

Springer
Handbook *of*
**Speech
Processing**



Benesty
Sondhi
Huang
Editors

 Springer

Springer Handbook of Speech Processing

Springer Handbooks provide a concise compilation of approved key information on methods of research, general principles, and functional relationships in physical sciences and engineering. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned editors to write the chapters comprising each volume. The content is selected by these experts from Springer sources (books, journals, online content) and other systematic and approved recent publications of physical and technical information.

The volumes are designed to be useful as readable desk reference books to give a fast and comprehensive overview and easy retrieval of essential reliable key information, including tables, graphs, and bibliographies. References to extensive sources are provided.

Springer Handbook of Speech Processing

Jacob Benesty, M. Mohan Sondhi, Yiteng Huang
(Eds.)

With DVD-ROM, 456 Figures and 113 Tables



Springer

Editors:

Jacob Benesty
INRS-EMT, University of Quebec
800 de la Gauchetiere Ouest, Suite 6900
Montreal, Quebec, H5A 1K6, Canada
benesty@emt.inrs.ca

M. Mohan Sondhi
Avayalabs Research
233 Mount Airy Road
Basking Ridge, NJ 07920, USA
mms@research.avayalabs.com

Yiteng Huang
Bell Laboratories, Alcatel-Lucent
600 Mountain Avenue
Murray Hill, NJ 07974, USA
arden_huang@ieee.org

Library of Congress Control Number: 2007931999

ISBN: 978-3-540-49125-5 e-ISBN: 978-3-540-49127-9

This work is subject to copyright. All rights reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September, 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2008

The use of designations, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Product liability: The publisher cannot guarantee the accuracy of any information about dosage and application contained in this book. In every individual case the user must check such information by consulting the relevant literature.

Typesetting and production:
LE- $\TeX$ Jelonek, Schmidt&Vöckler GbR, Leipzig
Senior Manager Springer Handbook: Dr. W. Skolaut, Heidelberg
Typography and layout: schreiberVIS, Seeheim
Illustrations: Hippmann GbR, Schwarzenbruck
Cover design: eStudio Calamar Steinen, Barcelona
Cover production: WMXDesign GmbH, Heidelberg
Printing and binding: Stürtz GmbH, Würzburg

Printed on acid free paper

SPIN 11544036 60/3180/YL 5 4 3 2 1 0

Foreword

Over the past three decades digital signal processing has emerged as a recognized discipline. Much of the impetus for this advance stems from research in representation, coding, transmission, storage and reproduction of speech and image information. In particular, interest in voice communication has stimulated central contributions to digital filtering and discrete-time spectral transforms.

This dynamic development was built upon the convergence of three then-evolving technologies: (i) sampled-data theory and representation of information signals (which led directly to digital telecommunication that provides signal quality independent of transmission distance); (ii) electronic binary computation (aided in early implementation by pulse-circuit techniques from radar design); and, (iii) invention of solid-state devices for exquisite control of electronic current (transistors – which now, through microelectronic materials, scale to systems of enormous size and complexity). This timely convergence was soon followed by optical fiber methods for broadband information transport.

These advances impact an important aspect of human activity – information exchange. And, over man's existence, speech has played a principal role in human communication. Now, speech is playing an increasing role in human interaction with complex information systems. Automatic services of great variety exploit the comfort of voice exchange, and, in the corporate sector, sophisticated audio/video teleconferencing is reducing the necessity of expensive, time-consuming business travel. In each instance an overarching target is a user environment that captures some of the naturalness and spatial realism of face-to-face communication. Again, speech is a core element, and new understanding from diverse research sectors can be brought to bear.

Editors-in-Chief Benesty, Sondhi and Huang have organized a timely engineering handbook to answer this need. They have assembled a remarkable compendium of current knowledge in speech processing. And, this accumulated understanding can be focused upon enlarging the human capacity to deal with a world ever increasing in complexity. Benesty, Sondhi and Huang are renowned researchers in their own right, and they have attracted an international cadre of over 80 fellow authors and collaborators who constitute a veritable *Who's Who* of world leaders in speech processing research. The resulting book provides under one cover authoritative treatments that commence with the basic physics and psychophysics of speech and hearing, and range through the related topics of computational tools, coding, synthesis, recognition, and signal enhancement, concluding with discussions on capture and projection of sound in enclosures. The book can be expected to become a valuable resource for researchers, engineers and speech scientists throughout the global community. It should equally serve teachers and students in human communication, especially delimiting knowledge frontiers where graduate thesis research may be appropriate.

Warren, New Jersey
October 2007

Jim Flanagan



J. L. Flanagan
Professor Emeritus
Electrical and Computer
Engineering
Rutgers University

Preface

The achievement of this Springer Handbook is the result of a wonderful journey that started in March 2005 at the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Two of the editors-in-chief (Benesty and Huang) met in one of the long corridors of the Pennsylvania Convention Center in Philadelphia with Dr Dieter Merkle from Springer. Together we had a very nice discussion about the conference and immediately an idea came up for a handbook. After a short discussion we converged without too much hesitation on a handbook of *speech processing*. It was quite surprising to see that, even after 30 years of ICASSP and more than half a century of research in this fundamental area, there was still no major book summarizing the important aspects of speech processing. We thought that the time was ripe for such a large project. Soon after we got home, a third editor-in-chief (Sondhi) joined the efforts.

We had a very clear objective in our minds: to summarize, in a reasonable number of pages, the most important and useful aspects of speech processing. The content was then organized accordingly. This task was not easy since we had to find a good balance between feasible ideas and new trends. As we all know, practical ideas can be viewed as *old stuff* while emerging ideas can be criticized for not having passed the test of time; we hope that we have succeeded in finding a good compromise. For this we relied on many authors who are well established and are recognized as experts in their field, from all over the world, and from academia as well as from industry.

From *simple* consumer products such as cell phones and MP3 players to more-sophisticated projects such as human-machine interfaces and robots that can obey orders, speech technologies are now everywhere. We believe that it is just a matter of time before more applications of the science of speech become impossible to miss in our daily life. So we believe that this Springer Handbook will play a fundamental role in the sustainable progress of speech research and development.

This handbook is targeted at three categories of readers: graduate students of speech processing, professors and researchers in academia and research labs who are active in this field, and engineers in industry who need to understand or implement specific algorithms for their speech-related products. The handbook could also be used as a text for one or more graduate courses on signal processing for speech and various aspects of speech processing and applications.

For the completion of such an ambitious project we have many people to thank. First, we would like to thank the many authors who did a terrific job in delivering very high-quality chapters. Second, we are very grateful to the members of the editorial board who helped us so much in organizing the content and structure of this book, taking part in all phases of this project from conception to completion. Third, we would like to thank all the reviewers, who helped us to improve the quality of the material. Last, but not least, we would like to thank the Springer team for their availability and very professional work. In particular, we appreciated the help of Dieter Merkle, Christoph Baumann, Werner Skolaut, Petra Jantzen, and Claudia Rau.

We hope this Springer Handbook will inspire many great minds to find new research ideas or to implement algorithms in products.

Montreal, Basking Ridge, Murray Hill
October 2007

Jacob Benesty
M. Mohan Sondhi
Yiteng Huang



Jacob Benesty



M. Mohan Sondhi



Yiteng Huang

List of Editors

Editors-in-Chief

Jacob Benesty, Montreal
M. Mohan Sondhi, Basking Ridge
Yiteng (Arden) Huang, Murray Hill

Part Editors

Part A: Production, Perception, and Modeling of Speech

M. M. Sondhi, Basking Ridge

Part B: Signal Processing for Speech

Y. Huang, Murray Hill; J. Benesty, Montreal

Part C: Speech Coding

W. B. Kleijn, Stockholm

Part D: Text-to-Speech Synthesis

S. Narayanan, Los Angeles

Part E: Speech Recognition

L. Rabiner, Piscataway; B.-H. Juang, Atlanta

Part F: Speaker Recognition

S. Parthasarathy, Sunnyvale

Part G: Language Recognition

C.-H. Lee, Atlanta

Part H: Speech Enhancement

J. Chen, Murray Hill; S. Gannot, Ramat-Gan; J. Benesty, Montreal

Part I: Multichannel Speech Processing

J. Benesty, Montreal; I. Cohen, Haifa; Y. Huang, Murray Hill

List of Authors

Alex Acero

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
e-mail: alexac@microsoft.com

Jont B. Allen

University of Illinois
ECE
Urbana, IL 61801, USA
e-mail: JontAllen@ieee.org

Jacob Benesty

University of Quebec
INRS-EMT
800 de la Gauchetiere Ouest
Montreal, Quebec H5A 1K6, Canada
e-mail: benesty@emt.inrs.ca

Frédéric Bimbot

IRISA (CNRS & INRIA) – METISS
Pièce C 320 – Campus Universitaire de Beaulieu
35042 Rennes, France
e-mail: bimbot@irisa.fr

Thomas Brand

Carl von Ossietzky Universität Oldenburg
Sektion Medizinphysik
Haus des Hörens, Marie-Curie-Str. 2
26121 Oldenburg, Germany
e-mail: thomas.brand@uni-oldenburg.de

Nick Campbell

Knowledge Creating Communication Research
Centre
Acoustics & Speech Research Project, Spoken
Language Communication Group
2-2-2 Hikaridai
619-0288 Keihanna Science City, Japan
e-mail: nick@nict.go.jp

William M. Campbell

MIT Lincoln Laboratory
Information Systems Technology Group
244 Wood Street
Lexington, MA 02420-9108, USA
e-mail: wcampbell@ll.mit.edu

Rolf Carlson

Royal Institute of Technology (KTH)
Department of Speech, Music and Hearing
Lindstedtsvägen 24
10044 Stockholm, Sweden
e-mail: rolf@speech.kth.se

Jingdong Chen

Bell Laboratories
Alcatel-Lucent
600 Mountain Ave
Murray Hill, NJ 07974, USA
e-mail: jingdong@research.bell-labs.com

Juin-Hwey Chen

Broadcom Corp.
5300 California Avenue
Irvine, CA 92617, USA
e-mail: rchen@broadcom.com

Israel Cohen

Technion-Israel Institute of Technology
Department of Electrical Engineering
Technion City
Haifa 32000, Israel
e-mail: icohen@ee.technion.ac.il

Jordan Cohen

SRI International
300 Ravenswood Drive
Menlo Park, CA 94019, USA
e-mail: jrc@speech.sri.com

Corinna Cortes

Google, Inc.
Google Research
76 9th Avenue, 4th Floor
New York, NY 10011, USA
e-mail: corinna@google.com

Eric J. Diethorn

Avaya Labs Research
Multimedia Technologies Research Department
233 Mt. Airy Road
Basking Ridge, NJ 07920, USA
e-mail: ejd@avaya.com

Simon Doclo

Katholieke Universiteit Leuven
Department of Electrical Engineering (ESAT-SCD)
Kasteelpark Arenberg 10 bus 2446
3001 Leuven, Belgium
e-mail: *simon.doclo@esat.kuleuven.be*

Jasha Droppo

Microsoft Research
Speech Technology Group
One Microsoft Way
Redmond, WA 98052, USA
e-mail: *jdroppo@microsoft.com*

Thierry Dutoit

Faculté Polytechnique de Mons FPMs
TCTS Laboratory
Bvd Dolez, 31
7000 Mons, Belgium
e-mail: *thierry.dutoit@fpms.ac.be*

Gary W. Elko

mh acoustics LLC
25A Summit Ave
Summit, NJ 07901, USA
e-mail: *gwe@mhacoustics.com*

Sadaoki Furui

Tokyo Institute of Technology Street
Department of Computer Science
2-12-1 Ookayama, Meguro-ku
152-8552 Tokyo, Japan
e-mail: *furui@cs.titech.ac.jp*

Sharon Gannot

Bar-Ilan University
School of Electrical Engineering
Ramat-Gan 52900, Israel
e-mail: *gannot@eng.biu.ac.il*

Mazin E. Gilbert

AT&T Labs, Inc., Research
180 Park Ave.
Florham Park, NJ 07932, USA
e-mail: *mazin@research.att.com*

Michael M. Goodwin

Creative Advanced Technology Center
Audio Research
1500 Green Hills Road
Scotts Valley, CA 95066, USA
e-mail: *mgoodwin@atc.creative.com*

Volodya Grancharov

Multimedia Technologies
Ericsson Research, Ericsson AB
Torshamnsgatan 23, Kista, KI/EAB/TVA/A
16480 Stockholm, Sweden
e-mail: *volodya.grancharov@ericsson.com*

Björn Granström

Royal Institute of Technology (KTH)
Department for Speech, Music and Hearing
Lindstedsvägen 24
10044 Stockholm, Sweden
e-mail: *bjorn@speech.kth.se*

Patrick Haffner

AT&T Labs-Research
IP and Voice Services
200 S Laurel Ave.
Middletown, NJ 07748, USA
e-mail: *haffner@research.att.com*

Roar Hagen

Global IP Solutions
Magnus Ladulsgatan 63B
118 27 Stockholm, Sweden
e-mail: *roar.hagen@gipscorp.com*

Mary P. Harper

University of Maryland
Center for Advanced Study of Language
7005 52nd Avenue
College Park, MD 20742, USA
e-mail: *mharper@casl.umd.edu*

Jürgen Herre

Fraunhofer Institute for Integrated Circuits
(Fraunhofer IIS)
Audio and Multimedia
Am Wolfsmantel 33
91058 Erlangen, Germany
e-mail: *hrr@iis.fraunhofer.de*

Wolfgang J. Hess

University of Bonn
 Institute for Communication Sciences, Dept. of
 Communication, Language, and Speech
 Poppelsdorfer Allee 47
 53115 Bonn, Germany
 e-mail: *wgh@ifk.uni-bonn.de*

Kiyoshi Honda

Université de la Sorbonne Nouvelle-Paris III
 Laboratoire de Phonétique et de Phonologie, ATR
 Cognitive Information Laboratories
 UMR-7018-CNRS, 46, rue Barrault
 75634 Paris, France
 e-mail: *honda@atr.jp*

Yiteng (Arden) Huang

Bell Laboratories
 Alcatel-Lucent
 600 Mountain Avenue
 Murray Hill, NJ 07974, USA
 e-mail: *arden_huang@ieee.org*

Matthieu Hébert

Network ASR Core Technology
 Nuance Communications
 1500 Université
 Montréal, Québec H3A-3S7, Canada
 e-mail: *hebert@nuance.com*

Biing-Hwang Juang

Georgia Institute of Technology
 School of Electrical & Computer Engineering
 777 Atlantic Dr. NW
 Atlanta, GA 30332-0250, USA
 e-mail: *juang@ece.gatech.edu*

Tatsuya Kawahara

Kyoto University
 Academic Center for Computing and Media Studies
 Sakyo-ku
 606-8501 Kyoto, Japan
 e-mail: *kawahara@i.kyoto-u.ac.jp*

Ulrik Kjems

Oticon A/S
 9 Kongebakken
 2765 Smørum, Denmark
 e-mail: *uk@oticon.dk*

Esther Klabbers

Oregon Health & Science University
 Center for Spoken Language Understanding, OGI
 School of Science and Engineering
 20000 NW Walker Rd
 Beaverton, OR 97006, USA
 e-mail: *klabbers@cslu.ogi.edu*

W. Bastiaan Kleijn

Royal Institute of Technology (KTH)
 School of Electrical Engineering, Sound and Image
 Processing Lab
 Osquidas väg 10
 10044 Stockholm, Sweden
 e-mail: *bastiaan.kleijn@ee.kth.se*

Birger Kollmeier

Universität Oldenburg
 Medizinische Physik
 26111 Oldenburg, Germany
 e-mail: *birger.kollmeier@uni-oldenburg.de*

Ermin Kozica

Royal Institute of Technology (KTH)
 School of Electrical Engineering, Sound and Image
 Processing Laboratory
 Osquidas väg 10
 10044 Stockholm, Sweden
 e-mail: *ermin.kozica@ee.kth.se*

Sen M. Kuo

Northern Illinois University
 Department of Electrical Engineering
 DeKalb, IL 60115, USA
 e-mail: *kuo@ceet.niu.edu*

Jan Larsen

Technical University of Denmark
 Informatics and Mathematical Modelling
 Richard Petersens Plads
 2800 Kongens Lyngby, Denmark
 e-mail: *jl@imm.dtu.dk*

Chin-Hui Lee

Georgia Institute of Technology
 School of Electrical and Computer Engineering
 777 Atlantic Drive NW
 Atlanta, GA 30332-0250, USA
 e-mail: *chl@ece.gatech.edu*

Haizhou Li

Institute for Infocomm Research
Department of Human Language Technology
21 Heng Mui Keng Terrace
Singapore, 119613
e-mail: hli@i2r.a-star.edu.sg

Jan Linden

Global IP Solutions
301 Brannan Street
San Francisco, CA 94107, USA
e-mail: jan.linden@gipscorp.com

Manfred Lutzky

Fraunhofer Integrated Circuits (IIS)
Multimedia Realtime Systems
Am Wolfsmantel 33
91058 Erlangen, Germany
e-mail: manfred.lutzky@iis.fraunhofer.de

Bin Ma

Human Language Technology
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
e-mail: mabin@i2r.a-star.edu.sg

Michael Maxwell

University of Maryland
Center for Advanced Study of Language
Box 25
College Park, MD 20742, USA
e-mail: mmaxwell@casl.umd.edu

Alan V. McCree

MIT Lincoln Laboratory
Department of Information Systems Technology
244 Wood Street
Lexington, MA 02420-9185, USA
e-mail: mccree@ll.mit.edu

Bernd Meyer

Carl von Ossietzky Universität Oldenburg
Medical Physics Section, Haus des Hörens
Marie-Curie-Str. 2
26121 Oldenburg, Germany
e-mail: bernd.meyer@uni-oldenburg.de

Jens Meyer

mh acoustics
25A Summit Ave.
Summit, NJ 07901, USA
e-mail: jmm@mhacoustics.com

Taniya Mishra

Oregon Health and Science University
Center for Spoken Language Understanding,
Computer Science and Electrical Engineering, OGI
School of Science and Engineering
20000 NW Walker Road
Beaverton, OR 97006, USA
e-mail: mishra@cslu.ogi.edu

Mehryar Mohri

Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY 10012, USA
e-mail: mohri@cs.nyu.edu

Marc Moonen

Katholieke Universiteit Leuven
Electrical Engineering Department ESAT/SISTA
Arenberg 10
3001 Leuven, Belgium
e-mail: marc.moonen@esat.kuleuven.be

Dennis R. Morgan

Bell Laboratories, Alcatel-Lucent
700 Mountain Avenue 2D-537
Murray Hill, NJ 07974-0636, USA
e-mail: drmm@bell-labs.com

David Nahamoo

IBM Thomas J. Watson Research Center
PO BOX 218
Yorktown Heights, NY 10598, USA
e-mail: nahamoo@us.ibm.com

Douglas O'Shaughnessy

Université du Québec
INRS Énergie, Matériaux et Télécommunications
(INRS-EMT)
800, de la Gauchetière Ouest
Montréal, Québec H5A 1K6, Canada
e-mail: dougo@emt.inrs.ca

Lucas C. Parra

Steinman Hall, The City College of New York
 Department of Biomedical Engineering
 140th and Convent Ave
 New York, NY 10031, USA
 e-mail: parra@ccny.cuny.edu

Sarangarajan Parthasarathy

Yahoo!, Applied Research
 1MC 743, 701 First Avenue
 Sunnyvale, CA 94089-0703, USA
 e-mail: parthas@yahoo-inc.com

Michael Syskind Pedersen

Oticon A/S
 Kongebakken 9
 2765 Smørum, Denmark
 e-mail: msp@oticon.dk

Fernando Pereira

University of Pennsylvania
 Department of Computer and Information Science
 305 Levine Hall, 3330 Walnut Street
 Philadelphia, PA 19104, USA
 e-mail: pereira@cis.upenn.edu

Michael Picheny

IBM Thomas J. Watson Research Center
 Yorktown Heights, NY 10598, USA
 e-mail: Picheny@us.ibm.com

Rudolf Rabenstein

University Erlangen-Nuremberg
 Electrical Engineering, Electronics, and
 Information Technology
 Cauerstrasse 7/LMS
 91058 Erlangen, Germany
 e-mail: rabe@LNT.de

Lawrence Rabiner

Rutgers University
 Department of Electrical and Computer
 Engineering
 96 Frelinghuysen Road
 Piscataway, NJ 08854, USA
 e-mail: lrr@caip.rutgers.edu

Douglas A. Reynolds

Massachusetts Institute of Technology
 Lincoln Laboratory, Information Systems
 Technology Group
 244 Wood Street
 Lexington, MA 02420-9108, USA
 e-mail: dar@ll.mit.edu

Michael Riley

Google, Inc., Research
 111 Eighth AV
 New York, NY 10011, USA
 e-mail: riley@google.com

Aaron E. Rosenberg

Rutgers University
 Center for Advanced Information Processing
 96 Frelinghuysen Road
 Piscataway, NJ 08854-8088, USA
 e-mail: aer@caip.rutgers.edu

Salim Roukos

IBM T. J. Watson Research Center
 Multilingual NLP Technologies
 Yorktown Heights, NY 10598, USA
 e-mail: roukos@us.ibm.com

Jan van Santen

Oregon Health And Science University
 OGI School of Science and Engineering,
 Department of Computer Science and Electrical
 Engineering
 20000 NW Walker Rd
 Beaverton, OR 97006-8921, USA
 e-mail: vansanten@cslu.ogi.edu

Ronald W. Schafer

Hewlett-Packard Laboratories
 1501 Page Mill Road
 Palo Alto, CA 94304, USA
 e-mail: ron.schafer@hp.com

Juergen Schroeter

AT&T Labs - Research
 Department of Speech Algorithms and Engines
 180 Park Ave
 Florham Park, NJ 07932, USA
 e-mail: schroeter@att.com

Stephanie Seneff

Massachusetts Institute of Technology
Computer Science and Artificial
Intelligence Laboratory
32 Vassar Street
Cambridge, MA 02139, USA
e-mail: senef@csail.mit.edu

Wade Shen

Massachusetts Institute of Technology
Communication Systems, Information Systems
Technology, Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108, USA
e-mail: swade@ll.mit.edu

Elliot Singer

Massachusetts Institute of Technology
Information Systems Technology Group, Lincoln
Laboratory
244 Wood Street
Lexington, MA 02420-9108, USA
e-mail: es@ll.mit.edu

Jan Skoglund

Global IP Solutions
301 Brannan Street
San Francisco, CA 94107, USA
e-mail: jan.skoglund@gipscorp.com

M. Mohan Sondhi

Avayalabs Research
233 Mount Airy Road
Basking Ridge, NJ 07920, USA
e-mail: mms@research.avayalabs.com

Sascha Spors

Deutsche Telekom AG, Laboratories
Ernst-Reuter-Platz 7
10587 Berlin, Germany
e-mail: Sascha.Spors@telekom.de

Ann Spriet

ESAT-SCD/SISTA, K.U. Leuven
Department of Electrical Engineering
Kasteelpark Arenberg 10
3001 Leuven, Belgium
e-mail: ann.spriet@esat.kuleuven.be

Richard Sproat

University of Illinois at Urbana-Champaign
Department of Linguistics
Urbana, IL 61801, USA
e-mail: rws@uiuc.edu

Yannis Stylianou

Institute of Computer Science
Heraklion, Crete 700 13, Greece
e-mail: yannis@csd.uoc.gr

Jes Thyssen

Broadcom Corporation
5300 California Avenue
Irvine, CA 92617, USA
e-mail: jthyssen@broadcom.com

Jay Wilpon

Research AT&T Labs
Voice and IP Services
Florham Park, NJ 07932, USA
e-mail: jgw@research.att.com

Jan Wouters

ExpORL, Department of Neurosciences, K.U. Leuven
O.& N2, Herestraat 49
3000 Leuven, Belgium
e-mail: jan.wouters@med.kuleuven.be

Arie Yeredor

Tel-Aviv University
Electrical Engineering – Systems
Tel-Aviv 69978, Israel
e-mail: arie@eng.tau.ac.il

Steve Young

Cambridge University Engineering Dept
Cambridge, CB21PZ, UK
e-mail: sjy@eng.cam.ac.uk

Victor Zue

Massachusetts Institute of Technology
CSAI Laboratory
32 Vassar Street
Cambridge, MA 02139, USA
e-mail: zue@csail.mit.edu

Contents

List of Abbreviations	XXXI
1 Introduction to Speech Processing	
<i>J. Benesty, M. M. Sondhi, Y. Huang</i>	1
1.1 A Brief History of Speech Processing	1
1.2 Applications of Speech Processing	2
1.3 Organization of the Handbook	4
References	4
Part A Production, Perception, and Modeling of Speech	
2 Physiological Processes of Speech Production	
<i>K. Honda</i>	7
2.1 Overview of Speech Apparatus	7
2.2 Voice Production Mechanisms	8
2.3 Articulatory Mechanisms	14
2.4 Summary	24
References	25
3 Nonlinear Cochlear Signal Processing and Masking in Speech Perception	
<i>J. B. Allen</i>	27
3.1 Basics	27
3.2 The Nonlinear Cochlea	35
3.3 Neural Masking	45
3.4 Discussion and Summary	55
References	56
4 Perception of Speech and Sound	
<i>B. Kollmeier, T. Brand, B. Meyer</i>	61
4.1 Basic Psychoacoustic Quantities	62
4.2 Acoustical Information Required for Speech Perception	70
4.3 Speech Feature Perception	74
References	81
5 Speech Quality Assessment	
<i>V. Grancharov, W. B. Kleijn</i>	83
5.1 Degradation Factors Affecting Speech Quality	84
5.2 Subjective Tests	85
5.3 Objective Measures	90
5.4 Conclusions	95
References	96

Part B Signal Processing for Speech

6 Wiener and Adaptive Filters

<i>J. Benesty, Y. Huang, J. Chen</i>	103
6.1 Overview.....	103
6.2 Signal Models.....	104
6.3 Derivation of the Wiener Filter.....	106
6.4 Impulse Response Tail Effect.....	107
6.5 Condition Number.....	108
6.6 Adaptive Algorithms.....	110
6.7 MIMO Wiener Filter.....	116
6.8 Conclusions.....	119
References	120

7 Linear Prediction

<i>J. Benesty, J. Chen, Y. Huang</i>	121
7.1 Fundamentals.....	121
7.2 Forward Linear Prediction.....	122
7.3 Backward Linear Prediction.....	123
7.4 Levinson–Durbin Algorithm.....	124
7.5 Lattice Predictor.....	126
7.6 Spectral Representation.....	127
7.7 Linear Interpolation.....	128
7.8 Line Spectrum Pair Representation.....	129
7.9 Multichannel Linear Prediction.....	130
7.10 Conclusions.....	133
References	133

8 The Kalman Filter

<i>S. Gannot, A. Yeredor</i>	135
8.1 Derivation of the Kalman Filter.....	136
8.2 Examples: Estimation of Parametric Stochastic Process from Noisy Observations.....	141
8.3 Extensions of the Kalman Filter.....	144
8.4 The Application of the Kalman Filter to Speech Processing.....	149
8.5 Summary.....	157
References	157

9 Homomorphic Systems and Cepstrum Analysis of Speech

<i>R. W. Schafer</i>	161
9.1 Definitions.....	161
9.2 Z-Transform Analysis.....	164
9.3 Discrete-Time Model for Speech Production.....	165
9.4 The Cepstrum of Speech.....	166
9.5 Relation to LPC.....	169
9.6 Application to Pitch Detection.....	171

9.7	Applications to Analysis/Synthesis Coding	172
9.8	Applications to Speech Pattern Recognition	176
9.9	Summary	180
	References	180
10	Pitch and Voicing Determination of Speech with an Extension Toward Music Signals	
	<i>W. J. Hess</i>	181
10.1	Pitch in Time-Variant Quasiperiodic Acoustic Signals	182
10.2	Short-Term Analysis PDAs	185
10.3	Selected Time-Domain Methods	192
10.4	A Short Look into Voicing Determination	195
10.5	Evaluation and Postprocessing	197
10.6	Applications in Speech and Music	201
10.7	Some New Challenges and Developments	203
10.8	Concluding Remarks	207
	References	208
11	Formant Estimation and Tracking	
	<i>D. O'Shaughnessy</i>	213
11.1	Historical	213
11.2	Vocal Tract Resonances	215
11.3	Speech Production	216
11.4	Acoustics of the Vocal Tract	218
11.5	Short-Time Speech Analysis	221
11.6	Formant Estimation	223
11.7	Summary	226
	References	226
12	The STFT, Sinusoidal Models, and Speech Modification	
	<i>M. M. Goodwin</i>	229
12.1	The Short-Time Fourier Transform	230
12.2	Sinusoidal Models	242
12.3	Speech Modification	253
	References	256
13	Adaptive Blind Multichannel Identification	
	<i>Y. Huang, J. Benesty, J. Chen</i>	259
13.1	Overview	259
13.2	Signal Model and Problem Formulation	260
13.3	Identifiability and Principle	261
13.4	Constrained Time-Domain Multichannel LMS and Newton Algorithms	262
13.5	Unconstrained Multichannel LMS Algorithm with Optimal Step-Size Control	266
13.6	Frequency-Domain Blind Multichannel Identification Algorithms	268
13.7	Adaptive Multichannel Exponentiated Gradient Algorithm	276

13.8 Summary	279
References	279

Part C Speech Coding

14 Principles of Speech Coding

<i>W. B. Kleijn</i>	283
14.1 The Objective of Speech Coding	283
14.2 Speech Coder Attributes	284
14.3 A Universal Coder for Speech	286
14.4 Coding with Autoregressive Models	293
14.5 Distortion Measures and Coding Architecture	296
14.6 Summary	302
References	303

15 Voice over IP: Speech Transmission over Packet Networks

<i>J. Skoglund, E. Kozica, J. Linden, R. Hagen, W. B. Kleijn</i>	307
15.1 Voice Communication	307
15.2 Properties of the Network	308
15.3 Outline of a VoIP System	313
15.4 Robust Encoding	317
15.5 Packet Loss Concealment	326
15.6 Conclusion	327
References	328

16 Low-Bit-Rate Speech Coding

<i>A. V. McCree</i>	331
16.1 Speech Coding	331
16.2 Fundamentals: Parametric Modeling of Speech Signals	332
16.3 Flexible Parametric Models	337
16.4 Efficient Quantization of Model Parameters	344
16.5 Low-Rate Speech Coding Standards	345
16.6 Summary	347
References	347

17 Analysis-by-Synthesis Speech Coding

<i>J.-H. Chen, J. Thyssen</i>	351
17.1 Overview	352
17.2 Basic Concepts of Analysis-by-Synthesis Coding	353
17.3 Overview of Prominent Analysis-by-Synthesis Speech Coders	357
17.4 Multipulse Linear Predictive Coding (MPLPC)	360
17.5 Regular-Pulse Excitation with Long-Term Prediction (RPE-LTP)	362
17.6 The Original Code Excited Linear Prediction (CELP) Coder	363
17.7 US Federal Standard FS1016 CELP	367
17.8 Vector Sum Excited Linear Prediction (VSELP)	368
17.9 Low-Delay CELP (LD-CELP)	370

17.10 Pitch Synchronous Innovation CELP (PSI-CELP)	371
17.11 Algebraic CELP (ACELP)	371
17.12 Conjugate Structure CELP (CS-CELP) and CS-ACELP	377
17.13 Relaxed CELP (RCELP) – Generalized Analysis by Synthesis	378
17.14 eX-CELP	381
17.15 iLBC	382
17.16 TSNFC	383
17.17 Embedded CELP	386
17.18 Summary of Analysis-by-Synthesis Speech Coders	388
17.19 Conclusion	390
References	390

18 Perceptual Audio Coding of Speech Signals

<i>J. Herre, M. Lutzky</i>	393
18.1 History of Audio Coding	393
18.2 Fundamentals of Perceptual Audio Coding	394
18.3 Some Successful Standardized Audio Coders	396
18.4 Perceptual Audio Coding for Real-Time Communication	398
18.5 Hybrid/Crossover Coders	403
18.6 Summary	409
References	409

Part D Text-to-Speech Synthesis

19 Basic Principles of Speech Synthesis

<i>J. Schroeter</i>	413
19.1 The Basic Components of a TTS System	413
19.2 Speech Representations and Signal Processing for Concatenative Synthesis	421
19.3 Speech Signal Transformation Principles	423
19.4 Speech Synthesis Evaluation	425
19.5 Conclusions	426
References	426

20 Rule-Based Speech Synthesis

<i>R. Carlson, B. Granström</i>	429
20.1 Background	429
20.2 Terminal Analog	429
20.3 Controlling the Synthesizer	432
20.4 Special Applications of Rule-Based Parametric Synthesis	434
20.5 Concluding Remarks	434
References	434

21 Corpus-Based Speech Synthesis

<i>T. Dutoit</i>	437
21.1 Basics	437

21.2	Concatenative Synthesis with a Fixed Inventory	438
21.3	Unit-Selection-Based Synthesis	447
21.4	Statistical Parametric Synthesis	450
21.5	Conclusion	453
	References	453
22	Linguistic Processing for Speech Synthesis	
	<i>R. Sproat</i>	457
22.1	Why Linguistic Processing is Hard	457
22.2	Fundamentals: Writing Systems and the Graphical Representation of Language	457
22.3	Problems to be Solved and Methods to Solve Them	458
22.4	Architectures for Multilingual Linguistic Processing	465
22.5	Document-Level Processing	465
22.6	Future Prospects	466
	References	467
23	Prosodic Processing	
	<i>J. van Santen, T. Mishra, E. Klabbers</i>	471
23.1	Overview	471
23.2	Historical Overview	475
23.3	Fundamental Challenges	476
23.4	A Survey of Current Approaches	477
23.5	Future Approaches	484
23.6	Conclusions	485
	References	485
24	Voice Transformation	
	<i>Y. Stylianou</i>	489
24.1	Background	489
24.2	Source-Filter Theory and Harmonic Models	490
24.3	Definitions	492
24.4	Source Modifications	494
24.5	Filter Modifications	498
24.6	Conversion Functions	499
24.7	Voice Conversion	500
24.8	Quality Issues in Voice Transformations	501
24.9	Summary	502
	References	502
25	Expressive/Affective Speech Synthesis	
	<i>N. Campbell</i>	505
25.1	Overview	505
25.2	Characteristics of Affective Speech	506
25.3	The Communicative Functionality of Speech	508
25.4	Approaches to Synthesizing Expressive Speech	510
25.5	Modeling Human Speech	512

25.6 Conclusion	515
References	515

Part E Speech Recognition

26 Historical Perspective of the Field of ASR/NLU	
<i>L. Rabiner, B.-H. Juang</i>	521
26.1 ASR Methodologies	521
26.2 Important Milestones in Speech Recognition History	523
26.3 Generation 1 – The Early History of Speech Recognition	524
26.4 Generation 2 – The First Working Systems for Speech Recognition....	524
26.5 Generation 3 – The Pattern Recognition Approach to Speech Recognition	525
26.6 Generation 4 – The Era of the Statistical Model.....	530
26.7 Generation 5 – The Future	534
26.8 Summary	534
References	535
27 HMMs and Related Speech Recognition Technologies	
<i>S. Young</i>	539
27.1 Basic Framework	539
27.2 Architecture of an HMM-Based Recognizer.....	540
27.3 HMM-Based Acoustic Modeling	547
27.4 Normalization	550
27.5 Adaptation.....	551
27.6 Multipass Recognition Architectures	554
27.7 Conclusions.....	554
References	555
28 Speech Recognition with Weighted Finite-State Transducers	
<i>M. Mohri, F. Pereira, M. Riley</i>	559
28.1 Definitions	559
28.2 Overview.....	560
28.3 Algorithms	567
28.4 Applications to Speech Recognition	574
28.5 Conclusion	582
References	582
29 A Machine Learning Framework for Spoken-Dialog Classification	
<i>C. Cortes, P. Haffner, M. Mohri</i>	585
29.1 Motivation	585
29.2 Introduction to Kernel Methods	586
29.3 Rational Kernels.....	587
29.4 Algorithms	589
29.5 Experiments.....	591
29.6 Theoretical Results for Rational Kernels	593

29.7	Conclusion	594
	References	595
30	Towards Superhuman Speech Recognition	
	<i>M. Picheny, D. Nahamoo</i>	597
30.1	Current Status	597
30.2	A Multidomain Conversational Test Set	598
30.3	Listening Experiments	599
30.4	Recognition Experiments	601
30.5	Speculation	607
	References	614
31	Natural Language Understanding	
	<i>S. Roukos</i>	617
31.1	Overview of NLU Applications	618
31.2	Natural Language Parsing	620
31.3	Practical Implementation	623
31.4	Speech Mining	623
31.5	Conclusion	625
	References	626
32	Transcription and Distillation of Spontaneous Speech	
	<i>S. Furui, T. Kawahara</i>	627
32.1	Background	627
32.2	Overview of Research Activities on Spontaneous Speech	628
32.3	Analysis for Spontaneous Speech Recognition	632
32.4	Approaches to Spontaneous Speech Recognition	635
32.5	Metadata and Structure Extraction of Spontaneous Speech	640
32.6	Speech Summarization	644
32.7	Conclusions	647
	References	647
33	Environmental Robustness	
	<i>J. Droppo, A. Acero</i>	653
33.1	Noise Robust Speech Recognition	653
33.2	Model Retraining and Adaptation	656
33.3	Feature Transformation and Normalization	657
33.4	A Model of the Environment	664
33.5	Structured Model Adaptation	667
33.6	Structured Feature Enhancement	671
33.7	Unifying Model and Feature Techniques	675
33.8	Conclusion	677
	References	677
34	The Business of Speech Technologies	
	<i>J. Wilpon, M. E. Gilbert, J. Cohen</i>	681
34.1	Introduction	682

34.2	Network-Based Speech Services	686
34.3	Device-Based Speech Applications	692
34.4	Vision/Predictions of Future Services – Fueling the Trends	697
34.5	Conclusion	701
	References	702

35 Spoken Dialogue Systems

	<i>V. Zue, S. Seneff</i>	705
35.1	Technology Components and System Development	707
35.2	Development Issues	712
35.3	Historical Perspectives	714
35.4	New Directions	715
35.5	Concluding Remarks	718
	References	718

Part F Speaker Recognition

36 Overview of Speaker Recognition

	<i>A. E. Rosenberg, F. Bimbot, S. Parthasarathy</i>	725
36.1	Speaker Recognition	725
36.2	Measuring Speaker Features	729
36.3	Constructing Speaker Models	731
36.4	Adaptation	735
36.5	Decision and Performance	735
36.6	Selected Applications for Automatic Speaker Recognition	737
36.7	Summary	739
	References	739

37 Text-Dependent Speaker Recognition

	<i>M. Hébert</i>	743
37.1	Brief Overview	743
37.2	Text-Dependent Challenges	747
37.3	Selected Results	750
37.4	Concluding Remarks	760
	References	760

38 Text-Independent Speaker Recognition

	<i>D. A. Reynolds, W. M. Campbell</i>	763
38.1	Introduction	763
38.2	Likelihood Ratio Detector	764
38.3	Features	766
38.4	Classifiers	767
38.5	Performance Assessment	776
38.6	Summary	778
	References	779

Part G Language Recognition

39 Principles of Spoken Language Recognition

<i>C.-H. Lee</i>	785
39.1 Spoken Language	785
39.2 Language Recognition Principles	786
39.3 Phone Recognition Followed by Language Modeling (PRLM)	788
39.4 Vector-Space Characterization (VSC)	789
39.5 Spoken Language Verification	790
39.6 Discriminative Classifier Design	791
39.7 Summary	793
References	793

40 Spoken Language Characterization

<i>M. P. Harper, M. Maxwell</i>	797
40.1 Language versus Dialect	798
40.2 Spoken Language Collections	800
40.3 Spoken Language Characteristics	800
40.4 Human Language Identification	804
40.5 Text as a Source of Information on Spoken Languages	806
40.6 Summary	807
References	807

41 Automatic Language Recognition Via Spectral and Token Based Approaches

<i>D. A. Reynolds, W. M. Campbell, W. Shen, E. Singer</i>	811
41.1 Automatic Language Recognition	811
41.2 Spectral Based Methods	812
41.3 Token-Based Methods	815
41.4 System Fusion	818
41.5 Performance Assessment	820
41.6 Summary	823
References	823

42 Vector-Based Spoken Language Classification

<i>H. Li, B. Ma, C.-H. Lee</i>	825
42.1 Vector Space Characterization	826
42.2 Unit Selection and Modeling	827
42.3 Front-End: Voice Tokenization and Spoken Document Vectorization	830
42.4 Back-End: Vector-Based Classifier Design	831
42.5 Language Classification Experiments and Discussion	835
42.6 Summary	838
References	839

Part H Speech Enhancement

43 Fundamentals of Noise Reduction

<i>J. Chen, J. Benesty, Y. Huang, E. J. Diethorn</i>	843
43.1 Noise	843
43.2 Signal Model and Problem Formulation	845
43.3 Evaluation of Noise Reduction	846
43.4 Noise Reduction via Filtering Techniques	847
43.5 Noise Reduction via Spectral Restoration	857
43.6 Speech-Model-Based Noise Reduction	863
43.7 Summary	868
References	869

44 Spectral Enhancement Methods

<i>I. Cohen, S. Gannot</i>	873
44.1 Spectral Enhancement	874
44.2 Problem Formulation	875
44.3 Statistical Models	876
44.4 Signal Estimation	879
44.5 Signal Presence Probability Estimation	881
44.6 A Priori SNR Estimation	882
44.7 Noise Spectrum Estimation	888
44.8 Summary of a Spectral Enhancement Algorithm	891
44.9 Selection of Spectral Enhancement Algorithms	896
44.10 Conclusions	898
References	899

45 Adaptive Echo Cancellation for Voice Signals

<i>M. M. Sondhi</i>	903
45.1 Network Echoes	904
45.2 Single-Channel Acoustic Echo Cancellation	915
45.3 Multichannel Acoustic Echo Cancellation	921
45.4 Summary	925
References	926

46 Dereverberation

<i>Y. Huang, J. Benesty, J. Chen</i>	929
46.1 Background and Overview	929
46.2 Signal Model and Problem Formulation	931
46.3 Source Model-Based Speech Dereverberation	932
46.4 Separation of Speech and Reverberation via Homomorphic Transformation	936
46.5 Channel Inversion and Equalization	937
46.6 Summary	941
References	942

47 Adaptive Beamforming and Postfiltering
S. Gannot, I. Cohen..... 945

47.1 Problem Formulation..... 947

47.2 Adaptive Beamforming..... 948

47.3 Fixed Beamformer and Blocking Matrix..... 953

47.4 Identification of the Acoustical Transfer Function..... 955

47.5 Robustness and Distortion Weighting..... 960

47.6 Multichannel Postfiltering..... 962

47.7 Performance Analysis..... 967

47.8 Experimental Results..... 972

47.9 Summary..... 972

47.A Appendix: Derivation of the Expected Noise Reduction
for a Coherent Noise Field..... 973

47.B Appendix: Equivalence Between Maximum SNR
and LCMV Beamformers..... 974

References..... 975

48 Feedback Control in Hearing Aids
A. Spriet, S. Doclo, M. Moonen, J. Wouters..... 979

48.1 Problem Statement..... 980

48.2 Standard Adaptive Feedback Canceller..... 982

48.3 Feedback Cancellation Based on Prior Knowledge
of the Acoustic Feedback Path..... 986

48.4 Feedback Cancellation Based on Closed-Loop System Identification..... 990

48.5 Comparison..... 995

48.6 Conclusions..... 997

References..... 997

49 Active Noise Control
S. M. Kuo, D. R. Morgan..... 1001

49.1 Broadband Feedforward Active Noise Control..... 1002

49.2 Narrowband Feedforward Active Noise Control..... 1006

49.3 Feedback Active Noise Control..... 1010

49.4 Multichannel ANC..... 1011

49.5 Summary..... 1015

References..... 1015

Part I Multichannel Speech Processing

50 Microphone Arrays
G. W. Elko, J. Meyer..... 1021

50.1 Microphone Array Beamforming..... 1021

50.2 Constant-Beamwidth Microphone Array System..... 1029

50.3 Constrained Optimization of the Directional Gain..... 1030

50.4 Differential Microphone Arrays..... 1031

50.5 Eigenbeamforming Arrays..... 1034

50.6	Adaptive Array Systems	1037
50.7	Conclusions	1040
	References	1040
51	Time Delay Estimation and Source Localization	
	<i>Y. Huang, J. Benesty, J. Chen</i>	1043
51.1	Technology Taxonomy	1043
51.2	Time Delay Estimation	1044
51.3	Source Localization	1054
51.4	Summary	1061
	References	1062
52	Convolutional Blind Source Separation Methods	
	<i>M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra</i>	1065
52.1	The Mixing Model	1066
52.2	The Separation Model	1068
52.3	Identification	1071
52.4	Separation Principle	1071
52.5	Time Versus Frequency Domain	1076
52.6	The Permutation Ambiguity	1078
52.7	Results	1084
52.8	Conclusion	1084
	References	1084
53	Sound Field Reproduction	
	<i>R. Rabenstein, S. Spors</i>	1095
53.1	Sound Field Synthesis	1095
53.2	Mathematical Representation of Sound Fields	1096
53.3	Stereophony	1100
53.4	Vector-Based Amplitude Panning	1103
53.5	Ambisonics	1104
53.6	Wave Field Synthesis	1109
	References	1113
	Acknowledgements	1115
	About the Authors	1117
	Detailed Contents	1133
	Subject Index	1161

47. Adaptive Beamforming and Postfiltering

S. Gannot, I. Cohen

In this chapter, we explore many of the basic concepts of array processing with an emphasis on adaptive beamforming for speech enhancement applications. We begin in Sect. 47.1 by formulating the problem of microphone array in a noisy and reverberant environment. In Sect. 47.2, we derive the frequency-domain linearly constrained minimum-variance (LCMV) beamformer, and its generalized sidelobe canceller (GSC) variant. The GSC components are explored in Sect. 47.3, and several commonly used special cases of these blocks are presented. As the GSC structure necessitates an estimation of the speech related acoustical transfer functions (ATFs), several alternative system identification methods are addressed in Sect. 47.4. Beamformers often suffer from sensitivity to signal mismatch. We analyze this phenomenon in Sect. 47.5 and explore several cures to this problem. Although the GSC beamformer yields a significant improvement in speech quality, when the noise field is spatially incoherent or diffuse, the noise reduction is insufficient and additional postfiltering is normally required. In Sect. 47.6, we present multi-microphone postfilters, based on either minimum mean-squared error (MMSE) or log-spectral amplitude estimate criteria. An interesting relation between the GSC and the Wiener filter is derived in this Section as well. In Sect. 47.7, we analyze the performance of the transfer-function GSC (TF-GSC), and in Sect. 47.8 demonstrate the advantage of multichannel postfiltering over single-channel postfiltering in nonstationary noise conditions.

47.1	Problem Formulation	947
47.2	Adaptive Beamforming	948
47.2.1	Frequency-Domain Frost Algorithm	948

47.2.2	Frequency-Domain Generalized Sidelobe Canceller	950
47.2.3	Time-Domain Generalized Sidelobe Canceller	952
47.3	Fixed Beamformer and Blocking Matrix ..	953
47.3.1	Using Acoustical Transfer Functions	953
47.3.2	Using Delay-Only Filters	954
47.3.3	Using Relative Transfer Functions ..	954
47.4	Identification of the Acoustical Transfer Function	955
47.4.1	Signal Subspace Method	955
47.4.2	Time Difference of Arrival	956
47.4.3	Relative Transfer Function Estimation	956
47.5	Robustness and Distortion Weighting	960
47.6	Multichannel Postfiltering	962
47.6.1	MMSE Postfiltering	963
47.6.2	Log-Spectral Amplitude Postfiltering	964
47.7	Performance Analysis	967
47.7.1	The Power Spectral Density of the Beamformer Output	967
47.7.2	Signal Distortion	968
47.7.3	Stationary Noise Reduction	968
47.8	Experimental Results	972
47.9	Summary	972
47.A	Appendix: Derivation of the Expected Noise Reduction for a Coherent Noise Field	973
47.B	Appendix: Equivalence Between Maximum SNR and LCMV Beamformers	974
	References	975

Over the last four decades, array processing has become a well-established discipline, see e.g., [47.1–14]. In the mid 1980s, array processing and beamforming methods were adopted by the speech community to deal with data received by microphone arrays. Since then, beamforming techniques for microphone arrays have been used in many applications, such as speaker separation, speaker localization, speech dereverberation, acoustic echo cancellation, and speech enhancement.

Adaptive beamforming for speech signals requires particular consideration of problems that are specific to speech signals and to the acoustic environment. The speech signal is wide-band, highly nonstationary, and has a very wide dynamic range. An acoustic enclosure is usually modeled as a filter with very long impulse response due to multiple reflections from the room walls. In a typical office, the length of the filters may reach several thousand taps. Furthermore, the impulse response is often time varying due to speaker and objects movements.

The term beamforming refers to the design of a spatiotemporal filter. Broadband arrays comprise a set of filters, applied to each received microphone signal, followed by a summation operation. The main objective of the beamformer is to extract a desired signal, impinging on the array from a specific position, out of noisy measurements thereof. Usually, the interference signals occupy the same frequency band as the desired signal, rendering temporal-only filtering useless. The simplest structure is the delay-and-sum beamformer, which first compensates for the relative delay between distinct microphone signals and then sums the steered signal to form a single output. This beamformer, which is still widely used, can be very effective in mitigating noncoherent, i.e., spatially white, noise sources, provided that the number of microphones is relatively high. However, if the noise source is coherent, the noise reduction (NR) is strongly dependent on the direction of arrival of the noise signal. Consequently, the performance of the delay and sum beamformer in reverberant environments is often insufficient. Jan and Flanagan [47.15, 16] and Rabinkin et al. [47.17] extended the delay and sum concept by introducing the filter-and-sum beamformer. This structure, designed for multipath environments, namely reverberant enclosures, replaces the simpler delay compensator with a matched filter.

The array beam pattern can generally be designed to have a specified response. This can be done by properly setting the values of the multichannel filters' weights. However, the application of data-independent design

methods is very limited in dynamic acoustical environments. Statistically optimal beamformers are designed based on the statistical properties of the desired and interference signals. In general, they aim at enhancing the desired signal, while rejecting the interference signal. Several criteria can be applied in the design of the beamformer, e.g., maximum signal-to-noise ratio (MSNR), minimum mean-squared error (MMSE), and linearly constrained minimum variance (LCMV). A summary of several design criteria can be found in [47.5, 7].

Beamforming methods use the signals' statistics (at least second-order statistics), which is usually not available and should be estimated from the data. Moreover, the acoustical environment is time varying, due to talker and objects movements, and abrupt changes in the noise characteristics (e.g., passing cars). Hence, adaptation mechanisms are required. An adaptive counterpart of each of the prespecified design criteria can be derived. Early contributions to the field of adaptive beamformers design can be attributed to Sondhi and Elko [47.18], to Kaneda and Ohga [47.19], and to Van Compernelle [47.20]. Kellermann [47.21] addressed the problem of joint echo cancellation and NR by incorporating echo cancellers into the beamformer design. Nordholm et al. [47.22, 23] used microphone arrays in a car environment, and designed a beamformer employing calibration signals to enhance the obtained performance. Martin [47.24] analyzed beamforming techniques for small microphone arrays. Many other applications of microphone arrays such as hearing aids, blind source separation (BSS), and dereverberation are addressed elsewhere in this handbook.

The minimization of the mean-squared error (MSE) in the context of array processing leads to the well-known multichannel Wiener filter [47.25]. Doclo and Moonen [47.26–28] proposed an efficient implementation of the Wiener filter based on the generalized singular-value decomposition (GSVD) of the microphone data matrix. This method yields an optimal estimation (in the MMSE sense) of the desired signal component of one of the microphone signals. The authors further proposed efficient schemes for recursive update of the GSVD. An optional, adaptive noise cancellation postfiltering stage is proposed as well. In that scheme, in addition to the optimal estimation of the desired speech signal, an optimal noise channel is also estimated. This estimated noise component can be used as a reference noise signal (similar to the one used in [47.25]), to further enhance the speech signal. Spriet et al. [47.29] proposed a subband implementation of the GSVD-based scheme, and Rombouts and

Moonen [47.30, 31] proposed to apply the efficient QR decomposition to the problem at hand.

In many adaptive array schemes the acoustical transfer-function (ATF) relating the speech source and the microphone should be known in advance, or at least estimated from the received data (note that in case of delay-only propagation, the acoustical transfer function reduces to a steering vector, consisting of phase-only components.) In contrast, the multichannel Wiener filter is uniquely based on estimates of the second-order statistics of the recorded noisy signal and the noise signal (estimated during noise-only segments), and does not make any a priori assumptions about the signal model. Unfortunately, as pointed by *Chen* et al. [47.32], the Wiener filter, which is optimal in the MMSE sense, cannot guarantee undistorted speech signal at its output. This drawback can however be mitigated by modifying the MMSE criterion to control the amount of imposed speech distortion. A method that employs this modification is presented in [47.33, 34]. It is also shown that the ATFs information (only a simple delay-only case is presented in the contributions) can be incorporated into the Wiener filter scheme (called the spatially preprocessed Wiener filter), resulting in an improved performance. The Wiener filter and its application to speech enhancement is addressed in a separate chapter of this handbook (6; 43).

In this chapter, we concentrate on a different adaptive structure based on the LCMV criterion. The LCMV beamformer, proposed by *Frost* [47.35], is aiming at minimizing the output power under linear constraints on the response of the array towards the desired speech signal. *Frost* proposed an adaptive scheme, which is based on a constrained least-mean-square (LMS)-type adaptation (for the LMS algorithm please refer to [47.25]). To avoid this constrained adaptation, *Griffiths* and *Jim* [47.36] proposed the GSC structure, which separates the output power minimization and the application of the constraint. The GSC structure is based on the assumption that the different sensors receive a delayed version of the desired signal, and therefore we refer to it as the

delay generalized sidelobe canceller (D-GSC). The GSC structure was rederived in the frequency domain, and extended to deal with, the more-complicated general ATFs case by *Affes* and *Grenier* [47.37] and later by *Gannot* et al. [47.38]. This frequency-domain version, which takes into account the reverberant nature of the enclosure, was nicknamed the transfer-function generalized sidelobe canceller (TF-GSC). The GSC comprises three blocks: a fixed beamformer (FBF), which aligns the desired signal components, a blocking matrix (BM), which blocks the desired speech components resulting in reference noise signals, and a multichannel adaptive noise canceller (ANC), which eliminates noise components that leak through the sidelobes of the FBF.

Nordholm and *Leung* [47.39] analyze the limits of the obtainable NR of the GSC in an isotropic noise field. *Bitzer* et al. address the problem in [47.40, 41] and [47.42]. In [47.40], the authors derive an expression for the NR as a function of the noise field and evaluate the degradation as a function of the reverberation time (T_{60}). The special two-microphone case is treated in [47.41]. The additional NR due to the ANC branch of the GSC, implemented by a closed-form Wiener filter rather than the adaptive Widrow least-mean-square (LMS) procedure, is presented in [47.42]. The frequency-band nested subarrays structure is presented and its NR is theoretically analyzed by *Marro* et al. [47.43]. A more-complex dual GSC structure employing calibration signals was suggested and analyzed by *Nordholm* et al. [47.44]. *Huang* and *Yeh* [47.45] addressed the distortion issue by evaluating the desired signal leakage into the reference noise branch of the GSC structure. However, the delay-only ATFs assumption is imposed and the expected degradation due to pointing errors alone is evaluated. The performance degradation due to constraining the Wiener filters to a finite impulse response (FIR) structure is demonstrated by *Nordholm* et al. in [47.46]. The resulting performance limits of the GSC structure strongly depend on the cross-correlation between the sensors' signals induced by the noise field, as shown in the above references and by *Cox* [47.47].

47.1 Problem Formulation

Consider an array of M sensors in a noisy and reverberant environment. The received signals generally include three components. The first is a desired speech signal, the second is some stationary interference signal, and the third is some nonstationary (transient) noise component.

Our goal is to reconstruct the speech component from the received signals. Let $s(t)$ denote the desired source signal, let $a_m(t)$ represent the room impulse response (RIR) of the m -th sensor to the desired source, and let $n_m(t)$ denote the noise component at the m -th sensor.

The observed signal at the m -th sensor ($m = 1, \dots, M$) is given by

$$\begin{aligned} z_m(t) &= a_m(t) * s(t) + n_m(t) \\ &= a_m(t) * s(t) + n_m^s(t) + n_m^t(t), \end{aligned} \quad (47.1)$$

where $n_m^s(t)$ and $n_m^t(t)$ represent the stationary and nonstationary noise components at the m -th sensor, respectively, and $*$ denotes convolution. We assume that both noise components may comprise coherent (directional) noise component and diffused noise component.

The observed signals are divided in time into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Assuming time-invariant transfer functions, we have in the time–frequency domain

$$\begin{aligned} Z_m(k, \ell) &\approx A_m(k)S(k, \ell) + N_m(k, \ell) \\ &\approx A_m(k)S(k, \ell) + N_m^s(k, \ell) + N_m^t(k, \ell), \end{aligned} \quad (47.2)$$

where ℓ is the frame index and $k = 1, 2, \dots, K$ represents the frequency bin index. (The equality in (47.2)

is only justified for segments which are longer than the RIR length. Since RIRs tend to be very long, the conditions allowing for this representation to hold cannot be exactly met. We assume, however, that the STFT relation is a reasonable approximation.) $Z_m(k, \ell)$, $S(k, \ell)$, $N_m(k, \ell)$, $N_m^s(k, \ell)$, and $N_m^t(k, \ell)$ are the STFT of the respective signals. $A_m(k)$ is the ATF relating the speech source with the m -th sensor. The vector formulation of the equation set (47.2) is

$$\begin{aligned} \mathbf{Z}(k, \ell) &= \mathbf{A}(k)S(k, \ell) + \mathbf{N}(k, \ell) \\ &= \mathbf{A}(k)S(k, \ell) + \mathbf{N}_s(k, \ell) + \mathbf{N}_t(k, \ell), \end{aligned} \quad (47.3)$$

where

$$\begin{aligned} \mathbf{Z}(k, \ell) &= (Z_1(k, \ell) \ Z_2(k, \ell) \ \dots \ Z_M(k, \ell))^T, \\ \mathbf{A}(k) &= (A_1(k) \ A_2(k) \ \dots \ A_M(k))^T, \\ \mathbf{N}(k, \ell) &= (N_1(k, \ell) \ N_2(k, \ell) \ \dots \ N_M(k, \ell))^T, \\ \mathbf{N}_s(k, \ell) &= (N_1^s(k, \ell) \ N_2^s(k, \ell) \ \dots \ N_M^s(k, \ell))^T, \\ \mathbf{N}_t(k, \ell) &= (N_1^t(k, \ell) \ N_2^t(k, \ell) \ \dots \ N_M^t(k, \ell))^T. \end{aligned}$$

47.2 Adaptive Beamforming

Frost [47.35] proposed a beamformer that relies on the assumption that the ATFs between the desired source and the array of sensors can be uniquely determined by gain and delay values. In this section, we follow Frost's approach in the STFT domain and derive a beamforming algorithm for the arbitrary ATF case. We first obtain a closed form of the LCMV beamformer, and subsequently derive an adaptive solution. The outcome is a constrained LMS-type algorithm. We proceed, following the seminal work of Griffiths and Jim [47.36], with the formulation of an unconstrained adaptive solution namely, the transfer-function generalized sidelobe canceller (TF-GSC). We initially assume that the ATFs are known. Later, in Sect. 47.4, we present several alternatives for estimating the ATFs.

47.2.1 Frequency-Domain Frost Algorithm

Optimal Solution

Let $W_m^*(k, \ell)$; $m = 1, \dots, M$ denote a set of M filters, and define

$$\mathbf{W}^H(k, \ell) = (W_1^*(k, \ell) \ W_2^*(k, \ell) \ \dots \ W_M^*(k, \ell)),$$

where the superscript H denotes conjugation transpose. A filter-and-sum beamformer, depicted in Fig. 47.1, is realized by filtering each sensor signal by $W_m^*(k, \ell)$ and summing the outputs,

$$\begin{aligned} Y(k, \ell) &= \mathbf{W}^H(k, \ell)\mathbf{Z}(k, \ell) \\ &= \mathbf{W}^H(k, \ell)\mathbf{A}(k)S(k, \ell) + \mathbf{W}^H(k, \ell)\mathbf{N}_s(k, \ell) \\ &\quad + \mathbf{W}^H(k, \ell)\mathbf{N}_t(k, \ell) \\ &\triangleq Y_s(k, \ell) + Y_{n,s}(k, \ell) + Y_{n,t}(k, \ell), \end{aligned} \quad (47.4)$$

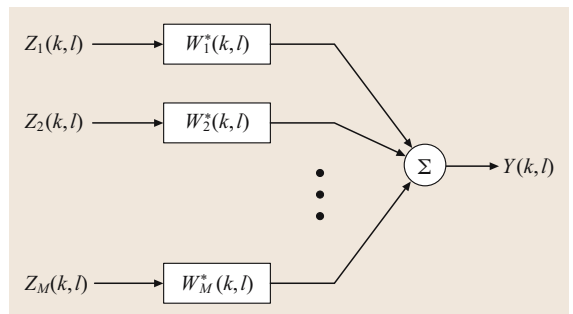


Fig. 47.1 Filter-and-sum beamformer

where $Y_s(k, \ell)$ is the signal component and $Y_{n,s}(k, \ell)$ and $Y_{n,t}(k, \ell)$ are the stationary and nonstationary noise components, respectively. The output power of the beamformer is given by

$$\begin{aligned} & E\{Y(k, \ell)Y^*(k, \ell)\} \\ &= E\{\mathbf{W}^H(k, \ell)\mathbf{Z}(k, \ell)\mathbf{Z}^H(k, \ell)\mathbf{W}(k, \ell)\} \\ &= \mathbf{W}^H(k, \ell)\boldsymbol{\Phi}_{ZZ}(k, \ell)\mathbf{W}(k, \ell), \end{aligned}$$

where $\boldsymbol{\Phi}_{ZZ}(k, \ell) \triangleq E\{\mathbf{Z}(k, \ell)\mathbf{Z}^H(k, \ell)\}$ is the power spectral density (PSD) matrix of the received signals. We want to minimize the output power subject to the following constraint on $Y_s(k, \ell)$:

$$\begin{aligned} Y_s(k, \ell) &= \mathbf{W}^H(k, \ell)\mathbf{A}(k)S(k, \ell) \\ &= \mathcal{F}^*(k, \ell)S(k, \ell), \end{aligned}$$

where $\mathcal{F}(k, \ell)$ is some prespecified filter, usually a simple delay. Without loss of generality we assume hereinafter that $\mathcal{F}(k, \ell) = 1$. Hence, the minimization problem can be stated as

$$\begin{aligned} & \min_{\mathbf{W}} \{ \mathbf{W}^H(k, \ell)\boldsymbol{\Phi}_{ZZ}(k, \ell)\mathbf{W}(k, \ell) \} \\ & \text{subject to } \mathbf{W}^H(k, \ell)\mathbf{A}(k) = 1. \end{aligned} \quad (47.5)$$

The minimization problem (47.5) is demonstrated in Fig. 47.2. The point where the equipower contours are tangent to the constraint plane is the optimum vector of beamforming filters. The perpendicular $\mathbf{F}(k)$ from the origin to the constraint plane will be calculated in the next section.

To solve (47.5) we first define the complex Lagrangian,

$$\begin{aligned} \mathcal{L}(\mathbf{W}) &= \mathbf{W}^H(k, \ell)\boldsymbol{\Phi}_{ZZ}(k, \ell)\mathbf{W}(k, \ell) \\ &+ \lambda[\mathbf{W}^H(k, \ell)\mathbf{A}(k) - 1] \\ &+ \lambda^*[\mathbf{A}^H(k)\mathbf{W}(k, \ell) - 1], \end{aligned} \quad (47.6)$$

where λ is a Lagrange multiplier. Setting the derivative with respect to $\mathbf{W}^*$ to 0 [47.48] yields

$$\nabla_{\mathbf{W}^*} \mathcal{L}(\mathbf{W})\boldsymbol{\Phi}_{ZZ}(k, \ell)\mathbf{W}(k, \ell) + \lambda\mathbf{A}(k) = 0.$$

Now, recalling the constraint in (47.5), we obtain the **LCMV** optimal filter

$$\mathbf{W}^{\text{LCMV}}(k, \ell) = \frac{\boldsymbol{\Phi}_{ZZ}^{-1}(k, \ell)\mathbf{A}(k)}{\mathbf{A}^H(k)\boldsymbol{\Phi}_{ZZ}^{-1}(k, \ell)\mathbf{A}(k)}. \quad (47.7)$$

This closed-form solution is difficult to implement, and is not suitable for time-varying environments. Therefore we often have to resort to an adaptive solution, which is derived in the sequel.

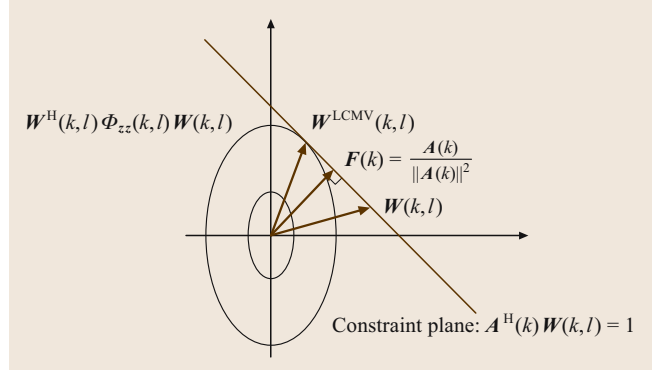


Fig. 47.2 Constrained minimization

It is interesting to show the equivalence between the **LCMV** solution (47.7) and the **MSNR** beamformer [47.7], which is obtained from

$$\max_{\mathbf{W}} \frac{|\mathbf{W}^H(k, \ell)\mathbf{A}(k)|^2}{\mathbf{W}^H(k, \ell)\boldsymbol{\Phi}_{NN}(k, \ell)\mathbf{W}(k, \ell)}. \quad (47.8)$$

The well-known solution to (47.8) is the (colored-noise) matched filter

$$\mathbf{W}(k, \ell) \propto \boldsymbol{\Phi}_{NN}^{-1}(k, \ell)\mathbf{A}(k).$$

If the array response is constrained to fulfil $\mathbf{W}^H(k, \ell)\mathbf{A}(k) = 1$, i. e., no distortion in the desired direction, we have

$$\mathbf{W}^{\text{MSNR}}(k, \ell) = \frac{\boldsymbol{\Phi}_{NN}^{-1}(k, \ell)\mathbf{A}(k)}{\mathbf{A}^H(k)\boldsymbol{\Phi}_{NN}^{-1}(k, \ell)\mathbf{A}(k)}. \quad (47.9)$$

Using (47.3) it can be verified that

$$\begin{aligned} \boldsymbol{\Phi}_{ZZ} &= \phi_{ss}(k, \ell)\mathbf{A}(k)\mathbf{A}^H(k) \\ &+ \boldsymbol{\Phi}_{N_s N_s}(k, \ell) + \boldsymbol{\Phi}_{N_t N_t}(k, \ell) \\ &= \phi_{ss}(k, \ell)\mathbf{A}(k)\mathbf{A}^H(k) + \boldsymbol{\Phi}_{NN}(k, \ell), \end{aligned} \quad (47.10)$$

where $\boldsymbol{\Phi}_{NN}(k, \ell) \triangleq \boldsymbol{\Phi}_{N_s N_s}(k, \ell) + \boldsymbol{\Phi}_{N_t N_t}(k, \ell)$, the overall noise PSD matrix. Using the matrix inversion lemma, it is shown in Appendix 47.B that

$$\mathbf{W}^{\text{LCMV}}(k, \ell) = \frac{\boldsymbol{\Phi}_{NN}^{-1}(k, \ell)\mathbf{A}(k)}{\mathbf{A}^H(k)\boldsymbol{\Phi}_{NN}^{-1}(k, \ell)\mathbf{A}(k)}. \quad (47.11)$$

This solution is identical to the solution of the **MSNR** beamformer.

While both methods are shown to be equal, provided that the **ATFs** $\mathbf{A}(k)$ are known, their behavior in the case of unknown **ATFs** is different. Analysis of these differences is given by Cox [47.47].

Note also that, due to the nonstationary noise component, the term $\Phi_{NN}(k, \ell)$ depends on the frame index. This time dependence is one of the major factors leading to performance degradation in beamforming. We address this problem by introducing the multichannel postfilter in Sect. 47.6.

Adaptive Solution

Consider the following steepest-descent adaptive algorithm:

$$\begin{aligned} \mathbf{W}(k, \ell + 1) &= \mathbf{W}(k, \ell) - \mu \nabla_{\mathbf{W}^*} \mathcal{L}(k, \ell) \\ &= \mathbf{W}(k, \ell) - \mu [\Phi_{ZZ}(k, \ell) \mathbf{W}(k, \ell) + \lambda \mathbf{A}(k)] . \end{aligned}$$

Imposing the look-direction constraint on $\mathbf{W}(\ell + 1, k)$ yields

$$\begin{aligned} 1 &= \mathbf{A}^H(k) \mathbf{W}(k, \ell + 1) \\ &= \mathbf{A}^H(k) \mathbf{W}(k, \ell) - \mu \mathbf{A}^H(k) \Phi_{ZZ}(k, \ell) \mathbf{W}(k, \ell) \\ &\quad - \mu \mathbf{A}^H(k) \mathbf{A}(k) \lambda . \end{aligned}$$

Solving for the Lagrange multiplier and applying further rearrangement of terms yields:

$$\begin{aligned} \mathbf{W}(k, \ell + 1) &= P(k) \mathbf{W}(k, \ell) - \mu P(k) \Phi_{ZZ}(k, \ell) \mathbf{W}(k, \ell) + \mathbf{F}(k) , \\ &\quad (47.12) \end{aligned}$$

where

$$P(k) \triangleq \mathbf{I} - \frac{\mathbf{A}(k) \mathbf{A}^H(k)}{\|\mathbf{A}(k)\|^2} \quad (47.13)$$

and

$$\mathbf{F}(k) \triangleq \frac{\mathbf{A}(k)}{\|\mathbf{A}(k)\|^2} . \quad (47.14)$$

Further simplification can be obtained by replacing $\Phi_{ZZ}(k, \ell)$ by its instantaneous estimator, $\mathbf{Z}(k, \ell) \mathbf{Z}^H(k, \ell)$, and recalling (47.4). We finally obtain,

$$\begin{aligned} \mathbf{W}(k, \ell + 1) &= P(k) [\mathbf{W}(k, \ell) - \mu \mathbf{Z}(k, \ell) \mathbf{Y}^*(k, \ell)] + \mathbf{F}(k) . \end{aligned}$$

The entire algorithm is summarized in Table 47.1.

Table 47.1 Frequency-domain Frost algorithm

$\mathbf{W}(\ell = 0, k) = \mathbf{F}(k)$
$\mathbf{W}(t + 1, k) = P(k) [\mathbf{W}(k, \ell) - \mu \mathbf{Z}(k, \ell) \mathbf{Y}^*(k, \ell)] + \mathbf{F}(k)$
$\ell = 0, 1, \dots$
$[P(k) \text{ and } \mathbf{F}(k) \text{ are defined by (47.13) and (47.14)}].$

47.2.2 Frequency-Domain Generalized Sidelobe Canceller

In [47.36], *Griffiths* and *Jim* considered the case where each ATF is a delay element (with gain). They obtained an unconstrained adaptive enhancement algorithm, using the same linear constraint imposed by *Frost* [47.35]. The unconstrained algorithm is more tractable, reliable, and computationally more efficient in comparison with its constrained counterpart. In the adaptive solution Section, we obtained an adaptive algorithm for the case where each ATF is represented by an arbitrary linear time-invariant system. We now repeat the arguments of Griffiths and Jim for the arbitrary ATFs case, and derive an unconstrained adaptive enhancement algorithm. A detailed description can be found in [47.38].

Derivation

Consider the null space of $\mathbf{A}(k)$, defined by

$$\mathcal{N}(k) \triangleq \{\mathbf{W} \mid \mathbf{A}^H(k) \mathbf{W} = 0\} .$$

The constraint hyperplane,

$$\Lambda(k) \triangleq \{\mathbf{W} \mid \mathbf{A}^H(k) \mathbf{W} = 1\}$$

is parallel to $\mathcal{N}(k)$. In addition, let

$$\mathcal{R}(k) \triangleq \{\kappa \mathbf{A}(k) \mid \text{for any real } \kappa\}$$

be the column space. By the fundamental theorem of linear algebra [47.49], $\mathcal{R}(k) \perp \mathcal{N}(k)$. In particular, $\mathbf{F}(k)$ is perpendicular to $\mathcal{N}(k)$, since $\mathbf{F}(k) = \frac{1}{\|\mathbf{A}(k)\|^2} \mathbf{A}(k) \in \mathcal{R}(k)$. Furthermore,

$$\mathbf{A}^H(k) \mathbf{F}(k) = \mathbf{A}^H(k) \mathbf{A}(k) (\mathbf{A}^H(k) \mathbf{A}(k))^{-1} = 1 .$$

Thus, $\mathbf{F}(k) \in \Lambda(k)$ and $\mathbf{F}(k) \perp \Lambda(k)$. Hence, $\mathbf{F}(k)$ is the perpendicular from the origin to the constraint hyperplane, $\Lambda(k)$. The matrix $P(k)$, defined in (47.13), is the projection matrix to the null space of $\mathbf{A}(k)$, $\mathcal{N}(k)$.

A vector in linear space can be uniquely split into a sum of two vectors in mutually orthogonal subspaces [47.49]. Hence,

$$\mathbf{W}(k, \ell) = \mathbf{W}_0(k, \ell) - \mathbf{V}(k, \ell) , \quad (47.15)$$

where $\mathbf{W}_0(k, \ell) \in \mathcal{R}(k)$ and $-\mathbf{V}(k, \ell) \in \mathcal{N}(k)$. By the definition of $\mathcal{N}(k)$,

$$\mathbf{V}(k, \ell) = \mathcal{H}(k) \mathbf{G}(k, \ell) , \quad (47.16)$$

where $\mathcal{H}(k)$ is a matrix such that its columns span the null space of $\mathbf{A}(k)$, i. e.,

$$\mathbf{A}^H(k) \mathcal{H}(k) = 0 , \quad \text{rank } \{\mathcal{H}(k)\} \leq M - 1 , \quad (47.17)$$

where $\mathcal{H}(k)$ is usually called a **BM** (blocking matrix). The outputs of the **BM** will be denoted, for reasons that will be clear in the sequel, noise reference signals $U(k, \ell)$, defined as

$$U(k, \ell) = \mathcal{H}^H(k) \mathbf{Z}(k, \ell), \quad (47.18)$$

where

$$U(k, \ell) = (U_2(k, \ell) \ U_3(k, \ell) \ \dots \ U_M(k, \ell))^T.$$

The vector $\mathbf{G}(k, \ell)$ is a $\text{rank}\{\mathcal{H}(k)\} \times 1$ vector of adjustable filters. We assume hereinafter that $\text{rank}\{\mathcal{H}(k)\} = M - 1$. Hence, the set of filters is defined as

$$\mathbf{G}(k, \ell) = (G_2(k, \ell) \ G_3(k, \ell) \ \dots \ G_M(k, \ell))^T. \quad (47.19)$$

By the geometrical interpretation of Frost's algorithm,

$$\mathbf{W}_0(k, \ell) = \mathbf{F}(k) = \frac{\mathbf{A}(k)}{\|\mathbf{A}(k)\|^2} \quad (47.20)$$

(Recall that $\mathbf{F}(k)$ is the perpendicular from the origin to the constraint hyperplane, $\mathbf{A}(k)$.) Now, using (47.4), (47.15), and (47.16) we obtain

$$Y(k, \ell) = Y_{\text{FBF}}(k, \ell) - Y_{\text{ANC}}(k, \ell), \quad (47.21)$$

where

$$\begin{aligned} Y_{\text{FBF}}(k, \ell) &= \mathbf{W}_0^H(k, \ell) \mathbf{Z}(k, \ell) \\ Y_{\text{ANC}}(k, \ell) &= \mathbf{G}^H(k, \ell) \mathcal{H}^H(k) \mathbf{Z}(k, \ell). \end{aligned} \quad (47.22)$$

The output of the constrained beamformer is a difference of two terms, both operating on the input signal $\mathbf{Z}(k, \ell)$. The first term, $Y_{\text{FBF}}(k, \ell)$, utilizes only fixed components (which depend on the **ATFs**), so it can be viewed as a **FBF**. The **FBF** coherently sums the desired speech components, while in general it destructively sums the noise components. Hence, it is expected that the signal-to-noise ratio (**SNR**) at the **FBF** output will be higher than the input **SNR**. However, this result cannot be guaranteed. We will elaborate on this issue while discussing the performance analysis in Sect. 47.7.

We now examine the second term $Y_{\text{ANC}}(k, \ell)$. Note that

$$\begin{aligned} U(k, \ell) &= \mathcal{H}^H(k) \mathbf{Z}(k, \ell) \\ &= \mathcal{H}^H(k) [\mathbf{A}(k) S(k, \ell) \\ &\quad + \mathbf{N}_s(k, \ell) + \mathbf{N}_t(k, \ell)] \\ &= \mathcal{H}^H(k) [\mathbf{N}_s(k, \ell) + \mathbf{N}_t(k, \ell)]. \end{aligned} \quad (47.23)$$

The last transition is due to (47.17). It is worth mentioning that, when a perfect **BM** is applied, $U(k, \ell)$ indeed

contains only noise components. In general, however, $\mathcal{H}^H(k) \mathbf{A}(k, \ell) \neq 0$, hence desired speech components may leak into the noise reference signals. If the speech component is indeed completely eliminated (blocked) by $\mathcal{H}(k)$, $Y_{\text{ANC}}(k, \ell)$ becomes a pure noise term. The residual noise term in $Y_{\text{FBF}}(k, \ell)$ can then be reduced by properly adjusting the filters $\mathbf{G}(k, \ell)$, using the minimum output power criterion. This minimization problem is in fact the classical multichannel noise cancellation problem. An adaptive **LMS** solution to the problem was proposed by Widrow [47.25].

To summarize, the beamformer is comprised of three parts. An **FBF** $\mathbf{W}_0$, which aligns the desired signal components, a **BM** $\mathcal{H}(k)$, which blocks the desired speech components resulting in the reference noise signals $U(k, \ell)$, and a multichannel **ANC** $\mathbf{G}(k, \ell)$, which eliminates the stationary noise that leaks through the sidelobes of the **FBF**.

Noise Canceller Adaptation

The reference noise signals are emphasized by the **ANC** and subtracted from the output of the **FBF**, yielding

$$Y(k, \ell) = [\mathbf{W}_0^H(k, \ell) - \mathbf{G}^H(k, \ell) \mathcal{H}^H(k)] \mathbf{Z}(k, \ell). \quad (47.24)$$

Let three hypotheses H_{0s} , H_{0t} , and H_1 indicate, respectively, the absence of transients, the presence of an interfering transient, and the presence of a desired source transient at the beamformer output. The optimal solution for the filters $\mathbf{G}(k, \ell)$ is obtained by minimizing the power of the beamformer output during the stationary noise frames (i. e., when H_{0s} is true) [47.2]. We note, however, that no adaptation should be carried out during abrupt changes in the characteristics of the noise signal (e.g., a passing car). When the noise source position is constant and the noise statistics is slowly varying, the **ANC** filters can track the changes.

Let $\Phi_{N_s N_s}(k, \ell) = E\{\mathbf{N}_s(k, \ell) \mathbf{N}_s^H(k, \ell)\}$ denote the **PSD** matrix of the input stationary noise. Then, the power of the stationary noise at the beamformer output is minimized by solving the unconstrained optimization problem:

$$\begin{aligned} \min_{\mathbf{G}} \{ & [\mathbf{W}_0(k, \ell) - \mathcal{H}(k, \ell) \mathbf{G}(k, \ell)]^H \\ & \times \Phi_{N_s N_s}(k, \ell) [\mathbf{W}_0(k, \ell) - \mathcal{H}(k, \ell) \mathbf{G}(k, \ell)] \}. \end{aligned} \quad (47.25)$$

A multichannel Wiener solution is given by (see also [47.42, 46])

$$\begin{aligned} \mathbf{G}(k, \ell) &= [\mathcal{H}^H(k, \ell) \Phi_{N_s N_s}(k, \ell) \mathcal{H}(k)]^{-1} \\ &\times \mathcal{H}^H(k, \ell) \Phi_{N_s N_s}(k, \ell) \mathbf{W}_0(k, \ell). \end{aligned} \quad (47.26)$$

In practice, this optimization problem is solved by using the normalized LMS algorithm [47.2]:

$$\mathbf{G}(k, \ell + 1) = \begin{cases} \mathbf{G}(k, \ell) + \frac{\mu_g}{P_{\text{est}}(k, \ell)} \mathbf{U}(k, \ell) \mathbf{Y}^*(k, \ell) & H_{0s} \text{ true,} \\ \mathbf{G}(k, \ell), & \text{otherwise,} \end{cases} \quad (47.27)$$

where

$$P_{\text{est}}(k, \ell) = \alpha_p P_{\text{est}}(k, \ell - 1) + (1 - \alpha_p) \|\mathbf{U}(k, \ell)\|^2 \quad (47.28)$$

represents the power of the noise reference signals, μ_g is a step size that regulates the convergence rate, and α_p is a smoothing parameter in the PSD estimation process.

To allow for the use of the STFT, we further assume that the ANC filters $\mathbf{g}_m$ have a time-varying finite impulse response (FIR) structure:

$$\mathbf{g}_m^T(t) = (g_{m, -K_L}(t) \ \dots \ g_{m, K_R}(t)). \quad (47.29)$$

Note, that the impulse responses are taken to be non-causal, to allow for relative delays between the FBF and the ANC branches.

In order to fulfill the FIR structure constraint in (47.29), the filters update is now given by

$$\tilde{\mathbf{G}}(\ell + 1, k) = \mathbf{G}(k, \ell) + \mu \frac{\mathbf{U}(k, \ell) \mathbf{Y}^*(k, \ell)}{P_{\text{est}}(k, \ell)},$$

$$\mathbf{G}(\ell + 1, k) \stackrel{\text{FIR}}{\leftarrow} \tilde{\mathbf{G}}(\ell + 1, k). \quad (47.30)$$

The operator $\stackrel{\text{FIR}}{\leftarrow}$ includes the following three stages, applied per filter: transformation of $\tilde{\mathbf{G}}_m(\ell + 1, k)$ to the time domain, truncation of the resulting impulse response to the interval $[-K_L, K_R]$ (i. e., imposing the FIR constraint), and transformation back to the frequency domain. The various filtering operations involved in the algorithm (multiplications in the transform domain) are realized using the overlap-and-save method [47.50, 51].

The resulting algorithm is merely an extension of the original Griffiths and Jim algorithm for the arbitrary ATF case. Figure 47.3 depicts a block diagram of the algorithm. The steps involved in the computation are summarized in Table 47.2. The matched beamformer $\mathbf{W}_0(k)$ and the BM $\mathcal{H}(k)$ are assumed to be known at this stage.

47.2.3 Time-Domain Generalized Sidelobe Canceller

The most commonly used GSC structure is the classical time-domain counterpart of the algorithm, proposed

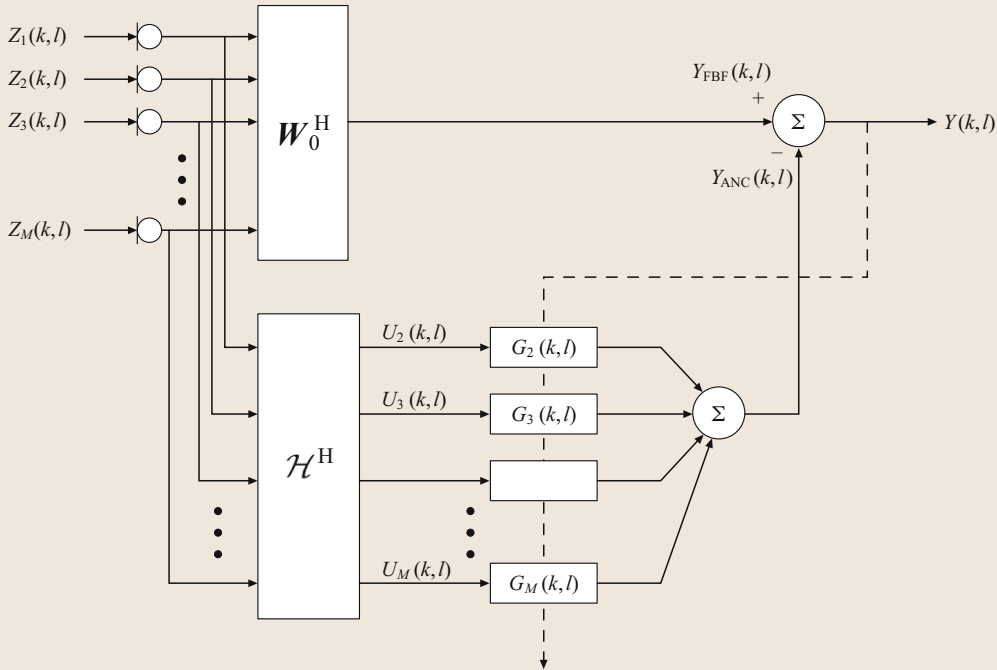


Fig. 47.3 Linearly constrained adaptive beamformer

Table 47.2 Frequency-domain GSC algorithm

1) Fixed beamformer:
$Y_{\text{FBF}}(k, \ell) = \mathbf{W}_0^H(k) \mathbf{Z}(k, \ell)$
2) Noise reference signals:
$\mathbf{U}(k, \ell) = \mathcal{H}^H(k) \mathbf{Z}(k, \ell)$
3) Output signal:
$Y(k, \ell) = Y_{\text{FBF}}(k, \ell) - \mathbf{G}^H(k, \ell) \mathbf{U}(k, \ell)$
4) Filters update
$\tilde{\mathbf{G}}(\ell+1, k) = \mathbf{G}(k, \ell) + \mu_g \frac{Y(k, \ell) Y^*(k, \ell)}{P_{\text{est}}(k, \ell)}$
$\mathbf{G}(\ell+1, k) \xleftarrow{\text{FIR}} \tilde{\mathbf{G}}(\ell+1, k)$,
where
$P_{\text{est}}(k, \ell) = \alpha_p P_{\text{est}}(\ell-1, k) + (1-\alpha_p) \sum_m Z_m(k, \ell) ^2$
5) Keep only nonaliased samples

by *Griffiths and Jim* [47.36]. For completeness of the exposition, we present now the time-domain algorithm.

Assuming that the array is steered towards the desired speech signal (refer to steering-related issues in Sect. 47.4), the FBF is given by

$$y_{\text{FBF}}(t) = \sum_{m=1}^M z_m(t),$$

which is the simple delay-and-sum beamformer. Under the same delay-only steered array assumptions, it is

evident that

$$u_m(t) = z_m(t) - z_1(t); \quad m = 2, \dots, M,$$

are noise-only signals and that the desired speech component is cancelled out.

The filters $\mathbf{g}_m$ are updated in the time domain. The error signal (which is also the output of the enhancement algorithm) is given by,

$$y(t) = y_{\text{FBF}}(t) - \sum_{m=2}^M \sum_{i=-K_L}^{K_R} g_{m,i}(t) u_m(t-i). \quad (47.31)$$

Define, for $m = 2, \dots, M$:

$$\begin{aligned} \mathbf{u}_m^T(t) &= (u_m(t+K_L) \cdots u_m(t) \cdots u_m(t-K_R)). \end{aligned}$$

Then, the adaptive normalized multichannel LMS solution is given by

$$\begin{aligned} \mathbf{g}_m(t+1) &= \mathbf{g}_m(t) + \frac{\mu}{p_{\text{est}}(t)} \mathbf{u}_m(t) y(t); \quad m = 2, \dots, M, \end{aligned}$$

where

$$p_{\text{est}}(t) = \sum_{m=1}^M \|\mathbf{u}_m(t)\|^2. \quad (47.32)$$

47.3 Fixed Beamformer and Blocking Matrix

In the previous section, we derived the generalized sidelobe canceller and showed that it includes a fixed beamformer, given in (47.20), a blocking matrix, given in (47.17), and a multichannel ANC, given in (47.26). Note that knowledge of the ATFs $A(k)$ (assumed to be slowly time variant) suffices to determine both the FBF and BM. In this section we present three methods for determination of the fixed beamformer and the blocking matrix.

47.3.1 Using Acoustical Transfer Functions

A typical RIR is depicted in Fig. 47.4. It can be seen that the impulse response can get very long (several thousand taps), which makes the estimation task quite cumbersome. This impulse response was generated by the image method [47.52] proposed by Allen and Berkley (The authors thank E. A. P. Habets from TU Eindhoven, The Netherlands, for providing an efficient implementation

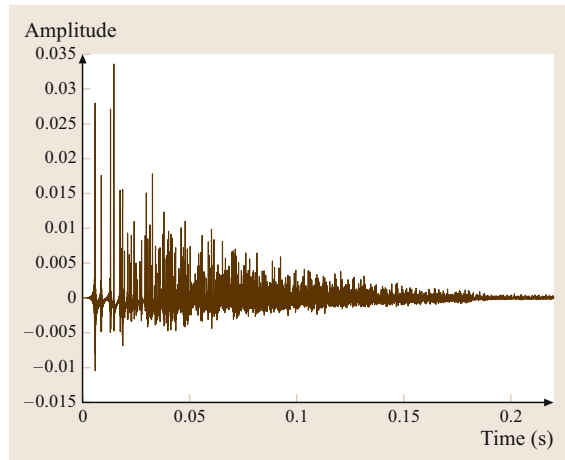


Fig. 47.4 A typical room impulse response with a reverberation time of 0.4 s

of the image method.). By this method, the **RIR** is generated by simulating multiple reflections of the sound source from the room walls. Two distinct segments of the impulse response can be observed. The first consists of the direct propagation path with a few early, distinguishable reflections. The second segment consists of overlapping random arrivals, with an exponentially decaying envelope. This segment is usually referred to as the tail of the **RIR**. Using this model we can estimate the various blocks of the **GSC**.

Assuming that $A(k)$ is known, we have by (47.22), (47.20), and (47.3)

$$Y_{\text{FBF}}(k, \ell) = S(k, \ell) + \frac{A^H(k)}{\|A(k)\|^2} [N_s(k, \ell) + N_t(k, \ell)]. \quad (47.33)$$

The first term on the right-hand side is the signal term, and the second is the noise term. The **FBF** in this case is hence a matched filter-and-sum beamformer (see also [47.16, 17]).

Considering the blocking matrix, there are many alternatives for blocking the desired speech signal in the reference channels. One alternative is calculation of

$$U_m(k, \ell) = A_m(k)Z_{m-1}(k, \ell) - A_{m-1}(k)Z_m(k, \ell)$$

for $m = 2, \dots, M$. Any other combination of the microphone signals is applicable.

47.3.2 Using Delay-Only Filters

The simplest and yet the most widely used model for the **ATF** is a delay-only model. Arbitrarily defining the first microphone as the reference microphone we have

$$A(k) = (1 \ e^{-i\frac{2\pi}{K}\tau_2} \ e^{-i\frac{2\pi}{K}\tau_3} \ \dots \ e^{-i\frac{2\pi}{K}\tau_M}),$$

where $\tau_2, \dots, \tau_M$ are the relative delays between each microphone and the reference microphone.

In the delay-only case, the **FBF** simplifies to the delay-and-sum beamformer, given by

$$W_0(k) = (1 \ e^{i\frac{2\pi}{K}\tau_2} \ e^{i\frac{2\pi}{K}\tau_3} \ \dots \ e^{i\frac{2\pi}{K}\tau_M}).$$

To avoid noncausal delays, a fixed amount of delay can be introduced.

It can easily be verified that the matrix

$$\mathcal{H}(k) = \begin{pmatrix} -e^{i\frac{2\pi}{K}\tau_2} & -e^{i\frac{2\pi}{K}\tau_2} & \dots & -e^{i\frac{2\pi}{K}\tau_2} \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & \dots & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix} \quad (47.34)$$

is a proper **BM** under the assumption of delay-only impulse response. This **BM**, originally proposed Griffiths and Jim [47.36], performs delay compensation and subtraction. It can be regarded as steering $M - 1$ null beams towards the desired speech signal.

47.3.3 Using Relative Transfer Functions

We have shown, on the one hand, that the **RIR** can be very long and hence difficult to estimate. On the other hand, the use of delay-only model suffers from severe undermodeling problems. A good compromise is the use of the relative transfer function (RTF) between sensors. Define the RTFs as the ratio

$$\tilde{A}^T(k) \triangleq \left(1 \ \frac{A_2(k)}{A_1(k)} \ \frac{A_3(k)}{A_1(k)} \ \dots \ \frac{A_M(k)}{A_1(k)} \right) = \frac{A^T(k)}{A_1(k)}. \quad (47.35)$$

Note that the **ATF** may have zeros outside the unit circle, as it is not necessarily a minimum-phase system. Thus to ensure stability of the RTFs we allow for noncausal systems. Therefore, we model the impulse response of the m -th ratio as

$$\tilde{a}_m^T(t) = (\tilde{a}_{m,-q_L}(t) \ \dots \ \tilde{a}_{m,q_R}(t)). \quad (47.36)$$

It was experimentally shown that RTFs are usually much shorter than the corresponding **ATFs** [47.38], hence the **FIR** assumption may be justified.

Replacing in (47.20) the actual **ATFs** by the RTFs, the **FBF** becomes

$$W_0(k) = \frac{\tilde{A}(k)}{\|\tilde{A}(k)\|^2}. \quad (47.37)$$

By (47.22) and (47.3) we then have

$$Y_{\text{FBF}}(k, \ell) = A_1(k)S(k, \ell) + \frac{\tilde{A}^H(k)}{\|\tilde{A}(k)\|^2} [N_s(k, \ell) + N_t(k, \ell)]. \quad (47.38)$$

Thus, when $W_0(k, \ell)$ is given by (47.37), the signal term of $Y_{\text{FBF}}(k, \ell)$ is the desired signal distorted only by the first **ATF**, $A_1(k)$. Note, however, that all the sensor outputs are summed coherently.

It can be easily verified that the use of the following **BM** suffices for completely eliminating the desired speech signal, provided that the RTFs are correctly modeled and estimated:

$$\mathcal{H}(k) = \begin{pmatrix} -\tilde{A}_2^*(k) & -\tilde{A}_3^*(k) & \dots & -\tilde{A}_M^*(k) \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \dots & \ddots \\ 0 & 0 & \dots & 1 \end{pmatrix}. \quad (47.39)$$

For this choice of the **BM**, the reference signals are given by,

$$\begin{aligned} U_m(k, \ell) &= Z_m(k, \ell) - \tilde{A}_m(k)Z_1(k, \ell); \\ m &= 2, 3, \dots, M. \end{aligned} \quad (47.40)$$

47.4 Identification of the Acoustical Transfer Function

The specific choice of the **ATFs** model governs the applicable estimation method.

47.4.1 Signal Subspace Method

Affes and *Grenier* [47.37] prove that the identification of source-to-array impulse responses is possible by subspace tracking. Assume that the approximation in (47.3) is valid, i. e., the multichannel correlation matrix of the received signals is given by

$$\Phi_{ZZ}(k, \ell) = \phi_{ss}(k, \ell)\mathbf{A}(k)\mathbf{A}^H(k) + \Phi_{NN}(k, \ell).$$

Assume also that the noise signal is spatially white, i. e., $\Phi_{NN}(k, \ell) = \sigma_n^2(k, \ell)\mathbf{I}$ (an extension for spatially nonwhite noise is described in [47.53]). The eigenvalues of the received signals correlation matrix are given by

$$\begin{aligned} \lambda_l &= \sigma_n^2(k, \ell) \quad l = 1, \dots, M-1 \\ \lambda_M &> \sigma_n^2(k, \ell) \quad \text{otherwise.} \end{aligned} \quad (47.41)$$

From the matrix structure we conclude that the most dominant eigenvector of $\Phi_{ZZ}(k, \ell)$ is given by $\mathbf{A}(k)$ (up to a scale factor). Hence, using the eigenvalue decomposition of $\Phi_{ZZ}(k, \ell)$, we can estimate the desired signal **ATFs**. In the case of spatially nonwhite noise signals the generalized eigenvalue decomposition (**GEVD**), using the known noise correlation matrix $\Phi_{NN}(k, \ell)$, can be applied instead [47.53].

Yang [47.54] proved that finding the most dominant eigenvectors is equivalent to minimizing a quadratic cost function. For the following derivation it is more convenient to deal with normalized terms. Let

$$\begin{aligned} \mathbf{Z}(k, \ell) &= \mathbf{A}(k)S(k, \ell) + \mathbf{N}(k, \ell) \\ &= \frac{\mathbf{A}(k)}{\|\mathbf{A}(k)\|} \|\mathbf{A}(k)\|S(k, \ell) + \mathbf{N}(k, \ell) \\ &\triangleq \tilde{\mathbf{A}}(k)\tilde{S}(k, \ell) + \mathbf{N}(k, \ell), \end{aligned} \quad (47.42)$$

where $\tilde{\mathbf{A}}(k) \triangleq \mathbf{A}(k)/\|\mathbf{A}(k)\|$ is the normalized **ATFs** vector, namely $\tilde{\mathbf{A}}^H(k)\tilde{\mathbf{A}}(k) = 1$, and $\tilde{S}(k, \ell) \triangleq \|\mathbf{A}(k)\|S(k, \ell)$ is the normalized desired speech signal.

When only the single most dominant eigenvector is required (as in the discussed case), *Affes* and *Grenier* [47.37] showed that Yang's criterion simplifies to the following minimization,

$$E\{\|\mathbf{I} - \tilde{\mathbf{A}}(k)\tilde{\mathbf{A}}^H(k)\mathbf{Z}(k, \ell)\|^2\}. \quad (47.43)$$

Similar to the approximation used in the derivation of **LMS** algorithm, the received signal correlation matrix is approximated by its instantaneous value $\Phi_{ZZ}(k, \ell) \approx \mathbf{Z}(k, \ell)\mathbf{Z}^H(k, \ell)$. Furthermore, approximating $\hat{\mathbf{A}}^H(k, \ell)\hat{\mathbf{A}}(k, \ell) \approx 1$, where $\hat{\mathbf{A}}(k, \ell)$ is an estimate of $\tilde{\mathbf{A}}(k)$ at the current time instant, the following sequential procedure can be derived:

$$\begin{aligned} &\hat{\mathbf{A}}(k, \ell+1) \\ &= \hat{\mathbf{A}}(k, \ell) + \mu_a(k, \ell)[\mathbf{Z}(k, \ell) \\ &\quad - \hat{\mathbf{A}}(k, \ell)\hat{\mathbf{A}}^H(k, \ell)\mathbf{Z}(k, \ell)][\hat{\mathbf{A}}^H(k, \ell)\mathbf{Z}(k, \ell)]^*. \end{aligned} \quad (47.44)$$

Define $\tilde{Y}_{\text{FBF}}(k, \ell) \triangleq \tilde{\mathbf{A}}^H(k, \ell)\mathbf{Z}(k, \ell)$, and observe that the desired signal component at $\tilde{Y}_{\text{FBF}}(k, \ell)$ is $\tilde{S}(k, \ell)$. Then,

$$\begin{aligned} &\hat{\mathbf{A}}(k, \ell+1) \\ &= \hat{\mathbf{A}}(k, \ell) + \mu_a(k, \ell)[\mathbf{Z}(k, \ell) \\ &\quad - \hat{\mathbf{A}}(k, \ell)\tilde{Y}_{\text{FBF}}(k, \ell)]\tilde{Y}_{\text{FBF}}^*(k, \ell). \end{aligned} \quad (47.45)$$

Now, it is easy to verify that

$$\mathbf{Z}(k, \ell) - \hat{\mathbf{A}}(k, \ell)\tilde{Y}_{\text{FBF}}(k, \ell)$$

are noise-only signals, provided that the estimate $\hat{\mathbf{A}}(k, \ell)$ converges to the true normalized **ATFs** vector $\tilde{\mathbf{A}}(k)$.

Hence, it can serve as a **BM** output, namely $U(k, \ell)$. (There is a slight difference between the proposed **BM** and the conventional **BM**, as the number of the components in $U(k, \ell)$ in the current scheme is M rather than $M - 1$ as with the latter.) Collecting all terms we finally have

$$\begin{aligned} & \hat{A}(k, \ell + 1) \\ &= \hat{A}(k, \ell) + \mu_a(k, \ell) U(k, \ell) \bar{Y}_{\text{FBF}}^*(k, \ell). \end{aligned} \quad (47.46)$$

Note that this procedure yields an estimate of the normalized **ATFs** rather than the **ATFs** themselves. *Affes* and *Grenier* [47.37] argue that $\|A(k)\|$ is invariant to small talker movements and could be estimated in advance.

As a final remark, we would like to point out that the estimation procedure in (47.46) has many similarities to the method of *Hoshuyama* et al. [47.55] for robust design of the **BM**, and with the decorrelation criterion presented by *Weinstein* et al. [47.56] and further adapted to the **GSC** structure by *Gannot* [47.57]. We will elaborate on this issue in Sect. 47.5 when discussing the robust beamformers.

47.4.2 Time Difference of Arrival

When a delay-only steering is applied, an estimation of the time difference of arrival between the microphones suffices to model the entire impulse response. It should be noted however, that this procedure usually undermodels the **RIR** and is not sufficient for the problem at hand. Many algorithms were proposed for estimation the time difference of arrival (**TDOA**) [47.58, 59]. A survey of state-of-the-art methods for **TDOA** estimation can be found in [47.60]. This topic is beyond the scope of this chapter.

47.4.3 Relative Transfer Function Estimation

We present two methods for RTF estimation. The first is based on speech nonstationarity [47.38], and the second employs the speech presence probability [47.61, 62].

Using Signal Nonstationarity

In this section, we review the system identification technique proposed by *Shalvi* and *Weinstein* [47.63] and later used by *Gannot* et al. [47.38] in the context of microphone arrays. This method relies on the assumptions that the background noise signal is stationary, that the desired signal $s(t)$ is nonstationary, and that the support of the relative impulse response between the sensors

is finite and slowly time varying. (Note that the relative impulse response between the sensors is generally of infinite length, since it represents the ratio of **ATFs**. However, in real environments, the energy of the relative impulse response often decays much faster than the corresponding **ATF** [47.38]. Therefore, the finite-support assumption is practically not very restrictive.)

Rearranging terms in (47.40) we have

$$Z_m(k, \ell) = \tilde{A}_m(k) Z_1(k, \ell) + U_m(k, \ell). \quad (47.47)$$

We assume that the RTFs are slowly changing in time compared to the time variations of the desired signal. We further assume that the statistics of the noise signal is slowly changing compared to the statistics of the desired signal. Consider some analysis interval during which the **ATFs** are assumed to be time invariant and the noise signal is assumed to be stationary. We divide that analysis interval into frames. Consider the i -th frame. By (47.47) we have

$$\begin{aligned} \Phi_{z_m z_1}^{(i)}(k) &= \tilde{A}_m(k) \Phi_{z_1 z_1}^{(i)}(k) + \Phi_{u_m z_1}(k), \\ i &= 1, \dots, I, \end{aligned} \quad (47.48)$$

where I is the number of frames, $\Phi_{z_i z_j}^{(i)}(k)$ is the cross-**PSD** between z_i and z_j at frequency bin k during the i -th frame, and $\Phi_{u_m z_1}(k)$ is the cross-**PSD** between u_m and z_1 at frequency bin k , which is independent of the frame index due to the noise stationarity. Now, equations (47.2) and (47.40) imply that, when the signal is present,

$$U_m(k, \ell) = N_m^s(k, \ell) - \tilde{A}_m(k) N_1^s(k, \ell) \quad (47.49)$$

$$Z_1(k, \ell) = A_1(k) S(k, \ell) + N_1^s(k, \ell). \quad (47.50)$$

Let $\hat{\Phi}_{z_1 z_1}^{(i)}(k)$, $\hat{\Phi}_{z_m z_1}^{(i)}(k)$, and $\hat{\Phi}_{u_m z_1}^{(i)}(k)$ be estimates of $\Phi_{z_1 z_1}^{(i)}(k)$, $\Phi_{z_m z_1}^{(i)}(k)$, and $\Phi_{u_m z_1}(k)$, respectively. The estimates are obtained by replacing expectations with averages. Note that (47.48) also holds for the estimated values. Let $\varepsilon_m^{(i)}(k) = \hat{\Phi}_{u_m z_1}^{(i)}(k) - \Phi_{u_m z_1}(k)$ denote the estimation error of the cross-**PSD** between z_1 and u_m in the i -th frame. We then obtain,

$$\begin{aligned} \hat{\Phi}_{z_m z_1}^{(i)}(k) &= \tilde{A}_m(k) \hat{\Phi}_{z_1 z_1}^{(i)}(k) + \Phi_{u_m z_1}(k) \\ &\quad + \varepsilon_m^{(i)}(k), \quad i = 1, \dots, I. \end{aligned} \quad (47.51)$$

If the noise reference signals $U_m(k, \ell)$, $m = 2, \dots, M$ were uncorrelated with $Z_1(k, \ell)$, then the standard system identification estimate, $\tilde{A}_m(k) = \hat{\Phi}_{z_m z_1}(k) / \hat{\Phi}_{z_1 z_1}(k)$, could be used to obtain an unbiased estimate of $A_m(k)$. Unfortunately, by (47.49) and (47.50), $U_m(k, \ell)$ and $Z_1(k, \ell)$ are in general correlated. Hence in [47.63] it is

proposed to obtain an unbiased estimate of $\tilde{A}_m(k)$ by applying the least-squares (LS) procedure to the following set of overdetermined equations

$$\begin{aligned} \mathbf{x} &\triangleq \begin{pmatrix} \hat{\Phi}_{z_m z_1}^{(1)}(k) \\ \hat{\Phi}_{z_m z_1}^{(2)}(k) \\ \vdots \\ \hat{\Phi}_{z_m z_1}^{(K)}(k) \end{pmatrix} \\ &= \begin{pmatrix} \hat{\Phi}_{z_1 z_1}^{(1)}(k) & 1 \\ \hat{\Phi}_{z_1 z_1}^{(2)}(k) & 1 \\ \vdots & \vdots \\ \hat{\Phi}_{z_1 z_1}^{(K)}(k) & 1 \end{pmatrix} \begin{pmatrix} \tilde{A}_m(k) \\ \Phi_{u_m z_1}(k) \end{pmatrix} + \begin{pmatrix} \varepsilon_m^{(1)}(k) \\ \varepsilon_m^{(2)}(k) \\ \vdots \\ \varepsilon_m^{(K)}(k) \end{pmatrix} \\ &\triangleq \mathbf{G} \boldsymbol{\theta} + \boldsymbol{\epsilon}, \end{aligned} \quad (47.52)$$

where a separate set of equations is used for each $m = 2, \dots, M$. The weighted least-squares (WLS) estimate of $\boldsymbol{\theta}$ is obtained by

$$\begin{aligned} \begin{pmatrix} \hat{A}_m(k) \\ \hat{\Phi}_{u_m z_1}(k) \end{pmatrix} &= \hat{\boldsymbol{\theta}} \\ &= \arg \min_{\boldsymbol{\theta}} (\mathbf{x} - \mathbf{G} \boldsymbol{\theta})^H \mathbf{W} (\mathbf{x} - \mathbf{G} \boldsymbol{\theta}) \\ &= (\mathbf{G}^H \mathbf{W} \mathbf{G})^{-1} \mathbf{G}^H \mathbf{W} \mathbf{x}, \end{aligned} \quad (47.53)$$

where $\mathbf{W}$ is a positive Hermitian weighting matrix, and $\mathbf{G}^H \mathbf{W} \mathbf{G}$ is required to be invertible.

Shalvi and Weinstein suggested two alternative weighting matrices. One alternative is given by

$$W_{ij} = \begin{cases} T_i, & i = j \\ 0, & i \neq j \end{cases}, \quad (47.54)$$

where T_i is the length of subinterval i , so that longer intervals have higher weights. In this case, (47.53) reduces to

$$\begin{aligned} \hat{A}(k) &= \frac{\langle \hat{\Phi}_{z_m z_1}(k) \hat{\Phi}_{z_1 z_1}(k) \rangle - \langle \hat{\Phi}_{z_m z_1}(k) \rangle \langle \hat{\Phi}_{z_1 z_1}(k) \rangle}{\langle \hat{\Phi}_{z_1 z_1}^2(k) \rangle - \langle \hat{\Phi}_{z_1 z_1}(k) \rangle^2} \end{aligned} \quad (47.55)$$

with the average operation defined by

$$\langle \varphi(k) \rangle \triangleq \frac{\sum_{i=1}^I T_i \varphi^{(i)}(k)}{\sum_{i=1}^I T_i}. \quad (47.56)$$

Another alternative for $\mathbf{W}$, that minimizes the covariance of $\hat{\boldsymbol{\theta}}$, is given by

$$W_{ij} = \frac{B}{\hat{\Phi}_{u_m u_m}(k)} \begin{cases} T_i / \hat{\Phi}_{z_1 z_1}^{(i)}(k), & i = j \\ 0, & i \neq j \end{cases}, \quad (47.57)$$

where B is related to the bandwidth of the window used for the cross-PSD estimation [47.63]. With this choice of the weighting function, (47.53) yields

$$\begin{aligned} \hat{A}(k) &= \frac{\langle 1 / \hat{\Phi}_{z_1 z_1}(k) \rangle \langle \hat{\Phi}_{z_m z_1}(k) \rangle}{\langle \hat{\Phi}_{z_1 z_1}(k) \rangle \langle 1 / \hat{\Phi}_{z_1 z_1}(k) \rangle - 1} \\ &\quad - \frac{\langle \hat{\Phi}_{z_m z_1}(k) / \hat{\Phi}_{z_1 z_1}(k) \rangle}{\langle \hat{\Phi}_{z_1 z_1}(k) \rangle \langle 1 / \hat{\Phi}_{z_1 z_1}(k) \rangle - 1} \end{aligned} \quad (47.58)$$

and the variance of $\hat{A}(k)$ is given by

$$\text{var}\{\hat{A}(k)\} = \frac{1}{BT} \frac{\Phi_{u_m u_m}(k) \langle 1 / \Phi_{z_1 z_1}(k) \rangle}{\langle \Phi_{z_1 z_1}(k) \rangle \langle 1 / \Phi_{z_1 z_1}(k) \rangle - 1}, \quad (47.59)$$

where $T \triangleq \sum_{i=1}^I T_i$ is the total observation interval. Special attention should be given to choosing the frame length. On the one hand, it should be longer than the correlation length of $z_m(t)$, which must be longer than the length of the filter $a_m(t)$. On the other hand, it should be short enough for the filter time invariance and the noise quasistationarity assumptions to hold.

A major limitation of the WLS optimization in (47.53) is that both the identification of $\tilde{A}(k)$ and the estimation of the cross-PSD $\Phi_{u_m z_1}(k)$ are carried out using the same weight matrix $\mathbf{W}$. That is, each subinterval i is given the same weight, whether we are trying to find an estimate for $\tilde{A}(k)$ or for $\Phi_{u_m z_1}(k)$. However, subintervals with higher SNR values are of greater importance when estimating $\tilde{A}(k)$, whereas the opposite is true when estimating $\Phi_{u_m z_1}(k)$. Consequently, the optimization criterion in (47.53) consists of two conflicting requirements: one is minimizing the error variance of $\tilde{A}(k)$, which pulls the weight up to higher values on higher SNR subintervals. The other requirement is minimizing the error variance of $\Phi_{u_m z_1}(k)$, which rather implies smaller weights on higher SNR subintervals. For instance, suppose we obtain observations on a relatively long low-SNR interval of length T_0 , and on a relatively short high-SNR interval of length T_1 ($T_1 \ll T_0$). Then, the variance of $\tilde{A}(k)$ in (47.59) is inversely proportional to the relative length of the high-SNR interval, $T_1 / (T_0 + T_1)$. That is, including in the observation interval additional segments that do not contain speech (i. e., increasing T_0) increases the variance of $\tilde{A}(k)$. This unnatural

consequence is a result of the desire to minimize the variance of $\phi_{u_m z_1}(k)$ by using larger weights on the segments that do not contain speech, while increasing the weights on such subintervals degrades the estimate for $\tilde{A}(k)$.

Another major limitation of RTF identification using nonstationarity is that the interfering signals are required to be stationary during the entire observation interval. The observation interval should include a certain number of subintervals that contain the desired signal, such that $\phi_{z_1 z_1}(k)$ is sufficiently nonstationary for all k . Unfortunately, if the desired signal is speech, the presence of the desired signal in the observed signals may be sparse in some frequency bands. This entails a very long observation interval, thus constraining the interfering signals to be stationary over long intervals. Furthermore, the RTF $\tilde{A}(k)$ is assumed to be constant during the observation interval. Hence, very long observation intervals also restrict the capability of the system identification technique to track varying $\tilde{A}(k)$ (e.g., tracking moving talkers in reverberant environments).

Using Speech Presence Probability

In this section, we present a system identification approach that is adapted to speech signals. Specifically, the presence of the desired speech signal in the time–frequency domain is uncertain, and the speech presence probability is utilized to separate the tasks of system identification and cross-PSD estimation. An estimate for $\tilde{A}(k)$ is derived based on subintervals that contain speech, while subintervals that do not contain speech are of more significance when estimating the components of $\phi_{u_m z_1}(k)$.

Let the observed signals be divided in time into overlapping frames by the application of a window function and analyzed using the STFT. Under the same considerations leading to the estimation procedure based on speech nonstationarity (and based on the assumption that the RTFs can be modelled by short filters), (47.48) is still valid. Now using (47.48)–(47.50) and the fact that the desired signal $s(t)$ is uncorrelated with the interfering signals $n_m^s(t)$; $m = 1, 2, \dots, M$, we have

$$\begin{aligned} \phi_{z_m z_1}(k, \ell) &= \tilde{A}_m(k) |A_1|^2(k) \phi_{s s}(k, \ell) \\ &\quad + \phi_{n_m^s n_1^s}(k, \ell). \end{aligned} \quad (47.60)$$

Writing this equation in terms of the PSD estimates, we obtain for $m = 2, 3, \dots, M$

$$\begin{aligned} \hat{\phi}_{z_m z_1}(k, \ell) &= \tilde{A}_m(k) |\hat{A}_1|^2(k) \hat{\phi}_{s s}(k, \ell) \\ &\quad + \hat{\phi}_{n_m^s n_1^s}(k, \ell) + \varepsilon_m(k, \ell) \\ &= \tilde{A}_m(k) \hat{\phi}_{s s}(k, \ell) + \hat{\phi}_{n_m^s n_1^s}(k, \ell) + \varepsilon_m(k, \ell), \end{aligned} \quad (47.61)$$

where $\varepsilon_m(k, \ell)$ denotes an estimation error and $\hat{\phi}_{s s}(k, \ell) = |\hat{A}_1|^2(k) \hat{\phi}_{s s}(k, \ell)$ represents the PSD of the speech signal component in microphone 1. This gives us L equations, which may be written in a matrix form as

$$\begin{aligned} \hat{\Psi}_m(k) &\triangleq \begin{pmatrix} \hat{\phi}_{z_m z_1}(k, 1) - \hat{\phi}_{n_m^s n_1^s}(k, 1) \\ \hat{\phi}_{z_m z_1}(k, 2) - \hat{\phi}_{n_m^s n_1^s}(k, 2) \\ \vdots \\ \hat{\phi}_{z_m z_1}(k, L) - \hat{\phi}_{n_m^s n_1^s}(k, L) \end{pmatrix} \\ &= \begin{pmatrix} \hat{\phi}_{s s}(k, 1) \\ \hat{\phi}_{s s}(k, 2) \\ \vdots \\ \hat{\phi}_{s s}(k, L) \end{pmatrix} \tilde{A}_m(k) + \begin{pmatrix} \varepsilon_m(k, 1) \\ \varepsilon_m(k, 2) \\ \vdots \\ \varepsilon_m(k, L) \end{pmatrix} \\ &\triangleq \hat{\phi}_{s s}(k) \tilde{A}_m(k) + \mathbf{\varepsilon}_m(k). \end{aligned} \quad (47.62)$$

Since the RTF $\tilde{A}_m(k)$ represents the coupling between the primary and reference sensors with respect to the *desired* source signal, the optimization criterion for the identification of $\tilde{A}_m(k)$ has to take into account only short-time frames which contain desired signal components. Specifically, let $I(k, \ell)$ denote an indicator function for the signal presence [$I(k, \ell) = 1$ if $\phi_{s s}(k, \ell) \neq 0$, i.e., during H_1 , and $I(k, \ell) = 0$ otherwise], and let $\mathbf{I}(k)$ represent a diagonal matrix with the elements $[I(k, 1), I(k, 2), \dots, I(k, L)]$ on its diagonal. Then the WLS estimate of $\tilde{A}(k)$ is obtained by

$$\begin{aligned} \hat{\tilde{A}}_m &= \arg \min_{\tilde{A}_m} \{ [\mathbf{I} \mathbf{\varepsilon}_m]^H \mathbf{W} [\mathbf{I} \mathbf{\varepsilon}_m] \} \\ &= \arg \min_{\tilde{A}_m} \{ [\hat{\Psi}_m - \hat{\phi}_{s s} \tilde{A}_m]^H \\ &\quad \mathbf{I} \mathbf{W} \mathbf{I} [\hat{\Psi}_m - \hat{\phi}_{s s} \tilde{A}_m] \} \\ &= [\hat{\phi}_{s s}^T \mathbf{I} \mathbf{W} \mathbf{I} \hat{\phi}_{s s}]^{-1} \hat{\phi}_{s s}^T \mathbf{I} \mathbf{W} \mathbf{I} \hat{\Psi}_m, \end{aligned} \quad (47.63)$$

where the argument k has been omitted for notational simplicity. Recognizing the product $\mathbf{I} \mathbf{W} \mathbf{I}$ as the equivalent weight matrix, the variance of $\hat{\tilde{A}}$ is given by ([47.64] p. 405)

$$\begin{aligned} \text{var}\{\hat{\tilde{A}}_m\} &= (\hat{\phi}_{s s}^T \mathbf{I} \mathbf{W} \mathbf{I} \hat{\phi}_{s s})^{-1} \hat{\phi}_{s s}^T \mathbf{I} \mathbf{W} \mathbf{I} \text{cov}(\mathbf{\varepsilon}_m) \\ &\quad \times \mathbf{I} \mathbf{W} \mathbf{I} \hat{\phi}_{s s} (\hat{\phi}_{s s}^T \mathbf{I} \mathbf{W} \mathbf{I} \hat{\phi}_{s s})^{-1}, \end{aligned} \quad (47.64)$$

where $\text{cov}(\mathbf{\varepsilon}_m)$ is the covariance matrix of $\mathbf{\varepsilon}_m$. The matrix $\mathbf{W}$ that minimizes the variance of $\hat{\tilde{A}}$ therefore satisfies ([47.64] prop. 8.2.4)

$$\mathbf{I} \mathbf{W} \mathbf{I} = \mathbf{I} [\text{cov}(\mathbf{\varepsilon}_m)]^{-1} \mathbf{I}. \quad (47.65)$$

This choice of $\mathbf{W}$ yields an asymptotically unbiased estimator

$$\hat{A} = \{\hat{\phi}_{\bar{s}\bar{s}}^T \mathbf{I}[\text{cov}(\mathbf{e}_m)]^{-1} \mathbf{I} \hat{\phi}_{\bar{s}\bar{s}}\}^{-1} \times \hat{\phi}_{\bar{s}\bar{s}}^T \mathbf{I}[\text{cov}(\mathbf{e}_m)]^{-1} \mathbf{I} \hat{\psi}_m, \quad (47.66)$$

which is known as the *minimum variance* or Gauss–Markov estimator. Substituting (47.65) into (47.64), we obtain the variance of the resulting estimator

$$\text{var}\{\hat{A}\} = \{\hat{\phi}_{\bar{s}\bar{s}}^T \mathbf{I}[\text{cov}(\mathbf{e}_m)]^{-1} \mathbf{I} \hat{\phi}_{\bar{s}\bar{s}}\}^{-1}. \quad (47.67)$$

The elements of $\text{cov}(\mathbf{e}_m)$ are asymptotically given by (see [47.61])

$$\begin{aligned} & \text{cov}[\varepsilon_m(k, \ell), \varepsilon_m(k, \ell')] \\ &= \begin{cases} \phi_{\bar{s}\bar{s}}(k, \ell) \phi_{u_m u_m}(k, \ell), & \text{if } \ell = \ell', \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (47.68)$$

Under the assumption that hypothesis H_{0t} is false, i. e., the noise is stationary, $\phi_{u_m u_m}(k, \ell)$ is independent of the frame index ℓ (in practice, it suffices that the statistics of the interfering signals is slowly changing compared with the statistics of the desired signal). Denoting by $\langle \cdot \rangle_\ell$ an average operation over the frame index ℓ

$$\langle \varphi(k, \ell) \rangle_\ell \triangleq \frac{1}{L} \sum_{\ell=1}^L \varphi(k, \ell), \quad (47.69)$$

and substituting (47.68) into (47.66) and (47.67) we obtain

$$\begin{aligned} \hat{A}_m(k) &= \frac{\langle I(k, \ell) [\hat{\phi}_{z_m z_1}(k, \ell) - \hat{\phi}_{n_m n_1^s}(k, \ell)] \rangle_\ell}{\langle I(k, \ell) \hat{\phi}_{\bar{s}\bar{s}}(k, \ell) \rangle_\ell}, \end{aligned} \quad (47.70)$$

$$\text{var}\{\hat{A}_m(k)\} = \frac{\phi_{u_m u_m}(k)}{L \langle I(k, \ell) \hat{\phi}_{\bar{s}\bar{s}}(k, \ell) \rangle_\ell}. \quad (47.71)$$

Note that, for a given frequency-bin index k , only frames that contain speech [$I(k, \ell) \neq 0$] influence the values of $\hat{A}_m(k)$ and $\text{var}\{\hat{A}_m(k)\}$. In contrast to the nonstationarity method, including in the observation interval additional segments that do not contain speech does not increase the variance of $\hat{A}_m(k)$ for any k . However, the proposed identification approach requires an estimate for $I(k, \ell)$, i. e. identifying which time–frequency bins (k, ℓ) contain the desired signal. In practice, the speech presence probability $p(k, \ell)$ can be estimated from the beamformer

output (see Sect. 47.6), and an estimate for the indicator function is obtained by

$$\hat{I}(k, \ell) = \begin{cases} 1, & \text{if } p(k, \ell) \geq p_0, \\ 0, & \text{otherwise,} \end{cases} \quad (47.72)$$

where p_0 ($0 \leq p_0 < 1$) is a predetermined threshold. Note, also that the output of the beamformer consists of the filtered version of the desired speech $\tilde{s}(t)$, when the RTFs is used in the FBF branch. The parameter p_0 controls the trade-off between the detection and false alarm probabilities, which are defined by $P_D \triangleq \mathcal{P}\{p(k, \ell) \geq p_0 \mid I(k, \ell) = 1\}$ and $P_{FA} \triangleq \mathcal{P}\{p(k, \ell) \geq p_0 \mid I(k, \ell) = 0\}$. A smaller value of p_0 increases the detection probability and allows for more short-time frames to be involved in the estimation of $\hat{A}(k)$. However, a smaller value of p_0 also increases the false alarm probability, which may cause a mismodification of $\hat{A}(k)$ due to frames that do not contain desired speech components.

The identification algorithm based on speech presence probability requires estimates for $\phi_{z_m z_1}(k, \ell)$, $\phi_{\bar{s}\bar{s}}(k, \ell)$, and $\phi_{n_m n_1^s}(k, \ell)$. An estimate for $\phi_{z_m z_1}(k, \ell)$ is obtained by applying a first-order recursive smoothing to the cross-periodogram of the observed signals, $Z_m(k, \ell) Z_1^*(k, \ell)$. Specifically,

$$\begin{aligned} \hat{\phi}_{z_m z_1}(k, \ell) &= \alpha_s \hat{\phi}_{z_m z_1}(k, \ell - 1) \\ &\quad + (1 - \alpha_s) Z_m(k, \ell) Z_1^*(k, \ell), \end{aligned} \quad (47.73)$$

where the smoothing parameter α_s ($0 \leq \alpha_s < 1$) determines the equivalent number of cross-periodograms that are averaged, $N_\ell \approx (1 + \alpha_s)/(1 - \alpha_s)$. Typically, speech periodograms are recursively smoothed with an equivalent rectangular window of $T_s = 0.2$ s long, which represents a good compromise between smoothing the noise and tracking the speech spectral variations [47.65]. Therefore, for a sampling rate of 8 kHz, an STFT window length of 256 samples and a frame update step of 128 samples, we use $\alpha_s = (T_s \cdot 8000/128 - 1)/(T_s \cdot 8000/128 + 1) \approx 0.85$.

To obtain an estimate for the PSD of the desired signal, we can use the output of the multichannel postfilter discussed in Sect. 47.6, during speech presence, i. e., H_1 is true.

$$\begin{aligned} \hat{\phi}_{\bar{s}\bar{s}}(k, \ell) &= \alpha_s \hat{\phi}_{\bar{s}\bar{s}}(k, \ell - 1) \\ &\quad + (1 - \alpha_s) I(k, \ell) |Y(k, \ell)|^2. \end{aligned} \quad (47.74)$$

The cross-PSD of the interfering signals, $n_m^s(t)$ and $n_1^s(t)$, is estimated by using the minima-controlled

recursive averaging (MCRA) approach [47.66, 67]. Specifically, past spectral cross-power values of the noisy observed signals are recursively averaged with a time-varying frequency-dependent smoothing parameter during periods for which H_{0s} is true

$$\begin{aligned}\hat{\phi}_{n_m n_1^s}(k, \ell) &= \tilde{\alpha}_u(k, \ell) \hat{\phi}_{n_m n_1^s}(k, \ell - 1) \\ &+ \beta [1 - \tilde{\alpha}_u(k, \ell)] Z_m(k, \ell) Z_1^*(k, \ell),\end{aligned}\quad (47.75)$$

where $\tilde{\alpha}_u(k, \ell)$ is the smoothing parameter ($0 < \tilde{\alpha}_u(k, \ell) \leq 1$), and β ($\beta \geq 1$) is a factor that compensates the bias when the desired signal is absent [47.67]. The smoothing parameter is determined by the signal presence probability, $p(k, \ell)$, and a constant α_u ($0 < \alpha_u < 1$)

that represents its minimal value

$$\tilde{\alpha}_u(k, \ell) = \alpha_u + (1 - \alpha_u)p(k, \ell). \quad (47.76)$$

The value of $\tilde{\alpha}_u$ is close to 1 when the desired signal is present to prevent the noise cross-PSD estimate from increasing as a result of signal components. It decreases linearly with the probability of signal presence to allow a faster update of the noise estimate. The value of α_u compromises between the tracking rate (response rate to abrupt changes in the noise statistics) and the variance of the noise estimate. Typically, in the case of high levels of nonstationary noise, a good compromise is obtained by $\alpha_u = 0.85$ [47.67]. Substituting these spectral estimates into (47.70) we obtain an estimate for $\hat{A}_m(k)$.

47.5 Robustness and Distortion Weighting

Beamformers often suffer from sensitivity to signal mismatch. The GSC in particular suffers from two basic problems. First, nonideal FBF can lead to noncoherent filter-and-sum operation. Doclo and Moonen [47.27] and Nordholm et al. [47.68] use spatial and frequency-domain constraints to improve the robustness of beamformers. The second problem, which is the concern of this survey, is the leakage phenomenon, caused by imperfect BM. If the desired speech leaks into the noise reference signals $U(k, \ell)$ the noise canceller filters will subtract speech components from the FBF output, causing self-cancellation of the desired speech, and hence a severe distortion. Note that, even when the ANC filters are adapted during noise-only periods, the self-cancellation is unavoidable. The goal of this section is to present several concepts for increasing the robustness of the GSC structure and reducing its sensitivity to signal mismatch.

Cox et al. [47.7] presented a thorough analysis of array sensitivity. The array output SNR is evidently given by

$$\text{SNR}_{\text{out}}(k, \ell) = \frac{\phi_{ss}(k, \ell) \mathbf{W}^H(k) \mathbf{A}(k) \mathbf{A}^H(k) \mathbf{W}(k)}{\mathbf{W}^H(k) \Phi_{NN}(k, \ell) \mathbf{W}(k)}.$$

Now assume that the signal's ATFs are different from the ATFs used for designing the LCMV beamformer, i. e., $\tilde{\mathbf{A}}(k) = \mathbf{A} + \epsilon(k)$. Assume also that the spatial correlation matrix of the perturbation $\epsilon(k)$ is given by $E\{\epsilon(k)\epsilon^H(k)\} = \sigma_\epsilon^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of dimensions $M \times M$. Namely, we assume that the perturbation components are uncorrelated. Hence the expected

output SNR is given by

$$\begin{aligned}E\{\text{SNR}_{\text{out}}(k, \ell)\} &= \frac{\phi_{ss}(k, \ell) \mathbf{W}^H(k) (\mathbf{A}(k) \mathbf{A}^H(k) + \sigma_\epsilon^2 \mathbf{I}) \mathbf{W}(k)}{\mathbf{W}^H(k) \Phi_{NN}(k, \ell) \mathbf{W}(k)}.\end{aligned}\quad (47.77)$$

Define the sensitivity of the array to the ATFs perturbation as $J(k, \ell)$

$$\begin{aligned}J(k, \ell) &\triangleq \frac{\frac{\partial}{\partial \sigma_\epsilon^2} E\{\text{SNR}_{\text{out}}(k, \ell)\}}{E\{\text{SNR}_{\text{out}}(k, \ell)\}_{|\sigma_\epsilon^2=0}} \\ &= \frac{\mathbf{W}^H(k) \mathbf{W}(k)}{\mathbf{W}^H(k) \mathbf{A}(k) \mathbf{A}^H(k) \mathbf{W}(k)}.\end{aligned}\quad (47.78)$$

The resulting expression is the reciprocal of the white-noise gain of the array. Using the array constraint $\mathbf{W}^H(k) \mathbf{A}(k) = 1$ we finally obtain the following expression for the sensitivity of the array to the ATFs perturbation,

$$J(k, \ell) = \mathbf{W}^H(k) \mathbf{W}(k). \quad (47.79)$$

Specifically, the array sensitivity is equal to the norm of the beamformer weights. Hence, reducing the sensitivity of the array is equivalent to constraining the norm of the array filter coefficients. Due to (47.15) the array filters can be decomposed into two orthogonal filters, $\mathbf{W}(k) = \mathbf{W}_0(k) - \mathbf{V}(k) = \mathbf{W}_0(k) - \mathcal{H}(k) \mathbf{G}(k, \ell)$. It is therefore sufficient to constrain the adaptive filter norm, namely $\mathbf{G}^H(k, \ell) \mathbf{G}(k, \ell) = \|\mathbf{G}(k, \ell)\|^2 \leq \Omega(k, \ell)$, where $\Omega(k, \ell)$ is a prespecified norm. The GSC structure is

modified to fulfil the norm constraint, as follows:

$$\tilde{\mathbf{G}}'(k) = \mathbf{G}(k, \ell) + \mu \frac{\mathbf{U}(k, \ell) \mathbf{Y}^*(k, \ell)}{P_{\text{est}}(k, \ell)}, \quad (47.80)$$

$$\tilde{\mathbf{G}}(\ell + 1, k) = \begin{cases} \tilde{\mathbf{G}}'(k) & \|\tilde{\mathbf{G}}'(k)\|^2 \leq \Omega(k, \ell + 1) \\ \frac{\sqrt{\Omega(k, \ell + 1)}}{\|\tilde{\mathbf{G}}'(k)\|} \tilde{\mathbf{G}}'(k) & \text{otherwise.} \end{cases} \quad (47.81)$$

Finally, the conventional **FIR** constraint is imposed on the norm-constrained filters

$$\mathbf{G}(\ell + 1, k) \stackrel{\text{FIR}}{\leftarrow} \tilde{\mathbf{G}}(\ell + 1, k).$$

Based on this concept, *Hoshuyama* et al. [47.55, 69] proposed several methods for addressing the robustness issue, concentrating on the self-cancellation phenomenon, caused by the leakage of the desired speech signal to the **BM** outputs $\mathbf{U}(k, \ell)$. This phenomenon is emphasized in reverberant environments, in the case where the **BM** only compensates for the relative delay [as in (47.34)]. In general there are two ways to mitigate this leakage problem. First, an improved spatial filtering can be incorporated into the design of the **BM**. *Claesson* and *Nordholm* [47.22] proposed to apply spatial high-pass filter to cancel out all signals within a specified frequency and angular range. *Huarnig* and *Yeh* [47.45] analyzed the leakage phenomenon and applied a derivative constraint on the array response, yielding wider tolerance to pointing errors.

A second cure for the leakage problems involves applying constraints on the **ANC** filters. *Hoshuyama* et al. [47.55] proposed several structures combining modifications for both the **BM** and the **ANC** blocks. The conventional delay-compensation **BM** is replaced by an adaptive **BM** based on signal cancellers. Two constraining strategies may be applied to the involved filters. The first strategy uses norm-constraint, and the second uses the leaky **LMS** adaptation scheme. *Haykin* [47.1] proved that both strategies are equivalent. The modified **BM** outputs, for $m = 1, \dots, M$, are given by

$$\mathbf{U}_m(k, \ell) = \mathbf{Z}_m(k, \ell) - \mathbf{H}_m^*(k, \ell) \mathbf{Y}_{\text{FBF}}(k, \ell), \quad (47.82)$$

where $\mathbf{H}_m(k, \ell)$ are updated as to minimize the power of $\mathbf{U}_m(k, \ell)$, by cancelling all desired speech components. Whenever, the **SNR** in $\mathbf{Y}_{\text{FBF}}(k, \ell)$ is sufficiently high, the blocking ability of the structure is improved.

Gannot [47.57] showed that this equation, in conjunction with the expression for the beamformer output

$$\mathbf{Y}(k, \ell) = \mathbf{Y}_{\text{FBF}}(k, \ell) - \mathbf{G}^H(k, \ell) \mathbf{U}(k, \ell),$$

is closely related to the decorrelation criterion proposed by *Weinstein* et al. [47.56]. A different decorrelation based structure was later proposed by *Fancourt* and *Parra* [47.70].

Two alternative schemes are proposed for adapting the filters $\mathbf{H}_m(k, \ell)$. (Originally, *Hoshuyama* et al. [47.55] stated their formulation in the time domain using the original **GSC** structure. Here we state the frequency-domain counterpart of the proposed algorithm. The first to propose frequency domain implementation of *Hoshuyama*'s concepts were *Herbordt* and *Kellermann* [47.71, 72].) The first scheme is the leaky **LMS**,

$$\begin{aligned} \mathbf{H}_m(k, \ell + 1) &= (1 - \delta) \mathbf{H}_m(k, \ell) \\ &+ \frac{\mu_h}{|\mathbf{Y}_{\text{FBF}}(k, \ell)|^2} \mathbf{U}_m(k, \ell) \mathbf{Y}_{\text{FBF}}^*(k, \ell) \end{aligned} \quad (47.83)$$

for $m = 1, \dots, M$. The regular **FIR** constraint, omitted for the clarity of the exposition, is then applied. The second scheme constrains the filter coefficients to a predefined mask, yielding for $m = 1, \dots, M$:

$$\begin{aligned} \mathbf{H}_m'(k, \ell + 1) &= \mathbf{H}_m(k, \ell) + \frac{\mu_h}{|\mathbf{Y}_{\text{FBF}}(k, \ell)|^2} \mathbf{U}_m(k, \ell) \mathbf{Y}_{\text{FBF}}^*(k, \ell) \end{aligned} \quad (47.84)$$

and

$$\begin{aligned} \mathbf{H}(\ell + 1, k) &= \begin{cases} \phi_{\text{low}}(k, \ell + 1) & \mathbf{H}_m'(k, \ell + 1) \geq \phi_{\text{low}}(k, \ell + 1), \\ \phi_{\text{up}}(k, \ell + 1) & \mathbf{H}_m'(k, \ell + 1) \leq \phi_{\text{up}}(k, \ell + 1), \\ \mathbf{H}_m'(k, \ell + 1) & \text{otherwise.} \end{cases} \end{aligned} \quad (47.85)$$

The **ANC** filter is either adapted by the leaky **LMS** algorithm or the norm-constrained adaptation mechanism proposed by *Cox* (see (47.81)). As a concluding remark summarizing *Hoshuyama*'s methods, we draw the reader attention to the resemblance of the proposed adaptation of the **BM** filters and the subspace tracking procedure presented by *Affes* and *Grenier* depicted in (47.46).

Spriet et al. [47.33] adopted a different approach to mitigating the leakage problem, by modifying the adaptation criterion for the **ANC** filters, $\mathbf{G}(k, \ell)$. The minimization criterion in (47.25) is altered to deal with the leakage problem. Let, $\mathbf{Y}_{\text{FBF}}^s(k, \ell)$ be the speech component at the **FBF** output. Let $\mathbf{U}^s(k, \ell)$ and $\mathbf{U}^n(k, \ell)$ be the speech and noise (without distinction between stationary and transient noise signals) components in the

reference signals, respectively. Then, the filters $\mathbf{G}(k, \ell)$ minimize the following expression

$$E\{\|Y_{\text{FBF}}^s(k, \ell) - \mathbf{G}^H(k, \ell)(\mathbf{U}^s(k, \ell) + \mathbf{U}^n(k, \ell))\|^2\}.$$

Since the speech and noise signals are uncorrelated, the above expression can be restated as

$$E\{\|\mathbf{G}^H(k, \ell)\mathbf{U}^n(k, \ell)\|^2\} + E\{\|Y_{\text{FBF}}^s(k, \ell) - \mathbf{G}^H(k, \ell)\mathbf{U}^s(k, \ell)\|^2\}. \quad (47.86)$$

Note, that the first term is related to the noise signal and the second term to the speech distortion. Hence, the Wiener filter design criterion can be easily generalized [47.26] to allow for a trade-off between speech distortion and NR, by incorporating a weighting factor $\mu \in [0, \infty)$. The resulting criterion is then given by

$$\mu E\{\|\mathbf{G}^H(k, \ell)\mathbf{U}^n(k, \ell)\|^2\} + E\{\|Y_{\text{FBF}}^s(k, \ell) - \mathbf{G}^H(k, \ell)\mathbf{U}^s(k, \ell)\|^2\}. \quad (47.87)$$

It is easily verified that the corresponding minimizer is

$$\mathbf{G}(k, \ell) = \left(\frac{1}{\mu} \Phi_{U^s U^s} + \Phi_{U^n U^n} \right)^{-1} \Phi_{U Y_{\text{FBF}}^s}, \quad (47.88)$$

where

$$\begin{aligned} \Phi_{U^s U^s} &= E\{\mathbf{U}^s(k, \ell)(\mathbf{U}^s(k, \ell))^H\}, \\ \Phi_{U^n U^n} &= E\{\mathbf{U}^n(k, \ell)(\mathbf{U}^n(k, \ell))^H\}, \\ \Phi_{U Y_{\text{FBF}}^s} &= E\{\mathbf{U}^s(k, \ell)(Y_{\text{FBF}}^s(k, \ell))^*\}. \end{aligned}$$

47.6 Multichannel Postfiltering

Postfiltering methods for multimicrophone speech enhancement algorithms have recently attracted an increased interest. It is well known that beamforming methods yield a significant improvement in speech quality [47.9]. However, when the noise field is spatially incoherent or diffuse, the NR is insufficient [47.77] and additional postfiltering is normally required [47.78]. Furthermore, as nonstationary noise cannot, in general, be distinguished from speech signals, a significant performance degradation is expected in nonstationary noise environment.

Most multimicrophone speech enhancement methods consist of a multichannel part (either delay and sum

Using the reference signals definitions we have

$$\begin{aligned} \Phi_{U^s U^s} &= \mathcal{H}^H(k, \ell) \Phi_{SS}(k, \ell) \mathcal{H}(k, \ell), \\ \Phi_{U^n U^n} &= \mathcal{H}^H(k, \ell) \Phi_{NN}(k, \ell) \mathcal{H}(k, \ell), \\ \Phi_{U^s Y_{\text{FBF}}^s} &= \mathcal{H}^H(k, \ell) \Phi_{SS}(k, \ell) \mathbf{W}_0(k, \ell). \end{aligned} \quad (47.89)$$

Since $\Phi_{SS}(k, \ell)$ is not available it can be evaluated using

$$\Phi_{SS}(k, \ell) = \Phi_{ZZ}(k, \ell) - \Phi_{NN}(k, \ell)$$

and $\Phi_{NN}(k, \ell)$ is estimated while the speech signal is absent. *Doclo* and *Moonen* [47.73] prove that the output SNR after NR with the above speech distortion weighted multichannel Wiener filter (SDW-MWF) is always larger than or equal to the input SNR, for any filter length, and for any value of the trade-off parameter μ between NR and speech distortion.

This solution for the ANC filters constitutes the speech distortion regularized generalized sidelobe canceller (SDR-GSC) structure. *Spriet* et al. [47.33] further proposed to incorporate a single-channel postfilter, which compensates for the distortion imposed by the structure in case of speech leakage into the reference signals. Further discussion of this structure is beyond the scope of this survey. In [47.74], the authors propose a stochastic gradient-based implementation of their criterion. The robustness of both the multichannel Wiener filter and the GSC structures are analyzed by *Spriet* et al. [47.75] in the context of hearing-aid application.

Improving the robustness of the BM is an ongoing research topic. An interesting direction was taken by *Low* et al. [47.76]. The authors propose to incorporate concepts adopted from the BSS discipline to improve the separation of the speech and noise signals and hence reducing the amount of leakage of the desired signal into the reference noise signals.

beamformer or GSC [47.36]) followed by a postfilter, which is based on Wiener filtering (sometimes in conjunction with spectral subtraction). Numerous articles have been published on the subject, e.g., [47.79–87] to mention just a few.

In general, the postfilters can be divided into two groups. The first is a single-channel postfilter which acts as a single-microphone speech enhancement algorithm on the beamformer output. Multichannel postfilters, on the other hand, explicitly use the spatial information, extracted by the GSC structure, to gain better distinction between the speech signal and the transient noise.

47.6.1 MMSE Postfiltering

Simmer et al. [47.78] address the general problem of the single-channel postfilter. They first derive the multichannel Wiener filter for estimating the speech signal, $S(k, \ell)$ from the microphone signals $\mathbf{Z}(k, \ell)$ given in (47.3). Then, they show that the Wiener filter can be factorized into a multiplication of the **LCMV** (Frost) beamformer given in (47.7) and a single-channel Wiener filter that depends on the output speech and noise signals.

To show this, we will start with the multichannel Wiener filter, which is given by

$$\mathbf{W}^{\text{Wiener}}(k, \ell) = \Phi_{\mathbf{Z}\mathbf{Z}}^{-1}(k, \ell) \Phi_{\mathbf{Z}S}(k, \ell), \quad (47.90)$$

where $\Phi_{\mathbf{Z}\mathbf{Z}}$ is given in (47.10) and

$$\begin{aligned} \Phi_{\mathbf{Z}S}(k, \ell) &= E\{\mathbf{Z}(k, \ell) S^*(k, \ell)\} \\ &= \mathbf{A}(k) \phi_{ss}(k, \ell). \end{aligned} \quad (47.91)$$

Hence,

$$\begin{aligned} \mathbf{W}^{\text{Wiener}}(k, \ell) &= (\phi_{ss}(k, \ell) \mathbf{A}(k) \mathbf{A}^H(k) + \Phi_{\mathbf{N}\mathbf{N}}(k, \ell))^{-1} \\ &\quad \times \phi_{ss}(k, \ell) \mathbf{A}(k). \end{aligned} \quad (47.92)$$

Omitting, for the clarity of the exposition, the explicit time- and frequency-domain dependence and using the matrix inversion lemma yields

$$\begin{aligned} \mathbf{W}^{\text{Wiener}}(k, \ell) &= \left(\Phi_{\mathbf{N}\mathbf{N}}^{-1} - \frac{\phi_{ss} \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A} \mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}} \right) \phi_{ss} \mathbf{A} \\ &= \left(1 - \frac{\phi_{ss} \mathbf{A} \mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}} \right) \Phi_{\mathbf{N}\mathbf{N}}^{-1} \phi_{ss} \mathbf{A} \\ &= \left(\frac{\phi_{ss}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}} \right) \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A} \\ &= \left(\frac{\phi_{ss}}{\phi_{ss} + (\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A})^{-1}} \right) \frac{\Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}}{\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}}. \end{aligned} \quad (47.93)$$

(The derivation here is a slight modification of the results introduced in [47.78].) The reader can easily identify the second multiplier as the **MSNR** beamformer (47.8) which was shown to be equivalent to the **LCMV** beamformer (47.11).

We turn now to analyzing the first term in the multiplicative expression. The **PSD** of desired signal component at the output of the **LCMV** beamformer is

given by

$$\begin{aligned} \phi_{Y_s Y_s}(k, \ell) &= \phi_{ss} (\mathbf{W}^{\text{LCMV}})^H \mathbf{A} \mathbf{A}^H \mathbf{W}^{\text{LCMV}} \\ &= \phi_{ss} \left(\frac{\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}}{\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}} \right)^2 = \phi_{ss}. \end{aligned} \quad (47.94)$$

As expected, the **LCMV** is a distortionless beamformer. The noise component at the beamformer output is given by,

$$\begin{aligned} \phi_{Y_n Y_n}(k, \ell) &= (\mathbf{W}^{\text{LCMV}})^H \Phi_{\mathbf{N}\mathbf{N}} \mathbf{W}^{\text{LCMV}} = \frac{\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}}{(\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A})^2} \\ &= \frac{1}{\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A}}. \end{aligned} \quad (47.95)$$

Using (47.94) and (47.95), the first term in the multichannel Wiener filter can be rewritten as

$$\begin{aligned} &\frac{\phi_{ss}}{\phi_{ss} + (\mathbf{A}^H \Phi_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{A})^{-1}} \\ &= \frac{\phi_{Y_s Y_s}(k, \ell)}{\phi_{Y_n Y_n}(k, \ell) + \phi_{Y_s Y_s}(k, \ell)}, \end{aligned} \quad (47.96)$$

which is evidently recognized as the single-channel Wiener filter applied to the **LCMV** beamformer output. It was therefore proven that the multichannel Wiener filter can be factorized into a product of the **LCMV** beamformer and a single-channel Wiener postfilter applied to the beamformer output,

$$\begin{aligned} \mathbf{W}^{\text{Wiener}}(k, \ell) &= \underbrace{\frac{\phi_{Y_s Y_s}(k, \ell)}{\phi_{Y_n Y_n}(k, \ell) + \phi_{Y_s Y_s}(k, \ell)}}_{\text{Wiener postfilter}} \\ &\quad \times \underbrace{\frac{\Phi_{\mathbf{N}\mathbf{N}}^{-1}(k, \ell) \mathbf{A}(k)}{\mathbf{A}^H(k) \Phi_{\mathbf{N}\mathbf{N}}^{-1}(k, \ell) \mathbf{A}(k)}}_{\text{LCMV beamformer}}. \end{aligned} \quad (47.97)$$

Several algorithms have been proposed for designing the postfilter; all differ in their treatment of the single-channel Wiener postfilter estimation. Zelinski [47.79] was probably the first to apply a postfilter to the output of a microphone array (delay and sum beamformer in his formulation). Zelinski proposed the following Wiener

filter estimation

$$\begin{aligned} W^{\text{Zelinski}}(k, \ell) &= \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^M \text{Re}[Z_i(k, \ell) Z_j^*(k, \ell)]}{\frac{1}{M} \sum_{i=1}^M |Z_i(k, \ell)|^2} \end{aligned} \quad (47.98)$$

Later, postfiltering was incorporated into the Griffiths and Jim **GSC** beamformer by Bitzer et al. [47.86, 87]. The authors proposed to use two postfilters in succession. The first is applied to the **FBF** branch, and the second to the **GSC** output. In directional noise source and in the low-frequency band of a diffused noise field, correlation between the noise components at each sensor exists. While the first postfilter is useless in this case, the latter suppresses the noise.

Simmer and *Wasiljeff* [47.88] showed that Zelinski's postfilter has two disadvantages. First, only minor **SNR** improvement can be expected in frequencies, for which the coherence function between the received noise signals is high (i.e., coherent noise sources). Second, for frequencies with a low coherence function, the noise **PSD** is overestimated by a factor M (the number of microphones). They propose to mitigate the second disadvantage by slightly modifying the noise estimation, to obtain an estimate of the noise **PSD** at the output of the beamformer, rather than at its input.

In diffused noise field the noise coherence function tends to be high in lower frequencies, whereas in higher-frequency bands it tends to be low. The cutoff frequency depends on the distance between microphones (see further discussion in Sect. 47.7). This property is the cause for the first drawback of Zelinski's postfilter. It was therefore proposed by *Fischer* and *Kammeyer* [47.83] to split the beamformer into three nonoverlapping subarrays (with different inter-microphone distances) for which the noise coherence function is kept low. To avoid grating lobes, bandpass filters with corresponding cutoff frequencies, are applied to the beamformer output. *Marro* et al. [47.43] improved this concept and further modified the Wiener postfilter estimation. A comprehensive survey of these postfiltering methods can be found in [47.78]. *McCowan* and *Boulevard* [47.89, 90] develop a more-general expression of the postfilter estimation, based on an assumed knowledge of the complex coherence function of the noise field. This general expression can be used to construct a more-appropriate postfilter in a variety of different noise fields.

47.6.2 Log-Spectral Amplitude Postfiltering

A major drawback of single-channel postfiltering techniques is that highly nonstationary noise components are not addressed. The time variation of the interfering signals is assumed to be sufficiently slow, such that the postfilter can track and adapt to the changes in the noise statistics. Unfortunately, transient interferences are often much too brief and abrupt for the conventional tracking methods.

Transient Beam-to-Reference Ratio

Generally, the **TF-GSC** output comprises three components: a nonstationary desired source component, a pseudostationary noise component, and a transient interference. Our objective is to determine which category a given time–frequency bin belongs to, based on the beamformer output and the reference signals.

Recall the three hypotheses H_{0s} , H_{0t} , and H_1 that indicate, respectively, the absence of transients, the presence of an interfering transient, and the presence of a desired source transient at the beamformer output (the pseudostationary interference is present in any case). Then, if transients have not been detected at the beamformer output and the reference signals, we can accept the H_{0s} hypothesis. If a transient is detected at the beamformer output but not at the reference signals, the transient is likely a source component and therefore we determine that H_1 is true. On the contrary, a transient that is detected at one of the reference signals but not at the beamformer output is likely an interfering component, which implies that H_{0t} is true. If a transient is simultaneously detected at the beamformer output and at one of the reference signals, a further test is required, which involves the ratio between the transient power at beamformer output and the transient power at the reference signals. The discussion here is partly based on [47.77]. A real-time version of the method that incorporates adaptive estimation of the **ATFs** is introduced in [47.91].

Let $\mathcal{S}$ be a smoothing operator in the power-spectral domain,

$$\begin{aligned} \mathcal{S}Y(k, \ell) &= \alpha_s \cdot \mathcal{S}Y(k, \ell - 1) \\ &\quad + (1 - \alpha_s) \sum_{i=-w}^w b_i |Y(k - i, \ell)|^2, \end{aligned} \quad (47.99)$$

where α_s ($0 \leq \alpha_s \leq 1$) is a forgetting factor for the smoothing in time, and b is a normalized window function ($\sum_{i=-w}^w b_i = 1$) that determines the order of

smoothing in frequency. Let $\mathcal{M}$ denote an estimator for the PSD of the background pseudostationary noise, derived using the MCRA approach [47.66,67]. The decision rules for detecting transients at the TF-GSC output and reference signals are

$$\Lambda_Y(k, \ell) \triangleq \mathcal{Y}(k, \ell) / \mathcal{M}Y(k, \ell) > \Lambda_0, \quad (47.100)$$

$$\Lambda_U(k, \ell) \triangleq \max_{2 \leq i \leq M} \left\{ \frac{\mathcal{Y}U_i(k, \ell)}{\mathcal{M}U_i(k, \ell)} \right\} > \Lambda_1, \quad (47.101)$$

respectively, where Λ_Y and Λ_U denote measures of the local nonstationarities, and Λ_0 and Λ_1 are the corresponding threshold values for detecting transients [47.92]. The transient beam-to-reference ratio (TBRR) is defined by the ratio between the transient power of the beamformer output and the transient power of the strongest reference signal

$$\Omega(k, \ell) = \frac{\mathcal{Y}(k, \ell) - \mathcal{M}Y(k, \ell)}{\max_{2 \leq i \leq M} \{\mathcal{Y}U_i(k, \ell) - \mathcal{M}U_i(k, \ell)\}}. \quad (47.102)$$

Transient signal components are relatively strong at the beamformer output, whereas transient noise components are relatively strong at one of the reference signals. Hence, we expect $\Omega(k, \ell)$ to be large for signal transients, and small for noise transients. Assuming there exist thresholds $\Omega_{\text{high}}(k)$ and $\Omega_{\text{low}}(k)$ such that

$$\Omega(k, \ell)|_{H_{0t}} \leq \Omega_{\text{low}}(k) \leq \Omega_{\text{high}}(k) \leq \Omega(k, \ell)|_{H_1} \quad (47.103)$$

the decision rule for differentiating desired signal components from the transient interference components is

$$\begin{aligned} H_{0t} : \gamma_s(k, \ell) \leq 1 \text{ or } \Omega(k, \ell) \leq \Omega_{\text{low}}(k), \\ H_1 : \gamma_s(k, \ell) \geq \gamma_0 \text{ and } \Omega(k, \ell) \geq \Omega_{\text{high}}(k), \\ H_r : \text{otherwise,} \end{aligned} \quad (47.104)$$

where

$$\gamma_s(k, \ell) \triangleq \frac{|Y(k, \ell)|^2}{\mathcal{M}Y(k, \ell)} \quad (47.105)$$

represents the a posteriori SNR at the beamformer output with respect to the pseudostationary noise, γ_0 denotes a constant satisfying $\mathcal{P}(\gamma_s(k, \ell) \geq \gamma_0 | H_{0s}) < \epsilon$ for a certain significance level ϵ , and H_r designates a reject option where the conditional error of making a decision between H_{0t} and H_1 is high.

Figure 47.5 summarizes a block diagram for the hypothesis testing. The hypothesis testing is carried out in the time–frequency plane for each frame and frequency bin. H_{0s} is accepted when transients have neither been detected at the beamformer output nor at the reference signals. If a transient is detected at the beamformer output but not at the reference signals, we accept H_1 . On the other hand, if a transient is detected at one of the reference signals but not at the beamformer output, we accept H_{0t} . If a transient is detected simultaneously at the beamformer output and at one of the reference signals, we compute the TBRR $\Omega(k, \ell)$ and the a posteriori SNR at the beamformer output with respect to the pseudostationary noise $\gamma_s(k, \ell)$ and decide on the hypothesis according to (47.104).

Log-Spectral Amplitude Estimation

We address now the problem of estimating the time-varying PSD of the TF-GSC output noise component, and present the multichannel postfiltering technique. Figure 47.6 describes a block diagram of the multichannel postfiltering. Following the hypothesis testing, an estimate $\hat{q}(k, \ell)$ for the a priori signal absence probability is produced. Subsequently, we derive an estimate $p(k, \ell) \triangleq \mathcal{P}(H_1 | Y, U)$ for the signal presence probability, and an estimate $\hat{\lambda}_d(k, \ell)$ for the noise PSD.

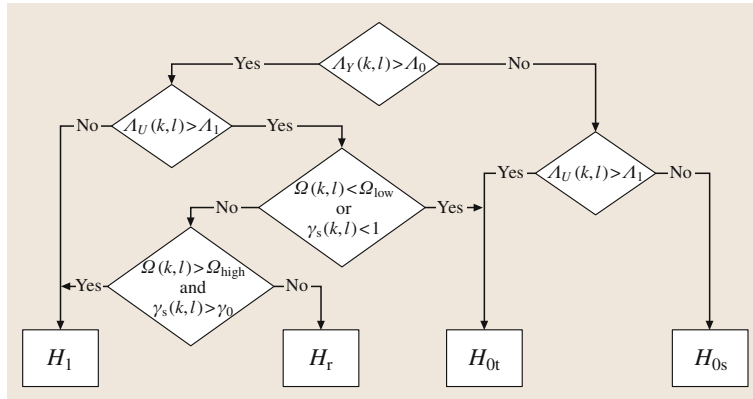


Fig. 47.5 Block diagram for hypothesis testing

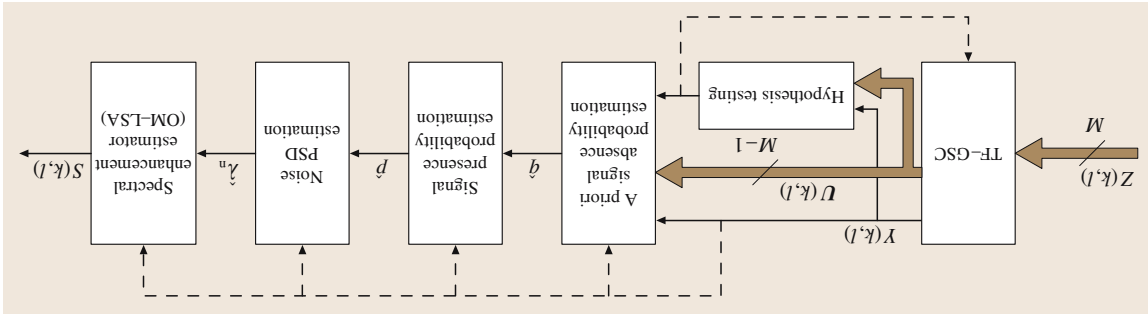


Fig. 47.6 Block diagram of multichannel postfiltering

Finally, spectral enhancement of the beamformer output is achieved by applying the optimally-modified log-spectral amplitude (OM-LSA) gain function [47.93], which minimizes the MSE of the log-spectral amplitude under signal presence uncertainty.

Based on a Gaussian statistical model [47.94], the signal presence probability is given by

$$p(k, \ell) = \left\{ 1 + \frac{q(k, \ell)}{1 - q(k, \ell)} [1 + \xi(k, \ell)] \exp[-v(k, \ell)] \right\}^{-1}, \quad (47.106)$$

where $\xi(k, \ell) \triangleq \lambda_s(k, \ell)/\lambda_n(k, \ell)$ is the a priori SNR, $\lambda_s(k, \ell)$ is the desired signal PSD at the beamformer output $\lambda_n(k, \ell)$ is the noise PSD at the beamformer output, $v(k, \ell) \triangleq \gamma(k, \ell)\xi(k, \ell)/[1 + \xi(k, \ell)]$, and $\gamma(k, \ell) \triangleq |Y(k, \ell)|^2/\lambda_n(k, \ell)$ is the a posteriori SNR. The a priori signal absence probability $\hat{q}(k, \ell)$ is set to 1 if the signal absence hypotheses (H_0 or H_0^i) are accepted, and is set to 0 if the signal presence hypothesis (H_1) is accepted. In the case of the reject hypothesis H_1 , a soft signal detection is accomplished by letting $\hat{q}(k, \ell)$ be inversely proportional to $\hat{\sigma}(k, \ell)$ and $\gamma_s(k, \ell)$:

$$\hat{q}(k, \ell) = \max \left\{ \gamma_0 - \gamma_s(k, \ell), \frac{\hat{\sigma}_{\text{high}} - \hat{\sigma}_{\text{low}}}{\gamma_0 - 1} \right\}.$$

The a priori SNR is estimated by [47.93]

$$\hat{\xi}(k, \ell) = \alpha K_2^{H_1}(k, \ell - 1) \gamma(k, \ell - 1) + (1 - \alpha) \max \{\gamma(k, \ell) - 1, 0\}, \quad (47.108)$$

where α is a weighting factor that controls the trade-off between NR and signal distortion, and

$$K_{H_1}(k, \ell) \triangleq \frac{\hat{\xi}(k, \ell)}{1 + \hat{\xi}(k, \ell)} \exp \left(\frac{1}{2} \int_{-\infty}^{v(k, \ell)} \frac{t}{e^{-t}} dt \right) \quad (47.109)$$

Table 47.3 Values of the parameters used in the implementation of the log-spectral amplitude postfiltering for a sampling rate of 8 kHz

Normalized LMS:	$\alpha^p = 0.9$	$\mu_n = 0.05$
ATF identification:	$N = 10$	$R = 10$
Hypothesis testing:	$\alpha_s = 0.9$	$\gamma_0 = 4.6$
	$\lambda_0 = 1.67$	$\lambda_1 = 1.81$
	$\hat{\sigma}_{\text{low}} = 1$	$\hat{\sigma}_{\text{high}} = 3$
Noise PSD estimation:	$\alpha_n = 0.85$	$\beta = 1.47$
Spectral enhancement:	$\alpha = 0.92$	$K_{\text{min}} = -20 \text{ dB}$

and β is a factor that compensates the bias when signal is absent. The constant α_n ($0 < \alpha_n < 1$) represents the minimal smoothing parameter value. The smoothing parameter is close to 1 when signal is present, to prevent an increase in the noise estimate as a result of signal components. It decreases when the probability of signal presence decreases, to allow a fast update of the noise estimate.

where the smoothing parameter $\hat{\alpha}_n(k, \ell)$ is determined by the signal presence probability $p(k, \ell)$,

$$\hat{\alpha}_n(k, \ell) \triangleq \alpha_n + (1 - \alpha_n) p(k, \ell), \quad (47.111)$$

$$\hat{\lambda}_n(k, \ell + 1) = \hat{\alpha}_n(k, \ell) \hat{\lambda}_n(k, \ell) + \beta [1 - \hat{\alpha}_n(k, \ell)] |Y(k, \ell)|^2, \quad (47.110)$$

by smoothing parameter. The recursive averaging is given by

The estimate of the clean signal **STFT** is finally given by

$$\hat{S}(k, \ell) = K(k, \ell)Y(k, \ell), \quad (47.112)$$

where

$$K(k, \ell) = \{K_{H_1}(k, \ell)\}^{p(k, \ell)} K_{\min}^{1-p(k, \ell)} \quad (47.113)$$

is the OM-LSA gain function and $K_{\min}$ denotes a lower bound constraint for the gain when signal is ab-

sent. The implementation of the integrated **TF-GSC** and multichannel postfiltering algorithm is summarized in Fig. 47.6.

Typical values of the respective parameters, for a sampling rate of 8 kHz, are given in Table 47.3. The **STFT** and its inverse are implemented with biorthogonal Hamming windows of 256 samples length (32 ms) and 64 samples frame update step (75% overlapping windows).

47.7 Performance Analysis

The use of actual signals (such as noisy speech recordings in room environment) demonstrates the ability of the **TF-GSC** algorithm to reduce the noise while maintaining the desired signal spectral content (Sect. 47.8). However, it is also beneficial to perform analytical evaluation of the expected performance, especially for determining the performance limits. While the **D-GSC** [47.36] is widely analyzed in the literature, the more-realistic arbitrary **ATFs** scenario, is only superficially treated. In more-complex environments such as reverberating room this assumption is not valid, and may result in severe degradation in the performance. Furthermore, most references address the NR obtained by the algorithm but do not present any measure of distortion imposed on the desired signal, even in the simple delay-only **ATFs**. In this section we will analyze both the NR of the **TF-GSC** structure (while using the RTFs rather than the **ATFs** themselves) and the distortion it imposes. For a thorough performance evaluation of the multichannel postfilter (for the two-channel case) please refer to [47.96].

47.7.1 The Power Spectral Density of the Beamformer Output

Using (47.24) and (47.37), the algorithm's output is given by

$$\begin{aligned} Y(k, \ell) \\ = \frac{\hat{\mathbf{A}}^H(k)}{\|\hat{\mathbf{A}}(k)\|^2} \mathbf{Z}(k, \ell) - \mathbf{G}^H(k, \ell) \hat{\mathcal{H}}^H(k) \mathbf{Z}(k, \ell), \end{aligned}$$

where only estimates of the RTFs, $\hat{\mathbf{A}}(k)$ [and $\hat{\mathcal{H}}(k)$], rather than their exact values, are assumed to be known. Using this expression, the **PSD** of the output signal is

given by

$$\begin{aligned} \Phi_{yy}(k, \ell) &= E[Y(k, \ell)Y^*(k, \ell)] \\ &= E \left\{ \left[\frac{1}{\|\hat{\mathbf{A}}(k)\|^2} \hat{\mathbf{A}}^H(k) \mathbf{Z}(k, \ell) \right. \right. \\ &\quad \left. \left. - \mathbf{G}^H(k, \ell) \hat{\mathcal{H}}^H(k) \mathbf{Z}(k, \ell), \right] \right. \\ &\quad \times \left[\frac{1}{\|\hat{\mathbf{A}}(k)\|^2} \hat{\mathbf{A}}^H(k) \mathbf{Z}(k, \ell) \right. \\ &\quad \left. \left. - \mathbf{G}^H(k, \ell) \hat{\mathcal{H}}^H(k) \mathbf{Z}(k, \ell), \right]^H \right\}. \end{aligned} \quad (47.114)$$

Opening brackets and using the **PSD** definition $\Phi_{ZZ}(k, \ell) = E\{\mathbf{Z}(k, \ell)\mathbf{Z}^H(k, \ell)\}$ yields,

$$\begin{aligned} \Phi_{yy}(k, \ell) &= \frac{1}{\|\hat{\mathbf{A}}(k)\|^4} \hat{\mathbf{A}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\mathbf{A}}(k) \\ &\quad - \frac{1}{\|\hat{\mathbf{A}}(k)\|^2} \mathbf{G}^H(k, \ell) \hat{\mathcal{H}}^H(k) \\ &\quad \times \Phi_{ZZ}(k, \ell) \hat{\mathbf{A}}(k) \\ &\quad - \frac{1}{\|\hat{\mathbf{A}}(k)\|^2} \hat{\mathbf{A}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\mathcal{H}}(k) \mathbf{G}(k) \\ &\quad + \mathbf{G}^H(k, \ell) \hat{\mathcal{H}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\mathcal{H}}(k) \mathbf{G}(k). \end{aligned} \quad (47.115)$$

The output **PSD** depends on the input signal $\mathbf{Z}(k, \ell)$ and the optimal multichannel Wiener filter, given by (47.26), calculated during frames for which hypothesis H_0 s is true. Using the independence of the desired signal and the noise signal, the NR and the distortion imposed by the algorithm can be calculated separately by deriving

expressions for the output PSD in the following two situations:

$$\Phi_{ZZ}(k, \ell) = \begin{cases} \Phi_{N_s N_s}(k, \ell) & H_{0s}, \\ \phi_{ss}(k, \ell) A(k) A^H(k) & H_1, \end{cases}$$

yielding

$$\Phi_{yy}(k, \ell) = \begin{cases} \Phi_{yy}^n(k, \ell) & H_{0s} \Rightarrow \text{noise reduction,} \\ \Phi_{yy}^s(k, \ell) & H_1 \Rightarrow \text{distortion.} \end{cases}$$

Note that for simplicity we calculate the NR only during H_{0s} , i.e., while only stationary noise signal is present. For a performance evaluation in the nonstationary case, please refer to [47.96].

Using (47.26) and (47.115), we obtain the output signal PSD:

$$\begin{aligned} \Phi_{yy}(k, \ell) = & \frac{1}{\|\hat{\hat{A}}(k)\|^4} \times \left\{ \hat{\hat{A}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\hat{A}}(k) \right. \\ & - \hat{\hat{A}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) \\ & \left[\hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) \right]^{-1} \\ & \hat{\mathcal{H}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\hat{A}}(k) \\ & - \hat{\hat{A}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\mathcal{H}}(k) \\ & \left[\hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) \right]^{-1} \\ & \hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\hat{A}}(k) \\ & + \hat{\hat{A}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) \\ & \left[\hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) \right]^{-1} \\ & \hat{\mathcal{H}}^H(k) \Phi_{ZZ}(k, \ell) \hat{\mathcal{H}}(k) \\ & \left. \times \left[\hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) \right]^{-1} \right. \\ & \left. \hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\hat{A}}(k) \right\}. \quad (47.116) \end{aligned}$$

This complicated expression depends on various parameters: the input signal PSD, $\Phi_{ZZ}(k, \ell)$, the stationary noise PSD used for calculating the optimal filters, $\Phi_{N_s N_s}(k, \ell)$, and the RTFs estimate $\hat{\hat{A}}(k)$ [which is also used for the BM $\hat{\mathcal{H}}(k)$]. This expression will be used in Subsects. 47.7.2 and 47.7.3 for deriving general expressions for the distortion imposed by the algorithm and the obtainable NR, respectively. These general expressions will be evaluated for several interesting cases.

47.7.2 Signal Distortion

The distortion imposed by the algorithm can be calculated by the general expression given in (47.116)

for a signal $Z(k, \ell) = A(k)s(k, \ell)$. Assume a perfect estimate of the RTFs $\hat{\hat{A}}(k)$, is available, i.e., $\hat{\hat{A}}(k) = \hat{A}(k) = \frac{A(k)}{A_1(k)}$. Thus, using the signal PSD expression $\Phi_{ZZ}(k, \ell) = \phi_{ss}(k, \ell) A(k) A^H(k)$ and the identity $\mathcal{H}^H(k) A(k) = 0$, expression (47.116) reduces to

$$\begin{aligned} \Phi_{yy}^s(k, \ell) &= \phi_{ss}(k, \ell) \frac{|\mathcal{F}(k)|^2}{\|\hat{\hat{A}}(k)\|^4} \hat{\hat{A}}^H(k) A(k) A^H(k) \hat{\hat{A}}(k) \\ &= \phi_{ss}(k, \ell) |A_1(k)|^2. \end{aligned} \quad (47.117)$$

The filter $A_1(k)$ is the ATF relating the source signal and the first (arbitrarily chosen as the reference) sensor. This distortion cannot be eliminated by the algorithm. Note that this distortion is due to the use of the RTFs, rather than the ATFs themselves. Using direct estimate of the ATFs will avoid this distortion affect. Actually, it imposes on the output signal the same amount of distortion imposed on the arbitrary reference sensor. Hence, we define the total distortion caused by the algorithm by normalizing the output,

$$\text{DIS}(k, \ell) = \frac{\Phi_{yy}^s(k, \ell)}{|A_1(k)|^2 \phi_{ss}(k, \ell)}. \quad (47.118)$$

Hence, a value of $\text{DIS}(k, \ell) = 1$ indicates a distortionless output. This value is obtained whenever an exact knowledge of the RTFs is available.

The distortion level demonstrates only weak dependence on the noise field, both for the delay-only and complex ATF cases. For details please refer to [47.77].

47.7.3 Stationary Noise Reduction

We calculate now the amount of obtainable stationary NR. When the noise is nonstationary, the use of a postfilter becomes more important. A performance analysis of the multichannel postfilter (for the two-channel case) in nonstationary noise environment can be found in [47.96].

We will use again the general expression for the output signal given by (47.116), this time with a noise signal as the input signal, i.e., $Z(k, \ell) = N_s(k, \ell)$ [the same noise signal used for calculating the optimal Wiener filter (47.26)]. The expression (47.116) now reduces to

$$\begin{aligned}
& \Phi_{yy}^n(k, \ell) \\
&= \frac{\hat{\tilde{A}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\tilde{A}}(k)}{\|\hat{\tilde{A}}(k)\|^4} - \frac{\hat{\tilde{A}}^H(k)}{\|\hat{\tilde{A}}(k)\|^4} \\
&\quad \times \Phi_{N_s N_s}(k, \ell) \hat{\mathcal{H}}(k) [\hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell)]^{-1} \\
&\quad \times \hat{\mathcal{H}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\tilde{A}}(k). \quad (47.119)
\end{aligned}$$

The first term of the equation can be identified as $\Phi_{\text{FBF}}^n(k, \ell)$, the FBF component of the output PSD when a noise-only signal is applied to the array,

$$\begin{aligned}
\Phi_{\text{FBF}}^n(k, \ell) &= E\{Y_{\text{FBF}}^n(k, \ell) [Y_{\text{FBF}}^n(k, \ell)]^*\} \\
&= \frac{\hat{\tilde{A}}^H(k) \Phi_{N_s N_s}(k, \ell) \hat{\tilde{A}}(k)}{\|\hat{\tilde{A}}(k)\|^4}. \quad (47.120)
\end{aligned}$$

Another interesting figure of merit is the extra NR obtained by the noise cancelling branch (see also [47.40]),

$$\text{NR}_{\text{ANC}}(k, \ell) = \frac{\Phi_{\text{FBF}}^n(k, \ell)}{\Phi_{yy}^n(k, \ell)}. \quad (47.121)$$

Expressions (47.119), (47.120), and (47.121) can be used for calculating the NR obtained by the algorithm and to determine the major contributor for this NR. Assuming small errors regime, the error in estimating $\tilde{A}(k)$ has only minor influence on the NR. Therefore, we will assume, throughout the NR analysis, perfect knowledge of the RTFs, i.e., $\hat{\tilde{A}}(k) = \tilde{A}(k)$.

The resulting expressions for the noise cancellation depends on the noise PSD at the sensors. We calculate now the expected NR of the algorithm for three important noise fields: coherent (point source), diffused (spherically isotropic), and incoherent (noise signals generated at the sensors; e.g., amplifier noise, are assumed to be uncorrelated).

Coherent Noise Field

Assume a single stationary point source noise signal with PSD $\Phi_{n_s n_s}(k, \ell)$ and assume that $b_m(t)$ are slowly time-varying ATFs relating the noise source and the m -th sensor. Define,

$$N_s(k, \ell) = \mathbf{B}(k) N_s(k, \ell),$$

where

$$\mathbf{B}^T(k) = (B_1(k) \ B_2(k) \ \cdots \ B_M(k)).$$

The PSD matrix of the noise component at the sensors' signals is given by,

$$\Phi_{N_s N_s}(k, \ell) = \Phi_{n_s n_s}(k, \ell) \mathbf{B}(k) \mathbf{B}^H(k) + \varepsilon \mathbf{I},$$

where $\mathbf{I}$ is an $M \times M$ identity matrix, and $\varepsilon \rightarrow 0$. The last term is added for stability reasons (see Appendix 47.A). For $\mathbf{B}(k) \neq \mathbf{A}(k)$, the achievable NR is infinite, i.e.,

$$\Phi_{yy}^n(k, \ell) = 0 \quad \text{for} \quad \mathbf{B}(k) \neq \mathbf{A}(k).$$

Thus, perfect noise cancellation is achieved. The derivation of this result is given in Appendix 47.A. Note that this is not a surprising result, since for $M \geq 2$ the Wiener filter can entirely eliminate the noise component. This result is valid for all ATFs $\mathbf{B}(k)$ provided that $\mathbf{B}(k) \neq \mathbf{A}(k)$, i.e., the noise and the signal do not originate from the same point. If $\mathbf{B}(k) = \mathbf{A}(k)$ the noise and desired signal are indistinguishable and no NR is expected. If $\mathbf{B}(k) \neq \mathbf{A}(k)$ the proposed algorithm can eliminate any point source noise signal as good as the D-GSC can eliminate a directional noise signal in the delay-only propagation case (see [47.40]).

It is also interesting to explore the contribution of the FBF block to the NR,

$$\begin{aligned}
\Phi_{\text{FBF}}^n(k, \ell) &= \frac{\tilde{\mathbf{A}}^H(k) \Phi_{N_s N_s}(k, \ell) \tilde{\mathbf{A}}(k)}{\|\tilde{\mathbf{A}}(k)\|^4} \\
&= \frac{\tilde{\mathbf{A}}^H(k) \Phi_{n_s n_s}(k, \ell) \mathbf{B}(k) \mathbf{B}^H(k) \tilde{\mathbf{A}}(k)}{\|\tilde{\mathbf{A}}(k)\|^4} \\
&= \frac{\Phi_{n_s n_s}(k, \ell)}{\|\tilde{\mathbf{A}}(k)\|^4} \tilde{\mathbf{A}}^H(k) \mathbf{B}(k) [\tilde{\mathbf{A}}^H(k) \mathbf{B}(k)]^H.
\end{aligned}$$

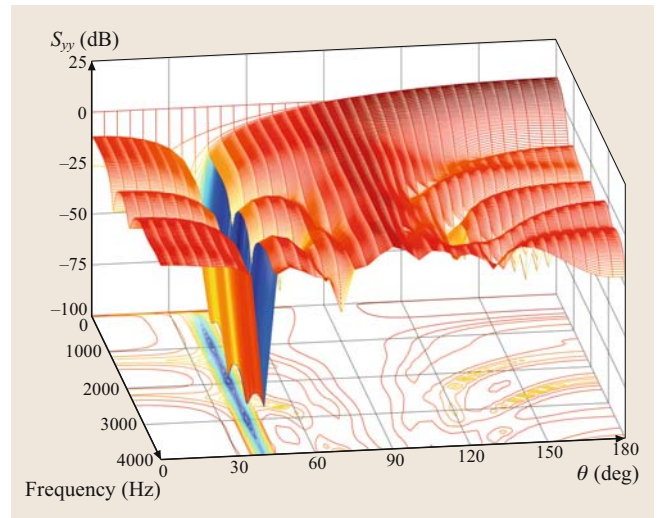


Fig. 47.7 Array output for directional noise field for linear array with $M = 5$ sensors for delay-only ATFs

Although the FBF combines the desired signal coherently, it does not necessarily improve the SNR. The infinite NR is due to the ANC branch.

The expected NR in the simple case of delay-only ATFs is shown in Fig. 47.7. We optimize a linear array of $M = 5$ microphones to cancel noise impinging on the array from the direction $\theta = 40^\circ$. We present the output signal of the array as a function of the frequency and the array steering angle θ . It is clearly shown that the main lobe is maintained (i. e., low distortion), while a null is constructed at all frequencies at the noise angle. The main lobe is wider in the lower-frequency band. This result is in good agreement with the theory, since at $\omega = 0$ rad/s there is no phase difference between the signals at the sensors. Similar results were obtained by Bitzer et al. [47.40,41]. The general ATFs case is further explored in [47.77,97].

Diffused Noise Field

In highly reverberant acoustical environment, such as a car enclosure, the noise field tends to be diffused (see for instance [47.42,98]). A diffused noise source is assumed to be equidistributed on a sphere in the far field of the array. The cross-coherence function between signals received by two sensors (i, j) with distance d_{ij} is given by

$$\Gamma_{N_i N_j}(k) = \frac{\Phi_{N_i N_j}(k)}{\sqrt{\Phi_{N_i N_i}(k) \Phi_{N_j N_j}(k)}} = \frac{\sin(kd_{ij}/c)}{kd_{ij}/c},$$

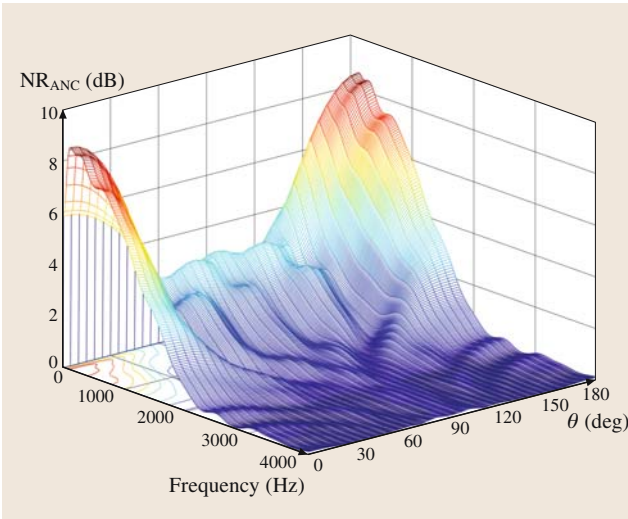


Fig. 47.8 Extra noise reduction of noise cancelling branch for diffused noise field using a linear array with $M = 5$ sensors and delay-only ATFs

where c is the speed of sound [47.98]. Thus, the coherence matrix is given by,

$$\Gamma(k) = \begin{pmatrix} 1 & \Gamma_{N_1 N_2}(k) & \cdots & \Gamma_{N_1 N_M}(k) \\ \Gamma_{N_2 N_1}(k) & 1 & \cdots & \Gamma_{N_2 N_M}(k) \\ \vdots & & \ddots & \vdots \\ \Gamma_{N_M N_1}(k) & & & 1 \end{pmatrix}.$$

The noise PSD at the sensors input is

$$\Phi_{N_s N_s}(k, \ell) = \Phi_{n_s n_s}(k, \ell) \Gamma(k).$$

Using (47.120), the noise PSD at the FBF output is given by

$$\Phi_{\text{FBF}}^n(k, \ell) = \frac{\phi_{nn}(k, \ell) \tilde{\mathbf{A}}^H(k) \Gamma(k) \tilde{\mathbf{A}}(k)}{\|\tilde{\mathbf{A}}(k)\|^4}.$$

The extra NR obtained by the ANC is given by

$$\begin{aligned} \text{NR}_{\text{ANC}}(k, \ell) &= \{1 - [\tilde{\mathbf{A}}^H(k) \Gamma(k) \mathcal{H}(k) [\mathcal{H}^H(k) \Gamma(k) \mathcal{H}(k)]^{-1} \\ &\quad \times \mathcal{H}^H(k) \Gamma(k) \tilde{\mathbf{A}}(k) [\tilde{\mathbf{A}}^H(k) \Gamma(k) \tilde{\mathbf{A}}(k)]^{-1}]^{-1}\}^{-1}. \end{aligned} \quad (47.122)$$

This expression depends on the RTFs $\tilde{\mathbf{A}}(k)$, assumed to be error free, and on the coherence function $\Gamma(k)$.

The same $M = 5$ microphone array used for the directional noise case is now used for the diffused noise field. In Fig. 47.8 we show the extra NR obtained by the ANC for various steering angles and for the entire frequency band. It is clear that almost no NR is obtained in the high-frequency band and only relatively low NR in the low-frequency band. The obtained results are in accordance with the results in [47.39,42].

Incoherent Noise Field

For incoherent noise field we assume that the noise at the sensors has no spatial correlation.

$$\Phi_{N_s N_s}(k, \ell) = \Phi_{n_s n_s}(k, \ell) \mathbf{I},$$

where $\mathbf{I}$ is an $M \times M$ identity matrix. Using (47.119) with perfect knowledge of the RTFs, i. e., $\hat{\tilde{\mathbf{A}}}(k) = \tilde{\mathbf{A}}(k)$, and with the prespecified $\Phi_{N_s N_s}(k, \ell)$ we obtain,

$$\begin{aligned} \Phi_{\text{yy}}^n(k, \ell) &= \frac{\Phi_{n_s n_s}(k, \ell)}{\|\tilde{\mathbf{A}}(k)\|^4} \tilde{\mathbf{A}}^H(k) \\ &\quad \times \{\mathbf{I} - \mathcal{H}(k) [\mathcal{H}^H(k) \mathcal{H}(k)]^{-1} \mathcal{H}^H(k)\} \tilde{\mathbf{A}}(k). \end{aligned}$$

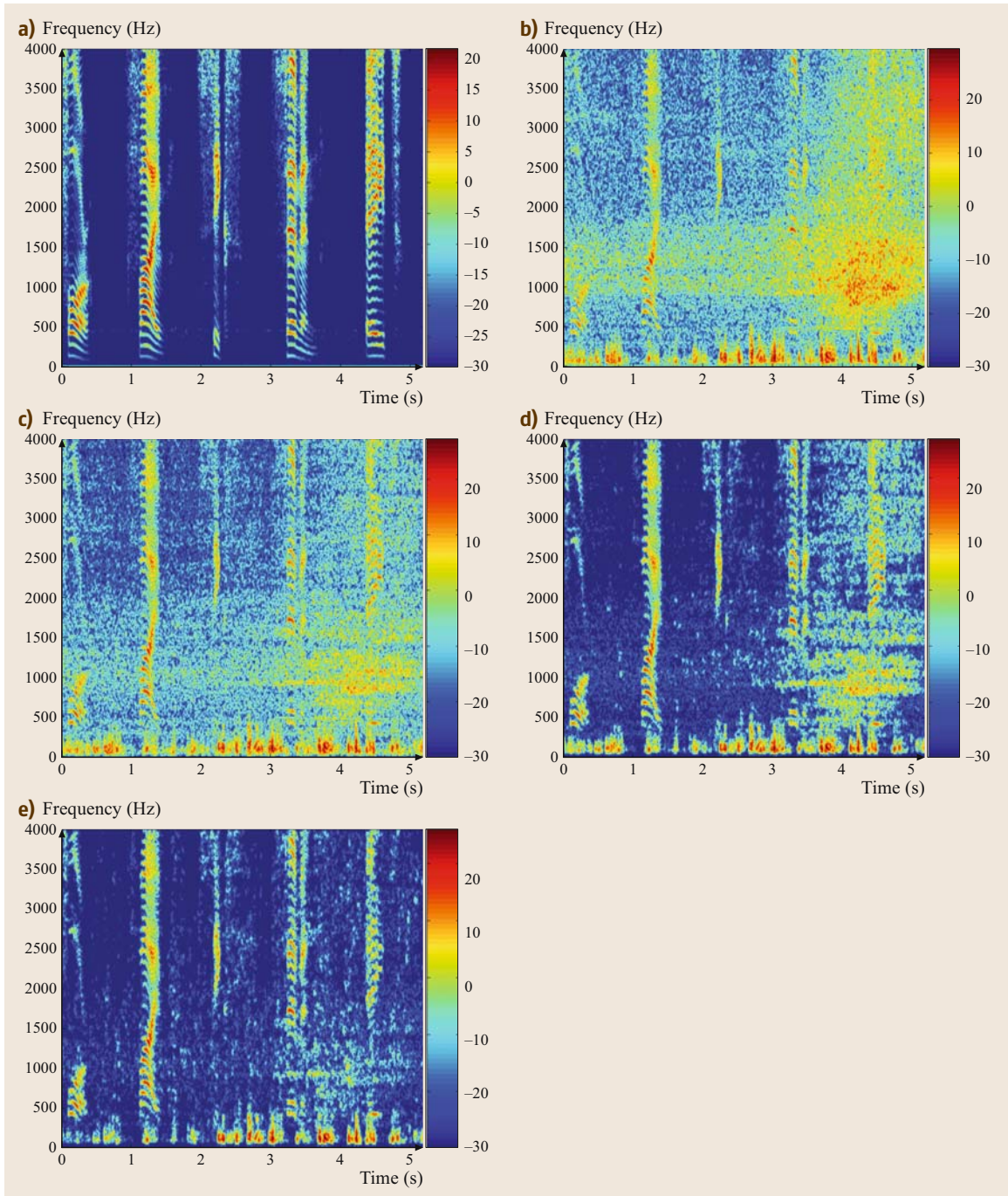


Fig. 47.9 (a) Original clean speech signal at microphone #1: "Five six seven eight nine". (b) Noisy signal at microphone #1. (c) TF-GSC output. (d) Single-channel postfiltering output. (e) Multichannel postfiltering.

It can be easily verified that,

$$\tilde{\mathbf{A}}^H(k)\mathcal{H}(k) = \mathbf{0}_{1 \times (M-1)}.$$

Furthermore, as $\mathcal{H}^H(k)\mathcal{H}(k)$ is a positive matrix, its inverse always exists. Thus, the contribution of the noise cancelling branch is zero, and the NR is only attributed to the FBF. The noise power at the output is thus,

$$\begin{aligned}\Phi_{yy}^n(k, \ell) &= \Phi_{\text{FBF}}^n(k, \ell) = \Phi_{n_s n_s}(k, \ell) \frac{\tilde{\mathbf{A}}^H(k)\tilde{\mathbf{A}}(k)}{\|\tilde{\mathbf{A}}(k)\|^4} \\ &= \frac{\Phi_{n_s n_s}(k, \ell)}{\|\tilde{\mathbf{A}}(k)\|^2}.\end{aligned}$$

Again, no NR is guaranteed by this structure, and the result depends on the RTFs involved.

In the case of delay-only ATFs, the FBF branch becomes a simple delay-and-sum beamformer, thus the expected NR is M , the number of sensors.

47.8 Experimental Results

In this section, we compare the performance of a system, consisting of the TF-GSC and multichannel postfilter, to a system consisting of a TF-GSC and a single-channel postfilter.

A linear array, consisting of four microphones with 5 cm spacing, is mounted in a car on the visor. Clean speech signals are recorded at a sampling rate of 8 kHz in the absence of background noise (standing car, silent environment). A car noise signal is recorded while the car speed is about 60 km/h, and the window next to the driver is slightly open (about 5 cm; the other windows are closed). The input microphone signals are generated by mixing the speech and noise signals at various SNR levels in the range $[-5, 10]$ dB.

Offline TF-GSC beamforming [47.38] is applied to the noisy multichannel signals, and its output is enhanced using the OM-LSA estimator [47.93]. The result is referred to as single-channel postfiltering output.

Alternatively, the proposed TF-GSC and multichannel postfiltering is applied to the noisy signals. A subjective comparison between multichannel and single-channel postfiltering was conducted using speech sonograms and validated by informal listening tests. Typical examples of speech sonograms are presented in Fig. 47.9. For audio samples please refer to [47.99]. The noise PSD at the beamformer output varies substantially due to the residual interfering components of speech, wind blows, and passing cars. The TF-GSC output is characterized by a high level of noise. Single-channel postfiltering suppresses pseudostationary noise components, but is inefficient at attenuating the transient noise components. By contrast, the system which consists a TF-GSC and multichannel postfilter achieves superior noise attenuation, while preserving the desired source components. This is verified by subjective informal listening tests.

47.9 Summary

In this chapter, we concentrated on the GSC beamformer, and presented a comprehensive study of its components. We showed, that the GSC structure is closely related to other array optimization criteria, such as the Wiener filter. We described multimicrophone postfilters, based on either the MMSE or the log-spectral estimation criteria, which are designed for improving the amount of obtainable NR, with minimal degradation

of speech quality. The robustness of the GSC structure to imperfect estimation of its components was analyzed. Various methods were proposed for increasing the robustness of the GSC, and especially, for avoiding leakage of the desired signal into the noise reference signals. Finally, the performance of the TF-GSC was theoretically analyzed and experimentally evaluated under nonstationary noise conditions.

47.A Appendix: Derivation of the Expected Noise Reduction for a Coherent Noise Field

For clarity of the exposition we will omit the time and frequency dependence in the derivation. Recall (47.119) and define

$$\begin{aligned}\mathcal{X}(k, \ell) &\triangleq \mathcal{X} \\ &= \Phi_{N_s N_s} \mathcal{H} (\mathcal{H}^H \Phi_{N_s N_s} \mathcal{H})^{-1} \mathcal{H}^H \Phi_{N_s N_s} .\end{aligned}\quad (47.A1)$$

Denote,

$$\mathcal{X} \triangleq \mathcal{K} \times \mathcal{L} \times \mathcal{M} ,$$

where, $\mathcal{K} = \Phi_{N_s N_s} \mathcal{H}$, $\mathcal{L} = (\mathcal{H}^H \Phi_{N_s N_s} \mathcal{H})^{-1}$, and $\mathcal{M} = \mathcal{H}^H \Phi_{N_s N_s}$. Thus, $\mathcal{X}$ is a multiplication of three terms.

Starting from $\mathcal{L}$ and using the detailed noise structure,

$$\begin{aligned}\mathcal{L} &= [\mathcal{H}^H (\phi_{nn} \mathbf{B} \mathbf{B}^H + \varepsilon \mathbf{I}) \mathcal{H}]^{-1} \\ &= [\phi_{nn} (\mathcal{H}^H \mathbf{B}) (\mathcal{H}^H \mathbf{B})^H + \varepsilon \mathcal{H}^H \mathcal{H}]^{-1} .\end{aligned}$$

If $\mathbf{B} = \mathbf{A}$, i.e., the noise source is located exactly at the desired signal position, then $\mathcal{H}^H \mathbf{B} = 0$, and the calculation of the inverse is straightforward, yielding $\mathcal{L} = (\varepsilon \mathcal{H}^H \mathcal{H})^{-1}$, $\mathcal{K} = \varepsilon \mathcal{H}$, and $\mathcal{M} = \varepsilon \mathcal{H}^H$. Collecting all terms we obtain, $\mathcal{X} = \frac{\varepsilon^2}{\varepsilon} \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \xrightarrow{\varepsilon \rightarrow 0} 0$, i.e., the signal at the BM output is zero, as expected. The total noise part of the output is given by,

$$\begin{aligned}\Phi_{yy}^n &= \Phi_{\text{FBF}}^n \\ &= \frac{\hat{\mathbf{A}}^H \Phi_{N_s N_s} \hat{\mathbf{A}}}{\|\hat{\mathbf{A}}\|^4} = \frac{\hat{\mathbf{A}}^H (\phi_{nn} \mathbf{B} \mathbf{B}^H + \varepsilon \mathbf{I}) \hat{\mathbf{A}}}{\|\hat{\mathbf{A}}\|^4} \\ &\xrightarrow{\varepsilon \rightarrow 0} \phi_{nn} |A_1|^2 ,\end{aligned}$$

where the last transition is due to $\mathbf{B} = \mathbf{A}$. This is exactly the no-distortion result obtained in Sect. 47.7.2, for the desired signal direction.

Table 47.4 Twelve terms used for calculating $\mathcal{X} = \mathcal{K} \mathcal{L} \mathcal{M}$

$\mathcal{K}_1 \mathcal{L}_1 \mathcal{M}_1$	$= \frac{1}{\varepsilon} \Phi_{nn}^2 \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H ;$	(I)
$\mathcal{K}_1 \mathcal{L}_2 \mathcal{M}_1$	$= -\frac{1}{\varepsilon} \Phi_{nn}^3 \times \frac{\mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathbf{B} \mathbf{B}^H}{\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}}$ $= -\frac{1}{\varepsilon} \Phi_{nn}^2 \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H ;$	(II)
$\mathcal{K}_1 \mathcal{L}_3 \mathcal{M}_1$	$= \frac{\Phi_{nn}^3 \mathbf{B} (\mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B})^2 \mathbf{B}^H}{(\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B})^2}$ $= \phi_{nn} \mathbf{B} \mathbf{B}^H ;$	(III)
$\mathcal{K}_2 \mathcal{L}_1 \mathcal{M}_1$	$= \phi_{nn} \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H ;$	(IV)
$\mathcal{K}_2 \mathcal{L}_2 \mathcal{M}_1$	$= -\Phi_{nn}^2 \frac{\mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathbf{B} \mathbf{B}^H}{\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}}$ $= -\phi_{nn} \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H ;$	(V)
$\mathcal{K}_2 \mathcal{L}_3 \mathcal{M}_1$	$= \varepsilon \frac{\Phi_{nn}^2 \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H}{(\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B})^2}$ $= \varepsilon \frac{\mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H}{\mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}} ;$	(VI)
$\mathcal{K}_1 \mathcal{L}_1 \mathcal{M}_2$	$= \phi_{nn} \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H ;$	(VII)
$\mathcal{K}_1 \mathcal{L}_2 \mathcal{M}_2$	$= -\Phi_{nn}^2 \frac{\mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H}{\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}}$ $= -\phi_{nn} \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H ;$	(VIII)
$\mathcal{K}_1 \mathcal{L}_3 \mathcal{M}_2$	$= \varepsilon \frac{\Phi_{nn}^2 \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H}{(\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B})^2}$ $= \varepsilon \frac{\mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H}{\mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}} ;$	(IX)
$\mathcal{K}_2 \mathcal{L}_1 \mathcal{M}_2$	$= \varepsilon \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H ;$	(X)
$\mathcal{K}_2 \mathcal{L}_2 \mathcal{M}_2$	$= -\varepsilon \frac{\mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H}{\mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}} ;$	(XI)
$\mathcal{K}_2 \mathcal{L}_3 \mathcal{M}_2$	$= \varepsilon^2 \frac{\phi_{nn} \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H}{(\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B})^2} .$	(XII)

For the general case, $\mathbf{B} \neq \mathbf{A}$, we use the *matrix inversion lemma*, yielding:

$$\mathcal{L} = \frac{1}{\varepsilon} (\mathcal{H}^H \mathcal{H})^{-1} - \frac{\frac{1}{\varepsilon^2} \phi_{nn} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1}}{1 + \frac{1}{\varepsilon} \phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}}.$$

Now, using the approximation $\frac{1}{1+\mu} \approx 1 - \mu$, for $\mu \rightarrow 0$ (μ properly defined), yields,

$$\begin{aligned} \mathcal{L} &= \frac{1}{\varepsilon} (\mathcal{H}^H \mathcal{H})^{-1} \\ &\quad - \frac{\frac{1}{\varepsilon} \phi_{nn} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1}}{\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B}} \\ &\quad + \frac{\phi_{nn} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1}}{(\phi_{nn} \mathbf{B}^H \mathcal{H} (\mathcal{H}^H \mathcal{H})^{-1} \mathcal{H}^H \mathbf{B})^2}. \end{aligned} \quad (47.A2)$$

Now, calculating $\mathcal{X}$,

$$\begin{aligned} \mathcal{X} &= \mathcal{K} \mathcal{L} \mathcal{M} = \Phi_{NN} \mathcal{H} \mathcal{L} \mathcal{H}^H \Phi_{NN} \\ &= (\phi_{nn} \mathbf{B} \mathbf{B}^H + \varepsilon \mathbf{I}) \mathcal{H} \mathcal{L} \mathcal{H}^H (\phi_{nn} \mathbf{B} \mathbf{B}^H + \varepsilon \mathbf{I}) \\ &\stackrel{\Delta}{=} (\mathcal{K}_1 + \mathcal{K}_2) \mathcal{H} (\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3) \mathcal{H}^H (\mathcal{M}_1 + \mathcal{M}_2), \end{aligned} \quad (47.A3)$$

with the obvious definitions of $\mathcal{K}_1$, $\mathcal{K}_2$, $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, $\mathcal{M}_1$, and $\mathcal{M}_2$. Opening the brackets we have twelve terms given in Table 47.4.

Note, that terms I and II, terms IV and V, and terms VII and VIII eliminate each other, and that terms VI, IX, X, XI, and XII vanish as ε approaches to zero. Only term III is left, i.e., $\mathcal{X} = \phi_{nn} \mathbf{B} \mathbf{B}^H$. Substituting $\mathcal{X}$ into (47.119) we have

$$\Phi_{yy}^n = \frac{\tilde{\mathbf{A}}^H \Phi_{N_s N_s} \tilde{\mathbf{A}}}{\|\tilde{\mathbf{A}}\|^4} - \frac{\tilde{\mathbf{A}}^H (\phi_{nn} \mathbf{B} \mathbf{B}^H) \tilde{\mathbf{A}}}{\|\tilde{\mathbf{A}}\|^4} = 0. \quad (47.A4)$$

47.B Appendix: Equivalence Between Maximum SNR and LCMV Beamformers

The LCMV beamformer is given by

$$\mathbf{W}^{\text{LCMV}} = \frac{(\phi_{ss} \mathbf{A} \mathbf{A}^H + \Phi_{NN})^{-1} \mathbf{A}}{\mathbf{A}^H (\phi_{ss} \mathbf{A} \mathbf{A}^H + \Phi_{NN})^{-1} \mathbf{A}}. \quad (47.B1)$$

Using the matrix inversion lemma we have in the denominator

$$\begin{aligned} &\mathbf{A}^H (\Phi_{NN} + \phi_{ss} \mathbf{A} \mathbf{A}^H)^{-1} \mathbf{A} \\ &= \mathbf{A}^H \left(\Phi_{NN}^{-1} - \frac{\phi_{ss} \Phi_{NN}^{-1} \mathbf{A} \mathbf{A}^H \Phi_{NN}^{-1}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \right) \mathbf{A} \\ &= \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A} - \mathbf{A}^H \frac{\phi_{ss} \Phi_{NN}^{-1} \mathbf{A} \mathbf{A}^H \Phi_{NN}^{-1}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \mathbf{A} \\ &= \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A} - \phi_{ss} \frac{(\mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A})^2}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \\ &= \frac{(1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A})(\mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A})}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \\ &\quad - \frac{\phi_{ss} (\mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A})^2}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \end{aligned}$$

$$= \frac{\mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \quad (47.B2)$$

Similarly, we have in the numerator

$$\begin{aligned} &(\phi_{ss} \mathbf{A} \mathbf{A}^H + \Phi_{NN})^{-1} \mathbf{A} \\ &= \Phi_{NN}^{-1} \mathbf{A} - \frac{\phi_{ss} \Phi_{NN}^{-1} \mathbf{A} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \\ &= \frac{\Phi_{NN}^{-1} \mathbf{A} (1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A})}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \\ &\quad - \frac{\phi_{ss} \Phi_{NN}^{-1} \mathbf{A} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \\ &= \frac{\Phi_{NN}^{-1} \mathbf{A}}{1 + \phi_{ss} \mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}}. \end{aligned} \quad (47.B3)$$

Dividing (47.B3) by (47.B2) we have

$$\mathbf{W}^{\text{LCMV}} = \frac{\Phi_{NN}^{-1} \mathbf{A}}{\mathbf{A}^H \Phi_{NN}^{-1} \mathbf{A}} \equiv \mathbf{W}^{\text{MSNR}}. \quad (47.B4)$$

References

- 47.1 S. Haykin: *Adaptive Filter Theory*, 4th edn. (Prentice Hall, Upper Saddle River 2002)
- 47.2 B. Widrow, S.D. Stearns: *Adaptive Signal Processing* (Prentice-Hall, Upper Saddle River 1985)
- 47.3 D. Monzingo, T. Miller: *Introduction to Adaptive Arrays* (Wiley, New York 1980)
- 47.4 H.L.V. Trees: *Detection, Estimation, and Modulation Theory*, Vol. IV (Wiley, New York 2002)
- 47.5 B.V. Veen, K. Buckley: Beamforming: A versatile approach to spatial filtering, *IEEE Acoust. Speech Signal Process. Mag.* **5**(2), 4–24 (1988)
- 47.6 H. Krim, M. Viberg: Two decades of array signal processing research: The parametric approach, *IEEE Signal Proc. Mag.* **13**, 67–94 (1996)
- 47.7 H. Cox, R. Zeskind, M. Owen: Robust Adaptive Beamforming, *IEEE Trans. Acoust. Speech Signal Process.* **35**(10), 1365–1376 (1987)
- 47.8 S.L. Gay, J. Benesty: *Acoustic Signal Processing for Telecommunication* (Kluwer Academic, Dordrecht 2001)
- 47.9 M.S. Brandstein, D.B. Ward: *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, Berlin, Heidelberg 2001)
- 47.10 J. Benesty, Y. Huang (Eds.): *Adaptive Signal Processing: Applications to Real-World Problems* (Springer, Berlin, Heidelberg 2003)
- 47.11 E. Hänsler, G. Schmidt: *Acoustic Echo and Noise Control: A Practical Approach* (Wiley, New York 2004)
- 47.12 J. Benesty, S. Makino, J. Chen: *Speech Enhancement* (Springer, Berlin, Heidelberg 2005)
- 47.13 W. Herbordt: *Sound Capture for Human/Machine Interfaces – Practical Aspects of Microphone Array Signal Processing* (Springer, Berlin, Heidelberg 2005)
- 47.14 S. Gannot, J. Benesty, J. Bitzer, I. Cohen, S. Doclo, R. Martin, S. Nordholm: Advances in multimicrophone speech processing, *EURASIP J. Appl. Signal Process.* **2006**(ID46357) (2006)
- 47.15 E. Jan, J. Flanagan: Microphone arrays for speech processing, *International Symposium on Signals, Systems, and Electronics (ISSSE) (URSI, San Francisco 1995)* pp. 373–376
- 47.16 E. Jan, J. Flanagan: Sound capture from spatial volumes: matched-filter processing of microphone arrays having randomly-distributed sensors, *Proc. ICASSP (Atlanta, 1996)* pp. 917–920
- 47.17 D. Rabinkin, R. Renomeron, J. Flanagan, D. Macomber: Optimal truncation time for matched filter array processing, *Proc. ICASSP (Seattle, 1998)* pp. 3269–3272
- 47.18 M. Sondhi, G. Elko: Adaptive optimization of microphone arrays under a nonlinear constraint, *Proc. ICASSP, Vol. 11 (Tokyo, 1986)* pp. 981–984
- 47.19 Y. Kaneda, J. Ohga: Adaptive microphone-array system for noise reduction, *IEEE Trans. Acoust. Speech Signal Process.* **34**(6), 1391–1400 (1986)
- 47.20 D.V. Compernelle: Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings, *Proc. ICASSP (Albuquerque, 1990)* pp. 833–836
- 47.21 W. Kellermann: Acoustic echo cancellation for beamforming microphone arrays. In: *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M.S. Brandstein, D.B. Ward (Springer, Berlin, Heidelberg 2001) pp. 281–306
- 47.22 I. Claesson, S. Nordholm: A spatial filtering approach to robust adaptive beamforming, *IEEE Trans. Antennas Propag.* **40**(1), 1093–1096 (1992)
- 47.23 S. Nordholm, I. Claesson, N. Grbić: Optimal and adaptive microphone arrays for speech input in automobiles. In: *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M.S. Brandstein, D.B. Ward (Springer, Berlin, Heidelberg 2001) pp. 307–330
- 47.24 R. Martin: Small microphone arrays with postfilters for noise and acoustic echo reduction. In: *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M.S. Brandstein, D.B. Ward (Springer, Berlin, Heidelberg 2001) pp. 255–280
- 47.25 B. Widrow, J.R. Glover Jr., J.M. McCool, J. Kautitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong Jr, R.C. Goodlin: Adaptive noise cancelling: Principles and applications, *Proc. IEEE* **63**(12), 1692–1716 (1975)
- 47.26 S. Doclo, M. Moonen: GSVD-based optimal filtering for single and multimicrophone speech enhancement, *IEEE Trans. Speech Audio Process.* **50**(9), 2230–2244 (2002)
- 47.27 S. Doclo, M. Moonen: Design of far-field and near-field broadband beamformers using eigenfilters, *Signal Process.* **83**(12), 2641–2673 (2003)
- 47.28 S. Doclo, M. Moonen: Multi-microphone noise reduction using recursive GSVD-based optimal filtering with ANC postprocessing stage, *IEEE Trans. Speech Audio Process.* **13**(1), 53–69 (2005)
- 47.29 A. Spriet, M. Moonen, J. Wouters: A multichannel subband GSVD approach for speech enhancement, *Eur. Trans. Telecommun.* **13**(2), 149–158 (2002), Special Issue on Acoustic Echo and Noise Control
- 47.30 G. Rombouts, M. Moonen: QRD-based unconstrained optimal filtering for acoustic noise reduction, *Signal Process.* **83**(9), 1889–1904 (2003)
- 47.31 G. Rombouts, M. Moonen: Fast QRD-lattice-based unconstrained optimal filtering for acoustic noise reduction, *IEEE Trans. Speech Audio Process.* **13**(6), 1130–1143 (2005)

- 47.32 J. Chen, J. Benesty, Y. Huang, S. Doclo: New insights into the noise reduction Wiener filter, *IEEE Trans. Speech Audio Process.* **14**(4), 1218–1234 (2006)
- 47.33 A. Spriet, M. Moonen, J. Wouters: Spatially pre-processed speech distortion weighted multichannel Wiener filtering for noise reduction, *Signal Process.* **84**(12), 2367–2387 (2004)
- 47.34 S. Doclo, A. Spriet, J. Wouters, M. Moonen: Speech distortion weighted multichannel wiener filtering techniques for noise reduction. In: *Speech Enhancement*, Signals and Communication Technology, ed. by J. Benesty, S. Makino, J. Chen (Springer, Berlin, Heidelberg 2005) pp. 199–228
- 47.35 O.L. Frost: An algorithm for linearly constrained adaptive array processing, *Proc. IEEE* **60**(8), 926–935 (1972)
- 47.36 L.J. Griffiths, C.W. Jim: An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. Antennas Propag.* **30**(1), 27–34 (1982)
- 47.37 S. Affes, Y. Grenier: A source subspace tracking array of microphones for double talk situations, *Proc. ICASSP (Munich, 1997)* pp. 269–272
- 47.38 S. Gannot, D. Burshtein, E. Weinstein: Signal enhancement using beamforming and nonstationarity with applications to speech, *IEEE Trans. Signal Proces.* **49**(8), 1614–1626 (2001)
- 47.39 S. Nordholm, Y.H. Leung: Performance limits of the broadband generalized sidelobe cancelling structure in an isotropic noise field, *J. Acoust. Soc. Am.* **107**(2), 1057–1060 (2000)
- 47.40 J. Bitzer, K.U. Simmer, K.D. Kammeyer: Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement, *Proc. ICASSP (Phoenix, 1999)* pp. 2965–2968
- 47.41 J. Bitzer, K. Simmer, K. Kammeyer: Multichannel noise reduction – algorithms and theoretical limits, *Proc. EUSIPCO, Vol. I (Rhodes, 1998)* pp. 105–108
- 47.42 J. Bitzer, K.-D. Kammeyer, K. Simmer: An alternative implementation of the superdirective beamformer, *Workshop on Application of Signal Processing to Audio and Acoustics (IEEE, New Paltz, NY 1999)*
- 47.43 C. Marro, Y. Mahieux, K. Simmer: Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering, *IEEE Trans. Speech Audio Process.* **6**(3), 240–259 (1998)
- 47.44 S. Nordholm, I. Claesson, M. Dahl: Adaptive microphone array employing calibration signals: an analytical evaluation, *IEEE Trans. Speech Audio Process.* **7**(3), 241–252 (1999)
- 47.45 K.-C. Huarng, C.-C. Yeh: Performance analysis of derivative constraint adaptive arrays with pointing errors, *IEEE Trans. Acoust. Speech Signal Process.* **38**(2), 209–219 (1990)
- 47.46 S. Nordholm, I. Claesson, P. Eriksson: The broadband wiener solution for griffiths–jim beamformers, *IEEE Trans. Signal Process.* **40**(2), 474–478 (1992)
- 47.47 H. Cox: Resolving power and sensitivity to mismatch of optimum array processors, *J. Acoust. Soc. Am.* **54**(3), 771–785 (1973)
- 47.48 D. Brandwood: A complex gradient operator and its application in adaptive array theory, *Proc. IEEE* **130**(1 Parts F and H), 11–16 (1983)
- 47.49 G. Strang: *Linear Algebra and its Application*, 2nd edn. (Academic, New York 1980)
- 47.50 R. Crochiere: A weighted overlap-add method for short-time Fourier analysis/synthesis, *IEEE Trans. Acoust. Speech Signal Process.* **28**(1), 99–102 (1980)
- 47.51 J. Shynk: Frequency-domain and multirate and adaptive filtering, *IEEE Signal Process. Mag.* **9**(1), 14–37 (1992)
- 47.52 J. Allen, D. Berkley: Image method for efficiently simulating small-room acoustics, *J. Acoust. Soc. Am.* **65**(4), 943–950 (1979)
- 47.53 S. Doclo, M. Moonen: Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement, *Int. Workshop Acoust. Echo Noise Control (IWAENC) (Darmstadt 2001)* pp. 31–34
- 47.54 B. Yang: Projection approximation subspace tracking, *IEEE Trans. Signal Process.* **43**(1), 95–107 (1995)
- 47.55 O. Hoshuyama, A. Sugiyama: Robust adaptive beamforming. In: *Microphone Arrays*, ed. by M. Brandstein, D. Ward (Springer, Berlin, Heidelberg 2001) pp. 87–109
- 47.56 E. Weinstein, M. Feder, A. Oppenheim: Multichannel signal separation by decorrelation, *IEEE Trans. Speech Audio Process.* **1**(4), 405–413 (1993)
- 47.57 S. Gannot: *Array Processing of Nonstationary Signals with Application to Speech* (Tel-Aviv University, Tel-Aviv 2000), available on line, <http://www.biu.ac.il/gannot>
- 47.58 C. Knapp, G. Carter: The generalized correlation method for estimation of time delay, *IEEE Trans. Acoust. Speech Signal Process.* **24**(4), 320–327 (1976)
- 47.59 T. Dvorkind, S. Gannot: Time difference of arrival estimation of speech source in a noisy and reverberant environment, *Signal Process.* **85**(1), 177–204 (2005)
- 47.60 J. Chen, J. Benesty, Y. Huang: Time delay estimation in room acoustic environments: an overview, *EURASIP J. Appl. Signal Process.* **26**(503), 19 (2006)
- 47.61 I. Cohen: Relative transfer function identification using speech signals, *IEEE Trans. Speech Audio Process.* **12**(5), 451–459 (2004)
- 47.62 I. Cohen: Identification of speech source coupling between sensors in reverberant noisy environments, *IEEE Signal Process. Lett.* **11**(7), 613–616 (2004)
- 47.63 O. Shalvi, E. Weinstein: System identification using nonstationary signals, *IEEE Trans. Signal Process.* **44**(8), 2055–2063 (1996)
- 47.64 D.G. Manolakis, V.K. Ingle, S.M. Kogan: *Statistical and Adaptive Signal Processing: Spectral Estima-*

- tion, *Signal Modeling, Adaptive Filtering and Array Processing* (McGraw-Hill, Singapore 2000)
- 47.65 R. Martin: Noise power spectral density estimation based on optimal smoothing and minimum statistics, *IEEE Trans. Speech Audio Process.* **9**(5), 504–512 (2001)
- 47.66 I. Cohen, B. Berdugo: Noise estimation by minima controlled recursive averaging for robust speech enhancement, *IEEE Signal Process. Lett.* **9**(1), 12–15 (2002)
- 47.67 I. Cohen: Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging, *IEEE Trans. Speech Audio Process.* **11**(5), 466–475 (2003)
- 47.68 S. Nordholm, H.Q. Dam, N. Grbić, S.Y. Low: Adaptive microphone array employing spatial quadratic soft constraints and spectral shaping. In: *Speech Enhancement, Signals and Communication Technology*, ed. by J. Benesty, S. Makino, J. Chen (Springer, Berlin, Heidelberg 2005) pp. 229–246
- 47.69 O. Hoshuyama, A. Sugiyama, A. Hirano: A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, *IEEE Trans. Signal Process.* **47**(10), 2677–2684 (1999)
- 47.70 C. Fancourt, L. Parra: The generalized sidelobe decorrelator, *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoustics* (New Paltz, NY 2001) pp. 167–170
- 47.71 W. Herbordt, W. Kellermann: Computationally efficient frequency-domain robust generalized sidelobe canceller, *Int. Workshop Acoust. Echo Noise Control (IWAENC)* (Darmstadt 2001) pp. 51–54
- 47.72 W. Herbordt, W. Kellermann: Adaptive beamforming for audio signal acquisition. In: *Adaptive Signal Processing: Applications to Real-World Problems*, Signals and Communication Technology, ed. by J. Benesty, Y. Huang (Springer, Berlin, Heidelberg 2003) pp. 155–194
- 47.73 S. Doclo, M. Moonen: On the output SNR of the speech-distortion weighted multichannel Wiener filter, *IEEE Signal Process. Lett.* **11**(12), 809–811 (2005)
- 47.74 A. Spriet, M. Moonen, J. Wouters: Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel wiener filtering for noise reduction in hearing aids, *IEEE Trans. Signal Process.* **53**(3), 911–925 (2005)
- 47.75 A. Spriet, M. Moonen, J. Wouters: Robustness analysis of multichannel wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications, *IEEE Trans. Speech Audio Process.* **13**(4), 487–503 (2005)
- 47.76 S. Low, S. Nordholm, R. Togneri: Convolutional blind signal separation with post-processing, *IEEE Trans. Speech Audio Process.* **12**(5), 539–548 (2004)
- 47.77 S. Gannot, D. Burshtein, E. Weinstein: Analysis of the power spectral deviation of the general transfer function GSC, *IEEE Trans. Signal Process.* **52**(4), 1115–1121 (2004)
- 47.78 K.U. Simmer, J. Bitzer, C. Marro: Post-filtering techniques. In: *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M.S. Brandstein, D.B. Ward (Springer, Berlin, Heidelberg 2001) pp. 39–60
- 47.79 R. Zelinski: A Microphone array with adaptive post-filtering for noise reduction in reverberant rooms, *IEEE ICASSP* (New York, 1988) pp. 2578–2581
- 47.80 R. Zelinski: Noise reduction based on microphone array with LMS adaptive post-filtering, *Electron. Lett.* **26**(24), 2036–2581 (1990)
- 47.81 S. Fischer, K.U. Simmer: An adaptive microphone array for hands-free communication, *Int. Workshop Acoust. Echo Noise Control (IWAENC)* (Røros, 1995) pp. 44–47
- 47.82 S. Fischer, K.U. Simmer: Beamforming microphone arrays for speech acquisition in noisy environments, *Speech Commun.* **20**(3–4), 215–227 (1996)
- 47.83 S. Fischer, K.-D. Kammeyer: Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environment, *Proc. ICASSP*, Vol. 1 (Munich, 1997) pp. 359–362
- 47.84 J. Meyer, K.U. Simmer: Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction, *Proc. ICASSP* (Munich, 1997) pp. 21–24
- 47.85 K.U. Simmer, S. Fischer, A. Wasiljeff: Suppression of coherent and incoherent noise using a microphone array, *Annales des Télécommun.* **49**(7–8), 439–446 (1994)
- 47.86 J. Bitzer, K. Simmer, K.-D. Kammeyer: Multi-microphone noise reduction by post-filter and superdirective beamformer, *Int. Workshop Acoust. Echo Noise Control (IWAENC)* (Pocono Manor, 1999) pp. 100–103
- 47.87 J. Bitzer, K.U. Simmer, K.-D. Kammeyer: Multi-microphone noise reduction techniques as front-end devices for speech recognition, *Speech Commun.* **34**, 3–12 (2001)
- 47.88 K.U. Simmer, A. Wasiljeff: Adaptive microphone arrays for noise suppression in the frequency domain, *Proc. Second Cost 229 Workshop on Adaptive Algorithms in Communications* (Bordeaux, 1992) pp. 185–194
- 47.89 I. McCowan, H. Bourlard: Microphone array post-filter for diffuse noise field, *Proc. ICASSP*, Vol. 1. (Orlando, 2002) pp. 905–908
- 47.90 I. McCowan, H. Bourlard: Microphone array post-filter based on noise field coherence, *IEEE Trans. Speech Audio Process.* **11**(6), 709–716 (2003)
- 47.91 I. Cohen, S. Gannot, B. Berdugo: An integrated real-time beamforming and postfiltering system for non-stationary noise environments, *EURASIP J. Appl. Signal Process.* **2003**(11), 1064–1073 (2003), special issue on Signal Processing for Acoustic Communication Systems

- 47.92 I. Cohen, B. Berdugo: Microphone array post-filtering for non-stationary noise suppression, Proc. ICASSP (Orlando, 2002) pp. 901–904
- 47.93 I. Cohen, B. Berdugo: Speech enhancement for non-stationary noise environments, Signal Process. **81**(11), 2403–2418 (2001)
- 47.94 Y. Ephraim, D. Malah: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. **ASSP-32**(6), 1109–1121 (1984)
- 47.95 Y. Ephraim, D. Malah: Speech enhancement using a minimum mean square error log-spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)
- 47.96 I. Cohen: Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering, IEEE Trans. Speech Audio Process. **11**(6), 684–699 (2003)
- 47.97 S. Gannot, D. Burshtein, E. Weinstein: Theoretical analysis of the general transfer function GSC, Int. Workshop Acoust. Echo Noise Control (IWAENC) (Darmstadt 2001) pp. 27–30
- 47.98 N. Dal-Degan, C. Prati: Acoustic noise analysis and speech enhancement techniques for mobile radio application, Signal Process. **15**(4), 43–56 (1988)
- 47.99 S. Gannot, I. Cohen: Audio Sample Files <http://www.biu.ac.il/~gannot> (2005)

51. Time Delay Estimation and Source Localization

Y. Huang, J. Benesty, J. Chen

A fundamental requirement of microphone arrays is the capability of instantaneously locating and continuously tracking a speech sound source. The problem is challenging in practice due to the fact that speech is a nonstationary random process with a wideband spectrum, and because of the simultaneous presence of noise, room reverberation, and other interfering speech sources. This Chapter presents an overview of the research and development on this technology in the last three decades. Focusing on a two-stage framework for speech source localization, we survey and analyze the state-of-the-art time delay estimation (TDE) and source localization algorithms.

This chapter is organized into two sections. In Sect. 51.2, we will study the TDE problem and review a number of cutting-edge TDE algorithms, ranging from the generalized cross-correlation methods to blind multichannel-identification-based algorithms and the second-order statistics-based multichannel cross-correlation coefficient method to the higher-order statistics-based entropy-minimization approach. In Sect. 51.3, we will investigate the source localization problem from the perspective of estimation theory. The emphasis is on least-squares estimators with closed-form estimates. The spherical intersection, spherical interpolation, and linear-correction spherical interpolation algorithms will be presented.

51.1	Technology Taxonomy	1043
51.2	Time Delay Estimation	1044
51.2.1	Problem Formulation and Signal Models	1044
51.2.2	The Family of the Generalized Cross-Correlation Methods	1045
51.2.3	Adaptive Eigenvalue Decomposition Algorithm	1047
51.2.4	Adaptive Blind Multichannel Identification Based Methods	1048
51.2.5	Multichannel Spatial Prediction and Interpolation Methods	1049
51.2.6	Multichannel Cross-Correlation Coefficient Algorithm	1050
51.2.7	Minimum-Entropy Method	1052
51.3	Source Localization	1054
51.3.1	Problem Formulation	1054
51.3.2	Measurement Model and Cramér–Rao Lower Bound	1054
51.3.3	Maximum-Likelihood Estimator	1055
51.3.4	Least-Squares Estimators	1056
51.3.5	Least-Squares Error Criteria	1056
51.3.6	Spherical Intersection (SX) Estimator	1057
51.3.7	Spherical Interpolation (SI) Estimator	1057
51.3.8	Linear-Correction Least-Squares Estimator	1058
51.4	Summary	1061
	References	1062

51.1 Technology Taxonomy

Thanks to the explosive growth in computing power and the continuous drop in the cost and size of acoustic electric transducers, a growing number of speech processing and communication systems can afford to employ multiple microphones. There are two major application areas that are driving the interest in microphone array technologies:

- distant speech acquisition that permits good-quality sound pickup due to the capability to mitigate the

effects of room reverberation, ambient noise, and other interfering sound sources; and

- acoustic surveillance and monitoring where sound caused by the events of particular interest is detected, localized, and tracked.

In both applications, there is a primary need to localize a speech source by analyzing the acoustic signals captured by various microphones placed at different positions. A good example of the former is microphone

array beam steering [51.1–4] and a good example of the latter is automatic camera pointing for videoconferencing [51.5–8].

The problem of locating radiative point sources using sensor arrays has long been of great research interest given its theoretical as well as practical importance in a great variety of applications, e.g., radar [51.9, 10], underwater sonar [51.11], and seismology [51.12]. In these applications, source localization is more commonly referred to as direction of arrival (DOA) estimation. A class of celebrated approaches is based on high-resolution spectral analysis, including Capon's minimum variance (MV) spectral estimation method [51.13], and eigenanalysis-based techniques, such as the popular multiple signal classification (MUSIC) algorithm [51.14]. These methods perform statistical fit for DOA with respect to a spatio-spectral correlation matrix derived from the signals received at the sensors. By assumption, the source signal needs to be statistically stationary and narrowband, and the source is located in the far field of the sensor array. However, these basic premises are barely met by speech sources. Moreover, the multipath effect is not taken into account in the formulation of DOA estimation. Therefore, the high-resolution DOA estimation algorithms were found to be incompetent for speech source localization, particularly in reverberant acoustic environments.

For speech source localization, microphone array beam scanning and time delay estimation (TDE)-based localization methods are the two most widely used approaches. A beamformer is a spatial filter that operates on the microphone outputs in order to enhance the signal coming from one direction while suppressing noise and interference from other directions. Therefore, steering a microphone array beamformer and scanning across a room for the highest-energy output leads to an estimate of the direction of an active speech source. Using two arrays and by intersecting the two corresponding direction estimates, the source location

is determined [51.15, 16]. While the beam scanning method has the remarkable advantage that it can easily be extended to the case of simultaneously localizing multiple speech sources [51.17], it has some intrinsic drawbacks. Since speech is a typical broadband signal, ideally a broadband beamformer whose beam pattern would be the same across the entire speech spectrum needs to be developed, which is a technically challenging problem. In addition, most beamforming algorithms assume that the speech sound source is in the far field in order to make the design problem analytically tractable. When the speech source is close to the microphone array, performance degradation can be expected. On the contrary, if the speech source is in the far field, the localization resolution may not be satisfactory as the source direction is only selected from a set of discrete beam scanning angles. The number of beam scanning angles is therefore a trade-off between resolution and computational complexity.

Alternatively, a TDE-based source localization algorithm involves a two-step procedure [51.18]. In the first step, a set of relative time differences of arrival (TDOAs) for different microphone pairs are calculated. In the second step, the acoustic source location is estimated from the TDOAs and using the a priori knowledge about the locations of the microphones. TDE-based source localization algorithms have many merits: they cope well with both narrowband and broadband source signals; their localization resolution can be flexibly adjusted by varying the sampling rate and microphone array size; the effect of room reverberation is only of concern in the first-step TDE and several recently developed robust TDE algorithms are promising for future practical use; and in general they are computationally efficient. Therefore, after continuous investigation over the last two decades, TDE-based speech source localization schemes have become the technique of choice, especially in recent digital systems. In this chapter, an overview of the state of the art of this class of source localization methods is presented.

51.2 Time Delay Estimation

51.2.1 Problem Formulation and Signal Models

Suppose that there is one speech sound source and N microphones. The TDE problem is concerned with the computation of the relative time difference of arrival (TDOA) between different microphone signals.

Depending on the surrounding acoustic environment, there are two signal models for the TDE problem: the ideal free-field model and the real reverberant model. The former assumes no room reverberation while the latter uses an acoustic channel impulse response, usually a finite impulse response (FIR) filter, to describe the effect of room reverberation.

Ideal Free-Field Model

In an anechoic open space, as shown in Fig. 51.1a, the speech source signal $s(k)$ propagates radiatively and the sound level falls off as a function of distance from the source. Then the signal captured by the n -th microphone at time k can be expressed as follows:

$$x_n(k) = \alpha_n s(k - \tau_n) + b_n(k), \quad n = 1, 2, \dots, N, \quad (51.1)$$

where α_n ($0 \leq \alpha_n \leq 1$) is an attenuation factor due to propagation loss, τ_n is the propagation time, and $b_n(k)$ is additive noise. The time difference of arrival (TDOA) between the i -th and j -th microphones is defined as,

$$\tau_{ij} \triangleq \tau_i - \tau_j, \quad i, j = 1, 2, \dots, N. \quad (51.2)$$

The noise signal $b_n(k)$ is presumed to be a zero-mean, white, Gaussian random process, and is uncorrelated with the source signal as well as the noise signals at other microphones.

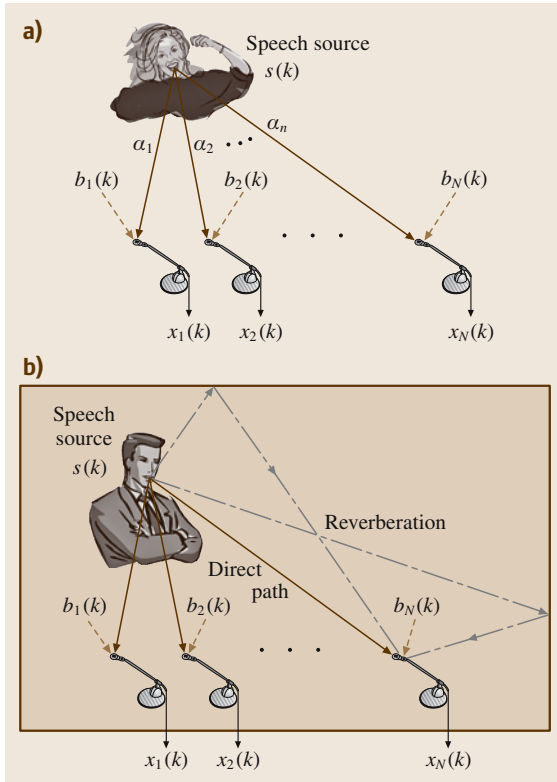


Fig. 51.1a,b Illustration of the two acoustic signal models for time delay estimation: (a) ideal free-field model and (b) real reverberant model

Real Reverberant Model

While the ideal free-field model has the merit of being simple, it does not take into account room reverberation. Therefore, in a real reverberant environment, the ideal free-field model is inadequate and problematic, and we need a more-comprehensive and more-informative alternative to describe the effect of multipath propagation. The real reverberant model treats the acoustic impulse response with an FIR filter, as illustrated by Fig. 51.1b. In such a single-input multiple-output (SIMO) system, the n -th microphone signal is given by

$$x_n(k) = h_n * s(k) + b_n(k), \quad n = 1, 2, \dots, N, \quad (51.3)$$

where h_n is the n -th channel impulse response, and the symbol $*$ denotes the linear convolution operator. In vector/matrix form, (51.3) can be rewritten as

$$\mathbf{x}_n(k) = \mathbf{H}_n \cdot \mathbf{s}(k) + \mathbf{b}_n(k), \quad n = 1, 2, \dots, N, \quad (51.4)$$

where

$$\begin{aligned} \mathbf{x}_n(k) &= [x_n(k) \ \cdots \ x_n(k-L+1)]^T, \\ \mathbf{H}_n &= \begin{pmatrix} h_{n,0} & \cdots & h_{n,L-1} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & h_{n,0} & \cdots & h_{n,L-1} \end{pmatrix}, \\ \mathbf{s}(k) &= [s(k) \quad s(k-1) \quad \cdots \quad s(k-L+1) \\ &\quad \cdots \quad s(k-2L+2)]^T, \\ \mathbf{b}_n(k) &= [b_n(k) \ \cdots \ b_n(k-L+1)]^T, \end{aligned}$$

where $[\cdot]^T$ denotes the transpose of a vector or a matrix, and L is the length of the longest channel impulse response in this SIMO system.

In contrast to the ideal free-field model, the time delay τ_n in the real reverberant model is an implicit or hidden parameter. Using such a model, TDE can be performed only after the SIMO system is *blindly* identified (since the source signal is unknown), which looks like a more-difficult problem but is fortunately not insurmountable.

51.2.2 The Family of the Generalized Cross-Correlation Methods

The generalized cross-correlation (GCC) algorithm proposed by Knapp and Carter [51.19] is so far the most widely used approach to TDE. It employs the ideal free-field model (51.1) and considers only two microphones,

i. e., $N = 2$. Then the TDOA estimate between the two microphones is obtained as the time lag that maximizes the cross-correlation between the filtered signals of the microphone outputs:

$$\hat{\tau}_{12}^{\text{GCC}} = \arg \max_{\tau} r_{x_1 x_2}^{\text{GCC}}(\tau), \quad (51.5)$$

where

$$\begin{aligned} r_{x_1 x_2}^{\text{GCC}}(\tau) &\triangleq \mathcal{F}^{-1} \{ \psi_{x_1 x_2}(f) \} \\ &= \int_{-\infty}^{\infty} \psi_{x_1 x_2}(f) e^{i2\pi f \tau} df \\ &= \int_{-\infty}^{\infty} \Phi(f) S_{x_1 x_2}(f) e^{i2\pi f \tau} df \end{aligned} \quad (51.6)$$

is the GCC function, $\mathcal{F}^{-1}\{\cdot\}$ stands for the inverse discrete-time Fourier transform (IDTFT),

$$S_{x_1 x_2}(f) \triangleq E \{ X_1(f) X_2^*(f) \} \quad (51.7)$$

is the cross spectrum, $\Phi(f)$ is a frequency-domain weighting function,

$$\psi_{x_1 x_2}(f) = \Phi(f) S_{x_1 x_2}(f) \quad (51.8)$$

is the generalized cross spectrum,

$$X_n(f) = \sum_k x_n(k) e^{-i2\pi f k}, \quad n = 1, 2,$$

where $E\{\cdot\}$ denotes mathematical expectation, and $(\cdot)^*$ denotes complex conjugate.

There are many different choices of the frequency-domain weighting function $\Phi(f)$, leading to a variety of different GCC methods for TDE. In the following, several most known and practically most useful algorithms in this family will be discussed.

Classical Cross-Correlation

The classical cross-correlation (CCC) method takes $\Phi(f) = 1$. We know from the ideal free-field model (51.1) that

$$X_n(f) = \alpha_n e^{-i2\pi f \tau_n} S(f) + B_n(f), \quad n = 1, 2. \quad (51.9)$$

Substituting (51.9) into (51.8) and noting that the noise signal at one microphone is uncorrelated with the source signal and the noise signal at the other microphone by assumption, we have

$$\psi_{x_1 x_2}^{\text{CCC}}(f) = \alpha_1 \alpha_2 e^{-i2\pi f \tau_{12}} E \{ |S(f)|^2 \}. \quad (51.10)$$

The fact that $\psi_{x_1 x_2}^{\text{CCC}}(f)$ depends on the source signal is detrimental for TDE since speech is inherently nonstationary.

Smoothed Coherence Transform

In order to overcome the impact of fluctuating levels of the speech source signal on TDE, an effective way is to pre-whiten the microphone outputs before their cross spectrum is computed. This is equivalent to choosing

$$\Phi(f) = \frac{1}{\sqrt{E\{|X_1(f)|^2\} E\{|X_2(f)|^2\}}}, \quad (51.11)$$

which leads to the so-called smoothed coherence transform (SCOT) method [51.20]. Substituting (51.9) and (51.11) into (51.8) produces the SCOT cross-spectrum:

$$\begin{aligned} \psi_{x_1 x_2}^{\text{SCOT}}(f) &= \frac{\alpha_1 \alpha_2 e^{-i2\pi f \tau_{12}} E\{|S(f)|^2\}}{\sqrt{E\{|X_1(f)|^2\} E\{|X_2(f)|^2\}}} \\ &= \frac{\alpha_1 \alpha_2 e^{-i2\pi f \tau_{12}} E\{|S(f)|^2\}}{\sqrt{\alpha_1^2 E\{|S(f)|^2\} + \sigma_{b_1}^2}} \cdot \\ &\quad \times \frac{1}{\sqrt{\alpha_2^2 E\{|S(f)|^2\} + \sigma_{b_2}^2}} \cdot \\ &= \frac{e^{-i2\pi f \tau_{12}}}{\sqrt{1 + \frac{1}{\text{SNR}_1(f)}} \cdot \sqrt{1 + \frac{1}{\text{SNR}_2(f)}}}, \end{aligned} \quad (51.12)$$

where

$$\begin{aligned} \sigma_{b_n}^2(f) &= E\{|B_n(f)|^2\}, \\ \text{SNR}_n(f) &= \frac{\alpha_n E\{|S(f)|^2\}}{E\{|B_n(f)|^2\}}, \quad n = 1, 2. \end{aligned}$$

If the signal-to-noise ratios (SNRs) are the same at the two microphones, then we get

$$\psi_{x_1 x_2}^{\text{SCOT}}(f) = \left(\frac{\text{SNR}(f)}{1 + \text{SNR}(f)} \right) \cdot e^{-i2\pi f \tau_{12}}. \quad (51.13)$$

Therefore, the performance of the SCOT algorithm for TDE would vary with the SNR. But when the SNR is large enough,

$$\psi_{x_1 x_2}^{\text{SCOT}}(f) \approx e^{-i2\pi f \tau_{12}},$$

which implies independent TDE performance of the power of the source signal. The SCOT method is theoretically superior to the CCC method only when the noise level is low.

Phase Transform

It becomes clear by examining (51.6) that the TDOA information is conveyed in the phase rather than the amplitude of the cross-spectrum. Therefore, we can simply discard the amplitude and only keep the phase. By setting

$$\Phi(f) = \frac{1}{|S_{x_1 x_2}(f)|}, \quad (51.14)$$

we get the phase transform (PHAT) method [51.19]. In this case, the generalized cross spectrum is given by

$$\Psi_{x_1 x_2}^{\text{PHAT}}(f) = e^{-i2\pi f \tau_{12}}, \quad (51.15)$$

which depends only on the TDOA τ_{12} . Substituting (51.15) into (51.6), we obtain an ideal GCC function:

$$r_{x_1 x_2}^{\text{PHAT}}(\tau) = \int_{-\infty}^{\infty} e^{i2\pi f(\tau - \tau_{12})} df = \begin{cases} \infty, & \tau = \tau_{12}, \\ 0, & \text{otherwise.} \end{cases} \quad (51.16)$$

As a result, the PHAT method in general performs better than the CCC and SCOT methods for TDE with respect to a speech sound source.

Discussion: The GCC methods are computationally efficient. They induce very short decision delays and hence have good tracking capability: an estimate is produced almost instantaneously. The GCC methods have been well studied and are found to perform fairly well in moderately noisy and nonreverberant environments [51.21, 22]. In order to improve their robustness to additive noise, many amendments have been proposed [51.6, 23–25]. However, these methods still tend to break down when room reverberation is high. This is insightfully explained by the fact that the GCC methods model the surrounding acoustic environment as an ideal free field and thus have a fundamental weakness in their inability to cope with room reverberation. Cepstral prefiltering was suggested to overcome the multipath propagation effect of room reverberation in [51.26]. However, the improvement is limited and shortcomings still remain. Therefore, the main focus of current research in TDE is on combating the effect of room reverberation.

51.2.3 Adaptive Eigenvalue Decomposition Algorithm

The adaptive eigenvalue decomposition (AED) algorithm [51.27] approaches the TDE problem from a different point of view from the traditional GCC methods. While the AED also considers only two microphones, it adopts the real reverberant rather than the

ideal free-field model. It first identifies the two channel impulse responses, and then measures the time difference between the two recognized direct paths as the TDOA estimate. Since the source signal is unknown, the channel identification has to be a blind method.

Following the real reverberant model (51.3) and the fact that, in the absence of additive noise,

$$x_1(k) * h_2 = s(k) * h_1 * h_2 = x_2(k) * h_1, \quad (51.17)$$

we deduce the following cross relation in vector/matrix form at time k :

$$\mathbf{x}^T(k) \mathbf{u} = \mathbf{x}_1^T(k) \mathbf{h}_2 - \mathbf{x}_2^T(k) \mathbf{h}_1 = 0, \quad (51.18)$$

where

$$\begin{aligned} \mathbf{x}(k) &= \begin{bmatrix} \mathbf{x}_1^T(k) & \mathbf{x}_2^T(k) \end{bmatrix}^T, \\ \mathbf{u} &= \begin{bmatrix} \mathbf{h}_2^T & -\mathbf{h}_1^T \end{bmatrix}^T, \\ \mathbf{h}_n &= \begin{bmatrix} h_{n,0} & h_{n,1} & \cdots & h_{n,L-1} \end{bmatrix}^T, \quad n = 1, 2. \end{aligned}$$

Multiplying (51.18) by $\mathbf{x}(k)$ from the left-hand side and taking the expectation yields

$$\mathbf{R}_{xx} \mathbf{u} = \mathbf{0}, \quad (51.19)$$

where $\mathbf{R}_{xx} \triangleq E \{ \mathbf{x}(k) \mathbf{x}^T(k) \}$ is the covariance matrix of the two microphone signals. This indicates that the vector $\mathbf{u}$, which consists of the two impulse responses, is in the null space of $\mathbf{R}$. More specifically, $\mathbf{u}$ is the eigenvector of $\mathbf{R}_x$ corresponding to its eigenvalue 0. If $\mathbf{R}_x$ is rank deficient by 1, $\mathbf{u}$ can be uniquely determined up to a scale, which is equivalent to saying that the two-channel SIMO system can be blindly identified. Using what has been proved in [51.28], we know that such a two-channel acoustic SIMO system is blindly identifiable using only the second-order statistics (SOS) of the microphone outputs if and only if the following two conditions hold:

- The polynomials formed from $\mathbf{h}_1$ and $\mathbf{h}_2$ are coprime, i. e., their channel transfer functions share no common zeros,
- The autocorrelation matrix $\mathbf{R}_{ss} = E \{ s(k) s^T(k) \}$ of the source signal is of full rank (such that the SIMO system can be fully excited).

In practice, noise always exists and the covariance matrix $\mathbf{R}_{xx}$ is positively definite rather than positively semidefinite. As a consequence, $\mathbf{u}$ is found as the normalized eigenvector of $\mathbf{R}_{xx}$ corresponding to its smallest eigenvalue:

$$\hat{\mathbf{u}} = \arg \min_{\mathbf{u}} \mathbf{u}^T \mathbf{R}_{xx} \mathbf{u}, \quad \text{subject to } \|\mathbf{u}\|^2 = 1. \quad (51.20)$$

In the AED algorithm, solving (51.20) is carried out adaptively using a constrained LMS algorithm:

Initialize

$$\hat{\mathbf{h}}_n(0) = \left[\frac{\sqrt{2}}{2} \ 0 \ \dots \ 0 \right]^T, \quad n = 1, 2;$$

$$\hat{\mathbf{u}}(0) = \left[\hat{\mathbf{h}}_2^T(0) \ -\hat{\mathbf{h}}_1^T(0) \right]^T;$$

Compute, $k = 0, 1, \dots$

$$e(k) = \hat{\mathbf{u}}^T(k)\mathbf{x}(k),$$

$$\hat{\mathbf{u}}(k+1) = \frac{\hat{\mathbf{u}}(k) - \mu e(k)\mathbf{x}(k)}{\|\hat{\mathbf{u}}(k) - \mu e(k)\mathbf{x}(k)\|}, \quad (51.21)$$

where the adaptation step size μ is a small positive constant.

After the AED algorithm converges, the time difference between the direct paths of the two identified channel impulse responses $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ is measured as the TDOA estimate:

$$\hat{\tau}_{12}^{\text{AED}} = \arg \max_l |\hat{\mathbf{h}}_{1,l}| - \arg \max_l |\hat{\mathbf{h}}_{2,l}|. \quad (51.22)$$

51.2.4 Adaptive Blind Multichannel Identification Based Methods

The AED algorithm provides us with a new way to look at the TDE problem, which was found to be particularly useful to combat room reverberation. It applies the more-realistic real-reverberant model to a two-microphone acoustic system at a time and attempts to blindly identify the two-channel impulse responses, from which the embedded TDOA information of interest is then extracted. Clearly the blind two-channel identification technique plays a central role in such an approach. The more accurately the two impulse responses are blindly estimated, the more precisely the TDOA can be inferred. However, for a two-channel system, the zeros of the two channels can be close, especially when their impulse responses are long, which leads to an ill-conditioned system that is difficult to identify. If they share some common zeros, the system becomes unidentifiable (using only second-order statistics) and the AED algorithm may completely fail. It was suggested in [51.29] that this problem can be alleviated by employing more microphones. When more microphones are employed, it is less likely that all channels will share a common zero. As such, blind identification deals with a more well-conditioned SIMO system and the solutions can be globally optimized over all channels. The resulting algorithm is referred to as the adaptive

blind multichannel identification (ABMCI)-based TDE algorithm.

The generalization of blind SIMO identification from two channels to multiple (> 2) channels is not straightforward and in [51.30] a systematic way was proposed. Consider a SIMO system with N channels whose outputs are described by (51.3). Each pair of the system outputs has a cross-relation similar to (51.18) in the absence of noise:

$$\mathbf{x}_i^T(k)\mathbf{h}_j = \mathbf{x}_j^T(k)\mathbf{h}_i, \quad i, j = 1, 2, \dots, N. \quad (51.23)$$

When noise is present or the channel impulse responses are improperly modeled, the cross-relation does not hold and the a priori error signal can be defined as

$$e_{ij}(k+1) = \frac{\mathbf{x}_i^T(k+1)\hat{\mathbf{h}}_j(k) - \mathbf{x}_j^T(k+1)\hat{\mathbf{h}}_i(k)}{\|\hat{\mathbf{h}}(k)\|}, \quad i, j = 1, 2, \dots, N, \quad (51.24)$$

where $\hat{\mathbf{h}}_i(k)$ is the model filter for the i -th channel at time k and

$$\hat{\mathbf{h}}(k) = \left[\hat{\mathbf{h}}_1^T(k) \ \hat{\mathbf{h}}_2^T(k) \ \dots \ \hat{\mathbf{h}}_N^T(k) \right]^T.$$

The model filters are normalized to avoid a trivial solution whose elements are all zeros. Based on the error signal defined here, a cost function at time $k+1$ is given by

$$J(k+1) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N e_{ij}^2(k+1). \quad (51.25)$$

The multichannel LMS (MCLMS) algorithm updates the estimate of the channel impulse responses as follows:

$$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) - \mu \nabla J(k+1), \quad (51.26)$$

where μ is again a small positive step size. As shown in [51.30], the gradient of $J(k+1)$ is computed as:

$$\begin{aligned} \nabla J(k+1) &= \frac{\partial J(k+1)}{\partial \hat{\mathbf{h}}(k)} \\ &= \frac{2 \left[\tilde{\mathbf{R}}_{x+}(k+1)\hat{\mathbf{h}}(k) - J(k+1)\hat{\mathbf{h}}(k) \right]}{\|\hat{\mathbf{h}}(k)\|^2}, \end{aligned} \quad (51.27)$$

where

$$\tilde{\mathbf{R}}_{x+}(k) \triangleq \begin{pmatrix} \sum_{n \neq 1} \tilde{\mathbf{R}}_{x_n x_n}(k) & -\tilde{\mathbf{R}}_{x_2 x_1}(k) \\ -\tilde{\mathbf{R}}_{x_1 x_2}(k) & \sum_{n \neq 2} \tilde{\mathbf{R}}_{x_n x_n}(k) \\ \vdots & \vdots \\ -\tilde{\mathbf{R}}_{x_1 x_N}(k) & -\tilde{\mathbf{R}}_{x_2 x_N}(k) \\ \cdots & -\tilde{\mathbf{R}}_{x_N x_1}(k) \\ \cdots & -\tilde{\mathbf{R}}_{x_N x_2}(k) \\ \vdots & \vdots \\ \cdots & \sum_{n \neq N} \tilde{\mathbf{R}}_{x_n x_n}(k) \end{pmatrix}$$

and

$$\tilde{\mathbf{R}}_{x_i x_j}(k) \triangleq \mathbf{x}_i(k) \mathbf{x}_j^T(k), \quad i, j = 1, 2, \dots, N.$$

If the model filters are always normalized after each update, then a simplified MCLMS algorithm is obtained:

$$\begin{aligned} \hat{\mathbf{h}}(k+1) \\ = \frac{\hat{\mathbf{h}}(k) - 2\mu \left[\tilde{\mathbf{R}}_{x+}(k+1) \hat{\mathbf{h}}(k) - J(k+1) \hat{\mathbf{h}}(k) \right]}{\left\| \hat{\mathbf{h}}(k) - 2\mu \left[\tilde{\mathbf{R}}_{x+}(k+1) \hat{\mathbf{h}}(k) - J(k+1) \hat{\mathbf{h}}(k) \right] \right\|}. \end{aligned} \quad (51.28)$$

A number of other adaptive blind **SIMO** identification algorithms with faster convergence and lower computational complexity have also been developed, e.g., [51.31, 32]. Due to space limitations, we choose not to present the development of those algorithms and refer the reader to [51.33] and references therein for more details.

After the adaptive algorithm converges, the **TDOA** between any two microphones can be inferred as

$$\begin{aligned} \hat{\tau}_{ij}^{\text{ABMCI}} &= \arg \max_l |\hat{h}_{i,l}| - \arg \max_l |\hat{h}_{j,l}|, \\ i, j &= 1, 2, \dots, N, \end{aligned} \quad (51.29)$$

which implies that in every channel the direct path is always dominant. This is generally true for acoustic waves, which would be considerably attenuated by wall reflection. But sometimes two or more reverberant signals via multipaths of equal delay could add coherently such that the direct-path component no longer dominates the impulse response. Therefore a more-robust way to pick the direct-path component is to identify the Q ($Q > 1$) strongest elements in the impulse responses and choose

the one with the smallest delay [51.29]:

$$\begin{aligned} \hat{\tau}_{ij}^{\text{ABMCI}} &= \min \left\{ \arg \max_l^q |\hat{h}_{i,l}| \right\} \\ &\quad - \min \left\{ \arg \max_l^q |\hat{h}_{j,l}| \right\}, \\ i, j &= 1, 2, \dots, N, \quad q = 1, 2, \dots, Q, \end{aligned} \quad (51.30)$$

where $\max^q$ computes the q -th largest element.

51.2.5 Multichannel Spatial Prediction and Interpolation Methods

While the blind **SIMO** identification-based TDE algorithms cope well with room reverberation, they have relatively poor tracking capability. As a result, in a rapidly varying acoustic environment, they will have a problem. From this perspective, cross-correlation-based methods do have some advantages and so there have been continuous efforts to improve their robustness against noise and reverberation.

In Sect. 51.2, only two microphones was considered. In this section, we explore the possibility of using multiple microphones (greater than two) to improve the estimate accuracy of the **TDOA** between the first two microphones in adverse acoustic environments. For a three-microphone system, the three **TDOAs**, namely τ_{12} , τ_{13} , and τ_{23} , are apparently not independent but obey the following relation: $\tau_{13} = \tau_{12} + \tau_{23}$. The idea of exploiting **TDOA** redundancy in sensor arrays was first conceived in [51.34] and a Kalman-filtering-based two-stage TDE algorithm was proposed. Recently, following a similar line of thoughts, several fusion algorithms have been developed [51.35–37]. In what follows, we present a multichannel TDE algorithm using spatial prediction and interpolation [51.38, 39], which takes advantage of the **TDOA** redundancy among multiple microphones in a more-intuitive way.

Suppose that there are N ($N > 2$) microphones in an open space to which the ideal free field can be properly applied. The N microphones are positioned in a prespecified geometric pattern such that the **TDOA** of a microphone pair can be expressed as a function of the **TDOAs** of a *small* number of other microphone pairs. The number depends on the distance of the source from the microphone array and on whether the microphones constitute a two-dimensional (2-D) or three-dimensional (3-D) array. For ease of presentation, let us consider a linear array and a far-field speech source. Then we

have

$$\tau_{1n} = \tau_1 - \tau_n = f_n(\tau_{12}), \quad n = 1, 2, \dots, N, \quad (51.31)$$

where

$$f_1(\tau_{12}) = 0, \quad f_2(\tau_{12}) = \tau_{12}.$$

Substituting (51.31) into (51.1) gives

$$x_n(k) = \alpha_n s[k - \tau_1 + f_n(\tau_{12})] + b_n(k). \quad (51.32)$$

In the absence of noise, since it can be easily checked that

$$x_n[k + f_N(\tau_{12}) - f_n(\tau_{12})] = \alpha_n s(k - \tau_N), \quad \forall n = 1, 2, \dots, N, \quad (51.33)$$

we know that $x_1[k + f_N(\tau_{12})]$ is aligned with $x_n[k + f_N(\tau_{12}) - f_n(\tau_{12})]$. Therefore, a forward spatial prediction error signal can be defined as:

$$e_1(k, \tau) \triangleq x_1[k + f_N(\tau)] - \mathbf{x}_{a,2:N}^T(k, \tau) \mathbf{a}_{2:N}(\tau), \quad (51.34)$$

where τ is a dummy variable for the hypothesized TDOA τ_{12} ,

$$\mathbf{x}_{a,2:N}(k, \tau) = \begin{bmatrix} x_2[k + f_N(\tau) - f_2(\tau)] & \cdots & x_N(k) \end{bmatrix}^T \quad (51.35)$$

is the aligned (subscript 'a') signal vector, and

$$\mathbf{a}_{2:N}(\tau) = \begin{bmatrix} a_2(\tau) & a_3(\tau) & \cdots & a_N(\tau) \end{bmatrix}^T$$

contains the forward spatial linear prediction coefficients. Minimizing the mean-square value of the prediction error signal

$$J_1(\tau) \triangleq E \left\{ e_1^2(k, \tau) \right\} \quad (51.36)$$

leads to the Wiener–Hopf equation

$$\mathbf{R}_{a,2:N}(\tau) \mathbf{a}_{2:N}(\tau) = \mathbf{r}_{a,2:N}(\tau), \quad (51.37)$$

where

$$\begin{aligned} \mathbf{R}_{a,2:N}(\tau) &\triangleq E \left\{ \mathbf{x}_{a,2:N}(k, \tau) \mathbf{x}_{a,2:N}^T(k, \tau) \right\} \\ &= \begin{pmatrix} \sigma_{x_2}^2 & r_{a,x_2x_3}(\tau) & \cdots & r_{a,x_2x_N}(\tau) \\ r_{a,x_3x_2}(\tau) & \sigma_{x_3}^2 & \cdots & r_{a,x_3x_N}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ r_{a,x_Nx_2}(\tau) & r_{a,x_Nx_3}(\tau) & \cdots & \sigma_{x_N}^2 \end{pmatrix} \end{aligned}$$

is the spatial correlation matrix of the aligned signals with

$$\begin{aligned} \sigma_{x_n}^2 &= E \left\{ x_n^2(k) \right\}, \quad n = 1, 2, \dots, N, \\ r_{a,x_i x_j}(\tau) &= E \left\{ x_i[k + f_N(\tau) - f_i(\tau)] \right. \\ &\quad \left. \times x_j[k + f_N(\tau) - f_j(\tau)] \right\} \\ &= E \left\{ x_i[k - f_i(\tau)] x_j[k - f_j(\tau)] \right\}, \\ i, j &= 1, 2, \dots, N, \end{aligned}$$

and

$$\mathbf{r}_{a,2:N}(\tau) \triangleq \begin{bmatrix} r_{a,x_1x_2}(\tau) & r_{a,x_1x_3}(\tau) & \cdots & r_{a,x_1x_N}(\tau) \end{bmatrix}^T.$$

Substituting the solution of the Wiener–Hopf (51.37), which is

$$\mathbf{a}_{2:N}(\tau) = \mathbf{R}_{a,2:N}^{-1}(\tau) \mathbf{r}_{a,2:N}(\tau),$$

into (51.34) gives the minimum forward prediction error

$$\begin{aligned} e_{1,\min}(k, \tau) &= x_1[k + f_N(\tau)] - \mathbf{x}_{a,2:N}^T(k, \tau) \mathbf{R}_{a,2:N}^{-1}(\tau) \mathbf{r}_{a,2:N}(\tau). \end{aligned} \quad (51.38)$$

Accordingly, we have

$$\begin{aligned} J_{1,\min}(\tau) &\triangleq E \left\{ e_{1,\min}^2(k, \tau) \right\} \\ &= \sigma_{x_1}^2 - \mathbf{r}_{a,2:N}^T(\tau) \mathbf{R}_{a,2:N}^{-1}(\tau) \mathbf{r}_{a,2:N}(\tau). \end{aligned} \quad (51.39)$$

Then we can argue that the time lag τ that induces a minimum $J_{1,\min}(\tau)$ would be the TDOA between the first two microphones:

$$\hat{\tau}_{12}^{\text{FSP}} = \arg \min_{\tau} J_{1,\min}(\tau), \quad (51.40)$$

where the superscript 'FSP' stands for forward spatial predication.

Similarly, the multichannel TDE algorithm can be developed using backward prediction or interpolation with any one of the N microphone outputs being regarded as the reference signal.

51.2.6 Multichannel Cross-Correlation Coefficient Algorithm

The idea of using the spatial prediction (or interpolation) error to measure the correlation among multiple signals is intuitive and helpful. However in statistics and signal processing, the cross-correlation coefficient is a more widely used concept for this purpose. Since the cross-correlation coefficient is defined traditionally only for

two random processes, it first needs to be generalized to multiple random processes before being employed in the development of a multichannel TDE algorithm.

The definition of a multiple coherence function, derived from the concepts of the ordinary coherence function between two signals and the partial (conditioned) coherence function, was presented in [51.40] to measure the correlation among the output of a multiple-input signal-output (MISO) system and its inputs. In [51.38] and [51.39], the multichannel cross-correlation coefficient (MCCC) was constructed in a new way from the multichannel correlation matrix; this provides a seamless generalization of the classical cross-correlation coefficient to the case where there are multiple random processes.

Following (51.35), we denote

$$\mathbf{x}_a(k, \tau) \triangleq \begin{bmatrix} x_1[k + f_N(\tau)] & \mathbf{x}_{a,2:N}^T(k, \tau) \end{bmatrix}^T, \quad (51.41)$$

and

$$\begin{aligned} \mathbf{R}_a(\tau) &\triangleq E \left\{ \mathbf{x}_a(k, \tau) \mathbf{x}_a^T(k, \tau) \right\} \\ &= \begin{pmatrix} \sigma_{x_1}^2 & r_{a,x_1x_2}(\tau) & \cdots & r_{a,x_1x_N}(\tau) \\ r_{a,x_2x_1}(\tau) & \sigma_{x_2}^2 & \cdots & r_{a,x_2x_N}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ r_{a,x_Nx_1}(\tau) & r_{a,x_Nx_2}(\tau) & \cdots & \sigma_{x_N}^2 \end{pmatrix}. \end{aligned} \quad (51.42)$$

The spatial correlation matrix $\mathbf{R}_a(\tau)$ can be factored as follows:

$$\mathbf{R}_a(\tau) = \mathbf{\Sigma} \tilde{\mathbf{R}}_a(\tau) \mathbf{\Sigma}, \quad (51.43)$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{x_1} & 0 & \cdots & 0 \\ 0 & \sigma_{x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_{x_N} \end{pmatrix}$$

is a diagonal matrix,

$$\tilde{\mathbf{R}}_a(\tau) = \begin{pmatrix} 1 & \rho_{a,x_1x_2}(\tau) & \cdots & \rho_{a,x_1x_N}(\tau) \\ \rho_{a,x_2x_1}(\tau) & 1 & \cdots & \rho_{a,x_2x_N}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{a,x_Nx_1}(\tau) & \rho_{a,x_Nx_2}(\tau) & \cdots & 1 \end{pmatrix}$$

is a symmetric matrix, and

$$\rho_{a,x_i x_j}(\tau) \triangleq \frac{r_{a,x_i x_j}(\tau)}{\sigma_i \sigma_j}, \quad i, j = 1, 2, \dots, N,$$

is the cross-correlation coefficient between the i -th and j -th aligned microphone signals.

Since matrix $\tilde{\mathbf{R}}_a(\tau)$ is symmetric, positively semidefinite, and its diagonal elements are all equal to one, it was shown in [51.38] and [51.39] that

$$0 \leq \det[\tilde{\mathbf{R}}_a(\tau)] \leq 1, \quad (51.44)$$

where $\det(\cdot)$ stands for determinant.

In the two-channel case, it can easily be checked that the squared cross-correlation coefficient is linked to the normalized spatial correlation matrix by

$$\rho_{a,x_1x_2}^2(\tau) = 1 - \det[\tilde{\mathbf{R}}_{a,1:2}(\tau)]. \quad (51.45)$$

Then by analogy the squared MCCC among the N aligned signals $x_n[k + f_N(\tau) - f_n(\tau)]$, $n = 1, 2, \dots, N$, is constructed as:

$$\begin{aligned} \rho_{a,x_1:x_N}^2(\tau) &\triangleq 1 - \det[\tilde{\mathbf{R}}_a(\tau)] \\ &= 1 - \frac{\det[\mathbf{R}_a(\tau)]}{\prod_{n=1}^N \sigma_{x_n}^2}. \end{aligned} \quad (51.46)$$

The MCCC has the following properties (presented without proof) [51.41].

1. $0 \leq \rho_{a,x_1:x_N}^2(\tau) \leq 1$.
2. If two or more signals are perfectly correlated, then $\rho_{a,x_1:x_N}^2(\tau) = 1$.
3. If all the signals are completely uncorrelated with each other, then $\rho_{a,x_1:x_N}^2(\tau) = 0$.
4. If one of the signals is completely uncorrelated with the $N - 1$ other signals, then the MCCC will measure the correlation among those $N - 1$ remaining signals.

Using the definition of MCCC, we deduce an estimate of the TDOA between the first two microphone signals as

$$\hat{\tau}_{12}^{\text{MCCC}} = \arg \max_{\tau} \rho_{a,x_1:x_N}^2(\tau), \quad (51.47)$$

which is equivalent to computing

$$\begin{aligned} \hat{\tau}_{12}^{\text{MCCC}} &= \arg \max_{\tau} \left\{ 1 - \frac{\det[\mathbf{R}_a(\tau)]}{\prod_{n=1}^N \sigma_{x_n}^2} \right\} \\ &= \arg \min_{\tau} \det[\mathbf{R}_a(\tau)]. \end{aligned} \quad (51.48)$$

To investigate the link of the MCCC method to the FSP method, let us revisit the spatial prediction error function given by (51.39). We define

$$\begin{aligned} \mathbf{a}(\tau) &\triangleq \begin{bmatrix} a_1(\tau) & a_2(\tau) & \cdots & a_N(\tau) \end{bmatrix}^T \\ &= \begin{bmatrix} a_1(\tau) & \mathbf{a}_{2:N}^T(\tau) \end{bmatrix}^T. \end{aligned} \quad (51.49)$$

Then, for $a_1(\tau) = -1$, the forward spatial prediction error signal (51.34) can be written as

$$e_1(k, \tau) = -\mathbf{x}_a^T(k, \tau)\mathbf{a}(\tau), \quad (51.50)$$

and (51.36) can be expressed as

$$\begin{aligned} J_1(\tau) &= E \left\{ e_1^2(k, \tau) \right\} + \kappa \left[\mathbf{v}_1^T \mathbf{a}(\tau) + 1 \right] \\ &= \mathbf{a}^T(\tau) \mathbf{R}_a(\tau) \mathbf{a}(\tau) + \kappa \left[\mathbf{v}_1^T \mathbf{a}(\tau) + 1 \right], \end{aligned} \quad (51.51)$$

where κ is a Lagrange multiplier introduced to force $a_1(\tau)$ to have value -1 and

$$\mathbf{v}_1 \triangleq \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T.$$

Taking the derivative of (51.51) with respect to $\mathbf{a}(\tau)$ and setting the result to zero yields

$$\frac{\partial J_1(\tau)}{\partial \mathbf{a}(\tau)} = 2\mathbf{R}_a(\tau)\mathbf{a}(\tau) + \kappa\mathbf{v}_1 = \mathbf{0}. \quad (51.52)$$

Solving (51.52) for $\mathbf{a}(\tau)$ produces

$$\mathbf{a}(\tau) = -\frac{\kappa \mathbf{R}_a^{-1}(\tau) \mathbf{v}_1}{2}. \quad (51.53)$$

Substituting (51.53) into (51.51) leads to

$$J_1(\tau) = \kappa \left(1 - \frac{\mathbf{v}_1^T \mathbf{R}_a^{-1}(\tau) \mathbf{v}_1}{4} \kappa \right), \quad (51.54)$$

from which we know that

$$J_{1,\min}(\tau) = \frac{1}{\mathbf{v}_1^T \mathbf{R}_a^{-1}(\tau) \mathbf{v}_1}. \quad (51.55)$$

Substituting (51.43) into (51.55) and using the fact that

$$\Sigma^{-1} \mathbf{v}_1 = \frac{\mathbf{v}_1}{\sigma_{x_1}},$$

we have

$$J_{1,\min}(\tau) = \frac{\sigma_{x_1}^2}{\mathbf{v}_1^T \tilde{\mathbf{R}}_a^{-1}(\tau) \mathbf{v}_1}. \quad (51.56)$$

Note that $\mathbf{v}_1^T \tilde{\mathbf{R}}_a^{-1}(\tau) \mathbf{v}_1$ is the (1, 1)-th element of matrix $\tilde{\mathbf{R}}_a^{-1}(\tau)$, which is computed using the adjoint method as the (1, 1)-th cofactor of $\tilde{\mathbf{R}}_a(\tau)$ divided by the determinant of $\tilde{\mathbf{R}}_a(\tau)$, i. e.,

$$\mathbf{v}_1^T \tilde{\mathbf{R}}_a^{-1}(\tau) \mathbf{v}_1 = \frac{\det[\tilde{\mathbf{R}}_{a,2:N}(\tau)]}{\det[\tilde{\mathbf{R}}_a(\tau)]}, \quad (51.57)$$

where $\tilde{\mathbf{R}}_{a,2:N}(\tau)$ is the lower-right submatrix of $\tilde{\mathbf{R}}_a(\tau)$ formed by removing the first row and the first column. By substituting (51.57) into (51.56), we get

$$J_{1,\min}(\tau) = \sigma_{x_1}^2 \cdot \frac{\det[\tilde{\mathbf{R}}_a(\tau)]}{\det[\tilde{\mathbf{R}}_{a,2:N}(\tau)]}. \quad (51.58)$$

Therefore, the FSP estimate of τ_{12} is found as

$$\begin{aligned} \hat{\tau}_{12}^{\text{FSP}} &= \arg \min_{\tau} J_{1,\min}(\tau) \\ &= \arg \min_{\tau} \frac{\det[\tilde{\mathbf{R}}_a(\tau)]}{\det[\tilde{\mathbf{R}}_{a,2:N}(\tau)]}. \end{aligned} \quad (51.59)$$

Comparing (51.48) to (51.59) reveals a clear distinction between the two methods in spite of the strong similarity in principle. In practice, the FSP method has a numerical stability problem since the calculation of the FSP cost function (51.58) involves division by $\det[\tilde{\mathbf{R}}_{a,2:N}(\tau)]$, while the MCCC method is found to be fairly stable.

It is worth pointing out that the microphone outputs can be pre-whitened before computing their MCCC as was done in the SCOT algorithm in the two-channel scenario. Using this approach, the TDE algorithms become more robust to volume variations of the speech source signal.

51.2.7 Minimum-Entropy Method

While the MCCC-based TDE algorithm performs well in the presence of noise and reverberation, the MCCC is by no means the only choice for developing the concept of multichannel TDE. MCCC is a second-order-statistics (SOS) measure of dependence among multiple random variables and is ideal for Gaussian source signals. However for non-Gaussian source signals such as speech, MCCC is not sufficient and higher-order-statistics (HOS) have more to say about their dependence.

The concept of entropy, which is a statistical (apparently HOS) measure of randomness or uncertainty of a random variable, was introduced by *Shannon* in the context of communication theory [51.42]. For a random variable x with a probability density function (PDF) $p(x)$ (in this chapter, we choose not to distinguish between random variables and their realizations for the conciseness of presentation), the entropy is defined as [51.43]:

$$\begin{aligned} H(x) &= - \int p(x) \ln p(x) dx \\ &= -E \{ \ln p(x) \}. \end{aligned} \quad (51.60)$$

The entropy (in the continuous case) is a measure of the structure contained in the PDF [51.44]. As far as the

multivariate random variable $\mathbf{x}_a(k, \tau)$ given by (51.41) is concerned, the joint entropy is:

$$H[\mathbf{x}_a(k, \tau)] = - \int p[\mathbf{x}_a(k, \tau)] \ln p[\mathbf{x}_a(k, \tau)] d\mathbf{x}_a(k, \tau). \quad (51.61)$$

It was then argued in [51.45] that the time lag τ that gives the minimum of $H[\mathbf{x}_a(k, \tau)]$ corresponds to the TDOA between the two microphones:

$$\hat{\tau}_{12}^{\text{ME}} = \arg \min_{\tau} H[\mathbf{x}_a(k, \tau)], \quad (51.62)$$

where the superscript ‘ME’ refers to the minimum-entropy method.

Gaussian Source Signal

If the source is Gaussian, so are the microphone outputs in the absence of noise. Suppose that the aligned microphone signals are zero-mean and joint Gaussian random signals. Their joint PDF is then given by

$$p[\mathbf{x}_a(k, \tau)] = \frac{\exp[-\eta_a(k, \tau)/2]}{\sqrt{(2\pi)^N \det[\mathbf{R}_a(\tau)]}}, \quad (51.63)$$

where

$$\eta_a(k, \tau) \triangleq \mathbf{x}_a^T(k, \tau) \mathbf{R}_a^{-1}(\tau) \mathbf{x}_a(k, \tau). \quad (51.64)$$

By substituting (51.63) into (51.61), the joint entropy can be computed [51.45] as:

$$H[\mathbf{x}_a(k, \tau)] = \frac{1}{2} \ln \left\{ (2\pi e)^N \det[\mathbf{R}_a(\tau)] \right\}. \quad (51.65)$$

Consequently, (51.62) becomes

$$\hat{\tau}_{12}^{\text{ME}} = \arg \min_{\tau} \det[\mathbf{R}_a(\tau)]. \quad (51.66)$$

It is clear from (51.48) and (51.66) that minimizing the entropy is equivalent to maximizing the MCCC for Gaussian source signals.

Speech Source Signal

Speech is a complicated random process and there is no rigorous mathematical formula for its entropy. However, in speech research it was found that speech can be fairly well modeled by a Laplace distribution [51.46, 47].

The univariate Laplace distribution with zero mean and variance σ_x^2 is given by

$$p(x) = \frac{\sqrt{2}}{2\sigma_x} e^{-\sqrt{2}|x|/\sigma_x}, \quad (51.67)$$

and the corresponding entropy is [51.43],

$$H(x) = 1 + \ln(\sqrt{2}\sigma_x). \quad (51.68)$$

Suppose that $\mathbf{x}_a(k, \tau)$ has a multivariate Laplace distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_a(\tau)$ [51.48], [51.49]:

$$p[\mathbf{x}_a(k, \tau)] = \frac{2[\eta_a(k, \tau)/2]^{P/2} K_P[\sqrt{2\eta_a(k, \tau)}]}{\sqrt{(2\pi)^N \det[\mathbf{R}_a(\tau)]}}, \quad (51.69)$$

where $P = (2 - N)/2$ and $K_P(\cdot)$ is the modified Bessel function of the third kind (also called the modified Bessel function of the second kind) given by

$$K_P(a) = \frac{1}{2} \left(\frac{a}{2}\right)^P \int_0^\infty z^{-P-1} \exp\left(-z - \frac{a^2}{4z}\right) dz, \quad a > 0. \quad (51.70)$$

The joint entropy is

$$H[\mathbf{x}_a(k, \tau)] = \frac{1}{2} \ln \left(\frac{(2\pi)^N \det[\mathbf{R}_a(\tau)]}{4} \right) - \frac{P}{2} E \left\{ \ln \left(\frac{\eta_a(k, \tau)}{2} \right) \right\} - E \left\{ \ln K_P \left[\sqrt{2\eta_a(k, \tau)} \right] \right\}. \quad (51.71)$$

The two quantities $E\{\ln[\eta_a(k, \tau)/2]\}$ and $E\{\ln K_P[\sqrt{2\eta_a(k, \tau)}]\}$ do not seem to have a closed form. So a numerical scheme needs to be developed to estimate them. One possibility to do this is the following. Assume that all processes are ergodic. As a result, ensemble averages can be replaced by time averages. If there are K samples for each element of the observation vector $\mathbf{x}_a(k, \tau)$, the following estimators were proposed in [51.45]:

$$E\{\ln[\eta_a(k, \tau)/2]\} \approx \frac{1}{K} \sum_{k=0}^{K-1} \ln[\eta_a(k, \tau)/2], \quad (51.72)$$

$$E\left\{\ln K_P\left[\sqrt{2\eta_a(k, \tau)}\right]\right\} \approx \frac{1}{K} \sum_{k=0}^{K-1} \ln K_P\left[\sqrt{2\eta_a(k, \tau)}\right]. \quad (51.73)$$

The simulation results presented in [51.45] show that the ME algorithm performs in general comparably to or better than the MCCC algorithm. Apparently the ME algorithm is computationally intensive, but the idea of using entropy expands our horizon of knowledge in pursuit of new TDE algorithms.

51.3 Source Localization

51.3.1 Problem Formulation

The problem addressed here is the estimation of the location of an acoustic point source given the array geometry and the relative TDOA measurements between different microphone pairs. The problem is stated mathematically as follows.

Consider an array that consists of N microphones. As shown in Fig. 51.2, these microphones are located in a 3-D Cartesian coordinate system at positions

$$\mathbf{y}_n \triangleq \begin{bmatrix} x_n & y_n & z_n \end{bmatrix}^T, \quad n = 1, 2, \dots, N. \quad (51.74)$$

Without loss of generality, the first microphone ($n = 1$) is regarded as the reference and is placed at the origin of the coordinate system, i. e., $\mathbf{y}_1 = [0 \ 0 \ 0]^T$. The acoustic source is located at $\mathbf{y}_s \triangleq [x_s \ y_s \ z_s]^T$. The distances from the origin to the n -th microphone and the source are denoted by Ω_n and Ω_s , respectively, where

$$\Omega_n \triangleq \|\mathbf{y}_n\| = \sqrt{x_n^2 + y_n^2 + z_n^2}, \quad n = 1, 2, \dots, N, \quad (51.75)$$

$$\Omega_s \triangleq \|\mathbf{y}_s\| = \sqrt{x_s^2 + y_s^2 + z_s^2}. \quad (51.76)$$

The distance between the source and the n -th microphone is denoted by

$$D_n \triangleq \|\mathbf{y}_n - \mathbf{y}_s\| = \sqrt{(x_n - x_s)^2 + (y_n - y_s)^2 + (z_n - z_s)^2}. \quad (51.77)$$

The difference in the distances of microphones i and j from the source is given by

$$d_{ij} \triangleq D_i - D_j, \quad i, j = 1, 2, \dots, N. \quad (51.78)$$

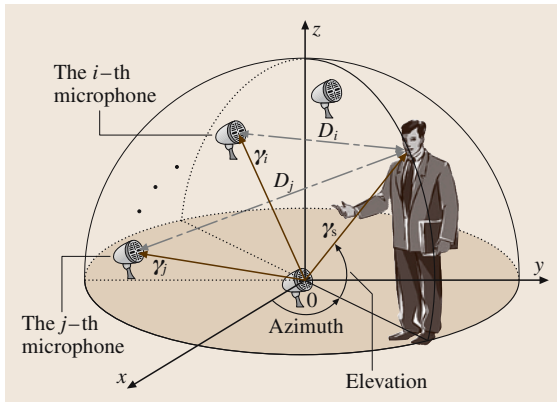


Fig. 51.2 Spatial diagram illustrating variables defined in the source localization problem

This difference is usually termed the range difference; it is proportional to the time difference of arrival τ_{ij} . If the speed of sound is c , then

$$d_{ij} = c \cdot \tau_{ij}. \quad (51.79)$$

The speed of sound (in m/s) can be estimated from the air temperature t_{air} (in degrees Celsius) according to the following approximate (first-order) formula,

$$c \approx 331 + 0.610 t_{\text{air}}. \quad (51.80)$$

The localization problem is then to estimate $\mathbf{y}_s$ given the set of $\mathbf{y}_n$ and τ_{ij} . Note that there are $N(N-1)/2$ distinct TDOA estimates τ_{ij} , which exclude the case $i = j$ and counts the $\tau_{ij} = -\tau_{ji}$ pair only once. However, in the absence of noise, the space spanned by these TDOA estimates is $(N-1)$ -dimensional. Any $N-1$ linearly independent TDOAs determine all of the others. In a noisy environment, the TDOA redundancy can be used to improve the accuracy of the source localization algorithms, but this would increase their computational complexity. For simplicity and also without loss of generality, we choose $\tau_{n1}, n = 2, \dots, N$ as the basis for this $\mathbb{R}^{N-1}$ space in this Chapter.

51.3.2 Measurement Model and Cramér–Rao Lower Bound

When the source localization problem is examined using estimation theory, the measurements of the range differences are modeled by

$$d_{n1} = g_n(\mathbf{y}_s) + \epsilon_n, \quad n = 2, \dots, N, \quad (51.81)$$

where

$$g_n(\mathbf{y}_s) = \|\mathbf{y}_n - \mathbf{y}_s\| - \|\mathbf{y}_s\|,$$

and the ϵ_n are measurement errors. In vector form, such an additive measurement error model becomes,

$$\mathbf{d} = \mathbf{g}(\mathbf{y}_s) + \boldsymbol{\epsilon}, \quad (51.82)$$

where

$$\begin{aligned} \mathbf{d} &= \begin{bmatrix} d_{21} & d_{31} & \dots & d_{N1} \end{bmatrix}^T, \\ \mathbf{g}(\mathbf{y}_s) &= \begin{bmatrix} g_2(\mathbf{y}_s) & g_3(\mathbf{y}_s) & \dots & g_N(\mathbf{y}_s) \end{bmatrix}^T, \\ \boldsymbol{\epsilon} &= \begin{bmatrix} \epsilon_2 & \epsilon_3 & \dots & \epsilon_N \end{bmatrix}^T. \end{aligned}$$

Furthermore, it was postulated that the additive measurement errors have zero mean and are independent of

the range difference observation, as well as the source location $\mathbf{y}_s$ [51.50]. For a continuous-time estimator, the corrupting noise, as indicated in [51.16], is jointly Gaussian distributed. The PDF of $\mathbf{d}$ conditioned on $\mathbf{y}_s$ is subsequently given by,

$$p(\mathbf{d}|\mathbf{y}_s) = \frac{\exp\left\{-\frac{1}{2}[\mathbf{d} - \mathbf{g}(\mathbf{y}_s)]^T \mathbf{R}_{\epsilon\epsilon}^{-1}[\mathbf{d} - \mathbf{g}(\mathbf{y}_s)]\right\}}{\sqrt{(2\pi)^N \det(\mathbf{R}_{\epsilon\epsilon})}}, \quad (51.83)$$

where $\mathbf{R}_{\epsilon\epsilon} \triangleq E\{\epsilon\epsilon^T\}$ is the covariance matrix of ϵ .

Note that $\mathbf{R}_{\epsilon\epsilon}$ is independent of $\mathbf{y}_s$ by assumption. Since digital equipment is used to sample the microphone waveforms and estimate the TDOAs, the error introduced by discrete-time processing also has to be taken into account. When this is done, the measurement error is no longer Gaussian and is more-properly modeled as a mixture of Gaussian noise and noise that is uniformly distributed over $[-T_s c/2, T_s c/2]$, where T_s is the sampling period. As an example, for a digital source location estimator with an 8 KHz sampling rate operating at room temperature (25 °C, implying that $c \approx 346.25$ m/s), the maximum error in range difference estimates due to sampling is about ± 2.164 cm, which leads to considerable errors in the location estimate, especially when the source is far from the microphone array.

Under the measurement model given by (51.82), we are now faced with the parameter estimation problem of extracting the source location information from the mis-measured range differences or the equivalent TDOAs. For an unbiased estimator, a Cramér–Rao lower bound (CRLB) can be placed on the variance of each estimated coordinate of the source location. However, since the range difference function $\mathbf{g}(\mathbf{y}_s)$ in the measurement model is nonlinear in the parameters under estimation, it is very difficult (or even impossible) to find an unbiased estimator that is mathematically simple and attains the CRLB. The CRLB is usually used as a benchmark against which the statistical efficiency of any unbiased estimators can be compared.

In general, without any assumptions made about the PDF of the measurement error ϵ , the CRLB of the i -th ($i = 1, 2, 3$) parameter variance is found as the $[i, i]$ element of the inverse of the Fisher information matrix defined by [51.51]:

$$[\mathbf{I}(\mathbf{y}_s)]_{ij} \triangleq -E\left(\frac{\partial^2 \ln p(\mathbf{d}|\mathbf{y}_s)}{\partial y_{s,i} \partial y_{s,j}}\right), \quad i, j = 1, 2, 3, \quad (51.84)$$

where the three parameters of $\mathbf{y}_s$, i. e., $y_{s,1}$, $y_{s,2}$, and $y_{s,3}$, are the x , y , and z coordinates of the source location, respectively.

In the case of a Gaussian measurement error, the Fisher information matrix becomes [51.52]:

$$\mathbf{I}(\mathbf{y}_s) = \left(\frac{\partial \mathbf{g}(\mathbf{y}_s)}{\partial \mathbf{y}_s}\right)^T \mathbf{R}_{\epsilon\epsilon}^{-1} \left(\frac{\partial \mathbf{g}(\mathbf{y}_s)}{\partial \mathbf{y}_s}\right), \quad (51.85)$$

where $\partial \mathbf{g}(\mathbf{y}_s)/\partial \mathbf{y}_s$ is an $(N-1) \times 3$ Jacobian matrix defined as,

$$\begin{aligned} \frac{\partial \mathbf{g}(\mathbf{y}_s)}{\partial \mathbf{y}_s} &= \begin{pmatrix} \frac{\partial g_2(\mathbf{y}_s)}{\partial x_s} & \frac{\partial g_2(\mathbf{y}_s)}{\partial y_s} & \frac{\partial g_2(\mathbf{y}_s)}{\partial z_s} \\ \frac{\partial g_3(\mathbf{y}_s)}{\partial x_s} & \frac{\partial g_3(\mathbf{y}_s)}{\partial y_s} & \frac{\partial g_3(\mathbf{y}_s)}{\partial z_s} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_N(\mathbf{y}_s)}{\partial x_s} & \frac{\partial g_N(\mathbf{y}_s)}{\partial y_s} & \frac{\partial g_N(\mathbf{y}_s)}{\partial z_s} \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{u}_{2 \rightarrow s} - \mathbf{u}_{1 \rightarrow s})^T \\ (\mathbf{u}_{3 \rightarrow s} - \mathbf{u}_{1 \rightarrow s})^T \\ \vdots \\ (\mathbf{u}_{N \rightarrow s} - \mathbf{u}_{1 \rightarrow s})^T \end{pmatrix}, \end{aligned} \quad (51.86)$$

and

$$\mathbf{u}_{n \rightarrow s} \triangleq \frac{\mathbf{y}_s - \mathbf{y}_n}{\|\mathbf{y}_s - \mathbf{y}_n\|} = \frac{\mathbf{y}_s - \mathbf{y}_n}{D_n}, \quad n = 1, 2, \dots, N, \quad (51.87)$$

is the normalized vector of unit length pointing from the n -th microphone to the sound source.

51.3.3 Maximum-Likelihood Estimator

The measurement model for the source localization problem is apparently nonlinear; an efficient estimator that attains the CRLB may not exist or otherwise might be impossible to identify. In practice, the maximum-likelihood estimator is often used since it has the well-proven advantage of asymptotic efficiency for a large sample space.

To apply the maximum-likelihood principle, the statistical characteristics of the measurements need to be known or properly assumed prior to any processing. From the central limit theorem and also for mathematical simplicity, the measurement error is usually modeled as Gaussian and the likelihood function is given by (51.83), which is considered as a function of the source position $\mathbf{y}_s$ under estimation.

Since the exponential function is monotonically increasing, the MLE is equivalent to minimizing a (log-

likelihood) cost function defined as,

$$\mathcal{L}(\mathbf{y}_s) \triangleq [\mathbf{d} - \mathbf{g}(\mathbf{y}_s)]^T \mathbf{R}_{\epsilon\epsilon}^{-1} [\mathbf{d} - \mathbf{g}(\mathbf{y}_s)] . \quad (51.88)$$

Direct estimation of the minimizer is generally not practical. If the noise signals at different microphones are assumed to be uncorrelated, the covariance matrix is diagonal

$$\mathbf{R}_{\epsilon\epsilon} = \text{diag}(\sigma_{\epsilon_2}^2, \sigma_{\epsilon_3}^2, \dots, \sigma_{\epsilon_N}^2), \quad (51.89)$$

where $\sigma_{\epsilon_n}^2 = E\{\epsilon_n^2\}$ ($n = 2, 3, \dots, N$) is the variance of ϵ_n , and the cost function (51.88) becomes,

$$\mathcal{L}(\mathbf{y}_s) = \sum_{n=2}^N \frac{[d_{n1} - g_n(\mathbf{y}_s)]^2}{\sigma_{\epsilon_n}^2}. \quad (51.90)$$

Among other approaches, the steepest-descent algorithm can be used to find $\hat{\mathbf{y}}_s^{\text{MLE}}$ iteratively with

$$\hat{\mathbf{y}}_s(k+1) = \hat{\mathbf{y}}_s(k) - \frac{1}{2}\mu \nabla \mathcal{L}[\hat{\mathbf{y}}_s(k)], \quad (51.91)$$

where μ is the step size.

The foregoing MLE can be determined and is asymptotically optimal for this problem only if its two assumptions (Gaussian and uncorrelated measurement noise) hold. However, this is not the case in practice, as discussed in Sect. 51.3. Furthermore, the number of microphones in an array for camera pointing or beamformer steering is always limited, which makes the source localization a finite-sample rather than a large-sample problem. In addition, the cost function (51.90) is generally not strictly convex. In order to avoid a local minimum with the steepest-descent algorithm, we need to select a good initial guess/estimate of the source location, which is difficult to do in practice, and convergence of the iterative algorithm to the desired solution cannot be guaranteed.

51.3.4 Least-Squares Estimators

Two limitations of the MLE are that probabilistic assumptions have to be made about the measured range differences and that the iterative algorithm to find the solution is computationally intensive. An alternative method is the well-known least-squares estimator (LSE). The LSE makes no probabilistic assumptions about the data and hence can be applied to the source localization problem in which a precise statistical characterization of the data is hard to determine. Furthermore, an LSE usually produces a closed-form estimate, which is desirable in real-time applications.

51.3.5 Least-Squares Error Criteria

In the least-squares (LS) approach, we attempt to minimize a squared error function that is zero in the absence of noise and model inaccuracies. Various error functions can be defined for closeness from the assumed (noiseless) signal based on hypothesized parameters for the observed data. When these are applied, various LSEs can be derived. For the source localization problem two LS error criteria have been proposed.

Hyperbolic LS Error Function

The first LS error function is defined as the difference between the observed range difference and that generated by a signal model depending upon the unknown parameters:

$$\mathbf{e}_h(\mathbf{y}_s) \triangleq \mathbf{d} - \mathbf{g}(\mathbf{y}_s), \quad (51.92)$$

and the corresponding LS criterion is given by

$$J_h = \mathbf{e}_h^T \mathbf{e}_h = [\mathbf{d} - \mathbf{g}(\mathbf{y}_s)]^T [\mathbf{d} - \mathbf{g}(\mathbf{y}_s)]. \quad (51.93)$$

In the source localization problem, an observed range difference d_{n1} defines a hyperboloid in 3-D space. All points lying on such a hyperboloid are potential source locations and all have the same range difference d_{n1} to the two microphones n and 1. Therefore, a sound source that is located by minimizing (51.93) has the shortest distance to all hyperboloids associated with different microphone pairs and specified by the estimated range differences, after which fact the error criterion (51.93) is termed the hyperbolic error criterion (subscript 'h').

In (51.92), the signal model $\mathbf{g}(\mathbf{y}_s)$ consists of a set of hyperbolic functions. Since they are nonlinear, minimizing (51.93) leads to a mathematically intractable solution as N becomes large. Moreover, the hyperbolic function is very sensitive to noise, especially for far-field sources. As a result, it is rarely used in practice.

When the statistical characteristics of the corrupting noise are unknown, uncorrelated *white* Gaussian noise is one reasonable assumption. In this case, it is not surprising that the hyperbolic LSE and the MLE minimize (maximize) similar criteria.

Spherical LS Error Function

The second LS criterion is based on the errors found in the distances from a hypothesized source location to the microphones. In the absence of measurement errors, the correct source location is preferably at the intersection of a group of spheres centered at the microphones. When

measurement errors are present, the best estimate of the source location would be the point that yields the shortest distance to those spheres defined by the range differences and the hypothesized source range.

Consider the distance D_n from the n -th microphone to the source. From the definition of the range difference (51.78) and the fact that $D_1 = \Omega_s$, we have:

$$\hat{D}_n = \Omega_s + d_{n1}, \quad (51.94)$$

where $\hat{D}_n$ denotes an observation based on the measured range difference. From the inner product, we can derive the true value for D_n^2 , the square of the noise-free distance generated by a spherical signal model:

$$D_n^2 = \|\mathbf{y}_n - \mathbf{y}_s\|^2 = \Omega_s^2 - 2\mathbf{y}_n^T \mathbf{y}_s + \Omega_s^2. \quad (51.95)$$

The spherical (subscript 'sp') LS error function is then defined as the difference between the measured and hypothesized values

$$\begin{aligned} e_{\text{sp},n}(\mathbf{y}_s) &\triangleq \frac{1}{2} (\hat{D}_n^2 - D_n^2) \\ &= \mathbf{y}_n^T \mathbf{y}_s + d_{n1} \Omega_s - \frac{1}{2} (\Omega_s^2 - d_{n1}^2), \\ n &= 2, 3, \dots, N. \end{aligned} \quad (51.96)$$

Putting the $N - 1$ errors together and writing them in vector form gives,

$$\mathbf{e}_{\text{sp}}(\mathbf{r}_s) = \mathbf{A}\boldsymbol{\theta} - \boldsymbol{\xi}, \quad (51.97)$$

where

$$\begin{aligned} \mathbf{A} &\triangleq [\mathbf{S} \mid \mathbf{d}], \quad \mathbf{S} \triangleq \begin{pmatrix} x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{pmatrix}, \\ \boldsymbol{\theta} &\triangleq \begin{pmatrix} x_s \\ y_s \\ z_s \\ \Omega_s \end{pmatrix}, \quad \boldsymbol{\xi} \triangleq \frac{1}{2} \begin{pmatrix} \Omega_s^2 - d_{21}^2 \\ \Omega_s^2 - d_{31}^2 \\ \vdots \\ \Omega_s^2 - d_{N1}^2 \end{pmatrix}, \end{aligned}$$

and $[\mathbf{S} \mid \mathbf{d}]$ indicates that $\mathbf{S}$ and $\mathbf{d}$ are stacked side by side. The corresponding LS criterion is then given by:

$$J_{\text{sp}} = \mathbf{e}_{\text{sp}}^T \mathbf{e}_{\text{sp}} = [\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\xi}]^T [\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\xi}]. \quad (51.98)$$

In contrast to the hyperbolic error function (51.92), the spherical error function (51.97) is linear in $\mathbf{y}_s$ given Ω_s and vice versa. Therefore, the computational complexity to find a solution will *not* dramatically increase as N gets large.

51.3.6 Spherical Intersection (SX) Estimator

The SX source location estimator employs the spherical error and solves the problem in two steps [51.53]. It first finds the least-squares solution for $\mathbf{y}_s$ in terms of Ω_s ,

$$\mathbf{y}_s = \mathbf{S}^\dagger (\boldsymbol{\xi} - \Omega_s \mathbf{d}), \quad (51.99)$$

where $\mathbf{S}^\dagger \triangleq (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T$ is the pseudo-inverse of the matrix $\mathbf{S}$. Then, substituting (51.99) into the constraint $\Omega_s^2 = \mathbf{y}_s^T \mathbf{y}_s$ yields a quadratic equation as follows

$$\Omega_s^2 = \left[\mathbf{S}^\dagger (\boldsymbol{\xi} - \Omega_s \mathbf{d}) \right]^T \left[\mathbf{S}^\dagger (\boldsymbol{\xi} - \Omega_s \mathbf{d}) \right]. \quad (51.100)$$

After expansion, this becomes

$$\alpha_1 \Omega_s^2 + \alpha_2 \Omega_s + \alpha_3 = 0, \quad (51.101)$$

where

$$\begin{aligned} \alpha_1 &= 1 - \|\mathbf{S}^\dagger \mathbf{d}\|^2, \quad \alpha_2 = 2\boldsymbol{\xi}^T \mathbf{S}^\dagger \mathbf{S}^\dagger \mathbf{d}, \\ \alpha_3 &= -\|\mathbf{S}^\dagger \boldsymbol{\xi}\|^2. \end{aligned}$$

The valid (real, positive) root is taken as an estimate of the source range Ω_s and is then substituted into (51.99) to calculate the SX estimate $\hat{\mathbf{y}}_s^{\text{SX}}$ of the source location.

In the SX estimation procedure, the solution of the quadratic equation (51.101) for the source range Ω_s is required. This solution must be a positive value. If a real positive root is not available, the SX solution does not *exist*. On the contrary, if both of the roots are real and greater than 0, then the SX solution is not *unique*. In both cases, the SX source location estimator fails to produce a reliable estimate, which is not desirable for a real-time implementation.

51.3.7 Spherical Interpolation (SI) Estimator

In order to overcome this drawback of the SX algorithm, a spherical interpolation estimator was proposed in [51.54], which attempts to relax the restriction $\Omega_s = \|\mathbf{y}_s\|$ by estimating Ω_s in the least-squares sense.

To begin, we substitute the least-squares solution (51.99) into the original spherical equation $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\xi}$ to obtain

$$\Omega_s \mathbf{P}_{\mathbf{S}^\perp} \mathbf{d} = \mathbf{P}_{\mathbf{S}^\perp} \boldsymbol{\xi}, \quad (51.102)$$

where

$$\mathbf{P}_{\mathbf{S}^\perp} \triangleq \mathbf{I}_{N \times N} - \mathbf{S} \mathbf{S}^\dagger, \quad (51.103)$$

and $\mathbf{I}_{N \times N}$ is the $N \times N$ identity matrix. The matrix $\mathbf{P}_{\mathbf{S}^\perp}$ is a projection matrix that projects a vector, when multiplied by the matrix, onto a space that is orthogonal

to the column space of S . Such a projection matrix is symmetric (i.e., $P_{S^\perp} = P_{S^\perp}^T$) and idempotent (i.e., $P_{S^\perp} = P_{S^\perp} \cdot P_{S^\perp}$). Then the least-squares solution to (51.102) is given by

$$\hat{\mathbf{d}}_s^{\text{SI}} = \frac{d^T P_{S^\perp} \xi}{d^T P_{S^\perp} d}. \quad (51.104)$$

Substituting this solution into (51.99) yields the SI estimate

$$\hat{\mathbf{y}}_s^{\text{SI}} = S^\dagger \left[I_{N \times N} - \left(\frac{d d^T P_{S^\perp}}{d^T P_{S^\perp} d} \right) \right] \xi. \quad (51.105)$$

In practice, the SI estimator performs better, but is computationally slightly more complex, than the SX estimator.

51.3.8 Linear-Correction Least-Squares Estimator

Finding the LSE based on the spherical error criterion (51.98) is a linear minimization problem, i.e.,

$$\hat{\theta}^{\text{LSE}} = \arg \min_{\theta} (\mathbf{A}\theta - \xi)^T (\mathbf{A}\theta - \xi) \quad (51.106)$$

subject to a quadratic constraint

$$\theta^T \Xi \theta = 0, \quad (51.107)$$

where $\Xi \triangleq \text{diag}(1, 1, 1, -1)$ is a diagonal and orthonormal matrix.

For such a constrained minimization problem, the technique of Lagrange multipliers will be used and the source location is determined by minimizing the Lagrangian

$$\begin{aligned} \mathcal{L}(\theta, \kappa) &= J_{\text{sp}} + \kappa \theta^T \Xi \theta \\ &= (\mathbf{A}\theta - \xi)^T (\mathbf{A}\theta - \xi) + \kappa \theta^T \Xi \theta, \end{aligned} \quad (51.108)$$

where κ is a Lagrange multiplier. Expanding this expression yields

$$\mathcal{L}(\theta, \kappa) = \theta^T (\mathbf{A}^T \mathbf{A} + \kappa \Xi) \theta - 2\xi^T \mathbf{A}\theta + \xi^T \xi. \quad (51.109)$$

The necessary conditions for minimizing (51.109) can be obtained by taking the gradient of $\mathcal{L}(\theta, \kappa)$ with respect to θ and equating the result to zero. This produces:

$$\frac{\partial \mathcal{L}(\theta, \kappa)}{\partial \theta} = 2 (\mathbf{A}^T \mathbf{A} + \kappa \Xi) \theta - 2\mathbf{A}^T \xi = \mathbf{0}. \quad (51.110)$$

Solving (51.110) for θ yields the constrained least-squares estimate

$$\hat{\theta} = (\mathbf{A}^T \mathbf{A} + \kappa \Xi)^{-1} \mathbf{A}^T \xi, \quad (51.111)$$

where κ is yet to be determined.

In order to find κ , we can impose the quadratic constraint directly by substituting (51.111) into (51.107), which leads to

$$\xi^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \kappa \Xi)^{-1} \Xi (\mathbf{A}^T \mathbf{A} + \kappa \Xi)^{-1} \mathbf{A}^T \xi = 0. \quad (51.112)$$

With eigenvalue analysis, the matrix $\mathbf{A}^T \mathbf{A} \Xi$ can be decomposed as

$$\mathbf{A}^T \mathbf{A} \Xi = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1}, \quad (51.113)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_4)$ and $\lambda_i, i = 1, \dots, 4$, are the eigenvalues of the matrix $\mathbf{A}^T \mathbf{A} \Xi$. Substituting (51.113) into (51.112), we may rewrite the constraint as:

$$\mathbf{p}^T (\mathbf{\Lambda} + \kappa \mathbf{I})^{-2} \mathbf{q} = 0, \quad (51.114)$$

where

$$\begin{aligned} \mathbf{p} &= \mathbf{U}^T \Xi \mathbf{A}^T \xi, \\ \mathbf{q} &= \mathbf{U}^T \mathbf{A}^T \xi. \end{aligned}$$

Let us define a function of the Lagrange multiplier as follows

$$f(\kappa) \triangleq \mathbf{p}^T (\mathbf{\Lambda} + \kappa \mathbf{I})^{-2} \mathbf{q} = \sum_{i=1}^4 \frac{p_i q_i}{(\kappa + \lambda_i)^2}. \quad (51.115)$$

This is a polynomial of degree eight and, because of its complexity, numerical methods need to be used for root searching. Since the root of (51.115) for κ is not unique, a two-step procedure will be followed such that the desired source location could be found.

Unconstrained Spherical Least-Squares Estimator

In the first step, we assume that x_s, y_s, z_s , and Ω_s are mutually independent or equivalently we disregard the quadratic constraint (51.107) intentionally. Then the LS solution minimizing (51.98) for θ (the source location as well as its range) is given by

$$\hat{\theta}_1 = \mathbf{A}^\dagger \xi, \quad (51.116)$$

where $\mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is the pseudo-inverse of the matrix $\mathbf{A}$.

A good parameter estimator first and foremost needs to be unbiased. For such an unconstrained spherical least-squares estimator, the bias and covariance matrix can be approximated by using the following perturbation analysis method.

When measurement errors are present in the range differences, $\mathbf{A}$, $\boldsymbol{\xi}$, and the parameter estimate $\hat{\boldsymbol{\theta}}_1$ deviate from their true values and can be expressed as:

$$\mathbf{A} = \mathbf{A}^t + \Delta\mathbf{A}, \quad \boldsymbol{\xi} = \boldsymbol{\xi}^t + \Delta\boldsymbol{\xi}, \quad \hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}, \quad (51.117)$$

where variables with superscript 't' denote the true values, which also satisfy

$$\boldsymbol{\theta}^t = \mathbf{A}^{t\dagger} \boldsymbol{\xi}^t. \quad (51.118)$$

If the magnitudes of the perturbations are small, the second-order errors are insignificant compared to their first-order counterparts and therefore can be neglected for simplicity, which then yields:

$$\Delta\mathbf{A} = [\mathbf{0} \ \boldsymbol{\epsilon}], \quad \Delta\boldsymbol{\xi} \approx -\mathbf{d}^t \odot \boldsymbol{\epsilon}, \quad (51.119)$$

where $\odot$ denotes the Schur (element-by-element) product. Substituting (51.117) into (51.116) gives,

$$\begin{aligned} & (\mathbf{A}^t + \Delta\mathbf{A})^T (\mathbf{A}^t + \Delta\mathbf{A}) (\boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}) \\ &= (\mathbf{A}^t + \Delta\mathbf{A})^T (\boldsymbol{\xi}^t + \Delta\boldsymbol{\xi}). \end{aligned} \quad (51.120)$$

Retaining only the linear perturbation terms and using (51.118) and (51.119) produces:

$$\Delta\boldsymbol{\theta} \approx -\mathbf{A}^{t\dagger} \mathbf{D} \boldsymbol{\epsilon}, \quad (51.121)$$

where $\mathbf{D} \triangleq \text{diag}(D_2, D_3, \dots, D_N)$ is a diagonal matrix. Since the measurement error $\boldsymbol{\epsilon}$ in the range differences has zero mean, $\hat{\boldsymbol{\theta}}_1$ is an unbiased estimate of $\boldsymbol{\theta}^t$ when the small-error assumption holds:

$$E\{\Delta\boldsymbol{\theta}\} \approx E\{-\mathbf{A}^{t\dagger} \mathbf{D} \boldsymbol{\epsilon}\} = \mathbf{0}_{4 \times 1}. \quad (51.122)$$

The covariance matrix of $\Delta\boldsymbol{\theta}$ is then found as,

$$\mathbf{R}_{\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}} = E\{\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T\} = \mathbf{A}^{t\dagger} \mathbf{D} \mathbf{R}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}} \mathbf{D} \mathbf{A}^{t\dagger T}, \quad (51.123)$$

where $\mathbf{R}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}$ is known or is properly assumed a priori. Theoretically, the covariance matrix $\mathbf{R}_{\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}}$ cannot be calculated since it contains true values. Nevertheless, it can be approximated by using the values in $\hat{\boldsymbol{\theta}}_1$ with sufficient accuracy, as suggested by numerical studies.

In the first unconstrained spherical LS estimate (51.116), the range information is redundant because of the independence assumption on the source location and range. If that information is simply discarded, the source location estimate is the same as the SI estimate but with less computational complexity [51.55]. To demonstrate this, we first write (51.116) in block form as

$$\hat{\boldsymbol{\theta}}_1 = \begin{pmatrix} \mathbf{S}^T \mathbf{S} & \mathbf{S}^T \mathbf{d} \\ \mathbf{d}^T \mathbf{S} & \mathbf{d}^T \mathbf{d} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}^T \\ \mathbf{d}^T \end{pmatrix} \boldsymbol{\xi}. \quad (51.124)$$

It can easily be shown that:

$$\begin{pmatrix} \mathbf{S}^T \mathbf{S} & \mathbf{S}^T \mathbf{d} \\ \mathbf{d}^T \mathbf{S} & \mathbf{d}^T \mathbf{d} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{Q} & \mathbf{v} \\ \mathbf{v}^T & \varrho \end{pmatrix}, \quad (51.125)$$

where

$$\begin{aligned} \mathbf{v} &= -\left(\mathbf{S}^T \mathbf{S} - \frac{\mathbf{S}^T \mathbf{d} \mathbf{d}^T \mathbf{S}}{\mathbf{d}^T \mathbf{d}}\right)^{-1} \frac{\mathbf{S}^T \mathbf{d}}{\mathbf{d}^T \mathbf{d}}, \\ \mathbf{Q} &= \left(\mathbf{S}^T \mathbf{S}\right)^{-1} \left[\mathbf{I} - \left(\mathbf{S}^T \mathbf{d}\right) \mathbf{v}^T\right], \\ \varrho &= \frac{1 - (\mathbf{d}^T \mathbf{S}) \mathbf{v}}{\mathbf{d}^T \mathbf{d}}. \end{aligned}$$

Next, we define another projection matrix $\mathbf{P}_{\mathbf{d}^\perp}$ associated with the $\mathbf{d}$ -orthogonal space:

$$\mathbf{P}_{\mathbf{d}^\perp} \triangleq \mathbf{I} - \frac{\mathbf{d} \mathbf{d}^T}{\mathbf{d}^T \mathbf{d}}, \quad (51.126)$$

and find

$$\mathbf{v} = -\left(\mathbf{S}^T \mathbf{P}_{\mathbf{d}^\perp} \mathbf{S}\right)^{-1} \frac{\mathbf{S}^T \mathbf{d}}{\mathbf{d}^T \mathbf{d}}, \quad (51.127)$$

$$\mathbf{Q} = \left(\mathbf{S}^T \mathbf{P}_{\mathbf{d}^\perp} \mathbf{S}\right)^{-1}. \quad (51.128)$$

Substituting (51.125) together with (51.127) and (51.128) into (51.124) yields the unconstrained spherical LS estimate for the source coordinates:

$$\hat{\mathbf{y}}_{s,1} = \left(\mathbf{S}^T \mathbf{P}_{\mathbf{d}^\perp} \mathbf{S}\right)^{-1} \mathbf{S}^T \mathbf{P}_{\mathbf{d}^\perp} \boldsymbol{\xi}, \quad (51.129)$$

which is the minimizer of

$$J_1(\mathbf{y}_s) = \|\mathbf{P}_{\mathbf{d}^\perp} \boldsymbol{\xi} - \mathbf{P}_{\mathbf{d}^\perp} \mathbf{S} \mathbf{y}_s\|^2, \quad (51.130)$$

or the least-squares solution to the linear equation

$$\mathbf{P}_{\mathbf{d}^\perp} \mathbf{S} \mathbf{y}_s = \mathbf{P}_{\mathbf{d}^\perp} \boldsymbol{\xi}. \quad (51.131)$$

In fact, the first unconstrained spherical LS estimator tries to approximate the projection of the observation vector $\boldsymbol{\xi}$ with the projections of the column vectors of the microphone location matrix $\mathbf{S}$ onto the $\mathbf{d}$ -orthogonal space. The source location estimate is the coefficient vector associated with the *best* approximation. Clearly from (51.131), this estimation procedure is the generalization of the plane intersection (PI) method proposed in [51.56].

By using the Sherman–Morrison formula [51.57]

$$\left(\mathbf{A} + \mathbf{x} \mathbf{y}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{x} \mathbf{y}^T \mathbf{A}^{-1}}{1 + \mathbf{y}^T \mathbf{A}^{-1} \mathbf{x}}, \quad (51.132)$$

we can expand the item in (51.129) as

$$\left(S^T P_{d^\perp} S \right)^{-1} = \left[S^T S - \left(\frac{S^T d}{d^T d} \right) (S^T d)^T \right]^{-1},$$

and finally show that the unconstrained spherical LS estimate (51.129) is equivalent to the SI estimate (51.105), i. e., $\hat{\mathbf{y}}_{s,1} \equiv \hat{\mathbf{y}}_s^{\text{SI}}$.

Although the unconstrained spherical LS and the SI estimators are mathematically equivalent, they are quite different in computational efficiency due to their different approaches to the source localization problem. The complexities of the SI and unconstrained spherical LS estimators are $\mathcal{O}(N^3)$ and $\mathcal{O}(N)$, respectively. In comparison, the unconstrained spherical LS estimator reduces the complexity of the SI estimator by a factor of N^2 , which is significant when N is large (more microphones are used).

Linear Correction

In the previous subsection, we developed the unconstrained spherical LS estimator (USLSE) for source localization and demonstrated that it is mathematically equivalent to the SI estimator but with less computational complexity. Although the USLSE/SI estimates can be accurate, as indicated in [51.55] among others, it is helpful to exploit the redundancy of source range to improve the statistical efficiency (i. e., to reduce the variance of source location estimates) of the overall estimation procedure. Therefore, in the second step, we intend to correct the USLS estimate $\hat{\boldsymbol{\theta}}_1$ to make a better estimate $\hat{\boldsymbol{\theta}}_2$ of $\boldsymbol{\theta}$. This new estimate should be in the neighborhood of $\hat{\boldsymbol{\theta}}_1$ and should obey the constraint (51.107). We expect that the corrected estimate will still be unbiased and have a smaller variance.

To begin, we substitute $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}$ into (51.110) and expand the expression to find

$$A^T A \hat{\boldsymbol{\theta}}_1 + \kappa \boldsymbol{\Xi} \hat{\boldsymbol{\theta}}_1 - (A^T A + \kappa \boldsymbol{\Xi}) \Delta\boldsymbol{\theta} = A^T \boldsymbol{\xi}. \quad (51.133)$$

Combined with (51.116), (51.133) becomes

$$(A^T A + \kappa \boldsymbol{\Xi}) \Delta\boldsymbol{\theta} = \kappa \boldsymbol{\Xi} \hat{\boldsymbol{\theta}}_1, \quad (51.134)$$

and hence

$$\Delta\boldsymbol{\theta} = \kappa \left(A^T A \right)^{-1} \boldsymbol{\Xi} \boldsymbol{\theta}^t. \quad (51.135)$$

Substituting (51.135) into $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}$ yields

$$\hat{\boldsymbol{\theta}}_1 = \left[I + \kappa \left(A^T A \right)^{-1} \boldsymbol{\Xi} \right] \boldsymbol{\theta}^t. \quad (51.136)$$

Solving (51.136) for $\boldsymbol{\theta}^t$ produces the corrected estimate $\hat{\boldsymbol{\theta}}_2$ and also the final output of the linear-correction least-squares (LCLS) estimator:

$$\hat{\boldsymbol{\theta}}_2 = \left[I + \kappa \left(A^T A \right)^{-1} \boldsymbol{\Xi} \right]^{-1} \hat{\boldsymbol{\theta}}_1. \quad (51.137)$$

Equation (51.137) suggests how the second-step processing updates the source location estimate based on the first unconstrained spherical least squares result, or equivalently the SI estimate. If the regularity condition [51.58]

$$\lim_{i \rightarrow \infty} \left[\kappa (A^T A)^{-1} \boldsymbol{\Xi} \right]^i = \mathbf{0} \quad (51.138)$$

is satisfied, then the estimate $\hat{\boldsymbol{\theta}}_2$ can be expanded in a Neumann series :

$$\begin{aligned} \hat{\boldsymbol{\theta}}_2 &= \left\{ I + \left[-\kappa \left(A^T A \right)^{-1} \boldsymbol{\Xi} \right] \right. \\ &\quad \left. + \left[-\kappa \left[A^T A \right]^{-1} \boldsymbol{\Xi} \right]^2 + \cdots \right\} \hat{\boldsymbol{\theta}}_1 \\ &= \hat{\boldsymbol{\theta}}_1 + \sum_{i=1}^{\infty} \left[-\kappa \left(A^T A \right)^{-1} \boldsymbol{\Xi} \right]^i \hat{\boldsymbol{\theta}}_1, \end{aligned} \quad (51.139)$$

where the second term is the linear correction. Equation (51.138) implies that, in order to avoid divergence, the Lagrange multiplier κ should be small. In addition, κ needs to be determined carefully such that $\hat{\boldsymbol{\theta}}_2$ obeys the quadratic constraint (51.107).

Because the function $f(\kappa)$ is smooth near $\kappa = 0$ (corresponding to the neighborhood of $\hat{\boldsymbol{\theta}}_1$), as suggested by numerical experiments, the secant method [51.59] can be used to determine its desired root. Two reasonable initial points can be chosen as:

$$\kappa_0 = 0, \quad \kappa_1 = \beta, \quad (51.140)$$

where the small number β depends on the array geometry. Five iterations should be sufficient to give an accurate approximation to the root.

The idea of exploiting the relationship between a sound source's range and its location coordinates to improve the estimation efficiency of the SI estimator was first suggested by Chan and Ho [51.52] with a quadratic correction. Accordingly, they constructed a quadratic data model for $\hat{\boldsymbol{\theta}}_1$.

$$\hat{\boldsymbol{\theta}}_1 \odot \hat{\boldsymbol{\theta}}_1 = T(\boldsymbol{y}_s \odot \boldsymbol{y}_s) + \mathbf{v}, \quad (51.141)$$

where

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

is a constant matrix, and $\mathbf{n}$ is the corrupting noise. In contrast to the linear correction technique based on the Lagrange multiplier, the quadratic counterpart needs to know the covariance matrix $\mathbf{R}_{\epsilon\epsilon}$ of measurement errors in the range differences a priori. In a real-time digital source localization system, a poorly estimated $\mathbf{R}_{\epsilon\epsilon}$ will lead to performance degradation. In addition, the quadratic-correction least-squares estimation procedure uses the perturbation approaches to linearly approximate $\Delta\theta$ and $\mathbf{v}$ in (51.117) and (51.141), respectively. Therefore, the approximations of their corresponding covariance matrices $\mathbf{R}_{\Delta\theta\Delta\theta}$ and $\mathbf{R}_{\mathbf{v}\mathbf{v}}$ can be good only when the noise level is low. When noise is at a practically high level, the quadratic-correction least squares estimate has a large bias and a high variance. Furthermore, since the true value of the source location that is necessary for calculating $\mathbf{R}_{\Delta\theta\Delta\theta}$ and $\mathbf{R}_{\mathbf{v}\mathbf{v}}$ cannot be known theoretically, the estimated source location has to be uti-

lized for approximation. It was suggested in [51.52] that several iterations in the second correction stage would improve estimation accuracy. However, while the bias is suppressed after iterations, the estimate is closer to the SI solution and the variance is boosted, as reported in [51.50]. Finally, the direct solutions of the quadratic-correction least-squares estimator are the squares of the source location coordinates $\mathbf{y}_s \odot \mathbf{y}_s$. In 3-D space, these correspond to eight positions, which introduce decision ambiguities. Other physical criteria, such as the domain of interest, were suggested, but these are hard to define in practical situations, particularly when one of the source coordinates is close to zero.

In comparison, the linear-correction method updates the source location estimate of the first unconstrained spherical LS estimator without making any assumption about the error covariance matrix and without resorting to a linear approximation. Even though we need to find a small root of function (51.115) for the Lagrange multiplier κ that satisfies the regularity condition (51.138), the function $f(\kappa)$ is smooth around zero and the solution can easily be determined using the secant method. The linear-correction method achieves a better balance between computational complexity and estimation accuracy.

51.4 Summary

In this chapter, we have explained why TDE-based two-stage approaches have become the technique of choice for real-time speech source localization using microphone arrays. We presented a comprehensive overview of the research and development on this technology over the last three decades.

The chapter consists of two parts. The first part was devoted to the TDE problem. We began with a close examination of two signal models used in this problem. Their strengths and drawbacks were discussed. We reviewed the well-known generalized cross-correlation (GCC) methods and intended to treat this class of methods as a benchmark for a number of recently developed new algorithms. We made it clear by analysis that the GCC methods cannot cope well with room reverberation and described two strategies to combat this effect. One strategy is to employ the more-realistic real reverberant model in the formulation of the TDE problem and try to blindly identify the acoustic channel impulse responses from which the time difference

of arrival (TDOA) is inferred. The adaptive eigenvalue decomposition algorithm and the adaptive blind multichannel identification algorithms were developed. The other strategy is to exploit the TDOA redundancy among different microphone pairs. This leads to the multichannel spatial predication and interpolation methods, the multichannel cross-correlation coefficient method, and the minimum-entropy method, which have all been thoroughly explored.

The second part of the Chapter surveyed the source localization techniques. The localization problem was postulated from the perspective of estimation theory, and the Cramér–Rao lower bound for unbiased location estimators was derived. After an insightful review of conventional approaches ranging from maximum-likelihood to least-squares estimators, we presented a recently developed linear-correction least-squares algorithm that is more robust to measurement errors and that is computationally as well as statistically more efficient.

References

- 51.1 D.R. Fischell, C.H. Coker: A speech direction finder, Proc. ICASSP (1984) pp. 19.8.1–19.8.4.
- 51.2 H.F. Silverman: Some analysis of microphone arrays for speech data analysis, IEEE Trans. ASSP **35**, 1699–1712 (1987)
- 51.3 J.L. Flanagan, A. Surendran, E. Jan: Spatially selective sound capture for speech and audio processing, Speech Commun. **13**, 207–222 (1993)
- 51.4 D.B. Ward, G.W. Elko: Mixed nearfield/farfield beamforming: a new technique for speech acquisition in a reverberant environment, Proc. IEEE ASSP Workshop Appl. Signal Process. Audio Acoust. (1997)
- 51.5 D.V. Rabinkin, R.J. Ranomeron, J.C. French, J.L. Flanagan: A DSP implementation of source location using microphone arrays, Proc. SPIE **2846**, 88–99 (1996)
- 51.6 H. Wang, P. Chu: Voice source localization for automatic camera pointing system in videoconferencing, Proc. IEEE ASSP Workshop Appl. Signal Process. Audio Acoust. (1997)
- 51.7 C. Wang, M.S. Brandstein: A hybrid real-time face tracking system, Proc. ICASSP, Vol. 6 (1998) pp. 3737–3741
- 51.8 Y. Huang, J. Benesty, G.W. Elko: Microphone arrays for video camera steering. In: *Acoustic Signal Processing for Telecommunication*, ed. by S.L. Gay, J. Benesty (Kluwer Academic, Boston 2000) pp. 239–259, chap. 11
- 51.9 S. Haykin: Radar array processing for angle of arrival estimation. In: *Array Signal Process*, ed. by S. Haykin (Prentice-Hall, Englewood Cliffs 1985)
- 51.10 H. Krim, M. Viberg: Two decades of array signal processing research: the parametric approach, IEEE Signal Process. Mag. **13**(4), 67–94 (1996)
- 51.11 R.J. Vaccaro: The past, present; future of underwater acoustic signal processing, IEEE Signal Process. Mag. **15**, 21–51 (1998)
- 51.12 D.V. Sidorovich, A.B. Gershman: Two-dimensional wideband interpolated root-MUSIC applied to measured seismic data, IEEE Trans. Signal Process. **46**(8), 2263–2267 (1998)
- 51.13 J. Capon: Maximum-likelihood spectral estimation, Proc. IEEE **57**, 1408–1418 (1969)
- 51.14 R.O. Schmidt: *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation* (Stanford University, Stanford 1981), Ph.D. thesis
- 51.15 W. Bangs, P. Schultheis: Space-time processing for optimal parameter estimation. In: *Signal Process*, ed. by J. Griffiths, P. Stocklin, C. Van Schooneveld (New York, Academic 1973) pp. 577–590
- 51.16 W.R. Hahn, S.A. Tretter: Optimum processing for delay-vector estimation in passive signal arrays, IEEE Trans. Inform. Theory **19**, 608–614 (1973)
- 51.17 M. Wax, T. Kailath: Optimum localization of multiple sources by passive arrays, IEEE Trans. ASSP **31**(5), 1210–1218 (1983)
- 51.18 M.S. Brandstein, H.F. Silverman: A practical methodology for speech source localization with microphone arrays, Comput. Speech Lang. **2**, 91–126 (1997)
- 51.19 C.H. Knapp, G.C. Carter: The generalized correlation method for estimation of time delay, IEEE Trans. ASSP **24**, 320–327 (1976)
- 51.20 G.C. Carter, A.H. Nuttall, P.G. Cable: The smoothed coherence transform, Proc. IEEE **61**, 1497–1498 (1973)
- 51.21 J.P. Ianniello: Time delay estimation via cross-correlation in the presence of large estimation errors, IEEE Trans. ASSP **30**, 998–1003 (1982)
- 51.22 B. Champagne, S. Bédard, A. Stéphenne: Performance of time-delay estimation in presence of room reverberation, IEEE Trans. Speech Audio Process. **4**, 148–152 (1996)
- 51.23 M. Omologo, P. Svaizer: Acoustic event localization using a crosspower-spectrum phase based technique, Proc. ICASSP, Vol. 2 (1994) pp. 273–276
- 51.24 M.S. Brandstein: A pitch-based approach to time-delay estimation of reverberant speech, Proc. IEEE ASSP Workshop Appl. Signal Process. Audio Acoustics (1997)
- 51.25 M. Omologo, P. Svaizer: Acoustic source location in noisy and reverberant environment using CSP analysis, ICASSP, Vol. 2 (1996) pp. 921–924
- 51.26 A. Stéphenne, B. Champagne: Cepstral prefiltering for time delay estimation in reverberant environments, Proc. ICASSP, Vol. 5. (1995) pp. 3055–3058
- 51.27 J. Benesty: Adaptive eigenvalue decomposition algorithm for passive acoustic source localization, J. Acoust. Soc. Am. **107**, 384–391 (2000)
- 51.28 G. Xu, H. Liu, L. Tong, T. Kailath: A least-squares approach to blind channel identification, IEEE Trans. Signal Process. **43**, 2982–2993 (1995)
- 51.29 Y. Huang, J. Benesty: Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization. In: *Adaptive Signal Processing: Application to Real-World Problems*, ed. by J. Benesty, Y. Huang (Springer, Berlin:Heidelberg 2003)
- 51.30 Y. Huang, J. Benesty: Adaptive multi-channel least mean square and Newton algorithms for blind channel identification, Signal Process. **82**, 1127–1138 (2002)
- 51.31 Y. Huang, J. Benesty: A class of frequency-domain adaptive approaches to blind multichannel identification, IEEE Trans. Signal Process. **51**, 11–24 (2003)
- 51.32 Y. Huang, J. Benesty, J. Chen: Optimal step size of the adaptive multichannel LMS algorithm for blind

- SIMO identification, IEEE Signal Process. Lett. **12**, 173–176 (2005)
- 51.33 Y. Huang, J. Benesty, J. Chen: *Acoustic MIMO Signal Process* (Berlin, Springer 2006)
- 51.34 R.L. Kirlin, D.F. Moore, R.F. Kubichek: Improvement of delay measurements from sonar arrays via sequential state estimation, IEEE Trans. ASSP **29**, 514–519 (1981)
- 51.35 T. Nishiura, T. Yamada, S. Nakamura, K. Shikano: Localization of multiple sound sources based on a CSP analysis with a microphone array, Proc. ICASSP (2000) pp. 1053–1055
- 51.36 S.M. Griebel, M.S. Brandstein: Microphone array source localization using realizable delay vectors, Proc. IEEE ASSP Workshop Appl. Signal Process. Audio Acoust. (2001) pp. 71–74
- 51.37 J. DiBiase, H. Silverman, M. Brandstein: Robust localization in reverberant rooms. In: *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M. Brandstein, D. Ward (Springer, Berlin 2001)
- 51.38 J. Chen, J. Benesty, Y. Huang: Robust time delay estimation exploiting redundancy among multiple microphones, IEEE Trans. Speech Audio Process. **11**, 549–557 (2003)
- 51.39 J. Benesty, J. Chen, Y. Huang: Time-delay estimation via linear interpolation and cross-correlation, IEEE Trans. Speech Audio Process. **12**, 509–519 (2004)
- 51.40 J.S. Bendat, A.G. Piersol: *Random Data Analysis and Measurement Procedures* (Wiley, New York 1986)
- 51.41 D. Cochran, H. Gish, D. Sinno: A geometric approach to multichannel signal detection, IEEE Trans. Signal Process. **43**, 2049–2057 (1995)
- 51.42 C.E. Shannon: A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
- 51.43 T.M. Cover, J.A. Thomas: *Elements of Information Theory* (Wiley, New York 1991)
- 51.44 I. Kojadinovic: On the use of mutual information in data analysis: an overview, Int. Symposium on Applied Stochastic Models and Data Analysis (2005)
- 51.45 J. Benesty, Y. Huang, J. Chen: Time delay estimation via minimum entropy, IEEE Signal Process. Lett. **14**, 157–160 (2006)
- 51.46 L.R. Rabiner, R.W. Schafer: *Digital Process. of Speech Signals* (Prentice-Hall, Englewood Cliffs 1978)
- 51.47 S. Gazor, W. Zhang: Speech probability distribution, IEEE Signal Process. Lett. **10**(7), 204–207 (2003)
- 51.48 S. Kotz, T.J. Kozubowski, K. Podgórski: *An asymmetric multivariate Laplace distribution*, Technical Report No. 367, Department of Statistics and Applied Probability (Univ. of California, Santa Barbara 2000)
- 51.49 T. Eltoft, T. Kim, T.-W. Lee: On the multivariate Laplace distribution, IEEE Signal Process. Lett. **13**, 300–303 (2006)
- 51.50 Y. Huang, J. Benesty, G.W. Elko, R.M. Mersereau: Real-time passive source localization: an unbiased linear-correction least-squares approach, IEEE Trans. Speech Audio Process. **9**, 943–956 (2001)
- 51.51 S.M. Kay: *Fundamentals of Statistical Signal Process.: Estimation Theory* (Prentice-Hall, Englewood Cliffs 1993)
- 51.52 Y.T. Chan, K.C. Ho: A simple and efficient estimator for hyperbolic location, IEEE Trans. Signal Process. **42**, 1905–1915 (1994)
- 51.53 H.C. Schau, A.Z. Robinson: Passive source localization employing intersecting spherical surfaces from time-of-arrival differences, IEEE Trans. ASSP **35**, 1223–1225 (1987)
- 51.54 J.S. Abel, J.O. Smith: The spherical interpolation method for closed-form passive source localization using range difference measurements echo cancelation, Proc. ICASSP (1987) pp. 471–474
- 51.55 Y. Huang, J. Benesty, G.W. Elko: Passive acoustic source localization for video camera steering, Proc. IEEE Int. Conf. ACSP, Vol. 2. (2000) pp. 909–912
- 51.56 R.O. Schmidt: A new approach to geometry of range difference location, IEEE Trans. Aerosp. Electron. **8**, 821–835 (1972)
- 51.57 T.K. Moon, W.C. Stirling: *Mathematical Methods and Algorithms* (Prentice-Hall, Upper Saddle River 1999)
- 51.58 C.D. Meyer: *Matrix Analysis and Applied Linear Algebra* (SIAM, Philadelphia 2000)
- 51.59 W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling: *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge 1988)