (Nobel Prize in Physiology or Medicine, 1979) developed the first prototype of a CT scanner in 1963, thanks to the availability of modern computers to solve the complex image reconstruction problems that ensued. Michael Hoffman, Michel Ter-Pogossian, and Michael E. Phelps built the first positron emission tomography (PET) camera in 1974. Underpinning these techniques, there are considerable signal and image processing problems, ranging from image reconstruction, image deconvolution, image denoising and restoration, image transformation, and multimodal image coregistration. Over the last three to four decades, several signal processing developments had major impacts on medical imaging. For example, wavelets and splines played a major role in medical image interpolation, denoising, and filtering. Mutual information and other information-theoretic metrics revolutionized multimodal image registration. Compressed sensing (CS) provided a novel approach to find solutions to underdetermined linear equations with a major impact on image reconstruction from projections (CT), from *k*-space (MRI), or from sensors (US), and, of course, the latest developments of deep learning and their impact across the board.

### *Neuroimaging: From images to connectomes*

Another flourishing outlet for biomedical signal and image processing has been the processing of structural and functional MRI (fMRI) data [9]. The concept of establishing connectivity between brain regions has been fundamental in many emerging methodologies [10] (Figure 4). Structural connectivity is defined by the strength of interregional axonal fiber pathways that can be revealed using tractography methods applied to diffusion-weighted MRI (dMRI) data. Functional connectivity relates to the statistical interdependency between two time series of blood oxygenation level-dependent (BOLD) activity. When established for all possible pairs of regions, defined by a brain atlas, this leads to the structural and functional connectomes, respectively. While hemodynamic imaging has been the most commonly used modality for constructing the functional connectome, neurophysiological signals such as magneto/electroencephalography (M/EEG) have also been adopted thanks to their high temporal resolution.

Both model-based and data-driven methods have been developed to quantify functional connectivity, including multivariate autoregressive models, graphical models, phase synchrony, and information-theoretic metrics. Functional connectivity is also intimately related to blind source separation. This, in turn, relies on decomposing the data matrix into components driven

by maximizing covariance [using techniques such as principal component analysis, singular value decomposition (SVD), or higher order SVD] or statistical independence (using independent component analysis), which has become part of the pipeline in fMRI software suites. During the past decade, dynamic functional connectivity has been introduced to acknowledge the changing patterns of co-fluctuations, either by sliding-window functional connectivity, instantaneous activation patterns, or autoregressive models.

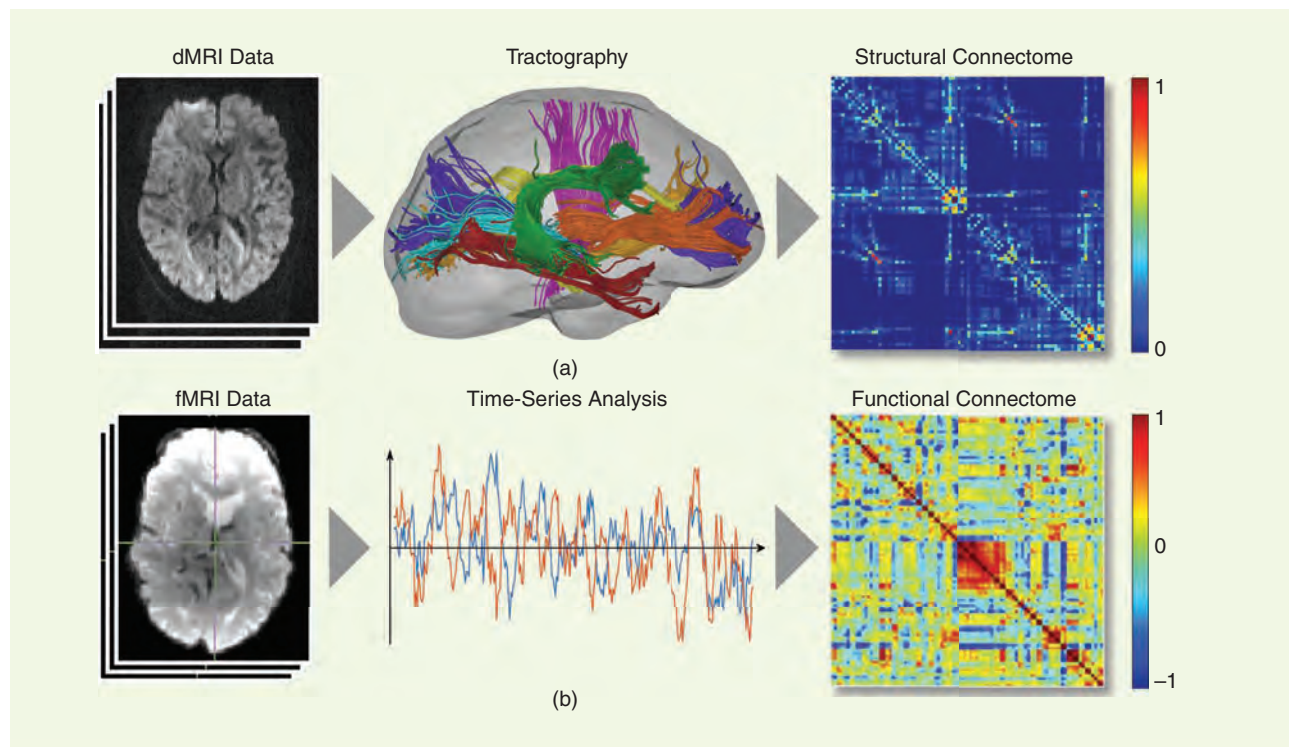
The resulting connectomes are commonly represented by graphs and analyzed to reveal organizational principles using, for instance, local clustering coefficients, efficiency, small-worldness, centrality, and phenotype behavior and disorder. These approaches have been applied to other species' connectomes obtained using different modalities, leading to a new field, network neuroscience, which further branched out to machine learning, information theory, and computational neuroscience. Finally, the emergence of graph signal processing has found its way into the neuroimaging field [11], providing a way to combine brain structure (i.e., a graph defined by the structural connectome) and brain function (i.e., graph signals obtained by fMRI snapshots of brain activity).

#### Physiological signal processing (M/EEG, electromyography, and electrocardiography)

With the advent of wearable sensors, physiological signals are collected for various applications and are central to

multiple new technologies, including brain-computer interfacing/human-machine interfacing (HMI), neurorehabilitation, and neuroprosthetics, in addition to medical diagnosis and monitoring [12]. The most employed physiological signals include electromyography (EMG), respiration, speech, heart rate variability, photoplethysmography (PPG), electrocardiography (ECG), and M/EEG. Some prominent applications that routinely rely on physiological signals include emotion recognition, autonomous driving, mental health, and assistive technologies. For example, physiological changes such as heart rate, skin conductance, and PPG signals are monitored for measuring human emotions as they are more reliable and harder to alter compared to explicit behaviors such as facial expressions and speech.

Similarly, the design of HMI systems requires the consistent and accurate decoding of motor intent with minimal training and calibration. The multimodal high-density sensing technology coupled with the nonstationary and nonlinear nature of biological signals requires the development of innovative signal processing and machine learning techniques to process, decompose, and decode these signals. Some methodologies employed in this area of research include blind source separation, time-frequency analysis, multimodal data fusion, supervised (or semisupervised) learning, and deep learning. Different tasks, such as event detection, prediction, and diagnosis, have been addressed using these tools.



**FIGURE 4.** Structural and functional connectomes play a key role in representing relationships between brain regions. (a) From diffusion-weighted MRI (dMRI), the orientation of axonal bundles in white matter can be extracted and processed by tractography to obtain the strength of structural connectivity between all pairs of regions. (b) Functional MRI (fMRI) provides a series of volumes where the blood oxygenation level-dependent (BOLD) signal is related to neuronal activity. Time series analysis exists in large diversity, but the functional connectome that reflects the statistical interdependencies between pairs of time series is one of them.

## Bioinformatics

Since the turn of the century, major advances in molecular biology, along with advances in genomic data acquisition technologies, led to the growth of biological data generated and shared by the scientific community; e.g., The Cancer Genome Atlas (TCGA). These data bring with themselves significant challenges in the identification of gene expression mechanisms; the determination of proteins encoded by the genes; understanding how these interact, i.e., gene regulatory networks; and marker identification.

BISP has contributed to this area by introducing a new line of research: genomic and proteomic signal processing [13]. While biomolecular sequence analysis has been addressed by computer scientists, physicists, and mathematicians, it was only at the turn of the century that signal processing started to play a role in this area. Genomic and proteomic data can be modeled as noisy, continuous, or discrete signals that represent the molecular structure and activities in cells. The high dimensionality, variability, and complexity of these data require the development of new signal processing methodologies that effectively deal with these challenges. By mapping the character strings corresponding to gene sequences into numerical sequences, signal processing offers a set of tools for solving highly relevant problems. For example, the magnitude and the phase of properly defined Fourier transforms can be used to predict important properties of protein-coding regions in DNA.

Similarly, concepts from digital filtering can be employed to analyze the mapping of DNA into proteins and the interdependence of two sequences. These and other signal processing methodologies, such as frequency domain analysis, high-dimensional data analysis, CS, and network inference, have played important roles in the advancement of this field. Genomic and proteomic signal processing both have had a substantial impact on different application areas, including sequence analysis, microarray analysis, structure identification, and regulatory networks.

From 2005 to 2013, the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSiPS) was organized annually and sponsored by the SPS. These workshops covered topics related to high-dimensional genomic data analysis, gene regulatory network inference, marker identification, drug screening, and proteomics.

## Methodological advances in biomedical image and signal processing

The field of biomedical image and signal processing has seen major methodological advances not only in how data are recorded, stored, and transmitted but also in how they are best represented, processed, analyzed, and modeled, depending on the application domain. Many paradigms have been proposed in recent decades by various schools of thought, resulting in a wide range of theories and methods for challenging problems, such as image and signal restoration, reconstruction, detection, segmentation, classification, pattern recognition, and statistical analysis, as documented in numerous textbooks and reviews. Given the limited space in this article, we only briefly

discuss some of the most impactful developments in recent years, including methods for computational imaging and deep learning-based image and signal analysis and efforts to stimulate reproducible research.

### Biomedical computational imaging

Most biomedical imaging modalities have a strong computational component as they systematically rely on signal processing to reconstruct the images from the raw imaging data. The data can take the form of 1) 2D projections of a 3D object, as in X-ray tomography, PET, and cryo-electron microscopy; 2) a series of blurred 2D slices of a specimen, as in fluorescence microscopy; or 3) samples of the Fourier transform of an object, as in MRI and optical diffraction tomography. By capitalizing on the knowledge of the imaging physics (linear forward model), the reconstruction task can then be formulated as an inverse problem. Until recently, classical imaging (MRI and CT) relied on a direct inversion of this forward model. This is achieved, for instance, by inverse Fourier transformation in MRI (with uniform sampling in  $k$ -space) or by inverse radon transformation (the celebrated filtered back projection algorithm) in CT. This works well when the measurements are sufficiently numerous and diverse and when the noise is negligible.

Besides the streamlining of the reconstruction process itself (improved nonuniform fast Fourier transform, optimization of sampling parameters, etc.), the earlier involvement of the signal processing community was to combat the effect of noise with the help of advanced statistical methods. One notable example of such success is the method of ordered subsets in PET and SPECT [15]. Another fruitful approach inspired by Wiener filtering is to inject prior information in a stochastic model (e.g., generalized Gaussian in a transformed domain), which makes a direct link between maximum a posteriori (MAP) reconstruction and regularization/energy minimization techniques [16].

The more significant revolution in imaging came with CS with theorists [17], [18] and then experimentalists [19], [20] showing the feasibility of image reconstruction from a reduced set of measurements. A milestone in this line of research was the development of efficient minimization methods under sparsity constraints, in particular the (fast) iterative soft thresholding algorithm and alternating direction method of multipliers [21]. The main benefit of CS is to enable faster imaging, which reduces not only cost but also radiation exposure (in the case of X-ray or PET/SPECT). This has led to a major revolution in MRI, with fast (CS-based) imaging protocols now offered by most vendors of MRI technology. While CS kept SPS researchers busy from 2005 to 2017, another wave then overtook the field—the incorporation of NNs in the image reconstruction pipeline. This led to further significant improvement in image quality (Figure 5), especially in extreme scenarios, e.g., low signal-to-noise ratio (SNR) and CS [14].

While image reconstruction based on convolutional NNs (CNNs) still has shortcomings—CNNs are poorly understood and can behave erratically (lack of stability and hallucination)—they demonstrate the potential for better reconstruction

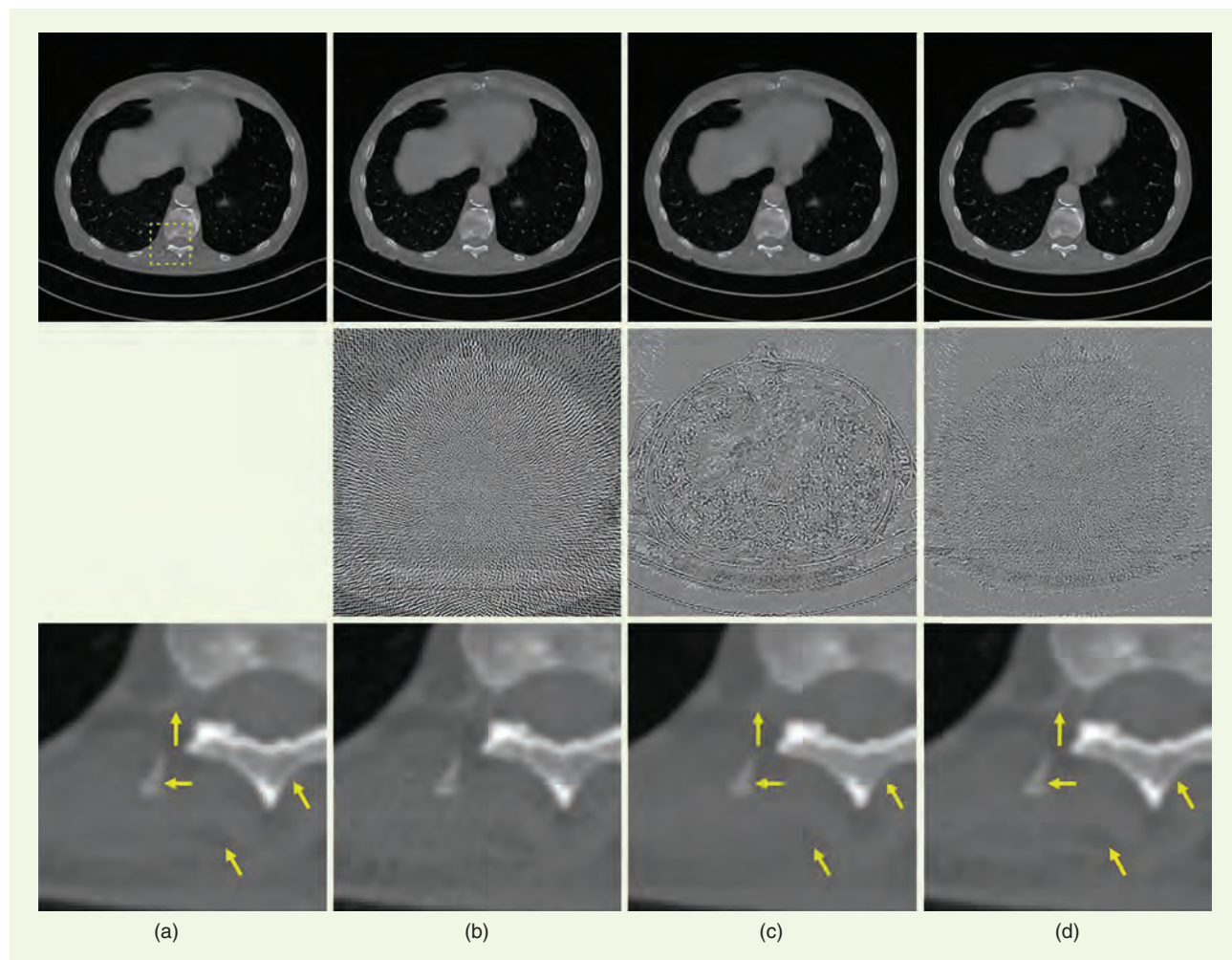
quality [22]. It is noteworthy that it took signal processing pioneers less than a year to tune their new CNN-based methods to the point where they would outperform sparsity-based methods for CS in public imaging challenges by the same margin (typically  $>4$  dB) that the latter had achieved over classical reconstruction during a whole decade of intense research activity. CNNs and learning-based techniques are presently at the center of attention of the research community. Recent trends include the development of more sophisticated iterative reconstruction schemes that rely on CNNs to regularize the solution—as enabled by the plug-and-play framework [23]—as well as the use of deep learning for the resolution of more challenging nonlinear inverse problems such as diffuse optical tomography [24] and diffraction tomography.

### Deep learning in biomedical image and signal processing

Traditionally, methods for image and signal processing have been based on carefully designed mathematical models of the

phenomena and anomalies of interest and their translation into efficient rules-based computational algorithms. Illustrative examples of this are mathematical point-spread function models of widefield or confocal microscopes based on physical (optical) principles, serving as the basis for various image restoration methods (in particular, deconvolution) [25] and object detection methods (such as single-molecule localization) [26]. However, as in many other fields, the demand for new and better methods from practitioners in biology and medical diagnostics outstrips the supply of researchers and developers in image and signal processing. That is, there are many more biologists and physicians in the world looking for tools to facilitate their data processing workflows than there are scientists and engineers looking to develop mathematical models and image/signal processing algorithms specifically for biomedical applications.

Moreover, especially in the biomedical field, many image and signal analysis tasks are notoriously difficult to model mathematically due to the complex nature of the problem,



**FIGURE 5.** A comparison of tomographic reconstruction algorithms for CS with a reduction of the number of views by seven. (a) Ground truth (high-quality reconstruction from 1,000 views). (b) Conventional reconstruction (filtered back projection) from a subset of 143 views. (c) CS reconstruction using total variation regularization. (d) CS reconstruction using a CNN (FBPConvNet). The middle panel displays the image residuals with the same contrast. The magnified images in the lower panel represent the corresponding region of interest overlaid in (a). (a) Ground truth. (b) Signal-to-noise ratio (SNR) = 24.06 dB. (c) SNR = 29.06 dB. (d) SNR = 35.38 dB. (Source: The figure is adapted from [14].)

the high ambiguity of the data, and the subjectivity of human experts who define the gold standards for interpreting the data. Thus, as imaging and measurement devices improved over the years and the number of potentially automated data processing tasks grew, the need for more generic, data-driven, and learning-based methods also increased.

In the past decade, the rapidly growing availability of large datasets, powerful computing capabilities, and open-access software libraries and frameworks has accelerated the development and adoption of machine learning and deep learning methods in biomedical image and signal processing [6], [27], [28], [29]. These methods show increasingly superior performance in benchmarking studies for various tasks, including reconstruction, restoration, detection, segmentation, classification, and tracking. In particular, deep learning of artificial NNs has become a popular approach for solving data analysis problems where multimodal, multi-dimensional, and multiparametric datasets need to be jointly processed, posing a clear challenge to traditional analysis methods. For the processing of biomedical images, CNNs in particular have become mainstream, a prominent example of this being the U-Net architecture [30], of which many variants exist for various tasks and applications, such as segmentation (Figure 6).

For biomedical signal processing, especially for dealing with time series, recurrent NNs such as the long short-term memory unit have seen widespread adoption. However, despite promising results, many challenges remain to be addressed

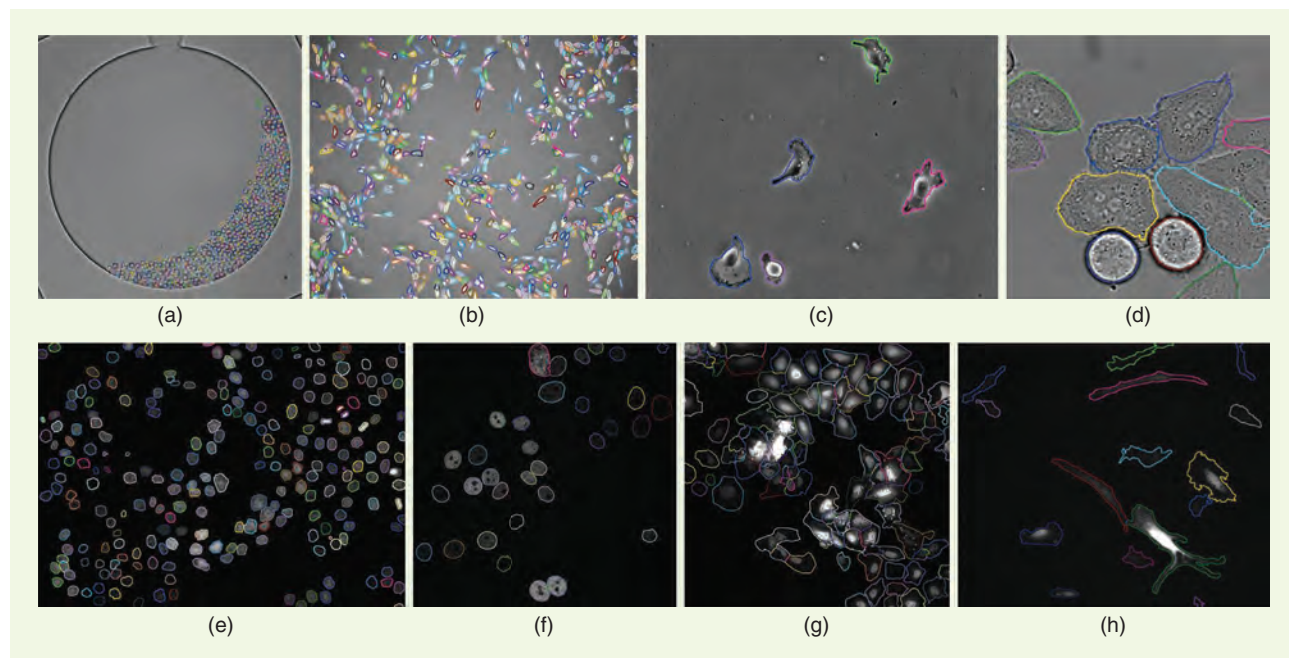
before deep learning solutions can be integrated with full confidence and accountability into the workflows of biomedical practitioners, such as developing ways to incorporate expert knowledge and improving the explainability and generalizability of the models (see the discussion of future directions in the last section).

### Reproducible research, open access, and code

Reproducing the results presented in a research work can be very challenging. For a computational algorithm, details such as the exact dataset, initialization or termination procedures, and precise parameter values are often omitted in the publication for various reasons. This makes it difficult, if not impossible, for someone else to obtain the same results [32]. In the early 2000s, the need to boost research by implementing reproducible research practices became apparent. Vandewalle et al. [32] published a seminal manuscript in *IEEE Signal Processing Magazine* in 2009, which defines reproducible research as follows:

“A research work is called reproducible if all information relevant to the work, including, but not limited to text, data, and code, is made available, so that an independent researcher can reproduce the results.”

The authors also distinguish six levels of reproducibility, from Level 5 (an independent researcher can easily reproduce results with at most 15 min of user effort, requiring only standard freely available tools—C compiler, etc.) to Level 0 (an independent researcher cannot reproduce results).



**FIGURE 6.** Examples of cell segmentation using deep NNs in diverse types of microscopy images. From (a) to (h), the images were captured using bright-field microscopy, phase-contrast microscopy (2 $\times$ ), differential interference contrast microscopy, and fluorescence microscopy (4 $\times$ ) and contain distinct types of cells in different spatial arrangements (densities and confluences). The segmentation results are the overlaid colored cell contours (arbitrary colors). These results were produced using a single deep learning framework with a U-Net-like macro-architecture consisting of various layers/blocks whose microarchitectures were optimized automatically using a neural architecture search approach [31]. The examples illustrate the power of deep learning and the level of automation that can be achieved nowadays in optimizing image segmentation results without requiring expert user input, other than manual annotations, to learn from.

The issue of reproducibility has been raised many times in the past few decades. In the 1980s, there was a growing awareness of poor building on the previous work of others [33]. Published algorithms were frequently evaluated with only select data and ad hoc metrics. The comparison of algorithms and software performance was difficult. In the 1990s, the first benchmark initiative in biomedical imaging, the Retrospective Image Registration Experiment, appeared at Vanderbilt University [34]. We needed to wait until the early 2010s for bioimaging reference datasets and challenges (benchmarks associated with competitions) to appear. ISBI 2012 in Barcelona was the first edition of the symposium to hold challenges on the following topics:

- 1) particle tracking [35]
- 2) segmentation of neuronal structures in electromagnetic (EM) stacks
- 3) vessel segmentation in the lung [36]
- 4) cardiac delayed-enhancement magnetic resonance image segmentation
- 5) high angular resolution diffusion imaging
- 6) Challenge US: Biometric Measurements from Fetal Ultrasound Images.

At ISBI 2015, Prof. Ronneberger's team won the Cell Tracking Challenge (third edition) [37] and the dental X-ray image segmentation challenge with their U-Net [30]. Figure 7

shows a word cloud of the challenge titles over the years; detection, images, and tracking occupy a prominent place in the cloud.

In the bioimaging community, the push for reproducibility led to several open software platforms, such as Cell Profiler (<https://cellprofiler.org/>), Fiji (<https://fiji.sc/>), and Icy (<https://icy.bioimageanalysis.org/>). They were made available to the community in the early 2000s to share the then state-of-the-art analytical methods, which are now used for integrated deep learning framework deployment [38].

Finally, imaging challenges foster collaboration between institutions and continents. Since 2012, more than 60 challenges have been organized, led by multiple institutions. Of these, 31 were organized by European institutions, 12 by organizations in the Americas, 10 by Asia or Oceania, and six involved cross-continental collaboration from the Americas, Asia, and Europe. These collaborations drove our community to learn from the strengths and pitfalls [39] in organizing challenges and interpreting their results [40] and thus developed best-practice guidelines for transparent reporting [41].

### Future directions

Advanced technologies for capturing biomedical images and signals have made a growing and lasting positive impact on clinical diagnostics and therapeutics, medical research,



**FIGURE 7.** The ISBI is the premier scientific venue for the BISP TC. Since 2012, our community has organized more than 60 challenges, where open datasets, well-specified tasks, and evaluation metrics have been made available for multiple groups to participate, compete, and learn from each other. Challenges have covered many imaging modalities and scales, image computing tasks, and organ systems.

and life sciences. They will continue to help improve our understanding of the conditions underlying human health and how to prevent and treat disease. Modern biomedical image and signal acquisition systems are based on a wide range of physical phenomena (electricity, magnetism, light, sound, force, etc.) capable of providing complementary information about the anatomical and functional properties of the human body and living organisms in general. Also, the sensitivity, resolution, and quality of these systems have improved dramatically over the years to the point where automated image and signal processing are now indispensable in virtually all clinical and biomedical research applications.

At this point in time, unlike in the past century, the chances of discovering totally new physical principles that could ultimately be used in biomedical practice have diminished, yet the challenges of fully exploiting existing technologies are far from having been solved. One of the main problems for the image and signal processing community in the years ahead will be to develop effective methods for data fusion and integration

[42] to maximize the potential of multimodal and correlative imaging as well as combining imaging and nonimaging (e.g., “omics”) data. This requires finding solutions to dealing with the fundamentally different nature of different data sources and the inevitable imbalances in the data but also with the huge volumes (terabytes and no doubt soon petabytes) of multimodal datasets.

Despite being comparatively young, the BISP community has already seen and contributed to major paradigm shifts in biomedical image and signal processing. Still, in addition to the data challenges mentioned previously, many fundamental technical challenges remain. Examples include some of the problems caused by the increasing emphasis on learning-based approaches. For starters, these approaches are typically very data hungry, while the human and time resources to produce high-quality annotated datasets are usually severely limited, especially in the biomedical domain, not to mention additional limiting factors due to privacy regulations. This requires the development of semi/unsupervised learning approaches, data modeling and simulation methods that can generate high-fidelity ground-truth data for training, and ways to integrate expert domain knowledge into the learning framework.

Furthermore, even if sufficient annotated data can be collected to train a machine or deep learning-based method for a given application, the resulting model is considered a black box in the eyes of practitioners, who remain fully accountable for any decisions based on the model’s predictions. Hence, there is a great need for explainable and interpretable machine and deep learning solutions. This is a fantastic opportunity for BISP researchers, many of whom traditionally are used to developing mathematical models based on sound physical principles, which by design are much more explainable and interpretable. Another challenge stemming from limited training data is the typically poor generalizability of the learned models. While

organized competitions in the field have done a great service by providing public datasets and benchmarks, it is now well known that models based on them do not always work on private datasets. This calls for continuing efforts to make public datasets less selective and more representative.

Given these and many other open challenges, the BISP TC will continue to play an important role in developing ever more advanced image and signal processing methodologies underpinning the next-generation technologies needed to improve the efficacy of biomedical practice and research. In this endeavor, we believe future advances will come not only from continuing research efforts but also from innovations in

education and how we train the next generation of scientists and engineers in our field. Clearly, biomedical image and signal processing has become increasingly multidisciplinary, requiring a deep understanding of not only the mathematics and algorithms of how to model and process digital images and signals but also of the underlying physical principles and limitations of data acquisition using various systems; the bio-

medical knowledge to properly interpret the data; the data science and informatics expertise to handle large datasets; and the experimental and statistical know-how to validate methods thoroughly. To this end, we envision the BISP TC strengthening ties with the relevant bodies in the respective disciplines and becoming more multidisciplinary in the future.

**We envision the BISP TC strengthening ties with the relevant bodies in the respective disciplines and becoming more multidisciplinary in the future.**

## Acknowledgment

The authors are the present and past BISP TC Chairs in reverse chronological order.

## Authors

*Selin Aviyente* (aviyente@egr.msu.edu) is with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA. She is a Senior Member of IEEE.

*Alejandro F. Frangi* (a.frangi@leeds.ac.uk) is with the Centre for Computational Imaging and Simulation Technologies in Biomedicine, Schools of Computing and Medicine, University of Leeds, LS2 9JT Leeds, U.K.; the Alan Turing Institute, LS2 9BW London, U.K.; and the Departments of Electrical Engineering and Cardiovascular Sciences, KU Leuven, 3000 Leuven, Belgium. He is a Fellow of IEEE.

*Erik Meijering* (erik.meijering@unsw.edu.au) is with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia. He is a Fellow of IEEE.

*Arrate Muñoz-Barrutia* (mamunozb@ing.uc3m.es) is with the Bioengineering Department, Universidad Carlos III de Madrid, Leganés, 28912 Madrid, Spain. She is a Senior Member of IEEE.

*Michael Liebling* (michael.liebling@idiap.ch) is with the Idiap Research Institute, 1920 Martigny, Switzerland. He is a Member of IEEE.



**Dimitri Van De Ville** (dimitri.vandeville@epfl.ch) is with the Neuro-X Institute, School of Engineering, EPFL, 1202 Geneva, Switzerland, and the Faculty of Medicine, University of Geneva, 1202 Geneva, Switzerland. He is a Fellow of IEEE.

**Jean-Christophe Olivo-Marin** (jcolivo@pasteur.fr) is with the Cell Biology and Infection Department, Institut Pasteur, F-75724 Paris, France. He is a Fellow of IEEE.

**Jelena Kovačević** (jelenak@nyu.edu) is with the New York University Tandon School of Engineering, Brooklyn, NY 11201 USA. She is a Fellow of IEEE.

**Michael Unser** (Michael.Unser@epfl.ch) is with the School of Engineering, EPFL, CH-1015 Lausanne, Switzerland. He is a Fellow of IEEE.

## References

[1] M. Unser and Z.-P. Liang, "Guest editorial: First IEEE symposium on biomedical imaging," *IEEE Trans. Med. Imag.*, vol. 21, no. 8, pp. 850–851, Aug. 2002, doi: 10.1109/TMI.2002.803604.

[2] R. F. Murphy, E. Meijering, and G. Danuser, "Guest editorial: Molecular and cellular bioimaging," *IEEE Trans. Image Process.*, vol. 14, no. 9, pp. 1233–1236, Sep. 2005, doi: 10.1109/TIP.2005.855701.

[3] J. Kovačević and R. F. Murphy, "Guest editorial: Molecular and cellular bioimaging," *IEEE Signal Process. Mag.*, vol. 23, no. 3, p. 19, May 2006, doi: 10.1109/MSP.2006.1628874.

[4] A. Muñoz-Barrutia, J. Kovačević, M. Kozubek, E. Meijering, and B. Parvin, "Guest editorial: Quantitative bioimaging: Signal processing in light microscopy," *IEEE Signal Process. Mag.*, vol. 32, no. 1, pp. 18–19, Jan. 2015, doi: 10.1109/MSP.2014.2359691.

[5] C. Kervrann, S. T. Acton, J.-C. Olivo-Marin, C. Ó. S. Sorzano, and M. Unser, "Introduction to the issue on advanced signal processing in microscopy and cell imaging," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 1, pp. 3–5, Feb. 2016, doi: 10.1109/JSTSP.2015.2511299.

[6] E. Meijering, V. Calhoun, G. Menegaz, D. Miller, and J. Ye, "Guest editorial: Deep learning in biological image and signal processing," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 24–26, Mar. 2022, doi: 10.1109/MSP.2021.3134525.

[7] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, Oct. 2009, doi: 10.1109/RBME.2009.2034865.

[8] X. Zhou et al., "A comprehensive review for breast histopathology image analysis using classical and deep neural networks," *IEEE Access*, vol. 8, pp. 90,931–90,956, May 2020, doi: 10.1109/ACCESS.2020.2993788.

[9] T. Adali et al., "Introduction to the issue on fMRI analysis for human brain mapping," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 6, pp. 813–816, Dec. 2008, doi: 10.1109/JSTSP.2008.2009263.

[10] D. Van De Ville, V. Jirsa, S. Strother, J. Richiardi, and A. Zalesky, "Introduction to the issue on advanced signal processing for brain networks," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 7, pp. 1131–1133, Oct. 2016, doi: 10.1109/JSTSP.2016.2602945.

[11] W. Huang, T. A. W. Bolton, J. D. Medaglia, D. S. Bassett, A. Ribeiro, and D. Van De Ville, "A graph signal processing perspective on functional brain imaging," *Proc. IEEE*, vol. 106, no. 5, pp. 868–885, May 2018, doi: 10.1109/JPROC.2018.2798928.

[12] D. Farina, A. Mohammadi, T. Adali, N. V. Thakor, and K. N. Plataniotis, "Signal processing for neurorehabilitation and assistive technologies," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 5–7, Jul. 2021, doi: 10.1109/MSP.2021.3076280.

[13] D. Schonfelda, J. Goutsias, I. Shmulevich, I. Tabus, and A. H. Tewfik, "Introduction to the issue on genomic and proteomic signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 257–260, Jun. 2008, doi: 10.1109/JSTSP.2008.925864.

[14] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017, doi: 10.1109/TIP.2017.2713099.

[15] S. Ahn and J. Fessler, "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms," *IEEE Trans. Med. Imag.*, vol. 22, no. 5, pp. 613–626, May 2003, doi: 10.1109/TMI.2003.812251.

[16] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 296–310, Jul. 1993, doi: 10.1109/83.236536.

[17] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006, doi: 10.1109/TVT.2006.871582.

[18] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008, doi: 10.1109/MSP.2007.914731.

[19] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007, doi: 10.1002/mrm.21391.

[20] M. Guerquin-Kern, M. Häberlin, K. Pruessmann, and M. Unser, "A fast wavelet-based reconstruction method for magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 30, no. 9, pp. 1649–1660, Sep. 2011, doi: 10.1109/TMI.2011.2140121.

[21] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003, doi: 10.1109/TIP.2003.814255.

[22] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image reconstruction is a new frontier of machine learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1289–1296, Jun. 2018, doi: 10.1109/TMI.2018.2833635.

[23] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 85–97, Jan. 2023, doi: 10.1109/MSP.2022.3199595.

[24] J. Yoo et al., "Deep learning diffuse optical tomography," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 877–887, Apr. 2020, doi: 10.1109/TMI.2019.2936522.

[25] P. Sarder and A. Nehorai, "Deconvolution methods for 3-D fluorescence microscopy images," *IEEE Signal Process. Mag.*, vol. 23, no. 3, pp. 32–45, May 2006, doi: 10.1109/MSP.2006.1628876.

[26] C. Vonesch, F. Aguet, J.-L. Vonesch, and M. Unser, "The colored revolution of bioimaging," *IEEE Signal Process. Mag.*, vol. 23, no. 3, pp. 20–31, May 2006, doi: 10.1109/MSP.2006.1628875.

[27] H. Greenspan, B. van Ginneken, and R. M. Summers, "Deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016, doi: 10.1109/TMI.2016.2553401.

[28] E. Meijering, "A bird's-eye view of deep learning in bioimage analysis," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 2312–2325, Jan. 2020, doi: 10.1016/j.csbj.2020.08.003.

[29] S. K. Zhou et al., "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021, doi: 10.1109/JPROC.2021.3054390.

[30] T. Falk et al., "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019, doi: 10.1038/s41592-018-0261-2.

[31] Y. Zhu and E. Meijering, "Automatic improvement of deep learning-based cell segmentation in time-lapse microscopy by neural architecture search," *Bioinformatics*, vol. 37, no. 24, pp. 4844–4850, Dec. 2021, doi: 10.1093/bioinformatics/btab556.

[32] P. Vandewalle, J. Kovačević, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 37–47, May 2009, doi: 10.1109/MSP.2009.932122.

[33] K. Price, "Anything you can do, I can do better (No you can't)..." *Comput. Vision, Graph., Image Process.*, vol. 36, nos. 2–3, pp. 387–391, Nov./Dec. 1986, doi: 10.1016/0734-189X(86)90083-6.

[34] J. West et al., "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assisted Tomogr.*, vol. 21, no. 4, pp. 554–568, Jul. 1997, doi: 10.1097/00004728-199707000-00007.

[35] N. Chenouard et al., "Objective comparison of particle tracking methods," *Nature Methods*, vol. 11, no. 3, pp. 281–289, Mar. 2014, doi: 10.1038/nmeth.2808.

[36] R. D. Rudyanto et al., "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: The VESSEL12 study," *Med. Image Anal.*, vol. 18, no. 7, pp. 1217–1232, Oct. 2014, doi: 10.1016/j.media.2014.07.003.

[37] M. Maška et al., "A benchmark for comparison of cell tracking algorithms," *Bioinformatics*, vol. 30, no. 11, pp. 1609–1617, Jun. 2014, doi: 10.1093/bioinformatics/btu080.

[38] E. Gómez-de Mariscal et al., "DeepImageJ: A user-friendly environment to run deep learning models in ImageJ," *Nature Methods*, vol. 39, no. 2, pp. 73–86, 2022.

[39] A. Reinke, M. D. Tizabi, M. Eisenmann, and L. Maier-Hein, "Common pitfalls and recommendations for grand challenges in medical artificial intelligence," *Eur. Urol. Focus*, vol. 7, no. 4, pp. 710–712, Jul. 2021, doi: 10.1016/j.euf.2021.05.008.

[40] L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Commun.*, vol. 9, no. 1, Dec. 2018, Art. no. 5217, doi: 10.1038/s41467-018-07619-7.

[41] L. Maier-Hein et al., "BIAS: Transparent reporting of biomedical image analysis challenges," *Med. Image Anal.*, vol. 66, Dec. 2020, Art. no. 101796, doi: 10.1016/j.media.2020.101796.

[42] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015, doi: 10.1109/JPROC.2015.2460697.

# Multimedia Signal Processing

*A history of the Multimedia Signal Processing Technical Committee*



©SHUTTERSTOCK.COM/TRIFF

**M**ultimedia signal processing (MMSP) refers to processing of signals from multiple media—speech, audio, images, text, graphics, point clouds, etc.—often jointly. This article reviews the history of MMSP and, in parallel, the history of the MMSP Technical Committee (TC), with a focus on the last three decades (Figure 1).

## Introduction

### *Overview of the MMSP TC*

The MMSP TC of the IEEE Signal Processing Society (SPS) promotes the advancement of MMSP technology. The TC was formed in 1996. The scope of the TC includes joint processing/representation of audio–visual and multimodal information, fusion/fission of sensor information or multimodal data, integration of media, art, and multimedia technology, and analysis and feature extraction of multimodal data. Other key areas encompass virtual reality and 3D imaging, multimedia communications and networking, human–machine interface and interaction, visual and auditory quality assessment, multimedia databases, and digital libraries. In this context, the TC also serves as an incubator of technologies that lie in the gaps between traditional areas. Each year, the MMSP TC organizes the IEEE International Workshop on Multimedia Signal Processing, which attracts researchers from the SPS and related communities that work on multimedia topics. The workshop typically receives around 150 paper submissions and has more than 100 attendees from all over the world.

### *Historical context: the 1980s*

Technological developments in the 1980s lay the foundation for the modern multimedia industry. The first CD appeared on the market in 1982 [1]. Personal computers gradually became more affordable throughout the decade and made their way into many homes. Video games, whose first prototypes appeared a few decades earlier, reached the level of popularity that made the gaming industry a notable segment of the tech sector. The first digital video coding standards, H.120 [2]

### Multimedia Signal Processing

- 1980's: First Digital Media Standards (CD, H.120, H.261) and WWW
- 1991: WWW Open to Public, 2G (GSM) Cellular Communications
- 1992: JPEG, MPEG-1, MP3
- 1993: First Movie Streaming and First Live Streaming Over the Internet
- 1996: MPEG-2/H.262
- 1997: Commercial On-Demand Streaming (VXtreme and RealNetworks)
- 1998: Digital Terrestrial TV
- 1999: MPEG-4
- 2000: JPEG2000
- 2001: 3G Cellular Communications
- 2002: ABR Over HTTP
- 2004: MPEG-4 AVC/H.264
- 2005: YouTube Launched
- 2006: Amazon Unbox Launched  
Google Buys YouTube
- 2007: Netflix and Hulu Streaming
- 2009: First 4G Cellular Service Launched in Norway and Sweden
- 2009: Apple Introduces HTTP Live Streaming; Also Multi-View Coding (MVC) Extension of H.264/AVC Released
- 2011: MPEG-DASH
- 2012: AlexNet Wins ILSVRC
- 2013: HEVC
- 2014: Amazon Echo Launched
- 2015: IMT-2020 Standard Requirements for 5G Communications
- 2016: HEVC-SCC
- 2018: Waymo Launches First Fully Autonomous Taxi Service in Phoenix, AZ
- 2020: VVC and MPEG-PCC
- 2021: MPEG-7 Part 17: Neural Network Compression

### Multimedia Signal Processing Technical Committee

- 1996: MMSP TC Formed
- 1997: First MMSP Workshop (Princeton, NJ)
- 1998: Second MMSP Workshop (Redondo Beach, CA)
- 1999: IEEE Transactions on Multimedia Launched
- 2000: First ICME Conference (New York, NY)
- 2001: MMSP (Cannes, Fra.)
- 2002: ICME (Lausanne)
- 2003: ICME (Baltimore)
- 2004: MMSP (Siena, Italy)
- 2005: ICME (Amsterdam)
- 2006: MMSP (Victoria, BC)
- 2007: ICME (Beijing)
- 2008: MMSP (Cairns, Australia), ICME (Hannover)
- 2009: MMSP (Rio de Janeiro)  
ICME (New York)
- 2010: ICME Introduces Double-Blind Review, 15% Acceptance for Orals, 30% Overall
- 2013: ICME (San Jose, CA)
- 2014: MMSP (Banff, AB)  
ICME (Chengdu)
- 2015: MMSP (Xiamen)  
ICME (Turin, Italy)
- 2017: MMSP (Luton, UK)  
ICME (Hong Kong)
- 2019: MMSP (Kuala Lumpur)  
ICME (Shanghai)
- 2020: First Virtual ICME and MMSP Workshop



**FIGURE 1.** MMSP timeline. See Table 1 for acronym definitions. ABR: adaptive bit rate; SCC: screen content coding.

and H.261 [3], and the first media platform World Wide Web (WWW, or simply the Web) [4] were developed in the 1980s, setting the stage for subsequent technological breakthroughs. To help the reader navigate the article, Table 1 gives a list of acronyms and their definitions, while Table 2 summarizes the multimedia standards mentioned in the text.

## Developments in the last three decades

### The 1990s

The 1990s were the decade of great milestones for digital multimedia. The Web was publicly released in 1991. As the first platform that enabled worldwide sharing of multimedia documents, combining text, images, graphics, audio, and video, it has since transformed the way we work, learn, shop, travel, keep in touch, and virtually all other aspects of our lives. Another pivotal event in 1991 was the launch of 2G cellular communications based on the global system for

mobile communications (GSM) standard in Finland. Besides voice communications, 2G systems enabled short message service text messages, which later evolved to multimedia messaging service messages and lay the foundation for the myriad of today's messaging services. Cellular communications have had an equally transformational effect on our lives, providing the infrastructure over which much of today's multimedia content is being shared.

The first widely used multimedia standards were released in 1992. The Joint Photographic Experts Group (JPEG) published Part 1 of the JPEG image coding standard [5], the most popular image coding standard to date. The Moving Picture Experts Group (MPEG) issued MPEG-1 [6], the audio-visual coding standard that formed the basis of video CD and early digital cable and satellite TV. The standard also introduced MPEG-1 audio layer III [7], more commonly known as MP3, a widely popular audio format for music sharing.

**The scope of the TC includes joint processing/representation of audio-visual and multimodal information, fusion/fission of sensor information or multimodal data, integration of media, art, and multimedia technology, and analysis and feature extraction of multimodal data.**

**Table 1. Acronyms and their definitions in alphabetical order.**

Acronym	Definition
AAC	Advanced audio coding
AOM	Alliance for open media
AVC	Advanced video coding
CD	Compact disc
CDVS	Compact descriptors for visual search
DASH	Dynamic adaptive streaming over HTTP
DVD	Digital video disc
GSM	Global system for mobile communications
HDTV	High-definition television
HEVC	High-efficiency video coding
HTTP	Hypertext transfer protocol
ICME	International Conference on Multimedia & Expo
JPEG	Joint Photographic Experts Group
Lidar	Light detection and ranging
MMS	Multimedia messaging service
MP3	MPEG-1 audio layer III
MPEG	Moving Picture Experts Group
MVC	Multiview video coding
P2P	Peer-to-peer communications
PCC	Point cloud compression
RGB+D	Red, green, blue plus depth
SMS	Short message service
VCM	Video coding for machines
VVC	Versatile video coding
WWW	World wide web

**Table 2. Select multimedia standards in chronological order of their release.**

Standard	Description	Initial Release
H.120	First digital video coding standard	1984
H.261	Video coding standard, targeted mainly at video telephony	1988
JPEG	First digital image coding standard	1992
MPEG-1	Audiovisual coding standard	1992
MPEG-2	Audiovisual coding standard	1996
MPEG-2 Part 2 (H.262)	Video coding standard	1996
MPEG-4	Audiovisual coding standard	1999
JPEG2000	A wavelet-based image coding standard	2000
MPEG-4 Part 10 (H.264/AVC)	Video coding standard	2004
MVC	Multiview video coding, amendment to MPEG-4 Part 10 (H.264/AVC)	2009
MPEG-DASH	MPEG dynamic adaptive streaming over HTTP, a video streaming standard	2011
MPEG-H Part 2 (H.265/HEVC)	Video coding standard	2013
3D-HEVC	HEVC-based coding standard for 3D video	2015
HEVC-SCC	HEVC-based coding standard for screen content video	2016
MPEG-I Part 3 (H.266/VVC)	Video coding standard	2020
MPEG-PCC	Point cloud compression standard	2020
MPEG-7 Part 17	A standard for compression of neural network models	2021

The year 1993 was a big year for video streaming. On 22 May, the movie called *Wax or the Discovery of Television Among the Bees* became the first movie to be streamed online, at half the standard definition resolution and a frame rate of only two frames per second. The first live streaming over the Internet occurred on 24 June 1993. This was a performance by the band called *Severe Tire Damage*, streamed from the Xerox Palo Alto Research Center [9]. The video resolution was only  $152 \times 76$ , the frame rate only 8 to 12 frames per second, and the audio quality no better than a telephone call, but it could be seen as far as Australia [10]. It was a historic event that demonstrated the potential of streaming technology and stimulated much research and development in the decades to follow. Propelled by this success, *Severe Tire Damage* opened for the *Rolling Stones* in the second live streaming of a musical event over the Internet on 18 November 1994 [11].

In 1994, DirecTV launched the first commercial digital satellite TV service in the United States. This marked the beginning of the transition of TV from analog to digital, which continued with the introduction of digital cable TV in 1996 and digital terrestrial TV in 1998. MPEG-2 was released in 1996 [6]. It was a very popular coding standard that was used in DVDs, digital TV, and HDTV. Its Part 7, advanced audio coding, was released in 1997.

As the Internet reached more users through the 1990s, with increased capacity and higher bit rates, the battle for streaming over the Internet would heat up, especially in the second part of the decade [12]. The big players in this area were Progressive Networks (which became RealNetworks in 1997) and Microsoft, along with a number of startups, including Vivo, Xing, VDOnet, and VXtreme. Among them, they are responsible for a number of firsts, including the first audio streaming service (RealAudio, 1995), first live audio webcast of a sports game (*Seattle Mariners* versus *New York Yankees*, RealNetworks, 1995), first commercial on-demand video streaming (RealNetworks and VXtreme, 1997), as well as the most popular media players of the time.

Amid all of these developments, the MMSP TC was formed in 1996, under the leadership of the first TC chair, Tsuhan Chen. The TC organized its first workshop, the MMSP Workshop, in June 1997 in Princeton, New Jersey. The workshop attracted 95 papers, including eight demonstrations. The next two workshops were held in Redondo Beach, California, (1998) and Copenhagen, Denmark (1999). The second TC chair (1999–2001) was K. J. Ray Liu, the 2022 IEEE president.

Because of the interdisciplinary nature of multimedia, the MMSP TC has collaborated with other TCs within the IEEE SPS and other IEEE societies since its inception. A notable result of such collaboration was the launch of the *IEEE Transactions on Multimedia* in 1999. With an impact factor of 8.18,

the journal is now considered among the top publication venues in the field of multimedia.

### The 2000s

If the 1990s demonstrated the potential of multimedia technologies, the 2000s were the decade when the technology reached the level of maturity that made it not only commercially viable, but highly successful. This was aided by the development of

3G cellular communications, first commercially launched in 2001 by NTT DoCoMo in Japan. In addition, new ideas emerged both from industry and academia. One of these ideas was peer-to-peer (P2P) file sharing, pioneered by Napster.

Napster initially launched in 1999 and quickly became a very popular platform for MP3 audio file sharing, especially among college students. It was soon sued over copyright infringement [13] and had to shut down in 2001. Despite its brief existence, Napster left a lasting legacy in the multimedia world. Its P2P distribution paradigm

generated enormous interest in the research community and became popular not only as a way to share files, but also in media streaming. At the same time, the music industry saw the potential for distributing content in digital form without physical media, which lay the foundation for subsequent online music stores, such as iTunes, and products such as iPod.

At the turn of the millennium, the MMSP TC was also busy launching new initiatives. The International Conference on Multimedia and Expo (ICME) was launched as a collaboration with the sister TCs in the IEEE Circuits and Systems, Communications, and Computer Societies. The first edition of the conference was held in New York in July–August 2000, and attracted over 400 papers. Since then, the ICME has established itself as a flagship IEEE conference in the field of multimedia: it has a rank of A according to the Computing, Research, and Education Association of Australia rankings and is among the top 10 venues (among both journals and conferences) in the field of multimedia, according to Google Scholar metrics.

An important milestone at the turn of the millennium was the standardization of JPEG2000 [14]. This was the first coding standard based on wavelets [15]. JPEG2000 introduced tools for resolution- and quality-scalable coding and decoding, region-of-interest coding, precise rate control, and a number of other features that made it suitable for high-quality imaging applications. A related image coding approach, called *ICER*, is used for encoding and sending back images from the Mars rovers [16]. In 2004, Motion JPEG2000, an extension of JPEG2000 to video, was adopted for digital cinema applications in the film industry.

Another major milestone was the development of the MPEG-4 Part 10 advanced video coding (AVC) standard, better known as *H.264/AVC* [17], in 2003. One of the main motivations behind *H.264/AVC* was to support various network-based video services, such as video streaming to heterogeneous

**As the first platform that enabled worldwide sharing of multimedia documents, combining text, images, graphics, audio, and video, it has since transformed the way we work, learn, shop, travel, keep in touch, and virtually all other aspects of our lives.**

clients. Hence, scalability also played an important role, and was materialized through the scalable extension of H.264/AVC [18], which enabled video coding and decoding at a number of resolutions, frame rates, and qualities to support a wide variety of client devices. H.264/AVC is used in Blu-ray discs and is still the most common format in online video streaming.

Although online video existed in various forms even in the 1990s, the first major video streaming service, YouTube, was launched in 2005, fueled by the development of H.264/AVC. YouTube allowed users to upload their own videos, which can then be searched and streamed to a wide audience. This quickly made YouTube very popular, leading to its purchase by Google for US\$1.65 billion in 2006, less than a year after its official launch. Commercial streaming services appeared around the same time. Amazon Unbox (now Amazon Prime video) launched in 2006, followed by Netflix and Hulu streaming services in 2007. As the customers' home Internet service speeds improved, the popularity of streaming services grew, and streaming now accounts for more viewing time than cable TV in the U.S. market. These streaming services, and many more that have followed since, became successful businesses, some even launching their own production studios to create exclusive content.

While certain forms of online games existed as far back as the 1970s, the era of massively multiplayer online gaming started in 2000s with the wider availability of fast Internet service. Gaming consoles such as Microsoft Xbox, Sony PlayStation, Nintendo, and Wii gradually became more popular, interfacing with cloud gaming platforms like Xbox Live and PlayStation Network. Increased interactivity in games and consoles' specialized hardware also incentivized the development of more sophisticated game controllers, like Microsoft Kinect, which would have a major impact on both gaming and MMSP research in the following decade.

The 2000s also saw the birth of social media, with the founding of LinkedIn in 2002, Facebook in 2004, and Twitter in 2006. A phenomenon that took the world by storm, social media allowed users to upload their own media content and share it with a circle of friends or a wider audience. Social media has since transformed marketing and market research, recruitment, the news industry, and many other aspects of our lives. It has also facilitated phenomena like trending, influencing, fake news, etc. On the technical side, the immense amount of user-supplied content ushered in the era of Big Data and set the stage for further technical developments in the coming decades. As an example, user-supplied photos and associated tags enabled Facebook to create a highly successful facial recognition system, which launched in 2010 but has since been scaled back due to ethical and privacy concerns.

The end of the decade was equally exciting in terms of technological developments. The 4G cellular service was first

launched in Norway and Sweden in 2009. With increased bit rates offered to the users, the demand for online media and streaming services will rapidly increase in the next decade. The same year, Apple introduced HTTP live Sstreaming, which is currently the most popular streaming format. Also, the multi-view video coding extension of H.264/AVC was introduced.

In the meantime, the MMSP TC was busy building up the MMSP community and organizing related events. MMSP

**JPEG2000 introduced tools for resolution- and quality-scalable coding and decoding, region-of-interest coding, precise rate control, and a number of other features that made it suitable for high-quality imaging applications.**

Workshops took place in Cannes, France (2001); St. Thomas, U.S. Virgin Islands (2002); Siena, Italy (2004); Shanghai, China (2005), Victoria, BC, Canada (2006); Chania, Greece (2007); Cairns, QLD, Australia (2008); and Rio de Janeiro, Brazil (2009). ICME conferences took place in New York, NY, USA (2000); Tokyo, Japan (2001); Lausanne, Switzerland (2002); Baltimore, MD, USA (2003); Taipei, Taiwan (2004); Amsterdam, The Netherlands (2005); Toronto, ON, Canada (2006); Beijing, China (2007); Hannover, Germany (2008); and again in New York, NY, USA (2009). During this period,

MMSP TC chairs were K.-J. Ray Liu (1999–2001), John. A. Sørensen (2002–2003), Yu Hen Hu (2004–2005), Ingemar J. Cox (2006–2007), and Anthony Vetro (2008–2009).

### *The 2010s*

During this decade, 4G communications spread throughout the world, increasing the demand for online media. Mobile screen resolutions increased sufficiently so that users could watch full HD video on their devices. Interactive media also became more popular; people could now have a reasonable videoconference on the go.

In 2011, MPEG dynamic adaptive streaming over HTTP (DASH) became an international standard. MPEG-DASH and related technologies, like Apple's HTTP live streaming, provided an incentive to consumer electronics companies to incorporate streaming apps into their devices, which in turn gave a boost to the streaming industry. Smart TVs and streaming devices like Apple TV, Amazon Fire TV, Roku, and many others, gradually started supplementing and then replacing traditional cable and satellite TV services.

Another type of application that became popular in the 2010s is mobile visual search [19], where users could take a photo of an object or a location and then retrieve additional information about it, possibly in the form of augmented reality. Audio search apps like Shazam were already established by that time, but an efficient mobile visual search required a good camera and sufficiently powerful hardware for fast feature extraction. All of it came together during the early 2010s. The MPEG compact descriptors for visual search standard [20] was released in 2015 and provided an interoperable way to compress and transmit visual features that facilitate image search and matching. While most multimedia compression standards code data for human consumption, this is a rare example of a standard for visual data coding for machine use, namely visual

search; the trend of coding for machines is becoming very popular at the time of the writing of this article.

The 2010s were a decade when immersive technologies took a big step forward. This was facilitated by improvements in sensing and display technologies over the years, but also computing infrastructure needed to process the increased amount of data required for a high-quality immersive experience. Representative technologies for 3D visual immersion include multiview video, red, green, blue plus depth (RGB+D), and point clouds, while audio counterparts include ambisonics and wave field synthesis. Haptic technologies also moved forward, finding new applications in wearable devices.

Another major event of the 2010s was a sharp rise in the popularity of deep learning with neural networks. Although the benefits of learning with many-layered models were already known in the 1960's [21] and the term *deep learning* dates back to the 1980s [22], it was the success of deep neural networks in acoustic modeling [23] and image classification [24], as well as the availability of large data sets and powerful computing infrastructure, that sparked the renewed interest in the topic, and subsequently transformed many technical fields, including MMSP. This is in part due to the ability of deep neural networks to effectively model relationships in multimodal data [25].

Among the emerging applications that were greatly facilitated by deep learning is autonomous driving, where multiple sensors—cameras, lidar, radar, microphones—collect information from the vehicle's surroundings to help it navigate the road. Processing signals from multiple modalities has traditionally been challenging. However, with the help of deep models, one can learn the complex relationships between different modalities from data, to enable their joint processing and analysis. In 2018, Waymo launched the first autonomous taxi service in Phoenix, Arizona. Another artificial intelligence (AI)/deep learning-driven trend is that of “smart” sensors and devices, such as smart speakers and cameras, whose capabilities have gone beyond capture and low-level processing of signals toward understanding and interaction with their environment.

On the video coding front, a major milestone was the 2013 release of the high-efficiency video coding (HEVC) standard, also known as *H.265* or *MPEG-H Part 2*. Beside the usual 50% coding efficiency gain over the predecessor (*H.264/AVC*), it was targeted at higher resolutions and allowed for higher bit-depth, thus facilitating high dynamic range display. It is currently the second most-widely used video coding format, after *H.264/AVC*. Despite the high adoption of standard codecs in various industries, especially by hardware developers, the research community felt that there is a strong need for royalty-free codecs. Hence, the Alliance for Open Media (AOM) was formed in 2015, with the goal of developing royalty-free video coding technology whose performance would be comparable to that of standard video codecs. Starting with Google's VP9

video codec, initially mainly used on YouTube, AOM released the AOMedia video 1 (AV1) video coding format in 2018. Royalty-free coding formats like VP9 and AV1 tend to be better supported in web browsers and streaming apps compared to standard coding formats.

Major revamping of the ICME conference took place during the 2010s. First, in 2010, ICME introduced double-blind review process, a departure from traditional single-blind review that is still common in signal processing. Moreover, the target acceptance rate was set to 30%, with the top 15% percent of papers being selected for oral presentation. Starting with 2012, ICME Workshops, which were introduced in 2009 to provide more focused satellite events and to foster new and emerging topics, have been published in separate proceedings. These innovations, and of course the hard work of many volunteers, helped the ICME become what it is today: a flagship IEEE conference in multimedia.

Besides the ICME, the MMSP TC has also been organizing MMSP Workshops, which took place in Saint Malo, France (2010); Hangzhou, China (2011); Banff, AB, Canada (2012); Pula, Italy (2013); Jakarta, Indonesia (2014); Xiamen, China (2015); Montréal, QC, Canada (2016); Luton, United Kingdom (2017); Vancouver, BC, Canada (2018); and Kuala Lumpur, Malaysia (2019). ICME conferences were held in Singapore (2010); Barcelona, Spain (2011); Melbourne, VIC, Australia (2012); San Jose, California, USA (2013); Chengdu, China (2014); Turin, Italy (2015); Seattle, Washington, USA (2016); Hong Kong (2017); San Diego, California, USA (2018); and Shanghai, China (2019). During this decade, the MMSP TC was chaired by Philip Chou (2010–2011), Oscar Au (2012–2013), Dinei Florencio (2014–2015), Enrico Magli (2016–2017), and Frédéric Dufaux (2018–2019).

## The 2020s

The current decade started with an event that impacted the world in many ways: the COVID-19 pandemic. As people retreated to their homes and started working remotely, the importance of multimedia suddenly grew. The demand for streaming services spiked, and videoconferencing became the norm for business meetings and presentations, education, and simply socializing and keeping in touch with friends and family. Before the pandemic, multimedia technology was mostly driven by entertainment. Now, it has become part of the infrastructure of our society. Even as the pandemic-related restrictions get removed, the concepts of remote work and collaboration are staying.

Although the decade is still young, several important technological milestones have already occurred. The latest video coding standard, versatile video coding (VVC) [26], also known as *H.266* or *MPEG-I Part 3*, was released in 2020. Besides the usual improvement in compression efficiency over its predecessor, VVC was developed to support a broad set

**As the customers' home Internet service speeds improved, the popularity of streaming services grew, and streaming now accounts for more viewing time than cable TV in the U.S. market.**

of resolutions, up to 16 K, a variety of color formats, as well as 360° video. Another important standard released in 2020 was MPEG point cloud compression (PCC). Targeting applications like augmented, virtual, and mixed reality, MPEG PCC provides compression technology for video-based and geometry-based PCC [27]. A related standard that is still being developed is JPEG Pleno [28], whose goal is to provide compression support for plenoptic imaging modalities, such as light fields, holography, and point clouds.

As noted earlier, learning-based technologies are playing an increasingly important role in many areas, including MMSP. But the benefit is mutual. In 2021, MPEG released a standard for neural network compression (MPEG-7 Part 17), whose purpose is to enable compression of neural network models for efficient storage and transport. While its purpose is to compress networks rather than multimedia signals, the standard was built upon the knowledge base developed over the years in image and video compression. Neural network compression is useful in federated learning, where model weights need to be transmitted between the clients and the server during the network training process.

Broader technological trends, such as the deployment of 5G communication systems, the growing Internet of Things, and advances in AI, are opening up possibilities for “smart” homes, buildings, factories, and cities. In applications like these, automation is a necessity, since the amount of data captured and communicated is far too much for humans to take note of. For example, most of the video captured by surveillance cameras will never be seen by humans, only “seen” by machines. As a result, several standardization efforts have been initiated to create media compression formats suitable for machine use, or combined human and machine use. One of these is JPEG AI [29], whose goal is to develop learning-based compression technology that supports conventional image decoding as well as a number of image processing and machine vision tasks. The other is MPEG video coding for machines [30], which targets both machine-only and human-machine tasks. Completion of these standards will facilitate improved efficiency of many technologies already in use, such as video monitoring, autonomous navigation, and multimedia database management, and create fertile ground for new and yet-to-be-imagined applications.

Due to the pandemic, MMSP TC activities in the 2020s have mostly been virtual so far. MMSP workshops took place virtually in Tampere, Finland in 2020, again in Tampere, Finland, as a hybrid event in 2021, and virtually in Shanghai, China, in 2022. ICME was a virtual event in London, United Kingdom, in 2020, and in Shenzhen, China, in 2021, and was organized as a hybrid event in Taipei, Taiwan, in 2022. During this period, MMSP TC chairs were Marta Mrak (2020–2021) and Ivan Bajić (2022–2023).

**On the technical side, the immense amount of user-supplied content ushered in the era of Big Data and set the stage for further technical developments in the coming decades.**

## Conclusions

Starting as a mostly entertainment-driven technology, MMSP has come a long way to become a part of the very fabric of our society. It has enabled highly successful businesses, provided critical infrastructure at the time of need, and reached virtually everyone in some form or another. The MMSP TC has been a part of that story over the last two and a half decades.

So what does the future of MMSP look like? As the saying goes, “making predictions is difficult, especially about the future.” In the near term, the trends are clear: data-driven approaches in the form of AI/deep learning are pushing the boundaries of what is possible with multimedia signals, and laying the foundation for the next generation of multimedia applications, products, and services. Beyond that, who knows: perhaps quantum multimedia?

## Acknowledgment

The authors would like to thank Dr. Philip Chou for his help and consultation during the writing of this article.

## Authors

**Ivan V. Bajić** (ibajic@ensc.sfu.ca) received his Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute. He is a professor of engineering Science at Simon Fraser University, Burnaby, BC V5A 1S6, Canada, and the current Chair of the Multimedia Signal Processing (MMSP) Technical Committee. He was an associate editor of the *IEEE Transactions on Multimedia* and served on the organizing and program committees of the main conferences in the field, having won several service awards in these roles. His group’s research has received awards at IEEE International Conference on Multimedia and Expo 2012, IEEE International Conference on Image Processing 2019, and IEEE MMSP 2022. He is a Senior Member of IEEE.

**Marta Mrak** (marta@ieee.org) received her Dipl. Ing. and M.Sc. electrical engineering degrees from the University of Zagreb, Croatia; and her Ph.D. degree from Queen Mary University of London, U.K. She is a senior AI research engineer at Helsing, WIT 3BL London, U.K. She has participated in the work of the Multimedia Signal Processing Technical Committee (MMSP TC) since 2014 and was MMSP TC chair from 2020–2021. During that time, she was a lead engineer at BBC R&D, where she ran various projects, ranging from video compression fundamentals to new content experiences powered by machine learning. Within the TC, her main contributions included serving as a general chair of IEEE International Conference on Multimedia and Expo (ICME) 2020 and lead Technical Program Committee chair for IEEE ICME 2019. She is a Senior Member of IEEE.

**Frédéric Dufaux** (frederic.dufaux@centralesupelec.fr) received his M.Sc. degree in physics and Ph.D. degree in electrical engineering from École Polytechnique Fédérale de Lausanne in 1990 and 1994, respectively. He is a



CNRS research director at Université Paris-Saclay, CNRS, CentraleSupélec, 91190 Gif-sur-Yvette, France. He was vice general chair of ICIP 2014, general chair of Multimedia Signal Processing (MMSp) 2018, and technical program co-chair of ICIP 2019 and ICIP 2021. He served as chair of the IEEE Signal Processing Society MMSp Technical Committee from 2018–2019. He is chair of the International Conference on Multimedia and Expo Steering Committee for 2022–2023. He was a founding member and the chair of the European Association for Signal Processing Technical Area Committee on Visual Information Processing from 2015 to 2021. He is a Fellow of IEEE.

**Enrico Magli** (enrico.magli@polito.it) received his Ph.D. degree from the Politecnico di Torino, Italy, in 2001. He is a professor with the Politecnico di Torino, 10129 Torino, Italy. He is a senior associate editor of *IEEE Journal on Selected Topics in Signal Processing*. He is a Fellow of the European Lab for Learning and Intelligent Systems Society for the advancement of artificial intelligence in Europe, and has been an IEEE distinguished lecturer from 2015 to 2016. He was the recipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award, the IEEE ICIP 2015 Best Student Paper Award (as senior author), the IEEE ICIP 2019 Best Paper Award, and the IEEE Multimedia 2019 Best Paper Award. He is a Fellow of the IEEE.

**Tsuhan Chen** (dprtchen@nus.edu.sg) is the deputy president for research and technology and distinguished professor at National University of Singapore, Singapore 119077. He also serves as the Chief Scientist of AI Singapore, a national program in artificial intelligence. He founded the Technical Committee on Multimedia Signal Processing in the IEEE Signal Processing Society, which later evolved into founding of the *IEEE Transactions on Multimedia* and the IEEE International Conference on Multimedia and Expo, joining efforts from multiple IEEE societies. He was appointed the editor-in-chief for *IEEE Transactions on Multimedia* from 2002 to 2004. He is a Fellow of IEEE.

## References

[1] "Compact disc." Wikipedia. Accessed: Jun. 24, 2022. [Online]. Available: [https://en.wikipedia.org/wiki/Compact\\_disc](https://en.wikipedia.org/wiki/Compact_disc)

[2] *H.120: Codecs for Videoconferencing Using Primary Digital Group Transmission*, ITU-T H.120, International Telecommunications Union, Geneva, Switzerland, 1984.

[3] *H.261: Video Codec for Audiovisual Services at p x 384 Kbit/s*, ITU-T H.261, International Telecommunications Union, Geneva, Switzerland, Nov. 1988. [Online]. Available: <https://www.itu.int/rec/T-REC-H.261-198811-S/en>

[4] D. H. Johnson, "Signal processing and the world wide web," *IEEE Signal Process. Mag.*, vol. 12, no. 5, pp. 53–57, Sep. 1995, doi: 10.1109/79.410440.

[5] *Information Technology – Digital Compression and Coding of Continuous-Tone Still Images – Requirements and Guidelines*, ITU-T T.81, International Organization for Standardization, Geneva, Switzerland, Sep. 1992.

[6] T. Sikora, "MPEG digital video-coding standards," *IEEE Signal Process. Mag.*, vol. 14, no. 5, pp. 82–100, Sep. 1997, doi: 10.1109/79.618010.

[7] P. Noll, "MPEG digital audio coding," *IEEE Signal Process. Mag.*, vol. 14, no. 5, pp. 59–81, Sep. 1997, doi: 10.1109/79.618009.

[8] J. Markoff, "Cult film is a first on internet," *NY Times*, May 1993. [Online]. Available: <https://www.nytimes.com/1993/05/24/business/cult-film-is-a-first-on-internet.html>

[9] K. Savetz, N. Randall, and Y. Lepage, *MBONE: Multicasting Tomorrow's Internet*. New York, NY, USA: Wiley, 1996.

[10] "Streaming media." Wikipedia. Accessed: Jul. 1, 2022. [Online]. Available: [https://en.wikipedia.org/wiki/Streaming\\_media](https://en.wikipedia.org/wiki/Streaming_media)

[11] N. Strauss, "Rolling stones live on internet: Both a big deal and a little deal," *NY Times*, Nov. 1994. [Online]. Available: <https://www.nytimes.com/1994/11/22/arts/rolling-stones-live-on-internet-both-a-big-deal-and-a-little-deal.html>

[12] D. Rayburn, "The early history of the streaming media industry and the battle between Microsoft and Realnetworks." Seeking Alpha. Accessed: Jul. 3, 2022. [Online]. Available: <https://seekingalpha.com/article/3957046-early-history-of-streaming-media-industry-and-battle-microsoft-and-realnetworks>

[13] R. Stern, "Napster: A walking copyright infringement?" *IEEE Micro*, vol. 20, no. 6, pp. 4–5, Nov./Dec. 2000, doi: 10.1109/40.888696.

[14] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 36–58, Sep. 2001, doi: 10.1109/79.952804.

[15] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Process. Mag.*, vol. 8, no. 4, pp. 14–38, Oct. 1991, doi: 10.1109/79.91217.

[16] A. Kieley and M. Klimesh, "The ICER progressive wavelet image compressor," Jet Propulsion Lab., Nat. Aeronaut. Space Admin., Washington, DC, USA, IPN Prog. Rep. 42-155, Nov. 2003. Accessed: Oct. 1, 2022. [Online]. Available: [https://ipnpr.jpl.nasa.gov/progress\\_report/42-155/1551.pdf](https://ipnpr.jpl.nasa.gov/progress_report/42-155/1551.pdf)

[17] T. Wiegand and G. J. Sullivan, "The H.264/AVC video coding standard [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 24, no. 2, pp. 148–153, Mar. 2007, doi: 10.1109/MSP.2007.323282.

[18] H. Schwarz and M. Wien, "The scalable video coding extension of the H.264/AVC standard [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 135–141, Mar. 2008, doi: 10.1109/MSP.2007.914712.

[19] B. Girod et al., "Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jul. 2011, doi: 10.1109/MSP.2011.940881.

[20] "Information technology - Multimedia content description interface - Part 13: Compact descriptors for visual search," ISO/IEC 15938-13:2015, Aug. 2015.

[21] A. G. Ivakhnenko, "Heuristic self-organization in problems of engineering cybernetics," *Automatica*, vol. 6, no. 2, pp. 207–219, Mar. 1970, doi: 10.1016/0005-1098(70)90092-0.

[22] R. Dechter, "Learning while searching in constraint-satisfaction-problems," in *Proc. AAAI Nat. Conf. Artif. Intell.*, Aug. 1986, pp. 178–183.

[23] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[24] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015, doi: 10.1007/s11263-015-0816-y.

[25] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017, doi: 10.1109/MSP.2017.2738401.

[26] B. Bross et al., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021, doi: 10.1109/TCSVT.2021.3101953.

[27] S. Schwarz et al., "Emerging MPEG standards for point cloud compression," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 133–148, Mar. 2019, doi: 10.1109/JETCAS.2018.2885981.

[28] P. Astola et al., "JPEG Pleno: Standardizing a coding framework and tools for Plenoptic imaging modalities," *ITU J., ICT Discoveries*, vol. 3, no. 1, Jun. 2020, Art. no. 10.

[29] "White paper on JPEG AI scope and framework v1.0," ISO/IEC JTC 1/SC29/WG1 N90049, 2021.

[30] "Use cases and requirements for video coding for machines," ISO/IEC JTC 1/SC29/WG2 N00190, Apr. 2022.



# Twenty-Five Years of Sensor Array and Multichannel Signal Processing

*A review of progress to date and potential research directions*



©SHUTTERSTOCK.COM/TRIFF

In this article, a general introduction to the area of sensor array and multichannel signal processing is provided, including associated activities of the IEEE Signal Processing Society (SPS) Sensor Array and Multichannel (SAM) Technical Committee (TC). The main technological advances in five SAM subareas made in the past 25 years are then presented in detail, including beamforming, direction-of-arrival (DOA) estimation, sensor location optimization, target/source localization based on sensor arrays, and multiple-input multiple-output (MIMO) arrays. Six recent developments are also provided at the end to indicate possible promising directions for future SAM research, which are graph signal processing (GSP) for sensor networks; tensor-based array signal processing, quaternion-valued array signal processing, 1-bit and noncoherent sensor array signal processing, machine learning and artificial intelligence (AI) for sensor arrays; and array signal processing for next-generation communication systems.

## Introduction

Sensor array and multichannel signal processing has a long history, with typical research topics including beamforming and DOA estimation at its early stage and corresponding representative algorithms, including the Capon beamformer/linearly constrained minimum variance (LCMV) beamformer and the MUSIC/ESPRIT algorithms [1], [2], [3], [4], [5]. The past 25 years have seen an explosive growth of research activities in this area, and significant progress has been made in a wide range of theoretical and application areas of sensor array and multichannel signal processing. Although, traditionally, the areas' applications have been mainly limited to the defense sector, such as radar and sonar, today, we can see their impact in everyday life, including beamforming for ultrasound imaging, synthetic aperture radar for remote sensing, vehicular radar (ultrasound and electromagnetic) for autonomous driving, microphone arrays for human-machine interfaces (a good example is the Amazon Echo), and MIMO antenna arrays for Wi-Fi and mobile communications standards (IEEE 802.11n, IEEE 802.11ac, 3G, WiMax, and LTE).

As a result, the sensor array and multichannel signal processing research area has expanded significantly in the past years, as reflected by the scope of the SPS SAM TC. The SAM TC, formed in 2000, aims to promote activities within the technical fields of sensor array processing and multichannel statistical signal processing [6], including beamforming and space-time adaptive processing; DOA estimation; source separation; target detection; localization and tracking; MIMO signal processing; array processing for radar, sonar, and communications; and many other applications of multisensor and synthetic aperture systems, as indicated by the list of editors' information classification schemes covered by the TC (<https://signalprocessingsociety.org/community-involvement/sensor-array-and-multichannel/edics>).

The SAM TC organizes two biennial workshops dedicated to the SAM research area: the IEEE International Workshop on Computational Advances in Multisensor Adaptive Processing (CAMSAP), organized in December every odd-numbered year since 2005, and the IEEE Sensor Array and Multichannel Signal Processing Workshop, organized in June/July every even-numbered year since 2002, each accepting 100–200 research papers. Due to the COVID-19 pandemic, CAMSAP 2021, originally scheduled for December 2021, in Costa Rica, was postponed to December 2023. The next SAM workshop (SAM 2024) will be held in the United States, with two possible venues: Oregon State University, Corvallis, OR, and Skamania Lodge, Stevenson, WA. Moreover, at each year's ICASSP conference, the SAM track also receives about 100–200 regular submissions. Currently, there is also the Synthetic Aperture Technical Working Group, which resides under the SAM TC, with the goal of “supporting the maturation of the theoretical framework and the associated empirical techniques that underpin the estimation of parameters of propagating waves through various media using synthetic apertures.”

In this article, as it is not possible to give an exhaustive list of all the advances made in the SAM area, we focus on five major topics and introduce the corresponding progress made in tackling their respective technical challenges: beamforming [including robust adaptive beamforming and frequency-invariant beamforming (FIB)], DOA estimation (including sparsity-based and underdetermined DOA estimation), sensor location optimization, target/source localization based on sensor arrays, and MIMO arrays (including MIMO radar and MIMO for wireless communications). The first two are classic SAM topics from the very beginning of SAM research, as mentioned earlier, while the latter three were studied systematically only in the past decades. Then, six new developments in the SAM area are presented to give an indication about possible future research directions, including GSP for sensor networks, tensor-based array signal processing, quaternion-valued array signal processing, 1-bit and noncoherent sensor array signal processing, machine learning and AI for sensor arrays, and array signal processing for next-generation communication systems.

This article is structured as follows. The five main technological advances are introduced in detail in the “Main Technological Advances in the SAM Area” section, followed

by the six new developments in the “New Developments in the SAM Area” section and some concluding remarks in the “Concluding Remarks” section.

## Main technological advances in the SAM area

In this section, advances made in the five major SAM research topics in the past 25 years are presented, including beamforming, DOA estimation, sensor location optimization, target/source localization based on sensor arrays, and MIMO arrays.

### Beamforming

Beamforming is a classic sensor array signal processing problem and a core SAM topic [1], [2], [3], [4], [5], and it has been studied extensively at least for a century. It can be classified into narrowband and wideband beamforming according to the relative bandwidth of the signals, adaptive and fixed beamforming according to its relationship with the received data, and analog and digital beamforming according to its circuits implementation. In the past 25 years, three main developments have been achieved, including robust adaptive beamforming [7], FIB [5], and hybrid beamforming [8], which is a combination of digital and analog beamforming techniques. In this section, we discuss the first two in detail and leave the topic of hybrid beamforming to the section about MIMO arrays.

### Robust adaptive beamforming

In general, for the narrowband case, for an  $M$ -sensor array with  $K$  impinging signals, the received array signals can be formulated into the following form:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \quad (1)$$

where  $\mathbf{x}(t) = [x_1(t), \dots, x_M(t)]^T$  is the received signal vector,  $\mathbf{A}$  is the steering matrix consisting of  $K$  steering vectors  $\mathbf{a}(\theta)$  corresponding to the  $K$  source signals [ $\theta$  represents the angle of arrival (AOA) of an arbitrary impinging signal], and  $\mathbf{n}(t)$  is the noise vector.

Then, the beamformer output  $y(t)$  is given by an instantaneous linear combination of the received spatial samples  $x_m(t)$ , as follows:

$$y(t) = \sum_{m=1}^M x_m(t)w_m^* = \mathbf{w}^H \mathbf{x}(t) \quad (2)$$

where  $w_m$  is the weight coefficient for the  $m$ th received sensor signal, with the weight vector  $\mathbf{w} = [w_1, \dots, w_M]^T$ .

The Capon beamformer, which can be considered a special case of the more general LCMV beamformer [1], [2], [3], [4], [5], can achieve effective adaptive beamforming when the DOA angle  $\theta_0$  of the desired signal is exactly known, and the following is the standard formulation:

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{a}(\theta_0) = 1 \quad (3)$$

where  $\mathbf{R} = E\{\mathbf{x}(t)\mathbf{x}^H(t)\}$  is the covariance matrix and  $\mathbf{a}(\theta_0)$  is the steering vector of the array at  $\theta_0$ . In practice, since  $\mathbf{R}$  is usually not available, as an approximation, it is replaced by the sample covariance matrix  $\hat{\mathbf{R}}$ , which is obtained through the finite number of data samples.

However, the Capon beamformer is very sensitive to model mismatch errors, such as DOA error for the desired signal, mutual coupling, general array manifold errors, and finite sample effects in covariance matrix estimation, and therefore, various robust adaptive beamforming techniques have been developed [7]. One well-known technique is diagonal loading, with the weight vector expressed as  $\alpha(\hat{\mathbf{R}} + \xi \mathbf{I})^{-1} \mathbf{a}(\theta_0)$ , with  $\alpha$  being a constant,  $\xi$  the diagonal loading factor, and  $\mathbf{I}$  the identity matrix.

One prominent development in this area in the past 25 years is the worst-case-based robust adaptive beamformer [9], where, instead of constraining the beamformer response to be unity at the desired signal direction, the response is forced to exceed unity within an uncertainty set of steering vectors, which can be expressed as

$$\begin{aligned} \min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{R}} \mathbf{w} \quad \text{subject to} \quad & |\mathbf{w}^H \tilde{\mathbf{a}}| \geq 1, \forall \tilde{\mathbf{a}} \in \mathcal{A} \\ & \mathcal{A} = \{\tilde{\mathbf{a}} | \tilde{\mathbf{a}} = \mathbf{a}(\theta_0) + \mathbf{e}, \|\mathbf{e}\| \leq \varepsilon\} \end{aligned} \quad (4)$$

where  $\tilde{\mathbf{a}}$  is the possible actual steering vector of the desired signal corresponding to the presumed steering vector  $\mathbf{a}(\theta_0)$ ,  $\mathcal{A}$  is the full set that  $\tilde{\mathbf{a}}$  belongs to, and  $\mathbf{e}$  is the steering vector error, with its norm bounded by  $\varepsilon$ . The problem is then converted to the following form using the worst-case optimization:

$$\begin{aligned} \min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{R}} \mathbf{w} \quad \text{subject to} \quad & \mathbf{w}^H \mathbf{a}(\theta_0) \geq \varepsilon \|\mathbf{w}\| + 1 \\ & \text{Im}\{\mathbf{w}^H \mathbf{a}(\theta_0)\} = 0 \end{aligned} \quad (5)$$

where  $\text{Im}\{\cdot\}$  denotes the imaginary part of its argument. Since the signal-to-interference-plus-noise ratio (SINR) of the beamformer output will not change by rotating the weight vector, an alternative formulation can be derived as

$$\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{R}} \mathbf{w} \quad \text{subject to} \quad \text{Re}\{\mathbf{w}^H \mathbf{a}(\theta_0)\} \geq \varepsilon \|\mathbf{w}\| + 1 \quad (6)$$

where  $\text{Re}\{\cdot\}$  takes the real part of its argument.

Both the Capon beamformer and the worst-case robust beamformer require estimation of the covariance matrix  $\mathbf{R}$ , and it is a challenging task when only a small number of snapshots is available; one solution to the problem is the family of iterative adaptive approach-based methods [10], which can still work for the extreme case with only one snapshot.

Another notable contribution for robust adaptive beamforming is based on interference covariance matrix reconstruction and steering vector estimation [11], which has attracted much attention recently, with follow-up works focusing on different ways of reconstructing either or both of the covariance matrices corresponding to the desired signal and interference plus noise, separately.

### Frequency-invariant beamforming

For wideband arrays, different from the data model in (1), the received array signals are expressed in the form of convolution (represented by  $\star$ ) [5]:

$$\mathbf{x}(t) = \tilde{\mathbf{A}} \star \mathbf{s}(t) + \mathbf{n}(t) \quad (7)$$

where the  $(m, k)$ th element of the matrix  $\tilde{\mathbf{A}}$  is given by  $\delta(t - \tau_{m,k})$ , with  $\tau_{m,k}$  being the time delay of the  $k$ th impinging signal at the  $m$ th sensor compared to some reference point.

As a result, wideband beamforming is achieved through a series of tapped delay lines (TDLs) or finite-impulse response/infinite-impulse response filters in its discrete form [5]. For wideband beamformers, in general, the beamwidth will increase with the decrease of frequency since the relative aperture of the array becomes smaller for lower frequencies, and therefore, one unique problem for wideband beamforming is how to design a beamformer with a frequency-invariant beam response or beam pattern.

To achieve a frequency-independent beam response, many methods were proposed in the past, and one typical solution is harmonic nesting, where, for a number of frequency bands, different subarrays with appropriate aperture and sensor spacing are operated [4]. In a design proposed in [12], each sensor in the array is followed by its own primary filter, and the outputs of these primary filters share a common secondary filter to form the final output; although the design for a 1D array is relatively simple due to the dilation property of the primary filters, for 2D and 3D arrays, this property is not guaranteed, which makes the general design case very complicated. In [5] and [13], based on a simple Fourier transform relationship, a systematic and consistent approach was developed to design fixed frequency-invariant beamformers for 1D, 2D, and 3D arrays and for both continuous and discrete apertures.

Furthermore, a series of least-squares-based frequency-invariant beamformer design methods were proposed with closed-form solutions and applicable to arbitrary array geometries [5]. In its very basic form, given the desired beam pattern  $P_d(\Omega, \theta)$  ( $\Omega$  is the normalized frequency) and designed response  $P(\Omega, \theta)$  (a quadratic function of the beamforming weight vector  $\mathbf{w}$ ) over the frequency range of interest  $\Omega_I$  and the range of the angle of interest  $\Theta$ , the design is to minimize the following cost function:

$$\begin{aligned} & \alpha \int_{\Theta} |P(\Omega_r, \theta) - P_d(\Omega_r, \theta)|^2 d\theta \\ & + (1 - \alpha) \int_{\Omega_I} \int_{\Theta} |P(\Omega, \theta) - P(\Omega_r, \theta)|^2 d\Omega d\theta \end{aligned} \quad (8)$$

where the first part is the traditional cost function for a least-squares-based design over one reference frequency  $\Omega_r$ ; the second part is the term for measuring the difference between the response of the designed beamformer and its response at the reference frequency  $\Omega_r$  over the full range of the angle of interest, i.e., the frequency variation of the response; and  $\alpha$  trades these off. Note that the first part of the cost function is calculated only at the reference frequency, not the whole  $\Omega_I$ , and the reason is that, if the response is frequency invariant, then as long as at one single frequency ( $\Omega_r$ ) the designed response is close to the desired one, the whole response will also be close to it. A design example for a frequency-invariant beamformer, over the normalized frequency range  $[0.3\pi, \pi]$ , based on a uniform linear array (ULA) of 10 sensors and a TDL length of 20 is shown in Figure 1.

The preceding FIB design techniques can be employed to design a FIB network, where multiple frequency-invariant beamformers pointing to different directions are placed in parallel to transform the wideband array signal processing problem into a narrowband one so that traditional narrowband beamforming and DOA estimation solutions can be applied directly to the output of the FIB network [5]; the second part of the cost function in (8) can also be incorporated into the adaptive beamforming process to realize adaptive FIB directly instead of relying on the FIB network [14].

Note that the TDL-based wideband beamforming structure could be replaced by the sensor delay line (SDL)-based structure [5], [15], where multiple sensors are placed behind the original array sensors in place of the delay lines for effective wideband beamforming; such an SDL-based structure may prove to be very important for the coming terahertz (THz) and sub-THz communication systems, where the delays required for effective wideband beamforming/beam steering may be too short to be implemented in practice.

### DOA estimation

DOA estimation is another core SAM research area. Originally, it was realized by various beamforming algorithms in its simplest form, such as the Butler matrix, the Capon beamformer, and the LCMV beamformer, and then more advanced super-resolution solutions were developed under the classic subspace framework. In the past 25 years, inspired by developments of compressive sensing (CS) [16], two important advances in this area are the sparsity-based DOA estimation framework [17], [18], which, unlike the subspace-based framework, can deal with coherent sources directly, and the underdetermined DOA estimation approach based on various signal properties (such as noncircularity and non-Gaussianity) and the coarray concept (both sum and difference coarrays) [17], [19], [20], [21]. (Here, “underdetermined” means that the number of signals is larger than or equal to the number of physical sensors.)

### Sparsity-based DOA estimation

To introduce the basic idea for sparsity-based DOA estimation, consider the following discrete version of the continuous model in (1):

$$\mathbf{x}[i] = \mathbf{A}\mathbf{s}[i] + \mathbf{n}[i] \quad (9)$$

where  $\mathbf{x}[i]$  is the array data vector for the  $i$ th snapshot,  $\mathbf{s}[i]$  is the source signal vector, and  $\mathbf{n}[i]$  is the noise vector.

For the  $i$ th snapshot, to exploit the spatial sparsity property of the source signals, a search grid of  $K_g$  ( $K_g \gg K$ ) potential incident angles  $\theta_{g,0}, \dots, \theta_{g,K_g-1}$  is first generated, and an overcomplete representation of  $\mathbf{A}$  is then constructed, given by

$$\mathbf{A}(\boldsymbol{\theta}_g) = [\mathbf{a}(\theta_{g,0}), \dots, \mathbf{a}(\theta_{g,K_g-1})]. \quad (10)$$

Here,  $\mathbf{A}(\boldsymbol{\theta}_g)$  is independent of the actual source directions  $\theta_k$ . We also construct a  $K_g \times 1$  column vector  $\mathbf{s}_g[i]$ , with each entry representing a potential source at the corresponding angle. Then, the model, from the perspective of sparse signal reconstruction, becomes

$$\mathbf{x}[i] = \mathbf{A}(\boldsymbol{\theta}_g)\mathbf{s}_g[i] + \mathbf{n}[i]. \quad (11)$$

Now the sparsity-based DOA estimation for a single snapshot can be formulated as

$$\begin{aligned} \min \quad & \|\mathbf{s}_g[i]\|_0 \\ \text{subject to} \quad & \|\mathbf{x}[i] - \mathbf{A}(\boldsymbol{\theta}_g)\mathbf{s}_g[i]\|_2 \leq \varepsilon \end{aligned} \quad (12)$$

where  $\|\cdot\|_0$  is the  $\ell_0$  norm to promote sparsity in  $\mathbf{s}_g[i]$ . Locations of the nonzero entries in the resultant  $\mathbf{s}_g[i]$  represent the corresponding DOA estimation results.

Since the  $\ell_0$  norm is nonconvex, in practice, it is normally replaced by the  $\ell_1$  norm as an approximation. Finally, the sparsity-based DOA estimation for a single snapshot is formulated as

$$\begin{aligned} \min \quad & \|\mathbf{s}_g[i]\|_1 \\ \text{subject to} \quad & \|\mathbf{x}[i] - \mathbf{A}(\boldsymbol{\theta}_g)\mathbf{s}_g[i]\|_2 \leq \varepsilon \end{aligned} \quad (13)$$

where  $\|\cdot\|_1$  is the  $\ell_1$  norm.

When multiple data snapshots are available, we could perform DOA estimation by (12) for each snapshot  $i$  separately. However, a more effective approach is to jointly estimate the DOAs of the impinging signals across multiple snapshots by employing the group sparsity concept since they all have the same spatial support.

Denote  $\mathbf{X} = [\mathbf{x}[0], \dots, \mathbf{x}[P-1]]$ , where  $P$  is the number of snapshots. Similarly, we can define  $\mathbf{S} = [\mathbf{s}[0], \dots, \mathbf{s}[P-1]]$  and  $\mathbf{N} = [\mathbf{n}[0], \dots, \mathbf{n}[P-1]]$ . Then, the signal model for multiple snapshots can be obtained by

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}. \quad (14)$$

To introduce spatial sparsity, similar to the single-snapshot case, we construct  $\mathbf{S}_g = [\mathbf{s}_g[0], \dots, \mathbf{s}_g[P-1]]$  and use the row vector  $\mathbf{s}_{g,k_g}$ ,  $0 \leq k_g \leq K_g - 1$  to represent the  $k_g$ th row of the matrix  $\mathbf{S}_g$ :

$$\mathbf{X} = \mathbf{A}(\boldsymbol{\theta}_g)\mathbf{S}_g + \tilde{\mathbf{N}}. \quad (15)$$

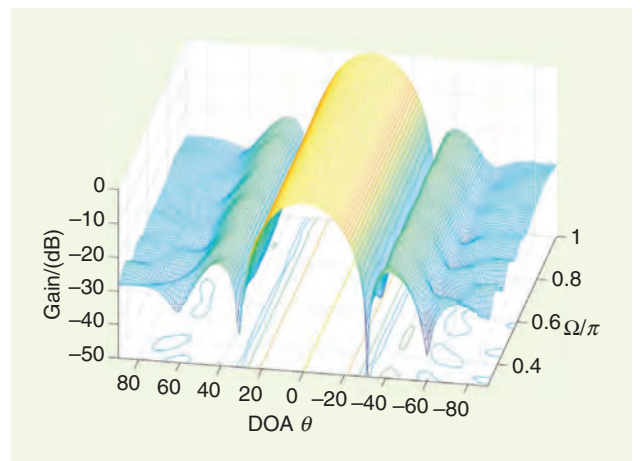


FIGURE 1. A frequency-invariant beamformer.

Then, a new  $K_g \times 1$  column vector is generated by computing the  $\ell_2$  norm of each row in  $\mathbf{S}_g$ , expressed as

$$\hat{\mathbf{s}}_g = [\|\mathbf{s}_{g,0}\|_2, \|\mathbf{s}_{g,1}\|_2, \dots, \|\mathbf{s}_{g,K_g-1}\|_2]^T. \quad (16)$$

Finally, the problem for multiple snapshots can be formulated as

$$\begin{aligned} \min_{\mathbf{S}_g} \quad & \|\hat{\mathbf{s}}_g\|_1 \\ \text{subject to} \quad & \|\mathbf{X} - \mathbf{A}(\theta_g)\mathbf{S}_g\|_F \leq \varepsilon \end{aligned} \quad (17)$$

where  $\|\cdot\|_F$  represents the Frobenius norm and  $\|\hat{\mathbf{s}}_g\|_1$  is also called the  $\ell_{2,1}$  norm of the matrix  $\mathbf{S}_g$ . Locations of the nonzero entries of the resultant column vector  $\hat{\mathbf{s}}_g$  are, then, the corresponding estimation results.

One problem with the preceding group sparsity-based formulation is its high computational complexity, especially when a large number of snapshots  $P$  is available. To reduce the complexity, we can perform singular value decomposition (SVD) to  $\mathbf{X}$  and project the data to a lower-dimension signal space, leading to the so-called  $\ell_1$ -SVD method [22], or use the covariance matrix of the data to form a virtual array directly [23].

## Underdetermined DOA estimation

For underdetermined DOA estimation, although it can be achieved by exploiting the non-Gaussianity, noncircularity, and nonstationarity of the signals, the most important development is through constructing various sparse array structures for virtual coarray generation, such as coprime arrays, nested arrays, and their numerous extensions [24], [25], [26].

For second-order statistics-based coarray generation, one common step is to vectorize the covariance matrix of the physical sparse array. Consider the covariance matrix

$$\mathbf{R}_{\mathbf{xx}} = \mathbb{E}\{\mathbf{x}[i]\mathbf{x}^H[i]\} = \sum_{k=1}^K \sigma_k^2 \mathbf{a}(\theta_k)\mathbf{a}^H(\theta_k) + \sigma_n^2 \mathbf{I}_N \quad (18)$$

where  $\sigma_k^2$  is the power of the  $k$ th impinging signal and  $\theta_k$  is its AOA.

By vectorizing  $\mathbf{R}_{\mathbf{xx}}$ , we obtain a virtual array model

$$\mathbf{z} = \text{vec}\{\mathbf{R}_{\mathbf{xx}}\} = \tilde{\mathbf{A}}(\boldsymbol{\theta})\tilde{\mathbf{s}} + \sigma_n^2 \tilde{\mathbf{I}}_{N^2} \quad (19)$$

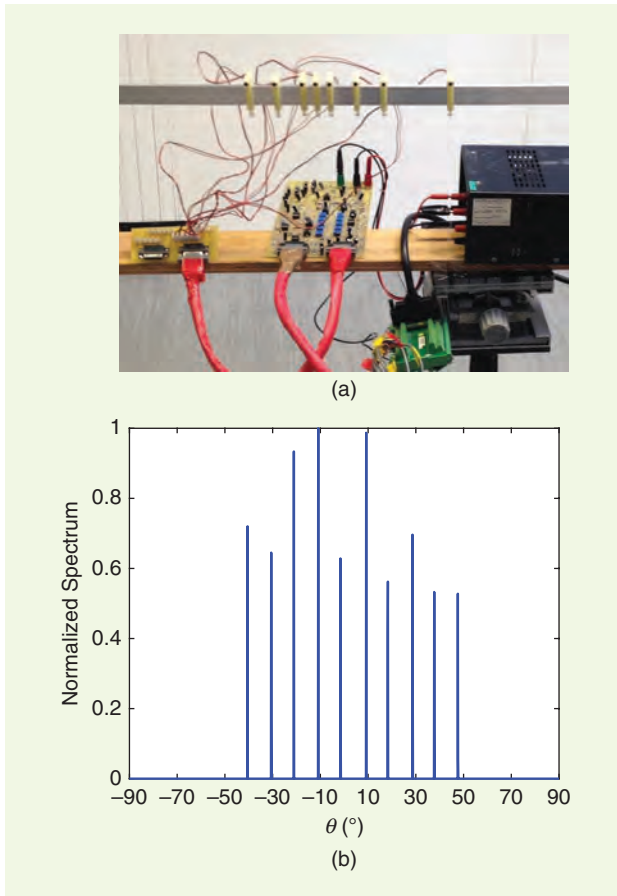
where  $\tilde{\mathbf{A}}(\boldsymbol{\theta}) = [\tilde{\mathbf{a}}(\theta_1), \dots, \tilde{\mathbf{a}}(\theta_K)]$  is the equivalent virtual steering matrix, with  $\tilde{\mathbf{a}}(\theta_k) = \mathbf{a}^*(\theta_k) \otimes \mathbf{a}(\theta_k)$  being the corresponding steering vector ( $\otimes$  denotes the Kronecker product);  $\tilde{\mathbf{s}} = [\sigma_1^2, \dots, \sigma_K^2]^T$  is the equivalent source signals; and  $\tilde{\mathbf{I}}_{N^2}$  is obtained by vectorizing  $\mathbf{I}_N$ .

In the preceding virtual array model, although there are repeated entries in  $\mathbf{R}_{\mathbf{xx}}$ , the number of virtual sensors corresponding to the difference coarray is much more than that of the physical sensors, and the equivalent source signals share the same spatial support with the original impinging signals. The virtual model in (19) is similar to the single-snapshot array model, and sparsity-based DOA estimation methods such as that in (13) can be applied here.

Instead of employing a sparse array, it is possible to extend the coarray concept to different frequencies, where a single ULA can be used with two continuous-wave signals of coprime or other different frequencies, and to the wideband case through frequency decomposition and employing multiple frequency pairs [17].

The group sparsity concept employed for the multiple-snapshot case can be applied to general underdetermined and overdetermined wideband DOA estimation [17]; as in traditional wideband DOA estimation, focusing can also be employed for sparsity-based wideband DOA estimation to simplify the problem to a single reference frequency. One interesting observation about the wideband case is that the sensor spacing can be larger than half the wavelength corresponding to the highest frequency of the signal, without causing the spatial aliasing problem; on the contrary, an improved estimation performance can be achieved for a larger spacing, due to an increased aperture.

Figure 2 gives a real experimental result based on an eight-microphone coprime array for estimating the directions of 10 speech signals, with a bandwidth from 5 to 10 kHz and sampling frequency of 20 kHz [27].



**FIGURE 2.** Group sparsity-based underdetermined wideband DOA estimation, where 10 uncorrelated acoustic source signals are distributed from around  $-40$  to  $50^\circ$ , with an approximate step size of  $10^\circ$  [27]. (a) The coprime microphone array system. (b) The estimation result for the 10 sources.

### Sensor location optimization

In many applications, the array's geometrical layout is assumed to be fixed and given in advance. However, it is possible to change the geometrical layout of the array, including the adjacent sensor spacing, and these additional spatial degrees of freedom (DOF) can be exploited to improve the performance in terms of beamforming, DOA estimation, or both. For the beamforming side, given the nonconvex nature of the optimization problem, traditionally, it is solved by genetic algorithms, simulated annealing, and similar approaches [28]. With the development of CS and the sparsity maximization framework, a new CS-based framework with a theoretically optimum solution (given the convex nature of the formulated problem) has been developed for sensor location optimization for fixed beamforming [29], followed by further work in adaptive beamforming [30], [31], with robustness against various array model errors considered, too. For the DOA estimation side, the main efforts have been focused on the coarray design to increase the DOF for underdetermined DOA estimation. As mentioned in the previous section, coprime arrays and nested arrays are two representative array structures [24], [25], based on which numerous second-order and fourth-order (and even higher) coarray construction methods have been developed. In this part, we focus on the sparse array design problem for beamforming.

To illustrate how the design works, consider a narrowband linear array structure consisting of  $M$  omnidirectional sensors, where the distance from the first sensor to subsequent sensors is denoted as  $d_m$ , for  $m = 1, 2, \dots, M$ , with  $d_1 = 0$ , i.e., the distance from the first sensor to itself. The output of the beamformer is a weighted sum of the received signals, and the weighting coefficients are denoted by  $w_m$  and  $m = 1, 2, \dots, M$ , which are placed together into the weighting vector  $\mathbf{w}$ . Then, the sparsity-based design for sensor location optimization can be described as follows.

First, consider the array geometry being a grid of potential active antenna locations. In this instance,  $d_M$  is the maxi-

imum aperture of the array, and the values of  $d_m$ , for  $m = 1, 2, \dots, M-1$ , are selected to give a uniform grid, with  $M$  being a very large number so that the spacing between adjacent antennas is very small. Through selecting the minimum number of nonzero-valued weight coefficients to generate a beam response close to the desired one, a sparse array design result is obtained. In other words, if a weight coefficient is zero valued, the corresponding sensor will be inactive and therefore can be removed, leading to a sparse or nonuniformly spaced sensor array.

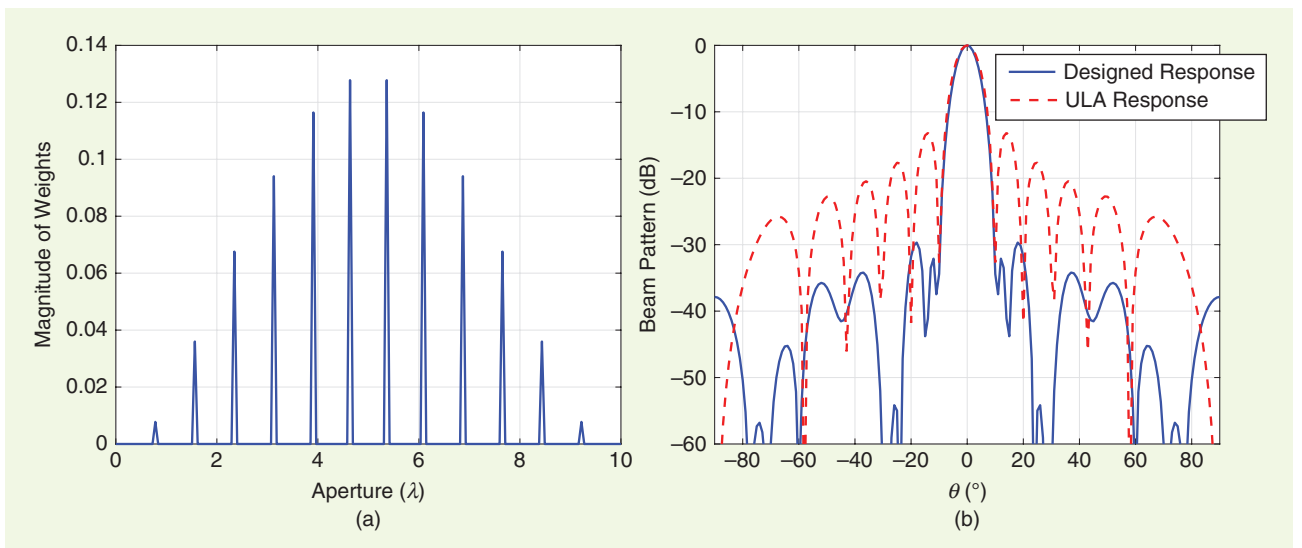
Mathematically, it is formulated as a constrained  $\ell_1$ -norm minimization problem

$$\min \quad \|\mathbf{w}\|_1 \quad (20)$$

$$\text{subject to} \quad \|\mathbf{p}_r - \mathbf{w}^H \mathbf{A}\|_2 \leq \varepsilon \quad (21)$$

where  $\mathbf{p}_r$  is the vector holding the desired beam responses at the sampled angular range of interest;  $\mathbf{A}$  is the steering matrix corresponding to those angles, with  $\mathbf{w}^H \mathbf{A}$  representing the designed beam responses; and  $\varepsilon$  is the allowed error between the designed and desired beam responses. The minimization of the  $\ell_1$  norm of the weight vector helps to promote sparsity in the weight vector, and the reweighted  $\ell_1$ -norm minimization could be used instead to have a closer approximation to the ideal  $\ell_0$ -norm minimization problem, where smaller weighting terms are added to the larger elements of the weight vector  $\mathbf{w}$  so that smaller values in  $\mathbf{w}$  are penalized more and become closer to zero after minimization [32].

A broadside main beam design example is provided in Figure 3, where the sensor locations are optimized over an overall aperture of  $d_M = 10\lambda$ , which is split into 181 potential sensor locations ( $M = 181$ ). It can be seen that the resultant weight vector is sparse, with only 12 nonzero-valued coefficients, leading to a sparse array of 12 sensors, and compared to the beam pattern of a standard 12-sensor ULA with



**FIGURE 3.** The location optimization result and comparison with the ULA. (a) The magnitude of the weights. (b) The array beam patterns.

half-wavelength spacing, the sparse array has a similar main beamwidth but a much lower sidelobe level.

Various constraints can be added to the preceding formulation to deal with more complicated application scenarios. For example, in the preceding formulation, the steering vector of the array is assumed to be known exactly, which may not be true due to various possible model perturbations, such as errors in sensor locations, mutual coupling, and discrepancies in individual sensor responses; then, robust designs can be achieved by applying a norm-bounded error constraint to the weight vector. In another case, it has been assumed that the sensors in the array are of zero size; however, this is not true in the real world, and various size constraints can be added to the design, and some postprocessing methods can be introduced to make sure the minimum spacing between adjacent sensors in the result is larger than the size of the sensor. Based on the concept of group sparsity, the design can also be extended to the wideband case with TDLs [29].

#### Target/source localization based on sensor arrays

This is another important problem in array signal processing, and significant progress has been made in this area in the past 25 years. Typical solutions include those based on the received signal strength [33]; those based on distance-related measurements, such as the time of arrival [34]; and those based on the AOA/DOA [35], [36]. The last is also called *bearing-only localization*, and it is an attractive candidate since synchronization among the distributed platforms is not required, and it can be used in both active and passive sensing networks and adopted in a wide range of applications, including multistatic radar, distributed massive MIMO, and wireless sensor networks. There are normally two steps in this bearing-only localization: the first is applying existing DOA estimation methods to find the AOAs at all distributed sensor arrays, while the

second is to find intersections of those estimated AOAs to localize the sources, and the maximum likelihood estimator has been adopted to minimize the total least-squares errors of the noise-corrupted angle measurements among all distributed sensor arrays. However, the performance of such a two-step localization approach is dependent on the accuracies of angle measurements obtained at all platforms, and even one bad AOA estimation result can lead to a serious performance degradation.

To tackle the shortcomings of the two-step approach, we could jointly process the collected information across the observation platforms in lieu of fusing the separate angle estimation results at all platforms. One recent advance in this direction is a group sparsity-based one-step approach

[37], where a common spatial sparsity support corresponding to all distributed sensor arrays is enforced, leading to a better estimation performance, which also avoids the possible pairing and ambiguity problems associated with the two-step AOA-based solution.

To show how this idea works, we consider a distributed narrowband sensor array network with  $M$  subarrays and  $K$  targets, as illustrated in Figure 4, where  $U_m(x_m, y_m)$  and  $T_k(x_{T_k}, y_{T_k})$  represent locations of the receiver platform and the  $k$ th target, respectively. For each receiver, a linear subarray with  $L_m$  sensors is employed.

For each target located at  $T_k(x_{T_k}, y_{T_k})$ , a unique incident angle  $\theta_{m,k}$  relative to the  $m$ th subarray can be obtained. Without loss of generality, a square area of interest in the Cartesian coordinate system is divided into a  $K_x \times K_y$  grid, with  $K_x$  and  $K_y$  being the number of grid points along the  $x$ -axis and the  $y$ -axis, respectively. Here,  $G(x_{k_s}, y_{k_s})$  represents the location of the  $(k_x, k_y)$ th search grid, and the signal originating from the possible source located at  $G(x_{k_s}, y_{k_s})$  will arrive at the  $m$ th subarray, with a DOA angle  $\theta_m^g(k_x, k_y)$ . Since  $(x_{k_s}, y_{k_s})$  is common to all subarrays and a source located at  $G(x_{k_s}, y_{k_s})$  will appear to come from the same location with respect to all subarrays, we can apply the group sparsity concept to all subarrays' source data.

For example, for the  $m$ th subarray, corresponding to the data model in (14), we can have the multiple-snapshot model as

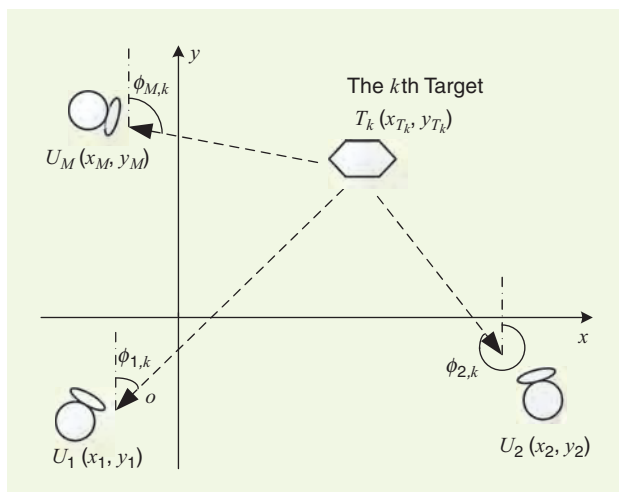
$$\mathbf{X}_m = \mathbf{A}_m \mathbf{S}_m + \mathbf{N}_m \quad (22)$$

with  $m = 1, 2, \dots, M$ . Applying the sparsity-based approach, we can construct the following overcomplete data model:

$$\mathbf{X}_m = \mathbf{A}_m^g \mathbf{S}_m^g + \mathbf{N}_m \quad (23)$$

where  $\mathbf{A}_m^g$  is the overcomplete steering matrix corresponding to the  $K_x K_y$  potential signal directions  $\theta_m^g(k_x, k_y)$  and  $\mathbf{S}_m^g$  is the potential source matrix. If there is no source located at a particular position  $G(x_{k_s}, y_{k_s})$ , then the corresponding row of  $\mathbf{S}_m^g$  will be zero valued for all  $m = 1, 2, \dots, M$ . We can place all the matrices  $\mathbf{S}_m^g$  together to form a new matrix  $\mathbf{S}^g$ , as follows:

**One unique problem for wideband beamforming is how to design a beamformer with a frequency-invariant beam response or beam pattern.**



**FIGURE 4.** A general target/source localization model based on distributed sensor arrays [37].



$$\mathbf{S}^g = [\mathbf{S}_1^g, \mathbf{S}_2^g, \dots, \mathbf{S}_M^g]. \quad (24)$$

Then, the group sparsity-based localization problem can be formulated by minimizing the  $\ell_{2,1}$  norm of  $\mathbf{S}^g$ , subject to limiting the overall reconstruction error for all subarrays to a small value. One main advantage of the group sparsity-based approach for direct target localization is that the different subarrays are not required to be synchronized and can work on different frequencies, the statistical properties of the sources can be different for different subarrays, and sensor numbers, rotation angles, and corresponding source signals of different subarrays do not need to be the same (as long as they come from the same set of target locations). This group sparsity-based one-step direct localization idea can be extended to the wideband, the underdetermined case, or both without difficulty [37].

Figure 5 displays a simulation result for underdetermined wideband localization, where the normalized signal frequency band is from  $0.75\pi$  to  $\pi$ , and there are six subarrays and five targets, with each subarray being a four-sensor minimum redundancy array [4].

### MIMO arrays

MIMO, which is, by its multichannel implementation at both the transmitter and the receiver, a natural fit within the SAM portfolio, represents another significant development in array signal processing in the past 25 years. There are mainly two totally different directions. One is MIMO radar, which exploits the orthogonality of the transmitted waveforms to increase the DOF of the system to improve the resolution and capacity of the array [38], [39], [40], which will play an important part in 4D auto radar imaging in addition to traditional radar detection applications. Note that nonorthogonal waveforms can also be employed for MIMO radar [41]. The other one is MIMO for wireless communications to exploit the spatial diversity of the channel to improve the performance and, in particular, the capacity of the communication system [42]. While MIMO has already been in use for both Wi-Fi and 4G communication systems, its new evolution, the so-called massive MIMO, or ultramassive MIMO (UM-MIMO), will play a crucial role in next-generation communication systems and beyond [43].

### MIMO radar

In a MIMO radar, multiple transmit antennas emit orthogonal waveforms and multiple receive antennas, then receive the echoes reflected by the targets. Antennas of the MIMO radar can be widely separated [38] and colocated [39], [40], with the latter more widely studied. For the case with colocated antennas, the transmitting side and the receiving side can be located either at the same site or far away from each other.

Consider a colocated narrowband MIMO radar system where the transmit and receive antennas are located at the same place. The transmitted multiple orthogonal waveforms are then reflected back by  $K$  present targets and received by the receive array. After matched filter processing, the output

signal vector  $\mathbf{x}[i]$  at the receiver at the  $i$ th snapshot can be expressed as

$$\begin{aligned} \mathbf{x}[i] &= \sum_{k=1}^K \mathbf{a}_t(\theta_k) \otimes \mathbf{a}_r(\theta_k) b_k[i] + \mathbf{n}[i] \\ &= [\mathbf{a}_t(\theta_1) \otimes \mathbf{a}_r(\theta_1), \dots, \mathbf{a}_t(\theta_K) \otimes \mathbf{a}_r(\theta_K)] \mathbf{b}[i] + \mathbf{n}[i] \end{aligned} \quad (25)$$

where  $\theta_k$  is the DOA of the  $k$ th target;  $\mathbf{a}_t(\theta_k)$  and  $\mathbf{a}_r(\theta_k)$  are the steering vectors of the transmit and receive arrays, respectively; and  $b_k[i] = \gamma_k e^{j2\pi f_k i}$ , with  $\gamma_k$  being the complex-valued reflection coefficient of the  $k$ th target and  $f_k$  being the Doppler frequency for moving targets.

It can be seen that with the MIMO radar configuration, a virtual array with a significantly increased aperture has been created due to the effect of the Kronecker product in (25). For example, if both the transmit array and receive array are three-sensor ULAs with a spacing of  $d$  and  $3d$ , respectively, the newly generated virtual ULA will consist of nine virtual sensors. In this way, by exploiting waveform diversity, a virtual array with a much larger aperture and significantly increased DOF is formed using a small number of physical sensors, providing enhanced spatial resolution, higher target detection capacity, and better performance.

### MIMO for wireless communications

On the other hand, MIMO for wireless communications is a huge research area, and numerous techniques have been developed centered around this concept, such as space-time coding, MIMO beamforming, spatial multiplexing, and spatial modulation. Today, an element of MIMO can be found in most of the publications in wireless communications. It is impossible to list all the important advances in the area, and in this section, we focus only on MIMO beamforming, which is playing an increasingly important role in the implementation of MIMO communication systems.

As well known by the array signal processing community and also presented in the ‘‘Beamforming’’ section, traditionally, beamforming is designed for line-of-sight (LOS) transmission

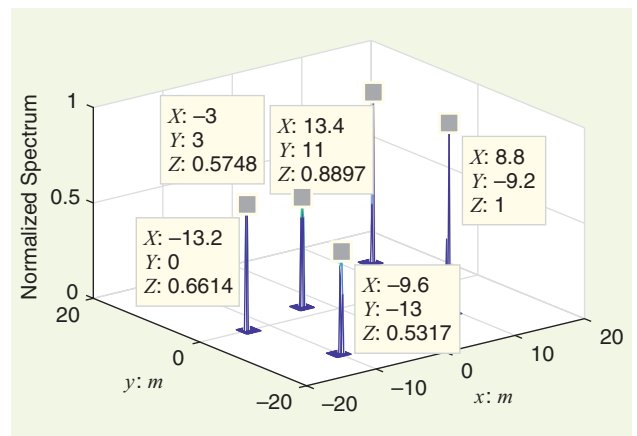


FIGURE 5. The localization results for five targets with six subarrays (20-dB signal-to-noise ratio, 500 snapshots).

and reception, and physically, a beam will be formed in the process, pointing to different directions around the array system. However, in MIMO beamforming, due to a very strong multipath effect, the result of beamforming between the transmitter and receiver will not necessarily form a beam in space but, rather, an overall enhanced signal transmission link between them. Decades of research in MIMO beamforming have pushed the boundaries of beamforming well beyond the technology's traditional meaning, and today, any process achieving enhancement of the desired signal while reducing the effect of interference can be considered beamforming. However, with the introduction of massive MIMO and millimeter-wave (mm-wave) communications in 5G and beyond, the LOS case is becoming more and more important again in MIMO beamforming, and one interesting development in this context is the hybrid beamforming structure proposed for massive MIMO systems [8].

Hybrid beamforming is a combination of analog beamforming and digital beamforming. Ideally, beamforming could be implemented completely in the digital domain for maximum flexibility and adaptability; however, for extremely large arrays, as in the case of massive/UM-MIMO, the extremely high cost associated with the large number of high-speed analog-to-digital converters (ADCs)/digital-to-analog converters (DACs) and the high-level power consumption will render it practically infeasible. For hybrid beamforming in the receive mode, analog beamforming is performed first to reduce the number of analog channels, which are then converted into digital via a reduced number of ADCs, and after that, digital beamforming can be performed; for the transmit mode, the process is simply reversed. There are many hybrid beamforming structures proposed in the literature, and one representative is the subaperture-based hybrid beamformer. An interesting recent development in this area is a new class of multibeam multiplexing designs, where the number of analog coefficients is the same as the number of antennas, independent of the number of parallel independent user beams generated, while the number of subarrays is the same as the number of beams; interested readers can refer to [44] and [45] for details.

### **New developments in the SAM area**

In the era of AI, multi-sensor-based systems and techniques are ubiquitous and will play an even greater role in the future. As a result, there has been an exponential increase in research activities in the SAM area in the past few years, and in the following, we introduce some new developments that may well indicate promising future research directions.

#### *GSP for sensor networks*

GSP is an emerging new mathematical tool for analysis of data resident on a largely irregular network of either physical or virtual sensor nodes, where the regular network can be considered a special case [46], [47]. Examples for the physical sensor

network include traffic networks, brain neural networks, and energy consumption sensor networks, while for the virtual one, a good example is social networks. In connection with classic signal processing, basic concepts, such as frequency, and operations, such as shift/delay and filtering, have been introduced.

However, there is still no unified framework for GSP, and it is still an open problem to find the best representations of a graph signal. However, this has not stopped the wide application of GSP, and it has been shown to be a powerful data analysis tool providing new insights into the studied problems; for example, brain signals can be mapped to a graph network to analyze cognitive behavior of the brain.

Application of GSP to traditional sensor array signal processing problems, such as direction finding and target localization, is an emerging but somewhat open area, as traditional sensor arrays and networks normally have a regular structure, and traditional sensor array signal processing tools have been extremely successful in tackling those associated problems. It is not clear yet whether GSP can bring any advantage to the traditional sensor array signal processing problems or not.

#### *Tensor-based array signal processing*

Tensors are extensions of matrices to higher dimensions and have been widely employed for multidimensional data analysis and processing with the aid of tensor decomposition tools and algorithms. Many sensor array signals and data can be transformed into a multidimensional form and viewed directly as a multidimensional structure [48], [49]. For example, the narrowband data received by a rectangular array and multiple subarrays are 3D, the data received by a wideband linear array can be transformed into the 3D space–time–frequency domain, and the data received by vector sensor arrays are naturally higher dimensional. For MIMO communication systems, the data can be placed into a tensor form by accounting for diversities in space, time, frequency (including Doppler frequency), and polarization. As a result, tensor processing can be applied to solve many array signal processing problems directly without much adaptation. However, although it is recognized that tensors can keep the inherent data structures and therefore have the potential to provide improved performance compared to classic array signal processing methods and algorithms, further research is needed to demonstrate the clear benefits of tensor processing and fully realize its potential.

#### *Quaternion-valued array signal processing*

As a higher-dimensional extension of complex numbers, a quaternion has one real part and three imaginary parts, and quaternion calculus has been applied to a range of signal processing problems related to 3D and 4D signals, such as color image processing, wind profile prediction, vector sensor array processing, and quaternion-valued wireless communications [50], [51]. In addition to solving the classic array signal processing

**Today, any process achieving enhancement of the desired signal while reducing the effect of interference can be considered beamforming.**

problems, such as DOA estimation and beamforming, one important development is the quaternion-valued MIMO array, where pairs of antennas with orthogonal polarization directions are employed at both the transmitter and receiver sides and a 4D modulation scheme across the two polarization diversity channels using a quaternion-valued representation is employed. Although the polarization states will change during transmission through the channel, and there may be interferences between these two states, we can employ a quaternion-valued adaptive algorithm to recover the original 4D signal, which inherently also performs an interference suppression operation to separate the original two 2D signals. For the MIMO array, reference signal-based and blind quaternion-valued adaptive algorithms can be employed for both channel estimation and beamforming. Signal processing has experienced a revolutionary change from real-valued processing to complex-valued processing, and we may be at the doorstep of a quaternion-valued world, and increasing interest in quaternion-valued sensor array signal processing is expected in the near future.

**Signal processing has experienced a revolutionary change from real-valued processing to complex-valued processing, and we may be at the doorstep of a quaternion-valued world.**

### *One-bit and noncoherent sensor array signal processing*

Given the extremely high data rate and storage requirements for a fully digital large sensor array system, there has been significant work aimed at achieving a reasonable sensor array processing performance with 1-bit representation of the array signals; i.e., only signs of the data samples are reserved, while the magnitude information is removed [52]. This problem can be simply considered the normal case but with extremely high quantization noise, and we can perform normal array processing irrespective of the number of bits per data sample; however, a more effective way is to try to achieve effective estimation of the statistics of the signals using the 1-bit data samples and then, based on the newly obtained statistics information, perform the corresponding tasks. Contrary to 1-bit array processing, the signs of the data samples are removed, and only the magnitude information is kept, which leads to the so-called noncoherent sensor array signal processing problem, with the advantage of being robust against array phase errors. One representative example is noncoherent DOA estimation and target localization [53], [54], [55], which can be cast into a phase retrieval problem; however, the difference is that there is usually only one snapshot in phase retrieval, while in array signal processing, multiple snapshots are available, which can be exploited by applying group sparsity to existing phase retrieval algorithms, such as the ToyBar and modified GESPAR algorithms [53], [54].

### *Machine learning and AI for sensor arrays*

Machine learning and AI have been applied to almost all areas of research in the signal processing community, and the SAM area is no exception. For example, machine learning and

AI have been applied to DOA estimation, beamforming, and source separation successfully [56]. There are strong topical connections among sparsity-inspired array processing (see the “DOA Estimation” section), compressed sensing (see the “Sensor Location Optimization” section), and machine learning.

Unlike in traditional machine learning and AI applications, where it is a challenge to acquire sufficient training data, in most of the array signal processing applications, the required training data can be obtained easily by simulation. Nonetheless, their application to array signal processing also faces some similar issues. For example, after training, the system may work very well for the targeted scenario, but it may struggle if

there is change to the system and the environment, while the traditional array signal processing methods and algorithms can cope with such changes well. Another challenge is how to apply machine learning and AI to distributed sensor arrays and networks effectively. As a hot topic, federated learning may prove to be a promising direction of research for the SAM community [57].

### *Array signal processing for next-generation communication systems*

Antenna array design and signal processing is one of the fundamental techniques in 5G (and beyond) wireless communication systems since the two underpinning 5G/6G technologies—massive MIMO/UM-MIMO and mm-wave/sub-THz/THz communications—are all based on antenna arrays [58]. It will continue to play a significant role in many other aspects in the future, such as the Internet of Things and integrated sensing and communication, both of which are hot topics for 6G wireless communications research, with extensive research activities in the community. Moreover, beamforming is essential to achieve effective communication over the THz and sub-THz frequency bands, as it is necessary to employ a large number of antennas for such high frequencies, while the widely studied reconfigurable intelligent surfaces can be considered semipassive antenna array systems [59]. To a great degree, array signal processing will be a main focus of research for next-generation communication systems and for the integration of sensing and communications, particularly at mm-waves [60].

### **Concluding remarks**

Accompanied by intensive research activities and the significant progress made in signal processing, the world now has stepped into the new era of AI, where multi-sensor-based systems and techniques have become ubiquitous and indispensable to our daily life and will play an even greater role in our society in the very near future. This is an exciting time for the SAM community, and we welcome new members at different levels to join the TC and work together to promote its activities, make a more extensive and deeper impact in the real world, and further enhance its standing in our wider society.

## Acknowledgment

This work is supported by the U.K. Engineering and Physical Sciences Research Council, under grant EP/V009419/1. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version arising.

## Authors

**Wei Liu** (w.liu@sheffield.ac.uk) received his Ph.D. degree from the University of Southampton, U.K., in 2003. He is currently with the Department of Electronic and Electrical Engineering, University of Sheffield, S1 3JD Sheffield, U.K. He has published 210+ journal papers, 170+ conference papers, and two research monographs [*Wideband Beamforming: Concepts and Techniques* (Wiley, 2010) and *Low-Cost Smart Antennas* (Wiley, 2019)] in the area of signal processing, with a particular focus on sensor array signal processing and its various applications, such as robotics and autonomous systems, radar, sonar, and wireless communications. He is a member of the IEEE Signal Processing Society Sensor Array and Multichannel Technical Committee (2021–2022 chair) and the IEEE Circuits and Systems Society Digital Signal Processing Technical Committee (2022–2024 chair), and he is an IEEE Aerospace and Electronic Systems Society Distinguished Lecturer for 2023–2024.

**Martin Haardt** (martin.haardt@tu-ilmenau.de) received his Ph.D. degree from Munich University of Technology in 1996. He is currently a full professor and the head of the Communications Research Laboratory, Ilmenau University of Technology, D-98684 Ilmenau, Germany. He chaired the IEEE Signal Processing Society (SPS) Sensor Array and Multichannel Technical Committee (TC) during 2017–2018, and he has been an elected member of the SPS Signal Processing Theory and Methods TC since 2020. He received a 2009 SPS Best Paper Award; the Vodafone Innovations Award for outstanding research in mobile communications; the Association of Electrical Engineering, Electronics, and Information Technology ITG Best Paper Award; and the Rohde & Schwarz Outstanding Dissertation Award. His research interests include wireless communications, array signal processing, high-resolution parameter estimation, and numerical linear and multilinear algebra. He is a Fellow of IEEE.

**Maria S. Greco** (maria.greco@unipi.it) received her Ph.D. degree from University of Pisa in 1998. She is a full professor in the Department of Information Engineering, University of Pisa, 56122 Pisa, Italy. She has coauthored many book chapters and more than 220 journal and conference papers and is the editor-in-chief of *EURASIP Journal of Advances in Signal Processing*. She is the president-elect of the IEEE Aerospace and Electronic Systems Society (AESS) for 2022–2023. Previously, she was the IEEE SPS director at large for Region 8 (2021–22), a member of the SPS Board of Governors (2015–2017), an SPS Distinguished Lecturer (2014–2015), and an AESS Distinguished Lecturer (2015–2020). She was a corecipient of the 2001 and 2012 AESS

Barry Carlton Award, the 2008 AESS Fred Nathanson Young Engineer of the Year award, and the AESS Board of Governors Exceptional Service Award. Her research interests include clutter models, cognitive radars, and the integration of sensing and communications.

**Christoph F. Mecklenbräuker** (cfm@tuwien.ac.at) received his Dr.-Ing. degree (with honors) from Ruhr University Bochum, Germany, in 1998. He is currently a full professor at the Institute of Telecommunications, TU Wien, 1040 Vienna, Austria. He chaired the IEEE Signal Processing Society Sensor Array and Multichannel Technical Committee during 2019–2020. His doctoral dissertation received the 1998 Gert-Massenberg Prize, and he has authored approximately 300 journal and conference papers and was granted several patents in the field of mobile cellular networks. His research interests include 5G/6G radio interfaces (vehicular connectivity and sensor networks) and antennas and propagation.

**Peter Willett** (peter.willett@uconn.edu) received his Ph.D. degree from Princeton University in 1986. He is currently a professor in the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA. He chaired the IEEE Signal Processing Society Sensor Array and Multichannel Technical Committee during 2015–2016. His research interests include statistical signal processing, detection, machine learning, communications, data fusion, and tracking. He is a Fellow of IEEE.

## References

- [1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988, doi: 10.1109/53.665.
- [2] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996, doi: 10.1109/79.526899.
- [3] L. C. Godara, "Application of antenna arrays to mobile communications, part II: Beam-forming and direction-of-arrival estimation," *Proc. IEEE*, vol. 85, no. 8, pp. 1195–1245, Aug. 1997, doi: 10.1109/5.622504.
- [4] H. L. Van Trees, *Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 2002.
- [5] W. Liu and S. Weiss, *Wideband Beamforming: Concepts and Techniques*. Chichester, U.K.: Wiley, 2010.
- [6] M. Haardt, C. Mecklenbrauker, and P. Willett, "Highlights from the sensor array and multichannel technical committee: Spotlight on the IEEE signal processing society technical committees," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 183–185, Sep. 2018, doi: 10.1109/MSP.2018.2835718.
- [7] J. Li and P. Stoica, Eds. *Robust Adaptive Beamforming*. Hoboken, NJ, USA: Wiley, 2005.
- [8] A. F. Molisch et al., "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017, doi: 10.1109/MCOM.2017.1600400.
- [9] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003, doi: 10.1109/TSP.2002.806865.
- [10] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, Jan. 2010, doi: 10.1109/TAES.2010.5417172.
- [11] Y. Gu and A. Leshem, "Robust adaptive beamforming based on interference covariance matrix reconstruction and steering vector estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3881–3885, Jul. 2012, doi: 10.1109/TSP.2012.2194289.
- [12] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns," *J. Acoust. Soc. Amer.*, vol. 97, no. 2, pp. 1023–1034, Feb. 1995, doi: 10.1121/1.412215.

- [13] W. Liu and S. Weiss, "Design of frequency invariant beamformers for broadband arrays," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 855–860, Feb. 2008, doi: 10.1109/TSP.2007.907872.
- [14] Y. Zhao, W. Liu, and R. J. Langley, "Adaptive wideband beamforming with frequency invariance constraints," *IEEE Trans. Antennas Propag.*, vol. 59, no. 4, pp. 1175–1184, Apr. 2011, doi: 10.1109/TAP.2011.2110630.
- [15] W. Liu, "Adaptive wideband beamforming with sensor delay-lines," *Signal Process.*, vol. 89, no. 5, pp. 876–882, May 2009, doi: 10.1016/j.sigpro.2008.11.005.
- [16] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006, doi: 10.1109/TIT.2005.862083.
- [17] Q. Shen, W. Liu, W. Cui, and S. L. Wu, "Underdetermined DOA estimation under the compressive sensing framework: A review," *IEEE Access*, vol. 4, pp. 8865–8878, Nov. 2016, doi: 10.1109/ACCESS.2016.2628869.
- [18] Z. Yang, J. Li, P. Stoica, and L. H. Xie, "Sparse methods for direction-of-arrival estimation," in *Array, Radar and Communications Engineering*, R. Chellappa and S. Theodoridis, Eds. New York, NY, USA: Academic, 2018, ch. 11, pp. 509–581.
- [19] P. Chevalier, L. Albera, A. Férréol, and P. Comon, "On the virtual array concept for higher order array processing," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1254–1271, Apr. 2005, doi: 10.1109/TSP.2005.843703.
- [20] P. Chevalier, A. Férréol, and L. Albera, "High-resolution direction finding from higher order statistics: The 2g-MUSIC algorithm," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 2986–2997, Aug. 2006, doi: 10.1109/TSP.2006.877661.
- [21] P. Pal and P. P. Vaidyanathan, "Multiple level nested array: An efficient geometry for 2gth order cumulant based array processing," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1253–1269, Mar. 2012, doi: 10.1109/TSP.2011.2178410.
- [22] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005, doi: 10.1109/TSP.2005.850882.
- [23] J. H. Yin and T. Q. Chen, "Direction-of-arrival estimation using a sparse representation of array covariance vectors," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4489–4493, Sep. 2011, doi: 10.1109/TSP.2011.2158425.
- [24] P. Pal and P. P. Vaidyanathan, "Nested arrays: A novel approach to array processing with enhanced degrees of freedom," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4167–4181, Aug. 2010, doi: 10.1109/TSP.2010.2049264.
- [25] P. P. Vaidyanathan and P. Pal, "Sparse sensing with co-prime samplers and arrays," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 573–586, Feb. 2011, doi: 10.1109/TSP.2010.2089682.
- [26] S. Qin, Y. D. Zhang, and M. G. Amin, "Generalized coprime array configurations for direction-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 6, pp. 1377–1390, Mar. 2015, doi: 10.1109/TSP.2015.2393838.
- [27] Q. Shen, W. Liu, W. Cui, S. L. Wu, Y. D. Zhang, and M. G. Amin, "Low-complexity direction-of-arrival estimation based on wideband co-prime arrays," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1445–1456, Sep. 2015, doi: 10.1109/TASLP.2015.2436214.
- [28] A. Trucco and V. Murino, "Stochastic optimization of linear sparse arrays," *IEEE J. Ocean. Eng.*, vol. 24, no. 3, pp. 291–299, Jul. 1999, doi: 10.1109/48.775291.
- [29] M. B. Hawes and W. Liu, "Sparse array design for wideband beamforming with reduced complexity in tapped delay-lines," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 8, pp. 1236–1247, Aug. 2014, doi: 10.1109/TASLP.2014.2327298.
- [30] X. R. Wang, M. G. Amin, and X. B. Cao, "Analysis and design of optimum sparse array configurations for adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 340–351, Jan. 2018, doi: 10.1109/TSP.2017.2760279.
- [31] S. A. Hamza and M. G. Amin, "Sparse array beamforming design for wideband signal models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 2, pp. 1211–1226, Apr. 2021, doi: 10.1109/TAES.2020.3037409.
- [32] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted  $l_1$  minimization," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 877–905, Oct. 2008, doi: 10.1007/s00041-008-9045-x.
- [33] R. X. Niu, A. Vempaty, and P. K. Varshney, "Received-signal-strength-based localization in wireless sensor networks," *Proc. IEEE*, vol. 106, no. 7, pp. 1166–1182, Jul. 2018, doi: 10.1109/JPROC.2018.2828858.
- [34] I. Guvenc and C.-C. Chong, "A survey on TOA based wireless localization and NLOS mitigation techniques," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 107–124, Third Quarter 2009, doi: 10.1109/SURV.2009.090308.
- [35] Z. Wang, J.-A. Luo, and X.-P. Zhang, "A novel location-penalized maximum likelihood estimator for bearing-only target localization," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6166–6181, Dec. 2012, doi: 10.1109/TSP.2012.2218809.
- [36] Y. Wang and K. C. Ho, "An asymptotically efficient estimator in closed-form for 3-D AOA localization using a sensor network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6524–6535, Dec. 2015, doi: 10.1109/TWC.2015.2456057.
- [37] Q. Shen, W. Liu, L. Wang, and Y. Liu, "Group sparsity based localization for far-field and near-field sources based on distributed sensor array networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 6493–6508, Nov. 2020, doi: 10.1109/TSP.2020.3037841.
- [38] A. M. Haimovich, R. S. Blum, and L. J. Cimini, "MIMO radar with widely separated antennas," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 116–129, 2008, doi: 10.1109/MSP.2008.4408448.
- [39] J. Li and P. Stoica, *MIMO Radar Signal Processing*. Hoboken, NJ, USA: Wiley, 2008.
- [40] S. Fortunati, L. Sanguinetti, F. Gini, M. S. Greco, and B. Himed, "Massive MIMO radar for target detection," *IEEE Trans. Signal Process.*, vol. 68, pp. 859–871, Jan. 2020, doi: 10.1109/TSP.2020.2967181.
- [41] B. Tang and J. Tang, "Joint design of transmit waveforms and receive filters for MIMO radar space-time adaptive processing," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4707–4722, Sep. 2016, doi: 10.1109/TSP.2016.2569431.
- [42] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2018.
- [43] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
- [44] M. Shimizu, "Millimeter-wave beam multiplexing method using subarray type hybrid beamforming of interleaved configuration with inter-subarray coding," *Int. J. Wireless Inf. Netw.*, vol. 24, no. 3, pp. 217–224, Sep. 2017, doi: 10.1007/s10776-017-0368-x.
- [45] J. Zhang, W. Liu, C. Gu, S. S. Gao, and Q. Luo, "Multi-beam multiplexing design for arbitrary directions based on the interleaved subarray architecture," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11,220–11,232, Oct. 2020, doi: 10.1109/TVT.2020.3008535.
- [46] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013, doi: 10.1109/MSP.2012.2235192.
- [47] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018, doi: 10.1109/JPROC.2018.2820126.
- [48] F. Roemer and M. Haardt, "Tensor-based channel estimation and iterative refinements for two-way relaying with multiple antennas and spatial reuse," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5720–5735, Nov. 2010, doi: 10.1109/TSP.2010.2062179.
- [49] S. Miron et al., "Tensor methods for multisensor signal processing," *IET Signal Process.*, vol. 14, no. 10, pp. 693–709, Dec. 2020, doi: 10.1049/iet-spr.2020.0373.
- [50] N. Le Bihan and J. Mars, "Singular value decomposition of quaternion matrices: A new tool for vector-sensor signal processing," *Signal Process.*, vol. 84, no. 7, pp. 1177–1199, Jul. 2004, doi: 10.1016/j.sigpro.2004.04.001.
- [51] W. Liu, "Channel equalization and beamforming for quaternion-valued wireless communication systems," *J. Franklin Inst.*, vol. 354, no. 18, pp. 8721–8733, Dec. 2017, doi: 10.1016/j.franklin.2016.10.043.
- [52] O. Bar-Shalom and A. J. Weiss, "DOA estimation using one-bit quantized measurements," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 38, no. 3, pp. 868–884, Jul. 2002, doi: 10.1109/TAES.2002.1039405.
- [53] H. Kim, A. M. Haimovich, and Y. C. Eldar, "Non-coherent direction of arrival estimation from magnitude only measurements," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 925–929, Jul. 2015, doi: 10.1109/LSP.2014.2377556.
- [54] Z. Y. Wan and W. Liu, "Non-coherent DOA estimation via proximal gradient based on a dual-array structure," *IEEE Access*, vol. 9, pp. 26,792–26,801, Feb. 2021, doi: 10.1109/ACCESS.2021.3058000.
- [55] Z. Wan, W. Liu, and P. Willett, "Target localization based on distributed array networks with magnitude-only measurements," *IEEE Trans. Aerosp. Electron. Syst.*, early access, 2023, doi: 10.1109/TAES.2023.3256359.
- [56] Z. M. Liu, C. W. Zhang, and P. S. Yu, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Trans. Antennas Propag.*, vol. 66, no. 12, pp. 7315–7327, Oct. 2018, doi: 10.1109/TAP.2018.2874430.
- [57] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 14–41, May 2022, doi: 10.1109/MSP.2021.3125282.
- [58] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/Jun. 2020, doi: 10.1109/MNET.001.1900287.
- [59] E. Björnson, H. Wymeersch, B. Matthieson, P. Popovski, L. Sanguinetti, and E. de Carvalho, "Reconfigurable intelligent surfaces: A signal processing perspective with wireless applications," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 135–158, Mar. 2022, doi: 10.1109/MSP.2021.3130549.
- [60] J. A. Zhang et al., "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1295–1315, Nov. 2021, doi: 10.1109/JSTSP.2021.3131210.

# Three More Decades in Array Signal Processing Research

*An optimization and structure exploitation perspective*



©SHUTTERSTOCK.COM/TRIFF

The signal processing community is currently witnessing the emergence of sensor array processing and direction-of-arrival (DoA) estimation in various modern applications, such as automotive radar, mobile user and millimeter wave indoor localization, and drone surveillance, as well as in new paradigms, such as joint sensing and communication in future wireless systems. This trend is further enhanced by technology leaps and the availability of powerful and affordable multiantenna hardware platforms.

## Introduction

New multiantenna technology has led to the widespread use of such systems in contemporary sensing and communication systems as well as a continuous evolution toward larger multiantenna systems in various application domains, such as massive multiple, input-multiple-output (MIMO) communications systems comprising hundreds of antenna elements. The massive increase of the antenna array dimension leads to unprecedented resolution capabilities, which opens new opportunities and challenges for signal processing. For example, in large MIMO systems, modern array processing methods can be used to estimate and track the physical path parameters, such as DoA, direction of departure, time delay of arrival, and Doppler shift, of tens or hundreds of multipath components with extremely high precision [1]. This parametric approach for massive MIMO channel estimation and characterization benefits from the enhanced resolution capabilities of large array systems and efficient array processing techniques. Direction-based MIMO channel estimation, which has not been possible in small MIMO systems due to the limited number of antennas, not only significantly reduces the complexity but also improves the quality of MIMO channel prediction as the physical channel parameters generally evolve on a much smaller timescale than the MIMO channel coefficients.

The history of advances in superresolution DoA estimation techniques is long, starting from the early parametric multisource methods, such as the computationally expensive

maximum likelihood (ML) techniques, to the early subspace-based techniques, such as Pisarenko and MUSIC [2]. Inspired by the seminal review article, “Two Decades of Array Signal Processing Research: The Parametric Approach” by Krim and Viberg, published in *IEEE Signal Processing Magazine* [3], we are looking back at another three decades in array signal processing research under the classical narrow-band array processing model based on second-order statistics. We revisit major trends in the field and retell the story of array signal processing from a modern optimization and structure exploitation perspective. In our overview, through prominent examples, we illustrate how different DoA estimation methods can be cast as optimization problems with side constraints originating from prior knowledge regarding the structure of the measurement system. Due to space limitations, our review of the DoA estimation research in the past three decades is by no means complete. For didactic reasons, we mainly focus on developments in the field that easily relate to the traditional multisource estimation criteria in [3] and choose simple illustrative examples.

As many optimization problems in sensor array processing are notoriously difficult to solve exactly due to their nonlinearity and multimodality, a common approach is to apply problem relaxation and approximation techniques in the development of computationally efficient and close-to-optimal DoA estimation methods. The DoA estimation approaches developed in the last 30 years differ in the prior information and model assumptions that are maintained and relaxed during the approximation and relaxation procedure in the optimization.

Along the line of constrained optimization, problem relaxation, and approximation, recently, the partial relaxation (PR) technique has been proposed as a new optimization-based DoA estimation framework that applies modern relaxation techniques to traditional multisource estimation criteria to achieve new estimators with excellent estimation performance at affordable computational complexity. In many senses, it can be observed that the estimators designed under the PR framework admit new insights into existing methods of this well-established field of research [4].

The introduction of sparse optimization techniques for DoA estimation and source localization in the late 2000s marks another methodological leap in the field [5], [6], [7], [8], [9]. These modern optimization-based methods became extremely popular due to their advantages in practically important scenarios where classical subspace-based techniques for DoA estimation often experience a performance breakdown, e.g., in the case of correlated sources, when the number of snapshots is low, or when the model order is unknown. Sparse representation-based methods have been successfully extended to incorporate and exploit various forms of structures, e.g., application-dependent row- and

rank-sparse structures [10], [11], that induce joint sparsity to enhance estimation performance in the case of multiple snapshots. In particular array geometries, additional structures, such as Vandermonde and shift invariance, can be used to obtain efficient parameterizations of the array sensing matrix that avoid the usual requirement of sparse reconstruction methods to sample the angular field of view (FoV) on a fine DoA grid [12], [13].

Despite the success of sparsity-based methods, it is, however, often neglected that these methods also have their limitations, such as estimation biases resulting from off-grid errors and the impact of the sparse regularization, high computational complexity, and memory demands as well as sensitivity to the choice of the so-called hyperparameters. In fact, for many practical estimation scenarios, sparse optimization techniques are often outperformed by classical subspace techniques in terms of both the resolution of sources and computational complexity. From the theoretical perspective, performance guarantees of sparse methods are generally available only under the condition of the minimum angular separation between the source signals [9]. Therefore, it is important to be

aware of these limitations and to appreciate the benefits of both traditional and modern optimization-based DoA estimation methods.

The narrow-band far-field point source signals with perfectly calibrated sensor arrays and centralized processing architectures have been fundamental assumptions in the past. With the trend of wider reception bandwidth on the one hand, and larger aperture and distributed array on the other hand, the aforementioned assumptions appeared restrictive and often impractical. Distributed sensor networks have emerged as a scalable solution for source localization where sensors exchange data locally within their neighborhood and in-network processing is used for distributed source localization with low communication overhead [14]. Furthermore, DoA estimation methods for partly calibrated subarray systems have been explored [15], [16].

Model structure, e.g., in the form of a favorable spatial sampling pattern, is exploited for various purposes: either to reduce the computational complexity and to make the estimation computationally tractable or to generally improve the estimation quality. In this article, we revisit the major trends of structure exploitation in sensor array signal processing. Along this line, we consider advanced spatial sampling concepts designed in recent years, including minimum redundancy [17], augmentable [18], nested [19], and coprime arrays [20], [21]. The aforementioned spatial sampling patterns were designed to facilitate new DoA estimation methods with the capability of resolving significantly more sources than sensors in the array. This is different from conventional sampling patterns, e.g., uniform linear

**Model structure, e.g., in the form of a favorable spatial sampling pattern, is exploited for various purposes: either to reduce the computational complexity and to make the estimation computationally tractable or to generally improve the estimation quality.**

array (ULA), where the number of identifiable sources is always smaller than the number of sensors.

## Signal model

In this overview article, we consider the narrow-band point source signal model. Under this signal model, we are interested in estimating the DoAs, i.e., the parameter vector  $\boldsymbol{\theta} = [\theta_1, \dots, \theta_N]^T$ , of  $N$  far-field narrow-band sources impinging on a sensor array composed of  $M$  sensors from noisy measurements.

We assume that the DoA  $\theta_n$  lies in the FoV  $\Theta$ , i.e.,  $\theta_n \in \Theta$ . Let  $\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{n}(t)$  denote the linear array measurement model at time instant  $t$  where  $\mathbf{s}(t)$  and  $\mathbf{n}(t)$  denote the signal waveform vector and the sensor noise vector, respectively. The sensor noise  $\mathbf{n}(t)$  is commonly assumed to be a zero-mean spatially white complex circular Gaussian random process with a covariance matrix  $\nu\mathbf{I}_M$ . The steering matrix  $\mathbf{A}(\boldsymbol{\theta}) \in \mathcal{A}_N$  lives on an  $N$ -dimensional array manifold  $\mathcal{A}_N$ , which is defined as

$$\mathcal{A}_N = \{ \mathbf{A} = [\mathbf{a}(\vartheta_1), \dots, \mathbf{a}(\vartheta_N)] \mid \vartheta_1 < \dots < \vartheta_N \text{ and } \vartheta_n \in \Theta \text{ for all } n = 1, \dots, N \}. \quad (1)$$

In (1), the steering vector  $\mathbf{a}(\theta) = [e^{-j\pi d_1 \cos(\theta)}, e^{-j\pi d_2 \cos(\theta)}, \dots, e^{-j\pi d_M \cos(\theta)}]^T$  denotes, e.g., the array response of a linear array with sensor positions  $d_1, \dots, d_M$  in half wavelength for a narrow-band signal impinging from the direction  $\theta$ . The steering matrix  $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1), \dots, \mathbf{a}(\theta_N)]$  must satisfy certain regularity conditions so that the estimated DoAs can be uniquely identifiable up to a permutation from the noiseless measurement. Mathematically, the unique identifiability condition requires that if  $\mathbf{A}(\boldsymbol{\theta}^{(1)})\mathbf{s}^{(1)}(t) = \mathbf{A}(\boldsymbol{\theta}^{(2)})\mathbf{s}^{(2)}(t)$  for  $t = 1, \dots, T$ , then  $\boldsymbol{\theta}^{(1)}$  is a permutation of  $\boldsymbol{\theta}^{(2)}$ . Generally, this condition must be verified for any sensor structure and the corresponding FoV. Specifically, it can be shown that if the array manifold is free from ambiguities, i.e., if any oversampled steering matrix  $\mathbf{A}(\boldsymbol{\theta}) \in \mathcal{A}_K$  of dimension  $M \times K$  with  $K \geq M$  has a Kruskal rank  $q(\mathbf{A}(\boldsymbol{\theta})) = M$ , then  $N$  DoAs with  $N < M$  can be uniquely determined from the noiseless measurement [3]. Equivalently, any set of  $M$  column vectors  $\{\mathbf{a}(\theta_1), \dots, \mathbf{a}(\theta_M)\}$  with  $M$  distinct DoAs  $\theta_1, \dots, \theta_M \in \Theta$  is linearly independent.

In the so-called conditional signal model, the waveform vector  $\mathbf{s}(t)$  is assumed to be deterministic such that  $\mathbf{x}(t) \sim \mathcal{N}_C(\mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t), \nu\mathbf{I}_M)$ . The unknown noise variance  $\nu$  and the signal waveform  $\mathbf{S} = [s(1), \dots, s(T)]$  are generally not of interest in the context of DoA estimation, but they are necessary components of the signal model. In contrast, in the unconditional signal model, the waveform is assumed to be zero-mean complex circular Gaussian such that  $\mathbf{x}(t) \sim \mathcal{N}_C(\mathbf{0}_M, \mathbf{A}(\boldsymbol{\theta})\mathbf{P}\mathbf{A}^H(\boldsymbol{\theta}) + \nu\mathbf{I}_M)$ , where the noise variance  $\nu$  and the waveform covariance matrix  $\mathbf{P} = \mathbb{E}\{\mathbf{s}(t)\mathbf{s}^H(t)\}$  are considered as unknown parameters. We assume, if not

stated otherwise, that the signals are not fully correlated, i.e.,  $\mathbf{P}$  is nonsingular.

Note that in practical wireless communication or radar applications, the received signal may be broadband. Such scenarios require extensions of the narrow-band signal model, e.g., to subband processing or the multidimensional harmonic retrieval, which is, however, out of scope of this article.

## Cost function and concentration

Parametric methods for DoA estimation can generally be cast as optimization problems with multivariate objective functions that depend on a particular data matrix  $\mathbf{Y}$  obtained from the array measurements  $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(T)]$  through a suitable mapping, the unknown DoA parameters of interest  $\boldsymbol{\theta}$ , and the unknown nuisance parameters, which we denote by the vector  $\boldsymbol{\alpha}$ . Hence, the parameter estimates are computed as the minimizer of the corresponding optimization problem with the objective function  $f(\mathbf{Y}|\mathbf{A}(\boldsymbol{\theta}), \boldsymbol{\alpha})$  as follows:

$$\mathbf{A}(\hat{\boldsymbol{\theta}}) = \underset{\mathbf{A}(\boldsymbol{\theta}) \in \mathcal{A}_N}{\operatorname{argmin}} \min_{\boldsymbol{\alpha}} f(\mathbf{Y}|\mathbf{A}(\boldsymbol{\theta}), \boldsymbol{\alpha}). \quad (2)$$

Remark that in (2), we make no restriction on how the data matrix  $\mathbf{Y}$  is constructed from the measurement matrix  $\mathbf{X}$ . For example, in the most trivial case, the data matrix  $\mathbf{Y}$  can directly represent the array measurement matrix, i.e.,  $\mathbf{Y} = \mathbf{X}$ . However, for other optimization criteria, the data matrix  $\mathbf{Y}$  can be the sample covariance matrix, i.e.,  $\mathbf{Y} = \hat{\mathbf{R}} = (1/T)\mathbf{X}\mathbf{X}^H$  as a sufficient statistics, or even the signal eigenvectors  $\mathbf{Y} = \hat{\mathbf{U}}_s$  (or the noise eigenvectors  $\mathbf{Y} = \hat{\mathbf{U}}_n$ ) obtained from the eigendecomposition  $\hat{\mathbf{R}} = \hat{\mathbf{U}}_s \hat{\mathbf{\Lambda}}_s \hat{\mathbf{U}}_s^H + \hat{\mathbf{U}}_n \hat{\mathbf{\Lambda}}_n \hat{\mathbf{U}}_n^H$  where  $\hat{\mathbf{\Lambda}}_s = \operatorname{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_N)$  contains the  $N$ -largest eigenvalues of  $\hat{\mathbf{R}}$ . In Table 1, some prominent examples of multisource estimation methods are listed: deterministic ML (DML) [2, Sec. 8.5.2], weighted subspace fitting (WSF) [22], and covariance matching estimation techniques (COMET) [23].

As we are primarily interested in estimating the DoA parameters  $\boldsymbol{\theta}$ , a common approach is to concentrate the objective function with respect to all (or only part of) the nuisance parameters  $\boldsymbol{\alpha}$ . In the case that a closed-form minimizer of the nuisance parameters w.r.t. the remaining parameters exists, the expression of this minimizer can be inserted back to the original objective function to obtain the concentrated optimization problem. More specifically, let  $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$  denote the minimizer of the full problem for the nuisance parameter vector  $\boldsymbol{\alpha}$  as a function of  $\boldsymbol{\theta}$ , i.e.,  $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \operatorname{argmin}_{\boldsymbol{\alpha}} f(\mathbf{Y}|\mathbf{A}(\boldsymbol{\theta}), \boldsymbol{\alpha})$ . The concentrated objective function  $g(\mathbf{Y}|\mathbf{A}(\boldsymbol{\theta})) = f(\mathbf{Y}|\mathbf{A}(\boldsymbol{\theta}), \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}))$  then depends only on the DoAs  $\boldsymbol{\theta}$ . Apart from the reduction of dimensionality, the concentrated versions of multisource optimization problems often admit appealing interpretations. In Table 1, the concentrated criteria corresponding to the previously considered full-parameter multisource criteria are provided. We observe, e.g., in the case of the concentrated

**The introduction of sparse optimization techniques for DoA estimation and source localization in the late 2000s marks another methodological leap in the field.**

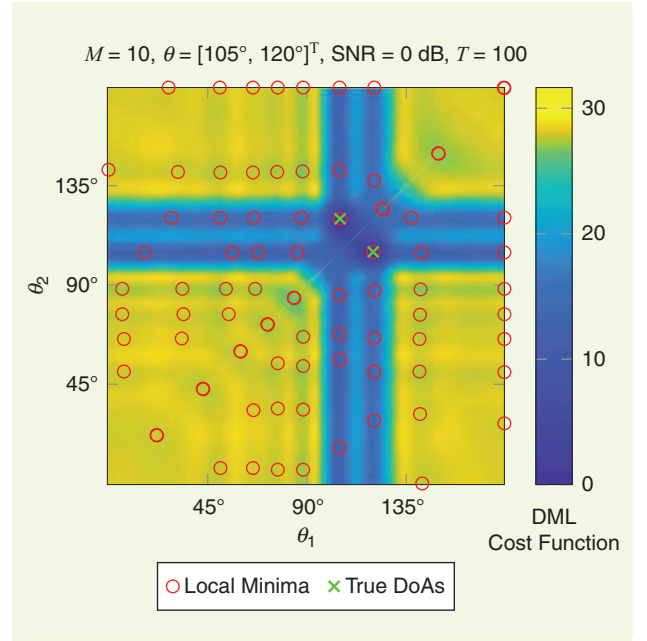


DML and the WSF criteria, that at the optimum, the residual signal energy contained in the nullspace of the steering matrix is minimized.

Due to the complicated structure of the array manifold  $\mathcal{A}_N$  in (1), the concentrated objective function  $g(Y|A(\theta))$  is, for common choices in Table 1, highly nonconvex and multimodal w.r.t. the DoA parameters  $\theta$ . Consequently, the concentrated cost function contains a large number of local minima in the vicinity of the global minimum. This can, e.g., be observed in Figure 1, where the cost function of the DML estimator is depicted. While multisource estimation criteria generally show unprecedented asymptotic as well as threshold performance for low sample size, signal-to-noise ratio (SNR), and closely spaced sources, their associated computational cost is unsuitable in many practical applications. The exact minimization generally requires an  $N$ -dimensional search over the FoV, which becomes computationally prohibitive even for low source numbers, e.g.,  $N = 3$ .

In the past three decades and beyond, significant efforts have been made to devise advanced DoA estimation algorithms that exhibit good tradeoffs between performance and complexity. While some very efficient methods have been proposed in a different context and based on pure heuristics, in this feature article, we focus on optimization-based estimators that stem, in some way or another, from multisource optimization problems for the classical array processing model (compare Table 1). Considering the array processing literature, a vast amount of estimators proposed in

the past years can be derived from multisource optimization problems. Optimization-based estimators have the advantage that they are not only well-motivated but also intuitively interpretable and flexible for generalization to more sophisticated



**FIGURE 1.** An example of the DML cost function for two sources evaluated over the FoV. Multiple local minima are observed. Consequently, local optimization search cannot guarantee to converge to the global minimum.

**Table 1. Conventional DoA estimators.**

		Full Dimension	Partial Relaxation	Single-Source Approximation
		<b>DML</b>	<b>PR-DML</b>	<b>Conventional Beamformer</b>
<b>Signal fitting</b>	Original	$\underset{\substack{A \in \mathbb{A}_N \\ S \in \mathbb{C}^{N \times T}}}{\operatorname{argmin}} \ X - AS\ _F^2$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \min_{s, B, J} \ X - as^T - BJ\ _F^2$	$\underset{a \in \mathcal{A}_1, s \in \mathbb{C}^T}{N \operatorname{argmin}} \min \ X - as^T\ _F^2$
	Concentrated	$\underset{A \in \mathcal{A}_N}{\operatorname{argmin}} \operatorname{tr}(\Pi_A^\perp X X^H)$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \sum_{k=1}^M \lambda_k(\Pi_a^\perp X X^H)$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \operatorname{tr}(\Pi_a^\perp X X^H)$
		<b>WSF</b>	<b>PR-WSF</b>	<b>Variant of Weighted MUSIC*</b>
<b>Subspace fitting</b>	Original	$\underset{\substack{A \in \mathcal{A}_N \\ V \in \mathbb{C}^{N \times N}}} {\operatorname{argmin}} \ \hat{U}_s W^{\frac{1}{2}} - AV\ _F^2$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \min_{s, B, Q} \ \hat{U}_s W^{\frac{1}{2}} - av^T - BQ\ _F^2$	$\underset{a \in \mathcal{A}_1, v \in \mathbb{C}^T}{N \operatorname{argmin}} \min \ \hat{U}_s W^{\frac{1}{2}} - av^T\ _F^2$
	Concentrated	$\underset{A \in \mathcal{A}_N}{\operatorname{argmin}} \operatorname{tr}(\Pi_A^\perp \hat{U}_s W \hat{U}_s^H)$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \sum_{k=1}^M \lambda_k(\Pi_a^\perp \hat{U}_s W \hat{U}_s^H)$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \operatorname{tr}(\Pi_a^\perp \hat{U}_s W \hat{U}_s^H)$
		<b>COMET</b>	<b>PR-CCF</b>	<b>Variant of Capon Beamformer**</b>
<b>Covariance fitting</b>	Original	$\underset{\substack{A \in \mathcal{A}_N \\ P, V}} {\operatorname{argmin}} \left\  \hat{R}^{-\frac{1}{2}} (\hat{R} - R) \hat{R}^{-\frac{1}{2}} \right\ _F^2$ subject to $R = APA^H + \nu I$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \min_{\sigma_i^2 \geq 0, G} \left\  \hat{R} - \sigma_a^2 aa^H - GG^H \right\ _F^2$ subject to $\hat{R} - \sigma_a^2 aa^H - GG^H \geq 0$ $\operatorname{rank}(G) \leq N - 1$	$\underset{a \in \mathcal{A}_1, \sigma_a^2 \geq 0}{N \operatorname{argmin}} \left\  \hat{R} - \sigma_a^2 aa^H \right\ _F^2$ subject to $\hat{R} - \sigma_a^2 aa^H \geq 0$
	Concentrated	See [23, eq. (35)]	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \sum_{k=1}^M \lambda_k \left( \hat{R} - \frac{1}{a^H \hat{R}^{-1} a} aa^H \right)$	$\underset{a \in \mathcal{A}_1}{N \operatorname{argmin}} \left\  \hat{R} - \frac{1}{a^H \hat{R}^{-1} a} aa^H \right\ _F^2$

$\Pi_A^\perp = I - A(A^H A)^{-1} A^H$  denotes the orthogonal projector onto the nullspace spanned by the columns of  $A$ . The code for the different variants of the PR methods can be downloaded at <https://github.com/PartialRelaxationMethods>.  
 \*Conventional weighted MUSIC algorithm (e.g., see [2, eq. (9.258)]) applies the weighting on the noise subspace. This variant applies the weighting on the signal subspace.  
 \*\*Note that the optimizer  $\hat{\sigma}_a^2$  is the spectrum of the Capon Beamformer. The null spectrum of this estimator contains both the spectra of the Conventional Beamformer and the Capon Beamformer.

realistic array signal models. Interestingly, some of the estimators in Table 1 were initially derived by heuristics and are reintroduced here from the perspective of multisource optimization problems.

### Modern convex optimization for DoA estimation

The progress in modern convex optimization theory and the emergence of efficient constrained optimization solvers with the turn of the millennium, such as, e.g., the SeDuMi software for solving semidefinite programs, had a significant impact on the research across disciplines in the signal processing, communication, and control communities. In fact, it comes as no surprise that the advances in sensor array signal processing of the past three decades are well aligned with this trend that facilitates advanced constrained optimization-based design approaches. Three closely related universal concepts have been intensively used in array signal processing to make optimization-based estimation procedures numerically stable and computationally feasible. These are: 1) structure exploitation, 2) approximation, and 3) relaxation.

- *Structure exploitation:* This refers to techniques that make use of particular redundancies in the measurement system to introduce convenient data reorganizations and reparameterizations. Examples are methods particularly designed for uniform, shift-invariant, and coprime array geometries.
- *Problem approximation:* These techniques provide local approximations of the multidimensional multimodal non-convex objective function with the goal to decompose a complex problem into several subproblems. Each subproblem, whose minimizer is much generally simpler to obtain than that of the original problem, is solved in parallel or sequentially, ideally in closed form. Examples are the expectation-maximization algorithm, the orthogonal matching pursuit (OMP), and the single-source approximation methods.
- *Problem relaxation:* Problem relaxation techniques in DoA estimation aim at simplifying the complicated manifold structure associated with the estimation problem. The manifold relaxation is carried out, e.g., to convexify the constraint sets in the associated optimization problems such that numerical methods can be applied.

Approximation and relaxation techniques have in common that they are used to deliberately ignore some parts of the problem structure at the expense of the optimality or performance of the solution. The objective is to simplify the problem so that efficient suboptimal solutions can be obtained that, in many cases, are close to optimal and often even admit performance guarantees. The DoA estimators reviewed in this overview article apply one or more of the aforementioned optimization concepts, as explained in more detail in the following sections.

### Single-source approximation

Spectral-based DoA estimation methods, like the popular MUSIC algorithm, belong to the class of single-source ap-

proximation methods. In contrast to the full parameter search of minimizing the multisource objective  $f(\mathbf{Y}|\mathbf{A}(\boldsymbol{\theta}), \boldsymbol{\alpha})$  over the  $N$ -source signal model with the array manifold  $\mathcal{A}_N$  and nuisance parameter vector  $\boldsymbol{\alpha}$ , the optimization problem in the single-source approximation approach is simplified, and the optimization is carried out only over a single-source model with array manifold, i.e.,  $\mathbf{A}(\boldsymbol{\theta}) \rightarrow \mathbf{a}(\theta) \in \mathcal{A}_1$  and nuisance parameters  $\boldsymbol{\alpha} \rightarrow \alpha_1$ . It is important to note that, while the number of signal components considered in the optimization is reduced in the single-source approximation approach, the data term  $\mathbf{Y}$  in the objective remains unchanged. The locations  $\mathbf{a}(\hat{\theta})$  of the  $N$ -deepest minima of the so-called null spectrum  $f(\mathbf{Y}|\mathbf{a}(\theta), \alpha_1)$  evaluated for all steering vectors  $\mathbf{a}(\theta) \in \mathcal{A}_1$  with angles in the FoV are considered as the steering vector of the estimated DoAs. By using the compact notation  ${}^N \text{argmin } g(\cdot)$  to represent the spectral search of the cost function  $g(\cdot)$  for the  $N$ -deepest local minima, the single-source approximation is formulated as follows:

$$\{\mathbf{a}(\hat{\theta})\} = {}^N \text{argmin}_{\mathbf{a}(\theta) \in \mathcal{A}_1} \min_{\alpha_1} f(\mathbf{Y}|\mathbf{a}(\theta), \alpha_1). \quad (3)$$

For clarity, the concept of the single-source approximation and the corresponding spectral search are visualized in Figure 2. As summarized in Table 1, classical spectral search methods, such as the conventional beamformer, Capon beamformer, and MUSIC, can be reformulated as single-source approximations of the corresponding multisource criteria.

### Partial Relaxation methods

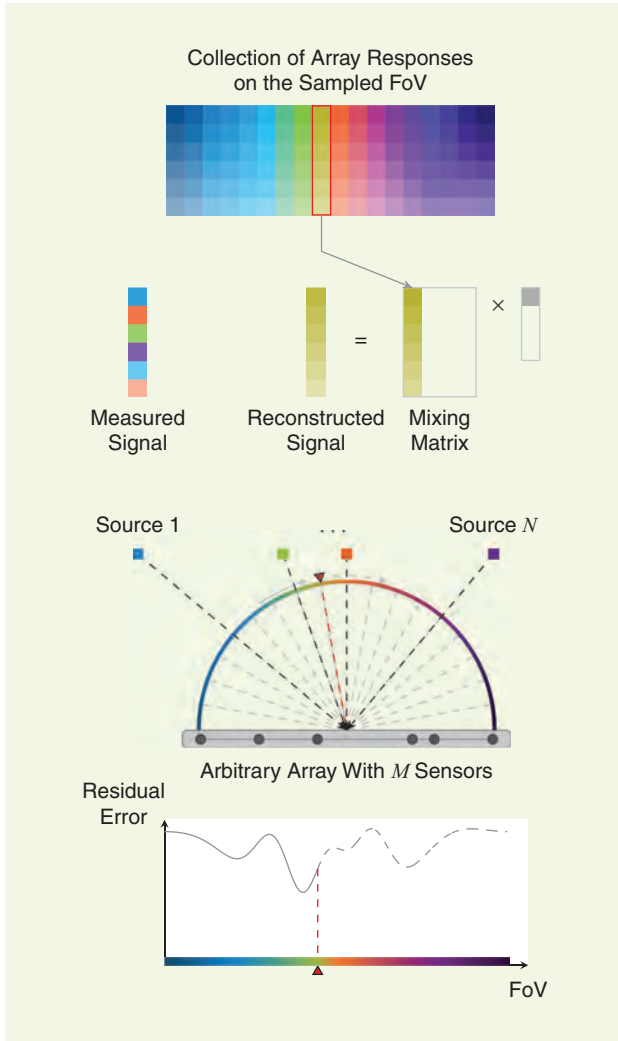
Similar to the conventional parametric methods, the PR approach considers the signals from all potential source directions in the multisource cost function. However, to make the problem tractable, the array structures of some signal components are relaxed. More precisely, instead of enforcing the steering matrix  $\mathbf{A} = [\mathbf{a}(\theta_1), \dots, \mathbf{a}(\theta_N)]$  to be an element in the highly structured array manifold  $\mathcal{A}_N$ , as in the multisource criteria in (2), without the loss of generality, we maintain the manifold structure of only the first column  $\mathbf{a}(\theta_1)$  of  $\mathbf{A}$ , which corresponds to the signal of consideration. On the other hand, the manifold structure of the remaining sources  $[\mathbf{a}(\theta_2), \dots, \mathbf{a}(\theta_N)]$ , which are considered as interfering sources, is relaxed to an arbitrary matrix  $\mathbf{B} \in \mathbb{C}^{M \times (N-1)}$  [4]. Mathematically, we assume that  $\mathbf{A} \in \tilde{\mathcal{A}}_N$  where the relaxed array manifold  $\tilde{\mathcal{A}}_N$  is parameterized as

$$\tilde{\mathcal{A}}_N = \{\mathbf{A}(\vartheta) = [\mathbf{a}(\vartheta), \mathbf{B}] | \mathbf{a}(\vartheta) \in \mathcal{A}_1, \mathbf{B} \in \mathbb{C}^{M \times (N-1)}\}. \quad (4)$$

We remark that every matrix element in the relaxed array manifold  $\tilde{\mathcal{A}}_N$  in (4) still retains the specific structure from the geometry of the sensor array in its first column, hence the name PR. However, only one DoA can be estimated from the first column of the matrix minimizer if the cost function of (2) is minimized on the relaxed array manifold  $\tilde{\mathcal{A}}_N$  of (4). Therefore, we perform the spectral search similarly to the single-source approximation in the ‘‘Single-Source Approximation’’ section as follows. First, we fix the data matrix

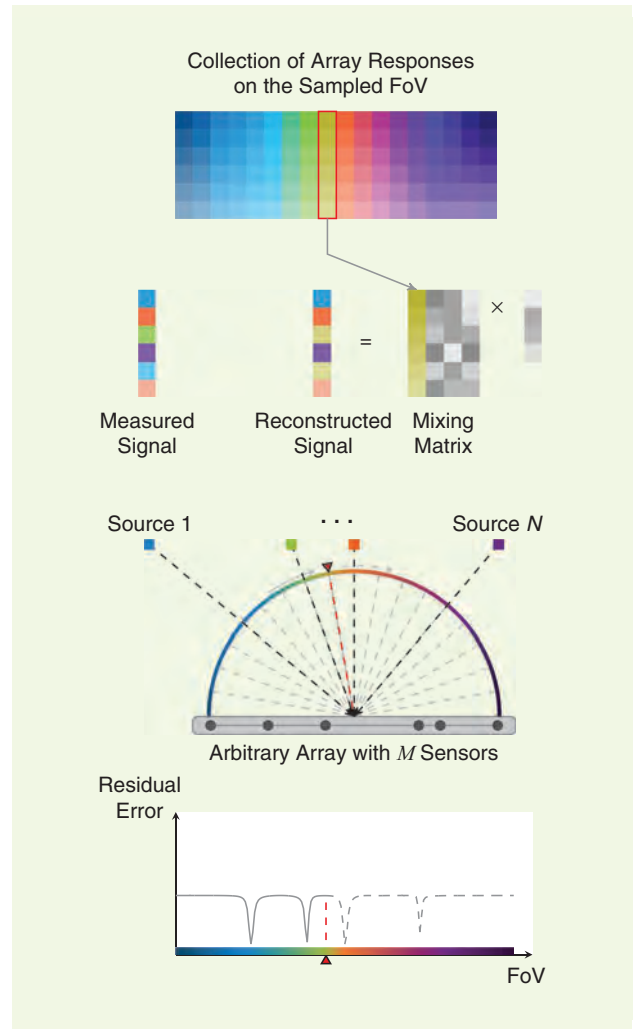
$\mathbf{Y}$  and minimize and concentrate the objective function in (2) with respect to  $\mathbf{B}$  and other nuisance parameters  $\alpha$  to obtain the concentrated cost function. Then, we evaluate the concentrated cost function for different values of  $\mathbf{a}(\theta) \in \mathcal{A}_1$  to determine the locations of the  $N$ -deepest local minima. The concept of the PR approach is illustrated in Figure 3. Using similar notation as in the single-source approximation approach, the PR approach admits the following general optimization problem:

$$\begin{aligned} \{\hat{\mathbf{a}}(\hat{\theta})\} &= \underset{\mathbf{A}(\theta) \in \tilde{\mathcal{A}}_N}{N} \operatorname{argmin} f(\mathbf{Y} | \mathbf{A}(\theta), \alpha) \\ &= \underset{\mathbf{a}(\theta) \in \mathcal{A}_1}{N} \operatorname{argmin} \min_{\mathbf{B} \in \mathbb{C}^{M \times (N-1)}} \min_{\alpha} f(\mathbf{Y} | [\mathbf{a}(\theta), \mathbf{B}], \alpha). \end{aligned} \quad (5)$$



**FIGURE 2.** An illustration of the single-source approximation concept. The optimization is carried out only over a single-source model with an array manifold, i.e.,  $\mathbf{A}(\theta) \rightarrow \mathbf{a}(\theta) \in \mathcal{A}_1$  (note that the mixing matrix has only one nonzero column corresponding to the candidate DoA  $\mathbf{a}(\theta)$ ). The influence of the remaining source signals during the spectral search is neglected, which is denoted by zero columns in the mixing matrix. The data term  $\mathbf{Y}$  in the objective function of the single-source approximation method is, however, identical to that of the corresponding multisource optimization problem.

The rationale for the PR approach lies in the fact that, when a candidate DoA  $\theta$  coincides with one of the true DoAs  $\theta_n$ , then with  $\mathbf{B}$  modeling the steering vectors of the remaining DoAs, a perfect fit to the data is attained at a high SNR or large number of snapshots  $T$ . When the candidate  $\theta$  is, however, different from all true DoAs  $\theta_n$ , the number of degrees of freedom in  $\mathbf{B}$  is not sufficiently large to represent the contribution of all  $N$ -source signals. By applying different cost functions to the general optimization problem in (5), multiple novel estimators in the PR framework are obtained in [4]. A summary of estimators under the PR framework and their relations with conventional multisource and single-source approximation-based DoA estimators are provided in Table 1.

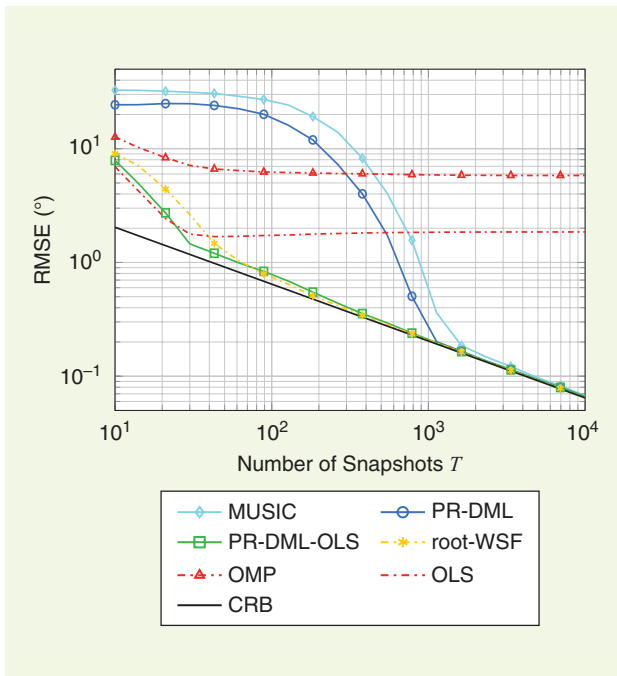


**FIGURE 3.** An illustration of the PR concept. The optimization is carried out over the relaxed array manifold  $\tilde{\mathcal{A}}_N$ , where the structure of the first column in  $\mathbf{a}(\theta_1)$  is maintained and the structure of the remaining columns is relaxed to an arbitrary complex matrix,  $[\mathbf{a}(\theta_2), \dots, \mathbf{a}(\theta_N)] \rightarrow \mathbf{B} \in \mathbb{C}^{M \times (N-1)}$ . Unlike the single-source approximation, the influence of the unstructured source signals during the spectral search is considered by the unstructured matrix  $\mathbf{B}$  (depicted by gray columns in the mixing matrix), which generally leads to an improvement of the DoA estimation when sources are closely spaced.

## Sequential techniques

While the PR methods show excellent threshold performance in scenarios with a low number of uncorrelated sources, their performance quickly degrades as the number of sources increases. This phenomenon can be explained in short as follows: the approximation error associated with the manifold relaxation of the interfering sources increases with the number of sources. The same holds true for the single-source approximation methods. As the approximation error increases, the capability of incorporating the influence of multiple structured source signals in the optimization problem decreases, and thus, a degradation in the estimation performance is observed. To overcome the degradation effect in scenarios with large source numbers, sequential estimation techniques have been proposed in which the parameters of multiple sources are estimated one after the other.

We revise three closely related and most widely known sequential estimation techniques: the MP technique, OMP, and the orthogonal least-squares (OLS) [24], [25] method. These methods have in common that the DoAs for  $N$ -sources are estimated sequentially and that the approximation is successively improved. In each iteration, the DoA of one additional source is estimated based on minimizing a function approximation of a given multisource criterion (compare Table 1), while the source DoAs estimated in the previous iterations are kept fixed at the value of their respective estimates.



**FIGURE 4.** A performance evaluation of the sequential DoA estimation techniques for four uncorrelated source signals at  $\theta = [90^\circ, 93^\circ, 135^\circ, 140^\circ]^\top$  with an array composed of  $M = 10$  sensors and SNR = 3 dB. OMP and OLS are biased as the first DoA is estimated according to the conventional beamformer, which cannot resolve two closely spaced sources at  $90^\circ$  and  $93^\circ$  regardless of the number of available snapshots  $T$ . On the other hand, PR-DML-OLS is asymptotically consistent, and its RMSE is close to the CRB.

Similar to the single-source approximation, the remaining sources whose DoA estimates have not yet been determined are ignored in the optimization.

The three methods differ, however, in the way the nuisance parameters corresponding to each source are treated in the optimization and in the corresponding parameter updating procedure. Concerning the MP algorithm, in each iteration, the nuisance parameters corresponding to the new source DoA are fixed and inserted as parameters in the objective in the following iterations. In contrast, in the OMP algorithm, the nuisance parameters of all estimated sources are updated in a refinement step after the DoA parameter of the current iteration is determined. The nuisance parameters are then inserted as parameters in the objective for the following estimation of the source DoA in the next iteration. The additional update step is generally associated with only a slight increase of the computational complexity. Nevertheless, this strategy effectively reduces error propagation effects.

OLS yields further performance improvements at the cost of more sophisticated estimate and update expressions. More precisely, in the OLS algorithm, the nuisance parameters corresponding to the sources of the previous and current iterations are treated as variables and optimized along with the DoA parameter of the new source in the current iteration. In Table 2, we provide the sequential estimation and update procedures of the MP, OMP, and OLS for the DML criterion (compare Table 1). At this point, we remark that the sequential estimation approach is general and can also be applied to other multisource criteria in Table 1, hence the WSF and the COMET criteria. Furthermore, sequential estimation can also be combined with the concept of PR that we introduced in the “PR Methods” section to further enhance the threshold performance and reduce the error propagation effects. As an example, we also provide the PR-DML-OLS method in Table 2. A numerical performance comparison of the sequential estimators is provided in Figure 4, where it can be observed that the OLS method shows improved performance as compared to OMP; however, both methods suffer from a bias. The PR-DML-OLS method is, in contrast, asymptotically consistent, and its root-mean-square error (RMSE) is close to the Cramér-Rao bound (CRB).

## Sparse reconstruction methods

The nonlinear LS DML problem in Table 1 generally requires a multidimensional grid search over the parameter space to obtain the global minimum. More precisely, the objective function is evaluated at all possible combinations of  $N$  DoAs on a particular discretized FoV. Clearly, the complexity of this brute-force multidimensional search strategy grows exponentially with the number of sources. To reduce the computational cost associated with the nonlinear LS optimization, convex approximation methods based on sparse regularization have been proposed.

We assume that for a particular FoV discretization  $\tilde{\theta} \in \mathbb{R}^K$  containing  $K \gg M$  angles, a so-called oversampled

dictionary matrix  $\tilde{\mathbf{A}} = \mathbf{A}(\tilde{\boldsymbol{\theta}})$  of dimensions  $M \times K$  is constructed and the DML problem is equivalently formulated as the linear LS minimization problem with a cardinality constraint, i.e.,

$$\min_{\tilde{\mathbf{S}} \in \mathbb{C}^{K \times T}} \|\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}\|_{\text{F}}^2 \text{ subject to } \|\tilde{\mathbf{S}}\|_{2,0} \leq N \quad (6)$$

where  $\|\mathbf{M}\|_{2,0} = \#\{k \mid \|\mathbf{m}_k\|_2 \neq 0\}$  denotes the  $\ell_{2,0}$  mixed pseudonorm of a matrix  $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_K]^T$ , i.e., the number of rows  $\mathbf{m}_k$  with nonzero Euclidean norm  $\|\mathbf{m}_k\|_2$  for  $k = 1, \dots, K$ . This is illustrated in Figure 5. Given a solution  $\tilde{\mathbf{S}}^*$  of the optimization problem in (6), the DoAs estimates  $\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}})$  are determined from the support of  $\tilde{\mathbf{S}}^*$ , i.e., the locations of the nonzero rows. Several approximation methods have been proposed to simplify the problem in (6) using sparse regularization. Sparse regularization approaches are directly devised from the Lagrangian function of the optimization problem in (6), i.e.,

$$\min_{\tilde{\mathbf{S}}} \|\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}\|_{\text{F}}^2 + \mu \|\tilde{\mathbf{S}}\|_{2,0} \quad (7)$$

where the hyperparameter  $\mu$  (also called the *regularization parameter*) balances the tradeoff between data matching and sparsity. For small values of  $\mu$ , the mismatch between the model and the measurements is emphasized in the minimization, whereas for larger values of  $\mu$ , the row sparsity of the solution is enhanced. Since the discretized FoV is given and, thus, the oversampled dictionary matrix  $\tilde{\mathbf{A}}$  is constant, the data-matching term in the objective function of (7) is a simple linear LS function. Nevertheless, the sparse regularization term is both nonsmooth and nonconvex w.r.t.  $\tilde{\mathbf{S}}$ , and thus, the problem in (7) is difficult to solve directly. In [26], a convergent iterative fixed-point algorithm is proposed that solves a sequence of the smooth approximation problems of (7).

To make the optimization in (7) more tractable, a common approach is to convexify the regularizer in (7) by approximating the  $\ell_{2,0}$  norm by the closest convex mixed-norm  $\ell_{2,1}$ , which is defined as  $\|\mathbf{M}\|_{2,1} = \sum_{k=1}^K \|\mathbf{m}_k\|_2$  for  $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_K]^T$ . The resulting multiple measurement problem (MMP)

$$\min_{\tilde{\mathbf{S}}} \|\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}\|_{\text{F}}^2 + \mu \|\tilde{\mathbf{S}}\|_{2,1} \quad (8)$$

is convex and thus can be solved efficiently [10]. One important drawback of the formulation in (8) is that the number

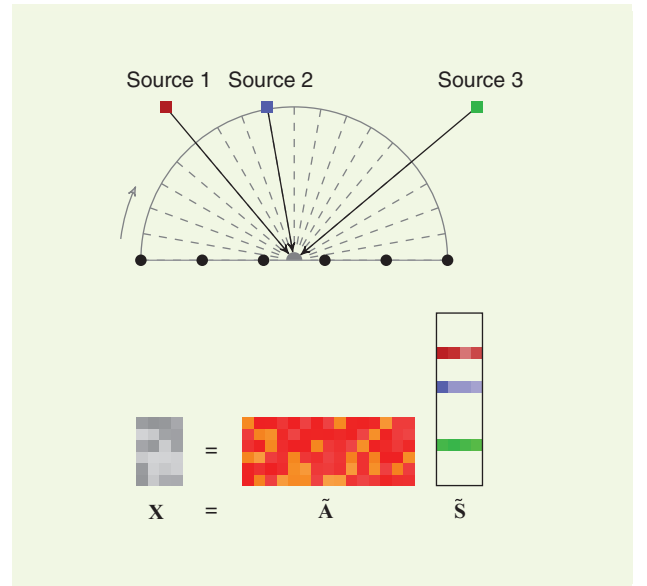
of optimization variables grows linearly with the number of snapshots  $T$  and, therefore, also the associated computational complexity. Interestingly, the MMP can be equivalently expressed as the Sparse Row-Norm Reconstruction (SPARROW) problem as follows [11]:

$$\min_{\tilde{\mathbf{D}} \in \mathcal{D}_+} \text{tr} \left( \left( \tilde{\mathbf{A}}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^H + \frac{\mu}{2\sqrt{T}} \mathbf{I}_M \right)^{-1} \hat{\mathbf{R}} \right) + \text{tr}(\tilde{\mathbf{D}}). \quad (9)$$

We remark that in (9), the optimizing variable  $\tilde{\mathbf{D}} = \text{diag}(\tilde{d}_1, \dots, \tilde{d}_K)$  is a nonnegative diagonal matrix whose

**Table 2. Sequential DoA estimators.**

	Iterative DoA Estimation Step	Nuisance Parameters Update
<b>MP</b>	$\hat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}}{\text{argmin}} \min_{s \in \mathbb{C}^T} \left\  \mathbf{X} - [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\boldsymbol{\theta})] \begin{bmatrix} \hat{\mathbf{S}}^{(k-1)} \\ s^T \end{bmatrix} \right\ _{\text{F}}^2$ $\hat{\mathbf{A}}^{(k)} = [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\hat{\boldsymbol{\theta}}^{(k)})]$	$\hat{\mathbf{s}}^{(k)} = \underset{s \in \mathbb{C}^T}{\text{argmin}} \left\  \mathbf{X} - [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\hat{\boldsymbol{\theta}}^{(k)})] \begin{bmatrix} \hat{\mathbf{S}}^{(k-1)} \\ s^T \end{bmatrix} \right\ _{\text{F}}^2$ $\hat{\mathbf{S}}^{(k)} = \begin{bmatrix} \hat{\mathbf{S}}^{(k-1)} \\ \hat{\mathbf{s}}^{(k)T} \end{bmatrix}$
<b>OMP</b>	$\hat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}}{\text{argmin}} \min_{s \in \mathbb{C}^T} \left\  \mathbf{X} - [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\boldsymbol{\theta})] \begin{bmatrix} \hat{\mathbf{S}}^{(k-1)} \\ s^T \end{bmatrix} \right\ _{\text{F}}^2$ $\hat{\mathbf{A}}^{(k)} = [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\hat{\boldsymbol{\theta}}^{(k)})]$	$\hat{\mathbf{S}}^{(k)} = \underset{S \in \mathbb{C}^{K \times T}}{\text{argmin}} \left\  \mathbf{X} - [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\hat{\boldsymbol{\theta}}^{(k)})] S \right\ _{\text{F}}^2$
<b>OLS</b>		$\hat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}}{\text{argmin}} \min_{S \in \mathbb{C}^{K \times T}} \left\  \mathbf{X} - [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\boldsymbol{\theta})] S \right\ _{\text{F}}^2$ $\hat{\mathbf{A}}^{(k)} = [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\hat{\boldsymbol{\theta}}^{(k)})]$
<b>PR-DML-OLS</b>		$\hat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}}{\text{argmin}} \min_{\substack{B \in \mathbb{C}^{M \times (K-1)} \\ S \in \mathbb{C}^{K \times T}}} \left\  \mathbf{X} - [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\boldsymbol{\theta}), \mathbf{B}] S \right\ _{\text{F}}^2$ $\hat{\mathbf{A}}^{(k)} = [\hat{\mathbf{A}}^{(k-1)}, \mathbf{a}(\hat{\boldsymbol{\theta}}^{(k)})]$



**FIGURE 5.** The concept of sparse reconstruction methods. In the noiseless case, the received signal  $\mathbf{X}$  is decomposable into a product of a fixed oversampled steering matrix  $\tilde{\mathbf{A}}$  and a row-sparse source signal matrix  $\tilde{\mathbf{S}}$ . The locations of the nonzero rows in  $\tilde{\mathbf{S}}$  correspond to the DoAs.

dimension does not depend on the number of snapshots  $T$ . The first term in (9) can be interpreted as a data-matching term, while the second term induces sparsity. The optima  $\tilde{\mathbf{S}}^*$  and  $\tilde{\mathbf{D}}^*$  of the respective problems share the same support, and the diagonal elements of  $\tilde{\mathbf{D}}^*$  represent the scaled  $\ell_2$  row norms of  $\tilde{\mathbf{S}}^*$ . The SPARROW problem is convex and can be efficiently solved using, e.g., a block-coordinate descent (BCD) algorithm. Remarkably, the support of the solution of (9) and therefore also the solution of the MMP (8) are fully encoded in the sample covariance matrix  $\hat{\mathbf{R}}$ , while the measurement matrix  $\mathbf{X}$  is not explicitly required for the estimation of the DoAs. Making use of the Schur complement, the optimization problem in (9) can further be reformulated as the semidefinite problem (SDP).

$$\begin{aligned} & \min_{\tilde{\mathbf{D}}, \mathbf{U}} \text{Tr}(\mathbf{U}\hat{\mathbf{R}}) + \frac{1}{M} \text{Tr}(\tilde{\mathbf{D}}) \\ & \text{subject to } \begin{bmatrix} \mathbf{U} & \mathbf{I}_M^H \\ \mathbf{I}_M & \tilde{\mathbf{A}}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^H + \frac{\mu}{2\sqrt{T}}\mathbf{I}_M \end{bmatrix} \succeq 0. \end{aligned} \quad (10)$$

An alternative SDP formulation also exists for the under-sampled case, when the number of snapshots is smaller than the number of sensors, i.e.,  $M \geq T$ . While from a computational point of view, the SDP formulations quickly become intractable when the dictionary  $\tilde{\mathbf{A}}$  becomes large and the BCD solution is preferable for large  $K$ , the SDP formulations admit interesting extensions for ULAs and other structured arrays that do not require the use of a sampling grid and the explicit formation of the dictionary  $\tilde{\mathbf{A}}$ . The so-called gridless sparse reconstruction methods are motivated by the following observation: in the ULA case, the dictionary  $\tilde{\mathbf{A}}$  is a Vandermonde matrix so that the matrix product  $\tilde{\mathbf{A}}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^H$  with any diagonal matrix  $\tilde{\mathbf{D}}$  can be substituted by a Toeplitz matrix  $\text{Toep}(\mathbf{u})$  with  $\mathbf{u}$  denoting its first column. Inserting the compact Toeplitz reparameterization in the SDP problem (10) and making use of the property  $\text{Tr}(\tilde{\mathbf{D}}) = 1/M \text{Tr}(\tilde{\mathbf{A}}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^H) = 1/M \text{Tr}(\text{Toep}(\mathbf{u}))$  the SDP reformulation becomes independent of a particular choice of the dictionary  $\tilde{\mathbf{A}}$ .

Consequently, the off-grid errors are avoided. An important question at this point is under which conditions the decomposition  $\text{Toep}(\mathbf{u}) = \tilde{\mathbf{A}}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^H$  holds and whether the solution is unique. If such a decomposition exists with a unique solution, the gridless reformulation of the SDP is equivalent to the original grid-based formulation. The answer to this question is provided by the well-known Carathéodory's theorem, which states that the Vandermonde decomposition of any positive semidefinite low-rank Toeplitz matrix is always unique. Hence, provided that the solution  $\text{Toep}(\mathbf{u}^*)$  is positive semidefinite and rank deficient, it can be uniquely factorized as  $\text{Toep}(\mathbf{u}^*) = \mathbf{A}^* \mathbf{D}^* (\mathbf{A}^*)^H$  [27]. Given the generator vector  $\mathbf{u}^*$  retrieved from a low-rank Toeplitz matrix, the DoA estimates can be uniquely

recovered, e.g., by solving the corresponding system of linear equations.

We remark that the gridless approach for sparse recovery in the MMP has first been introduced in the context of the atomic norm denoising problem [12], which can be considered as the continuous angle equivalent of the  $\ell_{2,1}$  norm regularized LS matching problem (8). The associated SDP formulation in the ULA case with Toeplitz parameterization can be shown to be equivalent to the gridless version of (10).

We further remark that gridless sparse reconstruction methods are not limited to contiguous ULA structures. Also, other redundant array geometries can be exploited, such as shift-invariant arrays or thinned ULAs, i.e., incomplete ULAs with missing sensors ("holes"). In thinned ULAs, ambiguities may arise in the array manifold, and the model parameters may no longer

be uniquely identifiable from the measurements. These ambiguities have, e.g., been characterized in [28], [29], and these references can provide guidelines for the choice of favorable thinned ULA geometries. Following a similar procedure as in the Toeplitz case, a substitution of the type  $\mathbf{Q} = \tilde{\mathbf{A}}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^H$  can be introduced where  $\mathbf{Q}$  is no longer perfectly Toeplitz but contains other structured redundancies that can be expressed in the form of linear equality constraints in the problem (10). In these cases, estimation of signal parameters via rotational invariance techniques (ESPRIT) or root-MUSIC can be employed to estimate the DoAs from the minimizer  $\mathbf{Q}^*$ . Even though unique factorization guarantees for  $\mathbf{Q}^*$  similar to Carathéodory's theorem do not exist, the generalized gridless recovery approach performs well in practice as long as the number of redundant entries in  $\mathbf{Q}$  is sufficiently large.

While we focused in our overview on sparse regularization methods that are based on the DML cost function in Table 1 as the data-matching term, there exist numerous alternative approaches that use other matching terms. See [9] for a comprehensive overview of sparse DoA estimation techniques. A particularly interesting sparse DoA estimation method is the Sparse Iterative Covariance-Based Estimation Approach (SPICE) [8], which, as the name suggests, stems from a weighted version of the covariance matching criterion in Table 1. Remarkably, the SPICE formulation does not contain any hyperparameters to trade off between the data-matching quality versus the sparsity of the solution, which makes SPICE an attractive candidate among sparse reconstruction methods.

As mentioned previously, the traditional superresolution methods, such as the multisource estimation methods of Table 1 as well as the PR methods and MUSIC for uncorrelated sources, are capable of resolving arbitrary closely spaced source even with a finite number of sensors as long as the number of snapshots or the SNR is sufficiently large. It is important to note that such guarantees generally do not exist in convex sparse optimization methods [9], [12], [27]. Furthermore, sparse regularization-based DoA estimation

**Considering the array processing literature, a vast amount of estimators proposed in the past years can be derived from multisource optimization problems.**

methods are known to suffer from bias, which marks one of their most important drawbacks. First, the bias can be originated from the grid mismatch of the source DoAs in the formation of the dictionary. Second, the sparse regularization term generally introduces a bias to the solution. While the former source of bias can be entirely avoided in the gridless sparse reconstruction formulations, the latter remains and can be reduced only by decreasing the regularization parameter  $\mu$ , e.g., in (7). This in turn leads to an enlarged support set whose sizes are much larger than the true number of sources  $N$ .

However, in the context of sparse regularization-based DoA estimation, if the model order  $N$  is known, it is often preferable to use comparably small values of the regularizers  $\mu$  and to perform a local search for the  $N$ -largest maxima of the recovered row-norm vector  $\vec{d}^* = [\vec{d}_1^*, \dots, \vec{d}_k^*]^T$  in (9) to determine the DoA estimates. More specifically, the DoA estimates are the  $N$  entries in the sample DoA vector  $\hat{\theta}$  that are indexed by  $\{i\} = \text{argmax}_k \vec{d}_k^*$ .

In conclusion, sparsity-based methods have their merit in difficult scenarios with low sample size or highly correlated and even coherent source signals where the sample covariance matrix does not exhibit the full signal rank  $N$ . In these scenarios, conventional subspace-based DoA estimation methods usually fail to resolve multiple sources. This is confirmed in the simulation example of Figure 6. Furthermore, sparsity-based methods can resolve multiple sources even in the single-snapshot case provided that the scene is sparse in the sense that the number of sources is small compared to the number of sensors in the array. A simple but interesting theorem for robust sparse estimation with noisy measurements regardless of the chosen sparse estimation approach is given by [30, Theorem 5].

Another benefit of sparse regularization methods is that, unlike parametric methods in DoA estimation, the knowledge of the number of sources  $N$  is not required for the estimation of the DoAs. In turn, the number of sources is implicitly determined from the sparsity of the solution. However, sparse reconstruction methods, with the exception of the hyperparameter-free SPICE method [8], are usually sensitive to a proper choice of the sparse regularization parameter  $\mu$ . Furthermore, the associated computational complexity of sparse regularization methods, in particular for the SDP formulations in the grid-based and gridless cases, is higher than that of conventional subspace-based methods.

### Exploitation of incomplete structural information

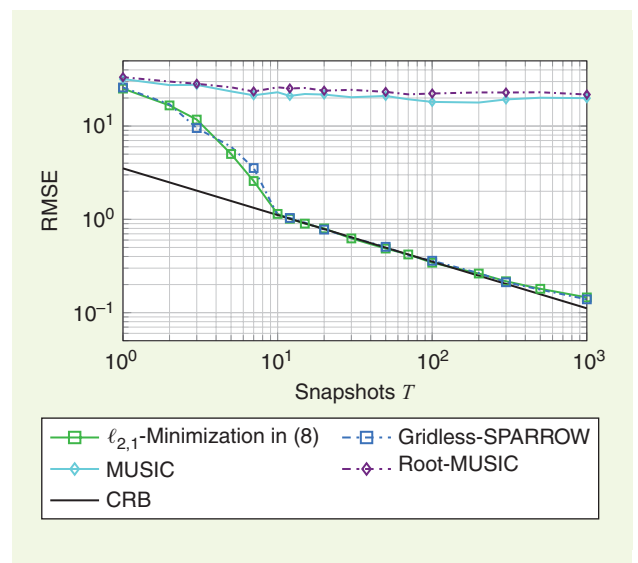
In many modern applications, such as the networks of aerial base stations, DoA estimation is carried out in a distributed fashion from signals measured at multiple subarrays, where the exact locations of subarrays are often unknown. Even in conventional centralized large sensor arrays, it is a challenging task to have a central synchronized clock among all sensors of the device and to maintain precise phase synchronization

due to large distances in the array. Therefore, in practice, large array systems are partitioned into local subarrays, where due to the proximity, each subarray can be considered as perfectly calibrated, whereas the relative phase differences between subarrays are considered as unknown. In this setup, it is commonly assumed that the narrow-band assumption remains valid; hence, the waveforms do not essentially decorrelate during the travel time over the array.

DoA estimation in partly calibrated sensor arrays has first been considered in shift-invariant sensor array systems, which are composed of two identically oriented identical subarrays separated by a known displacement  $\delta$ . For this configuration, the popular ESPRIT algorithm has been proposed [31]. In shift-invariant arrays, the overall array steering matrix  $A(\theta)$  can be partitioned into two potentially overlapping blocks,  $\underline{A}(\theta) \in \mathbb{C}^{M_1 \times N}$  and  $\overline{A}(\theta) \in \mathbb{C}^{M_2 \times N}$ , respectively, representing the array response of the reference subarray and the shifted subarray. For notational simplicity, we assume that  $M = 2M_1 = 2M_2$ . Due to the shifting structure, the two subarray steering matrices are related through right multiplication with a diagonal phase shift matrix  $D(z^\delta) = \text{diag}(z_1^\delta, \dots, z_N^\delta)$  with unit-magnitude generators  $z_n = e^{-j\pi \cos(\theta_n)}$  that account for the known displacement  $\delta$  measured in half wavelength, hence  $\overline{A}(\theta) = \underline{A}(\theta)D(z)$ .

Interestingly, ESPRIT as well as the enhanced Total-LS-ESPRIT (TLS-ESPRIT) method [31] can be reformulated as subspace-fitting techniques according to Table 1. Similar to the PR approach, a particular form of manifold relaxation is applied that maintains some part of the array structure and deliberately ignores other parts of the structure to admit a simple solution. The ESPRIT and TLS-ESPRIT estimators

**Similar to the conventional parametric methods, the Partial Relaxation approach considers the signals from all potential source directions in the multisource cost function.**



**FIGURE 6.** A performance evaluation of the sparse reconstruction-based DoA estimation techniques for two coherent sources at  $\theta = [90^\circ, 120^\circ]^T$  with an array composed of  $M = 6$  sensors and SNR = 10 dB.

are obtained from minimizing the subspace fitting functions  $\|\hat{U}_s V^{-1} - A(\boldsymbol{\theta})\|_F^2$  [22] and  $\|\hat{U}_s - A(\boldsymbol{\theta})V\|_F^2$ , respectively, for nonsingular  $V$ , where the array manifold  $\mathcal{A}_N$  of the fully calibrated subarray defined in (1) is relaxed to the ESPRIT manifold  $\mathcal{A}_N^{\text{ESPRIT}} = \{A \in \mathbb{C}^{M \times N} \mid A(\boldsymbol{\theta}) = [A \underline{A}^T, D(\mathbf{c})A \underline{A}^T]^T, \underline{A} \in \mathbb{C}^{M_1 \times N}, \mathbf{c} \in \mathbb{C}^N, \boldsymbol{\theta} = \cos^{-1}(-(\arg(\mathbf{c})/\pi\delta))\}$ .

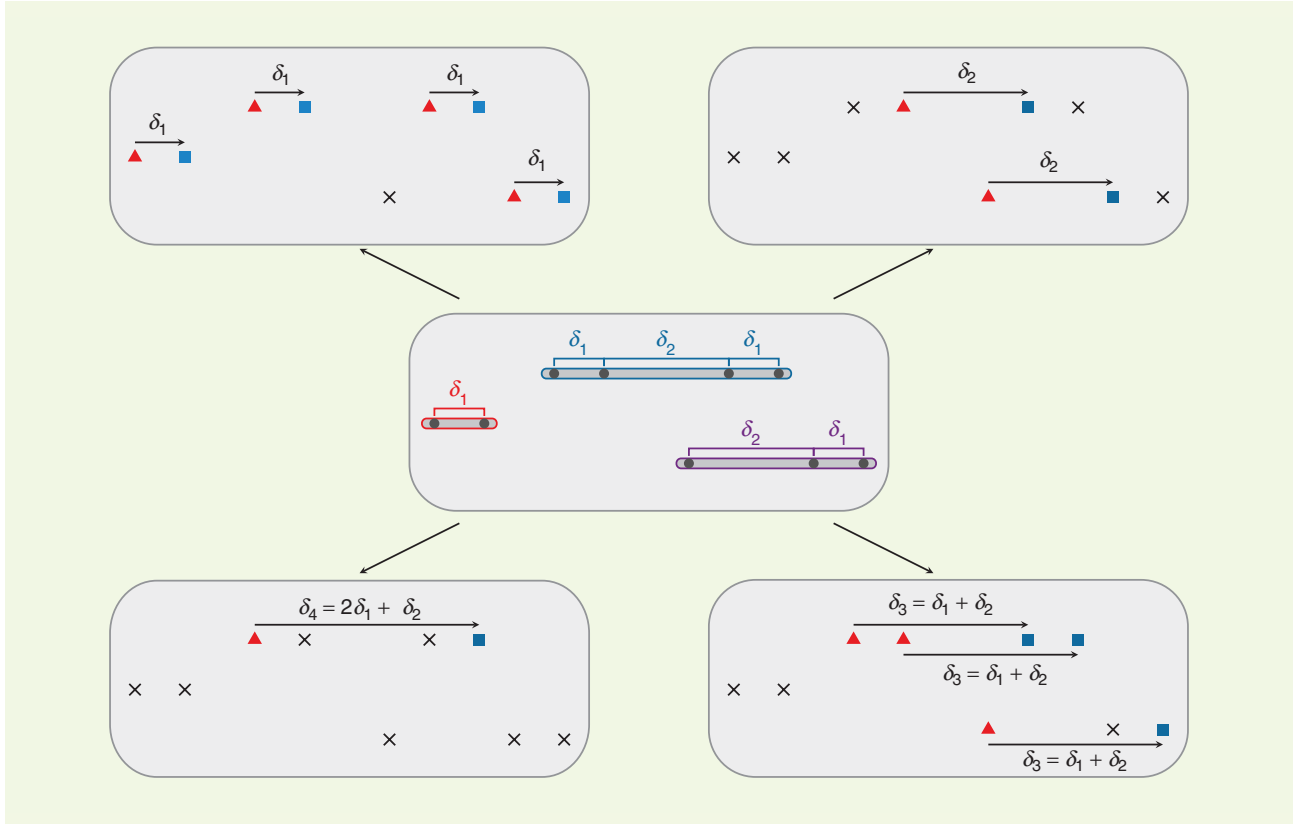
We remark that in the ESPRIT manifold  $\mathcal{A}_N^{\text{ESPRIT}}$ , only part of the shift-invariance structure of  $\mathcal{A}_N$  is maintained, and the particular subarray steering matrix structure in  $A(\boldsymbol{\theta}) = [a(\vartheta_1), \dots, a(\vartheta_N)]$  is relaxed to an arbitrary complex matrix  $\underline{A} \in \mathbb{C}^{M_1 \times N}$ . In addition, the magnitude-one structure of the diagonal shift matrix  $D(\mathbf{z}^\delta)$  is relaxed to an arbitrary diagonal matrix  $D(\mathbf{c})$ . This implies that in the ESPRIT and TLS-ESPRIT algorithms, neither the subarray geometry nor potential directional gain factors between the two subarrays need to be known as long as the subarrays are identical and subarray displacements are known. Due to this particular manifold relaxation, the subspace fitting problems admit efficient closed-form solutions.

The concept of DoA estimation in subarray structures has been generalized in [32] to cover the case of multiple shift invariance arrays. In more general partly calibrated array scenarios, we assume that the sensor positions are generally unknown. Nevertheless, only several displacements between selected pairs of sensors in the array, the so-called lags of the

array, are known. Let  $\delta_1, \dots, \delta_K$  denote known displacements in half wavelength in the array that are all pointing into the same direction. This is illustrated in Figure 7. The subarrays are flexibly defined by pairs of sensors that share a common lag  $\delta_k$  (or their summations). Depending on the number of known lags among the sensor arrays, one particular sensor can belong to one or more subarrays.

For all known lags, we consider again the subspace fitting approach and apply the ESPRIT manifold relaxation technique. Hence, defining  $T = V^{-1}$  and relaxing the structure of the subarrays, the objective becomes  $\sum_{k=1}^K \left\| [\hat{U}_{s,k} T, \hat{U}_{s,k} T] - \underline{A}_k [I, D(\mathbf{z}^{\delta_k})] \right\|_F^2$ , where  $\underline{A}_k$  is an arbitrary complex-valued matrix of known dimension that models the unknown subarray structure corresponding to the  $k$ th displacement  $\delta_k$  and  $\hat{U}_{s,k}$ . The matrix  $\hat{U}_{s,k}$  contains the corresponding rows of the signal eigenvectors in  $\hat{U}_{s,k}$ . Inserting the LS minimizers  $\hat{A}_{\text{LS},k} = (1/2)(\hat{U}_{s,k} T + \hat{U}_{s,k} T D^*(\mathbf{z}^{\delta_k}))$  back into the objective function, the concentrated objective function of the relaxed multiple shift-invariant ESPRIT is given by

$$\sum_{k=1}^K \text{Tr} \left( -T^H \hat{U}_{s,k}^H \hat{U}_{s,k} T D(\mathbf{z}^{-\delta_k}) + T^H (\hat{U}_{s,k}^H \hat{U}_{s,k} + \hat{U}_{s,k}^H \hat{U}_{s,k}) T - T^H \hat{U}_{s,k}^H \hat{U}_{s,k} T D(\mathbf{z}^{\delta_k}) \right). \quad (11)$$



**FIGURE 7.** The equivalence between the partly calibrated array setup and multiple shift-invariant setup. As depicted in the center, an exemplary partly calibrated array setup comprises three linear subarrays with unknown intersubarray displacements. The displacements between sensors in one subarray are, however, a priori known. From the known intrasubarray displacement, multiple shift-invariant structures between sensor pairs are exploited while formulating the optimization problem. Such exploitation allows reinterpretation of the RARE algorithm [15], [16] as a generalized multiple shift ESPRIT [31].



Due to the diagonal structure of  $\mathbf{D}(\mathbf{z}^{\delta_k})$ , the objective function in (11) is separable into  $N$  identical terms, one for each source. Hence the subspace fitting problem reduces to finding the  $N$  distinct minima of the rank reduction estimator (RARE) [15], [16] function  $f_{\text{RARE}}(\boldsymbol{\theta}) = \min_{\|\mathbf{t}\|=1} \mathbf{t}^H \mathbf{M}(\boldsymbol{\theta}) \mathbf{t}$  with respect to the DoAs  $\boldsymbol{\theta} \in \Theta$  where

$$\mathbf{M}(\boldsymbol{\theta}) = \sum_{k=1}^K \left( -\hat{\underline{\mathbf{U}}}_{s,k}^H \hat{\underline{\mathbf{U}}}_{s,k} e^{j\pi\delta_k \cos(\theta_n)} + \left( \hat{\underline{\mathbf{U}}}_{s,k}^H \hat{\underline{\mathbf{U}}}_{s,k} + \hat{\underline{\mathbf{U}}}_{s,k}^H \hat{\underline{\mathbf{U}}}_{s,k} \right) - \hat{\underline{\mathbf{U}}}_{s,k}^H \hat{\underline{\mathbf{U}}}_{s,k} e^{-j\pi\delta_k \cos(\theta_n)} \right) \quad (12)$$

and  $\mathbf{t}$  represents a particular column of  $\mathbf{T}$ . The unit-norm constraint in the RARE problem is introduced to ensure that the zero solution  $\mathbf{t} = \mathbf{0}$  is excluded since the zero solution  $\mathbf{t} = \mathbf{0}$  violates the constraint that  $\mathbf{T}$  and  $\mathbf{V}$  are nonsingular. The minimization of the RARE cost function w.r.t. to the vector  $\mathbf{t}$  admits the minor eigenvector of  $\mathbf{M}(\boldsymbol{\theta})$  as a minimizer. As a result, the concentrated RARE cost function is given by  $f_{\text{RARE}}(\boldsymbol{\theta}) = \lambda_{\min}(\mathbf{M}(\boldsymbol{\theta}))$ . Thus, similar to the single-source approximation approach, the DoAs are determined from the  $N$ -deepest minima of the RARE function. This ensures that the corresponding transformation matrices  $\mathbf{T}$  and  $\mathbf{V}$  are nonsingular. We remark that the RARE estimator has originally been derived from a relaxation of the MUSIC function, and the minimum eigenvalue function can equivalently be replaced by the determinant of the matrix  $\mathbf{M}(\boldsymbol{\theta})$ . The latter is, e.g., useful for developing a search-free variant of the spectral RARE algorithm based on matrix polynomial rooting in the case that the shifts  $\delta_1, \dots, \delta_K$  are integer multiples of a common baseline.

### Exploitation of array configuration

As mentioned in the ‘‘Signal Model’’ section, the number of signals  $N$  that can be uniquely recovered from DoA estimation methods with second-order statistics is strictly upper bounded by the Kruskal rank of the oversampled ( $K \geq M$ ) steering matrix  $\tilde{\mathbf{A}} \in \mathcal{A}_K$ , which, e.g., for ULA geometries is equal to the number of sensors. A direct consequence is that, using the conventional signal model in the ‘‘Signal Model’’ section, the number of uniquely identifiable sources  $N$  must be less than the number of sensors  $M$ . If further information on the source signals is available, e.g., that the source signals are uncorrelated, then the number of uniquely identifiable source signals can be improved.

This claim can be explained by comparing the number of equations and the number of unknowns, which are implied from the covariance model  $\mathbf{R} = \mathbf{A}(\boldsymbol{\theta})\mathbf{D}(\mathbf{p})\mathbf{A}^H(\boldsymbol{\theta}) + \nu\mathbf{I}_M$ , with  $\mathbf{D}(\mathbf{p}) = \text{diag}(p_1, \dots, p_N)$ . We assume that the number of snapshots  $T$  is sufficiently high such that the covariance matrix  $\mathbf{R}$  can be estimated with high accuracy. In addition, we remark that the structure of the covariance matrix depends on the geometry of the sensor array. For example, for a ULA, the covariance matrix is a Hermitian Toeplitz matrix. Conversely, if we assume that the sensor array does not exhibit any particular geometry, e.g., no ULA structure, then there is generally no relation between the elements in the covariance matrix. Conse-

quently, the covariance matrix  $\mathbf{R}$  is parameterized by maximally  $M^2 - M + 1$  independent real-valued variables (note that the diagonal entries are all identical). Thus, the number of independent equations from the covariance model is also  $M^2 - M + 1$ .

On the other hand, in the uncorrelated source case, the model on the right-hand side  $\mathbf{A}(\boldsymbol{\theta})\mathbf{D}(\mathbf{p})\mathbf{A}^H(\boldsymbol{\theta}) + \nu\mathbf{I}_M$  contains only  $2N + 1$  unknowns ( $N$  DoA parameters in vector  $\boldsymbol{\theta}$ ,  $N$ -source powers in  $\mathbf{p} = [p_1, \dots, p_N]^T$ , and the noise variance  $\nu$ ). This observation suggests that it is possible to significantly increase the number of uniquely identifiable sources in an array from  $\mathcal{O}(M)$  to  $\mathcal{O}(M^2)$  if the number of redundant entries in the covariance matrix is reduced. Therefore, from the viewpoint of improving the number of detectable sources for a fixed number of sensors, we should deviate from the conventional ULA array structure. The reason is that the covariance matrix in the case of a ULA and uncorrelated source signals is a Hermitian Toeplitz matrix, which contains only  $(2M - 1)$  real-valued independent entries.

In fact, the covariance matching approach in Table 1 combined with the concepts of sparse reconstruction in DoA estimation and positive definite Toeplitz matrix low-rank factorization has inspired an interesting line of research on nested and coprime arrays that aims at designing favorable nonredundant spatial sampling patterns [19]. These types of arrays include the class of minimum redundancy and augmentable arrays whose design approaches rely on thinned ULA geometries. One example is the sparse nonuniform arrays with intersensor spacings being integer multiples of a common baseline  $\delta$ . These geometries have the benefit over arrays with arbitrary noninteger spacings that they allow the use of search-free DoA estimation methods (compare the gridless sparse methods introduced in the previous section) and spatial smoothing techniques to build subspace estimates of the required rank. More precisely, given the stochastic signal model in the uncorrelated source case with source powers  $\mathbf{p}$ , hence,  $\mathbf{R} = \mathbf{A}(\boldsymbol{\theta})\mathbf{D}(\mathbf{p})\mathbf{A}^H(\boldsymbol{\theta}) + \nu\mathbf{I}_M$ , an equivalent single-snapshot model is obtained from vectorization.

Defining  $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{A}^*(\boldsymbol{\theta}) \odot \mathbf{A}(\boldsymbol{\theta})$  as the steering matrix of a so-called virtual difference coarray, where  $\odot$  stands for the Khatri-Rao product, i.e., column-wise Kronecker product, the vectorized covariance model reads  $\mathbf{r} = \text{vec}(\mathbf{R}) = \mathbf{C}(\boldsymbol{\theta})\mathbf{p} + \nu\text{vec}(\mathbf{I}_M)$  [19]. For this model, the  $\ell_{2,1}$ -norm regularized LS approach in (8) is a suitable candidate for DoA estimation. An interesting alternative approach that does not rely on sparsity but on the nonnegativity property of the source power vector  $\mathbf{p}$  is proposed in [33].

In the vectorized covariance model, the number of identifiable sources is fundamentally limited by the Kruskal rank of the difference coarray steering matrix  $\mathbf{C}(\boldsymbol{\theta})$ . Hence, the design objective for the physical array is to place the sensors such that, in the difference coarray, redundant rows of the difference coarray steering matrix  $\mathbf{C}(\boldsymbol{\theta})$  are avoided and the number of contiguous lags is maximized. Avoiding redundant rows is equivalent to maximizing the diversity of the coarray, i.e., the number of different lags in the coarray. Maximizing the number of contiguous lags in the coarray corresponds to

maximizing the size of the largest “hole-free” ULA partition of the coarray, which in turn is directly related to the Kruskal rank of the ULA partition (and therefore also the Kruskal rank of the entire difference coarray) due to the Vandermonde property of the ULA steering matrix.

Because of the Khatri-Rao structure of the difference coarray steering matrix  $\mathbf{C}(\boldsymbol{\theta})$ , the Kruskal rank and, thus, the number of uniquely identifiable sources, grows quadratically rather than linearly with the number of physical sensors  $M$ . While coarray designs allow one to significantly increase the number of detectable sources for a given number of physical sensors using standard DoA estimation algorithms, recent theoretical performance results reveal that, in the regime where the number of sources  $N$  exceeds the number of sensors  $M$ , the mean-square estimation error of the MUSIC algorithm applied to the coarray data does not vanish asymptotically with SNR [34].

In the minimum redundancy array design, the number of contiguous lags in the difference coarray, and hence the size of the largest ULA partition, is maximized by definition, which generally requires a computationally extensive combinatorial search over all possible spatial sampling patterns. The nested and coprime array designs, in contrast, represent systematic design approaches associated with computationally efficient analytic array design procedures [19]. Nested array and coprime arrays are composed of two uniform linear subarrays with different baselines. In the nested array structure, each subarray is composed of  $M_1$  and  $M_2$  sensors with baselines  $\delta$  and  $(M_1 + 1)\delta$ , respectively, where the first sensor of the first subarray lies at the origin and the first sensor of the second subarray is displaced by  $M_1$ . It can be shown that, with an equal split ( $M_1 = M_2$  and  $M_1 = M_2 + 1$  for even and odd  $M$ , respectively), the difference coarray becomes a ULA with  $2M_2(M_1 + 1) - 1$  elements.

Coprime arrays represent more general array structures and comprise, as the name suggests, two uniform linear subarrays with  $M_1$  and  $M_2 - 1$  sensors, respectively, with  $M_1$  and  $M_2$  being coprime numbers. The first and the second subarray have baselines  $L_1\delta$  and  $L_2\delta$ , respectively, where  $L_1 = M_2$  and  $L_2 = M_1/F$  are coprime numbers and integer  $F$  is a given array compression factor in the range  $1 \leq F \leq M_1$ .

Furthermore, the subarrays are displaced by integer multiples of the baseline  $\delta$ .

In Figure 8, the nested and coprime array structures and their respective virtual coarrays are illustrated for the case of three sensors in each subarray. While the nested array structure yields a coarray with a maximum number of contiguous lags, the coprime array structure may often be preferable in practice as it can achieve not only a larger number of unique lags, i.e., degrees of freedom of up to  $(M/2)^2 + M/2$ , but also a larger virtual coarray aperture as well as a larger minimum interelement spacing of the physical array to reduce, e.g., mutual coupling effects.

To further increase the estimation performance of DoA estimators in a coprime array structure, low-rank Toeplitz and Hankel matrix completion approaches have been proposed to fill the “holes” and augment the data in sparse virtual coarrays to the corresponding full virtual ULA [35]. This concept has been successfully applied in [36] in the context of bistatic automotive radar to improve the angular resolution without increasing the hardware costs. Similarly, in [37], matrix completion for data interpolation in coprime virtual arrays has been used for subspace estimation in hybrid analog and digital precoding with a reduced number of analog-to-digital converters and radio frequency chains in the hardware receives.

Conditions under which, in the noise-free case, the completion from a single temporal snapshot is exact have been derived in [38].

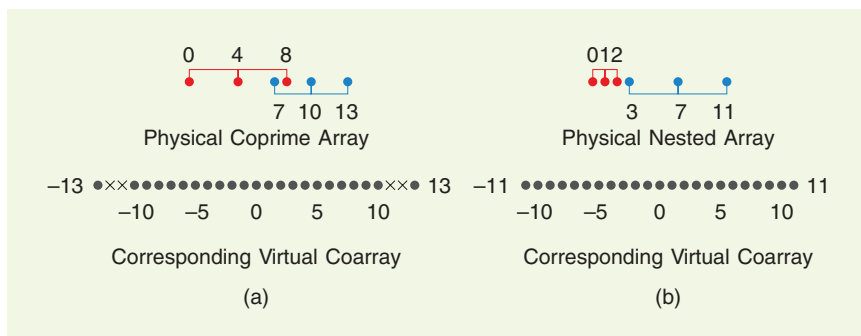
## Conclusions and future directions

In this review article, we revisit important developments in area sensor array processing for DoA estimation in the past three decades from a modern optimization and structure exploitation perspective. From several illustrative examples, we show how novel concepts and algorithms that have advanced the research field in the last decades are proposed to solve, in some way or the other, the same notoriously challenging multisource optimization problems, such as the well-known classical DML problem. Advances in convex optimization research and the development of efficient interior point solvers

for semidefinite programs made it possible to compute close-to-optimal approximate solutions to these problems with significantly reduced effort.

In addition, we also show how particular structure in the measurement model has been efficiently exploited to make the problems computationally tractable, both in terms of an affordable computational complexity as well as in terms of well posedness of the problem for identifying the parameters of interest. Nevertheless, we remark that our coverage of the sensor array processing

**Sparsity-based methods have their merit in difficult scenarios with low sample size or highly correlated and even coherent source signals.**



**FIGURE 8.** Examples of the coprime and nested array structure consisting of two subarrays, each with three sensors. The baseline of each physical subarray is a multiple of the baseline of the virtual coarray.

research of the past three decades is by no means meant to be exhaustive. Given the long history of array signal processing, by now, this field of research can certainly be considered mature. Despite all the progress that has been made over the past decades, many important and fundamental research problems in this area have not yet been solved completely and require new ideas and concepts, and some of these are outlined next.

### *Harmonic retrieval in large dimensional datasets*

One example is the extension of the parameter estimation in 1D spaces, such as in conventional DoA estimation, to higher dimensional spaces, e.g., as required in the aforementioned parametric MIMO channel estimation problem. With the trend to massive sensing systems and high dimensional datasets, the harmonic retrieval problem in extremely large dimensions gains significant interest. Due to the phenomenon known as the *curse of dimensionality*, where the computation workload increases exponentially with the number of dimensions, the extension of 1D DoA estimation methods to higher dimensions is not straightforward. Existing works on multidimensional harmonic retrieval either consider rather low dimensions or rely on dimensionality reduction approaches, i.e., projecting the multidimensional datasets into lower dimensions. This is, however, associated with a significant performance degradation if sources are not well separated in the projected domain.

### *Incorporation of signal properties as prior information*

The use of particular structures in the array manifold, which is considered in this review article, is only one form of incorporating additional prior information into the estimation problem. As more information is exploited, the parameter estimation task can be correspondingly simplified, and the estimation performance is enhanced. Theoretical investigations on the general use of additional side information incorporated in the DoA estimation problem as well as its estimation performance bound are addressed in [39]. In modern applications, the received signals as well as the waveforms often exhibit additional properties that can be exploited while designing novel DoA estimators. For example, constant modulus properties of the transmitted signals or signal waveforms with temporal dependence as, e.g., in radar chirp signals, enable coherent processing across multiple snapshots and dramatically enhance the resolution capabilities. Another example of signal exploitation is the DoA estimation with quantized or one-bit measurements, which has been studied in [40].

### *Robust sensor array processing*

In many real-world applications, the classical array signal model may be oversimplistic. This can lead to a severe performance degradation of conventional high-resolution DoA estimation methods, which are known to be very sensitive

to even small model mismatches. In recent years, significant efforts have been made to design DoA estimation methods that are robust to various model mismatches, including array imperfections due to miscalibration, impairments of the receiver front ends, mutual coupling between antennas, waveform decorrelation across the sensor array in inhomogeneous media, and multipath environments as well as impulsive and heavy-tailed noise [41].

### *Combining model-based with data-driven DoA estimation*

Recently, data-driven machine learning approaches have been successfully introduced in many areas of signal processing to overcome the existing limitations of traditional model-based approaches. Data-driven algorithms have the benefit that they naturally generalize to various statistics of the training data and, thus, are flexible to adapt to time-varying estimation scenarios. As such, data-driven algorithms are potential candidates to overcome the aforementioned challenges in DoA estimation.

However, typical off-the-shelf data-driven algorithms are known to be data hungry, which limits their practical use in many DoA estimation applications. Recently, hybrid model-and-data-driven methods were proposed in the context of deep algorithm unfolding, which combine the benefits of both approaches. The hybrid algorithms inherit the structure of existing model-based

algorithms in their learning architecture to reduce the number of learning parameters and therefore speed up the learning and improve the generalization capability of the algorithms [42].

### **Acknowledgment**

The work of Marius Pesavento was supported by the BMBF project Open6GHub under Grant 16KISK014 and the DFG PRIDE Project PE 2080/2-1 under Project No. 423747006.

### **Authors**

**Marius Pesavento** (pesavento@nt.tu-darmstadt.de) received his Dr.-Ing. degree in electrical engineering from Ruhr-University Bochum, Germany, in 2005. He is a professor at the Department of Electrical Engineering and Information Technology, TU Darmstadt, 64289 Darmstadt, Germany. He is the recipient of the 2005 Young Author Best Paper Award of *IEEE Transactions on Signal Processing* and chair of the technical area committee Signal Processing for Multisensor Systems of the EURASIP. His research interests are in statistical and array signal processing, multiuser communications, optimization, and learning.

**Minh Trinh-Hoang** (thminh@nt.tu-darmstadt.de) received his Dr. -Ing. in electrical engineering from TU Darmstadt, Germany in 2020. From 2020 to 2021 he was a postdoctoral researcher at the Department of Electrical Engineering and Information Technology at TU Darmstadt. He is currently a research and development engineer at Rohde & Schwarz, 81614 Munich, Germany. He is a recipient of the Best PhD

**In modern applications, the received signals as well as the waveforms often exhibit additional properties that can be exploited while designing novel DoA estimators.**

Thesis Award of the Association of Friends of TU Darmstadt. His research interests include statistical signal processing, array processing, parallel hardware architecture, numerical computation, and large-scale optimization.

**Mats Viberg** (mats.viberg@bth.se) received his Ph.D. degree from Linköping University, Sweden. During 1993–2018, he was a professor of signal processing at the Chalmers University of Technology, Sweden, and since 2018, he has served as the vice-chancellor at the Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. He has served in various capacities in the IEEE Signal Processing Society, including chair of two technical committees, and he has received two paper awards. His research interests are in statistical signal processing and its various applications. He is a Fellow of IEEE and a member of the Royal Swedish Academy of Sciences (KVA).

## References

- [1] F. Gao, Z. Tian, E. G. Larsson, M. Pesavento, and S. Jin, "Introduction to the special issue on array signal processing for angular models in massive MIMO communications," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 882–885, Sep. 2019, doi: 10.1109/JSTSP.2019.2938880.
- [2] H. L. Van Trees, *Optimum Array Processing*. New York, NY, USA: Wiley, 2002.
- [3] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996, doi: 10.1109/79.526899.
- [4] M. Trinh-Hoang, M. Viberg, and M. Pesavento, "Partial relaxation approach: An eigenvalue-based DOA estimator framework," *IEEE Trans. Signal Process.*, vol. 66, no. 23, pp. 6190–6203, Dec. 2018, doi: 10.1109/TSP.2018.2875853.
- [5] J.-J. Fuchs, "Detection and estimation of superimposed signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 1998, vol. 3, pp. 1649–1652, doi: 10.1109/ICASSP.1998.681771.
- [6] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004, doi: 10.1109/IT.2004.828141.
- [7] J. H. Ender, "On compressive sensing applied to radar," *Signal Process.*, vol. 90, no. 5, pp. 1402–1414, May 2010, doi: 10.1016/j.sigpro.2009.11.009.
- [8] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, Mar. 2011, doi: 10.1109/TSP.2010.2090525.
- [9] Z. Yang, J. Li, P. Stoica, and L. Xie, "Chapter 11 - Sparse methods for direction-of-arrival estimation," in *Academic Press Library in Signal Processing*, vol. 7, R. Chellappa and S. Theodoridis, Eds. New York, NY, USA: Academic, 2018, pp. 509–581.
- [10] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005, doi: 10.1109/TSP.2005.850882.
- [11] C. Steffens, M. Pesavento, and M. E. Pfetsch, "A compact formulation for the  $\ell_{2,1}$  mixed-norm minimization problem," *IEEE Trans. Signal Process.*, vol. 66, no. 6, pp. 1483–1497, Mar. 2018, doi: 10.1109/TSP.2017.2788431.
- [12] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7465–7490, Nov. 2013, doi: 10.1109/IT.2013.2277451.
- [13] Z. Yang, L. Xie, and P. Stoica, "Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3685–3701, Jun. 2016, doi: 10.1109/IT.2016.2553041.
- [14] A. Scaglione, R. Pagliari, and H. Krim, "The decentralized estimation of the sample covariance," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 1722–1726, doi: 10.1109/ACSSC.2008.5074720.
- [15] M. Pesavento, A. Gershman, and K. Wong, "Direction finding in partly calibrated sensor arrays composed of multiple subarrays," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2103–2115, Oct. 2002, doi: 10.1109/TSP.2002.801929.
- [16] C. See and A. Gershman, "Direction-of-arrival estimation in partly calibrated subarray-based sensor arrays," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 329–338, Feb. 2004, doi: 10.1109/TSP.2003.821101.
- [17] A. Moffet, "Minimum-redundancy linear arrays," *IEEE Trans. Antennas Propag.*, vol. 16, no. 2, pp. 172–175, Mar. 1968, doi: 10.1109/TAP.1968.1139138.
- [18] Y. Abramovich, D. Gray, A. Gorokhov, and N. Spencer, "Positive-definite Toeplitz completion in DOA estimation for nonuniform linear antenna arrays. I. Fully augmentable arrays," *IEEE Trans. Signal Process.*, vol. 46, no. 9, pp. 2458–2471, Sep. 1998, doi: 10.1109/78.709534.
- [19] P. Pal and P. P. Vaidyanathan, "Nested arrays: A novel approach to array processing with enhanced degrees of freedom," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4167–4181, Aug. 2010, doi: 10.1109/TSP.2010.2049264.
- [20] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using coprime arrays: A super resolution viewpoint," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5565–5576, Nov. 2014, doi: 10.1109/TSP.2014.2354316.
- [21] S. Qin, Y. D. Zhang, and M. G. Amin, "Generalized coprime array configurations for direction-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 6, pp. 1377–1390, Mar. 2015, doi: 10.1109/TSP.2015.2393838.
- [22] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1110–1121, May 1991, doi: 10.1109/78.80966.
- [23] B. Ottersten, P. Stoica, and R. Roy, "Covariance matching estimation techniques for array signal processing applications," *Digit. Signal Process.*, vol. 8, no. 3, pp. 185–210, Jul. 1998, doi: 10.1006/dspr.1998.0316.
- [24] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control.*, vol. 50, no. 5, pp. 1873–1896, May 2007, doi: 10.1080/00207178908953472.
- [25] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 10, pp. 1553–1560, Oct. 1988, doi: 10.1109/29.7543.
- [26] M. M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed  $\ell_{2,0}$  norm approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4646–4655, Sep. 2010, doi: 10.1109/TSP.2010.2050477.
- [27] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5145–5157, Oct. 2016, doi: 10.1109/TSP.2016.2576422.
- [28] A. Manikas and C. Proukakis, "Modeling and estimation of ambiguities in linear arrays," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2166–2179, Aug. 1998, doi: 10.1109/78.705428.
- [29] F. Matter, T. Fischer, M. Pesavento, and M. E. Pfetsch, "Ambiguities in DoA estimation with linear arrays," *IEEE Trans. Signal Process.*, vol. 70, pp. 4395–4407, Aug. 2022, doi: 10.1109/TSP.2022.3200548.
- [30] M. Babaie-Zadeh and C. Jutten, "On the stable recovery of the sparsest overcomplete representations in presence of noise," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5396–5400, Oct. 2010, doi: 10.1109/TSP.2010.2052357.
- [31] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989, doi: 10.1109/29.32276.
- [32] A. Swindlehurst, P. Stoica, and M. Jansson, "Exploiting arrays with multiple invariances using MUSIC and MODE," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2511–2521, Nov. 2001, doi: 10.1109/78.960398.
- [33] H. Qiao and P. Pal, "Guaranteed localization of more sources than sensors with finite snapshots in multiple measurement vector models using difference co-arrays," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5715–5729, Nov. 2019, doi: 10.1109/TSP.2019.2943224.
- [34] M. Wang and A. Nehorai, "Coarrays, MUSIC, and the Cramér–Rao bound," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 933–946, Feb. 2017, doi: 10.1109/TSP.2016.2626255.
- [35] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6576–6601, Oct. 2014, doi: 10.1109/IT.2014.2343623.
- [36] S. Sun and Y. D. Zhang, "4D automotive radar sensing for autonomous vehicles: A sparsity-oriented approach," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 879–891, Jun. 2021, doi: 10.1109/JSTSP.2021.3079626.
- [37] S. Haghshatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 303–318, Jan. 2017, doi: 10.1109/TSP.2016.2616336.
- [38] P. Sarangi, M. C. Hücümenoğlu, and P. Pal, "Single-snapshot nested virtual array completion: Necessary and sufficient conditions," *IEEE Signal Process. Lett.*, vol. 29, pp. 2113–2117, Oct. 2022, doi: 10.1109/LSP.2022.3213140.
- [39] T. J. Moore, B. M. Sadler, and R. J. Kozick, "Maximum-likelihood estimation, the Cramér–Rao bound, and the method of scoring with parameter constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 895–908, Mar. 2008, doi: 10.1109/TSP.2007.907814.
- [40] S. Sedighi, B. S. Mysore, R. M. Soltanian, and B. Ottersten, "On the performance of one-bit DoA estimation via sparse linear arrays," *IEEE Trans. Signal Process.*, vol. 69, pp. 6165–6182, Oct. 2021, doi: 10.1109/TSP.2021.3122290.
- [41] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robustness in Sensor Array Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2018, pp. 125–146.
- [42] J. P. Merkofer, G. Revach, N. Shlezinger, and R. J. G. van Sloun, "Deep augmented music algorithm for data-driven DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 3598–3602, doi: 10.1109/ICASSP43922.2022.9746637.

Emil Björnson<sup>1</sup>, Yonina C. Eldar<sup>2</sup>, Erik G. Larsson<sup>3</sup>,  
Angel Lozano<sup>4</sup>, and H. Vincent Poor<sup>5</sup>

# Twenty-Five Years of Signal Processing Advances for Multiantenna Communications

*From theory to mainstream technology*



©SHUTTERSTOCK.COM/TRIFF

Wireless communication technology has progressed dramatically over the past 25 years, in terms of societal adoption as well as technical sophistication. In 1998, mobile phones were still in the process of becoming compact and affordable devices that could be widely utilized in both developed and developing countries. There were “only” 300 million mobile subscribers in the world [1]. Cellular networks were among the first privatized telecommunication markets, and competition turned the devices into fashion accessories with attractive designs that could be individualized. The service was circumscribed to telephony and text messaging, but it was groundbreaking in that, for the first time, telecommunication was between people rather than locations.

There are now more than six billion subscribers worldwide, and the mobile phone remains the main wireless device, but much has changed. Traditional feature phones with physical keypads have been replaced by smartphones with large touchscreens. Telephony today constitutes a negligible fraction of the traffic, the vast majority of which amounts to packets bearing data for end-user applications. Video and audio streaming, social media, gaming, and a host of other apps, generate the bulk of the traffic. New services continue to arise and cement the smartphone’s central role in nearly every aspect of our lives. In parallel, nonhuman-operated devices are progressively coming online to form the Internet-of-Things (IoT) as society continues to be digitized.

Wireless networks have changed dramatically over the past few decades, enabling this revolution in service provisioning and making it possible to accommodate the ensuing dramatic growth in traffic. There are many contributing components, including new air interfaces for faster transmission, channel coding for enhanced reliability, improved source compression to remove redundancies, and leaner protocols to reduce overheads. Signal processing is at the core of these improvements, but nowhere has it played a bigger role than in the development of multiantenna communication. This article tells the story of how major signal processing advances have transformed the early multiantenna concepts into mainstream technology over

Digital Object Identifier 10.1109/MSP.2023.3261505  
Date of current version: 1 June 2023

the past 25 years. The story therefore begins somewhat arbitrarily in 1998. A broad account of the state-of-the-art signal processing techniques for wireless systems by 1998 can be found in [2], and its contrast with recent textbooks, such as [3], [4], and [5], reveals the dramatic leap forward that has taken place in the interim.

### Fundamentals of multiantenna communications

Traditionally, a base station (BS) at a cellular network site featured antenna panels connected to a baseband unit (BBU) that managed the digital signal processing. These panels, in turn, were tall and narrow, containing multiple vertically stacked radiating elements. By emitting the same signal from such elements, constructive superposition was leveraged to create a radiation pattern, vertically narrow and horizontally wide, that covered a swath of ground in a predefined manner. This is illustrated in Figure 1(a), with each panel's coverage region termed a *cell sector*.

At current BSs, the panels have been replaced with antenna arrays having a more symmetric aspect ratio, which results in radiated beams that can be narrow both horizontally and vertically. The signal transmitted from each antenna element is individually controlled by the BBU, which now has far stronger computational capabilities and can alter the physical shape of the produced beam over both time and frequency. Figure 1(b) illustrates such a setup, and how each beam is narrow enough to aim at a particular user. When these arrays are used in propagation environments with multiple widely spaced paths, each radiated signal loses its directional beam shape and is instead fine-tuned to make the paths superimpose coherently on a small region around the intended receiver.

Antenna arrays bring about three main categories of benefits:

1) *Beamforming gain*: The transmit beam is focused on the receiver, whereby a larger fraction of the radiated energy reaches it. Likewise, multiple receive antennas can collect

more energy from selected directions, reinforcing the beam at that end with a focus on the transmitter. The overall beamforming gain is proportional to the transmit and receive array sizes.

- 2) *Spatial diversity*: There are generally multiple paths via which signals travel between the transmitter and receiver, and the ensuing signal replicas can combine destructively. This causes signal fading, which antenna arrays can mitigate by observing multiple fading realizations simultaneously.
- 3) *Spatial multiplexing*: Multiple signals can be transmitted concurrently on different beams, either to a single user equipped with multiple antennas, or to multiple users, as in Figure 1(b). This provides a traffic multiplier or *multiplexing gain*, provided the interference among the signals can be kept at bay.

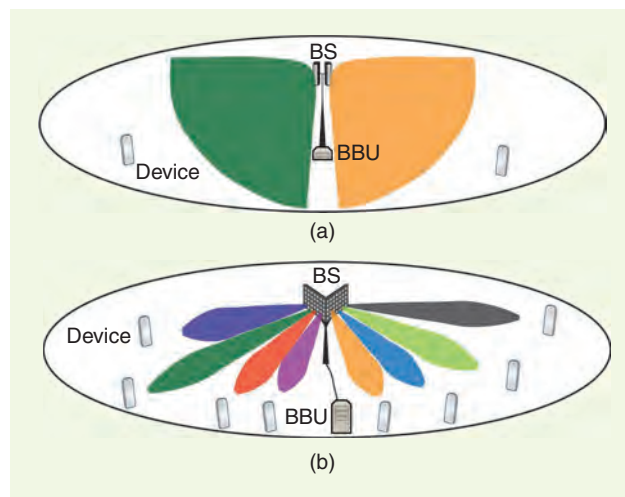
Above, and in the sequel, the beamforming gain is taken as the increase in signal power at the receiver, yet a more nuanced description would further include the reduction in interference to and from unintended users [5, Sec. 5.7]. With a careful design, beamforming can strike an optimum balance between increasing signal energy and reducing interference.

### State-of-the-art in 1998

Some of the benefits of antenna arrays were understood well before 1998, but their technology readiness levels were much different than today. Marconi himself famously capitalized on beamforming to enable wireless transatlantic communication in 1901. That experiment relied on an *array antenna*, which achieves beam directivity by connecting multiple elements to the same signal generator. The geometry of the array antenna determines the direction in which the radiated signals superimpose constructively. Hence, the beam direction is fixed and determined at the time of building and erecting the array. This is how the 2G antennas in Figure 1(a) were designed to cover a sector with a fixed beam.

A different beam direction than the one dictated by the array geometry can be realized by emitting the same signal from all of the elements, but with appropriate phase shifts. This concept was first observed experimentally by Ferdinand Braun in 1902, and it led to the phased array technology used for radar since World War II. The phase shifts can be varied over time, to scan for objects in different angular directions. Early field trials of phased arrays for 2G were conducted in 1996 [6]. The possibility of pointing the beams to user locations opened the door to stronger directivities and higher gains, since a beam no longer had to cover an entire sector. The difference between array antennas and phased arrays is illustrated in Figure 2, which also depicts the digital antenna arrays featured in 5G, where each element is connected to a separate signal generator.

In parallel with the refinement of phased arrays for beamforming over several decades, the use of multiple receive antennas for diversity also became commonplace. Spatial diversity was conceived for signal reception as far back as the 1930s [7] and builds on an intuitive principle: if the same signal reaches several physically separated antennas, it is unlikely



**FIGURE 1.** (a) A 2G deployment in 1998, consisting of fixed directive antennas that broadcast each signal into a sector. (b) A 5G deployment in 2023, entailing antenna arrays that can exploit the three main multiantenna benefits: beamforming gain, spatial diversity, and spatial multiplexing.

that the multipath propagation environment causes destructive superposition, hence signal fading, at all such antennas simultaneously. By decoding a combination of the observations at the various receive antennas, the communication becomes much more reliable.

A wireless network must of course provide an uplink connection from users to BSs, as well as a downlink connection from BSs to users. Thus, diversity is desirable in both link directions, yet transmit diversity did not emerge until the 1990s [8]. In 1998, the Alamouti space–time block code for two antennas was proposed [9] and a more general framework for space–time coding with multiple transmit antennas was published soon thereafter [10]. The principle is to repeatedly transmit a block of data symbols while varying the spatial directivity in a predetermined way (e.g., using different antennas); the receiver collects observations over a time interval and decodes them. Space–time codes are carefully crafted to not only enable decoding, but to strike a satisfactory tradeoff between high spectral efficiency (i.e., bits per second per Hertz of spectrum), high diversity, and low complexity.

Altogether, beamforming gains and spatial diversity were known by 1998, and it was largely thought that these were the two main benefits of antenna arrays: beamforming gains in the case of coherent arrays, associated with tight antenna spacings and cleanly defined directions of arrival and departure, and diversity gains in the case of arrays experiencing largely uncorrelated fading across the antennas, associated with wider spacings and rich multipath settings. The third, and ultimately the most powerful benefit of antenna arrays, spatial multiplex-

ing, was still largely under the radar. However, its seeds had already been planted in research efforts on interference-aware beamforming [11] and on communication concepts for linear

channels that couple multiple inputs into multiple outputs [12]. Unlike beamforming and diversity, which involved replicas of a single signal, these precursors of spatial multiplexing entailed the transmission and reception of distinct signals simultaneously and on the same bandwidth. Particularly prescient was the transmission and reception with two orthogonally polarized antennas, subsequently extended in a piece that featured multiantenna transmitters and

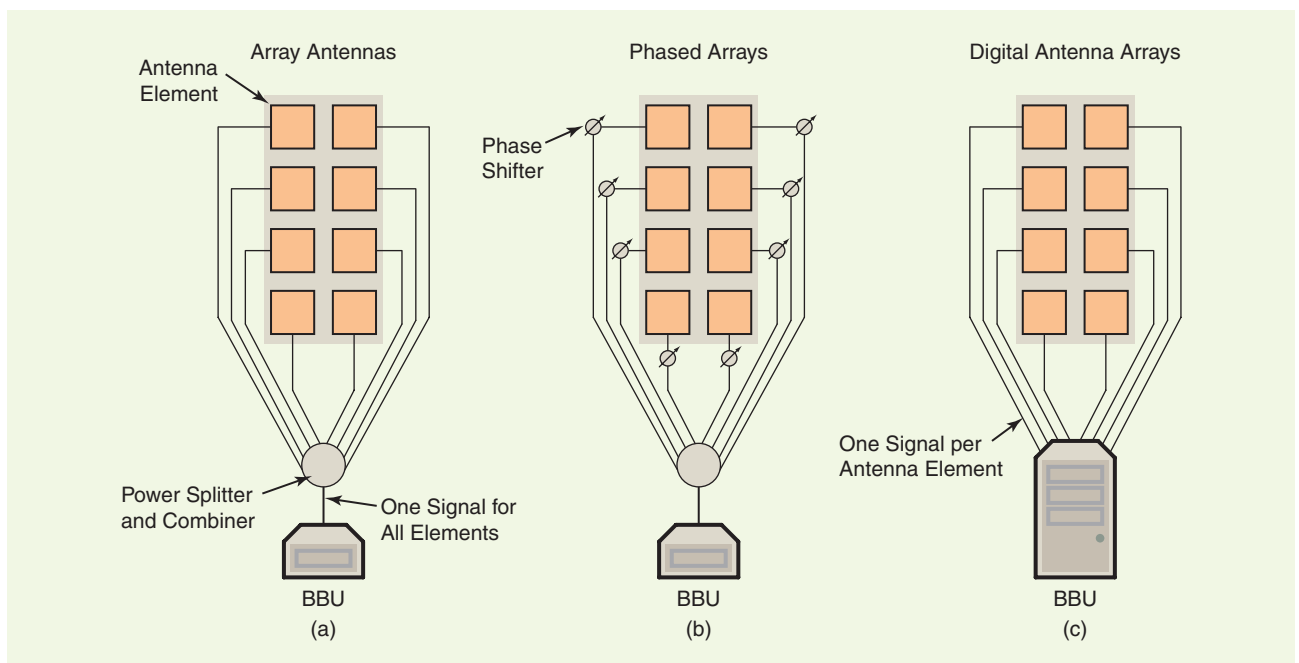
receivers with many of the ingredients required for true spatial multiplexing [13]. However, it was not until after 1998 that all of these pieces fell into place.

### External technology developments

Three external trends have heavily guided and influenced the evolution of multiantenna technology over the past decades.

- 1) The explosion in wireless traffic, which has doubled every 18 months as per Cooper’s law, along with a fundamental change in the nature of such traffic, was driven by new user behaviors and applications. An efficient network for telephony had to support many simultaneous fixed-rate connections, while today’s data networks aim at maximizing the bit rate per user device (to support certain applications) and the bit rate per unit area (to accommodate many devices).
- 2) The exponential improvement and size reduction of integrated circuits have led to systems-on-a-chip that combine radios, memory, and processors capable of advanced

**Wireless networks have changed dramatically over the past few decades, enabling this revolution in service provisioning and making it possible to accommodate the ensuing dramatic growth in traffic.**



**FIGURE 2.** There are three classical categories of arrays: (a) array antennas that generate fixed beams, (b) phased arrays that rely on phase shifters to control the beam direction, and (c) digital antenna arrays that have full control of the signal transmitted from each antenna element.

signal processing on a tiny piece of silicon. While in 1998, a digital antenna array with  $M = 2$  or  $M = 4$  elements would consist of  $M$  external antenna elements connected to  $M$  radio frequency (RF) units and one BBU, current 5G BSs can integrate  $M = 64$  elements and RF units into a single box. This development has also enabled smartphones to feature digital arrays, for now with  $M = 4$  elements.

- 3) The gradual change in the signal waveforms: There were multiple 2G standards based either on time-division multiple access (TDMA) or code-DMA (CDMA). The first versions of 3G, finalized precisely around 1998, were entirely based on CDMA, which won the battle against the competing orthogonal frequency-DMA (OFDMA). For 4G, the shift to OFDMA finally took place, and 5G retained this same waveform after a handful of alternatives were evaluated and discarded. While all waveforms are in principle compatible with antenna arrays, the choice does have a fundamental impact on what signal processing algorithms are required.

**Marconi himself famously capitalized on beamforming to enable wireless transatlantic communication in 1901.**

## Five key areas of signal processing advances

We have identified five stages of signal processing advances in the evolution of multiantenna technology from 2G to 5G and beyond. The background, new solutions, and specific insights are expounded on in the following sections.

### *From spatial diversity to spatial multiplexing*

One could argue that, all the way back to Marconi, beamforming was motivated by the interest in extending the range of coverage. In turn, diversity was motivated by the desire to increase reliability. By 1998, the exploding cost of radio spectrum ahead of 3G brought about a new and powerful necessity: increasing the spectral efficiency. The shift toward high bit-rate user applications further amplified this trend. The operational mode of antenna arrays that maximizes the spectral efficiency is spatial multiplexing and, after 1998, the atmosphere was therefore primed for it to finally come to the fore.

The prerequisite for spatial multiplexing is a multiple-input multiple-output (MIMO) communication channel, where each input/output refers to an antenna element in a digital antenna array. There are two MIMO categories: *single-user MIMO* entails a multiantenna BS and a multiantenna user device, while *multiuser MIMO* encompasses a multiantenna BS and multiple user devices.

Arguably, the main catalyst for single-user MIMO was the work in [14], which set out to design the perfect transceivers from an information-theoretic standpoint. Starting with transmit and receive digital antenna arrays and no preset conditions on how to employ them, it was found that, if the elements within each array exhibited uncorrelated fading, the optimum strategy was to have each radiate an independent data-carrying signal. This was radically novel in that it sought to exploit, rather than counter, multipath propagation; it

is the very existence of multiple paths that allows the receiver to observe a distinct linear combination of the transmit signals at each receive antenna, from where those transmit signals can be resolved. The number of signals that can be spatially multiplexed is then limited by the minimum of the number of

transmit and receive antenna elements. In follow-up work, a specific architecture was proposed to effect such spatial multiplexing, the so called layered architecture, which was remarkable in that it could be built with off-the-shelf encoders and decoders and did not require the transmitter to know anything about the channel [15]. Additional

results progressively solidified the theoretical underpinnings [16]. In particular, the idea of transmitting concurrent signals, one from each antenna element, was generalized to the transmission of concurrent beams from all elements at once. Phased arrays cannot achieve such spatial multiplexing because they only create one beam at a time, and digital antenna arrays are decidedly necessary.

Multiuser MIMO can be traced back to signal processing concepts for simultaneous uplink reception from multiple users [17] and simultaneous downlink beamforming to users in different angular directions [18]. Here, the number of signals that can be spatially multiplexed is not limited by the number of antenna elements per user, but rather across all users; even if each user features a single element, it is possible to spatially multiplex one signal to/from each one. This major advantage comes at the expense of the BS having to carefully arrange the transmit and receive beams, such that each one matches with the multipath characteristics of its intended user and there is minimal interference among them, as in Figure 2(b). With that, every user can transmit continuously and over the entire system bandwidth, rather than only in a time slot and/or frequency subband, reflecting the spatial multiplexing benefit. Multiuser MIMO is a generalization to multipath settings of classical space-DMA (SDMA), whereby users share a channel in space rather than in time or frequency. Interestingly, the SDMA concept is more than 20 years older than single-user MIMO [19], which showcases that establishing many simultaneous user connections was long perceived more important in wireless networks than achieving high data rate per connection.

The potential of MIMO, in both its single-user and multiuser fashions, sparked a chain reaction that spread rapidly through academia and industry, bringing much excitement by the early 2000s. Cellular standardization bodies, in particular, the 3G partnership project (3GPP) adopted it in a limited fashion for late 3G releases and then as an integral part of the designs beginning with 4G. Even faster was the adoption within Wi-Fi, with the first version including MIMO certified in 2007 and supported by a multitude of devices, including laptops, tablets, and smartphones.

MIMO harnesses the three dimensions of benefits shown in Figure 3: beamforming gain, spatial diversity, and spatial multiplexing. A clear understanding of how these benefits are



related has emerged over time. Fundamental tradeoffs have been identified, for given array configurations and channel conditions.

- *Beamforming* is a special case of spatial multiplexing where a single beam is transmitted to a single user. This is in fact the optimum strategy when the signal-to-noise ratio (SNR) is low; maximizing the signal energy is then of the essence, and the best recipe is to concentrate all the radiated energy on the strongest beam. At high SNR, in contrast, energy is plentiful and can be spread over multiple beams, to the point that it is optimum to activate as many beams as the channel and antenna counts allow. This is represented by the blue plane in Figure 3.
- *Spatial diversity*, roughly quantified as the number of independently faded signal replicas, and *spatial multiplexing*, meaning the number of concurrent beams, cannot be simultaneously maximized. These two quantities are rather subject to a tradeoff [20]. At the extreme points of this diversity-multiplexing tradeoff, one of the quantities is maximized while the other stands at a minimum. Various combinations are feasible at intermediate points of the tradeoff curve, which is cartooned on the yellow plane of Figure 3.

As mentioned, the choice between beamforming and spatial multiplexing is dictated by the SNR, hence, by its underlying parameters (e.g., transmit power, channel attenuation, noise power). In turn, the mix of diversity and multiplexing depends on whether the priority is to increase reliability or spectral efficiency. However, the operating point should be selected holistically, and over the years this has caused the mix to shift toward less diversity and more multiplexing. Indeed, as successive system generations have spanned ever broader bandwidths, more and more diversity has been reaped in the frequency domain. The rewards of additional diversity rapidly saturate, thus the need for spatial diversity has abated [21]. This of course does not apply to narrowband control channels or to low-power short-packet IoT communication, where spatial diversity remains important, but it does hold for the user data channels that carry the bulk of the traffic in cellular networks.

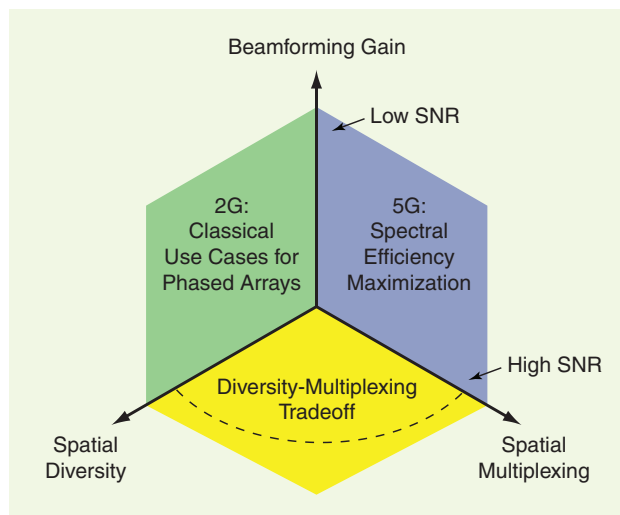
### From spatial multiplexing to massive MIMO

The basics of MIMO, developed under the premises of perfect channel state information (CSI) and rich scattering, indicate that an arbitrarily high spectral efficiency can be achieved by deploying sufficiently many antennas and serving many users at once. However, the practical challenges became apparent when the technology was first commercialized. The spatial multiplexing capability in single-user MIMO was often restricted by limited scattering, while multiuser MIMO is restrained by imperfect CSI. Massive MIMO, a new form of multiuser MIMO that originated from [22], was developed in the 2010s to address these issues and is now at the heart of 5G.

**While all waveforms are in principle compatible with antenna arrays, the choice does have a fundamental impact on what signal processing algorithms are required.**

The new aspects of massive MIMO are as follows. First, it relies on having many more BS antennas than spatially multiplexed users. This design choice renders the beams relatively narrow (e.g., in the sense of focusing on a small region around the intended receiver), hence there is likely to be little overlap among beams focused on distinct users. Moreover, by virtue of these favorable conditions, whatever little interference exists can be suppressed through low-complexity linear signal processing: for example, regularized zero-forcing that fine-tunes each beam's focal area to balance a strong beamforming gain with low interference [23]. At the same time, and again because of the excess BS antennas, the effective channels provided by these beams harden, meaning that they become very stable and subject to only minimal fading fluctuations.

Second, massive MIMO is tailored for resource-efficient CSI acquisition. The main estimation principle is to emit separate pre-defined pilot signals from each antenna element and then gauge the channel coefficients from the observations of these pilots at the receive elements. Massive MIMO adopted time-division duplexing (TDD), where the same bandwidth is utilized, in alternating fashion, for uplink and downlink. Since, by virtue of reciprocity, the channel is then identical in both directions, it suffices to estimate its coefficients in one direction. Specifically, the CSI required for both uplink and downlink is obtained from uplink pilots. The necessary pilot resources are thus determined by the number of multiplexed users, with no dependence on the number of BS antennas. In contrast, many previous commercial implementations of multiuser MIMO were based on frequency-division duplexing (FDD), where the uplink and downlink channels were

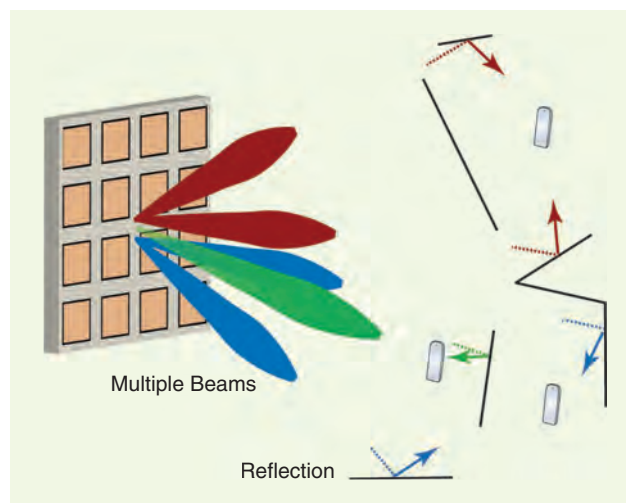


**FIGURE 3.** There are three benefit dimensions of multiantenna communication that have developed in past decades. From 2G to 5G, systems have shifted from the green plane to the blue plane; that is, spatial diversity has gradually been replaced by spatial multiplexing. Since spatial multiplexing requires high SNRs to be practically useful, beamforming gains remain essential at low SNRs.

entirely different, or TDD operation without using reciprocity. The downlink operation then required the BS to transmit as many pilots as it has antennas, except in specific propagation scenarios where the channels can be parametrized using a few angles. Moreover, each user needed to quantize and feedback its channel estimates to the BS. In a typical 5G setup with  $M = 64$  BS antennas that spatially multiplex  $K = 8$  users, the FDD alternative would require  $M/K = 8$  times as many pilots and a proportional amount of extra CSI feedback.

The TDD operation is particularly helpful in complex propagation environments with many paths per user, such as the one sketched in Figure 4, where the optimum downlink transmission spreads a user's signal energy in many directions to match the reflecting objects. CSI acquisition through uplink pilots automatically captures these fine characteristics, without any prior channel knowledge or array calibration. In FDD operation, besides requiring vastly more pilot resources, essential channel details are lost in the feedback quantization. In cellular networks, pilot signals must be reused with care across cells to avoid pilot contamination phenomena, whereby BSs inadvertently beamform toward pilot-sharing users in neighboring cells. This is particularly a concern in TDD operation, where uplink estimation errors also affect the downlink. A multitude of signal processing and resource allocation schemes have been developed over the past decade to alleviate pilot contamination [3], [4].

Massive MIMO provides a solid foundation for practical signal processing design. While sophisticated information theory for multiuser MIMO existed already by the 2000s [24], it was largely limited to scenarios with perfect CSI. As CSI quality is the main limiting factor of multiuser MIMO



**FIGURE 4.** The propagation channel consists of a multitude of specularly or diffusely reflecting objects. Such channels are very challenging to estimate with sufficient accuracy to enable multiuser MIMO communication, but the TDD CSI estimation approach in massive MIMO manages this.

**The potential of MIMO, in both its single-user and multiuser fashions, sparked a chain reaction that spread rapidly through academia and industry, bringing much excitement by the early 2000s.**

performance, this constrained the practical usefulness of the available theory. Thanks to the reliance on linear signal processing, massive MIMO analyses successfully handle imperfect CSI and hardware imperfections, resulting in rigorous and mathematically clean spectral efficiency expressions that not only predict actual performance accurately, but serve as effective tools for system optimization (e.g., pilot allocation, power control, and beamforming). Massive MIMO theory not

only turned multiuser MIMO into a practically feasible technology, the analytical elegance also expanded the way information theory for wireless communication can be taught [3], [4].

As mentioned, uplink–downlink reciprocity in TDD operation is important for massive MIMO. Reciprocity holds for the over-the-air propagation as long as the channel impulse response remains constant: that is, provided the duplexing takes place

within the channel coherence time. However, the transceiver hardware is generally not reciprocal between transmission and reception, for instance due to mismatches in the local oscillators. Such hardware nonreciprocity calls for a calibration procedure that phase-synchronizes the antennas within each array through occasional mutual measurements.

Today, 5G BSs feature almost exclusively massive MIMO configurations in TDD bands, with arrays of  $M = 32$  or  $M = 64$  antennas being the most common. Early on, there were concerns that the signal processing would entail an exceedingly high energy consumption, but this concern was later dispelled, and dedicated systems-on-chip are now available that implement clever signal processing algorithms for the entire BBU, including massive MIMO, at reasonable energy costs.

### *A quest for more bandwidth at higher frequencies*

Bit rate has long been the performance metric that users of wireless technology are most familiar with, and hence wireless technology has evolved to support higher values thereof. The bit rate enjoyed by a single device equals the product of the spectral efficiency and the spectral bandwidth. Therefore, besides being driven higher by single-user MIMO, bit rates have expanded over time thanks to the allocation of new frequency bands. A 2G network typically had access to 20 MHz at carrier frequencies around 1 GHz, while current 5G networks primarily span 100 MHz in the 3.5-GHz range, with the standard supporting in excess of 500 MHz.

The radio spectrum is a limited natural resource shared by a multitude of technologies, including those beyond the civilian wireless communication arena considered in this article. While a few sub-6-GHz bands have been refarmed from outdated technologies to cellular networks, the strive for fresh bandwidth inevitably pushes systems toward ever higher frequencies. In particular, millimeter-wave (mmWave) bands, nominally starting at 30 GHz, are now part of 5G. First-generation mmWave technology has been rolled out by a few telecom

operators, while the bulk of them wait for the hardware to mature and for the 3.5-GHz band to become congested.

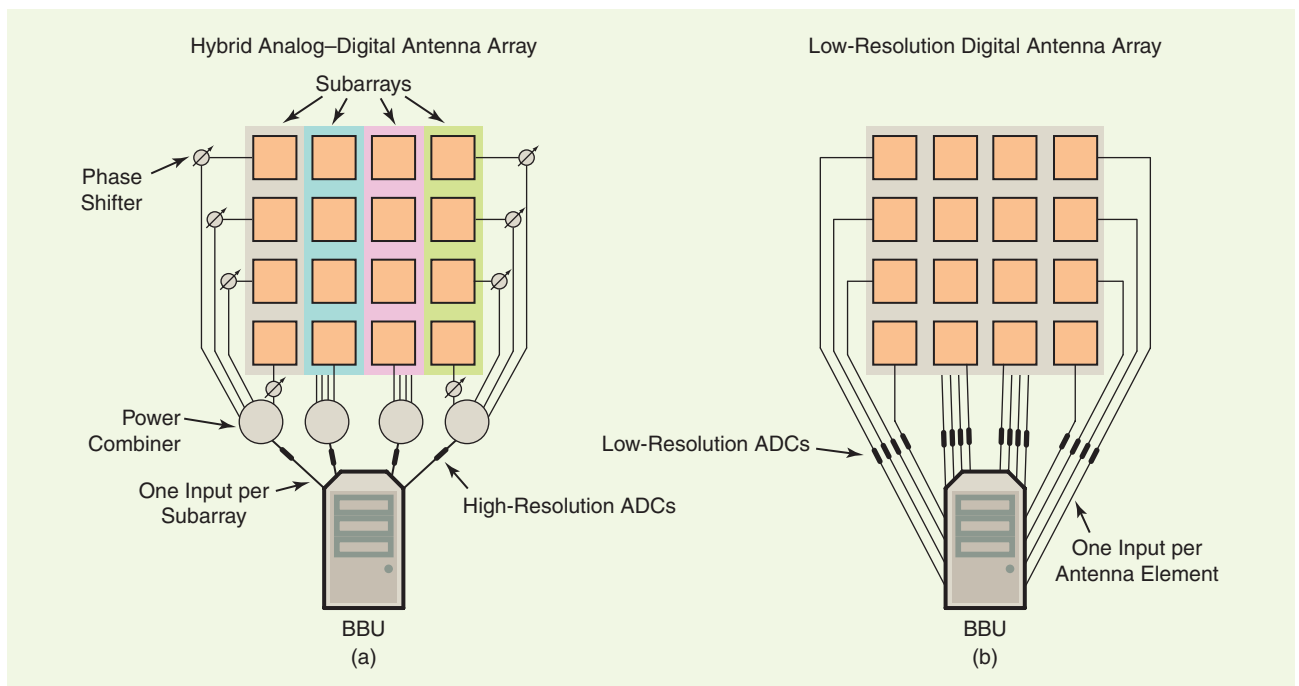
The field strength of a signal radiated from a point source in free space attenuates with distance in a frequency-independent manner. Then, the power captured from such an electric field is proportional to the receiver's aperture and, since the size of an antenna element shrinks with the wavelength, an increased carrier frequency necessitates further antenna elements to maintain the desired aperture. Moreover, the channel conditions in cellular networks become steadily more challenging as the frequency shifts up due to reduced scattering and diffraction, and steeper penetration losses, all of which call for beamforming gains. Multiantenna technology is therefore paramount at mmWave frequencies.

At the same time, implementation becomes difficult, and not only because of the added hardware components and the huge dimensionalities in digital signal processing. When moving to higher frequencies and broader bandwidths, hence to faster sampling rates, power amplifier efficiency and dissipation in analog-to-digital converters (ADCs) are further issues that need attention. The signal processing community has explored two main ways to deal with the hardware and algorithmic complexity [25].

The first option is to reduce the number of RF units, particularly converters, by designing transceivers as a mix of phased arrays and digital antenna arrays. The resulting hybrid analog-digital antenna array is illustrated in Figure 5, where each column is a phased subarray that is connected separately to the BBU, such that different signals can be transmitted and received. Each subarray can form a single beam and spatial

multiplexing can then be applied through different linear combinations of those beams. If the channel features a small number of propagation paths, each subarray can focus a beam on one of those paths, and the communication performance of a digital antenna array can be attained with fewer hardware components. In the multipath scenario illustrated in Figure 4, five distinct beam directions are sufficient to communicate effectively. Hybrid antenna arrays can take other forms, but generally entail a semianalog beamforming implementation with more antenna elements than digital ports [26]. There are several prices to pay for abandoning the digital antenna array paradigm. One can only transmit as many beams as there are digital ports and the beamforming fidelity is crippled in wide-band systems since combinations of the same beams must be used on all subcarriers. Channel estimation becomes more intricate since each phased array must sweep through as many beam directions as it has elements in order to excite all channel dimensions.

An alternative to reducing the number of RF units is digital antenna arrays with lowered ADC resolutions, as also illustrated in Figure 5. The energy consumption of an ADC grows exponentially with the resolution, hence enormous energy reductions are possible by moving from the conventional 15 bits per sample down to, say, five bits per sample. And yet, since an array with  $M$  elements and  $b$ -bit ADCs collects a total of  $bM$  bits per sample period, the total number of ADC bits can still be sizeable even if  $b$  is small, explaining why a high spectral efficiency can be maintained [27]. The extreme case of uplink massive MIMO with  $b = 1$  happens to be analytically tractable [28], which has facilitated the emergence of signal



**FIGURE 5.** Conventional digital antenna arrays incur a high energy consumption when implemented at mmWave frequencies. There are two main ways to circumvent this: (a) Reduce the number of components using hybrid analog-digital antenna arrays, consisting of multiple phased subarrays; (b) Design digital antenna arrays with simplified components, such as low-resolution ADCs.

processing algorithms that compensate for the ensuing quantization distortion. The downlink counterpart involves low-resolution digital-to-analog converters [29].

Reduced bit resolution is also viable for hybrid antenna arrays; if the analog signal combining prior to quantization and the digital postprocessing is properly optimized for the communication task at hand, the signal content becomes more amenable to a low-bit representation [30].

The initial 5G mmWave products are based on hybrid arrays, but there are indications that the low-resolution approach might eventually become the preferred solution [31].

### Further opportunities for dimensionality reduction

The joint evolution toward arrays with more antenna elements and wider bandwidths, requiring higher sampling rates, makes it essential not to overdesign the transceivers. As an alternative to scaling up a conventionally small digital array, Figure 5 showcases two ways of increasing the antenna element counts while reducing the hardware complexity per element. Both approaches capitalize on the massive MIMO philosophy of having more antenna elements than multiplexed beams, which enables a reduction in the beamforming exactness because the beams are so narrow that interuser interference is low anyway. Further dimensionality reductions are possible by exploiting the structure of the channel—say, sparsity in the angular, frequency, and time domains—or by exploiting the specific task the system is designed to address.

Besides being exponential in the resolution, the energy consumption of an ADC is proportional to the sampling rate. A host of ideas based on sub-Nyquist sampling and compressed sensing have been proposed to reduce the sampling rate by exploiting various forms of channel structure that exist in many scenarios [32]. In mmWave channels with a small num-

ber  $N$  of propagation paths, there might be only roughly  $N$  nonzero taps in the channel impulse response regardless of the bandwidth. Such time-domain sparsity in the channel response can be leveraged to reduce the sampling rate [33]. The  $N$  paths are likely distinct also in the angular domain, given that BSs

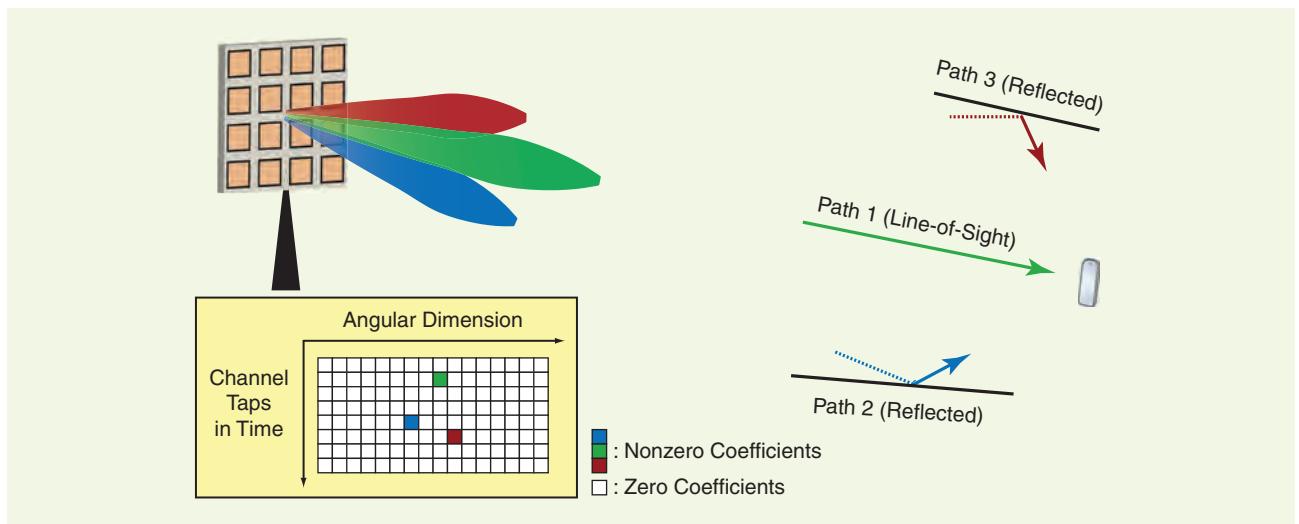
are usually deployed high above the environmental clutter; reflections take place only locally around each user, subtending a small angular spread at the distant BS. Figure 6 illustrates such a scenario with  $N = 3$  distinct paths, supporting three mul-

tiplexed signals. The line-of-sight path has a distinctly short delay, while the two remaining paths have similar delays but are clearly distinguishable in the angular realm. The joint channel sparsity is represented by the three colored entries in a time-angle matrix, with the vast majority of entries containing no propagation paths. Combining these forms of sparsity with modern compressed-sensing tools enables hefty reductions in sampling rates.

Many data services exhibit intermittent activity patterns; among the thousands of devices associated with a BS, only a small subset requires data transfers within a given time slot. Signal processing can enable these devices to transmit efficiently without requiring a preceding access procedure. The key is to assign each device with a unique but nonorthogonal pilot sequence and then utilize sparsity in the user domain along and the large number of spatial samples obtained over an antenna array to enable user identification and channel estimation [34]. The joint user and data detection problem has also been approached using compressed sensing methods [35].

Commercial massive MIMO products already exploit some elementary channel sparsity; for instance, the received signals over the many antennas might be transformed into an equal number of angular dimensions. The dimensions that contain little power are discarded in the early stages of the digital

**The MIMO technology is now present in every smartphone and BS that enters the market.**



**FIGURE 6.** The channel in mmWave systems with large spectral bandwidths and antenna counts might exhibit sparsity in the joint time-angle domain. In this example, there are  $N = 3$  paths that are distinct in both time and angle. The sparse impulse response can be exploited along with compressed sensing techniques to reduce the sampling rate, thereby lowering power consumption.

uplink processing to shrink the dimensionality of the remaining computations. However, the more radical compressed sensing solutions are yet to be brought to life.

### Machine learning-based algorithmic refinements

One of the most active areas in contemporary signal processing is machine learning (ML). While this is a many decades-old discipline, the increased availability of large amounts of data and processing power has, in recent years, greatly enhanced its potential to transform the implementation of many signal processing tasks from more traditional model-driven algorithms into data-driven ones. This transformation is also taking place in the context of signal processing for wireless communications, which has traditionally been very heavily (and successfully) model-based. Several trends are driving this transformation. One trend is that, with the vast amount of IoT and machine-type connections that coexist with human-type broadband connections, wireless network traffic is becoming increasingly intricate to model accurately, thereby making network operation difficult to optimize. Another trend is that antenna arrays and other sensors are becoming pervasive on smartphones and other connected devices, hence the volume of data available for learning is swelling dramatically. Yet a third trend is that the amount of processing power distributed throughout wireless networks is growing rapidly, giving rise to paradigms such as fog and edge computing.

There is a confluence of ML and communications in the optimization of wireless networks. This is a very natural application for ML since the operation of these networks involves a multitude of tasks that ML is good at addressing, including

inferential tasks, such as channel estimation, signal detection, and data decoding, as well as decision-making tasks such as routing, access control, and resource allocation. ML-based solutions can capture practical characteristics that were overlooked by the models, underpinning existing algorithms. However, to ensure that ML algorithms improve upon the existing, it is essential to initiate the training procedure judiciously.

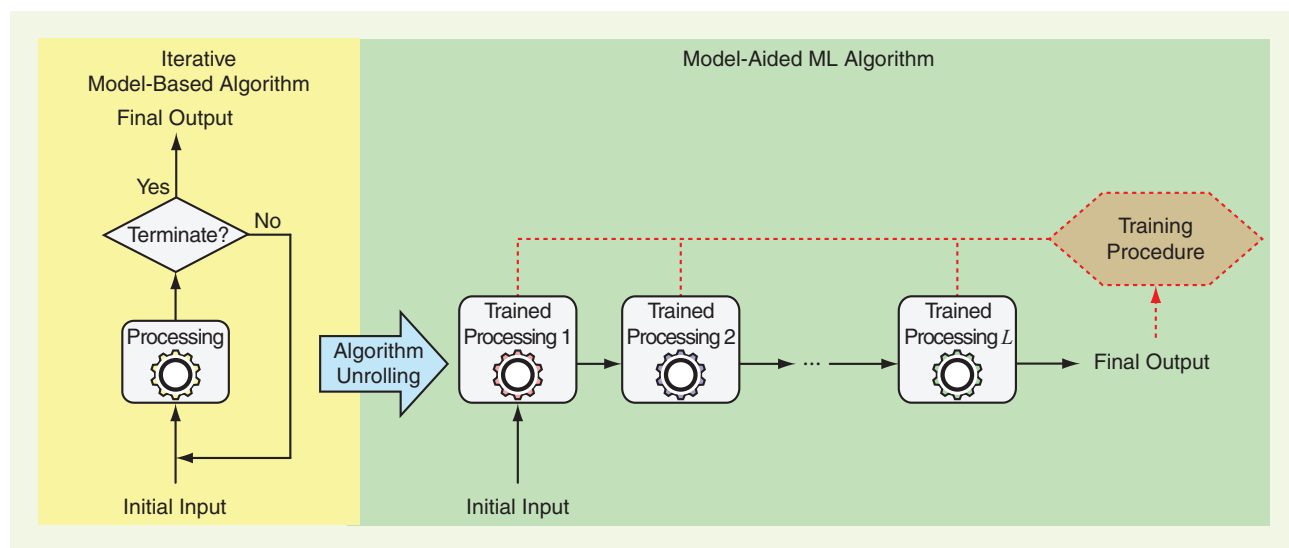
The model-aided ML paradigm provides a structured way to transfer classical know-how from the signal processing community onto new ML algorithms [36]. Figure 7 exemplifies how an existing iterative algorithm can be transformed into an enhanced ML algorithm. The existing algorithm takes an initial input signal and processes/updates it iteratively until a predefined termination criterion is satisfied, at which point the final output is obtained. The specific processing is normally obtained through model-based algorithm design. Instead of expressing the algorithm as a loop,  $L$  iterations of the algorithm can be expressed as a sequence of  $L$  identical processing layers. If a data-driven

training procedure is employed to fine-tune these processing layers, which no longer have to be identical, what ensues is an ML algorithm that is guaranteed to perform better than the original model-based algorithm. This procedure is called *algorithm unrolling* or *deep unfolding*, and it has in recent years been utilized to enhance various multiantenna tasks, including signal detection [37] and beamforming optimization for downlink multiuser MIMO [38].

### A peek into the future

Over the past 25 years, multiantenna techniques have gone from rudimentary designs for beamforming and diversity combining to a mainstream technology that uses massive

**Recent theoretical breakthroughs, including ML-based algorithms, are bound to continue sustaining the progress of the technology.**



**FIGURE 7.** Many conventional model-based algorithms for optimization in multiantenna communication are iterative. Suppose one such algorithm can be expressed as an iterative processing loop that continues until a termination criterion is satisfied. An enhanced ML algorithm can be developed through algorithm unrolling; that is, writing the iteration as  $L$  separate processing layers and fine-tuning these layers through a data-aided training procedure.

spatial multiplexing to multiply the capacity of 5G networks. The MIMO technology is now present in every smartphone and BS that enters the market. Fast and capable signal processing algorithms have enabled this leap forward, and are currently buttressing the emergence of low-power 5G mmWave transceivers where high-resolution hardware components are replaced with digital processing. Recent theoretical breakthroughs, including ML-based algorithms, are bound to continue sustaining the progress of the technology.

Indeed, we expect the multiantenna communication journey to continue. The insatiable growth in data traffic can only be met by deploying ever more antennas and ever more bandwidth. The massive MIMO philosophy prescribes that the number of BS antennas,  $M$ , must scale proportionally to the number of active users,  $K$ . As the complexity of algorithms, such as regularized zero-forcing is proportional to  $MK^2$  [4], a linear scaling in both  $K$  and  $M$  implies a cubic complexity growth. Moreover, once the bandwidth surpasses 1 GHz, the sampling rates approach the clock speed of existing processors, which renders the implementation even more demanding. The compressed sensing algorithms described earlier might be suitable to address these challenges, but there is likely room for many new signal processing advances.

When adding ever more antennas to BSs, practical size and weight constraints might make new deployment principles necessary, beyond the boxes-in-a-tower paradigm. One promising approach is to distribute the antennas over multiple physical locations while retaining the coherent transmission and reception processing, a concept rooted in cell cooperation and network MIMO ideas, as well as in the notion of remote RF units, and whose present embodiment is termed *cell-free massive MIMO* [39]. Apart from stronger beamforming gains, a distributed antenna deployment can provide improved spatial multiplexing capabilities and macroscopic diversity against the shadowing of large objects in the environment. The current trend of shifting baseband computations from BS sites to edge-cloud computers will ease the adoption of this deployment approach.

After two decades of smartphones ruling the wireless ecosystem, other devices, such as extended reality eyeglasses, are predicted to take center stage. New services will surface, with renewed standards for the bit rates, latency, and reliability that users expect wireless networks to deliver. Other performance metrics might arise to dictate future technology development, particularly related to sustainability, environmental impact, and deployment costs, as well as to the digital divide between the digitized and far-from-digitized regions of the world.

Beyond the signal processing advances captured in this article, two emerging research topics build on multiantenna technology. The first is integrated communication and sensing [40], which explores how large-scale antenna arrays can be simultaneously used for accurate radar sensing, localization, and communication. It seems natural that the deployment of massive antenna numbers for communication purposes can be the catalyst for other applications that benefit from wireless measurements. Another related research

direction is that of smart surfaces [41], whose signal reflection properties can be controlled by means of metamaterials with programmable impedance patterns. These reconfigurable intelligent surfaces provide a sort of passive beamforming that is particularly useful to enhance propagation conditions over wireless channels.

## Authors

**Emil Björnson** (emilbjo@kth.se) is a professor of wireless communication at the KTH Royal Institute of Technology, Stockholm, Sweden. He has authored three textbooks on multiple-input multiple-output technology, has received 23,000 citations, and has published a large amount of simulation code. He has received the 2018 and 2022 IEEE Marconi Prize Paper Awards in Wireless Communications, the 2019 EURASIP Early Career Award, the 2019 IEEE Communications Society Fred W. Ellersick Prize, the 2019 IEEE Signal Processing Magazine Best Column Award, the 2020 Pierre-Simon Laplace Early Career Technical Achievement Award, the 2020 CTTC Early Achievement Award, and the 2021 IEEE ComSoc RCC Early Achievement Award. He is an IEEE Fellow.

**Yonina C. Eldar** (yonina.eldar@weizmann.ac.il) is a professor in the Department of Math and Computer Science at the Weizmann Institute of Science, Rehovot, Israel, where she heads the Center for Biomedical Engineering and Signal Processing. She is also a visiting professor at the Massachusetts Institute of Technology and at the Broad Institute, Cambridge, MA 02142 USA, and an adjunct professor at Duke University, Durham, NC 27708 USA, and was a visiting professor at Stanford University, Stanford, CA USA. She is a member of the Israel Academy of Sciences and Humanities and a European Association for Signal Processing Fellow. She has received many awards for excellence in research and teaching and heads the Committee for Promoting Gender Fairness in Higher Education Institutions in Israel. She is an IEEE Fellow.

**Erik G. Larsson** (erik.g.larsson@liu.se) is a professor at Linköping University, Linköping, Sweden. He coauthored the textbook *Fundamentals of Massive MIMO* (Cambridge University Press, 2016). He received, among others, the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015, the IEEE ComSoc Leonard G. Abraham Prize in 2017, the IEEE ComSoc Best Tutorial Paper Award in 2018, and the IEEE ComSoc Fred W. Ellersick Prize in 2019. His interest include wireless communications, statistical signal processing, and networks. He is an IEEE Fellow.

**Angel Lozano** (angel.lozano@upf.edu) received his Ph.D. from Stanford University in 1999, worked for Bell Labs (Lucent Technologies, now Nokia) between 1999 and 2008, and served as an adjunct associate professor at Columbia University between 2005 and 2008. He is a professor at Universitat Pompeu Fabra, Barcelona, Spain. His papers have received several awards, including the 2009 Stephen O. Rice Prize, the 2016 Fred W. Ellersick Prize, and the 2016 Communications Society & Information Theory Society Joint Paper Award. He is also the recipient of a European Research

Council Advanced Grant for the period 2016–2021 and a 2017 Highly Cited Author. He is the coauthor of the textbook *Foundations of MIMO Communication*, published by Cambridge University Press in 2019. He is an IEEE Fellow.

**H. Vincent Poor** (poor@princeton.edu) is the Michael Henry Strater University Professor at Princeton University, Princeton, NJ USA, where his interests include information theory, machine learning, and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He is a member of the U.S. National Academies of Engineering and Sciences, and received the IEEE Alexander Graham Bell Medal in 2017. He is an IEEE Life Fellow.

## References

[1] "World telecommunication development report: Mobile cellular," International Telecommunication Union, Geneva, Switzerland, Tech. Rep. 5, 1999.

[2] H. V. Poor and G. W. Wornell, *Wireless Communications: Signal Processing Perspectives* (Prentice-Hall Signal Processing Series). Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

[3] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[4] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends® Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, Nov. 2017, doi: 10.1561/20000000093.

[5] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[6] S. Anderson, U. Forssen, J. Karlsson, T. Witzschel, P. Fischer, and A. Krug, "Ericsson/Mannesmann GSM field-trials with adaptive antennas," in *Proc. IEEE Colloq. Adv. TDMA Techn. Appl.*, 1996, pp. 1–6, doi: 10.1049/ic:19961235.

[7] H. O. Peterson, H. H. Beverage, and J. B. Moore, "Diversity telephone receiving system of R.C.A. communications, Inc.," in *Proc. Inst. Radio Eng. (IRE)*, 1931, vol. 19, no. 4, pp. 562–584, doi: 10.1109/JRPROC.1931.222363.

[8] A. Wittebne, "Basestation modulation diversity for digital simulcast," in *Proc. 41st IEEE Veh. Technol. Conf. (VTC)*, 1991, pp. 848–853, doi: 10.1109/VETEC.1991.140615.

[9] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998, doi: 10.1109/49.730453.

[10] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998, doi: 10.1109/18.661517.

[11] J. H. Winters, "Optimum combining in digital mobile radio with cochannel interference," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 4, pp. 528–539, Jul. 1984, doi: 10.1109/JSAC.1984.1146095.

[12] J. Salz, "Digital transmission over cross-coupled linear channels," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1147–1159, Jul./Aug. 1985, doi: 10.1002/j.1538-7305.1985.tb00269.x.

[13] J. Winters, "On the capacity of radio communication systems with diversity in a Rayleigh fading environment," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 5, pp. 871–878, Jun. 1987, doi: 10.1109/JSAC.1987.1146600.

[14] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998, doi: 10.1023/A:100888922784.

[15] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. URSI Int. Symp. Signals, Syst., Electron. Conf. Proc. (Cat. No.98EX167)*, Sep. 1998, pp. 295–300, doi: 10.1109/ISSSE.1998.738086.

[16] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999, doi: 10.1002/ett.4460100604.

[17] J. H. Winters, "Optimum combining for indoor radio systems with multiple users," *IEEE Trans. Commun.*, vol. 35, no. 11, pp. 1222–1230, Nov. 1987, doi: 10.1109/TCOM.1987.1096697.

[18] S. C. Swales, M. A. Beach, D. J. Edwards, and J. P. McGeehan, "The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems," *IEEE Trans. Veh. Technol.*, vol. 39, no. 1, pp. 56–67, Feb. 1990, doi: 10.1109/25.54956.

[19] Y. Tsuji and Y. Tada, "Transmit phase control system of synchronization burst for SDMA/TDMA satellite communication system," U.S. Patent 3 995 111, 1976.

[20] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003, doi: 10.1109/TIT.2003.810646.

[21] A. Lozano and N. Jindal, "Transmit diversity vs. spatial multiplexing in modern MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 186–197, Jan. 2010, doi: 10.1109/TWC.2010.01.081381.

[22] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010, doi: 10.1109/TWC.2010.092810.091092.

[23] M. Joham, W. Utschick, and J. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005, doi: 10.1109/TSP.2005.850331.

[24] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006, doi: 10.1109/TIT.2006.880064.

[25] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016, doi: 10.1109/JSTSP.2016.2523924.

[26] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, Jan. 2016, doi: 10.1109/ACCESS.2015.2514261.

[27] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Achievable uplink rates for massive MIMO with coarse quantization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 6488–6492, doi: 10.1109/ICASSP.2017.7953406.

[28] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017, doi: 10.1109/TSP.2017.2706179.

[29] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Mar. 2019, doi: 10.1109/TWC.2019.2894120.

[30] N. Shlezinger and Y. C. Eldar, "Task-based quantization with application to MIMO receivers," *Commun. Inf. Syst.*, vol. 20, no. 2, pp. 131–162, 2020, doi: 10.4310/CIS.2020.v20.n2.a3.

[31] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek, "A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect CSI," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 484–498, Jun. 2018, doi: 10.1109/JSTSP.2018.2813973.

[32] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[33] Z. Gao, L. Dai, S. Han, I. Chih-Lin, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018, doi: 10.1109/MWC.2017.1700147.

[34] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018, doi: 10.1109/MSP.2018.2844952.

[35] J. Yuan, Q. He, M. Matthaiou, T. Q. S. Quek, and S. Jin, "Toward massive connectivity for IoT in mixed-ADC distributed massive MIMO," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1841–1856, Mar. 2020, doi: 10.1109/JIOT.2019.2957281.

[36] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021, doi: 10.1109/MSP.2020.3016905.

[37] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, May 2019, doi: 10.1109/TSP.2019.2899805.

[38] L. Pellaco, M. Bengtsson, and J. Jaldén, "Matrix-inverse-free deep unfolding of the weighted MMSE beamforming algorithm," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 65–81, 2022, doi: 10.1109/OJCOMS.2021.3139858.

[39] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends® Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, Jan. 2021, doi: 10.1561/2000000109.

[40] J. A. Zhang et al., "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1295–1315, Nov. 2021, doi: 10.1109/JSTSP.2021.3113120.

[41] E. Björnson, H. Wymeersch, B. Matthieson, P. Popovski, L. Sanguinetti, and E. de Carvalho, "Reconfigurable intelligent surfaces: A signal processing perspective with wireless applications," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 135–158, Mar. 2022, doi: 10.1109/MSP.2021.3130549.



# Twenty-Five Years of Advances in Beamforming

*From convex and nonconvex optimization to learning techniques*



©SHUTTERSTOCK.COM/TRIFF

**B**eamforming is a signal processing technique to steer, shape, and focus an electromagnetic (EM) wave using an array of sensors toward a desired direction. It has been used in many engineering applications, such as radar, sonar, acoustics, astronomy, seismology, medical imaging, and communications. With the advent of multi-antenna technologies in, say, radar and communication, there has been a great interest in designing beamformers by exploiting convex or nonconvex optimization methods. Recently, machine learning (ML) is also leveraged for obtaining attractive solutions to more complex beamforming scenarios. This article captures the evolution of beamforming in the last 25 years from convex to nonconvex optimization and optimization to learning approaches. It provides a glimpse into these important signal processing algorithms for a variety of transmit–receive architectures, propagation zones, propagation paths, and multidisciplinary applications.

## Introduction

Beamforming is ubiquitous and essential to a multitude of array processing applications, such as radar, sonar, acoustics, astronomy, seismology, ultrasound, and communications [1]. Recent advances in mobile communications, usage of large arrays, high-frequency sensors, near-field signal recovery, and smart radio environments open up interesting and novel signal processing problems in beamforming. These applications are driving the need for higher robustness, flexible deployment, and low complexity in beamforming algorithms and an emphasis on advanced signal processing that should be tailored for emerging application-specific requirements.

Early experiments with beamforming could be traced back to Guglielmo Marconi, who used a circular array with four antennas to improve the gain of trans-Atlantic Morse code transmission in 1901 [2]. A similar early demonstration of gains provided by a phased array to direct radio waves was in 1905 by Karl Ferdinand Braun, who shared the Nobel Prize in Physics with Marconi in 1909 for their contributions to wireless telegraphy [3]. In the 1940s, antenna diversity as a technique to overcome fading was developed for phased array



radars and radio astronomy [4]. By the 1950s–1960s, with the development of phased arrays for sonars, the steering of signals with antenna arrays was no longer restricted to EM waves [5].

Adaptive beamforming [6], [7] emerged in the late 1960s, wherein a processor at the antenna back end updates and compensates the array weights. In particular, Bernard Widrow introduced the least mean square algorithm to update the weights at every iteration by estimating the gradient of the mean-square error (MSE) between the desired and received signals [7]. Subsequently, J. Capon proposed selecting the weight vectors, or beamformers, to minimize the array output power. The Capon beamformer is subjected to the linear constraint that the signal of interest (SoI) does not suffer from any distortion, e.g., direction mismatch, signal fading, local scattering, etc. [6], [8]. Hence, this technique is also usually referred to as the *minimum variance (MV) distortionless response (MVDR)* beamforming.

The performance of the Capon beamformer strongly depends on the knowledge of the SoI, which is imprecise in practice because of the differences between the assumed and true array responses. The beamforming performance is usually measured by the signal-to-interference-plus-noise ratio (SINR). This may severely degrade even in the presence of small errors or mismatches in the steering vector [8]. In the past, numerous approaches were proposed to improve the robustness against errors/mismatches in the look direction [9], [10]; array manifold [11]; and local scattering [12]. These techniques were limited to only the specific mismatch they treat [13], thereby giving rise to early generalization of robust beamforming approaches, such as the sample matrix inversion (SMI) algorithm [14], robust Capon beamforming [15], eigenspace-based beamformer [16], worst case performance optimization [13], and general-rank beamformer [17], [18].

In the late 1990s and early 2000s, significant progress was made toward robust beamformer design by exploiting convex optimization [19]. These methods typically consider minimizing the effect of mismatches in the array-steering vectors and the look direction based on the worst case performance optimization [13], [15], [20]. Here, the optimization problem is cast as a second-order cone (SOC) program and efficiently solved by interior-point methods. It may also be desirable to design a robust MVDR beamformer by including the uncertainty in the array manifold via an ellipsoid or a sphere model for a particular look direction [15], [20].

During the late 2000s, certain applications of beamforming that have nonconvex objective functions or constraints gained salience. These included robust adaptive beamforming with additional constraints related to the positive semidefiniteness (PSD) of the signal covariance matrix [18], norm of the steering vectors [21], [22], [23], [24], and stochastic distortionless response [25], [26]; multicast transmit beamforming [27]; and hybrid (analog/digital) beamforming [28]. The solution to these nonconvex optimization problems usually requires

recasting the problem into a tractable form through the use of, for example, semidefinite relaxation (SDR), compressed sensing (CS) [28], and alternating optimization [19]. Solving for beamforming weights is generally considered as a continuous optimization problem. However, there is a smaller body of literature [29], [30] on discrete/combinatorial techniques. Here, the beamforming weights are selected from a set of exponentials with discretized angles.

In the last decade, with the advent of new cellular communications technologies, beamforming has been extensively investigated for multiantenna systems [28]. The 4G networks (2009 to present) operating at 2.2–4.9 GHz use up to 32 antennas in a multiple-input, multiple-output (MIMO) configuration. The 5G systems (2019 to present) offer support for larger antenna arrays as well as communication at frequencies above 24 GHz. Support for larger arrays is essential in millimeter-wave (mm-wave) systems to overcome shrinking antenna sizes [31]. To reduce the hardware, cost, power, and area in mm-wave massive MIMO systems, hybrid (analog and digital)

beamforming has been introduced [28], [31]. Unlike a conventional digital beamformer employing a single radio-frequency (RF) chain dedicated to each antenna, hybrid approaches employ a few (large) RF chains (analog components, e.g., phase shifters) to reduce the hardware cost. The hybrid beamformer design is also nonconvex because of the unit-modulus constraint owing to the use of phase shifters in the analog beamformers. This problem has been addressed through techniques such as sparse matrix reconstruction via CS [28], optimization over Riemannian manifolds [32], phase extraction [33], and Gram–Schmidt orthogonalization [34].

Very recently, data-driven methods, such as ML, have been leveraged to obtain beamformers. ML is a subset of artificial intelligence (AI) that allows neural networks (NNs) to learn directly from precedents, data, and examples without being explicitly programmed. Many beamformers involve nonlinear operations. In this context, NNs are particularly attractive because they successfully approximate nonlinear functions or predict the class of a function that is divided by a nonlinear decision boundary. Compared to the model-based techniques, ML has lower posttraining computational complexity, expedited design procedure, and robustness against imperfections/mismatches [35], [36], [37]. The ML-based hybrid beamforming is also envisioned as a key to realize massive MIMO architectures beyond 5G communications [38], such as 6G systems operating at terahertz (THz) bands. This is largely because ML is helpful in processing copious amounts of antenna array data generated by massive MIMO systems employed at higher frequencies.

To shed light on the evolution of beamforming techniques, this article presents an overview of the aforementioned approaches while focusing on major breakthroughs during the last 25 years. Specifically, the article aims at 1) highlighting the two significant leaps in this research, i.e.,

**Beamforming has been used in many engineering applications, such as radar, sonar, acoustics, astronomy, seismology, medical imaging, and communications.**

convex to nonconvex optimization, and optimization- to learning-based beamforming; 2) depicting in detail the analytical background and the relevance of signal processing tools for beamforming; and 3) introducing the major challenges and emerging signal processing applications of beamforming. Figure 1 summarizes some important classes of beamformers discussed in this article.

### Notation

Throughout this article, uppercase and lowercase bold letters denote matrices and vectors, respectively. Also,  $(\cdot)^T$  and  $(\cdot)^H$  denote the transpose and conjugate transpose operations, respectively. For a matrix  $\mathbf{A} \in \mathbb{C}^{M \times N}$  and a vector  $\mathbf{a} \in \mathbb{C}^N$ ,  $[\mathbf{A}]_{ij}$ ,  $[\mathbf{A}]_k$ ,  $\Re\{\mathbf{A}\}$  and  $\Im\{\mathbf{A}\}$ , and  $a_i$  correspond to the  $(i, j)$  th entry,  $k$ th column, real and imaginary parts of  $\mathbf{A}$ , and  $i$ th entry of  $\mathbf{a}$ , respectively, while  $\mathbf{A}^\dagger$  denotes the Moore–Penrose pseudo-inverse of  $\mathbf{A}$ , and  $\mathbf{I}$  is the identity matrix of proper size.  $\|\mathbf{a}\|_2 = (\sum_{i=1}^N |a_i|^2)^{1/2}$  and  $\|\mathbf{A}\|_F = (\sum_{i=1}^M \sum_{j=1}^N |[\mathbf{A}]_{ij}|^2)^{1/2}$  denote the  $l_2$  norm and Frobenius norm, respectively.

### Convex optimization for beamforming

Convex optimization recasts originally difficult-to-design beamformers as computationally attractive problems that yield exact or approximate solutions through algorithms, such as interior-point methods. Its applications have traditionally advanced from the simple exact Capon approach to more complex transmit, multicast, network, and distributed beamformers; see, e.g., [19] and the references therein for details. In the following, we summarize the techniques that yield exact solutions. The

approximate solutions are considered under nonconvex beamformers in the sequel.

### Capon beamformer

Consider an antenna array with  $N$  elements. Define  $\mathbf{a}(\theta) \in \mathbb{C}^N$  as the array response to a plane-wave narrowband SoI  $s(t_i)$ ,  $i = 1, \dots, T$ , where  $T$  is the number of snapshots arriving from the direction of arrival (DoA) angle  $\theta$ . In particular, the steering vector  $\mathbf{a}(\theta)$  is

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N}} \left[ 1, e^{-j2\pi \frac{d}{\lambda} \sin\theta}, \dots, e^{-j2\pi \frac{(N-1)d}{\lambda} \sin\theta} \right]^T \quad (1)$$

where  $d$  is the element spacing, and  $\lambda$  is the wavelength. Then, the  $N \times 1$  antenna array output is

$$\mathbf{y}(t_i) = \mathbf{a}(\theta)s(t_i) + \mathbf{e}(t_i) \quad (2)$$

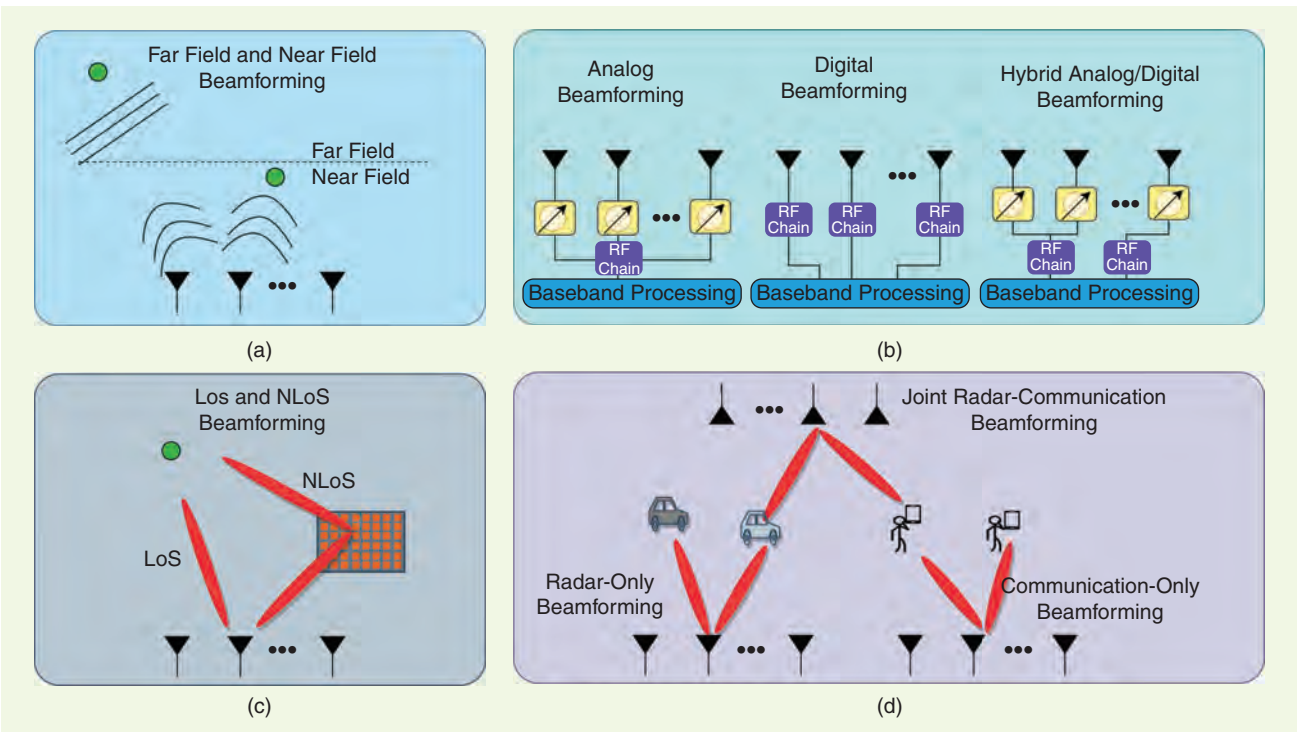
where  $\mathbf{e}(t_i) \in \mathbb{C}^N$  denotes the temporarily and spatially white Gaussian noise vector with variance  $\sigma^2$ .

The received signals are multiplied by the beamforming weights, i.e.,  $w_1, \dots, w_N \in \mathbb{C}$ . Therefore, the combined beamformer output becomes

$$y_o(t_i) = \mathbf{w}^H \mathbf{y}(t_i) = \mathbf{w}^H \mathbf{a}(\theta)s(t_i) + \mathbf{w}^H \mathbf{e}(t_i) \quad (3)$$

where  $\mathbf{w} = [w_1, \dots, w_N]^T$  includes the beamformer weights. To recover the signal  $s(t_i)$ , the beamformer weights are optimized via

$$\underset{\mathbf{w}}{\text{minimize}} \mathbf{w}^H \mathbf{R}_y \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{a}(\theta) = 1 \quad (4)$$



**FIGURE 1.** The major classes of beamforming methods by (a) transmission range: far and near fields; (b) transceiver architectures: analog, digital, and hybrid beamforming; (c) paths: LoS and NLoS beamforming, wherein the NLoS path is controlled via joint active (transmitter) and passive (intelligent reflecting surface) devices; (d) applications: radar, communications, and joint radar-communications. LoS: line-of-sight; NLoS: non-line-of-sight.

where  $\mathbf{R}_y = (1/T)\sum_{t=1}^T \mathbf{y}(t_i)\mathbf{y}^H(t_i)$  is the sample covariance matrix of the array output. The optimal solution for (4) yields the Capon beamformer [6]:

$$\mathbf{w}_{\text{opt}} = (\mathbf{a}^H(\theta)\mathbf{R}_y^{-1}\mathbf{a}(\theta))^{-1}\mathbf{R}_y^{-1}\mathbf{a}(\theta). \quad (5)$$

This beamformer requires the knowledge of  $\mathbf{a}(\theta)$  and  $\mathbf{R}_y$ . Therefore, its performance depends on the accuracy of the steering vector constructed from the estimate of  $\theta$  as well as the sample covariance matrix  $\mathbf{R}_y$ .

To stabilize the main beam response in the presence of a pointing error [9], additional constraints are added to the optimization problem as

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H\mathbf{R}_y\mathbf{w} \quad \text{subject to} \quad \mathbf{C}^H\mathbf{w} = \mathbf{u} \quad (6)$$

where  $L$  many constraints are represented by  $\mathbf{C} \in \mathbb{C}^{L \times N}$  and  $\mathbf{u} \in \mathbb{C}^L$ . For example, if it is desired to maximize the beam pattern at  $30^\circ$  and place a null at  $40^\circ$ , then  $\mathbf{C} = [\mathbf{a}(30^\circ), \mathbf{a}(40^\circ)]^T$  and  $\mathbf{u} = [1, 0]^T$ . The solution to this constrained problem is  $\mathbf{w}_C = \mathbf{R}_y^{-1}\mathbf{C}(\mathbf{C}^H\mathbf{R}_y^{-1}\mathbf{C})^{-1}\mathbf{u}$  [10].

### Loaded SMI beamformer

Even in the ideal case, wherein the SoI direction  $\theta$  is accurately known, beamforming performance may significantly deteriorate because of a small training sample size  $T$ . This is mitigated by adding a regularization term  $\gamma$  to the objective function in (4) leading, to loaded SMI (LSMI) beamforming [14]:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H\mathbf{R}_y\mathbf{w} + \gamma\|\mathbf{w}\|_2 \quad \text{subject to} \quad \mathbf{w}^H\mathbf{a}(\theta) = 1. \quad (7)$$

Its solution is  $\mathbf{w}_{\text{LSMI}} = \mathbf{R}_{\text{LSMI}}^{-1}\mathbf{a}(\theta)$ , where  $\mathbf{R}_{\text{LSMI}} = \mathbf{R}_y + \gamma\mathbf{I}_N$ .

### Robust Capon beamformer

The exact knowledge of the SoI direction  $\theta$  required by the Capon beamformer is not available in practice. This is addressed by robust beamforming, which provides tolerance against the inaccuracies in the estimated SoI direction and the corresponding steering vector. A robust variant of Capon beamforming was introduced in [15], wherein the convex optimization problem is

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H\mathbf{R}_y^{-1}\mathbf{w}, \quad \text{subject to} \quad \|\mathbf{w} - \bar{\mathbf{a}}\|_2 \leq \epsilon \quad (8)$$

where  $\bar{\mathbf{a}} = \mathbf{a}(\theta + \Delta_\theta)$  is the inaccurate steering vector for the mismatched direction  $\theta + \Delta_\theta$ .

### Beamforming with worst case performance optimization

A more general approach is considered in [13] by taking into account the distortions in the steering vector as  $\tilde{\mathbf{a}} = \mathbf{a}(\theta) + \Delta_{\mathbf{a}}$ , where  $\Delta_{\mathbf{a}} \in \mathbb{C}^N$  represents the steering vector distortions. As a result, the optimization problem is based on the worst case beamforming performance. Relying on the bounded Euclidean norm as  $\|\Delta_{\mathbf{a}}\|_2 \leq \epsilon$  corresponding to the case of spherical uncertainty [13], the following convex problem is formulated:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H\mathbf{R}_y\mathbf{w}, \quad \text{subject to} \quad |\mathbf{w}^H\tilde{\mathbf{a}}| \geq 1, \|\Delta_{\mathbf{a}}\|_2 \leq \epsilon \quad (9)$$

for which the LSMI-based solutions may also be obtained [8], [19]. A similar approach, called robust MV beamforming, introduced in [20], is based on ellipsoidal uncertainty. Both spherical [e.g.,  $\|\tilde{\mathbf{a}} - \mathbf{a}(\theta)\|_2 \leq \epsilon$  in (9)] and ellipsoidal [e.g.,  $(\tilde{\mathbf{a}} - \mathbf{a})^H\mathbf{V}(\tilde{\mathbf{a}} - \mathbf{a}) \leq \tilde{\epsilon}$ , where  $\mathbf{V} \in \mathbb{C}^{N \times N}$  is a PSD matrix] models are used to ensure robust solutions. The latter may naturally lead to a more accurate uncertainty description [20] than that with spherical models [20], [39] if more information than just the same uncertainty radius in all mismatch dimensions is available, and an uncertainty ball is replaced by an uncertainty ellipsoid. Assuming the availability of more information about the mismatch is, however, somewhat contradictory to the notion of robustness.

The structure of the beamformer design problem also depends on the noise model. Some beamforming techniques are based on the MV criterion mentioned earlier. However, this criterion is statistically optimal only when the SoI, interference, and noise are Gaussian. The non-Gaussian case leads to a nonconvex problem as

$$\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{Y}^H\mathbf{w}\|_p^p, \quad \text{subject to} \quad \mathbf{a}^H(\theta)\mathbf{w} = 1 \quad (10)$$

where  $\mathbf{Y} = [\mathbf{y}(t_1), \dots, \mathbf{y}(t_T)] \in \mathbb{C}^{N \times T}$ , and  $\|\mathbf{y}(t_i)\|_p^p = (\sum_{n=1}^N y_n(t_i))^{1/p}$  denotes the  $\ell_p$  norm for  $p \geq 1$ . Note that (10) reduces to Capon beamforming of (4) for  $p = 2$ . The solution for (10) is achieved via iterative reweighted MVDR techniques [40]. In addition to generalizing the noise model, a specific choice of priors over the distribution of the beamforming weights may also be used in, say, sparsity-driven beamforming [41].

### Beamforming for a general-rank source

In practice, the source signal is incoherently scattered such that the point-source assumption may not hold [17], and the array covariance matrix is no longer rank-1. Therefore, instead of a constraint on a single steering vector, the SoI covariance matrix is used. The corresponding MVDR-type optimization problem is

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H\mathbf{R}_y\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H\mathbf{R}_s\mathbf{w} = 1 \quad (11)$$

where  $\mathbf{R}_s$  is the SoI covariance matrix [18]. The optimal solution to (11) is  $\mathbf{w}_{\text{GR}} = \mathcal{P}[\mathbf{R}_y^{-1}\mathbf{R}_s]$ , where  $\mathcal{P}[\cdot]$  is the principal eigenvector operator.

### Nonconvex beamformer design

Nonconvex beamformers [21], [22], [23], [24], [25], [26], [27], [28], [42] tackle the design problem by recasting or relaxing it into tractable convex forms. This may be achieved by dropping the nonconvex constraints or decoupling the beamforming design into multiple convex subproblems.

### PSD-constrained beamforming

The general-rank beamforming solution in (11) requires the knowledge of signal covariance matrix  $\mathbf{R}_s$ , which is not always available [17], [18]. The actual signal correlation matrix is, then, not guaranteed to be PSD and usually modeled as  $\tilde{\mathbf{R}}_s = \mathbf{R}_s + \Delta_s$ . To guarantee the PSD-ness of  $\tilde{\mathbf{R}}_s$  decompose

it as  $\tilde{\mathbf{R}}_s = \mathbf{Q}\mathbf{Q}^H$  with the mismatch parameter  $\Delta_{\mathbf{Q}}$  bounded as  $\|\Delta_{\mathbf{Q}}\|_2 \leq \varepsilon_{\mathbf{Q}}$ . The resulting nonconvex problem is

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} \quad \max_{\|\Delta_y\|_2 \leq \varepsilon_y} \mathbf{w}^H (\mathbf{R}_y + \Delta_y) \mathbf{w} \\ & \text{subject to} \quad \min_{\|\Delta_{\mathbf{Q}}\|_2 \leq \varepsilon_{\mathbf{Q}}} \mathbf{w}^H (\mathbf{Q} + \Delta_{\mathbf{Q}})^H (\mathbf{Q} + \Delta_{\mathbf{Q}}) \mathbf{w} \geq 1 \end{aligned} \quad (12)$$

where  $\Delta_y$ , with  $\|\Delta_y\|_2 \leq \varepsilon_y$ , represents the mismatch in  $\mathbf{R}_y$ . The efficient solution to the nonconvex problem in (12) is obtained via the polynomial-time difference-of-convex functions algorithm [18].

### Norm-constrained beamforming based on steering vector estimation

Apart from the uncertainty constraint (8) of the robust Capon beamformer [15], [21] considers an additional norm constraint for beamformer weights in a more general setting as

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H \hat{\mathbf{R}}^{-1} \mathbf{w} \quad \text{subject to} \quad \|\mathbf{a} - \hat{\mathbf{a}}\|_2 \leq \varepsilon_a, \|\mathbf{a}\|_2^2 = N \quad (13)$$

which is identical to (8) and convex without the constraint  $\|\mathbf{a}\|_2^2 = N$ . The nonconvex problem in (13) is called *doubly constrained* robust Capon beamforming [21]. It is iteratively solved by interpreting the optimization as a covariance fitting problem. Thus, a robust beamformer is obtained by robustly estimating the array-steering vector. This formulation was further improved in [23], where the difference between the actual and presumed steering vectors is iteratively estimated without making any assumption on either the norm of the mismatch vector or its probability distribution.

The solution developed in [23] has led to a formulation in [24] of a new constraint, which guarantees that an estimate of the source steering vector does not converge to any steering vectors of interference signals as well as their linear combinations. This steering vector estimation problem is

$$\underset{\hat{\mathbf{a}}}{\text{minimize}} \quad \hat{\mathbf{a}}^H \hat{\mathbf{R}}^{-1} \hat{\mathbf{a}} \quad \text{subject to} \quad \|\hat{\mathbf{a}}\|_2^2 = N, \hat{\mathbf{a}}^H \tilde{\mathbf{C}} \hat{\mathbf{a}} \leq \Delta_0 \quad (14)$$

where the last constraint is new;  $\hat{\mathbf{a}} \in \mathbb{C}^N$  is the estimate of  $\mathbf{a}$ ;  $\tilde{\mathbf{C}} = \int_{\tilde{\Theta}} \mathbf{a}(\theta) \mathbf{a}^H(\theta) d\theta$   $\tilde{\Theta}$  is the complement of the angular sector  $\Theta = [\theta_{\min}, \theta_{\max}]$  where the desired signal is located; and  $\Delta_0$  is a uniquely selected value for a given  $\Theta$ , that is,  $\Delta_0 \triangleq \max_{\theta \in \Theta} \mathbf{a}^H(\theta) \tilde{\mathbf{C}} \mathbf{a}(\theta)$ , representing the boundary line to distinguish approximately whether or not the direction of  $\mathbf{a}$  is in the actual signal angular sector  $\Theta$ .

To account for gain perturbations in the steering vector, [22] added the double-sided norm constraint to the problem (14) as

$$\begin{aligned} & \underset{\hat{\mathbf{a}}}{\text{minimize}} \quad \hat{\mathbf{a}}^H \hat{\mathbf{R}}^{-1} \hat{\mathbf{a}} \quad \text{subject to} \quad \hat{\mathbf{a}}^H \hat{\mathbf{C}} \hat{\mathbf{a}} \geq \Delta_1, \\ & N(1 - \eta_1) \leq \|\hat{\mathbf{a}}\|_2^2 \leq N(1 + \eta_2), \|\mathbf{V}^H (\hat{\mathbf{a}} - \mathbf{a}_0)\|_2^2 \leq \varepsilon_u \end{aligned} \quad (15)$$

where  $\mathbf{a}_0 = \mathbf{a}(\theta_0)$ ,  $\theta_0 = (\theta_{\max} + \theta_{\min})/2$  is the middle value of the region  $\Theta$ ;  $\mathbf{V} \in \mathbb{C}^{N \times N}$  denotes a generalized similarity constraint together with  $\mathbf{a}_0$  and  $\varepsilon_u$ ;  $\mathbf{C} = \int_{\Theta} \mathbf{a}(\theta) \mathbf{a}^H(\theta) d\theta$ ; and  $\Delta_1$ ,

$\eta_1$ , and  $\eta_2$  are selected values. In (15), the generalized similarity condition implies that imperfect knowledge of the desired steering vector  $\hat{\mathbf{a}}$  is described as in a convex set (in particular, an ellipsoidal set when  $\mathbf{V}$  is of full row rank).

All of these problems are nonconvex but can be often exactly solved through SDR, iterative SOC program, quadratic matrix inequality, and bilinear matrix inequality approaches.

### Chance-constrained beamforming

In many applications, it is more natural that the distortionless constraint is satisfied with a certain probability. This leads to the chance-constrained robust adaptive beamforming problem [25]:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^H \hat{\mathbf{R}} \mathbf{w} \quad \text{subject to} \quad \Pr\{|\mathbf{w}^H \hat{\mathbf{a}}| \geq 1\} \geq p \quad (16)$$

where  $p$  is a certain preselected probability value, and  $\Pr\{\cdot\}$  stands for the probability operator. This problem corresponds to minimizing the beamformer output power subject to the stochastic constraint that the probability of the signal distortionless response is greater than or equal to some selected value  $p$ . The constraint may also be viewed as a nonoutage probability constraint where the outage probability  $p_{\text{out}} = 1 - p$  is defined as that of violating the inequality  $|\mathbf{w}^H \hat{\mathbf{a}}| \geq 1$  for a random  $\hat{\mathbf{a}}$  that consists of a presumptive steering vector and the mismatch that is assumed to be random. Problem (16) is nonconvex and specified by the mismatch distribution. The solutions of (16) for the case of Gaussian-distributed mismatch of the signal steering vector and for the worst case distribution are well approximated by the corresponding SOC programs [25].

In [26], a chance-constrained nonconvex formulation of robust adaptive beamforming considers a more practical scenario, wherein both interference-plus-noise covariance matrix  $\mathbf{R}_{i+n}$  and the true steering vector  $\mathbf{a}$  are not precisely known. It also shows the chance-constrained beamformer to have a higher output SINR than other convex (LSMI) and nonconvex (worst case optimization) beamformers [26]. Considering both  $\mathbf{R}_{i+n}$  and  $\mathbf{a}$  as random variables, the robust adaptive beamforming becomes

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} \quad \max_{G_1 \in \mathcal{S}_1} E_{G_1}\{\mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w}\} \\ & \text{subject to} \quad \min_{G_2 \in \mathcal{S}_2} E_{G_2}\{\mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}\} \geq 1 \end{aligned} \quad (17)$$

where  $E_{G_1}\{\cdot\}$  ( $E_{G_2}\{\cdot\}$ ) denotes the statistical expectation under the distribution  $G_1$  ( $G_2$ ), and  $\mathcal{S}_1$  ( $\mathcal{S}_2$ ) is a set of distributions  $G_1$  ( $G_2$ ) for random matrix  $\mathbf{R}_{i+n}$  (random vector  $\mathbf{a}$ ) as, respectively,

$$\mathcal{S}_1 = \left\{ G_1 \in \mathcal{M}_1 \left| \begin{array}{l} \Pr_{G_1}\{\mathbf{R}_{i+n} \in \mathcal{Z}_1\} = 1 \\ E_{G_1}\{\mathbf{R}_{i+n}\} \geq \mathbf{0} \\ \|E_{G_1}\{\mathbf{R}_{i+n}\} - \mathbf{S}_0\|_{\mathcal{F}} \leq \rho_1 \end{array} \right. \right\} \quad (18)$$

and

$$\mathcal{S}_2 = \left\{ G_2 \in \mathcal{M}_2 \left| \begin{array}{l} \Pr_{G_2}\{\mathbf{a} \in \mathcal{Z}_2\} = 1 \\ E_{G_2}\{\mathbf{a}\} = \mathbf{a}_0 \\ E_{G_2}\{\mathbf{a} \mathbf{a}^H\} = \Sigma + \mathbf{a}_0 \mathbf{a}_0^H \end{array} \right. \right\} \quad (19)$$

where  $\mathcal{M}_1$  and  $\mathcal{M}_2$  are sets of all probability measures;  $\mathcal{Z}_1$  and  $\mathcal{Z}_2$  are Borel sets;  $\mathbf{S}_0$  is the empirical mean of  $\mathbf{R}_{i+n}$ , that is, the sample covariance matrix  $\mathbf{R}_y$ ; and  $\Pr_{G_1}\{\cdot\}$  is the probability of an event under the distribution  $G_1$ . Assume the mean  $\mathbf{a}_0$  and covariance matrix  $\mathbf{\Sigma} \succ \mathbf{0}$  of random vector  $\mathbf{a}$  under the true distribution  $\tilde{G}_2$  are known. Then, the set  $\mathcal{S}_2$  includes all probability distributions on  $\mathcal{Z}_2$  that have the same first- and second-order moments as  $\tilde{G}_2$ . This problem is called *distributionally robust beamforming* because it considers distributional uncertainty in both the steering vector and  $\mathbf{R}_{i+n}$ .

### Multicast transmit beamforming

In wireless communications, multicast beamforming is used for broadcasting data streams  $s(t_i)$  toward multiple radio receivers. Consider a transmitter with an  $N$ -element antenna array that aims to deliver a signal to  $U$  single-antenna users. Denote the wireless channel between the transmitter and the  $u$ th receiver by  $\mathbf{h}_u \in \mathbb{C}^N$ . Then, for the beamformed transmitted signal  $\mathbf{x}(t_i) = \mathbf{w}\mathbf{s}(t_i)$ , the received signal at the  $u$ th user is  $y_u(t_i) = \mathbf{h}_u^H \mathbf{x}(t_i) + e_u(t_i)$ , where  $e_u(t_i)$  is the noise signal with variance  $\sigma_u^2$ . Then, the multicast beamforming problem is [27]

$$\underset{\mathbf{w}}{\text{minimize}} \|\mathbf{w}\|_2 \quad \text{subject to} \quad |\mathbf{w}^H \tilde{\mathbf{h}}_u| \geq 1, \quad u \in \{1, \dots, U\} \quad (20)$$

where  $\tilde{\mathbf{h}}_u = \mathbf{h}_u / \sqrt{\rho_{\min,u} \sigma_u^2}$  is the normalized channel vector with the minimum received signal-to-noise ratio (SNR)  $\rho_{\min,u}$

and the noise variance  $\sigma_u^2$  for the  $u$ th receiver. The optimization in (20) is a quadratically constrained quadratic programming problem with nonconvex constraints. A rigorous solution is based on reformulating the problem using SDR. To this end, define an  $N \times N$  rank-one matrix  $\mathbf{M} = \mathbf{w}\mathbf{w}^H$ . Then, the rank constraint is removed to recast the problem in a convex form as

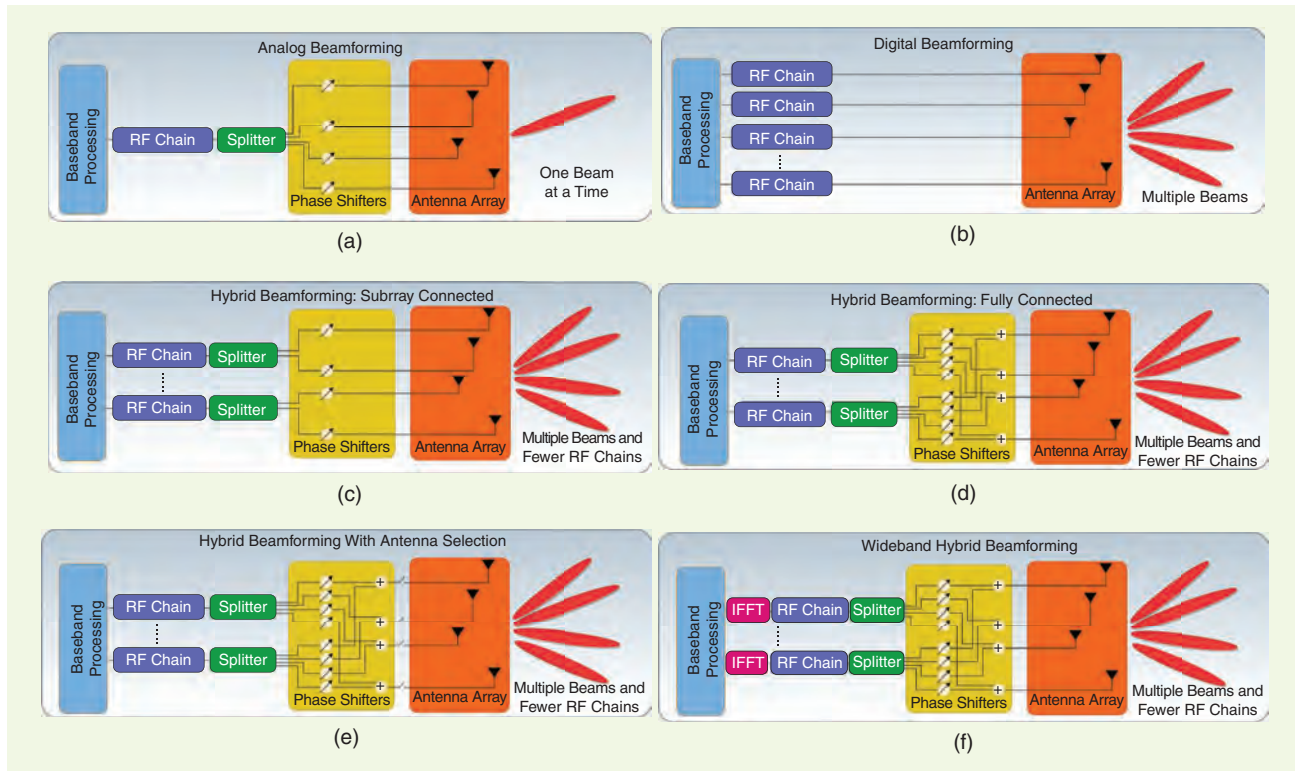
$$\underset{\mathbf{M}}{\text{minimize}} \text{trace}\{\mathbf{M}\} \quad \text{subject to} \quad \text{trace}\{\mathbf{M}\mathbf{D}_u\} \geq 1, \quad \mathbf{M} \succeq \mathbf{0} \quad (21)$$

where  $\mathbf{D}_u = \tilde{\mathbf{h}}_u \tilde{\mathbf{h}}_u^H$ , and the beamformer weight is obtained via eigenvalue decomposition of  $\mathbf{M}$ . A more accurate solution to (20) is obtained by rewriting  $\mathbf{M} = \mathbf{w}_1 \mathbf{w}_2^H$  and then alternately solving for  $\mathbf{w}_1$  and  $\mathbf{w}_2$  using an iterative procedure until convergence [30].

### Hybrid analog/digital beamforming

Compared to analog- and digital-only beamformers, hybrid analog/digital beamforming architecture may have a lower hardware cost while also providing satisfactory spectral efficiency (SE) and multiple beams (Figure 2). In fact, for massive antenna array processing applications, such as 5G communications, hybrid beamforming has emerged as the preferred means to realize large arrays with only a moderate increase in baseband signal processing [31], [33].

Consider a hybrid beamforming scenario, wherein the transmitter employs  $N$  antennas and  $N_{\text{RF}}$  RF chains to send  $N_s$



**FIGURE 2.** The transmitter architectures for (a) analog, (b) digital, and (c)–(f) hybrid beamforming. Analog beamforming generates only one beam because it employs a single RF chain. On the other hand, multiple beams are obtained via digital beamformers but at the cost of multiple RF chains. It is possible to generate multiple beams with fewer RF chains in the hybrid approach through configurations such as (c) subarray connected, (d) fully connected, (e) sparse antenna-selective, and (f) wideband architectures. IFFT: inverse fast Fourier transform.

data streams. Denote the analog and digital beamformers by matrices  $\mathbf{F}_{\text{RF}} \in \mathbb{C}^{N \times N_{\text{RF}}}$  and  $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{N_{\text{RF}} \times N_s}$ , respectively. Here, each element of  $\mathbf{F}_{\text{RF}}$  has a constant modulus because they are realized by phase shifters, i.e.,  $[\mathbf{F}_{\text{RF}}]_{i,j} = 1/\sqrt{N}$  for  $i = 1, \dots, N$ ,  $j = 1, \dots, N_{\text{RF}}$ . The transmitted signal is  $\mathbf{x} = \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{s}$ . The goal is to maximize mutual information

$$\mathcal{I}(\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}) = \log_2 \det \left( \mathbf{I}_{N_s} + \frac{\kappa}{N_s \sigma_n^2} \mathbf{H} \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}} \mathbf{F}_{\text{BB}}^H \mathbf{F}_{\text{RF}}^H \mathbf{H}^H \right)$$

where  $\mathbf{H} \in \mathbb{C}^{N \times N_{\text{R}}}$  is the wireless channel matrix,  $N_{\text{R}}$  is the number of antennas at the receiver,  $\kappa$  is the average received power, and  $\sigma_n^2$  is the noise power [28]. The hybrid beamforming problem is

$$\begin{aligned} & \underset{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}}{\text{maximize}} \quad \mathcal{I}(\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}) \\ & \text{subject to} \quad \|\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}\|_{\mathcal{F}} = N_s, |[\mathbf{F}_{\text{RF}}]_{i,j}| = \frac{1}{\sqrt{N}} \end{aligned} \quad (22)$$

which is nonconvex because of the constant-modulus constraint. The product  $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}$  also makes this problem nonlinear. Recast (22) to an equivalent form by minimizing the Euclidean cost between the hybrid beamformer  $\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}$  and the unconstrained baseband-only beamformer  $\mathbf{F}_{\text{C}} \in \mathbb{C}^{N \times N_s}$  as

$$\begin{aligned} & \underset{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}}{\text{minimize}} \quad \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}} - \mathbf{F}_{\text{C}}\|_{\mathcal{F}} \\ & \text{subject to} \quad \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_{\mathcal{F}} = N_s, |[\mathbf{F}_{\text{RF}}]_{i,j}| = \frac{1}{\sqrt{N}} \end{aligned} \quad (23)$$

where  $\mathbf{F}_{\text{C}}$  is obtained from singular value decomposition of the channel matrix  $\mathbf{H}$  [31]. In the wideband scenario, subcarrier-dependent (SD) digital beamformers are used, and the resulting signal is transformed to the time domain via the inverse fast Fourier transform (Figure 2). Then, subcarrier-independent analog beamformers are employed for all subcarriers because the direction of the generated beam does not change significantly with respect to subcarriers in the mm-wave band [31], [43]. The hybrid beamforming problem for a wideband system with  $M$  subcarriers is

$$\begin{aligned} & \underset{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}[m]}{\text{minimize}} \quad \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}[m] - \mathbf{F}_{\text{C}}[m]\|_{\mathcal{F}} \\ & \text{subject to} \quad \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}[m]\|_{\mathcal{F}} = MN_s, |[\mathbf{F}_{\text{RF}}]_{i,j}| = \frac{1}{\sqrt{N}} \end{aligned} \quad (24)$$

where  $\mathbf{F}_{\text{BB}}[m]$  is the SD digital beamformer that corresponds to the  $m$ th subcarrier,  $m \in \mathcal{M} = \{1, \dots, M\}$ .

For the nonconvex hybrid beamforming formulated in (23), the traditional route is to alternately optimize each ( $\mathbf{F}_{\text{RF}}$  and  $\mathbf{F}_{\text{BB}}$ ) beamformer iteratively while keeping the other one fixed [28], [32], [33]. This has been shown to provide satisfactory SE performance, often close to that of digital-only beamformers, i.e.,  $\mathbf{F}_{\text{C}}$  [28], [32]. During these alternations, while estimation of digital beamformer  $\mathbf{F}_{\text{BB}}$  is straightforward as  $\mathbf{F}_{\text{BB}} = \mathbf{F}_{\text{RF}}^\dagger \mathbf{F}_{\text{C}}$ , the analog beamformer  $\mathbf{F}_{\text{RF}}$  is difficult to obtain. Often  $\mathbf{F}_{\text{RF}}$  is obtained in terms of the steering vectors via CS-based techniques, e.g., orthogo-

nal matching pursuit (OMP). Here, a dictionary of possible steering vectors or atoms is employed, and the beamformers are iteratively selected from these atoms based on the similarity between the dictionary and the measurements (i.e., channel data) [28]. In manifold optimization (MO)-based approaches [32], the search space of  $\mathbf{F}_{\text{RF}}$  is regarded as a Riemannian submanifold of  $\mathbb{C}^N$  with a complex circle manifold to account for the constant-modulus constraint. Then, the analog and digital beamformers are alternately optimized. This method aims to solve the unconstrained optimization problem  $\min_{\mathbf{x}} f(\mathbf{x})$ ,  $\mathbf{x} \in \mathbb{C}^n$  where  $f(\mathbf{x})$  is the cost function, and vector  $\mathbf{x} = \text{vec}(\mathbf{F}_{\text{RF}})$ . To ensure global convergence, the cost function is defined over the Riemannian manifold  $\mathcal{M} = \{\mathbf{x} \in \mathbb{C}^N | x_n^* x_n = 1, n = 1, \dots, N\}$ . Then,  $\mathbf{x}$  is iteratively computed and the solution becomes  $\mathbf{x}_{k+1} = \text{Retr}_{\mathbf{x}_k}(-\alpha_k \text{grad} f(\mathbf{x}_k))$ , where  $\text{Retr}$  is the retraction on  $\mathcal{M}$ , and  $\text{grad} f(\mathbf{x}_k)$  denotes the Riemannian gradient [32].

The implementation of hybrid analog/digital beamforming imposes another constraint in the system design: a limited number of phase shifters and analog-to-digital converters (ADCs). Although the power consumption of phase shifters is typically lower than that of baseband beamformers, their number increases with the number of antennas. The implementation of hybrid analog/digital beamformers becomes more complex and expensive at higher frequencies (e.g., the upper mm-wave and THz). As an alternative, lens-based beamformers have been proposed [44]. Instead of using a phase shifter network, they use lenses to generate a directional beam from the EM sources placed at the focal points of the lenses. Thus, lens-based beamformers offer reduced computational complexity when compared with phase shifter-based architectures. Lens-based beamformers, though, only realize directional beams and not more sophisticated beam patterns, as may be useful in a spatial multiplexing or interference cancellation setting. A low-power design in [45] suggests using Butler matrices, which consist of an  $N \times N$  matrix of hybrid couplers and fixed phase shifters.

#### Low-resolution ADCs

Low-resolution (1–3-bit) ADCs for digital beamformers bring down the overall power consumption and hardware cost. In particular, 1-bit ADCs do not require hardware components, such as automatic gain control and linear amplifiers. Hence, the corresponding RF chain is implemented cost-efficiently [46]. Denote the received signal at the receiver and the corresponding beamformer matrix to be  $\mathbf{r} \in \mathbb{C}^{N_{\text{R}}}$  and  $\mathbf{W}_{\text{RF}} \in \mathbb{C}^{N_{\text{R}} \times N_s}$ , respectively. Then, the received signal sampled by low-resolution ADCs is  $\mathbf{r}_q = Q_b(\mathbf{W}_{\text{RF}}^H \mathbf{r})$ , where  $Q_b(\cdot)$  is the quantization operator with  $b$ -bit resolution. The received signal  $\mathbf{r}_q$  is then used to design the receiver via zero-forcing or maximum-rate-combining techniques [42], [46].

#### Finite-resolution phase shifters

In practice, continuous-valued phase angles are expensive to implement, and finite-resolution phase shifters may be used

with low-resolution ADCs. Here, the beamformer weights are selected from the finite set  $\mathcal{W} = \{1, \omega, \omega^2, \dots, \omega^{2^b-1}\}$ , where

$$\omega = \frac{1}{\sqrt{N}} e^{j\frac{2\pi}{2^b}}$$

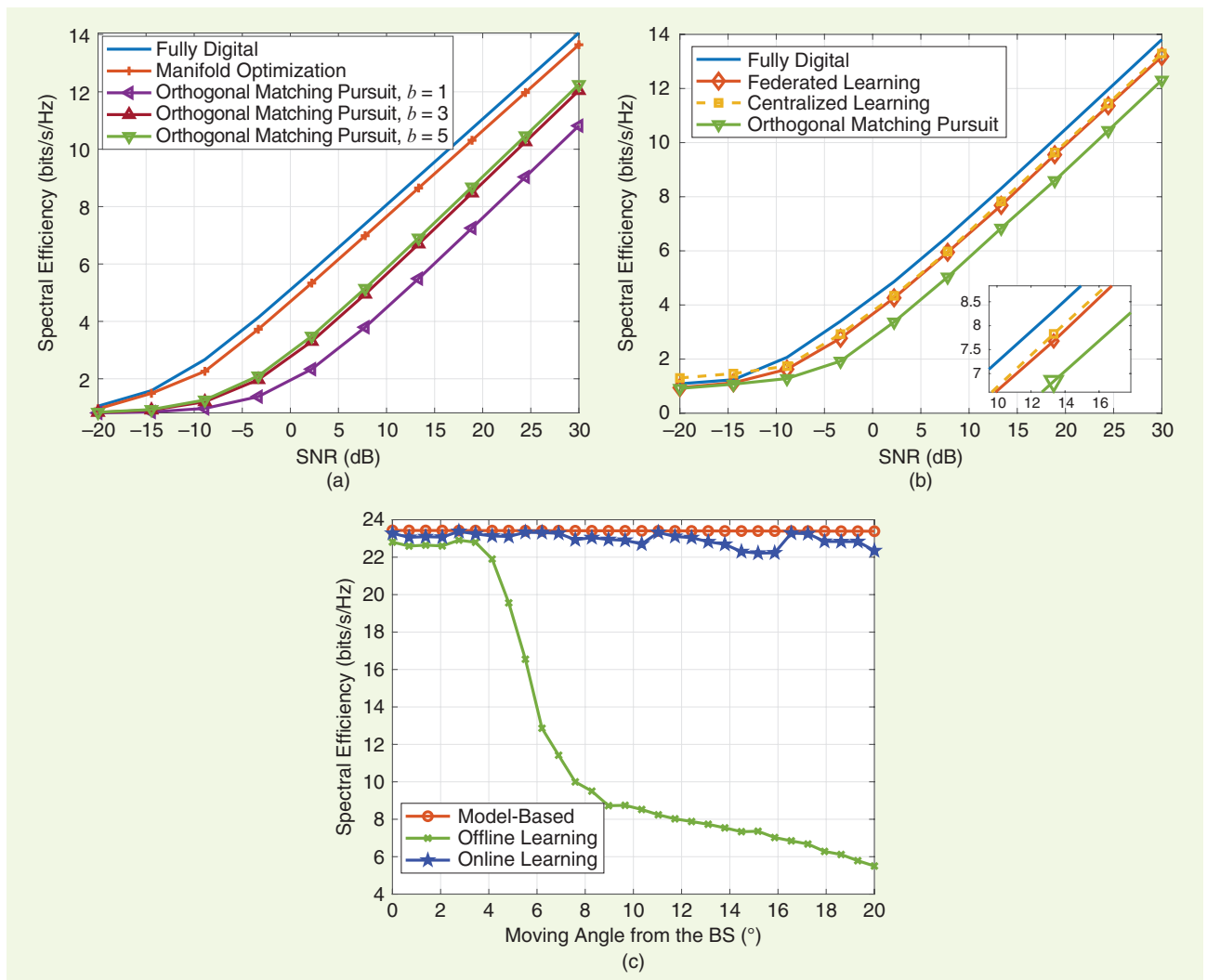
and  $b$  is the number of bits. Then, the constant-modulus constraint in (23) is replaced by  $[\mathbf{F}_{RF}]_{i,j} \in \mathcal{W}$ . A feasible solution to hybrid beamforming with finite resolution is to first solve (23) under the infinite resolution assumption and then quantize the phase elements of the beamformers [33].

Figure 3(a) shows the comparison of fully digital beamforming and hybrid beamforming with low-resolution phase shifters. The hybrid architecture with MO-based design has a performance very close to that of fully digital beamformers. The OMP with  $b = 5$ -bit phase shifters performs closest to infinite-resolution phase shifters. The gap from the fully digital performance is larger for OMP-based techniques compared to MO-based beamforming.

## Learning-based beamforming

Lately, as has been the case with many signal processing problems, beamforming has also not remained untouched by ML techniques. In learning-based hybrid beamforming, the problem is approached from a model-free viewpoint by constructing a nonlinear mapping between the input data (e.g., the channel matrix and array output) and output (beamformers) of a learning model [35], [36], [37]. This method has the following advantages over model-based techniques:

- The model-free/data-driven structure of a learning-based approach yields a robust performance in terms of SE against the corruptions (e.g., a mismatched number of received paths or imperfectly estimated channel gain and path directions [36], [37]) in the input.
- Learning techniques extract feature patterns in the data. Hence, they easily update incoming/future data and adapt in response to environmental changes. The model-based beamformers lack these abilities and may employ statistical predictive algorithms [see Figure 3(c)].



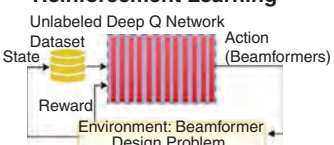
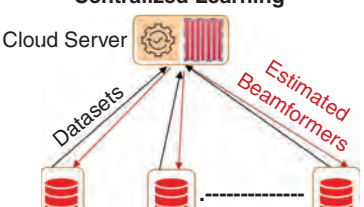
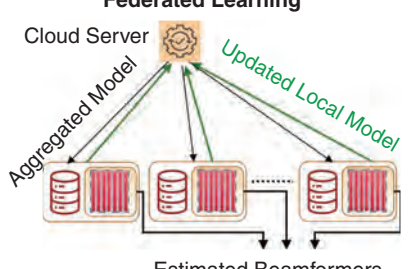
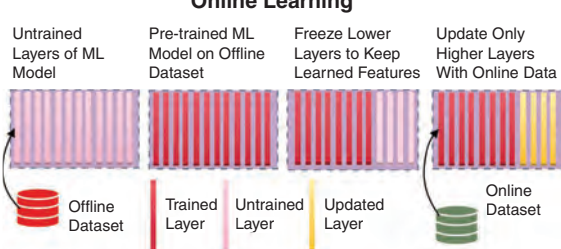


**FIGURE 3.** The SE performance of various hybrid beamforming approaches: (a) low-resolution phase shifters, (b) learning-based and model-based techniques, and (c) offline and online learning. Here, the channel is realized with three paths, the number of BS antenna elements  $N = 100$ , the number of users  $U = 8$ , and the number of user antennas  $M_r = 16$ .

■ Learning exhibits lower computational complexity in the prediction stage than optimization. Through parallel processing, ML significantly (~10-fold [36]) reduces the computational times. On the other hand, a parallel implementation of conventional convex/nonconvex optimization-based beamforming is not straightforward. Beginning from the earlier simpler networks, such as mul-

tilayer perception, to more complex deep learning models like convolutional NNs (CNNs), ML has come a long way in successfully performing feature extraction for analog and digital beamformers [47]. Table 1 summarizes various learning models, including the well-known unsupervised/supervised learning (UL/SL) and the more recent federated learning (FL).

**Table 1. Learning models.**

Network Model	Data	Application in Beamforming
<p><b>Unsupervised Learning</b> Unlabeled Neural Network</p> 	Unlabeled	<i>Fast beamforming:</i> Minimize a given optimization objective to implicitly obtain the beamformers. Unsupervised learning is useful, especially for mobile transmitters, where labels are not available.
<p><b>Supervised Learning</b> Labeled Neural Network</p> 	Labeled	<i>Uplink/downlink beamforming:</i> The network is trained to construct a nonlinear relationship between the input and the labeled data (beamformers).
<p><b>Reinforcement Learning</b> Unlabeled Deep Q Network</p> 	Unlabeled	<i>Uplink/downlink beamforming:</i> The network learns the beamformers based on a reward/punishment mechanism in accordance with optimizing the overall system's SE.
<p><b>Centralized Learning</b></p> 	Labeled/ unlabeled	<i>Uplink multiuser beamforming:</i> The training datasets are transmitted to a centralized cloud server, wherein the model is trained. Posttraining, each user sends the input data (channel) to the server that sends the output (estimated beamformers) to the users.
<p><b>Federated Learning</b></p> 	Labeled/ unlabeled	<i>Downlink multiuser beamforming:</i> Instead of transmitting the whole dataset to the cloud server, each user processes its own local dataset, computes the corresponding model update, and transmits only the updates to the server. Then, the server broadcasts the aggregated model updates to the users, which can estimate their own beamformers.
<p><b>Online Learning</b></p> 	Labeled	<i>Adaptive beamforming:</i> The learning model is updated when the prediction performance degrades because of deviations in the input compared to the training data.



## UL, SL, and semi-supervised learning

UL studies the clustering of unlabeled data into smaller sets by exploiting the hidden features/patterns derived from the dataset, for which an answer key (label) is not provided beforehand. Hence, the “distance” between the training data samples is optimized without prior knowledge of the “meaning” of each clustered set. In SL, however, the labeled data are used for model training while minimizing the error between the label and the model’s response. The cost function of the training is generally the MSE, but other functions (e.g., the mean error, mean absolute error, cross entropy, and Kullback–Leibler divergence) may also be used. Note that beamforming may be cast as either a regression (the output is the beamformer weights) or a classification (the output is an index of a vector from a predefined set of possible beamformers) problem. SL is widely used for several applications of beamformer design in radar and communications [43].

Define  $\mathcal{X} \in \mathbb{R}^{N_{\text{in}}}$  and  $\mathcal{Y} \in \mathbb{R}^{N_{\text{out}}}$  as the input and label data of a learning model whose real-valued learnable parameters are stacked into the vector  $\Theta \in \mathbb{R}^Q$ . Then, the relationship between the input  $\mathcal{X} \in \mathbb{R}^{N_{\text{in}}}$  and output  $\mathcal{Y} \in \mathbb{R}^{N_{\text{out}}}$  is represented by a nonlinear function  $f(\Theta, \mathcal{X}): \mathbb{R}^{N_{\text{in}}} \rightarrow \mathbb{R}^{N_{\text{out}}}$  such that  $\mathcal{Y} = f(\mathcal{X}|\Theta)$ . The input data are, say, the vectorized elements of the channel matrix  $\mathbf{H}$  as  $\mathcal{X} = [\text{vec}\{\Re\{\mathbf{H}\}^T\}, \text{vec}\{\Im\{\mathbf{H}\}^T\}]^T$ , and the labels are beamformers. In the case of the unit-modulus constraint, it suffices to represent the beamformers in terms of only the angle, i.e.,  $\mathcal{Y} = \angle\{\mathbf{F}_{\text{RF}}\}$ . Note that the baseband beamformers are readily computed as  $\mathbf{F}_{\text{BB}} = \mathbf{F}_{\text{RF}}^\dagger \mathbf{F}_{\text{C}}$  [28].

Apart from hybrid beamforming, ML techniques have been applied to other applications, such as robust beamformers [35]. Here, the sample covariance matrix is fed to a CNN whose output is the beamformer weights. The labels are obtained by solving the robust Capon beamformer problem in (8). The training dataset was  $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_J\}$ , where  $\mathcal{D}_i = (\mathcal{X}_i, \mathcal{Y}_i)$  denotes the  $i$ th input–output sample for  $i = 1, \dots, J$ . The model is trained by minimizing the MSE cost

$$\frac{1}{J} \sum_{i=1}^J \|\mathcal{Y}_i - f(\mathcal{X}_i|\Theta)\|_2^2$$

over  $\Theta$ . Posttraining, the learned parameters are used for pre-design purposes for beamforming.

The acoustic beamformers in [48] are obtained via semi-supervised learning (SSL), where both labeled and unlabeled data are used. When a small set of labeled data are available in addition to a large volume of unlabeled data, using both sets in SSL is more advantageous than SL alone.

## Reinforcement learning

In reinforcement learning (RL), the learning model is initialized from a random state, and the algorithms learn to react to the channel conditions on their own [49]. The model accepts the analog and baseband beamformers of the previous state as input and then updates the model parameters by taking into account

the corresponding average rate as a reward. In general, RL has autonomous AI agents that gather their own data and improve based on their trial-and-error interactions with the environment. It shows a lot of promise in basic research. However, so far, RL has been harder to use in real-world beamformer applications

because its dataset does not include labels. Consequently, RL requires longer training times for learning the features of wireless channels, especially in dynamic, short-coherence time scenarios.

## Online learning

The online learning (OL) algorithm involves a learning model whose parameters are updated when there is a significant change in the received input data. For example, consider the beamformer design for a wireless communications system [Figure 3(c)], wherein the user is moving away in the DoA domain from the base station (BS). Then, the received array data become significantly different from the collected offline training data, thereby degrading the network performance. Here, hybrid beamforming and channel estimation may be performed jointly because the beamformer weights are directly related to the channel matrix. Moreover, OL is a suitable choice for this problem [36]; it updates the model parameters when the normalized MSE of channel estimates is higher than a predetermined threshold. From Figure 3(c), the learning model requires retraining every  $\sim 4^\circ$  for a massive MIMO scenario.

FL is more suited for multiuser scenarios. Using the same NN structures, CL has a better performance than FL because the former has access to the whole dataset at once, whereas the latter employs decentralized training. The FL is ideal for downlink, wherein the trained model is available to the user at the network edge. As an example, consider a downlink scenario wherein  $U$  communications users collaborate to train a model with learnable parameters  $\Theta$  with local datasets  $\mathcal{D}^{(u)} = (\mathcal{X}^{(u)}, \mathcal{Y}^{(u)})$  for  $u = 1, \dots, U$ . Here, the output data  $\mathcal{Y}^{(u)}$  are the beamformer weights corresponding to the  $u$ th user. The FL-based training problem minimizes the averaged local cost

$$\min_{\Theta} \frac{1}{U} \sum_{u=1}^U \mathcal{L}_u(\Theta)$$

where  $i = 1, \dots, J_u$  and  $J_u = |\mathcal{D}^{(u)}|$  denotes the number of samples in  $\mathcal{D}^{(u)}$ , over  $\Theta$ . Different than the cost in the “UL, SL, and Semi-supervised Learning” section, the local cost here is

$$\mathcal{L}_u(\Theta) = \frac{1}{J_u} \sum_{i=1}^{J_u} \|f(\mathcal{X}_i^{(u)}|\Theta) - \mathcal{Y}_i^{(u)}\|_2^2$$

for the  $u$ th user. This is efficiently solved by iteratively applying gradient descent, which updates the model parameter at the  $t$ th iteration as

**Using the same NN structures, CL has a better performance than FL because the former has access to the whole dataset at once, whereas the latter employs decentralized training.**

$$\Theta_{t+1} = \Theta_t - \eta \frac{1}{U} \sum_{u=1}^U \beta_u(\Theta_t)$$

where  $\Theta_t$  is the computed model parameter vector at iteration  $t$ ,  $\beta_u(\Theta_t) = \nabla \mathcal{L}_u(\Theta_t) \in \mathbb{R}^Q$  is the gradient vector, and  $\eta$  is the learning rate. Figure 3(b) compares the performance of FL and CL with model-based techniques, such as OMP, and the fully digital beamformer in terms of SE [50]. Both CL and FL outperform OMP, but the performance gap between CL and FL increases with the nonuniformity of the local dataset.

## Emerging applications

Research in beamforming continues to be highly active in light of emerging applications and theoretical advances. For example, the hybrid approach of a model-driven network or deep unfolding for beamforming [51] allows for bounding the complexity of algorithms while also retaining their performance. Convolutional beamformers are gaining salience in acoustics [52] and ultrasound [53] as a means to combine multiple, usually nonlinear, operations with beamforming. There is also recent interest in beamforming for biomimetic antenna arrays that are based on the direction binaural mechanism of humans or animals [54], [55]. Synthetic apertures across a wide variety of applications, including quantum Rydberg sensing, present unique beamforming challenges [56]. Holographic beamformers [57] are currently investigated as attractive solutions for multibeam steering for future wireless applications. In the following, we illustrate a few major applications in the context of radar and communications.

### Joint radar communications

For several decades, sensing and communications systems have exclusively operated in different frequency bands to minimize interference with each other at all times. However, this conservative approach for spectrum access is no longer viable because of the demand for wider bandwidth for the improved performance of both systems. In the last few years, there has been substantial interest in designing joint radar and communications (JRC) [58] to share the spectrum. From a beamformer design perspective, the problem settings of communications and sensing are combined in JRC. Recall the hybrid beamforming for a communications-only problem as explained in (23). The sensing-only beamformer composed of the steering vectors corresponding to, say,  $K$  sensing targets is  $\mathbf{F}_R \in \mathbb{C}^{N_r \times K}$  [43]. Then, similar to (23), the hybrid beamformer for a sensing-only system is obtained by minimizing the Euclidean distance between  $\mathbf{F}_{RF} \mathbf{F}_{BB}$  and  $\mathbf{F}_R \mathbf{P}$  as

$$\begin{aligned} & \underset{\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{P}}{\text{minimize}} \|\mathbf{F}_{RF} \mathbf{F}_{BB} - \mathbf{F}_R \mathbf{P}\|_{\mathcal{F}} \\ & \text{subject to } \|\mathbf{F}_{RF} \mathbf{F}_{BB}\|_{\mathcal{F}} = N_s, \quad |[\mathbf{F}_{RF}]_{i,j}| = \frac{1}{\sqrt{N}}, \quad \forall i, j, \quad \mathbf{P} \mathbf{P}^H = \mathbf{I}_K \end{aligned} \quad (25)$$

where the unitary matrix  $\mathbf{P} \in \mathbb{C}^{K \times N_s}$  is an auxiliary variable to account for different dimensions of  $\mathbf{F}_{RF} \mathbf{F}_{BB}$  and  $\mathbf{F}_R$  without causing any distortion in the radar beam pattern. Define  $\mathbf{F}_{CR} \in \mathbb{C}^{N_t \times N_s}$  as the unconstrained JRC beamformer

$\mathbf{F}_{CR} = \zeta \mathbf{F}_C + (1 - \zeta) \mathbf{F}_R \mathbf{P}$ , where  $0 \leq \zeta \leq 1$  provides a tradeoff between radar and communications performance. Then, the JRC hybrid beamformer is obtained by solving the following optimization problem [43]:

$$\begin{aligned} & \underset{\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{P}}{\text{minimize}} \|\mathbf{F}_{RF} \mathbf{F}_{BB} - \mathbf{F}_{CR}\|_{\mathcal{F}} \\ & \text{subject to } \|\mathbf{F}_{RF} \mathbf{F}_{BB}\|_{\mathcal{F}} = N_s, \quad |[\mathbf{F}_{RF}]_{i,j}| = \frac{1}{\sqrt{N}}, \quad \forall i, j, \quad \mathbf{P} \mathbf{P}^H = \mathbf{I}_K. \end{aligned} \quad (26)$$

Radar and communications can be combined in other ways, for example, leveraging the radar information in a different band to reduce the overheads of configuring the beamforming for communication [59].

### THz communications

THz-band (0.1–10-THz) wireless systems have ultrawide bandwidth and very narrow beamwidth. The signal processing for these systems must address several unique THz challenges, including severe path loss arising from scattering and molecular absorption. In general, THz communications systems employ ultramassive antenna arrays, which may be variously configured as an array of subarrays or group of subarrays [43] (Figure 4) to achieve even higher beamforming gain than mm-wave systems. The wideband beamforming required at THz uses a single analog beamformer for all subcarriers for a hardware-efficient and computationally inexpensive design. However, this leads to beams generated at the lower and higher subcarriers pointing at different directions, resulting in the beam-squint phenomenon [43]. For comparison's sake, the angular deviation in the beam space due to beam squint is approximately  $6^\circ$  ( $0.4^\circ$ ) for 0.3 THz with a 30-GHz (60 GHz with a 1-GHz) bandwidth, respectively. One approach to deal with beam squint is to use time-delayer networks, which is classically known as *space-time filtering*. Alternatively, one may design a single analog beamformer while passing the effect of beam squint into the subcarrier digital beamformers.

Consider the problem in (24), where the analog beamformers are subcarrier independent but the mitigation of beam squint implies their SD-ness. Define  $\tilde{\mathbf{F}}_{BB}[m]$  as a beam-squint-aware digital beamformer. This is obtained via  $\tilde{\mathbf{F}}_{BB}[m] = \mathbf{F}_{RF}^\dagger \tilde{\mathbf{F}}_{RF}[m] \mathbf{F}_{BB}[m]$ , where  $\tilde{\mathbf{F}}_{RF}[m]$  is the SD analog beamformer derived from  $\mathbf{F}_{RF}$  for  $m \in \mathcal{M}$  [43].

### Intelligent reflecting surfaces

An intelligent reflecting surface (IRS) is composed of a large number of (usually passive) metamaterial elements, which reflect the incoming signal by introducing a predetermined phase shift [60]. Thus, IRS-assisted beamforming allows the BS to reach distant/blocked users/targets with low power consumption (Figure 4). Here, joint optimization of the beamformers at the BS as well as the phase shifts of IRS elements is necessary. Consider an IRS-assisted scenario, wherein the IRS is equipped with  $N_{IRS}$  elements, and the BS has  $N$  antennas. The transmitted data symbol  $s \in \mathbb{C}$  is received at the user as  $y_{IRS} = (\mathbf{h}_{IRS}^H \boldsymbol{\Psi} \mathbf{H}_{BS} + \mathbf{h}_D^H) \mathbf{f}_s + e$ , where  $\mathbf{h}_{IRS} \in \mathbb{C}^{N_{IRS}}$ ,  $\mathbf{h}_D \in \mathbb{C}^N$ , and  $\mathbf{H}_{BS} \in \mathbb{C}^{N_{IRS} \times N}$  are the user-IRS, user-BS,

and BS-IRS channels, respectively; the diagonal matrix  $\boldsymbol{\psi} = \text{diag}\{\psi_1, \dots, \psi_{N_{\text{IRS}}}\} \in \mathbb{C}^{N_{\text{IRS}} \times N_{\text{IRS}}}$  represents the IRS phase elements;  $\mathbf{f} \in \mathbb{C}^N$  is the beamformer vector at the BS; and  $e \in \mathbb{C}$  is additive noise. The joint active/passive beamformer design becomes

$$\begin{aligned} & \underset{\boldsymbol{\psi}, \mathbf{f}}{\text{maximize}} \quad |(\mathbf{h}_{\text{IRS}}^H \boldsymbol{\psi} \mathbf{H}_{\text{BS}} + \mathbf{h}_{\text{D}}^H) \mathbf{f}|^2 \\ & \text{subject to} \quad \|\mathbf{f}\|_2 \leq \bar{p}, \quad 0 \leq \psi_n \leq 2\pi \end{aligned} \quad (27)$$

where  $\bar{p}$  denotes the maximum transmit power, and  $n = 1, \dots, N_{\text{IRS}}$ .

### Near-field beamforming

Depending on the operating frequency, the wavefront of the transmitted signal appears to have different shapes in accordance with the observation distance. The wavefront is a plane wave in the far-field region. In the near field (Figure 4), where the transmission range is shorter than the Fraunhofer distance, i.e.,

$$R_{\text{NF}} = \frac{2A^2 f_c}{c_0}$$

with  $A$  being the array aperture, the wavefront takes a spherical form. As a result, unlike the far field, the near-field beam pattern is range dependent. For example, the array response vector for uniform linear array (ULA) is a function of both direction  $\theta$  and range  $r$  as

$$\mathbf{a}(\theta, r) = \frac{1}{\sqrt{N}} \left[ e^{-j\frac{2\pi}{\lambda} r^{(1)}}, \dots, e^{-j\frac{2\pi}{\lambda} r^{(N)}} \right]^T$$

where  $r^{(n)} = [r^2 + ((n-1)d)^2 - 2(n-1)dr \sin\theta]^{1/2} \approx r - (n-1)d \sin\theta$ , ( $n = 1, \dots, N$ ) is a range-dependent parameter corresponding to the receiver and the  $n$ th transmit antenna.

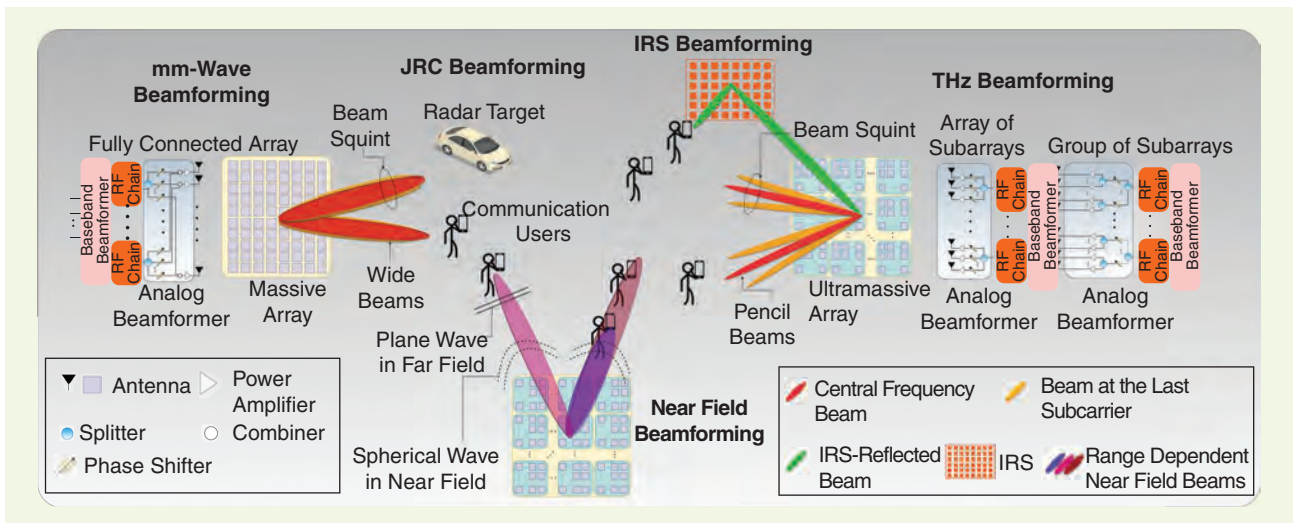
Hence, the beamformer design needs to account for this spherical model.

### Summary

The many beamforming algorithms, their possible variants, and their relative advantages provide a Swiss-knife approach to choosing the most appropriate technique for a specific application. We presented an overview of those algorithms that had a considerable impact on signal processing and system design during the last 25 years. We focused on radar and communications applications while also mentioning in passing the developments in beamforming for ultrasound, acoustics, synthetic apertures, and optics.

A typical use case of convex beamforming is to allow robustness against various sources of uncertainties, such as a small number of snapshots, mismatched SoI direction, and mismatched steering vectors. In nonconvex beamforming, each of the problem settings imposes different constraints on, e.g., PSD-ness (general-rank beamforming), the probability distribution (chance-constrained robust beamforming), constant-modulus (hybrid beamforming), and received SNR (multicast beamforming).

Each learning algorithm offers specific advantages of its own. The most common SL (UL and RL) admits labeled (unlabeled) datasets. Furthermore, the inherent reward/punishment mechanism in RL to optimize the learning model for a pre-defined cost function yields better performance than UL. FL is particularly helpful for multiuser scenarios, whereas CL is preferred if the dataset is small compared to the size of the learning model. When data are updated over time, then OL is beneficial. Note that SL, UL, and RL may also be combined with FL, CL, and OL depending on the problem and data; examples abound, such as federated RL, online RL, online CL, centralized RL, and so on.



**FIGURE 4.** A summary of beamforming in emerging applications. In mm-wave wideband beamforming, the generated beams are squinted while pointing to the same direction, but, at THz, these beams are squinted in considerably different directions. In a JRC scenario, a joint optimization of the beam pattern for both communications users and radar targets should be considered. For IRS-assisted wireless systems, the beamformer weights at the transmitter and the phase shifts of the IRS elements are jointly designed. When the users are in the near-field region of the transmitter, range-dependent beamforming is considered for spatial multiplexing.

## Acknowledgment

Kumar Vijay Mishra acknowledges support from the U.S. National Academies of Sciences, Engineering, and Medicine via an Army Research Laboratory Harry Diamond Distinguished Fellowship.

## Authors

**Ahmet M. Elbir** (ahmetmelbir@ieee.org) received his Ph.D. degree from the Middle East Technical University (METU), Turkey, in 2016 in electrical engineering. He is a research fellow at the University of Luxembourg, L-1855 Luxembourg City, Luxembourg. He serves as an associate editor for *IEEE Access* and a lead guest editor for *IEEE Journal of Selected Topics in Signal Processing* and *IEEE Wireless Communications*. He is the recipient of the 2016 METU Best Ph.D. Thesis Award for his doctoral studies and the IET Radar, Sonar, and Navigation Best Paper Award in 2022. His research interests include array signal processing for radar and communications as well as deep learning for multiantenna systems. He is a Senior Member of IEEE.

**Kumar Vijay Mishra** (kvm@ieee.org) received his Ph.D. degree in electrical and computer engineering from the University of Iowa while working on the NASA Global Precipitation Measurement Mission ground validation radars. He is a senior fellow at the U.S. DEVCOM Army Research Laboratory, Adelphi MD 20783 USA, and technical advisor to start-ups Hertzwell, Singapore, and Aura Intelligent Systems, Boston, MA USA. He is the recipient of the U.S. National Academies Harry Diamond Distinguished Fellowship and has won many best paper awards. His research interests include radar, remote sensing, signal processing, and electromagnetics. He is a Senior Member of IEEE.

**Sergiy A. Vorobyov** (svor@ieee.org) received his Ph.D. degree in systems and control from the National University of Radio Electronics, Kharkiv, Ukraine. He is a professor with the Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland. He was the recipient of the 2004 IEEE Signal Processing Society Best Paper Award, 2007 Alberta Ingenuity New Faculty Award, 2011 Carl Zeiss Award, 2012 NSERC Discovery Accelerator Award, and other awards. He is currently serving as the general cochair for EUSIPCO 2023, Helsinki, Finland. His research interests include optimization and multilinear algebra methods in signal processing and data analysis; statistical and array signal processing; sparse signal processing; estimation; detection and learning theory and methods; and multiantenna, large-scale, and cognitive systems. He is a Fellow of IEEE.

**Robert W. Heath Jr.** (rwheathjr@ncsu.edu) received his Ph.D. degree from Stanford University in electrical engineering. He is the Lampe Distinguished Professor at North Carolina State University, Raleigh, NC 27695 USA, and is president and CEO of MIMO Wireless Inc. He has authored or coauthored several books, including *Introduction to Wireless Digital Communication* (Prentice Hall, 2017) and *Foundations of MIMO Communication* (Cambridge

University Press, 2018). He is the recipient or corecipient of several awards, including the 2019 IEEE Kiyo Tomiyasu Award, 2020 IEEE Signal Processing Society Donald G. Fink Overview Paper Award, 2020 North Carolina State University Innovator of the Year Award, 2021 IEEE Vehicular Technology Society James Evans Avant Garde Award, and 2022 IEEE Vehicular Technology Society Best Vehicular Electronics Paper Award. He was editor-in-chief of *IEEE Signal Processing Magazine* from 2018 to 2020. He is a fellow of the National Academy of Inventors and a Fellow of IEEE.

## References

- [1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988, doi: 10.1109/53.665.
- [2] R. Simons, "Guglielmo Marconi and early systems of wireless communication," *Gec Rev.*, vol. 11, no. 1, pp. 37–55, Jan. 1996.
- [3] T. K. Sarkar, R. Mailloux, A. A. Oliner, M. Salazar-Palma, and D. L. Sengupta, *History of Wireless*. Hoboken, NJ, USA: Wiley, 2006.
- [4] F. Bartlett, "A dual diversity preselector," *QST*, vol. 25, pp. 37–39, Apr. 1941.
- [5] J. C. Chen and K. Yao, "Beamforming," in *Distributed Sensor Networks: Image and Sensor Signal Processing*, vol. 2, S. S. Iyengar and R. R. Brooks, Eds. Boca Raton, FL, USA: CRC Press, 2016, pp. 335–371.
- [6] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969, doi: 10.1109/PROC.1969.7278.
- [7] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, "Adaptive antenna systems," *Proc. IEEE*, vol. 55, no. 12, pp. 2143–2159, Dec. 1967, doi: 10.1109/PROC.1967.6092.
- [8] S. A. Vorobyov, "Adaptive and robust beamforming," in *Array and Statistical Signal Processing* (Academic Press Library in Signal Processing), vol. 3, A. M. Zoubir, M. Viberg, R. Chellappa, and S. Theodoridis, Eds. New York, NY, USA: Academic Press, 2014, pp. 503–552.
- [9] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, 2005, Art. no. 771, doi: 10.1121/1.1913659.
- [10] N. Jablon, "Adaptive beamforming with the generalized sidelobe canceller in the presence of array imperfections," *IEEE Trans. Antennas Propag.*, vol. 34, no. 8, pp. 996–1012, Aug. 1986, doi: 10.1109/TAP.1986.1143936.
- [11] A. B. Gershman, V. I. Turchin, and V. A. Zverev, "Experimental results of localization of moving underwater signal by adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2249–2257, Oct. 1995, doi: 10.1109/78.469863.
- [12] D. Astely and B. Ottersten, "The effects of local scattering on direction of arrival estimation with MUSIC," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3220–3234, Dec. 1999, doi: 10.1109/78.806068.
- [13] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003, doi: 10.1109/TSP.2002.806865.
- [14] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987, doi: 10.1109/TASSP.1987.1165054.
- [15] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003, doi: 10.1109/TSP.2003.812831.
- [16] D. D. Feldman and L. J. Griffiths, "A projection approach for robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 867–876, Apr. 1994, doi: 10.1109/78.285650.
- [17] S. Shahbazpanahi, A. B. Gershman, Z.-Q. Luo, and K. M. Wong, "Robust adaptive beamforming for general-rank signal models," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2257–2269, Sep. 2003, doi: 10.1109/TSP.2003.815395.
- [18] A. Khabbazi-basmenj and S. A. Vorobyov, "Robust adaptive beamforming for general-rank signal model with positive semi-definite constraint via POTDC," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 6103–6117, Dec. 2013, doi: 10.1109/TSP.2013.2281301.
- [19] A. B. Gershman, N. D. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten, "Convex optimization-based beamforming," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 62–75, May 2010, doi: 10.1109/MSP.2010.936015.
- [20] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1684–1696, May 2005, doi: 10.1109/TSP.2005.845436.

- [21] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust Capon beamformer," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2407–2423, Sep. 2004, doi: 10.1109/TSP.2004.831998.
- [22] Y. Huang, M. Zhou, and S. Vorobyov, "New designs on MVDR robust adaptive beamforming based on optimal steering vector estimation," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3624–3638, Jul. 2019, doi: 10.1109/TSP.2019.2918997.
- [23] A. Hassaniien, S. Vorobyov, and K. Wong, "Robust adaptive beamforming using sequential quadratic programming: An iterative solution to the mismatch problem," *IEEE Signal Process. Lett.*, vol. 15, pp. 733–736, Nov. 2008, doi: 10.1109/LSP.2008.2001115.
- [24] A. Khabbazi-basmenj, A. Hassaniien, and S. Vorobyov, "Robust adaptive beamforming based on steering vector estimation with as little as possible prior information," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2974–2987, Jun. 2012, doi: 10.1109/TSP.2012.2189389.
- [25] S. Vorobyov, H. Chen, and A. Gershman, "On the relationship between robust minimum variance beamformers with probabilistic and worst-case distortionless response constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5719–5724, Nov. 2008, doi: 10.1109/TSP.2008.929866.
- [26] Y. Huang, W. Yang, and S. A. Vorobyov, "Robust adaptive beamforming maximizing the worst-case SINR over distributional uncertainty sets for random INC matrix and signal steering vector," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 4918–4922, doi: 10.1109/ICASSP43922.2022.9746616.
- [27] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jul. 2006, doi: 10.1109/TSP.2006.872578.
- [28] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014, doi: 10.1109/TWC.2014.011714.130846.
- [29] Y. Savas, E. Noorani, A. Koppel, J. Baras, U. Topcu, and B. M. Sadler, "Collaborative one-shot beamforming under localization errors: A discrete optimization approach," *Signal Process.*, vol. 200, Nov. 2022, Art. no. 108647, doi: 10.1016/j.sigpro.2022.108647.
- [30] Ö. T. Demir and T. E. Tuncer, "Alternating maximization algorithm for the broadcast beamforming," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, 2014, pp. 1915–1919.
- [31] R. W. Heath Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016, doi: 10.1109/JSTSP.2016.2523924.
- [32] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016, doi: 10.1109/JSTSP.2016.2523903.
- [33] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016, doi: 10.1109/JSTSP.2016.2520912.
- [34] A. Alkhateeb and R. W. Heath Jr., "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, May 2016, doi: 10.1109/TCOMM.2016.2549517.
- [35] S. Mohammadzadeh, V. H. Nascimento, R. C. de Lamare, and N. Hajarolasvadi, "Robust beamforming based on complex-valued convolutional neural networks for sensor arrays," *IEEE Signal Process. Lett.*, vol. 29, pp. 2108–2112, Oct. 2022, doi: 10.1109/LSP.2022.3212637.
- [36] A. M. Elbir, K. V. Mishra, M. R. B. Shankar, and B. Ottersten, "A family of deep learning architectures for channel estimation and hybrid beamforming in multicarrier mm-Wave massive MIMO," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 642–656, Jun. 2022, doi: 10.1109/TCCN.2021.3132609.
- [37] A. M. Elbir and K. V. Mishra, "Joint antenna selection and hybrid beamformer design using unquantized and quantized deep learning networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1677–1688, Mar. 2020, doi: 10.1109/TWC.2019.2956146.
- [38] P. Dong, H. Zhang, and G. Y. Li, "Framework on deep learning-based joint hybrid processing for mmWave massive MIMO systems," *IEEE Access*, vol. 8, pp. 106,023–106,035, Jun. 2020, doi: 10.1109/ACCESS.2020.3000601.
- [39] A. Beck and Y. C. Eldar, "Doubly constrained robust capon beamformer with ellipsoidal uncertainty sets," *IEEE Trans. Signal Process.*, vol. 55, no. 2, pp. 753–758, Jan. 2007, doi: 10.1109/TSP.2006.885729.
- [40] X. Jiang, W.-J. Zeng, A. Yasotharan, H. C. So, and T. Kirubarajan, "Minimum dispersion beamforming for non-gaussian signals," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1879–1893, Apr. 2014, doi: 10.1109/TSP.2014.2305639.
- [41] A. Parayil, A. S. Bedi, and A. Koppel, "Joint position and beamforming control via alternating nonlinear least-squares with a hierarchical gamma prior," in *Proc. Amer. Control Conf. (ACC)*, 2021, pp. 3513–3518, doi: 10.23919/ACC50511.2021.9482851.
- [42] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017, doi: 10.1109/TSP.2017.2706179.
- [43] A. M. Elbir, K. V. Mishra, and S. Chatzinotas, "Terahertz-band joint ultra-massive MIMO radar-communications: Model-based and model-free hybrid beamforming," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1468–1483, Nov. 2021, doi: 10.1109/JSTSP.2021.3117410.
- [44] M. A. B. Abbasi, V. F. Fusco, H. Tataria, and M. Matthaiou, "Constant- $\epsilon_r$  lens beamformer for low-complexity millimeter-wave hybrid MIMO," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 2894–2903, Jul. 2019, doi: 10.1109/TMTT.2019.2903790.
- [45] E.-A. Fazal, C. C. Cavalcante, F. Antreich, A. L. F. De Almeida, and J. A. Nossek, "Efficient hybrid A/D beamforming for millimeter-wave systems using butler matrices," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1001–1013, Feb. 2023, doi: 10.1109/TWC.2022.3200298.
- [46] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath Jr., "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014, doi: 10.1109/MCOM.2014.6979963.
- [47] A. M. Elbir, "CNN-based precoder and combiner design in mmWave MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1240–1243, Jul. 2019, doi: 10.1109/LCOMM.2019.2915977.
- [48] S. Wager, A. Khare, M. Wu, K. Kumtani, and S. Sundaram, "Fully learnable front-end for multi-channel acoustic modeling using semi-supervised learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 6864–6868, doi: 10.1109/ICASSP40776.2020.9053367.
- [49] Q. Wang, K. Feng, X. Li, and S. Jin, "PrecoderNet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1677–1681, Oct. 2020, doi: 10.1109/LWC.2020.3001121.
- [50] A. M. Elbir and S. Coleri, "Federated learning for hybrid beamforming in mm-Wave massive MIMO," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2795–2799, Dec. 2020, doi: 10.1109/LCOMM.2020.3019312.
- [51] S. Shi, Y. Cai, Q. Hu, B. Champagne, and L. Hanzo, "Deep-unfolding neural-network aided hybrid beamforming based on symbol-error probability minimization," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 529–545, Jan. 2023, doi: 10.1109/TVT.2022.3201961.
- [52] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, Jun. 2019, doi: 10.1109/LSP.2019.2911179.
- [53] B. Heriard-Dubreuil, A. Besson, F. Wintzenrieth, J.-P. Thiran, and C. Cohen-Bacrie, "Sparse convolutional plane-wave compounding for ultrasound imaging," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, 2020, pp. 1–4, doi: 10.1109/IUS46767.2020.9251493.
- [54] A. R. Masoumi, Y. Yusuf, and N. Behdad, "Biomimetic antenna arrays based on the directional hearing mechanism of the parasitoid fly *Ormia ochracea*," *IEEE Trans. Antennas Propag.*, vol. 61, no. 5, pp. 2500–2510, May 2013, doi: 10.1109/TAP.2013.2245091.
- [55] A. R. Masoumi and N. Behdad, "An improved architecture for two-element biomimetic antenna arrays," *IEEE Trans. Antennas Propag.*, vol. 61, no. 12, pp. 6224–6228, Dec. 2013, doi: 10.1109/TAP.2013.2281352.
- [56] P. Vouras et al., "An overview of advances in signal processing techniques for classical and quantum wideband synthetic apertures," *IEEE J. Sel. Topics Signal Process.*, early access, Mar. 2023, doi: 10.1109/JSTSP.2023.3262443.
- [57] R. Deng, B. Di, H. Zhang, Y. Tan, and L. Song, "Reconfigurable holographic surface: Holographic beamforming for metasurface-aided wireless communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6255–6259, Jun. 2021, doi: 10.1109/TVT.2021.3079465.
- [58] K. V. Mishra, M. R. Bhavani Shankar, V. Koivunen, B. Ottersten, and S. A. Vorobyov, "Toward millimeter wave joint radar-communications: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 100–114, Sep. 2019, doi: 10.1109/MSP.2019.2913173.
- [59] A. Ali, N. Gonzalez-Prelcic, and A. Ghosh, "Passive radar at the roadside unit to configure millimeter wave vehicle-to-infrastructure links," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14,903–14,917, Dec. 2020, doi: 10.1109/TVT.2020.3027636.
- [60] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019, doi: 10.1109/TWC.2019.2936025.

# DATES AHEAD

Please send calendar submissions to:  
Dates Ahead, At: Samantha Walter, Email: [walter.samantha@ieee.org](mailto:walter.samantha@ieee.org)

## 2023

### JUNE

#### IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

4–9 June, Rhodes Island, Greece.  
General Chairs: Petros Maragos,  
Kostas Berberidis, and Petros Boufounos  
URL: <https://2023.ieeeicassp.org>

#### IEEE Conference on Artificial Intelligence (CAI)

7–8 June, Santa Clara, CA, USA.  
General Chairs: Piero Bonissone and Gary Fogel  
URL: <https://cai.ieee.org/2023/>

#### 15th International Conference on Quality of Multimedia Experience (QoMEX 2023)

20–22 June, Ghent, Belgium.  
General Chairs: Maria Torres Vega  
and Katrien De Moor  
URL: <https://sites.google.com/view/qomex2023?pli=1>

### JULY

#### IEEE Statistical Signal Processing Workshop (SSP)

2–5 July, Hanoi, Vietnam.  
General Chairs: Karim Abed-Meraim and  
Nguyen Linh Trung  
URL: <https://avitech.uet.vnu.edu.vn/ssp2023/>

#### IEEE International Conference on Multimedia and Expo (ICME 2023)

10–14 July, Brisbane, Australia.  
General Chairs: Ambarish Natu, Shan Liu,  
and Zhu Li  
URL: <https://www.2023.ieeeicme.org/>

### SEPTEMBER

#### 31st European Signal Processing Conference (EUSIPCO)

4–8 September, Helsinki, Finland.  
General Chairs: Esa Ollila and Sergiy A. Vorobyov  
URL: <http://eusipco2023.org/>

#### 12th Conference of the Sensor Signal Processing for Defense (SSPD) Series

12–13 September, Edinburgh, UK.  
General Chairs: Mike Davies, Stephen  
McLaughlin, Jordi Barr, and Gary Heald  
<https://sspd.eng.ed.ac.uk/>

Digital Object Identifier 10.1109/MSP.2023.3262459  
Date of current version: 1 June 2023



©SHUTTERSTOCK.COM/PHILIPPOS PHILIPPOU

The IEEE Signal Processing Society will celebrate its 75th Anniversary during the International Conference on Acoustic, Speech and Signal Processing, to be held in Rhodes Island, Greece, 4–10 June 2023.

#### IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP 2023)

17–20 September, Rome, Italy.  
General Chairs: Danilo Comminiello, Tulya  
Adali, and Aurelio Uncini  
URL: <https://2023.ieeemlsp.org/>

#### IEEE International Conference on Quantum Computing and Engineering (QCE23)

17–22 September, Bellevue, WA, USA.  
General Chair: Hausi Müller  
URL: <https://qce.quantum.ieee.org/2023/>

#### International Symposium on Image and Signal Processing and Analysis (ISPA 2023)

18–19 September, Rome, Italy.  
General Chairs: Marco Carli, Federica Battisti,  
and Sven Lončarić  
URL: <https://www.isipa.org/home>

#### IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2023)

25–28 September, Shanghai, China.  
General Chairs: Zhi Tian and Xin Wang  
URL: <http://2023.ieeespawc.org/>

### OCTOBER

#### IEEE International Conference on Image Processing (ICIP 2023)

8–11 October, Kuala Lumpur, Malaysia.  
General Co-Chairs: Norliza Mohd, Gaurav  
Sharma, and Mohan Kankanhalli  
URL: <https://2023.ieeeicip.org/>

#### IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2023)

22–25 October, New Paltz, NY, USA.  
General Chairs: Minje Kim  
and Nicholas J. Bryan  
<https://waspa.com/>

#### Asilomar Conference on Signals, Systems, and Computers (ACSSC 2023)

29 October–1 November,  
Pacific Grove, CA, USA.  
General Chair: Marco F. Duarte  
URL: <https://www.asilomarssc.org/>

#### Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2023)

October 31–November 3, Taipei, Taiwan.  
General Chairs: JIng-Ming Guo, Gwo-Giun Lee,  
Shih-Fu Chang, and Anthony Kuh  
URL: <https://www.apsipa2023.org>

### DECEMBER

#### 9th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2023)

10–13 December, Costa Rica.  
General Chairs: M. Haardt and André de Almeida  
<https://www.tuwien.at/eti/tc/en/camsap-2023/>





*New benefit from the IEEE Signal Processing Society*

# SPS Resource Center

The SPS Resource Center is the new home for the IEEE Signal Processing Society's online library of tutorials, lectures, presentations, and more. Unrestricted access to our fast-growing archive is now included with your SPS membership.

<http://rc.signalprocessingsociety.org>

**We accept submissions, too!**  
**Interested in submitting your educational materials?**

[sps-resourcecenter@ieee.org](mailto:sps-resourcecenter@ieee.org)

# MATLAB FOR AI

Boost system design and simulation with explainable and scalable AI. With MATLAB and Simulink, you can easily train and deploy AI models.

[mathworks.com/ai](https://mathworks.com/ai)

