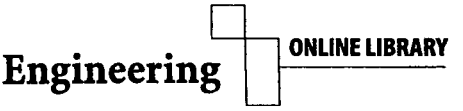
Digital Signal Processing

Michael Brandstein · Darren Ward

Microphone Arrays

Luxottica et al. v. Solos
IPR2026-00512
Petitioners Ex. 1015

Springer-Verlag Berlin Heidelberg GmbH



http://www.springer.de/engine/

Michael Brandstein · Darren Ward (Eds.)

Microphone Arrays

Signal Processing Techniques and Applications

With 149 Figures



Springer

Series Editors

Prof. Dr.-Ing. ARILD LACROIX
Johann-Wolfgang-Goethe-Universität
Institut für angewandte Physik
Robert-Mayer-Str. 2-4
D-60325 Frankfurt

Prof. Dr.-Ing.
ANASTASIOS VENETSANOPOULOS
University of Toronto
Dept. of Electrical and Computer Engineering
10 King's College Road
M5S 3G4 Toronto, Ontario
Canada

Editors

Prof. MICHAEL BRANDSTEIN
Harvard University,
Div. of Eng. and Applied Scciences
33 Oxford Street
MA 02138 Cambridge
USA
e-mail: msb@hrl.harvard.edu

Dr. DARREN WARD
Imperial College, Dept. of Electrical Engineering
Exhibition Road
SW7 2AZ London
GB
e-mail: d.ward@ic.ac.uk

ISBN 978-3-642-07547-6 ISBN 978-3-662-04619-7 (eBook)
DOI 10.1007/978-3-662-04619-7

Cip data applied for

This work is subject to copyright. All rights are 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 other ways, 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 Berlin Heidelberg GmbH.
Violations are liable for prosecution act under German Copyright Law.

http://www.springer.de

© Springer-Verlag Berlin Heidelberg 2001
Originally published by Springer-Verlag Berlin Heidelberg New York in 2001
Softcover reprint of the hardcover 1st edition 2001

The use of general descriptive names, registered names, 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.

Typesetting: Camera-ready copy by authors
Cover-Design: de'blik, Berlin
SPIN: 10836055 62/3020 5 4 3 2 1 0 Printed on acid-free paper

Preface

The study and implementation of microphone arrays originated over 20 years ago. Thanks to the research and experimental developments pursued to the present day, the field has matured to the point that array-based technology now has immediate applicability to a number of current systems and a vast potential for the improvement of existing products and the creation of future devices.

In putting this book together, our goal was to provide, for the first time, a single complete reference on microphone arrays. We invited the top researchers in the field to contribute articles addressing their specific topic(s) of study. The reception we received from our colleagues was quite enthusiastic and very encouraging. There was the general consensus that a work of this kind was well overdue. The results provided in this collection cover the current state of the art in microphone array research, development, and technological application.

This text is organized into four sections which roughly follow the major areas of microphone array research today. Parts I and II are primarily theoretical in nature and emphasize the use of microphone arrays for speech enhancement and source localization, respectively. Part III presents a number of specific applications of array-based technology. Part IV addresses some open questions and explores the future of the field.

Part I concerns the problem of enhancing the speech signal acquired by an array of microphones. For a variety of applications, including human-computer interaction and hands-free telephony, the goal is to allow users to roam unfettered in diverse environments while still providing a high quality speech signal and robustness against background noise, interfering sources, and reverberation effects. The use of microphone arrays gives one the opportunity to exploit the fact that the source of the desired speech signal and the noise sources are physically separated in space. Conventional array processing techniques, typically developed for applications such as radar and sonar, were initially applied to the hands-free speech acquisition problem. However, the environment in which microphone arrays is used is significantly different from that of conventional array applications. Firstly, the desired speech signal has an extremely wide bandwidth relative to its center frequency, meaning that conventional narrowband techniques are not suitable. Secondly, there

is significant multipath interference caused by room reverberation. Finally, the speech source and noise signals may located close to the array, meaning that the conventional far-field assumption is typically not valid. These differences (amongst others) have meant that new array techniques have had to be formulated for microphone array applications. Chapter 1 describes the design of an array whose spatial response does not change appreciably over a wide bandwidth. Such a design ensures that the spatial filtering performed by the array is uniform across the entire bandwidth of the speech signal. The main problem with many array designs is that a very large physical array is required to obtain reasonable spatial resolution, especially at low frequencies. This problem is addressed in Chapter 2, which reviews so-called superdirective arrays. These arrays are designed to achieve spatial directivity that is significantly higher than a standard delay-and-sum beamformer. Chapter 3 describes the use of a single-channel noise suppression filter on the output of a microphone array. The design of such a post-filter typically requires information about the correlation of the noise between different microphones. The spatial correlation functions for various directional microphones are investigated in Chapter 4, which also describes the use of these functions in adaptive noise cancellation applications. Chapter 5 reviews adaptive techniques for microphone arrays, focusing on algorithms that are robust and perform well in real environments. Chapter 6 presents optimal spatial filtering algorithms based on the generalized singular-value decomposition. These techniques require a large number of computations, so the chapter presents techniques to reduce the computational complexity and thereby permit real-time implementation. Chapter 7 advocates a new approach that combines explicit modeling of the speech signal (a technique which is well-known in single-channel speech enhancement applications) with the spatial filtering afforded by multi-channel array processing.

Part II is devoted to the source localization problem. The ability to locate and track one or more speech sources is an essential requirement of microphone array systems. For speech enhancement applications, an accurate fix on the primary talker, as well as knowledge of any interfering talkers or coherent noise sources, is necessary to effectively steer the array, enhancing a given source while simultaneously attenuating those deemed undesirable. Location data may be used as a guide for discriminating individual speakers in a multi-source scenario. With this information available, it would then be possible to automatically focus upon and follow a given source on an extended basis. Of particular interest lately, is the application of the speaker location estimates for aiming a camera or series of cameras in a video-conferencing system. In this regard, the automated localization information eliminates the need for a human or number of human camera operators. Several existing commercial products apply microphone-array technology in small-room environments to steer a robotic camera and frame active talkers. Chapter 8 summarizes the various approaches which have been explored to accurately locate an individ-

ual in a practical acoustic environment. The emphasis is on precision in the face of adverse conditions, with an appropriate method presented in detail. Chapter 9 extends the problem to the case of multiple active sources. While again considering realistic environments, the issue is complicated by the presence of several talkers. Chapter 10 further generalizes the source localization scenario to include knowledge derived from non-acoustic sensor modalities. In this case both audio and video signals are effectively combined to track the motion of a talker.

Part III of this text details some specific applications of microphone array technology available today. Microphone arrays have been deployed for a variety of practical applications thus far and their utility and presence in our daily lives is increasing rapidly. At one extreme are large aperture arrays with tens to hundreds of elements designed for large rooms, distant talkers, and adverse acoustic conditions. Examples include the two-dimensional, harmonic array installed in the main auditorium of Bell Laboratories, Murray Hill and the 512-element Huge Microphone Array (HMA) developed at Brown University. While these systems provide tremendous functionality in the environments for which they are intended, small arrays consisting of just a handful (usually 2 to 8) of microphones and encompassing only a few centimeters of space have become far more common and affordable. These systems are intended for sound capture in close-talking, low to moderate noise conditions (such as an individual dictating at a workstation or using a hands-free telephone in an automobile) and have exhibited a degree of effectiveness, especially when compared to their single microphone counterparts. The technology has developed to the point that microphone arrays are now available in off-the-shelf consumer electronic devices available for under $150. Because of their growing popularity and feasibility we have chosen to focus primarily on the issues associated with small-aperture devices. Chapter 11 addresses the incorporation of multiple microphones into hearing aid devices. The ability of beamforming methods to reduce background noise and interference has been shown to dramatically improve the speech understanding of the hearing impaired and to increase their overall satisfaction with the device. Chapter 12 focuses on the case of a simple two-element array combined with postfiltering to achieve noise and echo reduction. The performance of this configuration is analyzed under realistic acoustic conditions and its utility is demonstrated for desktop conferencing and intercom applications. Chapter 13 is concerned with the problem of acoustic feedback inherent in full-duplex communications involving loudspeakers and microphones. Existing single-channel echo cancellation methods are integrated within a beamforming context to achieve enhanced echo suppression. These results are applied to single- and multi-channel conferencing scenarios. Chapter 14 explores the use of microphone arrays for sound capture in automobiles. The issues of noise, interference, and echo cancellation specifically within the car environment are addressed and a particularly effective approach is detailed. Chapter 15 discusses the applica-

tion of microphone arrays to improve the performance of speech recognition systems in adverse conditions. Strategies for effectively coupling the acoustic signal enhancements afforded through beamforming with existing speech recognition techniques are presented. A specific adaptation of a recognizer to function with an array is presented. Finally, Chapter 16 presents an overview of the problem of separating blind mixtures of acoustic signals recorded at a microphone array. This represents a very new application for microphone arrays, and is a technique that is fundamentally different to the spatial filtering approaches detailed in earlier chapters.

In the final section of the book, Part IV presents expert summaries of current open problems in the field, as well as personal views of what the future of microphone array processing might hold. These summaries, presented in Chapters 17 and 18, describe both academically-oriented research problems, as well as industry-focused areas where microphone array research may be headed.

The individual chapters that we selected for the book were designed to be tutorial in nature with a specific emphasis on recent important results. We hope the result is a text that will be of utility to a large audience, from the student or practicing engineer just approaching the field to the advanced researcher with multi-channel signal processing experience.

Cambridge MA, USA
London, UK
January 2001

Michael Brandstein
Darren Ward

Contents

Part I. Speech Enhancement

1 Constant Directivity Beamforming
Darren B. Ward, Rodney A. Kennedy, Robert C. Williamson 3
1.1 Introduction 3
1.2 Problem Formulation 6
1.3 Theoretical Solution 7
1.3.1 Continuous sensor 7
1.3.2 Beam-shaping function 8
1.4 Practical Implementation 9
1.4.1 Dimension-reducing parameterization 9
1.4.2 Reference beam-shaping filter 11
1.4.3 Sensor placement 12
1.4.4 Summary of implementation 12
1.5 Examples 13
1.6 Conclusions 16
References 16

2 Superdirective Microphone Arrays
Joerg Bitzer, K. Uwe Simmer 19
2.1 Introduction 19
2.2 Evaluation of Beamformers 20
2.2.1 Array-Gain 21
2.2.2 Beampattern 22
2.2.3 Directivity 23
2.2.4 Front-to-Back Ratio 24
2.2.5 White Noise Gain 24
2.3 Design of Superdirective Beamformers 24
2.3.1 Delay-and-Sum Beamformer 26
2.3.2 Design for spherical isotropic noise 26
2.3.3 Design for Cylindrical Isotropic Noise 30
2.3.4 Design for an Optimal Front-to-Back Ratio 30
2.3.5 Design for Measured Noise Fields 32
2.4 Extensions and Details 33
2.4.1 Alternative Form 33

2.4.2 Comparison with Gradient Microphones 35
2.5 Conclusion 36
References 37

3 Post-Filtering Techniques
K. Uwe Simmer, Joerg Bitzer, Claude Marro 39
3.1 Introduction 39
3.2 Multi-channel Wiener Filtering in Subbands 41
3.2.1 Derivation of the Optimum Solution 41
3.2.2 Factorization of the Wiener Solution 42
3.2.3 Interpretation 45
3.3 Algorithms for Post-Filter Estimation 46
3.3.1 Analysis of Post-Filter Algorithms 47
3.3.2 Properties of Post-Filter Algorithms 49
3.3.3 A New Post-Filter Algorithm 50
3.4 Performance Evaluation 51
3.4.1 Simulation System 52
3.4.2 Objective Measures 52
3.4.3 Simulation Results 54
3.5 Conclusion 57

4 Spatial Coherence Functions for Differential Microphones in Isotropic Noise Fields
Gary W. Elko 61
4.1 Introduction 61
4.2 Adaptive Noise Cancellation 61
4.3 Spherically Isotropic Coherence 65
4.4 Cylindrically Isotropic Fields 73
4.5 Conclusions 77
References 84

5 Robust Adaptive Beamforming
Osamu Hoshuyama, Akihiko Sugiyama 87
5.1 Introduction 87
5.2 Adaptive Beamformers 88
5.3 Robustness Problem in the GJBF 90
5.4 Robust Adaptive Microphone Arrays — Solutions to Steering-Vector Errors 92
5.4.1 LAF-LAF Structure 92
5.4.2 CCAF-LAF Structure 94
5.4.3 CCAF-NCAF Structure 95
5.4.4 CCAF-NCAF Structure with an AMC 97
5.5 Software Evaluation of a Robust Adaptive Microphone Array 99
5.5.1 Simulated Anechoic Environment 99
5.5.2 Reverberant Environment 101

5.6 Hardware Evaluation of a Robust Adaptive Microphone Array 104
5.6.1 Implementation 104
5.6.2 Evaluation in a Real Environment 104
5.7 Conclusion 106
References 106

6 GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement
Simon Doclo, Marc Moonen 111
6.1 Introduction 111
6.2 GSVD-Based Optimal Filtering Technique 113
6.2.1 Optimal Filter Theory 114
6.2.2 General Class of Estimators 116
6.2.3 Symmetry Properties for Time-Series Filtering 117
6.3 Performance of GSVD-Based Optimal Filtering 118
6.3.1 Simulation Environment 118
6.3.2 Spatial Directivity Pattern 119
6.3.3 Noise Reduction Performance 121
6.3.4 Robustness Issues 121
6.4 Complexity Reduction 122
6.4.1 Linear Algebra Techniques for Computing GSVD 122
6.4.2 Recursive and Approximate GSVD-Updating Algorithms 123
6.4.3 Downsampling Techniques 125
6.4.4 Simulations 125
6.4.5 Computational Complexity 126
6.5 Combination with ANC Postprocessing Stage 127
6.5.1 Creation of Speech and Noise References 127
6.5.2 Noise Reduction Performance of ANC Postprocessing Stage . 128
6.5.3 Comparison with Standard Beamforming Techniques 129
6.6 Conclusion 129
References 130

7 Explicit Speech Modeling for Microphone Array Speech Acquisition
Michael Brandstein, Scott Griebel 133
7.1 Introduction 133
7.2 Model-Based Strategies 136
7.2.1 Example 1: A Frequency-Domain Model-Based Algorithm .. 137
7.2.2 Example 2: A Time-Domain Model-Based Algorithm 140
7.3 Conclusion 148
References 151

Part II. Source Localization

8 Robust Localization in Reverberant Rooms
Joseph H. DiBiase, Harvey F. Silverman, Michael S. Brandstein 157
8.1 Introduction 157
8.2 Source Localization Strategies 158
8.2.1 Steered-Beamformer-Based Locators 159
8.2.2 High-Resolution Spectral-Estimation-Based Locators 160
8.2.3 TDOA-Based Locators 161
8.3 A Robust Localization Algorithm 164
8.3.1 The Impulse Response Model 164
8.3.2 The GCC and PHAT Weighting Function 166
8.3.3 ML TDOA-Based Source Localization 167
8.3.4 SRP-Based Source Localization 169
8.3.5 The SRP-PHAT Algorithm 170
8.4 Experimental Comparison 172
References 178

9 Multi-Source Localization Strategies
Elio D. Di Claudio, Raffaele Parisi 181
9.1 Introduction 181
9.2 Background 184
9.2.1 Array Signal Model 184
9.2.2 Incoherent Approach 185
9.2.3 Coherent Signal Subspace Method (CSSM) 185
9.2.4 Wideband Weighted Subspace Fitting (WB-WSF) 186
9.3 The Issue of Coherent Multipath in Array Processing 187
9.4 Implementation Issues 188
9.5 Linear Prediction-ROOT-MUSIC TDOA Estimation 189
9.5.1 Signal Pre-Whitening 189
9.5.2 An Approximate Model for Multiple Sources in Reverberant Environments 191
9.5.3 Robust TDOA Estimation via ROOT-MUSIC 192
9.5.4 Estimation of the Number of Relevant Reflections 194
9.5.5 Source Clustering 195
9.5.6 Experimental Results 196
References 198

10 Joint Audio-Video Signal Processing for Object Localization and Tracking
Norbert Strobel, Sascha Spors, Rudolf Rabenstein 203
10.1 Introduction 203
10.2 Recursive State Estimation 205
10.2.1 Linear Kalman Filter 206
10.2.2 Extended Kalman Filter due to a Measurement Nonlinearity 210
10.2.3 Decentralized Kalman Filter 212
10.3 Implementation 218

10.3.1 System description 218
10.3.2 Results 219
10.4 Discussion and Conclusions 221
References 222

Part III. Applications

11 Microphone-Array Hearing Aids
Julie E. Greenberg, Patrick M. Zurek 229
11.1 Introduction 229
11.2 Implications for Design and Evaluation 230
11.2.1 Assumptions Regarding Sound Sources 230
11.2.2 Implementation Issues 231
11.2.3 Assessing Performance 232
11.3 Hearing Aids with Directional Microphones 233
11.4 Fixed-Beamforming Hearing Aids 234
11.5 Adaptive-Beamforming Hearing Aids 235
11.5.1 Generalized Sidelobe Canceler with Modifications 236
11.5.2 Scaled Projection Algorithm 242
11.5.3 Direction of Arrival Estimation 243
11.5.4 Other Adaptive Approaches and Devices 243
11.6 Physiologically-Motivated Algorithms 244
11.7 Beamformers with Binaural Outputs 245
11.8 Discussion 246
References 249

12 Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction
Rainer Martin 255
12.1 Introduction 255
12.2 Coherence of Speech and Noise 257
12.2.1 The Magnitude Squared Coherence 257
12.2.2 The Reverberation Distance 258
12.2.3 Coherence of Noise and Speech in Reverberant Enclosures .. 259
12.3 Analysis of the Wiener Filter with Symmetric Input Signals 263
12.3.1 No Near End Speech 265
12.3.2 High Signal to Noise Ratio 265
12.4 A Noise Reduction Application 266
12.4.1 An Implementation Based on the NLMS Algorithm 266
12.4.2 Processing in the 800 – 3600 Hz Band 268
12.4.3 Processing in the 240 – 800 Hz Band 269
12.4.4 Evaluation 269
12.4.5 Alternative Implementations of the Coherence Based Postfilter 271
12.5 Combined Noise and Acoustic Echo Reduction 271

12.5.1 Experimental Results 274
12.6 Conclusions 275
References 276

13 Acoustic Echo Cancellation for Beamforming Microphone Arrays
Walter L. Kellermann 281
13.1 Introduction 281
13.2 Acoustic Echo Cancellation 282
13.2.1 Adaptation algorithms 284
13.2.2 AEC for multi-channel sound reproduction 287
13.2.3 AEC for multi-channel acquisition 287
13.3 Beamforming 288
13.3.1 General structure 288
13.3.2 Time-invariant beamforming 290
13.3.3 Time-varying beamforming 291
13.3.4 Computational complexity 292
13.4 Generic structures for combining AEC with beamforming 292
13.4.1 Motivation 292
13.4.2 Basic options 293
13.4.3 'AEC first' 293
13.4.4 'Beamforming first' 296
13.5 Integration of AEC into time-varying beamforming 297
13.5.1 Cascading time-invariant and time-varying beamforming 297
13.5.2 AEC with GSC-type beamforming structures 301
13.6 Combined AEC and beamforming for multi-channel recording and multi-channel reproduction 302
13.7 Conclusions 303
References 303

14 Optimal and Adaptive Microphone Arrays for Speech Input in Automobiles
Sven Nordholm, Ingvar Claesson, Nedelko Grbić 307
14.1 Introduction: Hands-Free Telephony in Cars 307
14.2 Optimum and Adaptive Beamforming 309
14.2.1 Common Signal Modeling 309
14.2.2 Constrained Minimum Variance Beamforming and the Generalized Sidelobe Canceler 310
14.2.3 *In Situ* Calibrated Microphone Array (ICMA) 312
14.2.4 Time-Domain Minimum-Mean-Square-Error Solution 313
14.2.5 Frequency-Domain Minimum-Mean-Square-Error Solution 314
14.2.6 Optimal Near-Field Signal-to-Noise plus Interference Beamformer 316
14.3 Subband Implementation of the Microphone Array 317
14.3.1 Description of LS-Subband Beamforming 318

14.4 Multi-Resolution Time-Frequency Adaptive Beamforming 319
14.4.1 Memory Saving and Improvements 319
14.5 Evaluation and Examples 320
14.5.1 Car Environment 320
14.5.2 Microphone Configurations 321
14.5.3 Performance Measures 321
14.5.4 Spectral Performance Measures 322
14.5.5 Evaluation on car data 323
14.5.6 Evaluation Results 323
14.6 Summary and Conclusions 324
References 326

15 Speech Recognition with Microphone Arrays
Maurizio Omologo, Marco Matassoni, Piergiorgio Svaizer 331
15.1 Introduction 331
15.2 State of the Art 332
15.2.1 Automatic Speech Recognition 332
15.2.2 Robustness in ASR 336
15.2.3 Microphone Arrays and Related Processing for ASR 337
15.2.4 Distant-Talker Speech Recognition 339
15.3 A Microphone Array-Based ASR System 342
15.3.1 System Description 342
15.3.2 Speech Corpora and Task 345
15.3.3 Experiments and Results 346
15.4 Discussion and Future Trends 348
References 349

16 Blind Separation of Acoustic Signals
Scott C. Douglas 355
16.1 Introduction 355
16.1.1 The Cocktail Party Effect 355
16.1.2 Chapter Overview 356
16.2 Blind Signal Separation of Convolutive Mixtures 357
16.2.1 Problem Structure 357
16.2.2 Goal of Convolutive BSS 359
16.2.3 Relationship to Other Problems 360
16.3 Criteria for Blind Signal Separation 362
16.3.1 Overview of BSS Criteria 362
16.3.2 Density Modeling Criteria 362
16.3.3 Contrast Functions 364
16.3.4 Correlation-Based Criteria 366
16.4 Structures and Algorithms for Blind Signal Separation 367
16.4.1 Filter Structures 367
16.4.2 Density Matching BSS Using Natural Gradient Adaptation . 368
16.4.3 Contrast-Based BSS Under Prewhitening Constraints 370

16.4.4 Temporal Decorrelation BSS for Nonstationary Sources 372
16.5 Numerical Evaluations 373
16.6 Conclusions and Open Issues 375
References 378

Part IV. Open Problems and Future Directions

17 Future Directions for Microphone Arrays
Gary W. Elko 383
17.1 Introduction 383
17.2 Hands-Free Communication 383
17.3 The "Future" of Microphone Array Processing 385
17.4 Conclusions 387

18 Future Directions in Microphone Array Processing
Dirk Van Compernolle 389
18.1 Lessons From the Past 389
18.2 A Future Focused on Applications 391
18.2.1 Automotive 391
18.2.2 Desktop 392
18.2.3 Hearing Aids 393
18.2.4 Teleconferencing 393
18.2.5 Very Large Arrays 393
18.2.6 The Signal Subspace Approach - An Alternative to Spatial Filtering ? 393
18.3 Final Remarks 394

Index 395

List of Contributors

Joerg Bitzer
Houpert Digital Audio
Bremen, Germany

Michael S. Brandstein
Harvard Universtiy
Cambridge MA, USA

Ingvar Claesson
Blekinge Inst. of Technology
Ronneby, Sweden

Joseph H. DiBiase
Brown Universtiy
Providence RI, USA

Elio D. Di Claudio
University of Rome "La Sapienza"
Rome, Italy

Simon Doclo
Katholieke Universiteit Leuven
Leuven, Belgium

Scott C. Douglas
Southern Methodist University
Dallas TX, USA

Gary W. Elko
Agere Systems
Murray Hill NJ, USA

Nedelko Grbić
Blekinge Inst. of Technology
Ronneby, Sweden

Julie E. Greenberg
Massachusetts Inst. of Technology
Cambridge MA, USA

Scott M. Griebel
Harvard Universtiy
Cambridge MA, USA

Osamu Hoshuyama
NEC Media Research Labs
Kawasaki, Japan

Walter L. Kellermann
University Erlangen-Nuremberg
Erlangen, Germany

Rodney A. Kennedy
The Australian National University
Canberra, Australia

Claude Marro
France Télécom R&D
Lannion, France

Rainer Martin
Aachen University of Technology
Aachen, Germany

Marco Matassoni
Istituto per la Ricerca Scientifica e Tecnologica
Povo, Italy

Marc Moonen
Katholieke Universiteit Leuven
Leuven, Belgium

Sven Nordholm
Curtin University of Technology
Perth, Australia

Maurizio Omologo
Istituto per la Ricerca Scientifica e Tecnologica
Povo, Italy

Raffaele Parisi
University of Rome "La Sapienza"
Rome, Italy

Rudolf Rabenstein
University Erlangen-Nuremberg
Erlangen, Germany

Harvey F. Silverman
Brown Universtiy
Providence RI, USA

K. Uwe Simmer
Aureca GmbH
Bremen, Germany

Sascha Spors
University Erlangen-Nuremberg
Erlangen, Germany

Norbert Strobel
Siemens Medical Solutions
Erlangen, Germany

Akihiko Sugiyama
NEC Media Research Labs
Kawasaki, Japan

Piergiorgio Svaizer
Istituto per la Ricerca Scientifica e Tecnologica
Povo, Italy

Dirk Van Compernolle
Katholieke Universiteit Leuven
Leuven, Belgium

Darren B. Ward
Imperial College of Science, Technology and Medicine
London, UK

Robert C. Williamson
The Australian National University
Canberra, Australia

Patrick M. Zurek
Sensimetrics Corporation
Somerville MA, USA

Part I

Speech Enhancement

1 Constant Directivity Beamforming

Darren B. Ward[1], Rodney A. Kennedy[2], and Robert C. Williamson[2]

[1] Imperial College of Science, Technology and Medicine, London, UK
[2] The Australian National University, Canberra, Australia

Abstract. Beamforming, or spatial filtering, is one of the simplest methods for discriminating between different signals based on the physical location of the sources. Because speech is a very wideband signal, covering some four octaves, traditional narrowband beamforming techniques are inappropriate for hands-free speech acquisition. One class of broadband beamformers, called constant directivity beamformers, aim to produce a constant spatial response over a broad frequency range. In this chapter we review such beamformers, and discuss implementation issues related to their use in microphone arrays.

1.1 Introduction

Beamforming is one of the simplest and most robust means of *spatial filtering*, i.e., discriminating between signals based on the physical locations of the signal sources [1]. In a typical microphone array environment, the desired speech signal originates from a talker's mouth, and is corrupted by interfering signals such as other talkers and room reverberation. Spatial filtering can be useful in such an environment, since the interfering sources generally originate from points in space separate from the desired talker's mouth. By exploiting the spatial dimension of the problem, microphone arrays attempt to obtain a high-quality speech signal without requiring the talker to speak directly into a close-talking microphone.

In most beamforming applications two assumptions simplify the analysis: (i) the signals incident on the array are narrowband (the *narrowband assumption*); and (ii) the signal sources are located far enough away from the array that the wavefronts impinging on the array can be modeled as plane waves (the *farfield assumption*). For many microphone array applications, the farfield assumption is valid. However, the narrowband assumption is never valid, and it is this aspect of the beamforming problem that we focus on in this chapter (see [2] for techniques that also lift the nearfield assumption).

To understand the inherent problem in using a narrowband array for broadband signals, consider a linear array with a fixed number of elements separated by a fixed inter-element distance. The important dimension in measuring array performance is its size in terms of operating wavelength. Thus for high frequency signals (having a small wavelength) a fixed array will appear large and the main beam will be narrow. However, for low frequencies

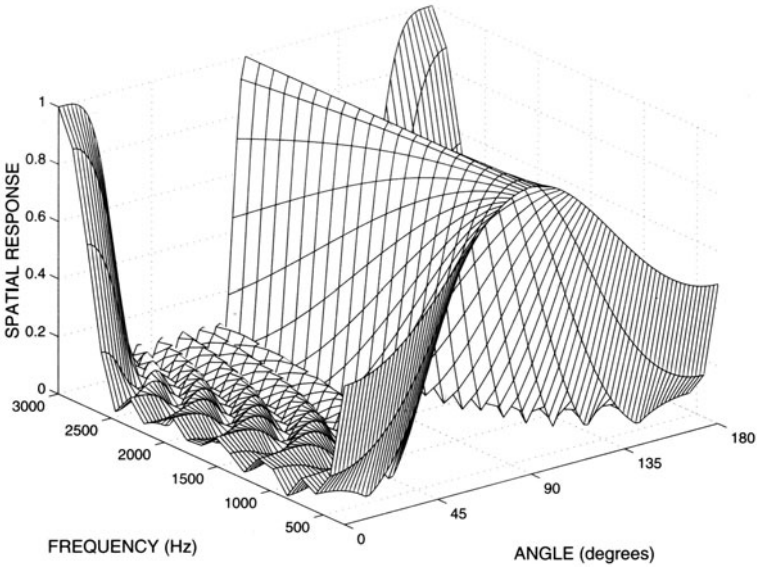


Fig. 1.1. Response of a narrowband array operated over a wide bandwidth.

(large wavelength) the same physical array appears small and the main beam will widen.

This is illustrated in Fig. 1.1 which shows the beampattern of an array designed for 1.5 kHz, but operated over a frequency range of 300 Hz to 3 kHz. If an interfering signal is present at, say, 60°, then ideally it should be attenuated completely by the array. However, because the beam is wider at low frequencies than at high frequencies, the interfering signal will be low-pass filtered rather than uniformly attenuated over its entire band. This "spectral tilt" results in a disturbing speech output if used for speech acquisition, and thus, such a narrowband array is unacceptable for speech applications. Another drawback of this narrowband design is that spatial aliasing is evident at high frequencies.[1]

To overcome this problem, one must use a beamformer that is designed specifically for broadband applications. In this chapter we focus on a specific class of broadband beamformers, called *constant directivity beamformers* (CDB), designed such that the spatial response is the same over a wide frequency band. The response of a typical CDB is shown in Fig. 1.6 on page 15.

There have been several techniques proposed to design a CDB. Most techniques are based on the idea that at different frequencies, a different array should be used that has total size and inter-sensor spacing appropriate for that particular frequency. An example of this idea is the use of harmonically-

[1] Spatial aliasing comes about if a sensor spacing wider than half a wavelength is used. It is analogous to temporal aliasing in discrete-time signal processing.

nested subarrays, e.g., [3–5]. In this case, the array is composed of a set of nested equally-spaced arrays, with each subarray being designed as a narrowband array. The outputs of the various subarrays are then combined by appropriate bandpass filtering. The idea of harmonic nesting is to reduce the beampattern variation to that which occurs within a single octave. This approach can be improved by using a set of subarray filters to interpolate to frequencies between the subarray design frequencies [6].

A novel approach to CDB design was proposed by Smith in [7]. Noting that, for a given array, the beamwidth narrows at high frequencies, Smith's idea was to form several beams and to steer each individual beam in such a way that the width of the overall multi-beam was kept constant. Thus, as the individual beams narrow at higher frequencies, they are progressively "fanned" outwards in an attempt to keep the overall beamwidth constant. Unless a very large number of beams are formed, at high frequencies this fanning will result in notches in the main beam where the progressively narrower beams no longer overlap. This approach was applied to the design of microphone arrays in [8].

The first approach to CDB design that attempted to keep a constant beampattern over the entire spatial region (not just for the main beam) was presented by Doles and Benedict [9]. Using the asymptotic theory of unequally-spaced arrays [10,11], they derived relationships between beampattern characteristics and functional requirements on sensor spacings and weightings. This results in a filter-and-sum array, with the sensor filters creating a space-tapered array: at each frequency the non-zero filter responses identify a subarray having total length and spacing appropriate for that frequency. Although this design technique results in a beampattern that is frequency-invariant over a specified frequency band, it is not a general design technique, since it is based on a specific array geometry and beampattern shape. Other recent techniques for CDB design include [12] (based on a two-dimensional Fourier transform property [13] which exists for equally-spaced arrays) and [14] (based on a beam space implementation).

Prompted by the work of Doles and Benedict, we derived in [15] a very general design method for CDB's, suitable for three-dimensional array geometries. In this chapter we outline this technique, and discuss implementation issues specific to microphone array applications.

Time-domain versus frequency-domain beamforming

There are two general methods of beamforming for broadband signals: time-domain beamforming and frequency-domain beamforming. In time-domain beamforming an FIR filter is used on each sensor, and the filter outputs summed to form the beamformer output. For an array with M sensors, each feeding a L tap filter, there are ML free parameters. In frequency-domain beamforming the signal received by each sensor is separated into narrowband frequency bins (either through bandpass filtering or data segmentation

and discrete Fourier transform), and the data in each frequency bin is processed separately using narrowband techniques. For an array with M sensors, with L frequency bins within the band of interest, there are again ML free parameters. As with most beamformers, the method that we describe in this chapter can be formulated in either domain. A time-domain formulation has previously been given in [16], and hence, we restrict our attention to frequency-domain processing here.

1.2 Problem Formulation

Consider a linear array of $M = 2N+1$ sensors located at $p_n, n = -N, \ldots, N$. Assume that the data received at the nth sensor is separated into narrowband frequency bins, each of width Δf. Let the center frequency of the ith bin be f_i, and denote the frequencies within the bin as

$$F_i = [f_i - \Delta f/2, f_i + \Delta f/2).$$

The array data received in the ith bin at time k, is given by the M-vector:

$$\boldsymbol{x}_i(k) = \boldsymbol{a}(\theta, f_i)s_i(k) + \boldsymbol{v}_i(k).$$

The desired source signal is represented by $s_i(k)$, and the M-vector $\boldsymbol{v}_i(k)$ represents the interfering noise (consisting of reverberation and other unwanted noise sources). The array vector $\boldsymbol{a}(\theta, f)$ represents the propagation of the signal source to the array, and its nth element is given by

$$a_n(\theta, f) = e^{-j2\pi f c^{-1} p_n \cos\theta},$$

where c is the speed of wave propagation, and θ is the direction to the desired source (measured relative to the array axis). To simplify notation we will drop the explicit dependence on k in the sequel.

The beamformer output is formed by applying a weight vector to the received array data, giving

$$y_i = \boldsymbol{w}_i^H \boldsymbol{x}_i, \tag{1.1}$$

where ^{H} denotes Hermitian transpose, and $\boldsymbol{w}_i$ is the M-vector of array weights to apply to the ith frequency bin.[2]

The spatial response of the beamformer is given by

$$b(\theta, f) = \boldsymbol{w}_i^H \boldsymbol{a}(\theta, f), \quad f \in F_i, \tag{1.2}$$

which defines the transfer function between a source at location $\theta \in [-\pi, \pi)$ and the beamformer output. Also of interest is the *beampattern*, defined as the squared magnitude of the spatial response.

[2] Note that it is a notational convention to use $\boldsymbol{w}^H$ rather than $\boldsymbol{w}^T$ [1].

The problem of designing a CDB can now be formulated as finding the array weights in each frequency bin such that the resulting spatial response remains constant over all frequency bins of interest.

One simple (but not very illuminating) approach to solving this problem is to perform a least-squares optimization in each frequency bin, i.e.,

$$\min_{\boldsymbol{w}_i} \int_{2\pi} |b_{\text{FI}}(\theta) - \boldsymbol{w}_i^H \boldsymbol{a}(\theta, f_i)|^2 \; d\theta, \tag{1.3}$$

where $b_{\text{FI}}(\theta)$ is the desired frequency-invariant response. Thus, in each frequency bin there are M free parameters to optimize. Although this is a standard least-squares optimization problem and the required array weights are easily found, the solution provides very little insight into the problem. Specifically, there is no suggestion of any inherent structure in the CDB, and many important questions are left unanswered, such as how many sensors are required, and what range of frequencies can be used.

In an attempt to provide some insight into the problem of designing a CDB, we take an alternative theoretical approach in the following section, and then relate these theoretical results back to the problem of finding the required filter coefficients. As we will see, there is in fact a very strong implicit structure in the CDB, and exploiting this structure enables us to reduce the number of design parameters and find efficient implementations.

1.3 Theoretical Solution

It is well known that the important dimension in determining the array response is the physical array size, measured in wavelengths. Thus, to obtain the same beampattern at different frequencies requires that the array size remains constant in terms of wavelength. Specifically, consider a linear array with N elements located at $p_n, n = 1, \ldots, N$, and assume the array weights are chosen to produce a desired beampattern $b(\theta)$ at a frequency f_1. Then, at a frequency f_2, the same beampattern $b(\theta)$ will be produced if the same array weights are used in an array with elements located at $p_n(f_1/f_2), n = 1, \ldots, N$. In other words, the size of the array must scale directly with frequency to obtain the same beampattern.[3] To obtain the same beampattern over a continuous range of frequencies would theoretically require a continuum of sensors.

1.3.1 Continuous sensor

Motivated by this interpretation, we consider the response of a theoretical continuous sensor. Assume that a signal $x(p, f)$ is received at a point p on

[3] This is precisely the idea used in the harmonically-nested subarray technique.

the sensor at frequency f, and a weight $w(p, f)$ is applied to the sensor at this point and frequency. The output of the sensor is

$$y(f) = \int w(p, f)\, x(p, f)\, dp,$$

and the spatial response for a source at angle θ is

$$b(\theta, f) = \int w(p, f)\, e^{-j2\pi f c^{-1} p \cos\theta}\, dp. \tag{1.4}$$

We assume that the aperture has finite support in p, and thus, the integration has infinite limits.

Let $u = c^{-1}\cos\theta$. The response of the continuous sensor can now be written

$$b_u(u, f) = \int w(p, f)\, e^{-j2\pi f p u}\, dp.$$

Let the sensor weighting function be given by

$$w(p, f) = fB(pf), \tag{1.5}$$

where $B(\cdot)$ is an arbitrary, absolutely-integrable, finite-support function. Substitution gives

$$b_u(u, f) = \int fB(pf)\, e^{-j2\pi f p u}\, dp. \tag{1.6}$$

With the change of variable $\zeta = pf$, and noting that $d\zeta = f dp$, it is easily seen that the resulting spatial response is now independent of frequency, i.e.,

$$b_u(u, f) = \int B(\zeta)\, e^{-j2\pi\zeta u}\, d\zeta = b_{\text{FI}}(u). \tag{1.7}$$

This is an important result, since it states that if the weighting function is given by (1.5), then the resulting spatial response will be independent of frequency. In other words, (1.5) defines the weighting function for a CDB. It was shown in [15], that not only does (1.5) provide a sufficient condition, but it is in fact the necessary condition for a frequency-invariant spatial response.

1.3.2 Beam-shaping function

Equation (1.7) defines a Fourier transform relationship between $B(\cdot)$ and $b_{\text{FI}}(\cdot)$. To achieve some desired spatial response, the required function $B(\zeta)$ is thus easily found by taking the inverse Fourier transform of $b(u)$. We will refer to $B(\cdot)$ as the *beam-shaping* (BS) function, since it has a fundamental role in determining the spatial response.

Because of its symmetry with respect to space and frequency, the BS function can be interpreted as either a filter response at a certain point, i.e., $H_p(f) = B(pf)$, or equivalently, as an aperture weighting function at a certain frequency, i.e., $A_f(p) = B(pf)$.

We will assume that the BS function is Hermitian symmetric, i.e., $B(-\zeta) = B^*(\zeta)$. This implies that the resulting spatial response is real-valued.

1.4 Practical Implementation

Whilst we have shown theoretically that it is possible to produce a beampattern that is exactly frequency-invariant using a continuous sensor, in practise we must attempt to approximate such a response using a finite array of discrete sensors. The problem of approximating a continuous aperture by a discrete array has been considered in [17]. One simple but effective technique is to approximate the integral in (1.6) using a Riemann sum—this is the approach we take here. In particular, we use trapezoidal integration to approximate the integral (1.6) by a summation of the form:

$$\hat{b}_{\mathrm{FI}}(u) = \sum_{n=-N}^{N} f\, B(p_n f)\, e^{-j2\pi f p_n u}\, \Delta_n \tag{1.8}$$

where p_n is the location of the nth discrete sensor, and $\hat{b}_{\mathrm{FI}}$ denotes an approximation of b_{FI}. We assume that the array is Hermitian symmetric about the origin, so that $B(-pf) = B(pf)^*$, and $p_{-n} = -p_n$. Although the technique is suitable for an arbitrary array geometry, a symmetric geometry simplifies implementation, and ensures that the position of the array phase center does not vary with frequency. The length of the nth subinterval is

$$\Delta_n = \frac{p_{n+1} - p_{n-1}}{2}, \tag{1.9}$$

which we refer to as the *spatial weighting term.*

Relating (1.8) to the response of a general array (1.2), we find that for a CDB the weight on the nth sensor in the ith frequency bin is

$$w_{i,n} = f_i\, \Delta_n\, B(p_n f_i), \tag{1.10}$$

where, recall, p_n is the location of the sensor, and f_i is the center frequency of the bin.

1.4.1 Dimension-reducing parameterization

Define the *reference beam-shaping filter response* as

$$H(f) = B(p_{\mathrm{ref}} f), \tag{1.11}$$

where p_{ref} is some reference location (to be defined later). Also define the *beam-shaping filter response* of the nth sensor as

$$H_n(f) = B(p_n f), \quad n = -N, \ldots, N.$$

It immediately follows that the BS filters satisfy the following dilation property:

$$H_n(f) = H(\gamma_n f), \tag{1.12}$$

where

$$\gamma_n = \frac{p_n}{p_{\text{ref}}}$$

is the dilation factor for the nth sensor. This is an extremely important property, since it shows that the filter responses on all sensors can be derived from the single filter response, $H(f)$, and enables the following efficient implementation of the CDB.

Let the reference BS filter response be given by its standard FIR filter representation:

$$H(f) = \sum_l h[l]\, e^{-j2\pi f/f_s l},$$

where f_s is the sampling frequency, and $h[l]$ is a L-vector of *beam-shaping coefficients*. From (1.12), the nth BS filter response is given by

$$\begin{aligned} H_n(f) &= \sum_l h[l]\, e^{-j2\pi f/f_s \gamma_n l} \\ &= \boldsymbol{h}^H \boldsymbol{d}_n(f), \end{aligned} \tag{1.13}$$

where $\boldsymbol{d}_n(f)$ is the L-dimensional *dilation vector* for the nth sensor. From (1.10), we see that the weight to use on the nth sensor in the ith bin is

$$w_{i,n} = \boldsymbol{h}^H \boldsymbol{t}_{i,n}, \tag{1.14}$$

where

$$\boldsymbol{t}_{i,n} = f_i \Delta_n \boldsymbol{d}_n(f_i) \tag{1.15}$$

is a L-dimensional *transformation vector*.

Equation (1.14) demonstrates the efficient parameterization afforded by this particular formulation of the CDB problem. Whereas the naive least-squares approach (1.3) requires an optimization of M parameters $\boldsymbol{w}_i$ in each frequency bin, we find that it is really only necessary to choose L frequency-independent BS parameters $\boldsymbol{h}$. Changing the beampattern shape only requires modification of these BS coefficients, and the implicit structure imposed by the transformation vectors ensures that the resulting response has constant directivity over the design band.

1.4.2 Reference beam-shaping filter

The underlying principle of the CDB is that the size and shape of the active array aperture should scale directly with frequency. This frequency scaling operation is performed by the BS filters. In deciding the coefficients of the reference BS filter, and the location of the reference point p_{ref}, we must consider this scaling property in more detail.

Let the chosen aperture size be Q wavelengths. Assuming the array is symmetric about the origin, this means that at any wavelength λ, sensors further from the origin than $Q\lambda/2$ should be inactive. In other words, the nth sensor should have a low-pass characteristic with a cutoff frequency of

$$f_n = \frac{Qc}{2|p_n|}. \tag{1.16}$$

From (1.13), note that $\gamma_n > 1$ results in compression in the frequency domain, whereas $\gamma_n < 1$ results in frequency expansion. Since the discrete-time frequency response $H(f)$ is periodic, it follows that frequency compression may cause aliasing; this is extremely undesirable. Aliasing can be avoided in one of two ways. First, choosing $p_{\text{ref}} = \max|p_n|$ ensures that $\gamma_n \leq 1, \forall n$, thus avoiding aliasing altogether—however, this requires additional constraints on the reference BS coefficients to impose the low-pass property (1.16). Alternatively, for sensors having $\gamma_n > 1$, the weights $w_{i,n}$ are set to zero for frequency bins $f_i > f_n$—the reference BS weights are now potentially unconstrained. Of these two approaches, the second is preferable, since it removes any constraints on the BS coefficients. Moreover, the requirement that the sensor weights within certain bins are always zero does not complicate implementation.

Assume that the frequency response of the reference BS filter is non-zero for all frequencies up to $f_s/2$, the Nyquist frequency; this is the most general case of $H(f)$. From (1.16), it follows that a sensor with non-zero frequency response up to $f_s/2$ would be positioned at $|p_n| = Qc/f_s$. Thus, for the most general case of $H(f)$ the reference location is chosen as

$$p_{\text{ref}} = \frac{Qc}{f_s}. \tag{1.17}$$

The reference BS coefficients can be found by using the Fourier transform relationship defined by (1.7). Specifically, the BS function $B(\zeta)$ is found by taking the Fourier transform of the desired frequency-invariant spatial response $b_{\text{FI}}(u)$. Setting $f = \zeta/p_{\text{ref}}$, $B(\zeta)$ now defines the frequency response of the reference BS filter. The BS coefficient vector $\boldsymbol{h}$ is found using any standard FIR filter design technique. In practise, low-order implementations of the reference BS filter are generally to be preferred; this point is demonstrated in the following section.

1.4.3 Sensor placement

The most common geometry for array processing applications is typically an equally-spaced array, usually with a spacing of one half-wavelength at the highest frequency of operation. Although such a geometry is valid for a CDB, less sensors are required if a logarithmically spaced array is used. In choosing an appropriate sensor geometry, the most important consideration is to ensure that at any frequency spatial aliasing is avoided.

The idea is to start with an equally-spaced array that is used at the highest frequency, and then progressively add more sensors with wider spacings as frequency decreases (and the wavelength increases). At any frequency f, the total active aperture size should be Qc/f, and the largest spacing within the active array should be $c/(2f)$. These requirements are met (using the least number of sensors) with the following symmetric array geometry:

$$p_n = n\,\frac{c}{2f_U}, \quad 0 \le n \le \frac{Q}{2} \tag{1.18a}$$

$$p_{n+1} = \frac{Q}{Q-1}\,p_n, \quad n > \frac{Q}{2},\ p_n < \frac{(Q-1)c}{2f_L} \tag{1.18b}$$

$$p_{-n} = -p_n. \tag{1.18c}$$

Note that a harmonically-nested subarray geometry is only produced if $Q = 2$.

1.4.4 Summary of implementation

1. Choose a set of L reference BS coefficients, $\boldsymbol{h}$.
2. Position the sensors according to (1.18a)–(1.18c).
3. In the ith frequency bin, the weight on the nth sensor is

$$w_{i,n} = \boldsymbol{h}^H \boldsymbol{t}_{i,n},$$

where

$$\boldsymbol{t}_{i,n} = \begin{cases} f_i \Delta_n \boldsymbol{d}_n(f_i), & f_i < f_n \\ \boldsymbol{0}, & \text{otherwise,} \end{cases}$$

$$f_n = \frac{Qc}{2|p_n|}$$

$$\Delta_n = \frac{p_{n+1} - p_{n-1}}{2}$$

$$\boldsymbol{d}_n(f_i) = \left[e^{j2\pi f/f_s\gamma_n(L-1)/2}, \ldots, e^{-j2\pi f/f_s\gamma_n(L-1)/2}\right]$$

$$\gamma_n = \frac{|p_n|}{p_{\text{ref}}}$$

$$p_{\text{ref}} = \frac{Qc}{f_s}$$

1.5 Examples

We now show an example of the CDB design technique. The design was for a bandwidth of 300–3000 Hz (i.e., the same bandwidth as used in Fig. 1.1), with an aperture size of $Q = 4$ wavelengths. Using an FFT size of 128 resulted in 44 bins within the design band, with each bin having a width of 62.5 Hz. The sensors were positioned according to (1.18a)–(1.18c), resulting in the $M = 25$ sensor array geometry shown in Fig. 1.2. For frequencies of 1000 Hz and 2000 Hz, the active sensors are also indicated in this figure.

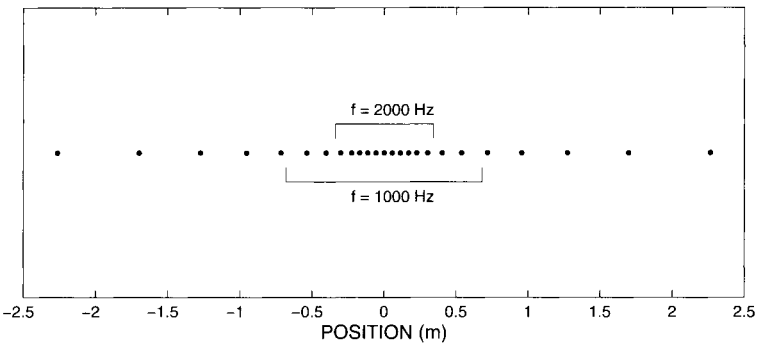


Fig. 1.2. Array geometry used for example CDB.

Assume we wish to design a standard sinc-like response (as produced by a uniformly weighted array). In this case it is known that the aperture function should be uniform. Thus, the BS function $B(\cdot)$ should ideally be a brick-wall low-pass filter. Assume we design the BS vector $\boldsymbol{h}$ to approximate an ideal low-pass filter using $L = 101$ filter coefficients. This results in the BS frequency responses shown in Fig. 1.3; for each sensor in the array, the weight required at each frequency is plotted. Note that these responses are all dilations of a single response, and that each has a low-pass characteristic.

Using these BS coefficients, the resulting spatial response of the CDB is shown in Fig. 1.4. Although the variation is not as great as for the narrowband design in Fig. 1.1, the spatial response in Fig. 1.4 is far from frequency invariant. Why is this? The answer lies in the fact that the BS frequency response has a very sharp cutoff. Consider a single sensor. At low frequencies the sensor is always on. As frequency increases, there will come a point where the sensor will suddenly turn off, and at this frequency the aperture abruptly changes size. This abrupt change in the active aperture causes the alp-like appearance of the spatial response in Fig. 1.4.

Now, returning to the problem of designing the BS coefficients for the desired uniform spatial response, assume we design the BS vector $\boldsymbol{h}$ to ap-

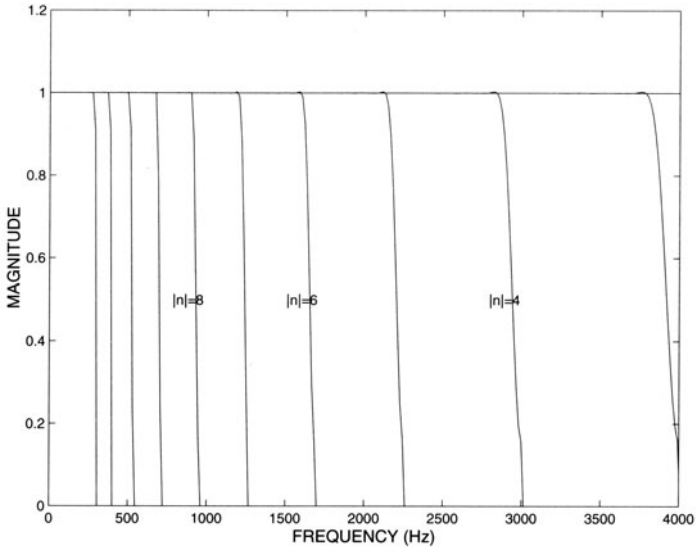


Fig. 1.3. Frequency responses of the weights on each sensor.

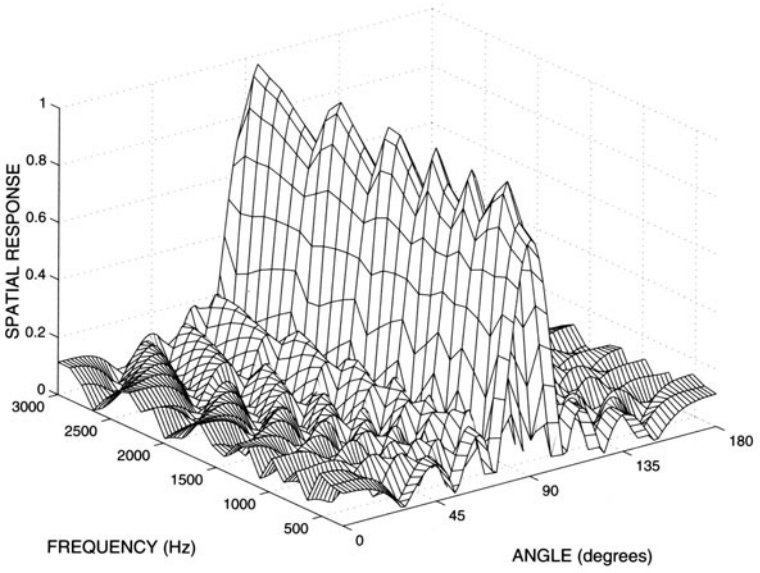


Fig. 1.4. Spatial response of example CDB.

proximate an ideal low-pass filter using only $L = 21$ filter coefficients. This results in the BS frequency responses shown in Fig. 1.5. In comparing this figure with Fig. 1.4, notice that the frequency responses exhibit a more gradual cutoff.

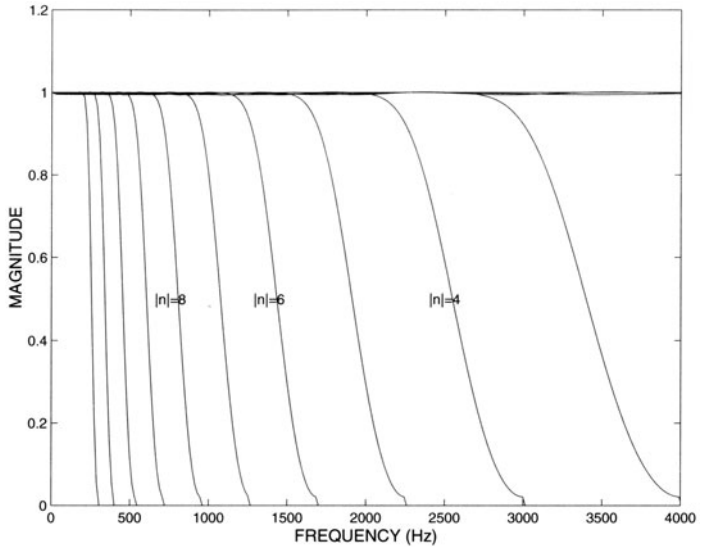


Fig. 1.5. Frequency responses of the weights on each sensor.

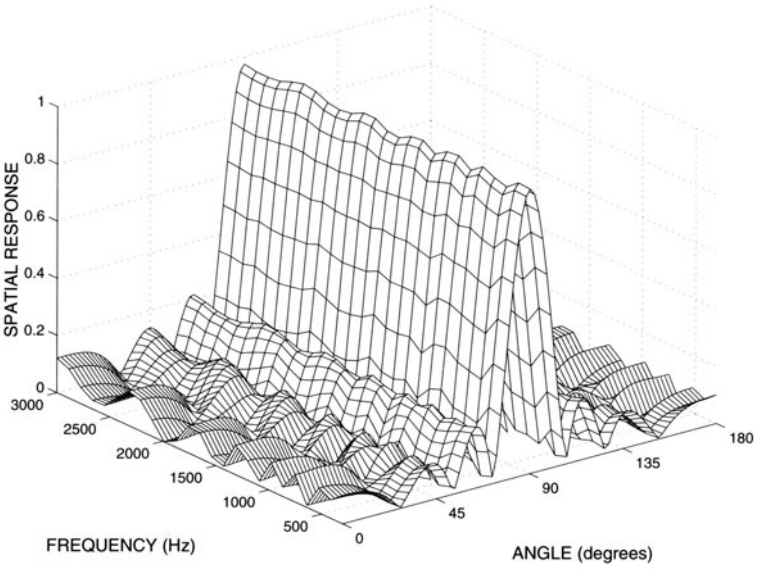


Fig. 1.6. Spatial response of example CDB.

Using these 21 BS coefficients, the resulting spatial response of the CDB is shown in Fig. 1.6. In this case the spatial response shows very little variation with frequency. This demonstrates that one should take careful consideration

of how well the underlying function can be approximated by the discrete array when choosing the required BS function.

1.6 Conclusions

Constant-directivity beamforming is a useful technique for spatial filtering in broadband signal environments in which the desired signal and the interference signals cover approximately the same bandwidth. In this chapter we have developed a technique for designing a CDB, and shown that there is an efficient parameterization and underlying structure exhibited by a CDB. The greatest drawback of a CDB in microphone array applications is that the size of the array is related to the lowest frequency of operation. Thus, producing an array that has a frequency-invariant spatial response down to, say, 300 Hz may require an array that is several meters long. In all but the largest rooms this is impractical. However, a constant spatial response can be readily achieved for mid and high frequencies (above say 1000 Hz) using an array with a total size of less than a meter. For the lower frequencies, other methods (such as the superdirective techniques described in the following chapter) are probably more appropriate.

References

1. B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
2. T.D. Abhayapala, R.A. Kennedy, and R.C. Williamson, "Nearfield broadband array design using a radially invariant modal expansion," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 392–403, Jan. 2000.
3. J.L. Flanagan, D.A. Berkeley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, 1991.
4. W. Kellermann, "A self-steering digital microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-91)*, 1991, vol. 5, pp. 3581–3584.
5. F. Khalil, J.P. Jullien, and A. Gilloire, "Microphone array for sound pickup in teleconference systems," *J. Audio Eng. Soc.*, vol. 42, no. 9, pp. 691–700, Sept. 1994.
6. J. Lardies, "Acoustic ring array with constant beamwidth over a very wide frequency range," *Acoust. Letters*, vol. 13, no. 5, pp. 77–81, 1989.
7. R. Smith, "Constant beamwidth receiving arrays for broad band sonar systems," *Acustica*, vol. 23, pp. 21–26, 1970.
8. M.M. Goodwin and G.W. Elko, "Constant beamwidth beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, 1993, vol. 1, pp. 169–172.
9. J.H. Doles III and F.D. Benedict, "Broad-band array design using the asymptotic theory of unequally spaced arrays," *IEEE Trans. Antennas Propagat.*, vol. 36, no. 1, pp. 27–33, Jan. 1988.

10. A. Ishimaru, "Theory of unequally-spaced arrays," *IRE Trans. Antennas Propagat.*, vol. AP-10, pp. 691–702, Nov. 1962.
11. A. Ishimaru and Y.S. Chen, "Thinning and broadbanding antenna arrays by unequal spacings," *IEEE Trans. Antennas Propagat.*, vol. AP-13, pp. 34–42, Jan. 1965.
12. T. Chou, "Frequency-independent beamformer with low response error," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit, USA, May 1995, pp. 2995–2998.
13. S. Haykin and J. Kesler, "Relation between the radiation pattern of an array and the two-dimensional discrete Fourier transform," *IEEE Trans. Antennas Propagat.*, vol. AP-23, no. 3, pp. 419–420, May 1975.
14. J.S. Marciano Jr. and T.B. Vu, "Reduced complexity beam space broadband frequency invariant beamforming," *Electronics Letters*, vol. 36, no. 7, pp. 682–683; Mar. 2000.
15. D.B. Ward, R.A. Kennedy, and R.C. Williamson, "Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns," *J. Acoust. Soc. Amer.*, vol. 97, no. 2, pp. 1023–1034, Feb. 1995.
16. D.B. Ward, R.A. Kennedy, and R.C. Williamson, "FIR filter design for frequency-invariant beamformers," *IEEE Signal Processing Lett.*, vol. 3, no. 3, pp. 69–71, Mar. 1996.
17. C. Winter, "Using continuous apertures discretely," *IEEE Trans. Antennas Propagat.*, vol. AP-25, pp. 695–700, Sept. 1977.

2 Superdirective Microphone Arrays

Joerg Bitzer[1] and K. Uwe Simmer[2]

[1] Houpert Digital Audio, Bremen, Germany
[2] Aureca GmbH, Bremen, Germany

Abstract. This chapter gives an overview of so-called superdirective beamformers, which can be derived by applying the minimum variance distortionless response (MVDR) principle to theoretically well-defined noise fields, as for example the diffuse noise field. We show that all relevant performance measures for beamformer designs are functions of the coherence matrix of the noise field. Additionally, we present unconstrained and constrained MVDR-solutions using modified coherence functions. Solutions for different choices of the optimization criterion are given including a new solution to optimize the front-to-back ratio. Finally, we present a comparison of superdirective beamformers to gradient microphones and an alternative generalized sidelobe canceler (GSC) implementation of the superdirective beamformer.

2.1 Introduction

What is "super" about a superdirective microphone array? Compared to the standard delay-and-sum beamformer a superdirective array achieves a higher directivity. Therefore, "super"-directivity indicates that summing is not the optimal choice for combining sensor signals, if optimal directivity is desired. The term directivity describes the ability of a beamformer to suppress noise coming from all directions without affecting a desired signal from one principal direction.

A short historical overview in [6] shows that superdirectivity (or supergain) in connection with array processing was first mentioned in the first half of the last century. The solutions provided at that time were of academic interest only, since a lot of practical problems occurred which restricted the use of the theoretical work. The main reasons for failure were the self-noise and the gain and phase errors of the microphones. In order to overcome these problems a first constrained solution was published by Gilbert and Morgan in 1955 [15]. Early applications with slight modifications were seismic and sonar techniques [5]. It was not until the 90's that supergain was connected to microphone applications. Research in hearing aids highlighted the advantages of fixed beamformers over adaptive solutions [17]. Modern designs of superdirective beamformers include nearfield assumptions and the possibility to adapt the constraining to the actual problem.

This chapter is organized as follows: Section 2.2 introduces the measures to judge the different designs. In section 2.3 the optimal design will be derived

with respect to the given problems. Further extensions and special details are given in section 2.4. Concluding remarks close this chapter.

2.2 Evaluation of Beamformers

In order to get a better understanding of the features of the different designs of optimal beamformers, we first need to derive the measures to analyze their performance.

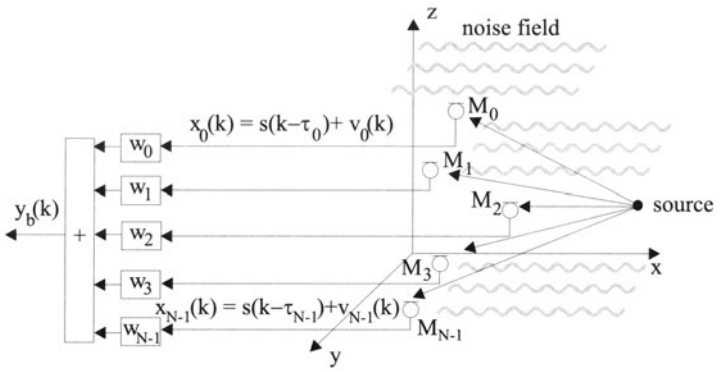


Fig. 2.1. Signal model consisting of noise field and desired source signal

The signal model is shown in Fig. 2.1. We assume that one sample of the discrete input sequence $x(k)$ at each sensor n consists of a delayed and attenuated version of the desired signal $a_i s(k-\tau_n)$ and a noise component $v_n(k)$ with arbitrary spatial statistics.

$$\begin{pmatrix} x_0(k) \\ x_1(k) \\ \vdots \\ x_{N-1}(k) \end{pmatrix} = \begin{pmatrix} a_0 s(k-\tau_0) \\ a_1 s(k-\tau_1) \\ \vdots \\ a_{N-1} s(k-\tau_{N-1}) \end{pmatrix} + \begin{pmatrix} v_0(k) \\ v_1(k) \\ \vdots \\ v_{N-1}(k) \end{pmatrix}$$

$$\boldsymbol{x}(k) = \boldsymbol{a}\boldsymbol{s}(k-\boldsymbol{\tau}) + \boldsymbol{v}(k)\,. \tag{2.1}$$

Since all relevant quantities and designs depend on the frequency, the following examinations are carried out in the frequency domain without any loss of generality. The Fourier-transform leads to

$$\boldsymbol{X}(e^{j\Omega}) = S(e^{j\Omega})\boldsymbol{d} + \boldsymbol{V}(e^{j\Omega})\,, \tag{2.2}$$

where $\boldsymbol{d}$ is the representation of the delays and the attenuation in the frequency domain which depends on the actual geometry of the array and the

direction of the source signal.

$$\boldsymbol{d}^T = [a_0 \exp(-j\Omega\tau_0), a_1 \exp(-j\Omega\tau_1), \cdots, a_{N-1} \exp(-j\Omega\tau_{N-1})] \ . \quad (2.3)$$

Finally, the output signal

$$Y_b(e^{j\Omega}) = \sum_{n=0}^{N-1} W_n{}^*(e^{j\Omega}) X_n(e^{j\Omega}) = \boldsymbol{W}^H \boldsymbol{X} \ , \quad (2.4)$$

where $W_n(e^{j\Omega})$ denotes the frequency-domain coefficients of the beamformer of sensor n at the frequency Ω and the operator ^{H} denotes a conjugated transposition (Hermitian operator). The inverse Fourier-transform results in the discrete-time output signal $y_b(k)$.

2.2.1 Array-Gain

The array-gain (AG) is the measure which shows the improvement of the signal-to-noise ratio (SNR) between one sensor and the output of the whole array [1]. Therefore,

$$G = \frac{SNR_{\text{Array}}}{SNR_{\text{Sensor}}} \ . \quad (2.5)$$

Assuming stationary signals, the SNR of one sensor is given by the ratio of the power spectral densities (PSD) of the signal Φ_{SS} and the average noise $\Phi_{V_a V_a}$.

The SNR at the output can be computed by deriving the PSD of the output signal

$$\Phi_{Y_b Y_b} = \boldsymbol{W}^H \boldsymbol{\Phi}_{\boldsymbol{XX}} \boldsymbol{W} \ , \quad (2.6)$$

where

$$\boldsymbol{\Phi}_{\boldsymbol{XX}} = \begin{pmatrix} \Phi_{X_0 X_0} & \Phi_{X_0 X_1} & \cdots & \Phi_{X_0 X_{N-1}} \\ \Phi_{X_1 X_0} & \Phi_{X_1 X_1} & \cdots & \Phi_{X_1 X_{N-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{X_{N-1} X_0} & \Phi_{X_{N-1} X_1} & \cdots & \Phi_{X_{N-1} X_{N-1}} \end{pmatrix} \quad (2.7)$$

is a power spectral density matrix of the array input signals. When the desired signal is present only, the output is

$$\Phi_{Y_b Y_b}\Big|_{\text{Signal}} = \Phi_{SS} \left| \boldsymbol{W}^H \boldsymbol{d} \right|^2 \ , \quad (2.8)$$

[1] The dependence on Ω is omitted for the sake of brevity and readability.

and for the noise-only case the output is

$$\boldsymbol{\Phi}_{Y_b Y_b}\Big|_{\text{Noise}} = \Phi_{V_a V_a} \boldsymbol{W}^H \boldsymbol{\Phi}_{\boldsymbol{VV}} \boldsymbol{W} \,, \tag{2.9}$$

where $\boldsymbol{\Phi}_{\boldsymbol{VV}}$ is a normalized cross power spectral density matrix of the noise[2]. Therefore,

$$G = \frac{\left|\boldsymbol{W}^H \boldsymbol{d}\right|^2}{\boldsymbol{W}^H \boldsymbol{\Phi}_{\boldsymbol{VV}} \boldsymbol{W}} \,. \tag{2.10}$$

Assuming a homogeneous noise field (2.10) can be expressed in terms of the coherence matrix

$$\boldsymbol{\Gamma}_{\boldsymbol{VV}} = \begin{pmatrix} 1 & \Gamma_{V_0 V_1} & \Gamma_{V_0 V_2} & \cdots & \Gamma_{V_0 V_{N-1}} \\ \Gamma_{V_1 V_0} & 1 & \Gamma_{V_1 V_2} & \cdots & \Gamma_{V_1 V_{N-1}} \\ \vdots & \vdots & \ddots & \vdots & \\ \Gamma_{V_{N-1} V_0} & \Gamma_{V_{N-1} V_1} & \Gamma_{V_{N-1} V_2} & \cdots & 1 \end{pmatrix} \,, \tag{2.11}$$

where

$$\Gamma_{V_n V_m}(e^{j\Omega}) = \frac{\Phi_{V_n V_m}(e^{j\Omega})}{\sqrt{\Phi_{V_n V_n}(e^{j\Omega}) \Phi_{V_m V_m}(e^{j\Omega})}} \tag{2.12}$$

is the coherence function [4].

Thus,

$$G = \frac{\left|\boldsymbol{W}^H \boldsymbol{d}\right|^2}{\boldsymbol{W}^H \boldsymbol{\Gamma}_{\boldsymbol{VV}} \boldsymbol{W}} \,. \tag{2.13}$$

This representation allows an easier examination of beamformers for different noise fields, since many theoretically defined noise fields can be expressed by their coherence function.

2.2.2 Beampattern

One way to evaluate beamformers is to compute the response of the array to a wavefront coming from a specific frequency and a specific angle, depending on azimuth φ and elevation θ in a spherical coordinate system. Computing this response over all angles and frequencies leads to the spatial-temporal transfer function

$$\left|H(\varphi,\theta)\right|^2\Big|_{\text{dB}} = -10 \log_{10} \left(\frac{\left|\boldsymbol{W}^H \boldsymbol{d}\right|^2}{\boldsymbol{W}^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}\Big|_{\text{Wavefront}} \boldsymbol{W}} \right) \tag{2.14}$$

[2] The normalization factor is set to force the trace of the matrix to equal N.

called the farfield beampattern, which is usually displayed on a logarithmic scale. It can be computed by using (2.13) and the knowledge of the coherence function of a single wavefront with frequency Ω and an angle of arrival φ, θ. Additionally, f_s denotes the sampling frequency, $c = 340$ m/s the speed of sound, and $l_{n\,m}$ the distances between the sensors in the Cartesian coordinate system

$$\Gamma_{V_n V_m}\Big|_{\text{Wavefront}} = \exp(j\Omega\tau_{n\,m}) \,, \tag{2.15}$$

where

$$\tau_{n\,m} = \frac{f_s}{c}\left(l_{x,n\,m}\sin(\theta)\cos(\varphi) + l_{y,n\,m}\sin(\theta)\sin(\varphi) + l_{z,n\,m}\cos(\theta)\right) \,. \tag{2.16}$$

Since the beampattern depends on three variables, it is not possible to display it in a single plot. Fortunately, line arrays aligned to the z-axis have a rotational symmetry and, therefore, the beampattern is independent of φ. Examples of beampatterns for line arrays will be shown in section 2.3.

2.2.3 Directivity

A common quantity to evaluate beamformers is the directivity factor, or its logarithmic equivalent the directivity index (DI) which describes the ability of the array to suppress a diffuse noise field. Therefore, we can compute the directivity factor by using (2.13) and inserting the coherence function of a diffuse noise field:

$$\begin{aligned}\Gamma_{V_n V_m}(e^{j\Omega})\Big|_{\text{Diffuse}} &= \frac{\sin\left(\Omega f_s l_{n\,m}/c\right)}{\Omega f_s l_{n\,m}/c} \\ &= \operatorname{sinc}\left\{\frac{\Omega f_s l_{n\,m}}{c}\right\}\end{aligned} \tag{2.17}$$

where $\operatorname{sinc}(x) = \sin(x)/x$. Thus, the DI is

$$\mathrm{DI}(e^{j\Omega}) = 10\log_{10}\left(\frac{\left|\boldsymbol{W}^H\boldsymbol{d}\right|^2}{\boldsymbol{W}^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}\Big|_{\text{Diffuse}}\boldsymbol{W}}\right) \,. \tag{2.18}$$

Another formal definition uses the transfer function (2.14) and describes the ratio of the transfer function of the look-direction θ_0, φ_0 of the array to the spatial integration over all directions of incoming signals.

$$\mathrm{DI}(e^{j\Omega}) = 10\log_{10}\left(\frac{\left|H(e^{j\Omega}, \varphi_0, \theta_0)\right|^2}{\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi}\left|H(e^{j\Omega}, \varphi, \theta)\right|^2 \sin(\theta)d\varphi d\theta}\right) \tag{2.19}$$

2.2.4 Front-to-Back Ratio

In many applications no principal look-direction exists, as for example in video-conferences or the recording of orchestras. Therefore, the DI is not the best quantity to describe the behavior of the array. In such applications a front-to-back ratio (FBR) is a better choice, since in most cases all desired sources are in front of the array and all unwanted disturbances are behind the array [19], [11]. The formal description utilizes the beampattern again:

$$\mathrm{FBR}(e^{j\Omega}) = \frac{\int_{\theta_0-\pi/2}^{\theta_0+\pi/2}\int_{\varphi_0-\pi/2}^{\varphi_0+\pi/2} \left|H(e^{j\Omega},\varphi,\theta)\right|^2 \sin(\theta)d\varphi d\theta}{\int_{\theta_0+\pi/2}^{\theta_0+3\pi/2}\int_{\varphi_0+\pi/2}^{\varphi_0+3\pi/2} \left|H(e^{j\Omega},\varphi,\theta)\right|^2 \sin(\theta)d\varphi d\theta} \tag{2.20}$$

2.2.5 White Noise Gain

This last quantity shows the ability of the array to suppress spatially uncorrelated noise, which can be caused by self-noise of the sensors. Inserting the coherence matrix for this noise field

$$\boldsymbol{\Gamma_{VV}}\Big|_{\text{uncorr}} = \boldsymbol{I} \tag{2.21}$$

into (2.13) results in the white noise gain:

$$\mathrm{WNG}(e^{j\Omega}) = \frac{\left|\boldsymbol{W}^H\boldsymbol{d}\right|^2}{\boldsymbol{W}^H\boldsymbol{W}} \,. \tag{2.22}$$

On a logarithmic scale positive values represent an attenuation of uncorrelated noise, whereas negative values show an amplification.

2.3 Design of Superdirective Beamformers

In order to design optimal beamformers, we have to minimize the power of the output signal $y_b(k)$ of the array. The output PSD is given by (2.6) and is a function of the input signal and the coefficients we want to determine. In order to avoid the trivial solution $W_n = 0$, the minimization is constrained to give an undistorted signal response in the desired look direction, i.e.,

$$\boldsymbol{W}^H\boldsymbol{d} = 1 \,. \tag{2.23}$$

Therefore, the following constrained minimization problem has to be solved:

$$\min_{\boldsymbol{W}} \boldsymbol{W}^H\boldsymbol{\Phi_{XX}}\boldsymbol{W} \quad \text{subject to} \quad \boldsymbol{W}^H\boldsymbol{d} = 1 \,. \tag{2.24}$$

Since we are only interested in the optimal suppression of the noise, and we assume a perfect correspondence between the direction of the desired signal and the look-direction of the array, only the noise PSD-matrix $\boldsymbol{\Phi_{VV}}$ is used.

The well-known solution for (2.24) is called the Minimum Variance Distortionless Response (MVDR) beamformer [6]. It is given by

$$\boldsymbol{W} = \frac{\boldsymbol{\Phi_{VV}}^{-1}\boldsymbol{d}}{\boldsymbol{d}^H\boldsymbol{\Phi_{VV}}^{-1}\boldsymbol{d}}\,, \tag{2.25}$$

and can be derived by using the Lagrange-multiplier [13] or gradient computation [20], [9]. Assuming a homogeneous noise field the solution is a function of the coherence matrix:

$$\boldsymbol{W} = \frac{\boldsymbol{\Gamma_{VV}}^{-1}\boldsymbol{d}}{\boldsymbol{d}^H\boldsymbol{\Gamma_{VV}}^{-1}\boldsymbol{d}}\,. \tag{2.26}$$

Equations (2.25) or (2.26) can be interpreted as a spatial decorrelation process followed by a matched filter for the desired signal. The normalization in the denominator leads to unity signal response for the look direction.

The design procedure reduces to the choice of theoretically well-defined noise-fields in order to get optimal designs for different applications. Furthermore, different models for the desired signal can be included, leading to farfield and nearfield designs.

Examples for desired signal models are:

- Standard farfield model for linear arrays with equidistant sensors:

$$\begin{aligned}\boldsymbol{d}^T = [&1, \exp(-j\Omega f_s c^{-1} l\cos(\theta_0)), \exp(-j\Omega f_s c^{-1} 2l\cos(\theta_0)),\\ &\cdots, \exp(-j\Omega f_s c^{-1}(N-1)l\cos(\theta_0))]\end{aligned} \tag{2.27}$$

  where l is the inter-sensor spacing.
- Nearfield design, including attenuation of the desired signal [14], [22]

$$\boldsymbol{d}^T = [a_0\exp(-j\omega\tau_0), a_1\exp(-j\omega\tau_1), \cdots, a_{N-1}\exp(-j\omega\tau_{N-1})]\,, \tag{2.28}$$

$$a_i = \frac{||\boldsymbol{q}-\boldsymbol{p}_{ref}||}{||\boldsymbol{q}-\boldsymbol{p}_i||}\,, \tag{2.29}$$

$$\tau_i = \frac{||\boldsymbol{q}-\boldsymbol{p}_{ref}|| - ||\boldsymbol{q}-\boldsymbol{p}_i||}{c}\,, \tag{2.30}$$

  where $||\boldsymbol{q}-\boldsymbol{p}_{ref}||$ and $||\boldsymbol{q}-\boldsymbol{p}_i||$ denote the distance between the vector location of the source q and a reference sensor p_{ref}, or the sensor p_i, respectively.

More elaborate examples for exact nearfield designs can be found in [18], [23]

2.3.1 Delay-and-Sum Beamformer

Although this chapter is called superdirective microphone arrays the well-known Delay-and-Sum Beamformer (DSB) is included for comparison purposes. It is an 'optimal' beamformer for optimizing the WNG. We can derive the coefficients from (2.26) by inserting the coherence matrix for spatial uncorrelated noise $\boldsymbol{\Gamma} = \boldsymbol{I}$. Thus,

$$\boldsymbol{W} = \frac{1}{N}\boldsymbol{d}\,. \tag{2.31}$$

The WNG is optimal in this case and reaches N. All other standard shading schemes like the Dolph-Chebycheff window [10] worsen the performance subject to WNG.

2.3.2 Design for spherical isotropic noise

In order to optimize the directivity factor, which depends on the noise-field of a spherical isotropic noise field (diffuse), we have to solve (2.26) by using the coherence matrix of the diffuse noise field, given by (2.17). The resulting coefficients represent the classic superdirective beamformer (SDB)[3].

Figure 2.2 shows the beampattern of a DSB and a superdirective beamformer, both using five linear equispaced microphones ($l = 5$ cm) in endfire steering direction ($\theta_0 = \pi$). The x-axis represents the incoming spatial angle ($[0 \cdots 2\pi]$) and the y-axes represents the frequency of the signal in kHz. The sampling-frequency was set to 8 kHz to cover the telephone bandwidth. The grey-scaled image represents the attenuation of the incoming signals in dB.

The look-direction is unmodified at all frequencies due to the linear constraint. Additionally, an unmodified region at higher frequencies can be seen caused by spatial aliasing, since our choice of the parameter does not fulfill the spatial sampling theorem, which is given by

$$l < \frac{\lambda}{2}\,, \tag{2.32}$$

where λ denotes the wavelength. The upper sampling frequency should therefore be restricted to $f_s = 6.8$ kHz, or the distance should not exceed $l = 4.25$ cm. However, in order to show some effects we will keep these parameters in all experiments.

Furthermore, the DSB is unable to suppress low frequency noise sources coming from any direction. In contrast, the superdirective beamformer attenuates very well sources coming from directions other than the look-direction

[3] In this chapter the term superdirective beamformer is used for the beamformer which optimizes the directivity factor, independent of the frequency or the ratio of the wavelength to the distance between the sensor elements. In the classic definition this is often restricted to the case where the wavelength is large with respect to the distance between sensors.

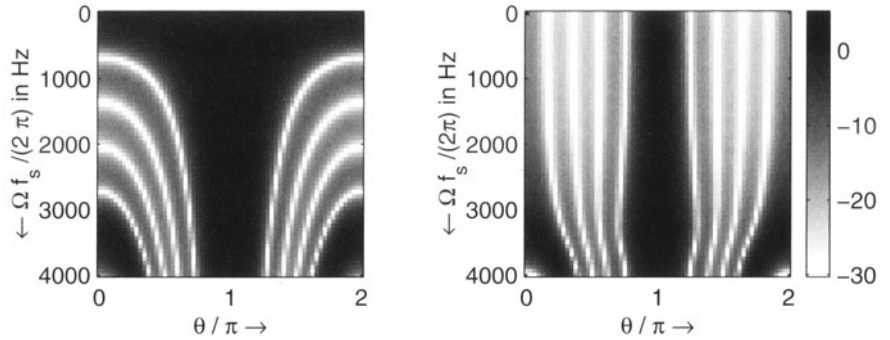


Fig. 2.2. Left: beampattern of a delay-and-sum beamformer. Right: beampattern of an optimal array for isotropic noise (superdirective beamformer). ($l = 5$ cm, $N = 5$, endfire steering direction)

over the whole frequency range. However, at higher frequencies the superdirective beamformer degrades to the DSB, since supergain can only be achieved if the signal wavelength is larger than two times the microphone distance.

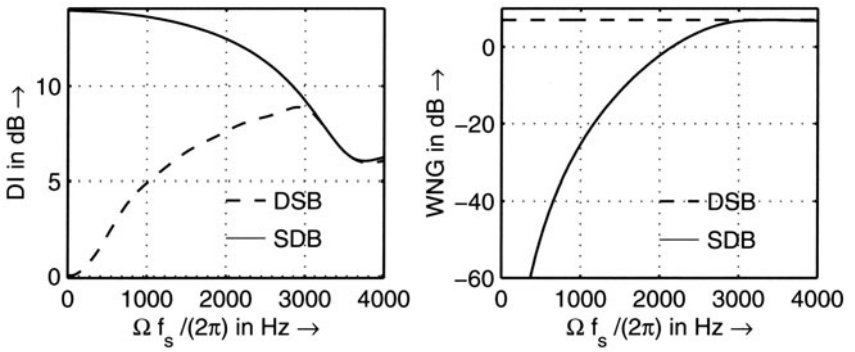


Fig. 2.3. Left: Directivity index (DI) for delay-and-sum beamformer and superdirective beamformer. Right: White noise gain (WNG) for delay-and-sum beamformer and superdirective beamformer. ($l = 5$ cm, $N = 5$, endfire steering direction)

Figure 2.3 shows the DI on the left side and the WNG on the right side for the same parameters as in the previous figure. The directivity index reaches zero at low frequencies for the DSB (as expected by analyzing the beampattern) and N^2 for the superdirective beamformer. The proof for this limit in the endfire steering case can be found in [11]. At higher frequencies the directivity for both designs is nearly the same and it is given by N, since the $\mathrm{sinc}\{\cdot\}$ function tends to zero, and the noise field is uncorrelated in this case.

If we now take a closer look at the WNG, we can see why this design is not suitable in real-world applications. Whereas the DSB suppresses uncorrelated noise equally at all frequencies, the SDB boosts uncorrelated noise at lower frequencies.

In order to give a deeper insight into how supergain works, we will compute the coefficients for an array of only two microphones. The distance is again 5 cm, and endfire steering is used.

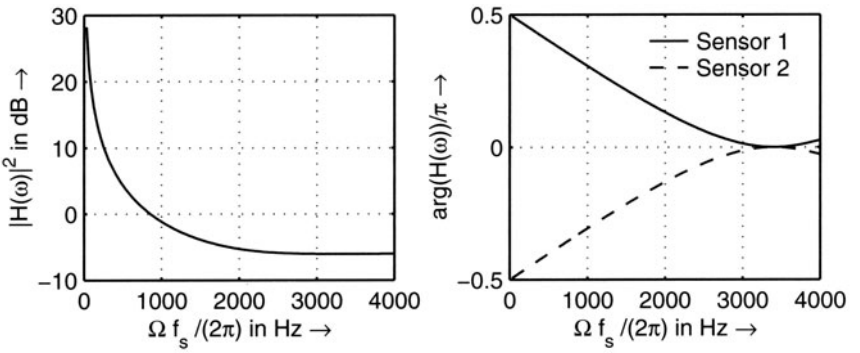


Fig. 2.4. Coefficients of a two channel SDB, left: Magnitude, right: Phase ($l = 5$ cm, $N = 2$, endfire steering direction)

In Fig. 2.4 the squared magnitude and the phase of the two coefficient vectors are shown. First of all, the coefficients are conjugate complex. Secondly, the filters force the phase between the noise components at each sensor to be π. Therefore, the correlated part of the noise will be compensated. Hence, the desired signal is also correlated, and therefore it is reduced as well. To fulfill the constraint of an undisturbed desired signal, the coefficients have to boost the input signals to compensate this behavior, which can be seen on the left side of Fig. 2.4. Therefore, uncorrelated noise will be amplified. At higher frequencies the correlation between the noise components vanishes and the beamformer degrades to the DSB. The magnitude of the coefficients reaches $1/2$.

In order to overcome the problem of self-noise amplification in superdirective designs, Gilbert and Morgan have proposed a method for solving (2.24) under a WNG constraint [15]. The method uses a small added scalar μ to the main diagonal of the normalized PSD or coherence matrix:

$$\boldsymbol{W}_c = \frac{(\boldsymbol{\Gamma}_{\boldsymbol{VV}} + \mu \boldsymbol{I})^{-1}\boldsymbol{d}}{\boldsymbol{d}^H(\boldsymbol{\Gamma}_{\boldsymbol{VV}} + \mu \boldsymbol{I}^{-1}\boldsymbol{d}}. \tag{2.33}$$

We prefer a mathematically equivalent form, which preserves the interpretation as a coherence matrix with elements smaller than one. Instead of adding

the scalar to the main diagonal, we divide each non-diagonal element by $1+\mu$. Therefore, μ can be interpreted as the ratio of the sensor noise σ^2 to the ambient noise power Φ_{VV}. For the diffuse noise field the non-diagonal elements are given by

$$\Gamma_{V_n V_m} = \frac{\operatorname{sinc}\left\{\dfrac{\Omega f_s l_{n\,m}}{c}\right\}}{1+\dfrac{\sigma^2}{\Phi_{VV}}} \quad \forall \quad n \neq m\,. \tag{2.34}$$

The factor μ can vary from zero to infinity, which results in the unconstrained SDB or the DSB respectively. The WNG changes as a monotonic function between the two limits [15]. Typical values for μ are in the range between $-10\,\mathrm{dB}$ to $-30\,\mathrm{dB}$. Unfortunately, there is no simple relation between μ and the resulting WNG. By using a frequency variant μ the WNG can be restricted at all frequencies, but not through direct computation.

There are two different iterative design schemes. The first one was published by Doerbecker [9]. It is a straightforward implementation of a trial-and-error strategy. Another iterative design method uses the scaled projection algorithm developed by Cox *et al.* for adaptive arrays [6]. Instead of the estimated PSD-matrix, the theoretically defined coherence or PSD-matrix is inserted in the scaled projection algorithm. This solution was presented in [17]. Both algorithms result in similar coefficients and can be implemented easily.

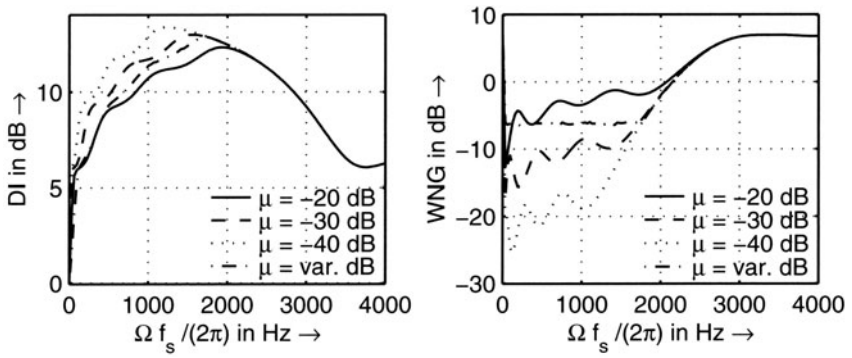


Fig. 2.5. Left: Directivity index (DI) for different constrained designs. Right: White noise gain (WNG) for different constrained designs. ($l = 5$ cm, $N = 5$, endfire steering direction)

Figure 2.5 depicts the effects for three fixed and one variable μ as constraining parameters. For the variable μ, the WNG constraint was set to $-6\,\mathrm{dB}$. The constrained design facilitates a good compromise between DI

and WNG. A careful design can optimize such arrays for a wide range of applications.

2.3.3 Design for Cylindrical Isotropic Noise

In some applications a spherical isotropic noise field is not the best choice or the best approximation of a given noise-field. Another well-defined noise-field can be used, if we reduce the three dimensions to two dimensions. We get a noise-field which is defined by infinite noise sources of a circle with an infinite radius. This kind of noise can arise if a lot of people speak in large rooms where the ceiling and the floor are damped well, or in the free-field (cocktail-party noise) [4]. The coherence between two sensors is given by [7]

$$\Gamma_{X_n X_m}(\omega) = J_0\left(\frac{\omega l_{n\,m}}{c}\right) , \tag{2.35}$$

where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind. This leads to the solution of [8] as an improved design for speech enhancement for a hearing-aid application. In order to constrain the coefficients, a similar technique as in (2.34) has to be carried out.

In comparison to the design for a diffuse noise-field the differences are not large, but at lower frequencies a better suppression of noise sources behind the look direction can be observed. Elko [11] has shown that the directivity factor is less and its limit is $2N-1$, in contrast to N^2 in the unconstrained case ($\mu = 0$). A design example will be given in the next section.

2.3.4 Design for an Optimal Front-to-Back Ratio

A last data-independent design tries to optimize the front-to-back ratio. In many applications the look direction of the desired signal cannot be predetermined, but in most cases the desired signal is in front of the array and all disturbances are at the rear, e.g. when recording an orchestra or in video-conferences.

Our suggestion for a different design strategy is not to use an isotropic noise field, but to restrict the assumed infinite noise sources to the back half of a circle or a sphere.

The resulting noise-field between two sensors separated by the distance l can be described by an integration over an infinite number of uncorrelated noise sources. The resulting function in the two-dimensional case is:

$$f(e^{j\Omega}, \theta_0) = \frac{1}{\pi}\int_{\theta_0+\pi/2}^{\theta_0+3\pi/2} \exp\left(j\Omega f_s c^{-1} l \cos(\theta)\right)\, d\theta. \tag{2.36}$$

[4] The origin of this cylindrical isotropic noise-field is the sonar application in shallow water.

Using numerical integration methods, inserting the resulting complex values in the coherence matrix and solving (2.26), results in a new design which suppresses noise sources from the rear very well.

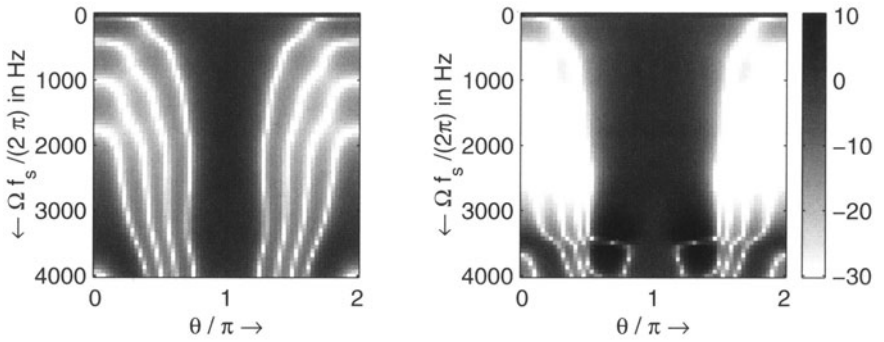


Fig. 2.6. Left: beampattern of a constrained superdirective beamformer. Right: beampattern of a constrained beamformer, designed with (2.36). ($l = 5$ cm, $N = 5$, $\mu = 0.01$, endfire steering direction)

Figure 2.6 shows beampatterns of two constrained beamformers ($\mu = 0.01$). The left side is computed with optimized coefficients for a diffuse noise-field, and the right side uses coefficients designed with the help of (2.36). At lower frequencies the constraining parameter is dominant and therefore, both designs do not perform well. From 300 Hz to 2800 Hz the new design suppresses all signals coming from the rear at the cost of a wider main lobe; this is sometimes an advantage, for example if the source is not exactly in endfire position.

At higher frequencies, especially if spatial aliasing occurs, the new design boosts signals coming from directions near the look direction, which can cause some unnatural coloring of the signal and the remaining noise. Therefore, special care has to be taken when choosing the parameters of the new design scheme.

In order to show the advantages of the new schemes, Fig. 2.7 depicts the DI and the FBR measure for the three different designs. At lower frequencies the small advantage of the cylindrical optimal design against the spherical design for the FBR can be seen, but the differences are very small over the whole frequency range. On the other hand, the behavior of the new design is completely different. Measuring the DI leads to much smaller values, but the FBR is very high, especially in the mid-frequency range.

Interestingly, we can transform between the optimal design for cylindrical isotropic noise and the new design by introducing a new variable which

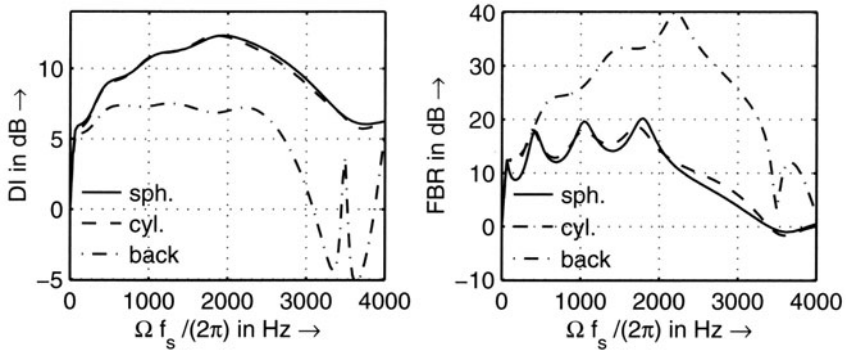


Fig. 2.7. Left: Directivity index (DI) for three optimal designs. Right: Front-to-back ratio (FBR) for three optimal designs. ($l = 5$ cm, $N = 5$, $\mu = 0.01$, endfire steering direction)

adjusts the limits of the integral, i.e.,

$$f(e^{j\Omega}, \theta_0, \delta) = \frac{1}{2(\pi - \delta)} \int_{\theta_0+\delta}^{\theta_0-\delta+2\pi} \exp\left(j\Omega f_s c^{-1} l \cos(\theta)\right) \, d\theta \quad 0 \leq \delta \leq \pi \tag{2.37}$$

Setting $\delta = 0$ corresponds to the isotropic noise case, and $\delta = \pi/2$ results in (2.36).

2.3.5 Design for Measured Noise Fields

So far, only data-independent designs have been considered. If *a priori* knowledge is available, however, it should be used to improve the performance. For example, this information could be a prescribed direction (θ = angle) of an incoming noise source. Assuming the noise source is in the far field of the microphone array, the complex coherence function between two sensors is given by

$$Re\{\Gamma_{X_n X_m}(\omega)\} = \cos\left(\frac{\Omega f_s \cos(\theta) l_{n\,m}}{c}\right) \tag{2.38}$$

$$Im\{\Gamma_{X_n X_m}(\omega)\} = -\sin\left(\frac{\Omega f_s \cos(\theta) l_{n\,m}}{c}\right) \quad . \tag{2.39}$$

Inserting the complete coherence matrix in (2.26) forms a null in that direction over the whole frequency range. In order to restrict the WNG a constrained design is necessary.

Furthermore, if we assume stationarity we can measure the actual noise-field and solve the design equation which results in the MVDR solution. Adaptive algorithms like the constrained projection by Cox [6], or the original

algorithm by Frost [13], will converge exactly to the same solution under the assumption of stationary noise and an infinitely small step-size.

2.4 Extensions and Details

After describing the main form of the MVDR beamformer and typical data-independent designs, we will compare them to their analogue counterparts, the gradient microphones. Furthermore, an alternative implementation structure will be given which can reduce the computational complexity and open superdirective designs for future extensions.

2.4.1 Alternative Form

Assuming a time-aligned input signal, the optimal weights are defined differently, since the look-direction vector $\boldsymbol{d}$ is replaced by the column-vector

$$\mathbf{1} = \underbrace{[1, 1, \cdots, 1]}_{N}^{T}$$

containing only ones, and the PSD-matrix or the coherence matrix contain the statistical information after time alignment (see Fig. 2.8). This gives

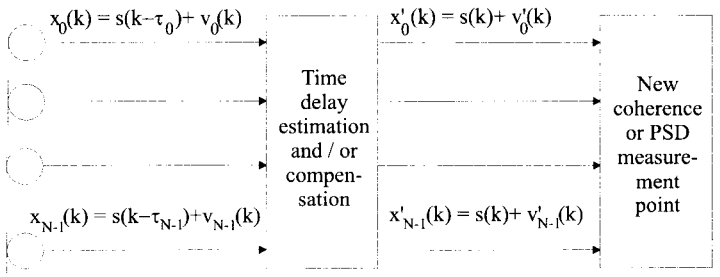


Fig. 2.8. Signal model after time delay compensation

$$\boldsymbol{W}\Big|_{\text{ta}} = \frac{\mathbf{1}^T(\boldsymbol{\Gamma}'_{\boldsymbol{VV}} + \mu\boldsymbol{I})^{-1}}{\mathbf{1}^T(\boldsymbol{\Gamma}'_{\boldsymbol{VV}} + \mu\boldsymbol{I})^{-1}\mathbf{1}}. \tag{2.40}$$

This solution of the constrained minimization problem can be decomposed into two orthogonal parts, following the ideas of Griffith and Jim [16]. One part represents the constraints only and the other part represents the unconstrained coefficients to minimize the output power of the noise.

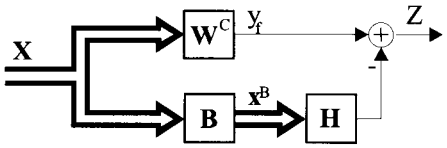


Fig. 2.9. Schematic description of the decomposition of the optimal weight vector into two orthogonal parts

The decomposed structure is depicted in Fig. 2.9. The multi-channel time-aligned input signal $\boldsymbol{X}$ is multiplied by $\boldsymbol{W}^C$ to fulfill the constraints. Furthermore, the input signal is projected onto the noise-only subspace[5] by a blocking matrix $\boldsymbol{B}$. The resulting vector $\boldsymbol{X}^B$ is multiplied by the optimal vector $\boldsymbol{H}$ and then subtracted from the output of the upper part of the structure to get the noise-reduced output signal Z. Several authors have shown the equivalence between this structure and the standard beamformer [16], [3], [12], if

$$\boldsymbol{W}^C = \frac{1}{N}\mathbf{1}\,,$$

which represents a delay-and-sum beamformer. Additionally, $\boldsymbol{B}$ has to fulfill the following properties:

- The size of the matrix is $(N-1) \times N$
- The sum of all values in one row is zero
- The matrix has to be of rank N-1.

An example for $N = 4$ is given by

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \tag{2.41}$$

Another well-known example is the original Griffith-Jim matrix which subtracts two adjacent channels only:

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{pmatrix}.$$

The last step to achieve a solution equivalent to (2.25) is the computation of the optimal filter $\boldsymbol{H}$. A closer look at Fig. 2.9 shows that Y_f, $\boldsymbol{X}^B$ and Z describe exactly the problem of a multiple input noise canceler, described by

[5] Which means that the desired signal is spatially filtered out (blocked).

Widrow and Stearns [24]. Therefore, this structure is called the generalized sidelobe canceler (GSC), if an adaptive implementation is used. The non-adaptive multi-channel Wiener solution of this problem can be found in [21]

$$\boldsymbol{H} = \boldsymbol{\Phi}_{\boldsymbol{X}^B\boldsymbol{X}^B}{}^{-1}\boldsymbol{\Phi}_{\boldsymbol{X}^B Y_f} \,, \tag{2.42}$$

where $\boldsymbol{\Phi}_{\boldsymbol{X}^B\boldsymbol{X}^B}$ denotes the PSD-matrix of all signals after the matrix $\boldsymbol{B}$, and $\boldsymbol{\Phi}_{\boldsymbol{X}^B Y_f}$ is the cross-PSD vector between the fixed beamformer output and the output signals $\boldsymbol{X}^B$. Additionally, the coefficient vector can be computed as a function of the input PSD-matrix:

$$\boldsymbol{H} = (\boldsymbol{B}\boldsymbol{\Phi}_{\boldsymbol{V}\boldsymbol{V}}{}'\boldsymbol{B}^H)^{-1}\boldsymbol{B}\boldsymbol{\Phi}_{\boldsymbol{V}\boldsymbol{V}}{}'\boldsymbol{W}^C \,. \tag{2.43}$$

If we now assume a homogeneous noise field, the PSD-matrix can be replaced by the coherence matrix of the delay-compensated noise field to compute the optimal coefficients:

$$\boldsymbol{H} = (\boldsymbol{B}\boldsymbol{\Gamma}_{\boldsymbol{V}\boldsymbol{V}}{}'\boldsymbol{B}^H)^{-1}\boldsymbol{B}\boldsymbol{\Gamma}_{\boldsymbol{V}\boldsymbol{V}}{}'\boldsymbol{W}^C. \tag{2.44}$$

Therefore, all designs presented in section 2.3 can be implemented by using the GSC-structure. However, why should we do that? First of all, the new structure needs one filter less than the direct implementation. Using the first blocking matrix (2.41) further reduces the number of filters [1]. Secondly, a DSB output is available which can be used for future extensions. Thirdly, the new structure allows us to combine superdirective beamformers with adaptive post-filters for further noise reduction [2], and the new structure gives a deeper insight into MVDR-beamforming. For example, we can see that optimal beamforming is an averaging process combined with noise compensation.

2.4.2 Comparison with Gradient Microphones

Other devices with superdirectional characteristics are optimized gradient microphones [11]. In Fig. 2.10 a typical structure of a first order gradient microphone and its technical equivalent (composed of two omni-directional microphones) is shown.

The acoustic delay between the two open parts of the microphone can be realized by placing the diaphragm not exactly in the middle, or by using a material with a slower speed of sound.

The output of such systems is given by

$$E(\omega,\theta) = P_0\big(1 - \exp(-j\omega[\tau + c^{-1}l\cos(\theta)])\big) \,, \tag{2.45}$$

where τ is the acoustic delay and P_0 denotes the amplitude of the source signal. If we now assume a small spacing with respect to the wavelength, an approximate solution can be derived:

$$E(\omega,\theta) \approx P_0\omega(\tau + c^{-1}l\cos(\theta)) \,. \tag{2.46}$$

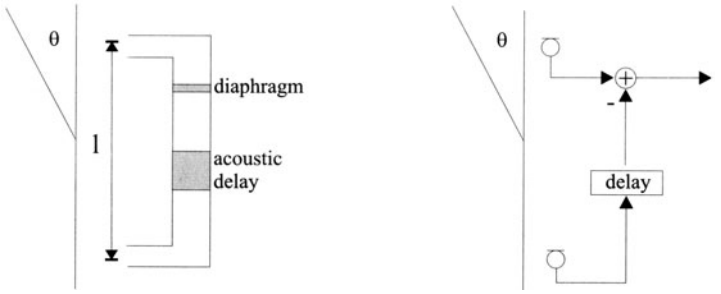


Fig. 2.10. Schematic description of a first order gradient microphone

A proper choice of τ leads to the different superdirective designs, called cardioid, supercardioid and hypercardioid. For example, the beampattern for a hypercardioid first order gradient microphone shows its zeros at $\approx \pm 109°$. This type of microphone is designed to optimize the directivity factor and therefore, it represents the analogue equivalent of a two-sensor superdirective array. For a deeper insight and a complete review of higher order gradient microphones see [11].

At lower frequencies the two systems react more or less equally. The advantages of the analogue system are the smaller size of the device, and that no analogue-to-digital conversion is necessary. The advantages of the digital array technique are its flexibility, the easy scaling for many microphones, and the possible extensions with post-filters or other adaptive techniques.

At higher frequencies, if the assumption of small spacing is not valid anymore, the differences become visible. Through careful manufacturing these frequencies are much higher than the covered bandwidth. However, at some high frequencies the analogue microphone cancels the desired signal completely. On the other hand the array system reacts like a DSB at these frequencies, and no cancellation occurs.

2.5 Conclusion

Designing a so-called superdirective array or an optimal array for theoretically well-defined noise fields can be reduced to solving a single equation. Even nearfield assumptions and measured noise fields can be easily included. We have shown that the spatial characteristic, described by the coherence function, plays a key role in designing arrays. Most of the evaluation tools like the beampattern or the directivity index are directly connected to the coherence function. Beamformer designs with optimized directivity or higher front-to-back ratio also use the coherence.

One of the new aspects included in this chapter was a new noise model to improve the front-to-back ratio. Furthermore, we emphasized the close

relationship between superdirective arrays and adaptive beamformers and their well-known implementation as a generalized sidelobe canceler.

References

1. J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "An alternative implementation of the superdirective beamformer", in *Proc. IEEE Workshop Applicat. Signal Processing to Audio Acoust.*, pp. 7–10, New Paltz, NY, USA, Oct 1999.
2. J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer", in *Proc. Int. Workshop Acoust. Echo and Noise Control*, pp. 100–103, Pocono Manor, USA, Sep 1999.
3. K. M. Buckley, "Broad-band beamforming and the generalized sidelobe canceller", *IEEE Trans. Acoust. Speech Signal Processing*, vol. 34, pp. 1322–1323, Oct 1986.
4. G. C. Carter, *Coherence and Time Delay Estimation*, IEEE Press, 1993.
5. H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain", *IEEE Trans. Acoust. Speech Signal Processing*, vol. 34, pp. 393–398, Jun 1986.
6. H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming", *IEEE Trans. Acoust. Speech Signal Processing*, vol. 35, pp. 1365–1375, Oct 1987.
7. B. F. Cron and C.H. Sherman, "Spatial-correlation functions for various noise models", *J. Acoust. Soc. Amer.*, vol. 34, pp. 1732–1736, Nov 1962.
8. M. Doerbecker, "Speech enhancement using small microphone arrays with optimized directivity", In *Proc. Int. Workshop Acoust. Echo and Noise Control*, pp. 100–103, London, UK, Sep 1997.
9. M. Doerbecker, *Mehrkanalige Signalverarbeitung zur Verbesserung akustisch gestörter Sprachsignale am Beispiel elektronischer Hörhilfen.* PhD thesis, Dept. of Telecommunications, University of TH Aachen, Verlag der Augustinus Buchhandlung, Aachen, Germany, Aug 1998.
10. C. L. Dolph, "A current distribution for broadside arrays which optimizes the relationship between beamwidth and sidelobe level", *Proc. IRE*, pp. 335–348, Jun 1946.
11. G. W. Elko, "Superdirectional microphone arrays", in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, eds, ch. 10, pp. 181–235, Kluwer Academic Press, 2000.
12. M.H. Er and A. Cantoni, "Transformation of linearly constrained broadband processors to unconstrained partitioned form", *IEE Proc. Pt. H*, vol. 133, pp. 209–212, June 1986.
13. O. L. Frost, "An algorithm for linearly constrained adaptive array processing", *Proc. IEEE*, vol. 60, pp. 926–935, Aug 1972.
14. J.G. Ryan and R. A. Goubran, "Optimal nearfield response for microphone arrays", in *Proc. IEEE Workshop Applicat. Signal Processing to Audio Acoust.*, New Paltz, NY, USA, Oct 1997.
15. E. N. Gilbert and S. P. Morgan, "Optimum design of directive antenna arrays subject to random variations", *Bell Syst. Tech. J.*, pp. 637–663, May 1955.
16. L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. Antennas Propagat.*, vol. 30, pp. 27–34, 1982.

17. J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques", *J. Acoust. Soc. Amer.*, vol. 99, pp. 3138–3148, May 1996.
18. R. A. Kennedy, T. Abhayapala, D. B. Ward, and R. C. Williamson, "Nearfield broadband frequency invariant beamforming", in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP-96)*, pp. 905–908, April 1996.
19. R. N. Marshall and W. R. Harry, "A new microphone providing uniform directivity over an extended frequency range", *J. Acoust. Soc. Amer.*, vol. 12, pp. 481–497, 1941.
20. R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*, John Wiley and Sons, New York, 1980.
21. S. Nordholm, I. Claesson, and P. Eriksson, "The broad-band Wiener solution for Griffith-Jim beamformers", *IEEE Trans. Signal Processing*, vol. 40, pp. 474–478, Feb 1992.
22. W. Taeger, "Near field superdirectivity (NFSD)", in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP-98)*, Seattle, WA, USA, 1998.
23. D. B. Ward and G. W. Elko, "Mixed nearfield/farfield beamforming: A new technique for speech acqusition in a reverberant environment", in *Proc. IEEE Workshop Applicat. Signal Processing to Audio Acoust.*, New Paltz, NY, USA, Oct 1997.
24. B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Englewood Cliffs, 1985.

3 Post-Filtering Techniques

K. Uwe Simmer[1], Joerg Bitzer[2], and Claude Marro[3]

[1] Aureca GmbH, Bremen, Germany
[2] Houpert Digital Audio, Bremen, Germany
[3] France Télécom R&D, Lannion, France

Abstract. In the context of microphone arrays, the term post-filtering denotes the post-processing of the array output by a single-channel noise suppression filter. A theoretical analysis shows that Wiener post-filtering of the output of an optimum distortionless beamformer provides a minimum mean squared error solution. We examine published methods for post-filter estimation and develop a new algorithm. A simulation system is presented to compare the performance of the discussed algorithms.

3.1 Introduction

What can be gained by additional post-filtering if the Minimum Variance Distortionless Response (MVDR) beamformer already provides the optimum solution for a given sound field?

Assuming that signal and noise are mutually uncorrelated the MVDR beamformer minimizes the noise power (or variance) subject to the constraint of a distortionless look direction response. The solution can be shown to be optimum in the Maximum Likelihood (ML) sense and produces the best possible Signal to Noise Ratio (SNR) for a narrowband input [1]. However, it does not maximize the SNR for a broadband input such as speech. Furthermore, the MVDR beamformer does not provide a broadband Minimum Mean Squared Error (MMSE) solution. The best possible linear filter in the MMSE sense is the multi-channel Wiener filter. As shown below the broadband multi-channel MMSE solution can be factorized into a MVDR beamformer followed by a single-channel Wiener post-filter. The multi-channel Wiener filter generally produces a higher output SNR than the MVDR filter. Therefore, additional post-filtering can significantly improve the SNR, which motivates this chapter.

The squared error minimized by the single-channel Wiener filter is the sum of residual noise and signal distortion components at the output of the filter. As a result, linear distortion of the desired signal cannot be avoided entirely if Wiener filtering is used. Additional Wiener filtering is advantageous in practice, however, because signal distortions can be masked by residual noise and a compromise between signal distortion and noise suppression can be found. Using MVDR beamforming alone often does not provide sufficient noise reduction due to its limited ability to reduce diffuse noise and reverberation.

The first concept of an electronic multi-microphone device to suppress diffuse reverberation was proposed by Danilenko in 1968 [2]. His research was motivated by Békésy's [3] observation that human listeners are able to suppress reverberation if sounds are presented binaurally. In Danilenko's reverberation suppressor a main microphone signal is multiplied by a broadband gain factor that is equal to the ratio of short-time cross-correlation and energy measurements. Two auxiliary microphones were used to measure correlation and energy. Danilenko already noted that such a system would also suppress incoherent acoustic noise. However, the proposed analog, electronic tube version of this system was not realized at that time. Another proposal in [2] was to evaluate squared sum and differences of two microphone signals, an idea that later was developed independently by Gierl and others in the context of digital multi-channel spectral subtraction algorithms [4], [5], [6], [7], [8].

According to Danilenko, his correlation-based concept was first realized during Blauert's stay at Bell Labs. In [9], Allen *et al.* presented a digital, two-microphone algorithm for dereverberation based on short-term Fourier-Transform and the overlap-add method. In 1984, Kaneda and Tohyama extended the application of the correlation based post-filters to noise reduction [10]. The first multi-microphone solution was published by Zelinski [11], [12]. Simmer and Wasiljeff showed that Zelinski's approach does not provide an optimum solution in the Wiener sense if the noise is spatially uncorrelated, and developed a slightly modified version [13]. A deeper analysis of the Zelinski and the Simmer post-filter can be found in [14], [15].

In the last decade, several new combinations and extensions of the post-filter approach were published. Le-Bouquin and Faucon used the coherence function as a post-filter [16], [17] and extended their system by a coherence subtraction method to overcome the problem of insufficient noise reduction at low frequencies [18], [19]. The problem of time delay estimation and further improvement of the estimation of the transfer function was independently addressed by Kuczynski *et al.* [20], [21] and Drews *et al.* [22], [23]. Fischer and Simmer gave a first solution by associating a post-filter and a generalized sidelobe canceler (GSC) to improve the noise reduction in case the noise field is dominated by coherent sources [24], [25]. Another system for the same task was introduced by Hussain *et al.* [26] and was based on switching between algorithms. The same strategy of switching between different algorithms, where the decision is based on the coherence between the sensors, can be found in [27], [28]. Furthermore, Mamhoudi and Drygajlo used the wavelet-transform in combination with different post-filters to improve the performance [29], [30]. Bitzer *et al.* [31], [32] proposed a solution with a super-directive array and McCowan *et al.* used a near-field super-directive approach [33].

Reading these papers we find that a theoretical basis for post-filtering seems to be missing. Therefore, an analysis based on optimum MMSE multi-channel filtering is presented in the following section.

3.2 Multi-channel Wiener Filtering in Subbands

We use matrix notation for a compact derivation. Signal vector $\mathbf{x}$ and weight vector $\mathbf{w}$ denote the multi-channel signal at the output of the N microphones and the multi-channel beamformer coefficients, respectively. We assume that the input signal vector $\mathbf{x}(k)$ is decomposed into M complex subband signals $\mathbf{x}(k,i)$ by means of an analysis filter-bank, where k is the discrete time index and i is the subband index. The optimum weight vector $\mathbf{w}_{opt}(k,i)$ for transforming the input signal vector $\mathbf{x}(k,i) = \mathbf{s}(k,i) + \mathbf{v}(k,i)$ corrupted by additive noise $\mathbf{v}(k,i)$ into the best possible MMSE approximation of the desired scalar signal $s(k,i)$ is referred to as multi-channel Wiener filter [34]. We assume that the relation between the desired scalar signal $s(k,i)$ and the signal vector $\mathbf{s}(k,i)$ is linear and that the N elements of the column vectors $\mathbf{s}(k,i)$ and $\mathbf{v}(k,i)$ are random processes. In the following, ^{T} denotes transposition, * denotes complex conjugation, ^{H} denotes Hermitian transposition, and $E[\cdot]$ denotes the statistical expectation operator.

3.2.1 Derivation of the Optimum Solution

The error in subband i for an arbitrary weight vector $\mathbf{w}(k,i)$ is defined as the difference of the filter output

$$y(k,i) = \mathbf{w}^H(k,i)\mathbf{x}(k,i) = \mathbf{w}^H(k,i)\left[\mathbf{s}(k,i) + \mathbf{v}(k,i)\right] \tag{3.1}$$

and the scalar desired signal $s(k,i)$, that is

$$e(k,i) = s(k,i) - \mathbf{w}^H(k,i)\mathbf{x}(k,i). \tag{3.2}$$

Using the definitions for the power of a complex signal

$$\phi_{xx}(k,i) = E\left[x(k,i)x(k,i)^*\right], \tag{3.3}$$

the cross-correlation vector

$$\boldsymbol{\phi}_{xy}(k,i) = E\left[\mathbf{x}(k,i)y^*(k,i)\right], \tag{3.4}$$

and the correlation matrix

$$\boldsymbol{\Phi}_{xx}(k,i) = E\left[\mathbf{x}(k,i)\mathbf{x}^H(k,i)\right], \tag{3.5}$$

the squared error at time k may be written as

$$\begin{aligned}\phi_{ee}(i) &= E\left[\{s(i) - \mathbf{w}^H(i)\mathbf{x}(i)\}\{s^*(i) - \mathbf{x}^H(i)\mathbf{w}(i)\}\right] \\ &= \phi_{ss}(i) - \mathbf{w}^H(i)\boldsymbol{\phi}_{xs}(i) - \boldsymbol{\phi}_{xs}^H(i)\mathbf{w}(i) + \mathbf{w}^H(i)\boldsymbol{\Phi}_{xx}(i)\mathbf{w}(i),\end{aligned} \tag{3.6}$$

where the time index k has been omitted without loss of generality. The optimum solution minimizes the sum of all error powers $\phi_{ee}(i)$:

$$\sum_{i=0}^{M}\left[\phi_{ss}(i) - \mathbf{w}^H(i)\boldsymbol{\phi}_{xs}(i) - \boldsymbol{\phi}_{xs}^H(i)\mathbf{w}(i) + \mathbf{w}^H(i)\boldsymbol{\Phi}_{xx}(i)\mathbf{w}(i)\right]. \tag{3.7}$$

Since the error power is necessarily real-valued and nonnegative for all subbands, the sum can be minimized for the weight vector $\mathbf{w}(i)$ by minimizing the error power $\phi_{ee}(i)$ for each subband. Therefore, the frequency index i may also be omitted without loss of generality.

The power ϕ_{ee} is a quadratic function of $\mathbf{w}$ and therefore has a single, global minimum. The optimum weight vector minimizing the squared error is obtained by setting the gradient of ϕ_{ee} with respect to $\mathbf{w}$ equal to the null vector [35]:

$$\nabla_{\mathbf{w}}(\phi_{ee}) = 2\frac{\partial \phi_{ee}}{\partial \mathbf{w}^*} = -2\boldsymbol{\phi}_{xs} + 2\boldsymbol{\Phi}_{xx}\mathbf{w} = \mathbf{0}. \tag{3.8}$$

The resulting expression is the subband version of the multi-channel Wiener-Hopf equation in its most general form

$$\boldsymbol{\Phi}_{xx}\mathbf{w}_{opt} = \boldsymbol{\phi}_{xs}, \tag{3.9}$$

where $\boldsymbol{\Phi}_{xx}$ is the correlation matrix of the noisy input vector and $\boldsymbol{\phi}_{xs}$ is the cross-correlation vector between the noisy input vector and the desired scalar signal s. Assuming $\boldsymbol{\Phi}_{xx}$ to be nonsingular, we may solve (3.9) for the optimum weight vector:

$$\mathbf{w}_{opt} = {\boldsymbol{\Phi}_{xx}}^{-1}\boldsymbol{\phi}_{xs}. \tag{3.10}$$

3.2.2 Factorization of the Wiener Solution

In our application, the received signal is assumed to consist of a single desired scalar signal that is transformed by the acoustic path $\mathbf{d}$ and additive noise:

$$\mathbf{x} = s\mathbf{d} + \mathbf{v}. \tag{3.11}$$

The noise vector $\mathbf{v}$ is given by

$$\mathbf{v} = [v_0, v_1, \cdots, v_{N-1}]^T \tag{3.12}$$

where v_n is a complex noise signal in subband i at microphone n. The complex propagation vector is

$$\mathbf{d} = [d_0, d_1, \cdots, d_{N-1}]^T \tag{3.13}$$

where d_n describes the acoustic path from the desired source to the microphone n for subband i. The propagation vector $\mathbf{d}$ may include time delays, near-field effects, and the transfer functions of enclosure and microphones. With the definitions (3.3), (3.4), (3.5) and assuming that signal and noise are uncorrelated, the cross-correlation vector may be reduced to

$$\boldsymbol{\phi}_{xs} = \phi_{ss}\mathbf{d} \tag{3.14}$$

and the correlation matrix may be expressed as

$$\boldsymbol{\Phi}_{xx} = \phi_{ss}\mathbf{d}\mathbf{d}^H + \boldsymbol{\Phi}_{vv}. \tag{3.15}$$

Consequently, the optimum weight vector may be written as

$$\mathbf{w}_{opt} = {\boldsymbol{\Phi}_{xx}}^{-1}\phi_{ss}\mathbf{d} = \left[\phi_{ss}\mathbf{d}\mathbf{d}^H + \boldsymbol{\Phi}_{vv}\right]^{-1}\phi_{ss}\mathbf{d}. \tag{3.16}$$

The multi-channel Wiener filter can now be factorized into an array processor and a single channel post-filter by applying the Sherman-Morrison-Woodbury formula

$$\left[\mathbf{A}^{-1} + \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^H\right]^{-1} \equiv \mathbf{A} - \mathbf{A}\mathbf{B}(\mathbf{C} + \mathbf{B}^H\mathbf{A}\mathbf{B})^{-1}\mathbf{B}^H\mathbf{A} \tag{3.17}$$

which is also known as the matrix inversion lemma [35]. Substituting

$$\mathbf{A} = {\boldsymbol{\Phi}_{vv}}^{-1}, \quad \mathbf{B} = \sqrt{\phi_{ss}}\mathbf{d}, \quad \text{and} \quad \mathbf{C} = 1 \tag{3.18}$$

into (3.17), and taking into account that the Hermitian form $\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}$ is scalar and real valued, the MMSE solution (3.16) can be transformed into

$$\begin{aligned}
\mathbf{w}_{opt} &= \left[{\boldsymbol{\Phi}_{vv}}^{-1} - \frac{\phi_{ss}{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}}{1 + \phi_{ss}\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}}\right]\phi_{ss}\mathbf{d} \\
&= \left[1 - \frac{\phi_{ss}\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}}{1 + \phi_{ss}\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}}\right]\phi_{ss}{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d} \\
&= \left[\frac{\phi_{ss}}{1 + \phi_{ss}\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}}\right]{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d} \\
&= \left[\frac{\phi_{ss}}{\phi_{ss} + \left(\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}\right)^{-1}}\right]\frac{{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}}{\mathbf{d}^H{\boldsymbol{\Phi}_{vv}}^{-1}\mathbf{d}}.
\end{aligned} \tag{3.19}$$

Equation (3.19) shows that the multi-channel Wiener filter (3.10) can be written as the product of the weight vector of the MVDR beamformer, (see Chapter 2) and a real-valued scalar factor. A similar result is used in [36] and [1] to show that the multi-channel Wiener and the MVDR solution yield the same SNR if the input is narrowband. In this case the MVDR beamformer is preferable since it is data independent (i.e. completely defined by the spatial configuration of signal and noise sources), whereas the Wiener solution is data dependent (ϕ_{ss} must be known or estimated) and is therefore much more difficult to handle. However, MVDR and Wiener solutions yield the same SNR only if the input consists of a single frequency. For the broadband case (which has already been discussed in [37]), the scalar factor becomes a subband or frequency domain post-filter that may significantly improve the SNR.

To show that the optimum post-filter is also a Wiener filter that operates on the single-channel output data, we evaluate the power of the desired signal at the output of the MVDR processor as

$$\phi_{s_o s_o} = \phi_{ss} \mathbf{w}_{\text{mvdr}}^H \mathbf{d}\mathbf{d}^H \mathbf{w}_{\text{mvdr}} = \phi_{ss} \left| \frac{\mathbf{d}^H \boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d}}{\mathbf{d}^H \boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d}} \right|^2 = \phi_{ss}. \tag{3.20}$$

This demonstrates the distortionless magnitude response. Furthermore, we determine the power of the output noise as

$$\phi_{v_o v_o} = \mathbf{w}_{\text{mvdr}}^H \boldsymbol{\Phi}_{vv} \mathbf{w}_{\text{mvdr}} = \frac{\mathbf{d}^H \boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d}}{(\mathbf{d}^H \boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d})^2} = \frac{1}{\mathbf{d}^H \boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d}}. \tag{3.21}$$

Substituting (3.20) and (3.21) into (3.19), we can finally factorize the optimum MMSE solution into the following expression:

$$\mathbf{w}_{\text{opt}} = \underbrace{\left[\frac{\phi_{s_o s_o}}{\phi_{s_o s_o} + \phi_{v_o v_o}} \right]}_{\text{Wiener post-filter}} \underbrace{\frac{\boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d}}{\mathbf{d}^H \boldsymbol{\Phi}_{vv}{}^{-1} \mathbf{d}}}_{\text{MVDR array}}. \tag{3.22}$$

Equation (3.22) includes the complex weight vector of the MVDR beamformer

$$\mathbf{w}_{\text{mvdr}}(k,i) = \frac{\boldsymbol{\Phi}_{vv}^{-1}(k,i)\, \mathbf{d}(k,i)}{\mathbf{d}^H(k,i)\, \boldsymbol{\Phi}_{vv}^{-1}(k,i)\, \mathbf{d}(k,i)}, \tag{3.23}$$

and the scalar, single channel Wiener post-filter that depends on the SNR at the output of the beamformer:

$$H_{\text{post}}(k,i) = \frac{\phi_{s_o s_o}(k,i)}{\phi_{s_o s_o}(k,i) + \phi_{v_o v_o}(k,i)} = \frac{SNR_{out}(k,i)}{1 + SNR_{out}(k,i)}. \tag{3.24}$$

The output signal $z(k,i)$ of the factorized MMSE filter is the product of the output signal $y(k,i)$ of the MVDR array:

$$y(k,i) = \mathbf{w}_{\text{mvdr}}^H(k,i)\, \mathbf{x}(k,i), \tag{3.25}$$

and the transfer function $H_{\text{post}}(k,i)$ of a single-channel post-filter:

$$z(k,i) = y(k,i)\, H_{\text{post}}(k,i). \tag{3.26}$$

The MVDR solution (3.23) maximizes the directivity index if $\boldsymbol{\Phi}_{vv}$ equals the correlation matrix of the diffuse sound field. The resulting system may therefore be called 'superdirective array with Wiener post-filter' (although the term superdirectivity originated in the context of analog microphones). Since the definition (3.13) of the propagation vector does not include any far-field assumptions, (3.23) may also be used to design a near-field superdirective array.

3.2.3 Interpretation

Although the above results are clearly related to Wiener's work on optimum filtering [38], some basic assumptions were different. First of all, Wiener considered continuous time signals which leads to the Wiener-Hopf integral equation. The corresponding equation in matrix form (3.10) usually determines the filter coefficients for an optimum discrete time FIR filter of order N. In our case, the delay line is defined by the spatial arrangement of the acoustic sensor and the taps are realized by the N microphones. The array and the weight vector form a spatial filter. Wiener assumed that signal and noise are ergodic and stationary random processes and he used the Fourier-transform to find a solution for the optimum filter. This leads to a linear, time invariant filter. Such a filter is not appropriate for speech signals that may be modeled as short-time stationary processes only. The derivation used here is based on ensemble averages (expectations) and does not assume stationarity. In practice, however, only an approximate realization of such a filter is possible.

There are two main sources of errors: the analysis and synthesis filter-bank, and the procedures to estimate the time-varying signal and noise powers in the individual subbands. For the design of the filter-banks, a compromise between frequency and time resolution has to be made. High resolution in the frequency domain leads to poor resolution in the time domain and vice versa. Therefore, the highest possible frequency resolution that does not violate the short-term stationarity of speech should be chosen. Furthermore, the minimum error in the time-domain is only reached if the filters have non-overlapping frequency regions (see the discussion of subband methods in [39]). Since such filters are physically unrealizable, overlapping of subbands cannot be avoided. As a result, the suppression of a noise-only subband may affect adjacent subbands containing desired signal components. In the following, we will use windowing, Fast Fourier Transform (FFT) and the overlap-add method to implement the filter-bank. However, (3.22) is general enough to allow any complex or real valued filter-bank method. If overlap-add is used, circular convolution should be avoided by zero padding and by constraints imposed on the estimated transfer function.

In the derivation of the optimum filter, expectations are used to estimate the parameters. This is a theoretical construction since the ensemble averages cannot be computed in practice. An approximation proposed in [9] is the recursive Welsh periodogram:

$$\hat{\phi}_{xy}(k,i) = \alpha\, \hat{\phi}_{xy}(k-1,i) + (1-\alpha)x(k,i)y^*(k,i), \tag{3.27}$$

where $\alpha = \exp(-D/[\tau_\alpha f_s])$ is defined by the decimation factor D of the filter-bank, the time-constant τ_α (ms), and the sampling frequency f_s (kHz). The time constant is again a compromise. If τ_α is low, artifacts may occur due to the variation of the transfer function estimate. On the other hand, if a high time constant τ_α is chosen, the assumption of short time stationarity is violated and the output speech signal may sound reverberant.

Unfortunately, the factorized result (3.22) does not give any indication of how the Wiener post-filter could be estimated. A possible solution, which we discuss in the next section, is based on the observation that the correlation between two microphone signals is low if the sound field is diffuse and the microphone distance is large enough.

3.3 Algorithms for Post-Filter Estimation

Figure 3.1 shows the block diagram of the studied algorithms. The microphone signals are time aligned and decomposed by a frequency subband transform (FT). The coefficients w_n represent the weight vector $\mathbf{w}$ of the beamformer and H represents the post-filter. The inverse subband transform (IFT) synthesizes the output signal. The coefficients f_n for post-filter estimation form a vector $\mathbf{f}$. Unless otherwise noted we assume that $\mathbf{f} = \mathbf{w}$. We begin

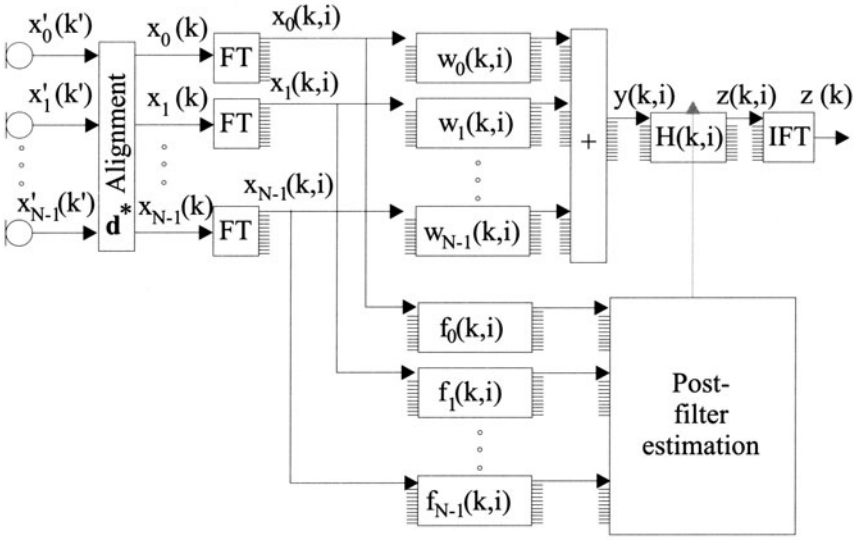


Fig. 3.1. General block diagram of the examined post-filters.

our analysis on multi-microphone post-filters by recalling some results on the performance of arrays from Chapter 2 since these results are needed later. We generally assume that the coefficients are normalized so that $\mathbf{w}^H\mathbf{1}\mathbf{1}^H\mathbf{w} = 1$ and $\mathbf{f}^H\mathbf{1}\mathbf{1}^H\mathbf{f} = 1$, where $\mathbf{1}$ is the N-vector of ones. Therefore, the array gain equals the noise reduction of the array. For convenience, we define a noise power attenuation factor that equals the inverse of the array gain:

$$A_\Gamma = \mathbf{w}^H\Gamma_{vv}\mathbf{w} = G^{-1}, \tag{3.28}$$

where the coherence matrix Γ_{vv} is the normalized noise correlation matrix $\Gamma_{vv} = \Phi_{vv}N/\text{trace}[\Phi_{vv}]$, and all quantities are assumed to be frequency dependent.

An examination of (3.28) shows that the noise attenuation of the array is the weighted sum of the complex coherence functions of all sensor pairs. Thus, all products appear in conjugate pairs $\Gamma_{mn} + \Gamma_{nm} = 2Re\{\Gamma_{nm}\}$. As a result, the noise reduction of the array is actually a function of the real part of the complex coherence between the sensors. The knowledge of the magnitude squared coherence is not sufficient.

The white noise gain is the array gain for spatially uncorrelated noise, where $\Gamma_{vv} = \mathbf{I}$. Thus, the attenuation factor for spatially white noise is

$$A_{\mathbf{I}} = \mathbf{w}^H\mathbf{w} = \text{WNG}^{-1}. \tag{3.29}$$

The additional noise attenuation of the post-filter is given by

$$A_{\text{post}} = |H_{\text{post}}|^2. \tag{3.30}$$

The total noise attenuation of the combined system is the product of the attenuation of the array and the attenuation of the post-filter, or the respective sum in dB:

$$A_{total}\Big|_{\text{dB}} = 10\log_{10}(A_\Gamma) + 10\log_{10}(A_{\text{post}}). \tag{3.31}$$

3.3.1 Analysis of Post-Filter Algorithms

The first method for post-filter estimation we study is a generalized version of Zelinski's algorithms that was discussed by Marro *et al.* [15]. It covers several other algorithms as a special case.

$$H_{\text{zm}}(i) = \frac{Re\left\{\sum_{n=0}^{N-2}\sum_{m=n+1}^{N-1} w_n(i)w_m^*(i)\Phi_{x_nx_m}(i)\right\}\sum_{n=0}^{N-1}|w_n(i)|^2}{Re\left\{\sum_{n=0}^{N-2}\sum_{m=n+1}^{N-1} w_n(i)w_m^*(i)\right\}\sum_{n=0}^{N-1}|w_n(i)|^2\Phi_{x_nx_n}(i)} \tag{3.32}$$

Equation (3.32) includes Danilenko's [2] idea to use the ratio of cross-correlation $\Phi_{x_nx_m}$ and power $\Phi_{x_nx_n}$ for suppressing incoherent noise, the complex subband approach of Allen *et al.* [9], Zelinski's proposal to average over all microphone pairs $m > n$ [11], and Marro's [40] extension to complex shading coefficients w_n. To write this algorithm in matrix notation, we note that

$$2Re\left\{\sum_{n=0}^{N-2}\sum_{m=n+1}^{N-1} w_nw_m^*\Phi_{x_nx_m}\right\} = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} w_nw_m^*\Phi_{x_nx_m} - \sum_{n=0}^{N-1} w_nw_n^*\Phi_{x_nx_n}.$$

This is a Hermitian form of the shading coefficients w_n and the correlation matrix Φ_{xx}, minus the weighted sum of diagonal elements of Φ_{xx}. The algorithm (3.32) requires that the relative time-delay differences and gain ratios between the microphone signals have been compensated in advance so that $\mathbf{d} = \mathbf{1}$. This leads to a modified noise correlation matrix Φ_{xx} (see Chapter 2). The transfer function of the post-filter (3.32) may now conveniently be written in matrix form as

$$H_{\mathrm{zm}} = \frac{\left(\mathbf{w}^H \Phi_{xx} \mathbf{w} - \mathbf{w}^H \Phi_{xx}^D \mathbf{w}\right) \mathbf{w}^H \mathbf{w}}{\left(\mathbf{w}^H \mathbf{1}\mathbf{1}^H \mathbf{w} - \mathbf{w}^H \mathbf{w}\right) \mathbf{w}^H \Phi_{xx}^D \mathbf{w}}, \tag{3.33}$$

where Φ_{xx}^D is a diagonal matrix of the diagonal elements of Φ_{xx}. If the sound field is homogeneous, we have the same input power at each microphone, i.e. $\Phi_{xx}^D = \phi_{xx}\mathbf{I}$, and may write

$$H_{\mathrm{zm}} = \frac{\left(\mathbf{w}^H \Phi_{xx} \mathbf{w} - \phi_{xx} \mathbf{w}^H \mathbf{w}\right)}{\phi_{xx} \left(\mathbf{w}^H \mathbf{1}\mathbf{1}^H \mathbf{w} - \mathbf{w}^H \mathbf{w}\right)}. \tag{3.34}$$

If signal and noise are uncorrelated we have $\Phi_{xx} = \Phi_{ss} + \Phi_{vv}$. Therefore,

$$H_{\mathrm{zm}} = \frac{\left(\mathbf{w}^H \Phi_{ss} \mathbf{w} - \phi_{ss} \mathbf{w}^H \mathbf{w}\right) + \left(\mathbf{w}^H \Phi_{vv} \mathbf{w} - \phi_{vv} \mathbf{w}^H \mathbf{w}\right)}{\left(\phi_{ss} + \phi_{vv}\right) \left(\mathbf{w}^H \mathbf{1}\mathbf{1}^H \mathbf{w} - \mathbf{w}^H \mathbf{w}\right)}. \tag{3.35}$$

Assuming that the coefficients are normalized such that $\mathbf{w}^H \mathbf{1}\mathbf{1}^H \mathbf{w} = 1$, the desired signal is coherent, i.e., $\Phi_{ss} = \phi_{ss}\mathbf{1}\mathbf{1}^H$. With the noise correlation matrix being $\Phi_{vv} = \phi_{vv}\Gamma_{vv}$, where $\phi_{vv} = \mathrm{trace}\left[\Phi_{vv}\right]/N$, we finally obtain

$$H_{\mathrm{zm}} = \frac{\phi_{ss}}{\phi_{ss} + \phi_{vv}} + \frac{\phi_{vv} \left(\mathbf{w}^H \Gamma_{vv} \mathbf{w} - \mathbf{w}^H \mathbf{w}\right)}{\left(\phi_{ss} + \phi_{vv}\right) \left(1 - \mathbf{w}^H \mathbf{w}\right)}. \tag{3.36}$$

Although the designs of the MVDR array and the post-filter estimation algorithm do not seem to have much in common, the transfer function of the post-filter may be expressed as a function of the attenuation factors of the array by substituting (3.28) and (3.29) into (3.36):

$$H_{\mathrm{zm}} = \frac{\phi_{ss}}{\phi_{ss} + \phi_{vv}} + \frac{\phi_{vv} \left(A_\Gamma - A_{\mathbf{I}}\right)}{\left(\phi_{ss} + \phi_{vv}\right) \left(1 - A_{\mathbf{I}}\right)}. \tag{3.37}$$

This is also true for the slightly modified version of Zelinski's algorithm [13]:

$$H_{\mathrm{sm}}(i) = \frac{Re\left\{\sum_{n=0}^{N-2} \sum_{m=n+1}^{N-1} w_n(i) w_m^*(i) \Phi_{x_n x_m}(i)\right\}}{Re\left\{\sum_{n=0}^{N-2} \sum_{m=n+1}^{N-1} w_n(i) w_m^*(i)\right\} \phi_{yy}(i)}, \tag{3.38}$$

where $\phi_{yy} = \phi_{ss} + \phi_{vv} A_\Gamma$ is the output power of the array. The modified post-filter can be expressed as

$$H_{\mathrm{sm}} = \frac{\phi_{ss}}{\phi_{ss} + \phi_{vv} A_\Gamma} + \frac{\phi_{vv} A_\Gamma \left(A_\Gamma - A_{\mathrm{I}}\right)}{\left(\phi_{ss} + \phi_{vv} A_\Gamma\right)\left(1 - A_{\mathrm{I}}\right)}. \tag{3.39}$$

These rather surprising results were first derived in [15]. They are used in the following section to discuss the properties of a large class of post-filtering algorithms.

3.3.2 Properties of Post-Filter Algorithms

First of all, we note that the shading coefficients w_n form a weight vector $\mathbf{w}$ that generally can be computed by using the design rule of the MVDR array. It is not necessary, however, to use the same design for array processor and post-filter (see Fig. 3.1). Both the MVDR weight vector and the array gain are functions of the noise correlation matrix. It should be noted that the correlation matrix that is used for the design may differ from the correlation matrix of the environment in which the array operates. Therefore, three different correlation matrices may be involved: a first one for the design of the array processor, a second one for the design of the post-filter, and a third one to determine the performance in the actual environment.

Analyzing (3.37) and (3.39) leads to the following conclusions:

- Optimum performance is only reached if $A_\Gamma = A_{\mathrm{I}}$:
  The difference of the two attenuation factors is zero only if the noise is spatially uncorrelated which was Danilenko's initial assumption in the design of his suppression system. In this case, (3.37) becomes a Wiener filter for the input signal of the beamformer. On the other hand, (3.39) becomes a Wiener filter for the beamformer output and therefore represents the MMSE solution for uncorrelated noise if the delay and sum beamformer is used. All other coefficient sets, including superdirective solutions, yield suboptimal performance. In a diffuse sound field, the noise is correlated at low frequencies which leads to poor performance for low frequency noise.
- Negative post-filter if $A_\Gamma < A_{\mathrm{I}}$:
  In a diffuse noise field, or if coherent sources are present, the difference of the attenuation factors $(A_\Gamma - A_{\mathrm{I}})$ may cause a negative transfer-function. If negative parts of the transfer functions are set to zero, which is a common strategy, signal cancellation may occur.
- Infinite post-filter if $A_{\mathrm{I}} = 1$:
  This is usually the case with superdirective designs which amplify uncorrelated noise at low frequencies.

To demonstrate the preceding results, we computed the theoretical performance of a four microphone end-fire array with 8 cm inter-microphone distance in a diffuse noise field ($\phi_{ss} = 0$). Figure 3.2 shows the attenuation

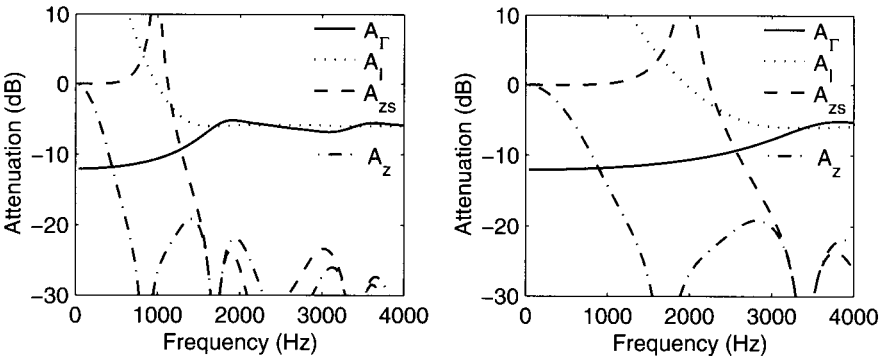


Fig. 3.2. Theoretical noise attenuation of an end-fire array for a diffuse noise field. Left: delay and sum beamformer coefficients. Right: superdirective coefficients.

factors A_Γ and A_I of the beamformer and the noise attenuation A_post of the post-filter (3.37). The left part depicts the attenuation for delay and sum beamformer coefficients ($\mathbf{f} = \mathbf{w} = \mathbf{1}/N$) and the right part depicts the attenuation for superdirective coefficients ($\mathbf{f} = \mathbf{w}_\mathrm{MVDR}$).

The performance of the delay and sum beamformer and the respective post-filter is poor at low frequencies. At high frequencies the coherence of a diffuse noise field is approaching zero. Therefore, A_Γ is close to A_I and both post-filters perform nearly optimally.

The superdirective beamformer performs particularly well at low frequencies. The respective post-filter, however, does not benefit from using superdirective coefficients. The performance gets even worse at low frequencies and the transfer function is infinite at the frequency where A_I crosses 0 dB.

3.3.3 A New Post-Filter Algorithm

To derive an improved algorithm we note that in all cases the subtraction of the white noise attenuation A_I in (3.37) is causing the trouble. It reduces the performance for superdirective coefficients and is responsible for negative or infinite post-filters. Our straightforward approach for solving these problems is to replace the difference $A_\Gamma - A_\mathrm{I}$ with A_Γ, since A_Γ is the parameter that is actually minimized by the design of the MVDR beamformer. Substituting $A_\mathrm{I} = 0$ in (3.37) results in

$$H_\mathrm{apab} = \frac{\phi_{ss}}{\phi_{ss} + \phi_{vv}} + \frac{\phi_{vv} A_\Gamma}{\phi_{ss} + \phi_{vv}} = \frac{\phi_{yy}}{\phi_{xx}}. \tag{3.40}$$

This new algorithm can be implemented easily by estimating the ratio of the output power ϕ_{yy} and the input power ϕ_{xx} of the beamformer for all subbands, where ϕ_{xx} is the power of the microphone closest to the desired source

or, alternatively, the average input power of the beamformer (see Fig. 3.3). This design is compatible with superdirective coefficients, is always positive, and provides good performance for low frequency noise. However, the new transfer function still approximates a Wiener filter for the input signal. It does not take into account that the noise has already been reduced by the MVDR beamformer. In order to correct this behavior, we may apply the following function to (3.40)

$$g(H, A) = \frac{H}{H + (1 - H) A}. \tag{3.41}$$

This transforms the Wiener filter for the input to a Wiener filter for the output of the beamformer:

$$g\left(\frac{\phi_{ss}}{\phi_{ss} + \phi_{vv}}, A_\Gamma\right) = \frac{\phi_{ss}}{\phi_{ss} + \phi_{vv} A_\Gamma}. \tag{3.42}$$

Since A_Γ is usually unknown, we may implement (3.40) directly and call this algorithm Adaptive Post-Filter for an Arbitrary Beamformer (APAB).

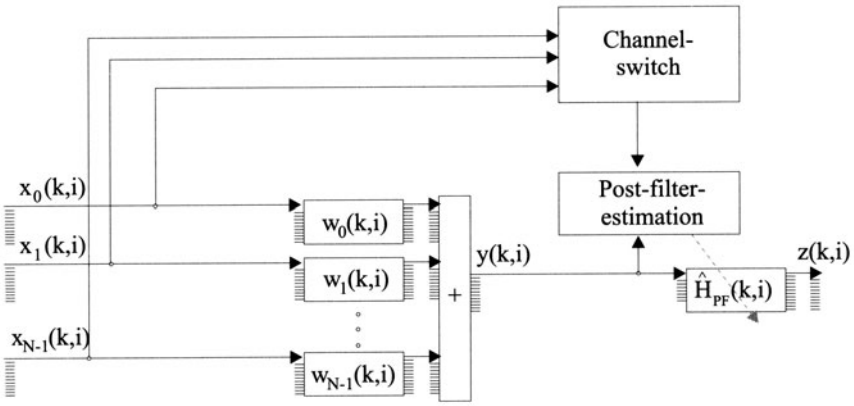


Fig. 3.3. Block diagram of the adaptive post-filter for an arbitrary beamformer (APAB).

3.4 Performance Evaluation

It is difficult to obtain reliable speech quality measures for the performance evaluation of noise reduction units. Subjective listening tests reach statistical significance only for a large number of trained listeners and are expensive and time-consuming. On the other hand, objective measures are often less

sensitive than the human auditory system to artifacts such as musical tones. Therefore, we did not rely exclusively on objective measures to optimize the noise reduction algorithms. Accompanying informal listening tests were conducted to validate the objective results.

3.4.1 Simulation System

Our simulation system consists of three parts: A signal generation module, the device or algorithm under test (DUT), and the evaluation unit. In a first step, clean speech $s(k)$ and a pure noise signal $v(k)$ are convolved with room impulse responses (RIR) that are computed using the image method of Allen and Berkley [41]. In Fig. 3.4, we show the room configuration used. Noise is added to the computed multi-channel signals to produce a given signal-to-noise ratio (SNR). The resulting noisy signal is fed into the DUT.

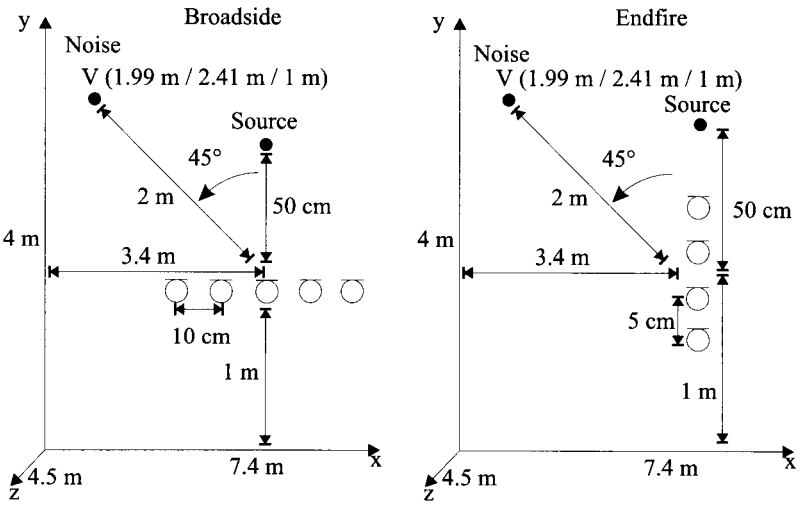


Fig. 3.4. Configuration of the simulated room.

The adaptive coefficients of the algorithm are copied to two slave algorithms which process speech or noise only. Thus, we have access to the processed speech signal $y_s(k)$, the processed noise signal $y_v(k)$, and a processed sum $y_{s+v}(k)$. Finally, these three output signals and the input signals are used in the evaluation unit to compute several speech quality measures. See Fig. 3.5 for a graphical description of the complete system.

3.4.2 Objective Measures

We are using three different quantities to obtain objective information about the tested algorithm. The first one is the segmental signal-to-noise ratio en-

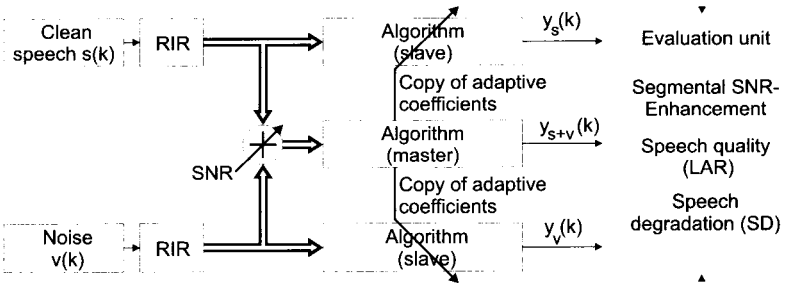


Fig. 3.5. Graphical description of the complete simulation system.

hancement (SNRE):

$$SNRE(l) = SNR_{in}(l) - SNR_{out}(l). \tag{3.43}$$

The segmental SNR is computed from consecutive samples with block-length $B = 256$ at a sampling frequency of 8 kHz:

$$SNR_{in}(l) = 10 \cdot \log_{10} \frac{\sum_{k=lB+1}^{(l+1)B} s^2(k)}{\sum_{k=lB+1}^{(l+1)B} v^2(k)}, \tag{3.44}$$

$$SNR_{out}(l) = 10 \cdot \log_{10} \frac{\sum_{k=lB+1}^{(l+1)B} y_s^2(k)}{\sum_{k=lB+1}^{(l+1)B} y_v^2(k)}. \tag{3.45}$$

The second objective measure is the log-area-ratio distance (LAR) which has been tested with good results in [42]. This quantity can be computed in three steps:

1. Estimate the PARtial CORrelation coefficients (PARCOR) of a block of samples. The block-size should be small enough to hold the assumption of stationarity but large enough to reduce bias and variance of the estimated values. A good choice is a block-size of 256 for a model order of $P = 12$. An algorithm for estimating PARCOR coefficients is the well-known Burg-algorithm [35].
2. Determine the area-coefficients by

$$g(p,l) = \frac{1 + k(p,l)}{1 - k(p,l)} \quad \forall \quad 1 \leq p \leq 12 \tag{3.46}$$

where $k(p,l)$ is the pth PARCOR coefficient of block l.

3. Compute the LAR of block l

$$LAR(l) = \sum_{p=1}^{12} 20\log_{10}\left|\frac{g_s(p,l)}{g_{y_{s+v}}(p,l)}\right| . \tag{3.47}$$

The final quantity we use is a speech degradation measure, which can be defined by the LAR of the input and the output speech signals only

$$SD(l) = \sum_{p=1}^{12} 20\log_{10}\left|\frac{g_s(p,l)}{g_{y_s}(p,l)}\right| . \tag{3.48}$$

It includes the room reverberation, the signal distortion caused by the tested algorithm, and the dereverberation features of the tested algorithm only. Finally, the average of all blocks containing speech is computed.

3.4.3 Simulation Results

The described simulation system was used to evaluate the performance of four different post-filter algorithms:

1. Zel88: The algorithm by Zelinski in the frequency-domain implementation [21].
2. Sim92: The algorithm by Simmer described in [13].
3. APAB: The adaptive post-filter for an arbitrary beamformer, described in section 3.3 with a constrained MVDR-beamformer designed for an isotropic noise field in three dimensions (superdirective beamformer). The constraining parameter is set to $\mu = 0.01$ (see Chapter 2).
4. APES: The adaptive post-filter extension for superdirective beamformers [32].

For comparison, we include the results of the case in which no algorithm is used (No NR).

The speech sample we used is the sentence "I am now speaking to you from a distance of 50 cm from the microphone" spoken by an adult male. The length of this file leads to 98 blocks containing speech. The noise file was white Gaussian noise used in order to give technical results which can be reproduced by other researchers. The input SNR was computed only for blocks containing speech by using the segmental SNR.

In the first experiment, the broadside array shown on the left side of Fig. 3.4 is examined. Figure 3.6 depicts the results for the SNRE. The left side shows the dependence on the input-SNR if the reverberation time is set to $\tau_{60} = 300$ ms. The right figure shows the results for SNR=5 dB as a function of the reverberation time. This provides information on the behavior of the algorithms for different spatial conditions. The noise-field is coherent for low reverberation time and approximately diffuse for high values.

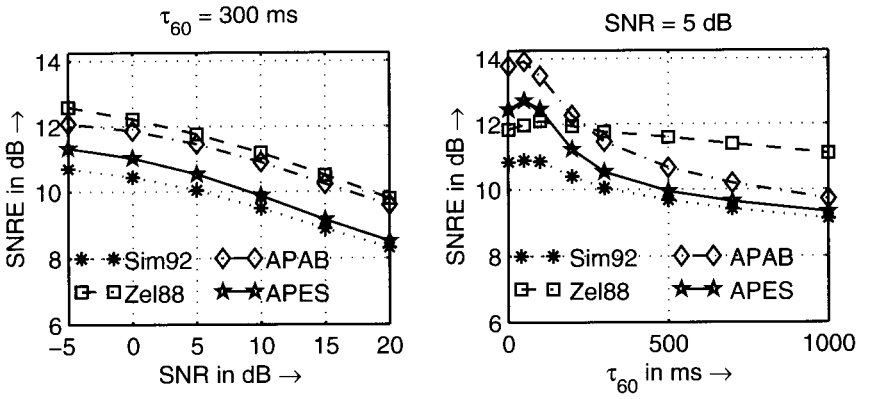


Fig. 3.6. Left: SNRE vs. input-SNR. Right: SNRE vs. reverberation time τ_{60} (Broadside).

Although not optimal the Zel88 algorithm performs quite well, especially for high reverberation times where it provides the best results of all tested algorithms (if only the SNRE is considered). At low reverberation times APAB and APES can benefit from the better suppression at low frequencies by using a superdirective beamformer instead of a standard delay and sum beamformer.

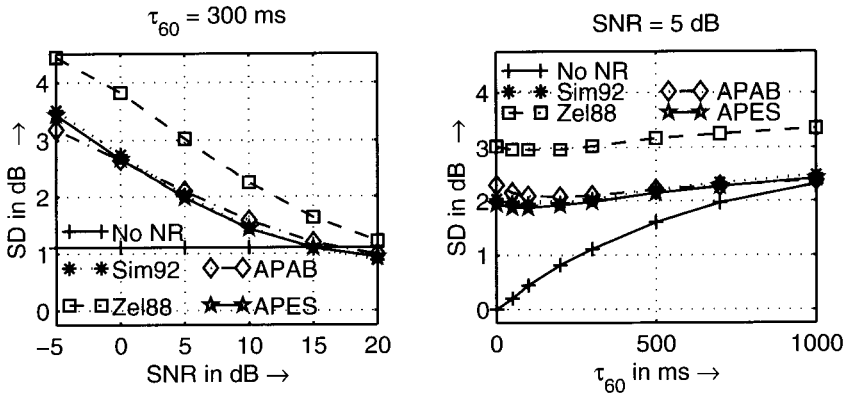


Fig. 3.7. Left: SD vs. input-SNR. Right: SD vs. reverberation time τ_{60} (Broadside).

If we take into account the next two measures shown in Fig. 3.7 and 3.8, which describe the performance in terms of speech quality, the results are different. All algorithms enhance the speech quality in comparison to the

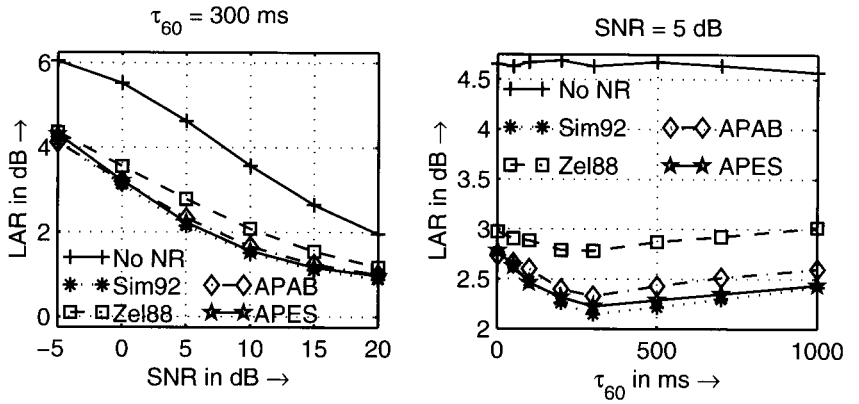


Fig. 3.8. Left: LAR vs. input-SNR. Right: LAR vs. reverberation time τ_{60} (Broadside).

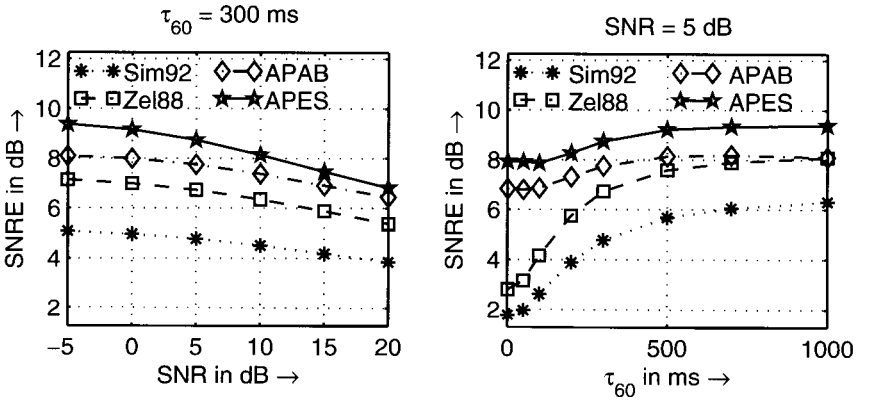


Fig. 3.9. Left: SNRE vs. input-SNR. Right: SNRE vs. reverberation time τ_{60} (Endfire).

unprocessed input signal [1]. However, the algorithm with the highest SNRE does not produce the best LAR. A closer look at Fig. 3.7 explains this behavior. Since these figures show the speech degradation only, the non-processed signal is constant versus the SNR and reduces to zero if no reverberation is added to the speech signal. The algorithms cause signal distortion at low SNR and the algorithm with the highest performance in SNRE induces the largest distortion, whereas APAB and APES provide the best speech quality (LAR). At very good conditions (SNR > 15 dB), these algorithms are able to suppress reverberation without introducing speech degradation. The lack of artifacts was corroborated through informal listening tests.

[1] Smaller values indicate better quality.

In a second experiment (right side of Fig. 3.4), we changed the orientation of the array and the inter-microphone distance. Additionally, only four microphones were used to reduced the array size. In Fig. 3.9 the SNRE results of the simulation are shown. The performance of the Sim92 and Zel88 algorithms degrades drastically, since the inherent delay and sum beamformer does not perform well at low frequencies due to the small array size. On the other hand, APAB and APES perform well under all conditions. The SNRE for APES at high reverberation time is close to the result for the broadside-experiment although the number of microphones is reduced. Thus, we conclude that end-fire steering is preferable for this algorithm.

3.5 Conclusion

Wiener post-filtering of the output signal of an MVDR beamformer provides an optimum MMSE solution for signal enhancement. A large number of published algorithms for post-filter estimation are based on the assumption of spatially uncorrelated noise. This assumption leads to post-filtering algorithms with suboptimal performance in coherent and diffuse noise fields. In this chapter we presented a new algorithm which performs considerably better in correlated noise fields by using the gain of an arbitrary array. Small size end-fire arrays comprising an MVDR beamformer and optimized post-filters showed the best performance in our simulations.

References

1. R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*, John Wiley and Sons, New York, 1980.
2. L. Danilenko, *Binaurales Hören im nichtstationären diffusen Schallfeld*, PhD thesis, RWTH Aachen, Aachen, Germany, 1968.
3. G. von Békésy, *Experiments in Hearing*, McGraw–Hill, New York, 1960.
4. S. Gierl, *Geräuschreduktion bei Sprachübertragung mit Hilfe von Mikrofonarraysystemen*, PhD thesis, Universität Karlsruhe, Karlsruhe, Germany, 1990.
5. S. Gierl, "Noise reduction for speech input systems using an adaptive microphone-array", in *Int. Symp. Automotive Tech. and Automation (ISATA)*, Florence, Italy, May 1990, pp. 517–524.
6. H.-Y. Kim, F. Asano, Y. Suzuki, and T. Sone, "Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer", *IEICE Trans. Fundament.*, vol. E79-A, no. 12, pp. 2151–2158, Dec. 1996.
7. M. Dörbecker and S. Ernst, "Combination of two–channel spectral subtraction and adaptive Wiener post–filtering for noise reduction and dereverberation", in *Proc. EURASIP European Signal Proc. Conf. (EUSIPCO)*, Trieste, Italy, Sept. 1996.
8. K. Kroschel, A. Czyzewksi, M. Ihle, and M. Kuropatwinski, "Adaptive noise cancellation of speech signals in a noise reduction system based on a microphone array", in *102nd Audio Eng. Soc. Conv., preprint 4450*, Munich, Germany, Mar. 1997.

9. J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals", *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, Oct. 1977.
10. Y. Kaneda and M. Tohyama, "Noise suppression signal processing using 2–point received signals", *Electron. Communicat. Japan*, vol. 67-A, no. 12, pp. 19–28, Apr. 1984.
11. R. Zelinski, "A microphone array with adaptive post–filtering for noise reduction in reverberant rooms", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*, New York, USA, Apr. 1988, pp. 2578–2581.
12. R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post–filtering", *Electron. Lett.*, vol. 26, no. 24, pp. 2036–2037, Nov. 1990.
13. K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain", in *Second Cost 229 Workshop Adapt. Alg. Communicat.*, Bordeaux, France, Oct. 1992, pp. 185–194.
14. Y. Mahieux and C. Marro, "Comparison of dereverberation techniques for videoconferencing applications", in *100th Audio Eng. Soc. Conv., preprint 4231*, Copenhagen, Denmark, May 1996.
15. C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering", *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.
16. R. Le Bouquin and G. Faucon, "On using the coherence function for noise reduction", in *Proc. EURASIP European Signal Proc. Conf. (EUSIPCO)*, Barcelona, Spain, Sept. 1990, pp. 1103–1106.
17. R. Le Bouquin and G. Faucon, "Study of a noise cancellation system based on the coherence function", in *Proc. EURASIP European Signal Proc. Conf. (EUSIPCO)*, Brussels, Belgium, Aug. 1992, pp. 1633–1636.
18. G. Faucon and R. Le Bouquin-Jeannes, "Optimization of speech enhancement techniques coping with uncorrelated and correlated noise", in *Proc. IEEE Int. Conf. on Communication Technology (ICCT–96)*, Beijing, China, May 1996, pp. 416–419.
19. R. Le Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 5, pp. 484–487, Sept. 1997.
20. P. Kuczynski, *Mehrkanal-Analyse von Sprachsignalen zur adaptiven Störunterdrückung*, PhD thesis, University of Bremen, Shaker Verlag, Aachen, Germany, Sept. 1995.
21. K. U. Simmer, P. Kuczynski, and A. Wasiljeff, "Time delay compensation for adaptive multichannel speech enhancement systems", in *Proc. Int. Symp. Signals, Syst. Electron. ISSSE–92*, Paris, France, Sept. 1992, pp. 660–663.
22. M. Drews and M. Streckfuß, "Multi–channel speech enhancement using an adaptive post–filter with channel selection and auditory constraints", in *Proc. Int. Workshop Acoust. Echo and Noise Control*, London, UK, Sept. 1997, pp. 77–80.
23. M. Drews, *Mikrofonarrays und mehrkanalige Signalverarbeitung zur Verbesserung gestörter Sprache*, PhD thesis, Technische Universität Berlin, Berlin, Germany, 1999.

24. K. U. Simmer, S. Fischer, and A. Wasiljeff, "Suppression of coherent and incoherent noise using a microphone array", *Annals of Telecommunications*, vol. 49, no. 7/8, pp. 439–446, July 1994.
25. S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments", *Speech Commun.*, vol. 20, no. 3-4, pp. 215–227, Dec. 1996.
26. A. Hussain, D.R. Campbell, and T.J. Moir, "A new metric for selecting sub–band processing in adaptive speech enhancement systems", in *Proc. ESCA European Conf. Speech Communicat. Tech. (EUROSPEECH)*, Rhodes, Greece, Sept. 1997, pp. 1489–1492.
27. R. Atay, E. Mandridake, D. Bastard, and M. Najim, "Spatial coherence exploitation which yields non–stationary noise reduction in subband domain", in *Proc. EURASIP European Signal Proc. Conf. (EUSIPCO)*, Rhodes, Greece, Sept. 1998, pp. 1489–1492.
28. J. Gonzales-Rodriquez, J. L. Sanchez-Bote, and J. Ortega-Garcia, "Speech dereverberation and noise reduction with a combined microphone array approach", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*, Istanbul, Turkey, Apr. 2000, pp. 1489–1492.
29. D. Mahmoudi and A. Drygajlo, "Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*, Atlanta, USA, May 1998, pp. 1489–1492.
30. D. Mahmoudi and A. Drygajlo, "Wavelet transform based coherence function for multi–channel speech enhancement", in *Proc. EURASIP European Signal Proc. Conf. (EUSIPCO)*, Rhodes, Greece, Sept. 1998, pp. 1489–1492.
31. J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "An alternative implementation of the superdirective beamformer", in *Proc. IEEE Workshop Applicat. Signal Processing to Audio and Acoust.*, New Paltz, New York, Oct. 1999, pp. 7–10.
32. J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer", in *Proc. Int. Workshop Acoust. Echo and Noise Control*, Pocono Manor, USA, Sept. 1999, pp. 100–103.
33. I. A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near–field superdirective beamforming with post–filtering", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*, Istanbul, Turkey, Apr. 2000.
34. J. P. Burg, "Three–dimensional filtering with an array of seismometers", *Geophysics*, vol. 29, no. 5, pp. 693–713, Oct. 1964.
35. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.
36. L. W. Brooks and I. S. Reed, "Equivalence of the likelihood ratio processor, the maximum signal–to–noise ratio filter, and the Wiener filter", *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES–8, no. 5, pp. 690–692, Sept. 1972.
37. D. J. Edelblute, J. M. Fisk, and G. L. Kinnison, "Criteria for optimum–signal–detection theory for arrays", *J. Acoust. Soc. Amer.*, vol. 41, no. 1, pp. 199–205, Jan. 1967.
38. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*, Wiley, New York, 1949.
39. W. Kellermann, "Analysis and design of multirate systems for cancelling of acoustic echoes", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*, Munich, Germany, Apr. 1988, pp. 2570–2573.

40. C. Marro, Y. Mahieux, and K. U. Simmer, "Performance of adaptive dereverberation techniques using directivity controlled arrays", in *Proc. EURASIP European Signal Proc. Conf. (EUSIPCO)*, Trieste, Italy, Sept. 1996, pp. 1127–1130.
41. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small–room acoustics", *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
42. S. R. Quakenbusch, T. P. Barnwell, and B. A. Clemens, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

4 Spatial Coherence Functions for Differential Microphones in Isotropic Noise Fields

Gary W. Elko

Media Signal Processing Research, Agere Systems, Murray Hill NJ, USA

Abstract. The spatial correlation function between directional microphones is useful in the design and analysis of the performance of these microphones in actual acoustic noise fields. These correlation functions are well known for omnidirectional receivers, but not well known for directional receivers. This chapter investigates the spatial correlation functions for Nth-order differential microphones in both spherically and cylindrically isotropic noise fields. The results are used to calculate the amount of achievable cancellation from an adaptive noise cancellation application using combinations of differential microphones to remove unwanted noise from a desired signal. The results are useful in determining signal-to-noise ratio gains from arbitrarily positioned differential microphone elements in microphone array applications.

4.1 Introduction

The spatial correlation function is important in the design of optimal beamformers that maximize the signal-to-noise ratio (SNR), source direction finding algorithms, the calculation of actual SNR gain from arrays, and other array signal processing areas. The space-time correlation functions are well known for omnidirectional receivers in two specific environments: spherically and cylindrically isotropic noise fields. One area of large concern that has been a topic of ongoing work has been the design and performance of directional differential microphone systems. One application of these systems is in adaptive noise cancellation schemes. In order to predict the expected performance gains of these adaptive cancellation systems, the spatial correlation functions between directional microphones are required. Results are presented here for the specific cases of general orientation for first-order differential microphones in both spherically and cylindrically isotropic fields. Specific results are given for the general Nth-order cases for differential arrays that have collinear axes.

4.2 Adaptive Noise Cancellation

The use of adaptive noise cancellation in communication devices has been under investigation for more than two decades [1], [2]. The early studies predicted SNR gains on the order of 10 dB and higher. However, it was

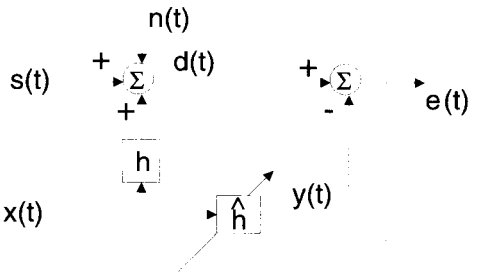


Fig. 4.1. Schematic model of adaptive noise cancellation system.

quickly learned that these predictions were not realized when devices were actually tested in real acoustic environments [2]. One of the problems that was encountered was the lack of coherence between a noise-alone sensor and the noise signal that was corrupting the desired signal. This lack of coherence was due to time-varying multipath, multiple uncorrelated noise sources, and nonlinearities in the transmission path to the signal channel [3].

Figure 4.1 shows the typical model of an adaptive noise cancellation system. It can be seen from this model that the adaptive noise cancellation problem is equivalent to the acoustic echo cancellation problem as described by Sondhi [4]. The desired output signal is $s(t)$. This signal is, however, corrupted by the noise signal $n(t)$, and the measured noise signal $x(t)$ convolved with the transmission path h from the measured noise channel to the signal pick-up channel.

The adaptive cancellation algorithm estimates the transmission path h and this estimated filter is represented by $\hat{h}$. It is assumed that the signals $s(t)$, $n(t)$, and $x(t)$ are uncorrelated stationary random processes. The output signal is $e(t)$, and if $h \approx \hat{h}$, the output signal $e(t) \approx s(t)$. If it is further assumed that the filter h is time-invariant, the optimum filter $\hat{H}_{opt}$ is the Wiener filter given by [1],

$$\hat{H}_{opt}(\omega) = \frac{S_{xd}(\omega)}{S_{xx}(\omega)} \tag{4.1}$$

where S_{xd} is the cross-spectrum between signals x and d, and S_{xx} is the autospectrum of signal x. If this filter is used in the model shown in Fig 4.1 then the output auto-spectrum is,

$$\begin{aligned} S_{ee}(\omega) &= S_{dd}(\omega) - \mid \hat{H}_{opt}(\omega) \mid^2 S_{xx}(\omega) \\ &= S_{dd}(\omega) \left[1 - \mid \gamma_{xd}(\omega) \mid^2\right] \end{aligned} \tag{4.2}$$

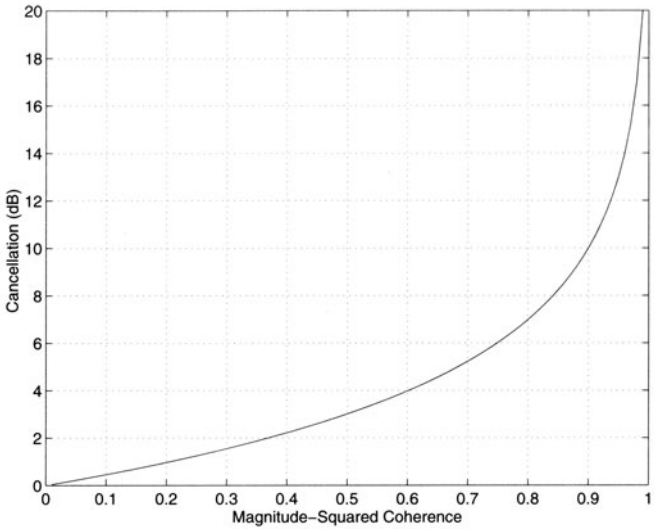


Fig. 4.2. Adaptive cancellation in dB versus the mean-square coherence between the noise signal $x(t)$ and the signal $d(t)$ as defined in Fig 4.1.

where γ_{xd} is the complex coherence function between the signals $x(t)$ and $d(t)$ and is defined as,

$$\gamma_{xd}(\omega) = \frac{S_{xd}(\omega)}{S_{xx}^{1/2}(\omega) S_{dd}^{1/2}(\omega)}. \tag{4.3}$$

The amount of cancellation is equal to the ratio of the primary corrupted signal power to the output signal power,

$$\begin{aligned} R(\omega) &= \frac{S_{dd}(\omega)}{S_{ee}(\omega)} \\ &= \frac{1}{1 - \mid \gamma_{xd}(\omega) \mid^2}. \end{aligned} \tag{4.4}$$

The results presented in (4.4) are well known [2] and a plot of this equation is shown in Fig. 4.2.

As can be seen in Fig. 4.2, the magnitude-squared coherence value must be greater than 0.9 if the cancellation R is to be larger than 10 dB.

In Fig. 4.1 it can be seen that if $s(t)$ and $n(t)$ are zero, then the cancellation will become infinite. However, in the case of a multipath field with many independent noise sources, the cancellation will be diminished since the coherence between the signals x and d will decrease. To see this, it is illustrative to examine the case of two independent noise sources n_1 and n_2 as shown in Fig. 4.3.

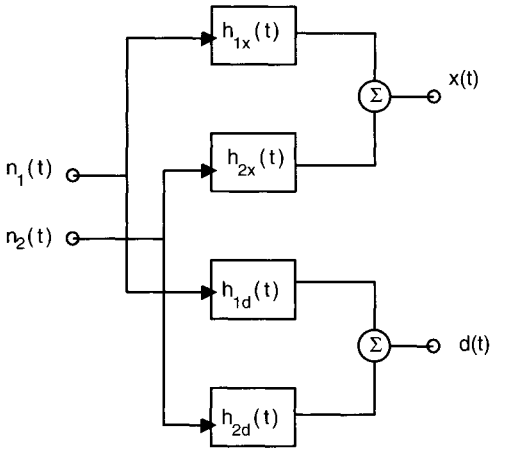


Fig. 4.3. Schematic model of two independent sources n_1 and n_2 combining through filters to form signals x and d.

For this case the autospectral densities are,

$$S_{xx}(\omega) = S_{11}(\omega) \mid H_{1x}(\omega) \mid^2 + S_{22}(\omega) \mid H_{2x}(\omega) \mid^2 \tag{4.5}$$

and

$$S_{dd}(\omega) = S_{11}(\omega) \mid H_{1d}(\omega) \mid^2 + S_{22}(\omega) \mid H_{2d}(\omega) \mid^2 . \tag{4.6}$$

The cross-spectral density is,

$$S_{xd}(\omega) = S_{11}(\omega) H_{1x}(\omega) H_{1d}^*(\omega) + S_{22}(\omega) H_{2x}(\omega) H_{2d}^*(\omega) \tag{4.7}$$

where the superscript * denotes the complex conjugate. The magnitude-squared coherence between x and d is therefore,

$$\mid \gamma_{xd}(\omega) \mid^2 = \frac{\left| \sum_{i=1}^{2} S_{ii}(\omega) H_{ix}(\omega) H_{id}^*(\omega) \right|^2}{\left[\sum_{i=1}^{2} S_{ii}(\omega) \mid H_{ix}(\omega) \mid^2 \right] \left[\sum_{i=1}^{2} S_{ii}(\omega) \mid H_{id}(\omega) \mid^2 \right]} \leq 1. \tag{4.8}$$

The coherence function given in (4.8) has a value of 1 only if $H_{1x} = H_{1d}$ and $H_{2x} = H_{2d}$. In general, for L independent sources the limit of the sums in (4.8) would be L. The model as explained above is a reasonable approximation to what is typically found in practice for acoustic environments in which people work and communicate. Thus, the loss of coherence between

sensors in adaptive noise cancellation will most likely be due to this multiple-independent-noise condition. As such, an analysis as to the loss of coherence between sensors for different acoustic noise fields is important. This chapter investigates the achievable cancellation for adaptive noise cancellation using differential sensors in both spherically and cylindrically isotropic noise fields. It is expected that these two types of fields will yield results that are representative of what can be obtained in real-world acoustic noise fields.

A practical example of interest in telephony is the use of adaptive noise cancellation for noise removal from the transmitter (microphone) in a telephone handset. A recent patent application [5] has explicitly proposed the use of a secondary directional microphone mounted on the handset such that the null of this noise-alone microphone is aimed in the direction of the "desired" signal. The output from this "noise-alone" microphone is then used to cancel the correlated noise in the microphone that is used to pick-up the desired signal. In order to predict the cancellation from this proposed arrangement of transducers, it is necessary to calculate the spatial coherence between these sensors.

In a typical adaptive noise cancellation implementation the transfer function H is approximated as an all-zero filter, i.e., the impulse response h is estimated by an adaptive finite-impulse response (FIR) filter. One advantage of making this system adaptive is to allow for the possibility of a time varying impulse response $h(t)$. There are several problems that occur in this implementation. One major problem is the presence of the desired signal and/or uncorrelated noise signal $n(t)$, when the adaptive filter is attempting to adapt to the measured noise-to-primary input transfer function. This problem is the same as the "double-talk" problem in the field of acoustic echo cancellation [4]. Another problem is that the signals $s(t)$, $n(t)$, and $x(t)$ are typically nonstationary. Finally, another problem that can limit the cancellation performance is low coherence between the signals $x(t)$ and the signal $d(t)$, even when $s(t)$ and $n(t)$ are small in signal power compared to the power of the noise signal $x(t)$. This lack of coherence has been postulated to be due to nonlinearities and strong nonstationary (time-varying) multipath environments [3],[10].

4.3 Spherically Isotropic Coherence

The spatio-temporal autocorrelation and cross-correlation functions are very useful quantities in sensor array processing. Perhaps the most simple and historically prominent calculation was the correlation between two omnidirectional microphones in an isotropic noise field. The initial calculation was published by R. K. Cook *et al.* [6]. For completeness and to develop the notation this well-known result will now be derived.

The space-time correlation function for stationary random processes p_1 and p_2 is defined as,

$$R_{12}(\mathbf{r},\tau) = E[p_1(\mathbf{s},t)p_2(\mathbf{s}-\mathbf{r},t-\tau)] \tag{4.9}$$

where E is the expectation operator, $\mathbf{s}$ is the position of the sensor measuring acoustic pressure p_1, and $\mathbf{r}$ is the displacement vector to the sensor measuring pressure p_2. For a plane-wave incident field with wavevector $\mathbf{k}$, ($||\ \mathbf{k}\ ||= k = \omega/c$ where c is the speed of sound), R_{12} can be written as

$$R_{12}(\mathbf{r},\tau) = R(\tau + \mathbf{k}\cdot\mathbf{r}) \tag{4.10}$$

where R is the temporal autocorrelation function of the acoustic pressure p. The cross-spectral density is the Fourier transform of the cross-correlation function,

$$S_{12}(\mathbf{r},\omega) = \int R_{12}(\mathbf{r},\tau)e^{j\omega\tau}d\tau. \tag{4.11}$$

If we assume that the acoustic field is spatially homogeneous (the correlation function is not dependent on the absolute position of the sensors), and also assume that the field is spherically isotropic (uncorrelated signals from all directions), the vector $\mathbf{r}$ can be replaced with a scalar variable r which is the spacing between the two measurement locations. Thus the cross-spectral density for an isotropic field is the average cross-spectral density for all spherical directions, θ, ϕ. Therefore,

$$\begin{aligned} S_{12}(r,\omega) &= \frac{N_o(\omega)}{4\pi}\int_0^{\pi}\int_0^{2\pi} e^{-jkr\cos\theta}\sin\theta d\theta d\phi \\ &= \frac{N_o(\omega)\sin(\omega r/c)}{\omega r/c} \\ &= \frac{N_o(\omega)\sin(kr)}{kr} \end{aligned} \tag{4.12}$$

where $N_o(\omega)$ is the power spectral density at the measurement locations and it has been assumed without loss in generality that the vector $\mathbf{r}$ lies along the z-axis. Note that the isotropic assumption implies that the autopower-spectral density is the same at each location. The complex coherence function γ is defined as the normalized cross spectral density,

$$\gamma_{12}(r,\omega) = \frac{S_{12}(r,\omega)}{[S_{11}(\omega)S_{22}(\omega)]^{1/2}}. \tag{4.13}$$

For spherically isotropic noise and omnidirectional receivers, the spatial coherence function is,

$$\gamma_{12}(r,\omega) = \frac{\sin(kr)}{kr}. \tag{4.14}$$

In general, the spatial coherence function can be determined as,

$$\gamma_{12}(r,\omega) = \frac{E\left[T_1(\theta,\phi,\omega)T_2^*(\theta,\phi,\omega)e^{-j\mathbf{k}\cdot\mathbf{r}}\right]}{E\left[\mid T_1(\theta,\phi,\omega)\mid^2\right]^{1/2} E\left[\mid T_2(\theta,\phi,\omega)\mid^2\right]^{1/2}} \tag{4.15}$$

where T_1 and T_2 are the directivity functions for the two directional sensors. In integral form for spherically isotropic fields, (4.15) can be written as,

$$\gamma_{12}(r,\omega) = \frac{N_{12}(r,\omega)}{D_{12}(\omega)}, \tag{4.16}$$

where

$$N_{12}(r,\omega) = \int_0^{\pi}\int_0^{2\pi} T_1(\theta,\phi,\omega)T_2^*(\theta,\phi,\omega)e^{-jkr\cos\theta}\sin\theta d\theta d\phi,$$

and

$$D_{12}(r,\omega) = \left(\int_0^{\pi}\int_0^{2\pi} \mid T_1(\theta,\phi,\omega)\mid^2 \sin\theta\, d\theta d\phi\right)^{1/2} \times \left(\int_0^{\pi}\int_0^{2\pi} \mid T_2(\theta,\phi,\omega)\mid^2 \sin\theta\, d\theta d\phi\right)^{1/2}.$$

The denominator is inversely proportional to the geometric mean of the two microphone directivity factors Q_1 and Q_2 [8]. Therefore the denominator D_{12} is,

$$D_{12}(\omega) = [Q_1(\omega)Q_2(\omega)]^{-1/2}. \tag{4.17}$$

A general closed-form solution for the spatial coherence between any Nth and Mth-order differential array if the differential axes are collinear has been found and is presented in a subsequent section. First, however, a general result for first-order differential arrays will be discussed. For this particular differential order, a solution is presented that allows the calculation of the spatial coherence for any arbitrary orientation of first-order differential arrays.

The directional response for a first-order differential microphone can be written as [8],

$$T_i(\psi_i) = \alpha_i + (1-\alpha_i)\cos\psi_i, \quad i \in \{1,2\} \tag{4.18}$$

where ψ_i is the angle between the incident wave and the axis of the ith first-order microphone. Defining $\mathbf{u}_i$ as the unit vector indicating the spatial orientation of differential microphone i, and defining $\hat{\mathbf{k}} = \mathbf{k}/\parallel\mathbf{k}\parallel$ as a unit vector, results in the following definitions in spherical coordinates:

$$\begin{aligned}\hat{\mathbf{k}} &= (\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)\\ \mathbf{u}_i &= (\cos\phi_i\sin\theta_i, \sin\phi_i\sin\theta_i, \cos\theta_i).\end{aligned} \tag{4.19}$$

Thus, the cosine term in (4.18) can be written as

$$\cos\psi_i = \hat{\mathbf{k}} \cdot u_i. \tag{4.20}$$

Using (4.15), (4.16), (4.18), (4.19), and (4.20) and again assuming, without loss of generality, that the microphones lie along the z-axis, yields

$$\begin{aligned} N_{12}(kr) = \frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} & \\ & [\alpha_1 + (1-\alpha_1)(x_1\cos\phi\sin\theta + y_1\sin\phi\sin\theta + z_1\cos\theta)] \\ & [\alpha_2 + (1-\alpha_2)(x_2\cos\phi\sin\theta + y_2\sin\phi\sin\theta + z_2\cos\theta)] \\ & \sin(\theta)e^{-jkr\cos\theta}d\theta d\phi, \end{aligned} \tag{4.21}$$

where

$$\begin{aligned} x_i &= \cos\phi_i\sin\theta_i, \\ y_i &= \sin\phi_i\sin\theta_i, \\ z_i &= \cos\theta_i, \quad i \in \{1,2\}. \end{aligned}$$

Note that since the directional response of the differential array is independent of ω, then the functional arguments for γ, N, and D, can be compressed into one variable that is the product of k and r. Thus, only the functional dependency for the product kr will be used in the remainder of this chapter and the functions that depend solely on frequency will be written without the frequency dependence. Solving the integral (4.21) yields

$$\begin{aligned} N_{12}(kr) = & \frac{\alpha_1\alpha_2\sin(kr)}{kr} \\ & + \frac{(1-\alpha_2)(1-\alpha_2)(x_1x_2+y_1y_2)}{(kr)^3}[\sin(kr) - kr\cos(kr)] \\ & + \frac{z_1z_2}{(kr)^3}\{[(kr)^2\sin(kr) + 2kr\cos(kr)](1-\alpha_1)(1-\alpha_2) \\ & \quad + 2\sin(kr)(1-\alpha_1)(\alpha_2-1)\} \\ & + \frac{z_1}{(kr)^3}[j(kr)^2\alpha_2\cos(kr)(\alpha_1-1) + jkr\alpha_2\sin(kr)(1-\alpha_1)] \\ & + \frac{z_2}{(kr)^3}[j(kr)^2\alpha_1\cos(kr)(\alpha_2-1) + jkr\alpha_1\sin(kr)(1-\alpha_2)]. \end{aligned} \tag{4.22}$$

For an Nth-order differential array whose directional response can be written as [8]

$$T(\theta) = a_0 + a_1\cos(\theta) + a_2\cos^2(\theta) + \cdots + a_N\cos^N(\theta), \tag{4.23}$$

the solution for the directivity factor is

$$Q(a_0, ..., a_N) = \left[\sum_{\substack{i=0 \\ i+j \text{ even}}}^{N} \sum_{j=0}^{N} \frac{a_i a_j}{1+i+j} \right]^{-1} \tag{4.24}$$

For a first-order differential microphone (4.24) reduces to

$$Q(a_0, a_1) = \frac{3}{3a_0^2 + a_1^2} = \frac{3}{3\alpha^2 + (1-\alpha)^2}. \tag{4.25}$$

Thus, for the first-order case the denominator term D_{12} is

$$D_{12} = \frac{\left[3\alpha_1^2 + (1-\alpha_1)^2\right]^{1/2} \left[3\alpha_2^2 + (1-\alpha_2)^2\right]^{1/2}}{3}. \tag{4.26}$$

The quotient of (4.22) and (4.26) yields the general result for the coherence function between any arbitrarily oriented first-order differential microphones spaced at a distance r . If the values of α_i are both equal to 1, the microphones are omnidirectional and the coherence from the ratio of (4.22) and (4.26) reduces to the well-known $\sin(kr)/kr$ result as given in (4.14). Figure 4.4 shows the coherence between a pair of omnidirectional microphones and the coherence between various orientations of pairs of dipole microphones spaced as a function of the dimensionless parameter kr. The suffix symbols in the plot legend indicate the orientation of the dipole microphone axes. The curve for the orthogonal dipole case runs along the abscissa and therefore can not be explicitly seen in the figure. The fact that the orthogonal dipoles have a zero coherence can be understood if symmetry is considered. The complex coherence for a wave impinging from one angle is of opposite sign to a wave impinging from the opposing angle ($\theta' = \theta$, $\phi' = -\phi$). The net coherence is therefore zero for the isotropic noise case. The parallel dipoles (denoted as a dash-dot line) have a higher coherence value for $kr < \pi$, compared to the omnidirectional and collinear dipoles since there is zero delay for signals propagating along the major axes of the microphones.

Figure 4.5 shows the amount of possible cancellation attainable with these various orientations of the dipole microphones calculated from (4.4). Figure 4.6 shows the coherence between various orientations of cardioid microphones as a function of kr. Figure 4.7 shows the amount of possible cancellation attainable with these various orientations of the dipole microphones. These results are the same as those calculated using explicit forms for the cardioid microphones in an earlier paper by Goulding and Bird [9].

Figure 4.8 shows the coherence between various orientations of omnidirectional microphones, dipole and cardioid microphones, as a function of kr. Figure 4.9 shows the amount of possible cancellation attainable with these various orientations of omnidirectional, dipole, and cardioid microphones. It is interesting to note (but not unexpected) that the maximum cancellation for

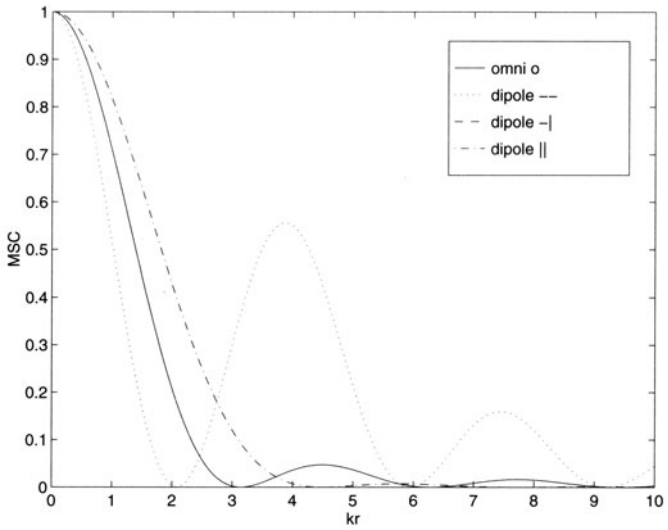


Fig. 4.4. Magnitude-squared coherence (MSC) for omnidirectional and dipole microphones in a spherically isotropic noise field. Note that the curve for the orthogonal dipoles lies along the abscissa.

the omnidirectional and the cardioid has a value of 6 dB, which is the maximum directional gain for a first-order differential microphone in an isotropic noise field. It is also interesting to note that the omnidirectional and the dipole are uncorrelated over all values of k and r. This result is again due to the symmetry argument that was made for the two orthogonal dipoles.

In order to find a closed-form solution for an arbitrary order differential microphone arrangement, it is necessary to confine the orientation of the arrays. A solution can be found if the axes of the two Mth-order and Nth-order differential microphones are collinear. To begin, recall that an Nth-order differential array directional response can be written as (4.23),

$$\begin{aligned} T_1(\theta) &= a_0 + a_1\cos(\theta) + \ \ldots\ + a_N\cos^N(\theta) \\ T_2(\theta) &= b_0 + b_1\cos(\theta) + \ \ldots\ + b_N\cos^N(\theta). \end{aligned} \tag{4.27}$$

Note that it is not necessary to have both differential elements of the same order.[1] The solution to (4.21) using directivity functions of the form of (4.27) requires the solution of the integral:

$$I_n(kr) = \frac{1}{2}\int_0^{\pi} \cos^n(\theta) e^{-jkr\cos\theta}\sin\theta d\theta. \tag{4.28}$$

[1] The number N chosen in (4.27) is the larger order of the individual microphones. Therefore, the coefficients of the lower order differential microphone are zero from the differential order of this microphone to the term N.

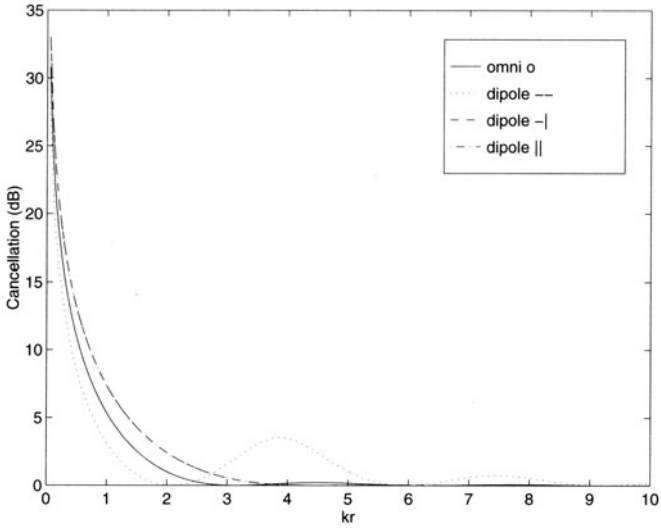


Fig. 4.5. Maximum cancellation (dB) for omnidirectional and of dipole microphones for spherically isotropic noise fields. Note that the curve for the orthogonal dipoles lies along the abscissa.

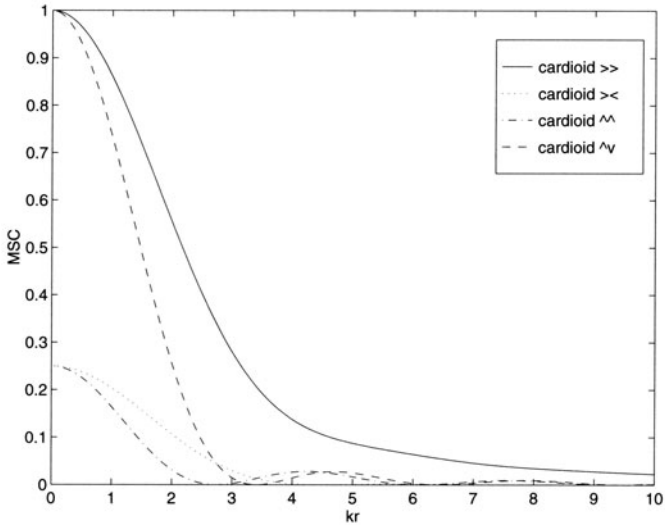


Fig. 4.6. Magnitude-squared coherence (MSC) for various orientations of cardioid microphones in a spherically isotropic noise field.

From Appendix A, the result is

$$I_n = \frac{n!}{2(jkr)^{n+1}} \left[e^{jkr} \sum_{m=0}^{n} \frac{(-jkr)^n}{m!} - e^{-jkr} \sum_{m=0}^{n} \frac{(jkr)^n}{m!} \right]. \tag{4.29}$$

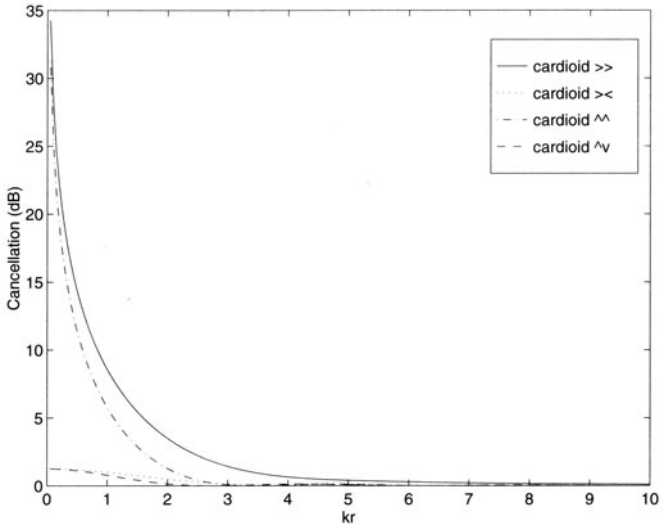


Fig. 4.7. Maximum cancellation (dB) for various orientations of cardioid microphones for spherically isotropic noise fields.

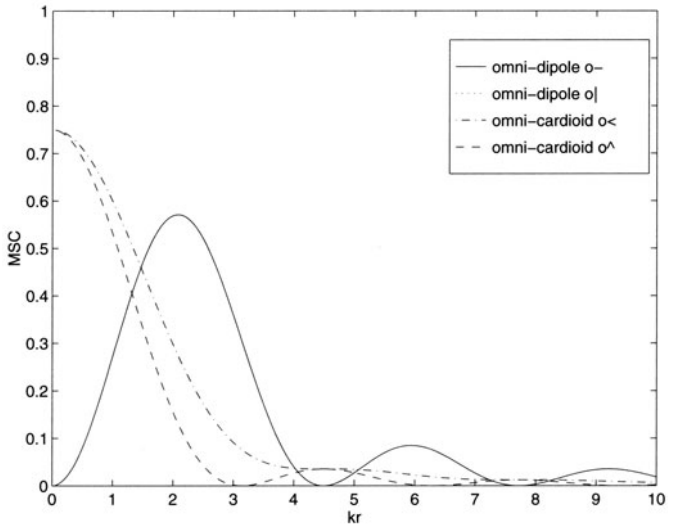


Fig. 4.8. Magnitude-squared coherence (MSC) for various orientations of omnidirectional and dipole and cardioid microphones in a spherically isotropic noise field.

The numerator of (4.21) is a sum of integrals given by (4.29). The denominator is inversely proportional to the square-root of the product of the directivity factors as given in (4.24). Therefore the solution to (4.21) for a general

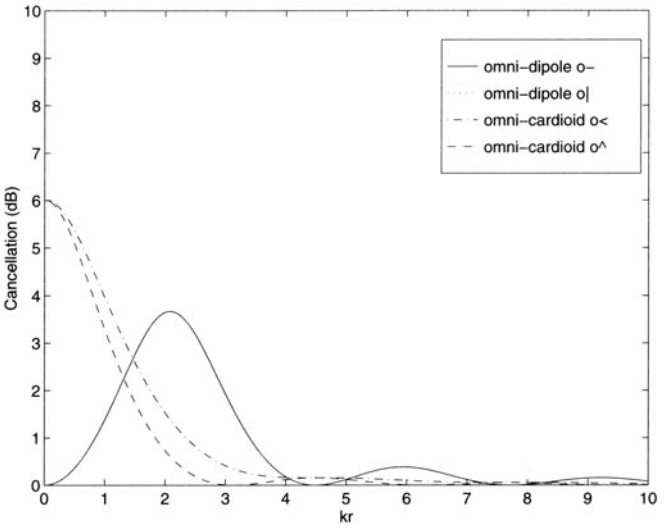


Fig. 4.9. Maximum cancellation (dB) for various orientations of omnidirectional and dipole and cardioid microphones for spherically isotropic noise fields.

combination of collinear differential arrays is

$$\gamma(kr) = \frac{\sum_{n=0}^{N} a_n b_{N-n} \frac{n!}{(jkr)^{n+1}} \left[e^{jkr} \sum_{m=0}^{n} \frac{(-jkr)^n}{m!} - e^{-jkr} \sum_{m=0}^{n} \frac{(jkr)^n}{m!} \right]}{2 \left[\sum_{\substack{i=0 \\ i+j \text{ even}}}^{N} \sum_{j=0}^{N} \frac{a_i a_j}{1+i+j} \right]^{1/2} \left[\sum_{\substack{i=0 \\ i+j \text{ even}}}^{N} \sum_{j=0}^{N} \frac{b_i b_j}{1+i+j} \right]^{1/2}} \tag{4.30}$$

Plots of the coherence function for second and third-order dipole and cardioid microphones are shown in Figs. 4.10 and 4.11.

4.4 Cylindrically Isotropic Fields

The previous section dealt with spherically isotropic acoustic noise fields. It has been proposed that some room acoustic fields may be more closely modeled as a cylindrically isotropic field [8]. As a result, it is useful to derive theoretical spatial coherence functions for this type of field. The coherence function for any general field was given in (4.15). To derive the forms for the cylindrical field the only difference from the previous development for the spherically isotropic case is the integration implied by the expectation operator E. For the cylindrically isotropic field the expectation involves only

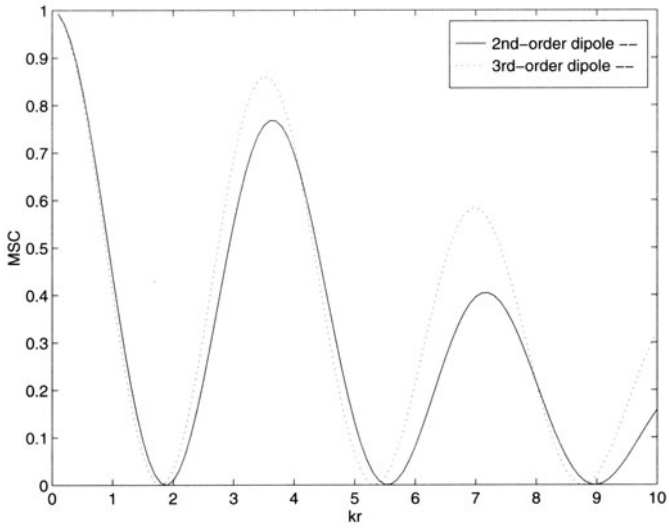


Fig. 4.10. Magnitude-squared coherence for second and third-order collinear dipoles in a spherically isotropic noise field.

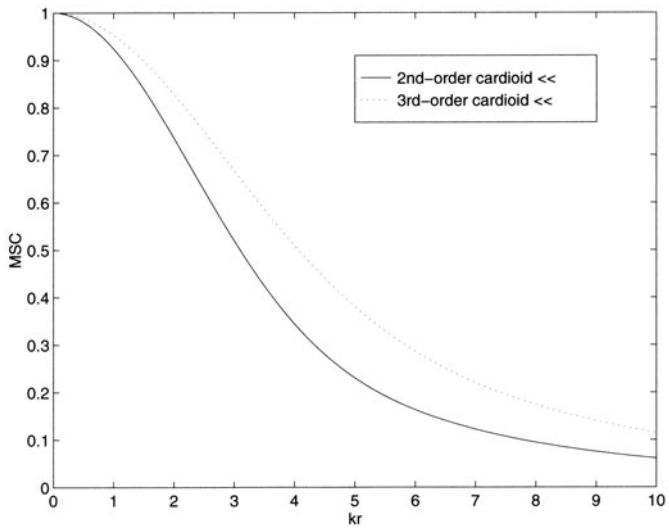


Fig. 4.11. Magnitude-squared coherence for second and third-order collinear cardioids in a spherically isotropic noise field.

the integration in one dimension, the cylindrical angle ϕ. The directional responses of the two first-order differential arrays with general orientation of

ϕ_1 and ϕ_2 are

$$T_i(\phi) = \alpha_i + (1 - \alpha_i)\cos(\phi - \phi_i), \quad i \in \{1, 2\}. \tag{4.31}$$

The numerator for the coherence function is the integral of the product of the two directional responses given in (4.31) and is (assuming without loss in generality that the microphones lie along the z-axis),

$$\begin{aligned} N_{12}(kr) = \frac{1}{2\pi}\int_0^{2\pi} & [\alpha_1 + (1 - \alpha_1)(x_1 \cos\phi\cos\phi_1 + \sin\phi\sin\phi_1)] \\ & \times [\alpha 2 + (1 - \alpha_2)(\cos\phi\cos\phi_2 + \sin\phi\sin\phi_2)] \\ & \times e^{-jkr\cos\phi} d\phi. \end{aligned} \tag{4.32}$$

The integration of (4.32) is rather tedious and is given in Appendix B. The resulting numerator for the coherence function is

$$\begin{aligned} N_{12}(kr) = {} & \alpha_1\alpha_2 J_0(kr) \\ & + (\alpha_1 - 1)(\alpha_2 - 1)\cos\phi_1\cos\phi_2[J_0(kr) - J_2(kr)]/2 \\ & + (\alpha_1 - 1)(\alpha_2 - 1)\sin\phi_1\sin\phi_2[J_0(kr) + J_2(kr)]/2 \\ & + j[\alpha_2\cos\phi_1(1 - \alpha_1) + \alpha_1\cos\phi_2(1 - \alpha_2)]J_1(kr) \end{aligned} \tag{4.33}$$

where J_n are the Bessel functions of the first-kind of integer order n. The denominator for the coherence function for first-order differential arrays is easily derived and is,

$$D_{12} = \left[(\alpha_1^2 + \frac{1}{2}(1 - \alpha_1)^2\right]^{1/2}\left[(\alpha_2^2 + \frac{1}{2}(1 - \alpha_2)^2\right]^{1/2}. \tag{4.34}$$

A closed-form solution can also be found for the general Nth-order differential array in a cylindrically correlated field if the differential microphones have axes that are collinear. The numerator for the coherence function is the integral of the product of the individual directional responses given in (4.27). This product of polynomials can itself be expressed as a polynomial of order equal to the sum of the two individual directivity polynomial orders. In general, the solution for the numerator requires the evaluation of the integral

$$I_n = \frac{1}{\pi}\int_0^{\pi} \cos^n\phi e^{-jkr\cos\phi} d\phi. \tag{4.35}$$

From Appendix C, I_n is,

$$\begin{aligned} I_n &= \frac{1}{2^{n-1}}\left[\sum_{m=0}^{n/2} \varepsilon_m(-j)^{n-2m}C(n,m)J_{n-2m}(kr)\right], \quad \text{for } n \text{ even} \\ I_n &= \frac{1}{2^{n-1}}\left[\sum_{m=0}^{(n-1)/2} (-j)^{n-2m}C(n,m)J_{n-2m}(kr)\right], \quad \text{for } n \text{ odd} \end{aligned} \tag{4.36}$$

where ε_m is defined as,

$$\begin{aligned} \varepsilon_m &= 1, \ m \neq n/2, \\ &= \frac{1}{2}, \ m = n/2, \end{aligned} \tag{4.37}$$

and the function C is the binomial coefficient [7]

$$C(n,m) = \frac{n!}{(n-m)!m!}. \tag{4.38}$$

The numerator of the coherence function is

$$N_{12}(kr) = \sum_{n=0}^{2N} d_n I_n, \tag{4.39}$$

where the coefficients d_n are components of the vector

$$\mathbf{d} = \mathbf{a} \star \mathbf{b}. \tag{4.40}$$

The symbol $\star$ indicates the convolution and the vectors $\mathbf{a}$ and $\mathbf{b}$ are from the directivity response polynomials as defined in (4.27).

The denominator term has previously been shown as equal to the inverse of the directivity factor. The directivity factor for a differential array in a cylindrically isotropic sound field is [8]

$$Q_{\text{cyl}}(a_0, ...a_{N-1}) = \frac{\mathbf{a}^T \mathbf{G} \mathbf{a}}{\mathbf{a}^T \mathbf{H} \mathbf{a}}, \tag{4.41}$$

where the superscript T denotes the transpose operator, the subscript on Q indicates a cylindrical field,

$$\mathbf{a}^T = \{a_0, a_1, ..., a_N\} \ , \tag{4.42}$$

$\mathbf{G}$ is an $(N+1) \times (N+1)$ matrix whose elements are

$$G_{ij} = 1, \tag{4.43}$$

and $\mathbf{H}$ is a Hankel matrix given by,

$$H_{i,j} = \begin{cases} \dfrac{(i+j-1)!!}{(i+j)!!}, & \text{if i+j even,} \\ 0, & \text{otherwise.} \end{cases} \tag{4.44}$$

The double factorial function is defined as [7]: $(2n)!! = 2 \cdot 4 ... \cdot (2n)$ for n even, and $(2n+1)!! = 1 \cdot 3 \cdot \ ... \cdot (2n+1)$ for n odd. The denominator D_{12} is

$$D_{12} = \left[Q_{\text{cyl1}} Q_{\text{cyl2}} \right]^{-1/2}. \tag{4.45}$$

The quotient of (4.39) and (4.45) yields the general result for the coherence function between any arbitrarily oriented first-order differential microphones spaced at a distance r . If the two values of α_i are both unity, the spatial coherence reduces to the well-known value for omnidirectional elements in a cylindrically isotropic noise field [6]

$$\gamma_{12}(kr) = J_0(kr), \tag{4.46}$$

where J_0 is the zero-order Bessel function of the first-kind. Figure 4.12 shows the coherence between a pair of omnidirectional microphones and various orientations of dipole microphones spaced as a function of the dimensionless parameter kr. Figure 4.13 shows the amount of possible cancellation attainable with these various orientations of the dipole microphones. In general the curves for the cylindrically isotropic noise fields are similar to those of the spherically isotropic fields except that the values are higher for the cylindrical case as a function of kr. This result should not be too surprising since the integration region has now been confined to a plane, and not over all spherical directions.

Figure 4.14 shows the coherence between various orientations of cardioid microphones and as a function of kr. Figure 4.15 shows the amount of possible cancellation attainable with these various orientations of the cardioid microphones. Figure 4.16 shows the coherence between various orientations of omnidirectional microphones and dipole and cardioid microphones as a function of kr. Figure 4.17 shows the amount of possible cancellation attainable with these various orientations of the omnidirectional and dipole and cardioid microphones. Plots of coherence function for second and third-order dipole and cardioid microphones are shown in Figs. 4.18 and 4.19. The coherence functions decay more slowly for higher-order differential arrays that are collinear. This is due to the narrower beamwidth and the commensurate higher weighting of the noise field in the direction along the microphone axes.

4.5 Conclusions

It has been shown that adaptive noise cancellation schemes that utilize low-order differential microphones in isotropic noise fields require care in the orientation of the sensors. As an example, the use of orthogonal dipole microphones or an omnidirectional and an appropriately rotated dipole microphone will yield *no* noise cancellation at all. In general, adaptive cancellation will occur only for small values of kr (frequency-spacing product). It has been argued that strong multipath (reverberant) acoustic fields exhibit statistics similar to isotropic fields [10]. As a result, it should be expected that adaptive noise cancellation schemes will show limited SNR improvements in isotropic fields over a wide bandwidth. There is also the the problem of signal cancellation that occurs with adaptive algorithms in multipath acoustic fields that further limits the performance of adaptive noise cancellation in reverberant acoustic fields. The results presented here can be used to predict the

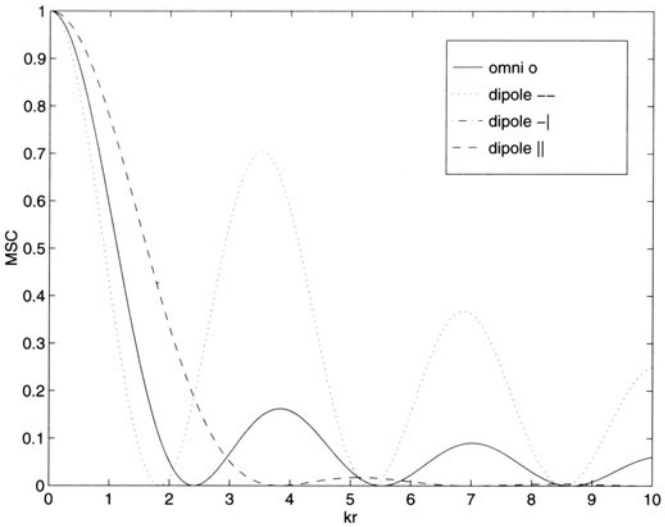


Fig. 4.12. Magnitude-squared coherence (MSC) for omnidirectional and various orientations of dipole microphones in a cylindrically isotropic noise field.

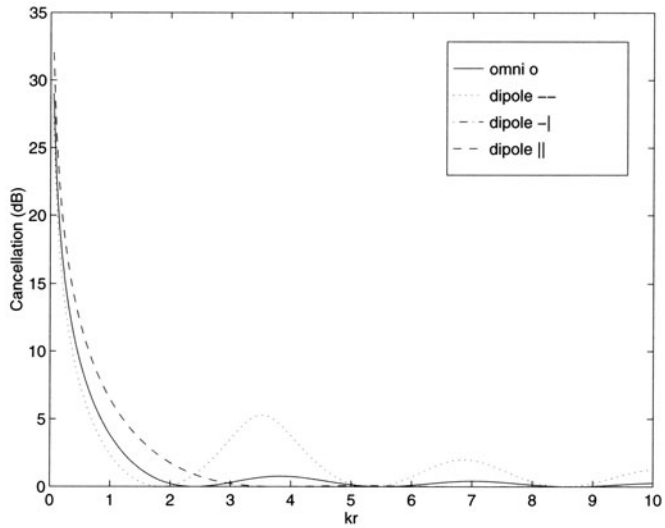


Fig. 4.13. Maximum cancellation (dB) for omnidirectional and various orientations of dipole microphones for cylindrically isotropic fields.

maximum attainable noise reduction for adaptive noise cancellation implementations in isotropic fields. If the field is significantly non-isotropic it can be expected that higher cancellation can be achieved. This is especially true

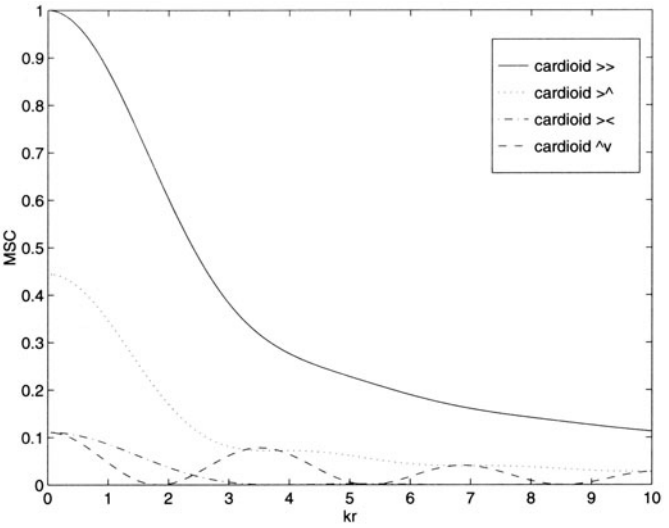


Fig. 4.14. Magnitude-squared coherence (MSC) for various orientations of cardioid microphones in a cylindrically isotropic noise field.

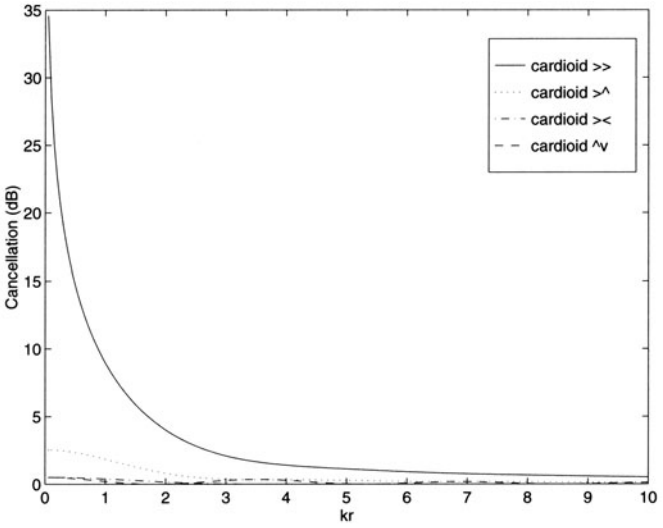


Fig. 4.15. Maximum cancellation (dB) for various orientations of cardioid microphones for cylindrically isotropic fields.

if the noise field is generated by a dominant noise source close to the microphone array, i.e., the direct field of the noise dominates.

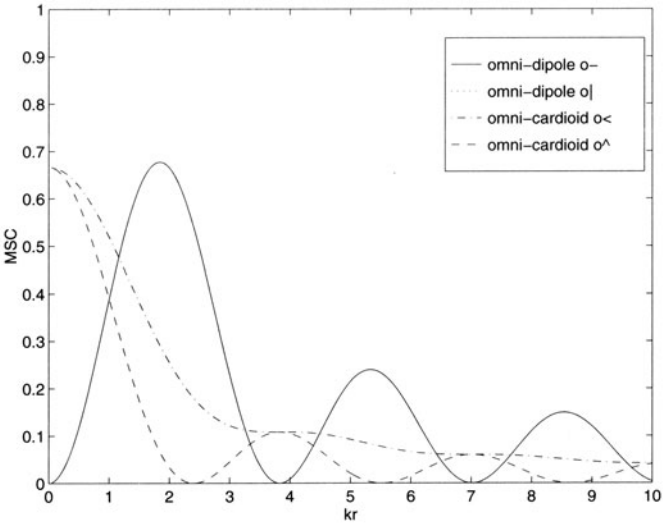


Fig. 4.16. Magnitude-squared coherence (MSC) for various orientations of omnidirectional and dipole and cardioid microphones in a cylindrically isotropic noise field.

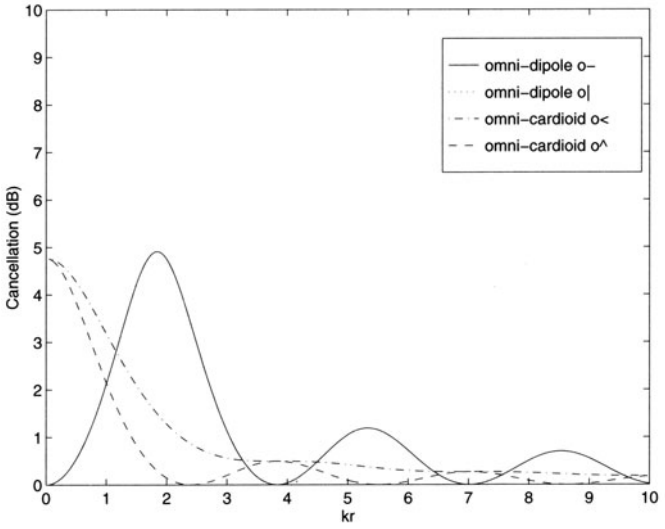


Fig. 4.17. Maximum cancellation (dB) for various orientations of omnidirectional and dipole and cardioid microphones for cylindrically isotropic fields.

Appendix A

The numerator term of the spatial coherence function for spherically isotropic noise fields contains a finite series containing integrals of the following type:

$$I_n = \frac{1}{2}\int_0^{\pi} \cos^n \theta e^{-jkr\cos\theta} \sin\theta d\theta. \tag{4.47}$$

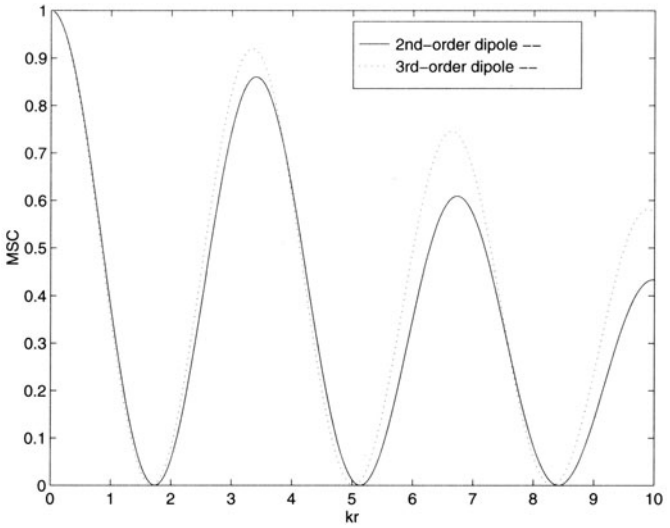


Fig. 4.18. Magnitude-squared coherence for second and third-order collinear dipoles in a cylindrically isotropic noise field.

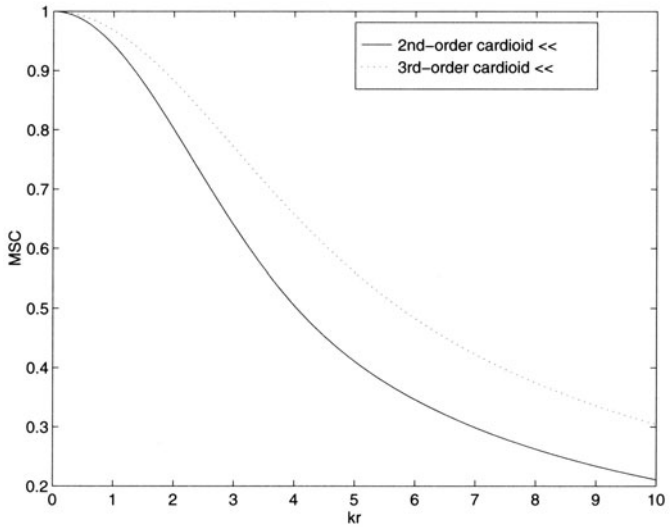


Fig. 4.19. Magnitude-squared coherence for second and third-order collinear cardioids in a cylindrically isotropic noise field.

This integral can be simplified by a change of variables,

$$t = jkr\cos\theta,$$
$$dt = -jkr\sin\theta\ d\theta.$$

Therefore,

$$I_n = \frac{1}{2(jkr)^{n+1}} \int_{-jkr}^{jkr} t^n e^{-t} dt. \tag{4.48}$$

The integral given in (4.48) can be found in Abramowitz [7] and is

$$I_n = \frac{1}{2(jkr)^{n+1}} [P(n+1, jkr) + (-1)^n P(n+1, -jkr)], \tag{4.49}$$

where the function P is the normalized incomplete Gamma function, which can be written in series form for integer values n as [7]

$$P(n+1, jkr) = 1 - (1 + jkr + \frac{(jkr)^2}{2!} + \cdots + \frac{(jkr)^n}{n!}) e^{-jkr}. \tag{4.50}$$

Therefore the integral I_n is

$$I_n = \frac{n!}{2(jkr)^{n+1}} \left[e^{jkr} \sum_{m=0}^{n} \frac{(-jkr)^m}{m!} - e^{-jkr} \sum_{m=0}^{n} \frac{(jkr)^m}{m!} \right]. \tag{4.51}$$

Appendix B

The numerator for the spatial coherence function for first-order microphones in a cylindrically isotropic noise field, and whose major axes are oriented at ϕ_1 and ϕ_2 with respect to the Cartesian coordinate system is

$$\begin{aligned} N(kr) = \frac{1}{2\pi} \int_0^{2\pi} & [\alpha_1 + (1-\alpha_1)\cos\phi_1 \cos\phi - \sin\phi_1 \sin\phi)] \\ & \times [\alpha_2 + (1-\alpha_2)\cos\phi_2 \cos\phi - \sin\phi_2 \sin\phi)]\, e^{-jkr\cos\phi} d\phi \end{aligned} \tag{4.52}$$

Expanding the integrand yields

$$\begin{aligned} N(kr) = \frac{1}{2\pi} \int_0^{2\pi} \Big\{ & \alpha_1\alpha_2 \\ & + [(\alpha_1 - 1)\alpha_2 \cos\phi_1 + (\alpha_2 - 1)\alpha_1 \cos\phi_2] \cos\phi \\ & + (\alpha_1 - 1)(\alpha_2 - 1)\cos\phi_1 \cos\phi_2 \cos^2\phi \\ & + (\alpha_1 - 1)(\alpha_2 - 1)\sin\phi_1 \sin\phi_2 \sin^2\phi \\ & + [(\alpha_1 - 1)\alpha_2 \sin\phi_1 + (\alpha_2 - 1)\alpha_1 \sin\phi_2] \sin\phi \\ & + [(\alpha_1 - 1)(\alpha_2 - 1)\cos\phi_2 \sin\phi_1 \\ & \quad + (\alpha_1 - 1)(\alpha_2 - 1)\cos\phi_1 \sin\phi_2] \sin\phi \Big\} \\ & e^{-jkr\cos\phi} d\phi. \end{aligned} \tag{4.53}$$

Only four of the terms in (4.53) that are symmetric in $\pm\phi$ survive in the integration. The first term in (4.53) involves

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-jkr\cos\phi}d\phi = J_0(kr). \tag{4.54}$$

The next term involves

$$\frac{1}{2\pi}\int_0^{2\pi} \cos\phi e^{-jkr\cos\phi}d\phi = jJ_1(kr). \tag{4.55}$$

The next term involves

$$\frac{1}{2\pi}\int_0^{2\pi} \cos^2\phi e^{-jkr\cos\phi}d\phi = \frac{J_0(kr) - J_2(kr)}{2}. \tag{4.56}$$

The final non-zero term involves

$$\frac{1}{2\pi}\int_0^{2\pi} \sin^2\phi e^{-jkr\cos\phi}d\phi = \frac{J_0(kr) + J_2(kr)}{2}. \tag{4.57}$$

The resulting numerator for the spatial coherence function is therefore

$$\begin{aligned} N_{12}(kr) = {} & \alpha_1\alpha_2 J_0(kr) \\ & +(\alpha_1 - 1)(\alpha_2 - 1)\cos\phi_1\cos\phi_2(J_0(kr) - J_2(kr))/2 \\ & +(\alpha_1 - 1)(\alpha_2 - 1)\sin\phi_1\sin\phi_2(J_0(kr) + J_2(kr))/2 \\ & +j[\alpha_2\cos\phi_1(1-\alpha_1) + \alpha_1\cos\phi_2(1-\alpha_2)]J_1(kr). \end{aligned} \tag{4.58}$$

Appendix C

The numerator of the spatial coherence function for Nth-order differential microphones in a cylindrically correlated sound field involves integrals of the following form:

$$I_n = \frac{1}{\pi}\int_0^{\pi} \cos^n\phi e^{-jkr\cos\phi}d\phi. \tag{4.59}$$

Abramowitz [7] defines the nth-order Bessel function of the first kind as

$$J_n(kr) = \frac{(-j)^n}{\pi}\int_0^{\pi} \cos(n\phi)e^{-jkr\cos\phi}d\phi. \tag{4.60}$$

Therefore it remains to find the relationship between $\cos^n\phi$ and $\cos(n\phi)$. This relationship can be easily obtained by using Euler's relation and the binomial theorem. Using Euler's relation,

$$\cos^n\phi = \frac{1}{2^n}\left(e^{j\phi} + e^{-j\phi}\right)^n. \tag{4.61}$$

Using the binomial theorem,

$$\begin{aligned}\cos^n\phi = \frac{1}{2^n}[&e^{jn\phi} + C(n,1)e^{j(n-1)\phi}e^{-j\phi} \\ &+C(n,2)e^{j(n-2)\phi}e^{-2j\phi} \\ &+\cdots \\ &+C(n,n-1)e^{-j(n-1)\phi}e^{j\phi} + e^{-jn\phi}],\end{aligned} \tag{4.62}$$

where the function C is the binomial coefficient [7]

$$C(n,m) = \frac{n!}{(n-m)!m!}. \tag{4.63}$$

Combining terms and invoking (4.60) yields

$$\begin{aligned}I_n &= \frac{1}{2^{n-1}}\left[\sum_{m=0}^{n/2}\varepsilon(n,m)(-j)^{n-2m}C(n,m)J_{n-2m}(kr)\right], \quad n \text{ even} \\ I_n &= \frac{1}{2^{n-1}}\left[\sum_{m=0}^{(n-1)/2}(-j)^{n-2m}C(n,m)J_{n-2m}(kr)\right], \quad n \text{ odd.}\end{aligned} \tag{4.64}$$

where $\varepsilon(n,m)$ is defined as,

$$\begin{aligned}\varepsilon(n,m) &= 1, \ m \neq n/2, \\ &= \frac{1}{2}, \ m = n/2.\end{aligned} \tag{4.65}$$

References

1. B. Widrow, et al., "Adaptive noise cancelling, principles and applications," *Proc. IEEE*, vol. 63, pp. 1692-1716, 1975.
2. J.J. Rodriquez, "Adaptive noise reduction in aircraft communication systems," MIT Lincoln Lab, Lexington, MA, Tech. Rep. 756, Jan. 1987.
3. T. J. Sutton, S. J. Elliott, I. Moore, "Use of nonlinear controllers in the active attenuation of road noise inside cars," *Proc. of Recent Advances in Active Control of Sound and Vibration* , C. A. Rogers and C. R. Fuller, Eds., Blacksburg VA, 1991, pp. 682-690.
4. M.M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol.46, pp. 497-511, 1966.
5. D. Andrea, and M. Topf, "Noise cancellation apparatus," US Patent US05673325, Sept. 1997.
6. R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 27, pp.1072-1077 1955.
7. M. Ambramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover Publications, NY, 1970.

8. G. W. Elko, "Superdirectional microphone arrays," in *Acoustic Signal Processing for Telecommuncation*, S.L. Gay and J. Benesty, Eds., Kluwer Academic Publishers, Chapter 10, pp. 181-237, 2000.
9. M. M. Goulding, and J. S. Bird, "Speech enhancement for mobile telephony," *IEEE Trans. Vehicular Tech.*, vol. 39, pp. 316-326, 1990.
10. C. T. Morrow, "Point-to-point correlation of sound pressures in reverberant chambers," *J. Sound Vib.*, vol. 16, pp. 29-42, 1971.

5 Robust Adaptive Beamforming

Osamu Hoshuyama and Akihiko Sugiyama

NEC Media Research Labs, Kawasaki, Japan

Abstract. This chapter presents robust adaptive beamforming techniques designed specifically for microphone array applications. The basics of adaptive beamformers are first reviewed with the Griffiths-Jim beamformer (GJBF). Its robustness problems caused by steering vector errors are then discussed with some conventionally proposed robust beamformers. As better solutions to the conventional robust beamformers, GJBFs with an adaptive blocking matrix are presented in the form of a microphone array. Simulation results and real-time evaluation data show that a new robust adaptive microphone array achieves improved robustness against steering vector errors. Good sound quality of the output signal is also confirmed by a subjective evaluation.

5.1 Introduction

Beamforming is a technique which extracts the desired signal contaminated by interference based on directivity, i.e. spatial signal selectivity [1]–[5]. This extraction is performed by processing the signals obtained by multiple sensors such as microphones, antennas, and sonar transducers located at different positions in the space. The principle of beamforming has been known for a long time. Because of the vast amount of necessary signal processing, most research and development effort has been focused on geological investigations and sonar, which can afford a higher cost. With the advent of LSI technology, the required amount of signal processing has become relatively small. As a result, a variety of research projects where acoustic beamforming is applied to consumer-oriented applications, have been carried out [6].

Applications of beamforming include microphone arrays for speech enhancement. The goal of speech enhancement is to remove undesirable signals such as noise and reverberation. Among research areas in the field of speech enhancement are teleconferencing [7]–[8], hands-free telephones [9]–[11], hearing aids [12]-[21], speech recognition [22]–[23], intelligibility improvement [24]–[25], and acoustic measurement [26].

Beamforming can be considered as multidimensional signal processing in space and time. Ideal conditions assumed in most theoretical discussions are not always maintained. The target DOA (direction of arrival), which is assumed to be stable, does change with the movement of the speaker. The sensor gains, which are assumed uniform, exhibit significant distribution. As a result, the performance obtained by beamforming may not be as good as

is expected. Therefore, robustness against steering-vector errors caused by these array imperfections are becoming more and more important.

This chapter presents robust adaptive beamforming with the emphasis on microphone arrays as its application. In Section 2, the basics of adaptive beamformers are reviewed with the Griffiths-Jim beamformer (GJBF). Section 3 discusses robustness problems in the GJBF. Robust adaptive microphone arrays as solutions to the robustness problem are presented in Section 4. Finally in Section 5 evaluations of a robust adaptive microphone array are presented with simulation results and real-time evaluation data.

5.2 Adaptive Beamformers

A beamformer which adaptively forms its directivity pattern is called an adaptive beamformer. It simultaneously performs beam steering and null steering. In most acoustic beamformers, however, only null steering is performed with an assumption that the target DOA is known *a priori*. Due to adaptive processing, deep nulls can be developed even when errors in the propagation model exist. As a result, adaptive beamformers naturally exhibit higher interference suppression capability than its fixed counterpart. Among various adaptive beamformers, the Griffiths-Jim beamformer (GJBF) [27], or the generalized sidelobe canceler, is most widely known.

Figure 5.1 depicts the structure of the GJBF. It comprises a fixed beamformer (FBF), a multiple-input canceler (MC), and a blocking matrix (BM). The FBF is designed to form a beam in the look direction so that the target signal is passed and all other signals are attenuated. On the contrary, the BM forms a null in the look direction so that the target signal is suppressed and all other signals are passed through.

The simplest structure for the BM is a delay-and-subtract beamformer which was described in the previous section. Assuming a look direction perpendicular to the array surface, no delay element is necessary. Thus, a set of subtracters which take the difference between the signals at the adjacent microphones can be used as a BM. This structure is actually the one shown in Fig. 5.1. The BM was named after its function, which is to block the target signal.

The MC is composed of multiple adaptive filters each of which is driven by a BM output, $z_n(k)$ (n=0, 1, $\cdots$, $N-2$). The BM outputs, $z_n(k)$, contain all the signal components except that in the look direction. Based on these signals, the adaptive filters generate replicas of components correlated with the interferences. All the replicas are subtracted from a delayed output signal, $b(k - L_1)$,[1] of the fixed beamformer which has an enhanced target signal component. As a result, in the subtracter output $y(k)$, the target signal is

[1] The L_1-sample delay is introduced to compensate for the signal processing delay in the BM and the MC.

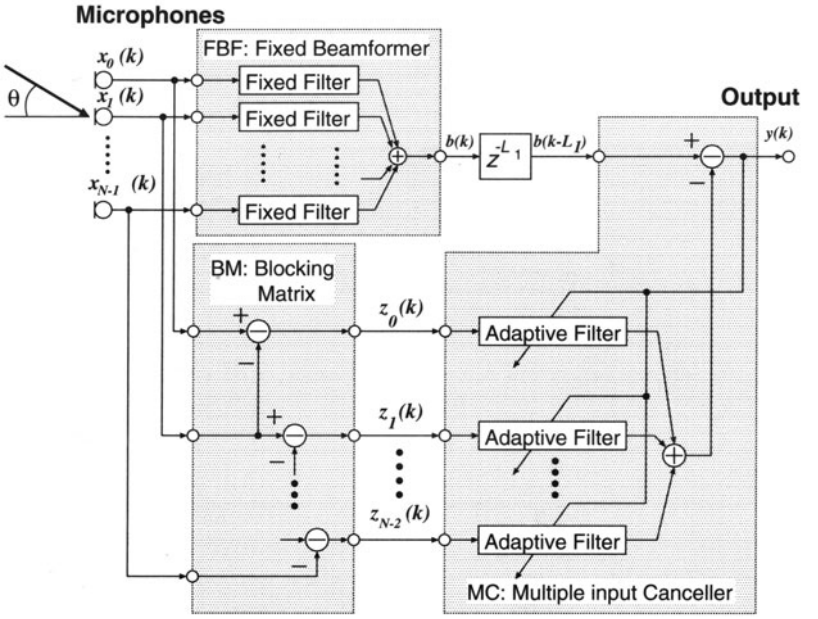


Fig. 5.1. Griffiths-Jim beamformer. It comprises a fixed beamformer (FBF), a multiple-input canceler (MC), and a blocking matrix (BM).

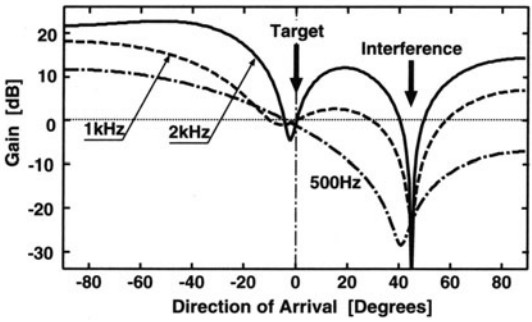


Fig. 5.2. Example directivity pattern of the Griffiths-Jim beamformer.

enhanced and undesirable signals such as ambient noise and interferences are suppressed.

The GJBF can be considered as an adaptive noise canceler with multiple reference signals, each of which is preprocessed by the BM. In an adaptive noise canceler, the auxiliary microphone is located close to the noise source to obtain a best possible noise reference. On the other hand, the BM in the

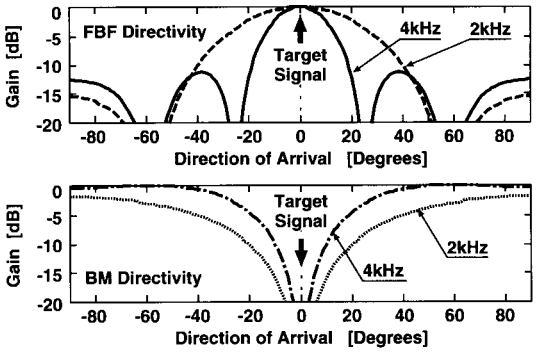


Fig. 5.3. Directivity pattern of a fixed beamformer (FBF) and a blocking matrix (BM).

GJBF extracts, with its directivity, the signal components correlated with the noise.

Figure 5.2 depicts an example directivity pattern obtained by the GJBF. In the direction of the target signal, almost constant gains close to 0 dB are obtained over a wide range of frequencies. On the contrary, in the direction of the interference, a deep null is formed. Although the directivity has frequency dependency, target signal extraction and interference suppression are simultaneously achieved.

With the same microphone array, adaptive beamformers generally achieve better interference suppression than fixed beamformers. This is because nulls are sharper than beams. The effect is demonstrated in Fig. 5.3, where directivity patterns of the FBF and the BM are illustrated. The null of the BM and the main lobe (beam) of the FBF are located in the target direction. It is also clear from the figure that they are orthogonal to each other. The BM in Fig. 5.1 has a simple delay-and-sum structure, however, a filter-and-sum beamformer [28,29] may also be employed.

5.3 Robustness Problem in the GJBF

The GJBF suffers from target-signal cancellation due to steering-vector errors, which is caused by an undesirable phase difference between $x_n(k)$ and $x_{n+1}(k)$ for the target. A phase error leads to target signal leakage into the BM output signal. As a result, blocking of the target becomes incomplete, which results in target signal cancellation at the microphone array output.

Steering-vector errors are inevitable because the propagation model does not always reflect the nonstationary physical environment. The steering vector is sensitive to errors in the microphone positions, those in the microphone characteristics, and those in the assumed target DOA (which is also known

as the look direction). For teleconferencing and hands-free communication in the car, the error in the assumed target DOA is the dominant factor.

A variety of techniques to reduce target-signal cancellation have been proposed mainly in the field of antennas and radars. The beamformers with these techniques are called robust beamformers. Typical approaches are reduction of the target-signal leakage in the BM outputs and restraint of coefficient growth in the MC. The former can be considered as a direct approach which reduces the target leakage in the BM output. The latter takes the form of an indirect approach. Even if there is target leakage in the BM output used as the MC input, the MC tries to minimize its influence.

Techniques to reduce target-signal leakage include:

- Target Tracking: The look direction is steered to the continuously estimated DOA [30]–[32]. Mistracking to interference may occur in the absence of a target signal.
- Multiple Constraints in BM: Multiple constraints are imposed on the BM so that signals from multiple DOAs are eliminated [33]. To compensate for the loss of the degrees of freedom for interference reduction with a large DOA error, additional microphones are needed.
- Constrained Gradient for Look-Direction Sensitivity: Gradient of the sensitivity at the look direction is constrained for a smaller variance of the sensitivity [34,35]. For a large error, loss in the degrees of freedom is inevitable.
- Improved Spatial Filter: A carefully designed spatial filter is used to eliminate the target signal [28]. Such a spatial filter also loses degrees of freedom.

Techniques that attempt to restraint excess coefficient growth include:

- Noise Injection: Artificially-generated noise is added to the error signal used to update the adaptive filters in the MC. This noise causes errors in the adaptive filter coefficients, preventing tap coefficients from growing excessively [36]. A higher noise level is needed to allow a larger look-direction error, resulting in less interference suppression.
- Norm Constraint: The coefficient norm of the adaptive filters in the MC is constrained by an inequality to suppress the growth of the tap coefficients [37]. In spite of its simplicity, interference reduction is degraded when the constraint is designed to allow a large error.
- Leaky Adaptive Algorithm: A leaky coefficient adaptation algorithm such as leaky LMS is used for the adaptive filters in the MC [28]. A large leakage is needed to allow a large look-direction error, leading to degraded interference-reduction.
- Adaptation Mode Control: Coefficient adaptation in the MC is controlled so that adaptation is carried out only when there is no target signal [38]. If there is no target signal when coefficients are adapted in the MC, the target leakage, if any, will have no effect on the performance of the beamformer.

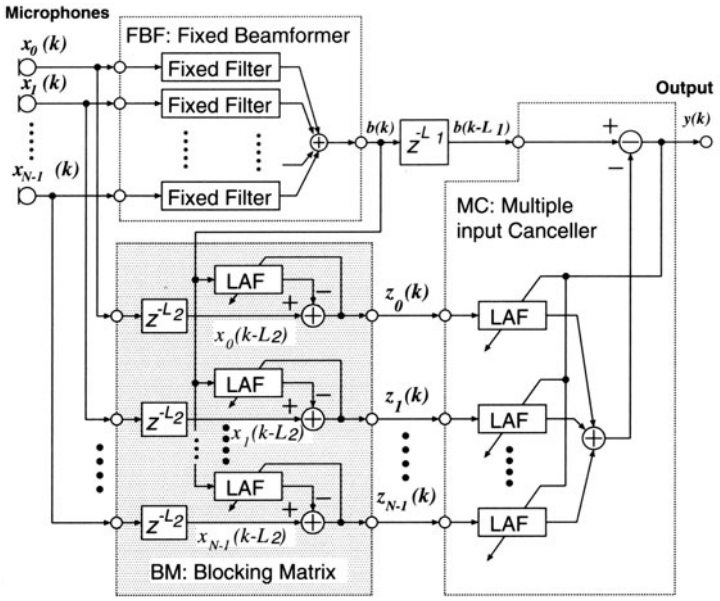


Fig. 5.4. GJBF with a LAF-LAF Structure.

These methods have been developed for a small look-direction error, typically less than 10 degrees. In the case of microphone arrays, the variance of the target DOA is typically much larger than in antennas and radar applications. No single conventional technique for robustness is sufficient for microphone arrays with a larger phase errors.

5.4 Robust Adaptive Microphone Arrays — Solutions to Steering-Vector Errors

5.4.1 LAF-LAF Structure

A target-tracking method with leaky adaptive filters (LAF) in the BM was proposed as a solution to target signal cancellation in [39]. It is combined with leaky adaptive filters in the MC [28], thereby called a LAF-LAF structure. Figure 5.4 depicts its block diagram. The leaky adaptive filters in the BM alleviate the influence of phase error, which results in the robustness. This structure can pick up a target signal with little distortion when the error between the actual and the assumed DOAs is not small. It does not need matrix products, and provides easy implementation. The nth output

$z_n(k)(n = 0, 1, \ldots, N-1)$ of the BM can be obtained as follows:

$$z_n(k) = x_n(k - L_2) - \mathbf{h}_n^T(k)\mathbf{b}(k), \tag{5.1}$$

$$\mathbf{h}_n(k) \triangleq [h_{n,0}(k), h_{n,1}(k), \ldots, h_{n,M_1-1}(k)]^T, \tag{5.2}$$

$$\mathbf{b}(k) \triangleq [b(k), b(k-1), \ldots, b(k - M_1 + 1)]^T, \tag{5.3}$$

where $[\cdot]^T$ denotes vector transpose and $x_n(k)$ is the nth microphone signal. L_2 is the number of delay samples for causality, $\mathbf{h}_n(k)$ is the coefficient vector of the nth LAF, and $\mathbf{b}(k)$ is the signal vector consisting of delayed signals of $b(k)$ (which is the FBF output). Each LAF is assumed to have M_1 taps. The adaptation by the normalized LMS (NLMS) algorithm [40] is described as follows:

$$\mathbf{h}_n(k+1) = \mathbf{h}_n(k) - \delta \cdot \mathbf{h}_n(k) + \alpha \frac{z_n(k)}{\mathbf{b}(k)^T\mathbf{b}(k)}\mathbf{b}(k), \tag{5.4}$$

where α is the step size for the adaptation algorithm, and $\delta, 0 \leq \delta \leq 1$, is the leakage constant.

LAFs are also used in the MC for enhancing the robustness obtained in the BM. The LAFs prevent undesirable target-signal cancellation caused by the remaining correlation with the target signal in $z_n(k)$. Tap coefficient vectors $\mathbf{w}_n(k)$ of the MC have M_2 taps and are updated by an equation similar to (5.4), where $\mathbf{h}_n$, $\mathbf{b}$, and $z_n(k)$ are replaced with $\mathbf{w}_n$, $\mathbf{z}_n$, and $y(k)$, respectively. The leakage constant δ and the step size α are replaced with γ and β respectively, and may take different values from those in (5.4).

With the LAFs in the BM, the LAF-LAF structure adaptively controls the look direction, which is fixed in the GJBF. Due to robustness by the adaptive control of the look direction, the LAF-LAF structure does not lose degrees of freedom for interference reduction. Thus, no additional microphones are required compared to the conventional robust beamformers. Target signal leakage in the BM is sufficiently small to use a minimum leakage constant γ in the MC even for a large look-direction error. Such a value of γ leads to a higher interference-reduction performance in the MC. The output of the LAFs are summed and subtracted from an L_1 sample delayed version of the FBF output to generate the microphone array output $y(k)$.

The width of the allowable DOA for the target is determined by the leaky constants and the step sizes in both the BM and the MC. Generally, smaller values of these parameters make the allowable target DOA wider. The allowable DOA width for the target is not a simple function of the parameters, however, and is not easy to prescribe. It is reported [39] that the interference is attenuated by more than 18 dB when it is designed, through simulations, to allow 20 degree directional error. Tracking may not be sufficiently precise for a large tracking range. Thus, there is a trade-off between the degree of target-signal cancellation and the amount of interference suppression.

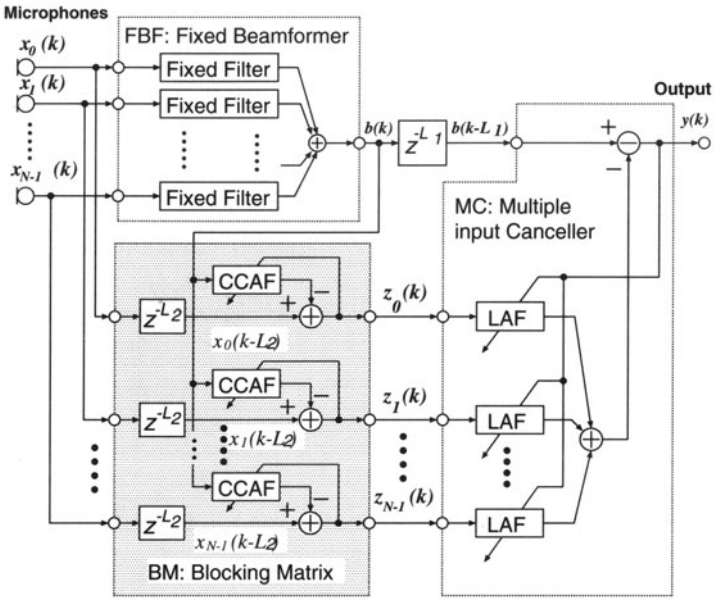


Fig. 5.5. GJBF with a CCAF-LAF Structure.

5.4.2 CCAF-LAF Structure

A more effective solution is to use coefficient-constrained adaptive filters (CCAFs) in the BM [41,42]. When combined with leaky adaptive filters in the MC as depicted in Fig. 5.5, the result is called a CCAF-LAF structure. CCAFs behave like adaptive noise cancelers. The input signal to each CCAF is the output of the FBF, and the output of the CCAF is subtracted from the delayed microphone signal. The CCAF coefficient vectors $\mathbf{h}_n(k)$ are adapted with constraints. Adaptation by the NLMS algorithm is described as follows:

$$\mathbf{h'}_n(k+1) = \mathbf{h}_n(k) + \alpha \frac{z_n(k)}{\mathbf{b}(k)^T \mathbf{b}(k)} \mathbf{b}(k), \tag{5.5}$$

$$\mathbf{h}_n(k+1) = \begin{cases} \boldsymbol{\phi}_n, & \text{for } \mathbf{h'}_n(k+1) > \boldsymbol{\phi}_n \\ \boldsymbol{\psi}_n, & \text{for } \mathbf{h'}_n(k+1) < \boldsymbol{\psi}_n \\ \mathbf{h'}_n(k+1), & \text{otherwise.} \end{cases} \tag{5.6}$$

$$\boldsymbol{\phi}_n \triangleq [\phi_{n,0}, \phi_{n,1}, \cdots, \phi_{n,M_1-1}]^T, \tag{5.7}$$

$$\boldsymbol{\psi}_n \triangleq [\psi_{n,0}, \psi_{n,1}, \cdots, \psi_{n,M_1-1}]^T, \tag{5.8}$$

where each CCAF is assumed to have M_1 taps and $\mathbf{h'}_n(k+1)$ is a temporal coefficient vector for limiting functions. $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$ are the upper and lower

bounds for coefficients. In the output signal $z_n(k)$, the components correlated with $\mathbf{b}(k)$ are cancelled by the CCAFs.

Each coefficient of the CCAFs is constrained based on the fact that filter coefficients for target-signal minimization vary significantly with the target DOA. An example of filter-coefficient variation is illustrated in Fig. 5.6. By the design of the constrained regions of the CCAF coefficients, the maximum allowable look-direction error can be specified. For example, when the CCAF coefficients are constrained in the hatched region in Fig. 5.6, up to 20° error in look direction could be allowed. Only the signal that arrives from a DOA in the limited DOA region is minimized at the outputs of the BM and remains at the output of the MC. If no interference exists in the region, which is common with microphone arrays, no mistracking occurs. For details on the design of upper and lower bounds, refer to [42].

Figure 5.7 illustrates a qualitative comparison between the LAF and the CCAF with respect to look-direction error and coefficient error from the optimum for signal blocking. Both the CCAF and the LAF give error characteristics approximating the ideal nonlinearity for target tracking. However, the coefficient error of the CCAF is a better approximation to the ideal nonlinearity than that of the LAF as shown by Fig. 5.7. The coefficient error of the CCAF becomes effective only when the look-direction error exceeds the threshold, otherwise it has no effect. On the other hand, the coefficient error of the LAF varies continuously with the look-direction error. Therefore, the CCAF leads to precise target tracking, which results in sharper spatial selectivity and less target-signal cancellation.

5.4.3 CCAF-NCAF Structure

It is possible to combine the BM with CCAFs [42] and the MC with norm-constrained adaptive filters (NCAFs) [37]. This is a CCAF-NCAF structure [43]. NCAFs subtract from $b(k-L_1)$ the components correlated with $z_n(k)$ $(n=0,...,N-1)$. Let M_2 be the number of taps in each NCAF, and let $\mathbf{w}_n(k)$ and $\mathbf{z}_n(k)$ be the coefficient vector and the signal vector of the nth NCAF, respectively. The signal processing in the MC is described by

$$y(k) = b(k-L_1) - \sum_{n=0}^{N-1} \mathbf{w}_n^T(k)\mathbf{z}_n(k), \tag{5.9}$$

where

$$\mathbf{w}_n(k) \triangleq [w_{n,0}(k), w_{n,1}(k), \ldots, w_{n,M_2-1}(k)]^T, \tag{5.10}$$

$$\mathbf{z}_n(k) \triangleq [z_n(k), z_n(k-1), \ldots, z_n(k-M_2+1)]^T. \tag{5.11}$$

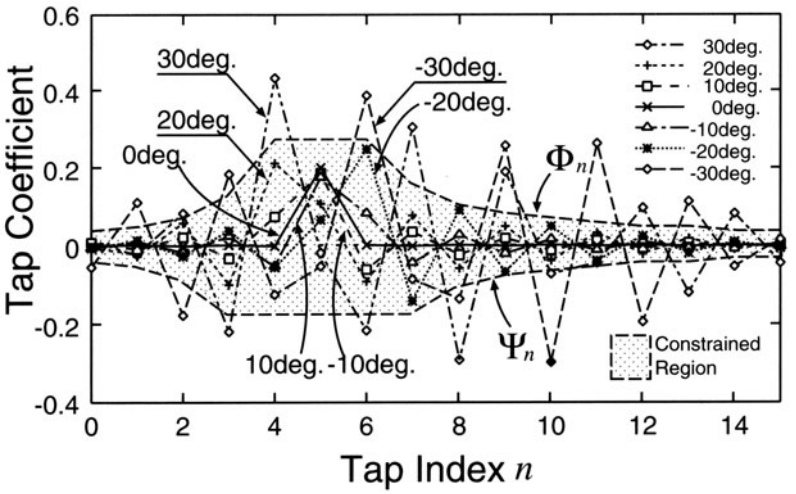


Fig. 5.6. An example of CCAF coefficients to minimize signals from different DOAs and their constraints. When the CCAF coefficients are constrained in the hatched region, up to 20° error in look direction could be allowed.

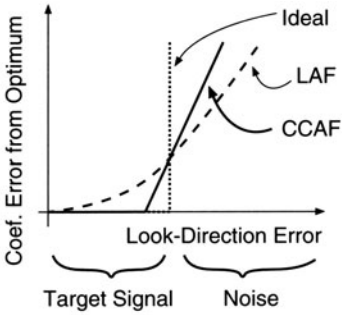


Fig. 5.7. Comparison of selectivity in LAF and CCAF.

Coefficients of the NCAFs are updated by an adaptive algorithm with a norm constraint. Adaptation with the NLMS algorithm is described as follows:

$$\mathbf{w}'_n = \mathbf{w}_n(k) + \beta \frac{y(k)}{\mathbf{z}_j(k)^T \mathbf{z}_j(k)} \mathbf{z}_n(k), \tag{5.12}$$

$$\Omega = \mathbf{w}'^T_n \mathbf{w}'_n, \tag{5.13}$$

$$\mathbf{w}_n(k+1) = \begin{cases} \sqrt{\frac{K}{\Omega}}\, \mathbf{w}'_n & \text{for } \Omega > K \\ \mathbf{w}'_n & \text{otherwise,} \end{cases} \tag{5.14}$$

where β and $\mathbf{w}'_n$ are a step size and a temporal vector for the constraint, respectively. Ω and K are the total squared-norm of $\mathbf{w}_n(k)$ and a threshold. If Ω exceeds K, $\mathbf{w}_n(k+1)$ are restrained by scaling. The norm constraint by scaling restrains excess growth of tap coefficients. The restraint inhibits the undesirable target cancellation when the target signal leaks into the NCAF inputs. If the outputs of the BM have no target signal, the MC cancels only the interference signals. In this ideal case, a norm constraint in the MC is not needed. However, complete rejection of the target signal is almost impossible in the BM, because actual environments have reflection and reverberation. To completely cancel the target signal in a reverberant environment, more than 1,000 taps are needed for each CCAF in the BM. Such a large number of taps leads to slow convergence, large misadjustment, and increased computation. Even with a high-speed processor and a fast convergence algorithm, misadjustment with the adaptive filters is inevitable. Adaptation with a low signal-to-interference ratio (SIR) causes additional misadjustment by the interference, which leads to leakage of the target signal at the BM outputs. Therefore, to avoid the target signal cancellation by leakage, a restraint with the MC such as the NCAF is essential. Because the CCAF-NCAF structure loses no degrees of freedom for interference reduction in the BM, it is robust to large look-direction errors with a small number of microphones.

5.4.4 CCAF-NCAF Structure with an AMC

Adaptations in the BM and in the MC should be performed alternately. This is because the relationship between the desired signal and the noise for the adaptation algorithm in the BM is contrary to that in the MC. For the adaptation algorithm in the BM, the target signal is the desired signal and the noise is the undesired signal. In the MC, however, the noise is the desired signal and the target signal is the undesired signal.

In the robust adaptive beamformers discussed so far, it was implicitly assumed that adaptive filters in the BM are adapted only when the target is active and those in the MC are adapted only when the target is inactive. In a real environment, however, the situation is not so simple, since incorrect adaptation of the BM may cause incomplete target blocking. As a result, the MC directivity may have a null in the direction of the target signal, resulting in target-signal cancellation. Combined with target tracking by the BM, adapting coefficients only when the target signal is absent is an effective strategy for adding robustness to adaptive beamforming [38]–[45]. In order to discriminate active and inactive periods of the target, an adaptation mode controller (AMC) is necessary.

The CCAF-NCAF structure with an AMC [46] depicted in Fig. 5.8 uses a mixed approach of the BM with CCAFs, the MC with NCAFs [37], and an AMC. A BM consisting of CCAFs provides a wider null for the target with sharper edges than leaky adaptive filters. An MC comprising NCAFs reduces

undesirable target-signal cancellation when the MC inputs have some leakage from the target signal.

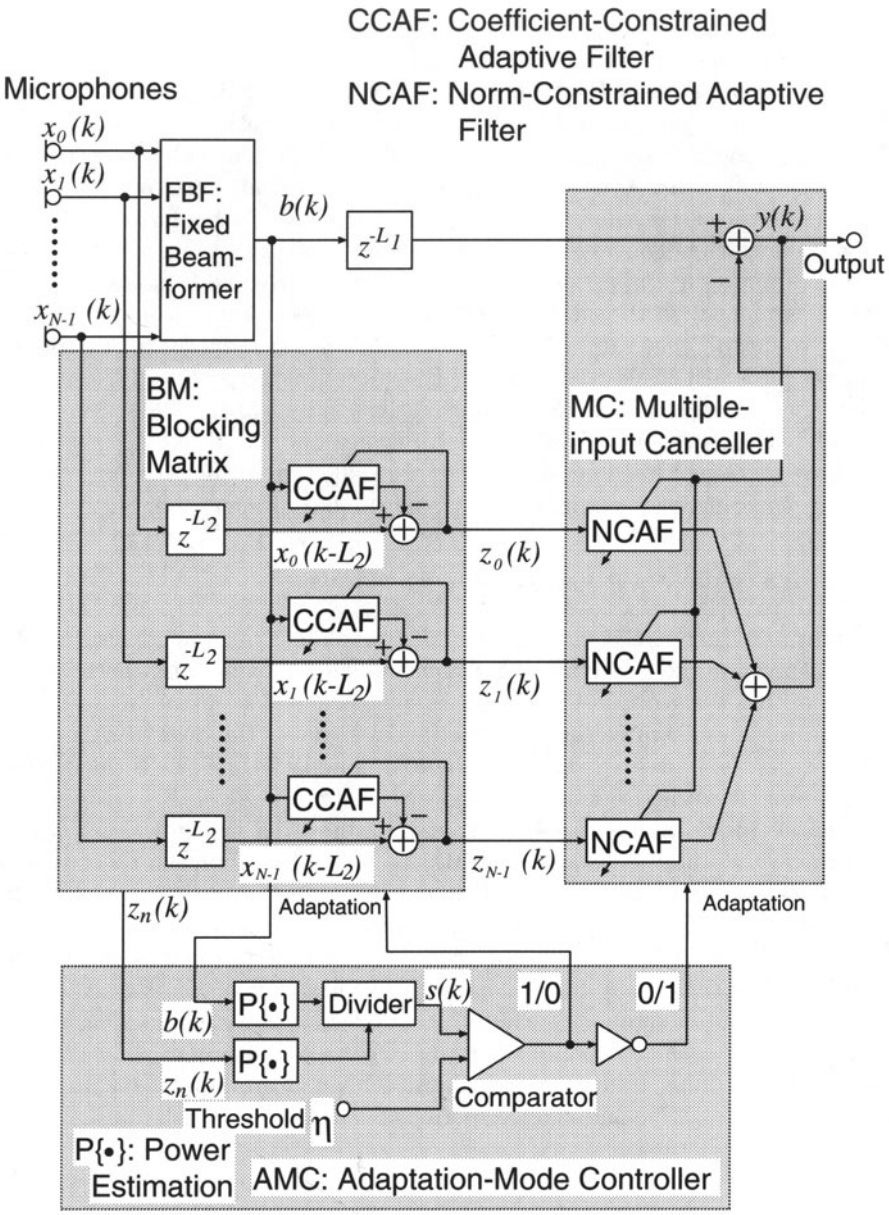


Fig. 5.8. CCAF-NCAF structure with an AMC.

The AMC controls adaptation of the BM and the MC by target-signal detection based on an estimate of the SIR [46]. The SIR is estimated as a power ratio of the output signal $b(k)$ of the FBF, to the output signal $z_n(k)$ of the BM. The main component in the FBF output is the target signal and that in the BM output is the noise. Therefore, the power ratio $s(k)$ can be considered as a direct estimate of the SIR. When the ratio is larger than a threshold η, the adaptation of the BM is performed. Otherwise, the MC is adapted.

5.5 Software Evaluation of a Robust Adaptive Microphone Array

The GJBF with CCAF-NCAF structure combined with an AMC (GJBF-CNA) was evaluated in a computer-simulated anechoic environment and in a real environment with reverberation. In the former environment, it was compared with conventional beamformers in terms of sensitivity pattern. In the latter environment, it was evaluated objectively by SIR and subjectively by mean opinion score (MOS).

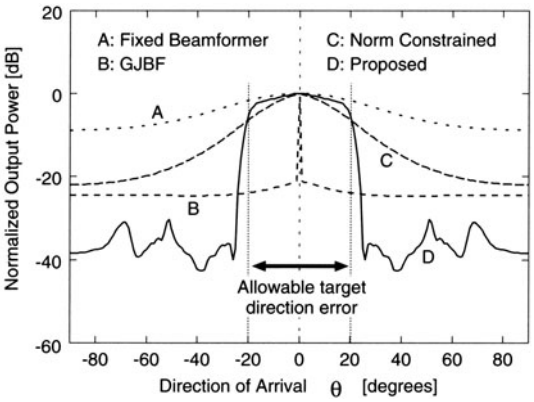


Fig. 5.9. Normalized output power after convergence as a function of DOA.

5.5.1 Simulated Anechoic Environment

A four-channel equi-spaced broadside array was used for these simulations. The spacing between microphones was 4.1 cm.The sampling rate was 8 kHz. The FBF used was a simple beamformer whose output is given by

$$b(k) = \frac{1}{N} \sum_{n=0}^{N-1} x_n(k). \tag{5.15}$$

The first simulation investigated sensitivity (after convergence) as a function of the single-sided DOA. Band-limited (0.3–3.7 kHz) Gaussian signals were used, and the assumed target direction was 0°. The maximum allowable target-direction error was 20°, unless otherwise stated. The number of coefficients for all the CCAFs and all the NCAFs was 16. The parameters were L_1=10, L_2=5, K=10.0, α=0.1, and β=0.2. The constraints of the CCAF were set based on the arrangement of the simulated array and maximum allowable target-direction errors. Total output powers after convergence, normalized by the power of the assumed target direction, are plotted in Fig. 5.9.

The plots are of the FBF (FBF), simple GJBF [27] (GJBF), norm constrained method [37] (Norm Constrained), and the GJBF-CNA (Proposed). The solid line D shows that the GJBF-CNA achieves both robustness against 20° target-direction error and high interference-reduction performance (which is 30 dB at $\theta = \pm 30°$). Similar results for a colored signal instead of the band-limited Gaussian signal have been obtained [43]. The directivity pattern of the GJBF-CNA is slightly degraded for a colored signal. However, the degradation by the norm-constrained method is more serious. This fact shows that the GJBF-CNA exhibits robustness to the power spectrum of input signal.

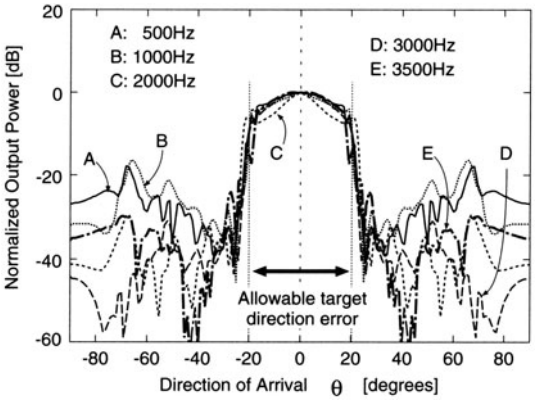


Fig. 5.10. Sensitivities after convergence as a function of DOA at different frequencies.

Frequency dependency of the directivity pattern is shown in Fig. 5.10. In this figure, sensitivities to the frequency component of the target signal are plotted. Frequency dependency of the GJBF-CNA is small, and thus, the GJBF-CNA is suitable for broadband applications such as microphone arrays. The widths of the high-sensitivity regions are almost the same as the allowable target-direction error ($-20° < \theta < 20°$) and the sensitivity in the region is constant.

In the second simulation, sensitivities for different SIRs were investigated. The simulation was performed with amplitude control that was similar to a realistic scenario. A target signal source generated a band-limited white Gaussian signal for the first 50,000 iterations and then stopped. This is a simple simulation of burst characteristics like speech. Another bandlimited white Gaussian signal, which imitates an interference like airconditioner noise, existed throughout the simulation. The SIR is defined as a power ratio of the two signals. The target signal source was placed about 10° off the assumed target DOA and the DOA of the interfering signal source was scanned.

Figure 5.11 shows normalized output power after convergence as a function of interference DOA. Lines G and H have a sharp peak at $\theta = 10°$, which indicates that the target-signal at the output of the BM is sufficiently minimized for the overall robustness. Therefore, when SIR is higher than about 10dB (which is lower than a typical SIR value expected in teleconference) the interference is suppressed even if it arrives from a direction in the allowable target DOA region. When the interference comes from outside the allowable target DOA region, even an SIR of 0 dB causes almost no problem in the GJBF-CNA.

Finally, Fig. 5.12 shows the total output powers for various coefficient constraints with the CCAFs. The signal was bandlimited white Gaussian noise. The allowable target-direction errors are approximately 4, 6, 9, 12, 16, and 20 degrees. These lines demonstrate that the allowable target-direction error can be specified by the user.

5.5.2 Reverberant Environment

Simulations with real sound data captured in a reverberant environment were also performed. The data were recorded with a broad-side linear array. Four omni-directional microphones without calibration were mounted on a universal printed circuit board with an equal spacing of 4.1 cm. The signal of each microphone was bandlimited between 0.3 and 3.4 kHz and sampled at 8 kHz. The number of taps was 16 for both the CCAFs and NCAFs.

Figure 5.13 illustrates the arrangement for sound-data acquisition. The target source was located in front of the array at a distance of 2.0 m. A white noise source was placed about $\theta = 45°$ off the target DOA at a distance of 2.0 m. The reverberation time of the room was about 0.3 second, which is common with actual small offices. All the parameters except the step-sizes were the same as those in the previous subsection. The target source was an English male speech signal.

Objective Evaluation

Output powers for the FBF, the GJBF [27] (GJBF), and the norm-constrained method [37] (Norm Constrained) after convergence are shown in Fig. 5.14. The step-size α for the CCAFs was 0.02 and β for the NCAFs was 0.004.

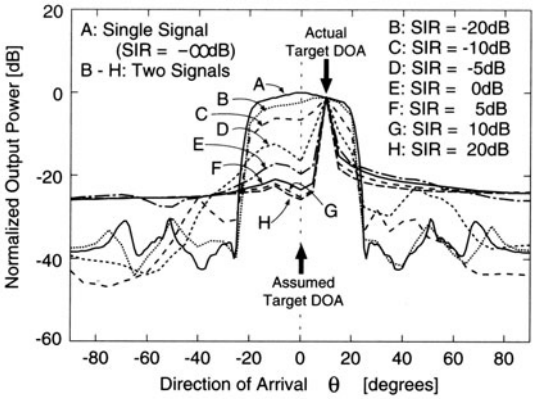


Fig. 5.11. Normalized output power after convergence as a function of DOA with different SIRs.

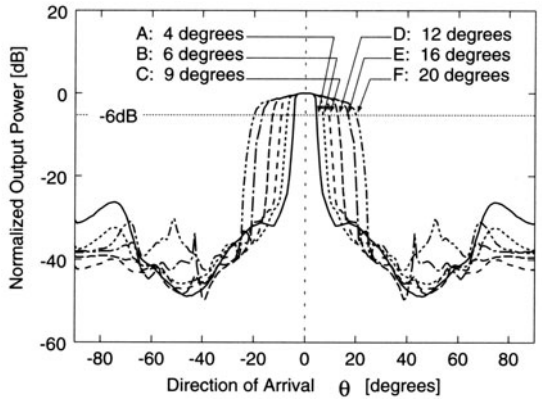


Fig. 5.12. Normalized output power after convergence for different allowable target directions.

These step-sizes were selected so that breathing noise and cancellation of the target signal are sufficiently small subjectively. All other parameters were selected based on the microphone arrangement. If there is any difference between trajectory A and any of B, C, D, E, or F when the voice is active (sample index from 1,720,000 to 1,740,000), the target signal corresponding to the trajectory is partially cancelled. The FBF (B) causes almost no target-signal cancellation. With the GJBF (C), cancellation of the target signal is serious. With the the norm-constrained method (D), and the GJBF-CNA (E), the cancellation of target signal was 2dB, which is subjectively small.

The output powers during voice absence (after sample index 1,760,000) indicate the interference-reduction ratio (IRR). The IRR of the FBF is 3dB,

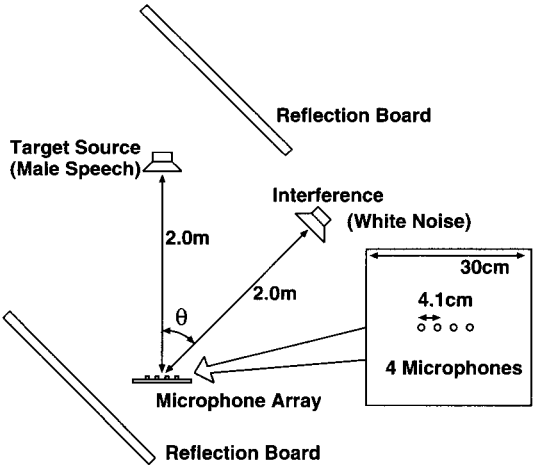


Fig. 5.13. Experimental set-up.

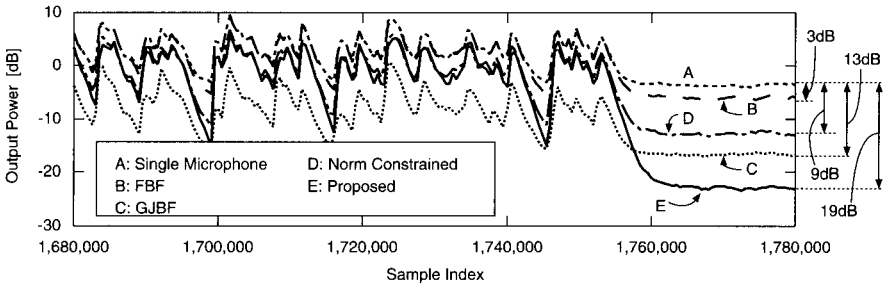


Fig. 5.14. Output Powers for a male speech signal and white noise.

and that of the norm-constrained method is 9dB. On the other hand, with the GJBF-CNA (F), the IRR is as much as 19dB.

Subjective Evaluation

MOS evaluation by 10 nonprofessional subjects was performed based on [47]. As anchors, the signal recorded by a single microphone was used for grade 1 and the original male speech without interference for grade 5. Subjects were instructed that target-signal cancellation should obtain a low score.

Evaluation results are shown in Fig. 5.15. The thick horizontal line on each bar and the number on it represent the score obtained by the corresponding method. The vertical hatched box on each bar indicates ± one standard deviation. The FBF obtained 1.7 points because the number of microphones is so small that its IRR is low. The GJBF reduced the interfer-

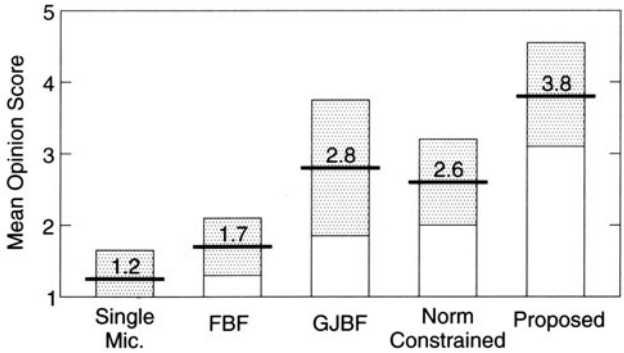


Fig. 5.15. Mean opinion score results.

ence considerably with serious target signal cancellation, thus, it was scored 2.8 points. The norm-constrained method was scored 2.6 points for its 9dB interference-reduction capability. The GJBF-CNA obtained 3.8 points, which is the highest of all the beamformers.

5.6 Hardware Evaluation of a Robust Adaptive Microphone Array

5.6.1 Implementation

The GJBF-CNA was implemented on a portable and flexible DSP system shown in Fig. 5.16 [48,49]. The system comprises a microphone array and a compact touch-panel personal computer which includes a floating point DSP, the ADSP-21062 [50]. The DSP contains a dual on-chip 2-Mbit SRAM and allows 32-bit IEEE floating-point computation. The sampling rate was software-programmed at 8 kHz.

The DSP board has a PCI (Peripheral Component Interconnect) interface, therefore, it can be connected to the PCI bus of any personal computer. A graphical interface has been developed to facilitate ease-of-use and monitoring of the implemented GJBF-CNA. It provides interactive parameter selection and displays the input and the output signals powers as well as the filter coefficients. This graphical display is useful for demonstrating the behavior of the GJBF-CNA and its performance. The system is shown in Fig. 5.16

5.6.2 Evaluation in a Real Environment

The GJBF-CNA in Fig. 5.16 was evaluated using the same linear microphone array as in the previous section. The selected step sizes were 0.02 for the ABM and 0.005 for the MC. The threshold $\eta = 0.65$ was used for the AMC. All other parameters were the same as those in the previous section.

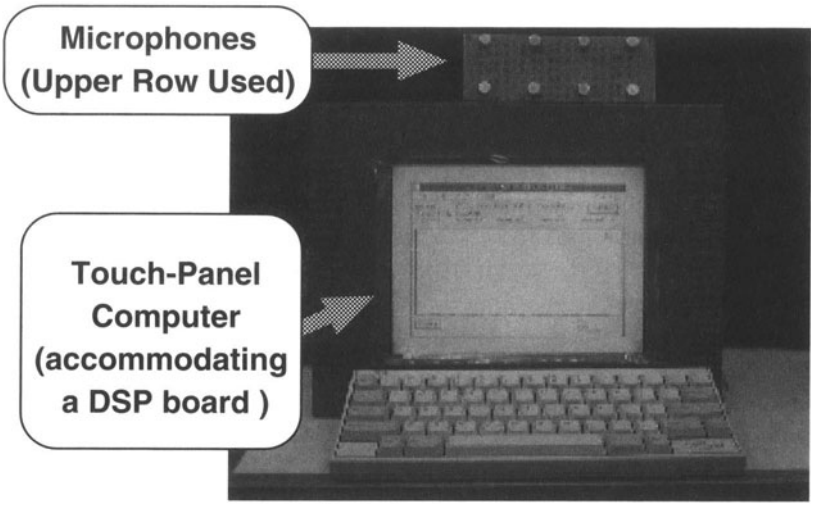


Fig. 5.16. Real-time DSP system.

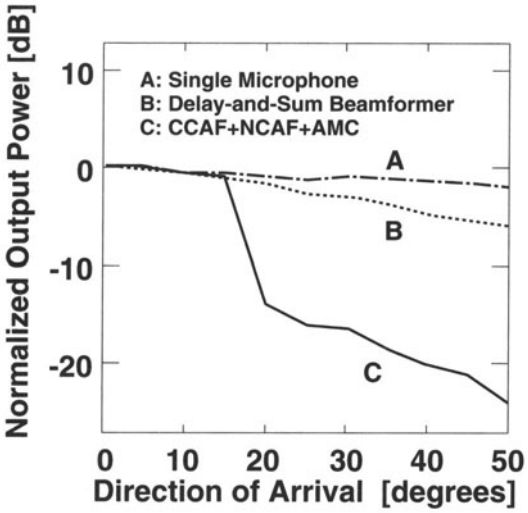


Fig. 5.17. Directivity patterns (i.e., the output powers normalized by the power at the center) measured in 5-degree intervals.

Directivity

Directivity for a single signal-source was measured. A white-noise source was scanned in two directions from 0° to 50° at a distance of 2.0 m from the array. Output powers of the system were measured in 5-degree intervals, and compared with those of a single microphone and an FBF (delay-and-sum beamformer). Figure 5.17 shows the output powers normalized by the power at the center. The figure indicates that the GJBF-CNA can suppress the interference at $\theta = 30°$ by as much as 15 dB when the allowable target DOA is set to ± 20 degrees.

Noise Reduction

Noise reduction capability was evaluated in the same room as that for directivity evaluation. There were several computers with noisy fans. In addition, two noise-generating loudspeakers were located on both sides of the array. Stereo music or white noise was used as the noise signal.

In the beginning, breathing noise due to adaptation was observed at almost every utterance. It disappeared in a second and caused almost no problem for conversation. Although the degree of noise reduction depends on the loudspeaker positions, it was typically 8 to 1 dB. These results confirm that the GJBF-CNA is a promising technique for voice communications.

5.7 Conclusion

An overview of robust adaptive beamforming techniques have been presented in this chapter, with an emphasis on systems that are robust to steering-vector errors. It has been shown that the GJBF with the CCAF-NCAF structure and an AMC (GJBF-CNA) is effective in a real environment. Integrated systems with a microphone array, a noise canceler, and an echo canceler will play a key role in future acoustic noise and echo control devices.

References

1. R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*, New York: Wiley, 1980.
2. B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, New York: Prentice-Hall, 1985.
3. S. Haykin ed., *Array Signal Processing*, Englewood Cliffs: Prentice-Hall, 1985.
4. B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, Apr. 1988.
5. D. H. Johnson and D. E. Dudgeon, *Array Signal Processing – Concepts and Techniques*, Englewood Cliffs: Prentice-Hall, 1993.
6. Special Session on Microphone Array Signal Processing, in *Proc. IEEE ICASSP'97*, vol. I, pp. 211–254, Apr. 1997.

7. J. L. Flanagan, D. A. Berkley, G. W. Elko, and M. M. Sondhi, "Autodirective microphone systems," *Acustica*, vol.73, pp. 58–71, Feb. 1991.
8. P. L. Chu, "Superdirective microphone array for a set-top video conferencing system," in *Proc. IEEE ICASSP'97*, vol. I, pp. 235–238, Apr. 1997.
9. I. Claessen, S. Nordholm, B. A. Bengtsson, and P. Eriksson, "A multi-DSP implementation of a broad-band adaptive beamformer for use in a hands-free mobile radio telephone," *IEEE Trans. Vehicular Tech.*, vol. 40, no. 1, pp. 194–202, Feb. 1991.
10. Y. Grenier, "A microphone array for car environments," *Speech Communicat.*, vol. 12, no. 1, pp. 25–39, Mar. 1993.
11. M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. IEEE ICASSP'97*, vol. I, pp. 239–242, Apr. 1997.
12. P. M. Peterson, "Using linearly-constrained adaptive beamforming to reduce interferrence in hearing aids from competing talkers in reverberant rooms," in *Proc. IEEE ICASSP'87*, 5.7.1, pp. 2364–2367, Apr. 1987.
13. P. M. Zurek, J. E. Greenberg, and P. M. Peterson, "Sensitivity to design parameters in an adaptive-beamforming hearing aid,' in *Proc. IEEE ICASSP'90*, A1.10, pp. 1129–1132, Apr. 1990.
14. W. Soede, A. J. Berkhout, and F. Bilson, "Development of a directional hearing instrument based on array technology," *J. Acoust. Soc. Amer.*, vol. 94, no. 2, pt.1, pp. 785–798, Aug. 1993.
15. W. Soede, F. Bilson, and A. J. Berkhout, "Assignment of a directional microphone array for hearing-impared listeners," *J. Acoust. Soc. Amer.*, vol. 94, no. 2, pt.1, pp. 799–808, Aug. 1993.
16. J. M. Kates, "Superdirective arrays for hearing aids," *J. Acoust. Soc. Amer.*, vol.94, no. 4, pp. 1930–1933, Oct. 1993.
17. M. W. Hoffman, T. D. Trine, K. M. Buckley, and D. J. Tasell, "Robust adaptive microphone array processing for hearing aids: Realistic speech enhancement," *J. Acoust. Soc. Amer.*, vol. 96, pp. 759–770, Aug. 1994.
18. F. Asano, Y. Suzuki, and T. Sone, "Weighted RLS adaptive beamforming with initial directivity," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 424–428, Sep. 1995.
19. J. M. Kates, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 3138–3148, May 1996.
20. E. D. McKinney, and V. E. DeBrunner, "A two-microphone adaptive broadband array for hearing aids," in *Proc. IEEE ICASSP'96*, pp. 933–936, May 1996.
21. A. Wang, K. Yao, R. E. Hudson, D. Korompis, S. F. Soli, and S. Gao, "A high performance microphone array system for hearing aid applications," in *Proc. IEEE ICASSP'96*, pp. 3197–3200, May 1996.
22. K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura, and J. Kojima, "A microphone array system for speech recognition," in *Proc. IEEE ICASSP'97*, vol. I, pp. 215–218, Apr. 1997.
23. M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. IEEE ICASSP'97*, vol. I, pp. 227–230, Apr. 1997.
24. T. Nishi, "Relation between objective criteria and subjective factors in a sound field, determined by multivariance analysis," *Acustica*, 76, pp. 153–162, 1992.

25. T. Nishi amd T. Inoue, 'Development of a multi-beam array microphone for multi-channel pickup of sound fields," *Acustica*, 76, pp. 163–172, 1992.
26. W. Täger and Y. Mahieux, "Reverberant sound field analysis using a microphone array," in *Proc. IEEE ICASSP'97*, vol. I, pp. 383–386, Apr. 1997.
27. L. J. Griffiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.
28. I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beamforming," *IEEE Trans. Antennas Propagat.*, pp. 1093–1096, Sep. 1992.
29. S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, pp. 215–227, Apr. 1996.
30. S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Processing* , vol. 5, no. 5, pp. 425–437, Sep. 1997.
31. M. H. Er and B. C. Ng, "A new approach to robust beamforming in the presence of steering vector errors," *IEEE Trans. Signal Processing*, vol. 42, no. 7, pp. 1826–1829, Jul. 1994.
32. G. L. Fudge and D. A. Linebarger, "A calibrated generalized sidelobe canceller for wideband beamforming," *IEEE Trans. Signal Processing*, pp. 2871–2875, Oct. 1994.
33. B. Widrow and M. McCool, "A comparison of adaptive algorithms based on the methods of steepest descent and random search," *IEEE Trans. Antennas Propagat.*, pp. 615–637, Sep. 1976.
34. M. H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust. Speech Signal Processing*, pp. 1378–1393, Dec. 1983.
35. M. H. Er and A. Cantoni, "An unconstrained partitioned realization for derivative constrained broad-band antenna array processors," *IEEE Trans. Acoust. Speech Signal Processing*, pp. 1376–1379, Dec. 1986
36. N. K. Jablon, "Adaptive beamforming with the generalized sidelobe canceller in the presence of array imperfections," *IEEE Trans. Antennas Propagat.*, pp.996–1012, Aug. 1986.
37. H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Processing*, pp.1365–1376, Oct. 1987.
38. J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1662–1676, Mar. 1992.
39. O. Hoshuyama and A. Sugiyama, "A robust generalized sidelobe canceller with a blocking matrix using leaky adaptive filters," *Trans. IEICE*, vol.J79-A, no.9, pp.1516–1524, Sep. 1996 (in Japanese). (English version available in *Electron. Communicat. Japan*, vol.80, no.8, pp.56–65, Aug. 1997.)
40. G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*, Englewood Cliffs: Prentice-Hall, 1984.
41. O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proc. IEEE ICASSP'96*, pp. 925–928, May 1996.
42. O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer with a blocking matrix using coefficient constrained adaptive filters," *Trans. IEICE*, vol.E82-A, no.4, pp.640–647, Apr. 1999.

43. O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.
44. O. Hoshuyama, B. Begasse, A. Sugiyama, and A. Hirano, "A robust adaptive microphone array control based on output signals of different beamformers," in *Proc. IEICE 12th DSP Symp.*, A3-4, Nov. 1997.
45. O. Hoshuyama, B. Begasse, and A. Sugiyama, "A new adaptation-mode control based on cross correlation for a robust adaptive microphone array," *IEICE Trans. Fundamentals*, vol. E84-A, Feb. 2001 (to appear).
46. O. Hoshuyama, B. Begasse, A. Sugiyama, and A. Hirano, "A real-time robust adaptive microphone array controlled by an SNR estimate," in *Proc. IEEE ICASSP'98*, pp. 3605–3678, May 1998.
47. H. R. Silbiger, "Audio Subjective Test Methods for Low Bit Rate Codec Evaluations," ISO/IEC JTC1/SC29/WG11/N0981, Jul. 1995.
48. O. Hoshuyama and A. Sugiyama, "An adaptive microphone array with good sound quality using auxiliary fixed beamformers and its DSP implementation," in *Proc. IEEE ICASSP'99*, pp. 949–952, Mar. 1999.
49. O. Hoshuyama and A. Sugiyama: "Realtime adaptive microphone array on a single DSP system," in *Proc. IWAENC'99*, pp. 92–95, Sep. 1999.
50. Analog Devices, *ADSP-2106x SHARC User's Manual*, Mar. 1995.

6 GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement

Simon Doclo and Marc Moonen

Katholieke Universiteit Leuven, Leuven, Belgium

Abstract. In this chapter a class of multi-microphone signal enhancement techniques is described, which are based on a Generalized Singular Value Decomposition (GSVD) and which amount to a specific optimal filtering problem when the so-called desired response signal cannot be observed. When applying this GSVD-based optimal filtering technique to noise reduction in multi-microphone speech recordings, simulations show that this technique has a better noise reduction performance than standard beamforming techniques for all reverberation times and that it is more robust to deviations from the nominal situation, e.g. encountered in uncalibrated microphone arrays. If the GSVD-based procedure is used to create speech as well as noise references, as in a Generalized Sidelobe Canceler (GSC) structure, and an adaptive noise canceler (ANC) is added, the noise reduction performance can still be improved. Because computational complexity may appear as a bottleneck for this algorithm, recursive GSVD-updating and downsampling techniques are needed to make this signal enhancement technique amenable to real-time implementation.

6.1 Introduction

In many speech communication applications, like audio-conferencing and hands-free mobile telephony, the recorded speech signals are corrupted by acoustic background noise and far-end echo signals (see Fig. 6.1). This is mainly due to the fact that the speaker is located at a distance from the recording microphones, allowing the microphones to record the noise and echo sources too. Background noise and far-end echo signals cause a signal degradation which can lead to total unintelligibility of the speech and which decreases the performance of speech coding and speech recognition systems. Therefore efficient acoustic noise and echo reduction algorithms are called for.

Several single-microphone techniques for noise reduction in speech have been proposed that are based on a (generalized) singular value decomposition. While early techniques only dealt with white noise [1][2], more advanced techniques also incorporated colored noise [3][4]. However, since these techniques are single-microphone speech enhancement techniques, they only perform a (signal-adaptive) frequency filtering on the noisy speech signal.

When performing multi-microphone speech enhancement, both the frequency and the spatial characteristics of the speech and the noise sources can be exploited. Recently a multi-microphone noise reduction technique for

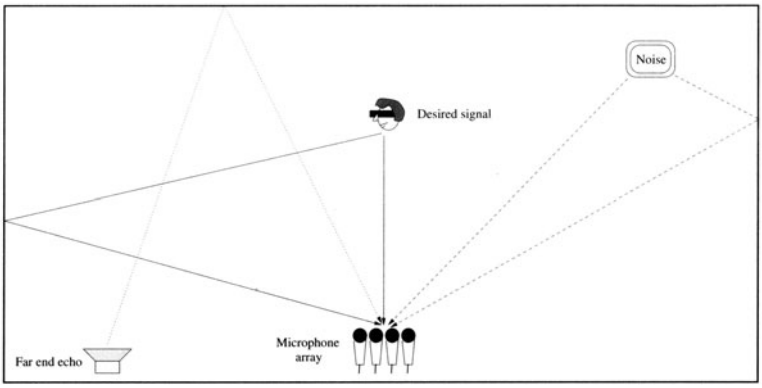


Fig. 6.1. Typical acoustic environment for hands-free speech communication

speech recognition has been proposed, based on coherent signal subspace processing [5]. In Section 6.2 a class of multi-microphone signal enhancement techniques is described, which are based on a Generalized Singular Value Decomposition (GSVD) and which amount to a specific optimal filtering problem when the so-called desired response signal cannot be observed. It is shown that the optimal filter can be written as a function of the generalized singular vectors and singular values of a so-called speech and noise data matrix. For single- and multi-microphone time-series filtering a number of symmetry properties can be derived for the optimal filter that are valid for the white noise case as well as for the colored noise case [6].

In Section 6.3 it is shown that when the optimal filtering technique is applied to multi-microphone acoustic noise reduction, it exhibits some kind of beamforming behavior for simple scenarios. It will also be shown that the GSVD-based optimal filtering technique has a better noise reduction performance than standard beamforming techniques (delay-and-sum beamformer, GSC) for all reverberation times and that it is more robust to deviations from the nominal situation, e.g. when incorrectly estimating the position of the speech source or when using uncalibrated microphone arrays [7].

In Section 6.4 the computational complexity of the GSVD-based optimal filtering technique is discussed. Since recalculating the GSVD and the optimal filter for every sample clearly requires too many computations, this section discusses several techniques for reducing the computational complexity. By using recursive and approximate GSVD-updating algorithms and by using downsampling techniques, the computational complexity can be reduced such that this technique becomes amenable to real-time implementation [8].

Although the GSVD-based optimal filtering technique as such significantly increases the signal-to-noise ratio, its noise reduction performance can still be improved by using it in a GSC-like structure, using a speech reference and a noise reference in an adaptive noise cancellation (ANC) algorithm. This ANC postprocessing stage is discussed in Section 6.5. The output of the

GSVD-based optimal filtering technique is used as a speech reference, whereas different possibilities exist for creating a noise reference [9]. The complete scheme can actually be considered as a GSC, where the simple time-delay compensation has been replaced by the GSVD-based optimal filter.

Since in many speech communication applications not only background noise but also far-end echo sources are present, a combined echo and noise reduction scheme is needed. In [10] two different schemes have been presented, one using a multi-channel adaptive echo canceler, the other incorporating the far-end echo reference directly into the GSVD-based signal enhancement technique. Because of its reduced computational complexity, the scheme using a multi-channel adaptive echo canceler is currently preferred.

6.2 GSVD-Based Optimal Filtering Technique

Consider N microphones where each microphone signal $x_n[k], n = 0 \ldots N-1$, consists of a filtered version of the speech signal $s[k]$ and additive noise,

$$x_n[k] = s_n[k] + v_n[k] = h_n^s[k] \star s[k] + v_n[k] \,, \tag{6.1}$$

with $s_n[k]$ and $v_n[k]$ respectively the speech and noise component received at the nth microphone at time k and $h_n^s[k]$ the acoustic room impulse response between the speech source and the nth microphone. The goal of speech enhancement is to compute the filters $\mathbf{w}_n[k]$ (see Fig. 6.2) such that the speech signal $s[k]$ or one of the received speech signals $s_n[k]$ is recovered. A Generalized Sidelobe Canceler (GSC) attempts to recover the speech signal $s[k]$ by constraining the array response to unity in the direction of the speech source and by minimizing the energy from all other directions. The GSVD-based optimal filtering technique attempts to recover one of the speech components $s_n[k]$ in an optimal way, using all the microphone signals $x_n[k]$.

If the filters $\mathbf{w}_n[k]$ have filter length L,

$$\mathbf{w}_n[k] = \left[\, w_n^0[k] \; w_n^1[k] \ldots w_n^{L-1}[k] \,\right]^T \,, \tag{6.2}$$

we can consider the data vectors $\mathbf{x}_n[k] \in \mathbb{R}^L$, the stacked filter $\mathbf{w}[k] \in \mathbb{R}^M$ (with $M = LN$) and the stacked data vector $\mathbf{x}[k] \in \mathbb{R}^M$, defined as

$$\mathbf{x}_n[k] = \left[\, x_n[k] \; x_n[k-1] \ldots \; x_n[k-L+1] \,\right]^T \,, \tag{6.3}$$

$$\mathbf{w}[k] = \left[\, \mathbf{w}_0^T[k] \; \mathbf{w}_1^T[k] \ldots \; \mathbf{w}_{N-1}^T[k] \,\right]^T \,, \tag{6.4}$$

$$\mathbf{x}[k] = \left[\, \mathbf{x}_0^T[k] \; \mathbf{x}_1^T[k] \ldots \; \mathbf{x}_{N-1}^T[k] \,\right]^T \,, \tag{6.5}$$

such that the output signal $y[k]$ can be written as

$$y[k] = \sum_{n=0}^{N-1} \mathbf{w}_n^T[k] \mathbf{x}_n[k] = \mathbf{w}^T[k] \mathbf{x}[k] \,. \tag{6.6}$$

In this section a method will be described for computing the stacked filter $\mathbf{w}[k]$ and some symmetry properties of this filter will be discussed.

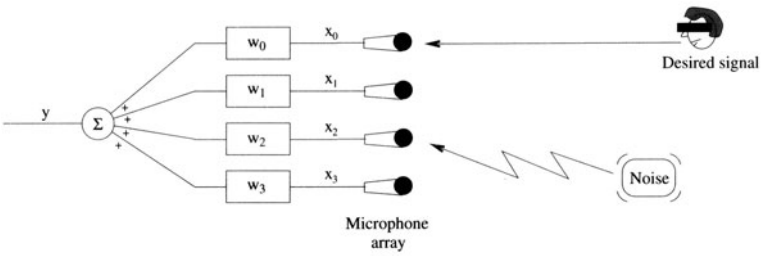


Fig. 6.2. Multi-channel filtering for speech enhancement

6.2.1 Optimal Filter Theory

Consider the general filtering problem in Fig. 6.3: $\mathbf{x} \in \mathbb{R}^M$ is the filter input vector, $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ is the filter output vector with $\mathbf{W} \in \mathbb{R}^{M\times M}$ the filter matrix. The vector $\mathbf{d} \in \mathbb{R}^M$ is the desired response and $\mathbf{e} = \mathbf{d} - \mathbf{y}$ is the error vector. The MSE (mean square error) cost function for optimal filtering is

$$\mathbf{J}_{MSE}(\mathbf{W}) = \mathcal{E}\{\|\mathbf{e}\|_2^2\} = \mathcal{E}\{\mathbf{d}^T\mathbf{d}\} - 2\mathcal{E}\{\mathbf{x}^T\mathbf{W}\mathbf{d}\} + \mathcal{E}\{\mathbf{x}^T\mathbf{W}\mathbf{W}^T\mathbf{x}\} , \tag{6.7}$$

with $\mathcal{E}$ the expected value operator. The optimal filter is found by setting the derivative $\frac{\partial \mathbf{J}_{MSE}(\mathbf{W})}{\partial \mathbf{W}}$ to zero. The optimal filter $\mathbf{W}_{WF}$ is the well-known multi-dimensional Wiener filter,

$$\mathbf{W}_{WF} = \mathbf{R}_{xx}^{-1} \cdot \mathbf{R}_{xd} , \tag{6.8}$$

with $\mathbf{R}_{xx} = \mathcal{E}\{\mathbf{x}\mathbf{x}^T\}$ the correlation matrix and $\mathbf{R}_{xd} = \mathcal{E}\{\mathbf{x}\mathbf{d}^T\}$ the cross-correlation matrix. If both matrices $\mathbf{R}_{xx}$ and $\mathbf{R}_{xd}$ are known, the problem is solved conceptually.

When considering acoustic background noise reduction in multi-microphone speech signals, the filter input vector $\mathbf{x}[k]$ at time k consists of a speech component $\mathbf{s}[k]$ and an additive noise component $\mathbf{v}[k]$,

$$\mathbf{x}[k] = \mathbf{s}[k] + \mathbf{v}[k] , \tag{6.9}$$

with $\mathbf{x}[k]$ defined in (6.5) and $\mathbf{s}[k]$ and $\mathbf{v}[k]$ similarly defined. If we use a robust voice activity detection algorithm [11][12], noise-only observations can

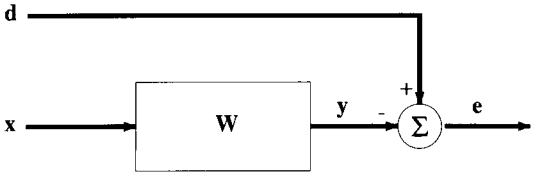


Fig. 6.3. Optimal filtering problem with unknown desired response vector $\mathbf{d}$

be made during speech pauses (time k'), $\mathbf{x}[k'] = \mathbf{v}[k']$, which allows one to estimate the spatial and temporal correlation properties of the noise signal. Our goal is to reconstruct the speech signal $\mathbf{s}[k]$ from $\mathbf{x}[k]$ during speech periods by means of the linear filter $\mathbf{W}$. In the optimal filtering context this means that the desired signal is equal to the signal-of-interest, $\mathbf{d}[k] = \mathbf{s}[k]$, but it also implies that the desired signal $\mathbf{d}[k]$ is in fact an unobservable signal. We now make two *assumptions*, namely that the noise is short-term stationary,

$$\mathbf{R}_{vv}[k] = \mathcal{E}\{\mathbf{v}[k]\mathbf{v}^T[k]\} = \mathcal{E}\{\mathbf{v}[k']\mathbf{v}^T[k']\} = \mathbf{R}_{vv}[k'] \,, \tag{6.10}$$

and furthermore statistical independent of the speech signal,

$$\mathbf{R}_{sv}[k] = \mathcal{E}\{\mathbf{s}[k]\mathbf{v}^T[k]\} = \mathbf{0} \,. \tag{6.11}$$

The first assumption allows one to estimate the noise correlation matrix $\mathbf{R}_{vv}[k]$ during speech pauses. From the second assumption it is easily verified that $\mathbf{R}_{xx}[k] = \mathbf{R}_{ss}[k] + \mathbf{R}_{vv}[k]$ and that $\mathbf{R}_{xs}[k] = \mathbf{R}_{ss}[k]$, such that the optimal filter can be written as

$$\mathbf{W}_{WF} = \mathbf{R}_{xx}^{-1}[k] \cdot (\mathbf{R}_{xx}[k] - \mathbf{R}_{vv}[k]) \,. \tag{6.12}$$

The optimal filter can then be calculated by using the joint diagonalization of the symmetric block-Toeplitz correlation matrices $\mathbf{R}_{xx}[k]$ and $\mathbf{R}_{vv}[k]$,

$$\begin{aligned} \mathbf{R}_{xx}[k] &= \mathbf{Q} \cdot \mathrm{diag}\{\sigma_i^2\} \cdot \mathbf{Q}^T \\ \mathbf{R}_{vv}[k] &= \mathbf{Q} \cdot \mathrm{diag}\{\eta_i^2\} \cdot \mathbf{Q}^T \,. \end{aligned} \tag{6.13}$$

In practice, $\mathbf{Q}$, σ_i^2 and η_i^2 are computed by means of a generalized singular value decomposition (GSVD) [13][14] of a speech data matrix $\mathbf{X}[k] \in \mathbb{R}^{p \times M}$, containing p speech data vectors, and a noise data matrix $\mathbf{V}[k] \in \mathbb{R}^{q \times M}$, containing q noise data vectors (with p and q typically larger than M),

$$\mathbf{X}[k] = \begin{bmatrix} \mathbf{x}^T[k-p+1] \\ \vdots \\ \mathbf{x}^T[k-1] \\ \mathbf{x}^T[k] \end{bmatrix} \qquad \mathbf{V}[k] = \begin{bmatrix} \mathbf{v}^T[k-q+1] \\ \vdots \\ \mathbf{v}^T[k-1] \\ \mathbf{v}^T[k] \end{bmatrix} \,, \tag{6.14}$$

such that $\mathbf{R}_{xx}[k] \simeq \mathbf{X}^T[k]\mathbf{X}[k]/p$ and $\mathbf{R}_{vv}[k] \simeq \mathbf{V}^T[k]\mathbf{V}[k]/q$. The GSVD of the two matrices $\mathbf{X}[k]$ and $\mathbf{V}[k]$ is defined as

$$\begin{aligned} \mathbf{X}[k] &= \mathbf{U}_X \cdot \Sigma_X \cdot \mathbf{Q}^T \\ \mathbf{V}[k] &= \mathbf{U}_V \cdot \Sigma_V \cdot \mathbf{Q}^T \,, \end{aligned} \tag{6.15}$$

with $\Sigma_X = \mathrm{diag}\{\sigma_i\}$, $\Sigma_V = \mathrm{diag}\{\eta_i\}$, $\mathbf{U}_X$ and $\mathbf{U}_V$ orthogonal matrices, $\mathbf{Q}$ an invertible matrix containing the generalized singular vectors and $\frac{\sigma_i}{\eta_i}$ the generalized singular values. Substituting these formulas into (6.12) gives

$$\mathbf{W}_{WF} = \mathbf{Q}^{-T} \cdot \mathrm{diag}\{1 - \frac{p}{q}\frac{\eta_i^2}{\sigma_i^2}\} \cdot \mathbf{Q}^T \,, \tag{6.16}$$

which shows that the optimal filter is a function of the generalized singular vectors and values. In the white noise case, the noise correlation matrix reduces to $\mathbf{R}_{vv}[k] = \eta^2 I$, with η^2 the white noise power. The GSVD then reduces to an SVD with $\mathbf{Q}$ an orthogonal matrix, such that $\mathbf{W}_{WF}$ is a symmetric matrix. The enhanced speech signal $\hat{s}[k]$ is computed as

$$\hat{\mathbf{s}}[k] = \begin{bmatrix} \hat{s}[k-p+1] \\ \vdots \\ \hat{s}[k-1] \\ \hat{s}[k] \end{bmatrix} = \mathbf{X}[k] \cdot \mathbf{w}_{WF}^{i} , \tag{6.17}$$

with $\mathbf{w}_{WF}^{i}$ the ith column of $\mathbf{W}_{WF}$. The signal $\hat{s}[k]$ is an optimal estimate for the speech component $s_l[k-\Delta]$ in the lth microphone signal, with

$$l = \mathrm{mod}(i-1, L) \tag{6.18}$$

$$\Delta = \mathrm{rem}(i-1, L) . \tag{6.19}$$

The estimation error $\mathbf{e}[k]$ is defined as $\mathbf{e}[k] = \mathbf{s}[k] - \mathbf{y}[k] = \mathbf{s}[k] - \mathbf{W}_{WF}^{T}\mathbf{x}[k]$, such that error covariance matrix $\mathbf{R}_{ee}[k]$ can be written as

$$\mathbf{R}_{ee}[k] = \mathcal{E}\{\mathbf{e}[k]\mathbf{e}^{T}[k]\} = \mathbf{R}_{vv}[k] \cdot \mathbf{W}_{WF} . \tag{6.20}$$

The elements on the main diagonal of the error covariance matrix $\{\mathbf{R}_{ee}[k]\}_{ii}$ indicate how well the ith component of $\mathbf{s}[k]$, i.e. $s_l[k-\Delta]$, is estimated. The smallest element on the diagonal of this matrix therefore corresponds to the best estimator, which is the corresponding column of $\mathbf{W}_{WF}$. However simulations indicate that e.g. taking $i = \frac{L}{2}$ instead of the optimal value does not decrease the performance considerably.

6.2.2 General Class of Estimators

The filter $\mathbf{W}_{WF}$ in fact belongs to a more general class of estimators, which can be described by

$$\mathbf{W} = \mathbf{Q}^{-T} \cdot \mathrm{diag}\{f(\sigma_i^2, \eta_i^2)\} \cdot \mathbf{Q}^{T} , \tag{6.21}$$

with $f(\sigma_i^2, \eta_i^2)$ an arbitrary function. This formula can be interpreted as an analysis filterbank $\mathbf{Q}^{-T}$ which performs a transformation from the time domain to a signal-dependent transform domain, a gain function $f(\sigma_i^2, \eta_i^2)$ which modifies the transform domain parameters and a synthesis filterbank $\mathbf{Q}^{T}$ which performs a transformation back to the time domain.

If the MSE-criterion is optimized, the filter $\mathbf{W}$ is equal to (6.16). If the signal-to-noise ratio (SNR) is optimized, only the principal generalized singular vector should be considered, such that the gain function $f(\sigma_i^2, \eta_i^2) =$

$[1\ 0\ \ldots\ 0]$. This will however introduce a significant amount of signal distortion. In the remainder of this chapter we will use the gain function

$$f(\sigma_i^2, \eta_i^2) = 1 - \alpha \cdot \frac{p}{q} \frac{\eta_i^2}{\sigma_i^2} , \tag{6.22}$$

with α a noise overestimation factor ($\alpha = 1$ corresponds to the MMSE estimator). By increasing the factor α, the SNR of the enhanced signal increases, but some signal distortion is introduced (see Section 6.5.2).

6.2.3 Symmetry Properties for Time-Series Filtering

When applying the GSVD-based optimal filtering to *single-microphone* noise reduction, the vector $\mathbf{x}[k]$ is taken from the time series $x[k]$, such that

$$\mathbf{x}[k] = \left[x[k]\ x[k-1]\ \ldots\ x[k-L+1]\right]^T . \tag{6.23}$$

The data matrices $\mathbf{X}[k]$ and $\mathbf{V}[k]$, defined in (6.14), are now Toeplitz matrices, such that the correlation matrices $\mathbf{R}_{xx}[k]$ and $\mathbf{R}_{vv}[k]$ are symmetric Toeplitz matrices. Symmetric Toeplitz matrices belong to the class of double symmetric matrices, which are symmetric about both the main diagonal and the secondary diagonal and whose eigenvectors have special symmetry properties [15], i.e. every eigenvector is either symmetric or skew-symmetric. Using these properties, one can prove the following symmetry property for the filter $\mathbf{W}$.

Theorem 1. *If* $\mathbf{W}$ *is constructed according to (6.21), then* $\mathbf{W}$ *satisfies the symmetry properties*

$$\mathbf{W} = \mathbf{J} \cdot \mathbf{W} \cdot \mathbf{J} \qquad \mathbf{W}^T = \mathbf{J} \cdot \mathbf{W}^T \cdot \mathbf{J} , \tag{6.24}$$

with $\mathbf{J}$ *the reverse identity matrix. These properties hold in the white noise case as well as in the colored noise case.*

Proof : Considering the joint diagonalization of $\mathbf{R}_{xx}[k]$ and $\mathbf{R}_{vv}[k]$ in (6.13), one can easily verify that

$$\mathbf{R}_{xx}^{-1}[k] \cdot \mathbf{R}_{vv}[k] = \mathbf{Q}^{-T} \cdot \mathrm{diag}\{\frac{\eta_i^2}{\sigma_i^2}\} \cdot \mathbf{Q}^T \tag{6.25}$$

is the eigenvector decomposition. Because

$$\mathbf{R}_{xx}^{-1}[k] \cdot \mathbf{R}_{vv}[k] = \mathbf{J} \cdot \mathbf{R}_{xx}^{-1}[k] \cdot \mathbf{R}_{vv}[k] \cdot \mathbf{J} , \tag{6.26}$$

the eigenvectors, which are the columns of $\mathbf{Q}^{-T}$, satisfy the property

$$\mathbf{J} \cdot \mathbf{Q}^{-T} = \mathbf{Q}^{-T} \cdot \mathrm{diag}\{\pm 1\} , \tag{6.27}$$

such that

$$\begin{aligned}\mathbf{J}\cdot\mathbf{W}\cdot\mathbf{J} &= \mathbf{J}\cdot\mathbf{Q}^{-T}\cdot\mathrm{diag}\{f(\sigma_i^2,\eta_i^2)\}\cdot\mathbf{Q}^{T}\cdot\mathbf{J} \\ &= \mathbf{Q}^{-T}\cdot\mathrm{diag}\{f(\sigma_i^2,\eta_i^2)\}\cdot\mathbf{Q}^{T} = \mathbf{W}\ .\end{aligned} \tag{6.28}$$

These symmetry properties imply that the ith row/column of $\mathbf{W}$ is equal to the $(L+1-i)$th row/column with its elements in reverse order. For L odd, the middle column in $\mathbf{W}$ is symmetric, and hence represents a linear phase filter. This linear phase property is an extension of the zero phase property that has already been attributed to SVD and rank truncation based estimators for the white noise case, if an additional averaging step is included [16]. The above linear phase property is however valid for the colored noise case as well as for an arbitrary gain function $f(\sigma_i^2,\eta_i^2)$.

For *multi-channel filtering* similar symmetry properties as for the single-channel case can be derived. When making additional assumptions for the noise correlation matrix, additional symmetry properties can be derived [17].

6.3 Performance of GSVD-Based Optimal Filtering

This section discusses the performance of the GSVD-based optimal filtering technique for noise reduction in multi-microphone speech signals. The used simulation environment is described and using simulations the spatial directivity pattern, noise reduction performance and robustness of the GSVD-based filtering technique are discussed.

6.3.1 Simulation Environment

The simulation room configuration is depicted in Fig. 6.1, with dimensions 6 m × 3 m × 2.5 m. It consists of a microphone array, a speech source $s[k]$ and a background noise source $v[k]$. In our simulations we use a linear equi-spaced microphone array with $N = 4$ microphones and the distance d between two adjacent microphones is 5 cm. The signals used are an 8 kHz clean speech signal and a temporally white noise source.

The room impulse responses are obtained using the image method [18], with a filter length of 1500 taps and for different reverberation times T_{60}. The reverberation time T_{60} can be expressed as a function of the reflection coefficient γ of the walls, according to Eyring's formula,

$$T_{60} = \frac{0.163V}{-S\log(1-\gamma)}\ , \tag{6.29}$$

with V the volume of the room and S the total surface of the room. Simulations have been performed at different signal-to-noise ratios. Since we are

using simulated data (and hence know the speech and noise component of each signal), the unbiased signal-to-noise ratio (SNR) can be computed as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum \tilde{s}^2[k]}{\sum \tilde{v}^2[k]}\,, \tag{6.30}$$

with $\tilde{s}[k]$ and $\tilde{v}[k]$ the speech and noise component of the considered signal.

6.3.2 Spatial Directivity Pattern

When considering localized sources and no multipath propagation, it can be shown by simulations that the GSVD-based filtering technique exhibits some kind of beamforming behavior. The spatial directivity pattern $H(f,\theta)$ of the GSVD-based optimal filter $\mathbf{w}[k] = \mathbf{w}_{WF}^{i}$ is defined as

$$H(f,\theta) = \sum_{n=0}^{N-1} W_n(f)\cdot \exp\left(j2\pi f\frac{nd\cos\theta}{c}\right)\,. \tag{6.31}$$

$H(f,\theta)$ is a function of frequency f and angle θ. $W_n(f)$ is the frequency response of the filter $\mathbf{w}_n[k]$, d is the distance between the microphones and c is the speed of sound wave propagation ($c \simeq 340$m/sec).

In the first case we consider spatio-temporal white noise, i.e. the noise component $v_n[k]$ present in every microphone signal $x_n[k]$ is temporally white and is uncorrelated with the noise component in every other microphone signal. We consider the situation where the speech source impinges on the

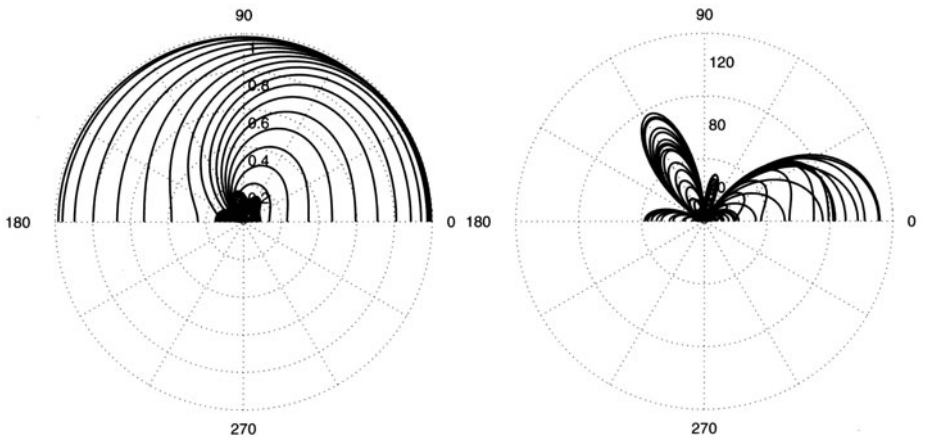


Fig. 6.4. Spatial directivity pattern $|H(f,\theta)|$ for (**a**) spatio-temporal white noise and speech source at $\theta = 45°$ ($N = 4$, $L = 10$, SNR= 0 dB) and (**b**) localized white noise sources at $\theta = 60°$ and $\theta = 150°$ and speech source at $\theta = 90°$ ($N = 4$, $L = 20$, SNR= 0 dB)

microphone array at an angle $\theta = 45°$. Fig. 6.4a shows the spatial directivity pattern for the frequencies $f = i \cdot 100, i = 1 \ldots 40$. For most frequencies the directivity gain is maximal for the direction $\theta = 45°$, which implies that the GSVD-based filtering technique automatically finds the direction of the speech source. However for low frequencies the spatial selectivity is poor.

In the second case we consider two localized white noise sources that impinge on the microphone array at angles $\theta = 60°$ and $\theta = 150°$, The speech source is located in front of the microphone array ($\theta = 90°$). Figure 6.4b shows the directivity pattern for the frequencies $f = i \cdot 100, i = 1 \ldots 40$. As can be seen, for all frequencies the directivity gain is approximately zero for $\theta = 60°$ and $\theta = 150°$, the directions of the two noise sources. Although difficult to see on this figure, the directivity gain in the direction of the speech source $\theta = 90°$ is not equal to unity, as is the case for a GSC, but depends on the frequency content of the speech and noise signals.

We can conclude that the GSVD-based optimal filtering technique has the desired beamforming behavior for both simple situations. For more realistic reverberant situations it is more difficult to interpret the spatial directivity plots, since the GSVD-based filtering technique computes an optimal estimate for the speech component of one microphone signal, thereby reducing the additive noise but not the reverberation of the speech signal.

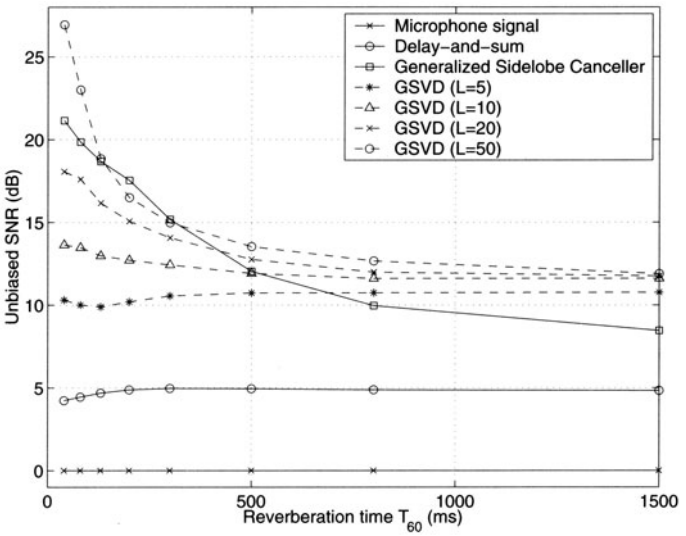


Fig. 6.5. Comparison of unbiased SNR for delay-and-sum, GSC and GSVD-based optimal filter ($N = 4$, SNR$= 0$ dB)

6.3.3 Noise Reduction Performance

This section compares the noise reduction performance of the GSVD-based optimal filtering technique for different filter lengths L and for different reverberation times T_{60}. Low reverberation corresponds to highly correlated signals, while high reverberation corresponds to highly uncorrelated (diffuse) signals. The noise reduction performance is also compared with standard adaptive beamforming techniques, i.e. the delay-and-sum beamformer and Generalized Sidelobe Canceler (GSC).

Fig. 6.5 shows the unbiased SNR of the enhanced signal for reverberation up to 1500 ms for the delay-and-sum beamformer, the GSC and the GSVD-based optimal filtering technique ($L = 5, 10, 20, 50$). The unbiased SNR of the original microphone signal is 0 dB. As can be seen, for small T_{60} the GSC performs much better than for high T_{60}. This is normal because the GSC is designed for correlated noise, not for diffuse noise. Unlike the GSC, the GSVD-based optimal filtering technique still performs well for high T_{60}. As can be seen, for all reverberation times, the GSVD-based filtering technique performs better than the GSC, if the filter length L is large enough.

6.3.4 Robustness Issues

Many multi-microphone noise reduction techniques, e.g. GSC, exploit *a priori* knowledge about the position of the speech source and the microphone array configuration. These techniques therefore tend to be rather sensitive

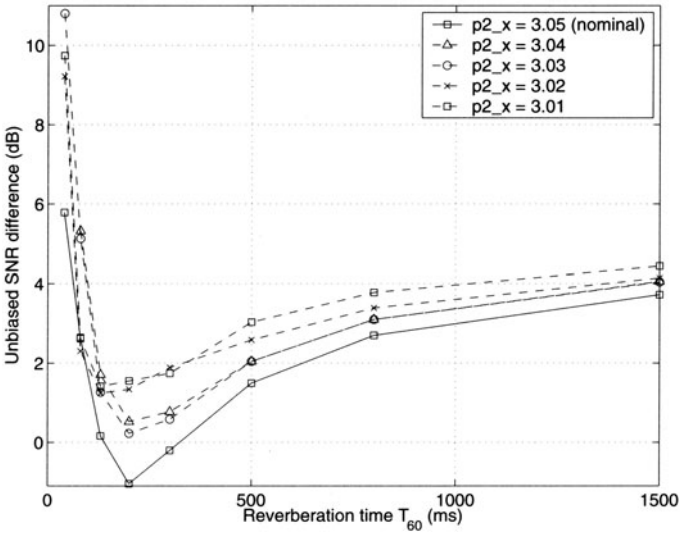


Fig. 6.6. Unbiased SNR-difference between GSVD-based optimal filtering and GSC for different microphone positions $\mathbf{p}_2$ ($N = 4$, $L = 50$, SNR= 0 dB)

to deviations from the nominal situation, e.g. encountered when incorrectly estimating the position of the speech source or when using uncalibrated microphone arrays. The GSVD-based optimal filtering technique does not make any assumptions of this kind. Therefore we can expect the GSVD-based filtering technique to be less sensitive to deviations from the nominal situation.

In [7] we have compared the robustness of the GSVD-based filtering technique with the GSC for 3 kinds of deviations from the nominal situation: (a) incorrect estimation of the position of the speech source, (b) microphone displacement and (c) different microphone gains.

Figure 6.6 shows the difference in noise reduction performance (unbiased SNR) between the GSVD-based optimal filtering technique and the GSC for different microphone positions $\mathbf{p}_2$. Because the difference in performance increases the more the microphone position $\mathbf{p}_2$ deviates from the nominal position $\mathbf{p}_2^{nom}$, we can conclude that for microphone displacement, the GSVD-based optimal filtering technique is more robust than the GSC. In [7] it has been shown that also for incorrect estimation of the position of the speech source and different microphone gains, the GSVD-based optimal filtering technique is more robust than the GSC. In fact it can be theoretically proven that the noise reduction performance of the GSVD-based optimal filtering technique is independent of different gains for the different microphones.

6.4 Complexity Reduction

This section discusses the computational complexity of the GSVD-based optimal filtering technique. Recalculating the GSVD and the optimal filter from scratch for every sample clearly requires too many computations In this section several techniques for reducing the complexity are discussed, enabling real-time implementation of the GSVD-based optimal filtering technique. For realistic parameters the computational complexity can be reduced from 34.2 Gflops to 55 Mflops (see Section 6.4.5).

6.4.1 Linear Algebra Techniques for Computing GSVD

The GSVD of the matrices $\mathbf{X}[k]$ and $\mathbf{V}[k]$ can be computed as follows (for details see [13][19][20]). First, the two matrices are reduced to upper triangular form by a QR-decomposition,

$$\mathbf{X}[k] = \underbrace{\mathbf{Q}_{X[k]}}_{p\times M} \cdot \underbrace{\mathbf{R}_{X[k]}}_{M\times M} \qquad \mathbf{V}[k] = \underbrace{\mathbf{Q}_{V[k]}}_{q\times M} \cdot \underbrace{\mathbf{R}_{V[k]}}_{M\times M}, \tag{6.32}$$

with $\mathbf{R}_{X[k]}$ and $\mathbf{R}_{V[k]}$ square upper triangular matrices and $\mathbf{Q}_{X[k]}$ and $\mathbf{Q}_{V[k]}$ orthogonal matrices. The GSVD of $\mathbf{X}[k]$ and $\mathbf{V}[k]$ readily follows from the GSVD of $\mathbf{R}_{X[k]}$ and $\mathbf{R}_{V[k]}$. This decomposition is computed by carrying

out an iterative procedure where a series of Jacobi-rotations is applied to $\mathbf{R}_{X[k]}$ and $\mathbf{R}_{V[k]}$ in order to yield upper triangular factors with parallel rows $\bar{\mathbf{R}}_{X[k]} = \Sigma_{X[k]} \cdot \mathbf{R}[k]$ and $\bar{\mathbf{R}}_{V[k]} = \Sigma_{V[k]} \cdot \mathbf{R}[k]$, with $\Sigma_{X[k]}$ and $\Sigma_{V[k]}$ diagonal matrices. The GSVD of $\mathbf{R}_{X[k]}$ and $\mathbf{R}_{V[k]}$ can be written as

$$\mathbf{R}_{X[k]} = \Theta_{X[k]} \cdot \Sigma_{X[k]} \cdot \mathbf{Q}^T[k] = \Theta_{X[k]} \cdot \underbrace{\Sigma_{X[k]} \cdot \mathbf{R}[k]}_{\bar{\mathbf{R}}_{X[k]}} \cdot \bar{\mathbf{Q}}^T[k] \tag{6.33}$$

$$\mathbf{R}_{V[k]} = \Phi_{V[k]} \cdot \Sigma_{V[k]} \cdot \mathbf{Q}^T[k] = \Phi_{V[k]} \cdot \underbrace{\Sigma_{V[k]} \cdot \mathbf{R}[k]}_{\bar{\mathbf{R}}_{V[k]}} \cdot \bar{\mathbf{Q}}^T[k] \,, \tag{6.34}$$

with $\Theta_{X[k]}$ and $\Phi_{V[k]}$ orthogonal matrices containing the rotation angles, defined in (6.36), and $\bar{\mathbf{Q}}^T[k]$ also an orthogonal matrix.

Each iteration essentially reduces to a GSVD of an elementary 2×2 block on the main diagonal, parallelizing the rows of $\{\mathbf{R}_{X[k]}\}_{i,i+1}$ and $\{\mathbf{R}_{V[k]}\}_{i,i+1}$. When the pivot index i repeatedly takes up all possible values, $i = 1 \ldots M-1$, this is called one sweep ($= M-1$ GSVD-steps). The GSVD of the 2×2 upper triangular blocks $\{\mathbf{R}_{X[k]}\}_{i,i+1}$ and $\{\mathbf{R}_{V[k]}\}_{i,i+1}$ corresponds to the SVD of the 2×2 upper triangular block

$$\{\mathbf{R}_{C[k]}\}_{i,i+1} = \{\mathbf{R}_{X[k]}\}_{i,i+1} \cdot \{\mathbf{R}_{V[k]}\}_{i,i+1}^{-1} \,, \tag{6.35}$$

followed by an orthogonal transformation to upper-triangularize $\{\mathbf{R}_{X[k]}\}_{i,i+1}$ and $\{\mathbf{R}_{V[k]}\}_{i,i+1}$. The SVD of the elementary 2×2 block $\{\mathbf{R}_{C[k]}\}_{i,i+1}$ comes down to calculating the Givens rotations θ and ϕ [13][21] such that

$$\begin{bmatrix} r_C^{i,i\,*} & 0 \\ 0 & r_C^{i+1,i+1\,*} \end{bmatrix} = \begin{bmatrix} -\sin\theta & \cos\theta \\ \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} r_C^{i,i} & r_C^{i,i+1} \\ 0 & r_C^{i+1,i+1} \end{bmatrix} \begin{bmatrix} -\sin\phi & \cos\phi \\ \cos\phi & \sin\phi \end{bmatrix} . \tag{6.36}$$

Computing a full GSVD requires βM sweeps (with β typically $3 \ldots 5$). Therefore the total complexity, which is defined as the total number of additions and multiplications, amounts to $3\beta M^2(p+q-2M/3)$ (QR-decompositions) $+ 18\beta M^3$ (GSVD computation). For typical values of p, q and M the complexity is too high for real-time implementation (see Section 6.4.5). Therefore more efficient recursive GSVD-updating algorithms will be considered.

6.4.2 Recursive and Approximate GSVD-Updating Algorithms

Instead of recomputing the GSVD from scratch for each time step, recursive GSVD-updating algorithms compute the GSVD at time k using the decomposition at time $k-1$. In [22][23] a Jacobi-type (G)SVD-updating algorithm is described. Suppose that at time $k-1$, the upper triangular factors are reduced to $\bar{\mathbf{R}}_{X[k-1]}$ and $\bar{\mathbf{R}}_{V[k-1]}$ with approximately parallel rows, such that

$$\begin{aligned} \mathbf{X}[k-1] &= \mathbf{U}_{X[k-1]} \cdot \Sigma_{X[k]} \cdot \mathbf{Q}^T[k] = \mathbf{U}_{X[k-1]} \cdot \bar{\mathbf{R}}_{X[k-1]} \cdot \bar{\mathbf{Q}}^T[k-1] \\ \mathbf{V}[k-1] &= \mathbf{U}_{V[k-1]} \cdot \Sigma_{V[k]} \cdot \mathbf{Q}^T[k] = \mathbf{U}_{V[k-1]} \cdot \bar{\mathbf{R}}_{V[k-1]} \cdot \bar{\mathbf{Q}}^T[k-1] \,, \end{aligned} \tag{6.37}$$

of which only the upper-triangular matrices $\bar{\mathbf{R}}_{X[k-1]}$, $\bar{\mathbf{R}}_{V[k-1]}$ and the orthogonal matrix $\bar{\mathbf{Q}}[k-1]$ are stored. At time k a new data vector $\mathbf{x}[k]$ (speech periods) or $\mathbf{v}[k]$ (noise periods) is present, such that we need to recompute the GSVD of $\mathbf{X}[k]$ and $\mathbf{V}[k]$, defined as

$$\mathbf{X}[k] = \begin{bmatrix} \lambda_x \cdot \mathbf{X}[k-1] \\ \mathbf{x}[k] \end{bmatrix} \qquad \mathbf{V}[k] = \begin{bmatrix} \lambda_v \cdot \mathbf{V}[k-1] \\ \mathbf{v}[k] \end{bmatrix}, \tag{6.38}$$

with λ_x an exponential weighting factor for speech and λ_v an exponential weighting factor for noise (if $\lambda = 1$ no weighting is performed). In fact, either $\mathbf{x}[k]$ or $\mathbf{v}[k]$ is equal to $\mathbf{0}$, depending on whether the signal component contains speech or only noise, which can lead to a further complexity reduction. For the general case we can rewrite $\mathbf{X}[k]$ as

$$\mathbf{X}[k] = \begin{bmatrix} \mathbf{U}_{X[k-1]} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 \ldots 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \lambda_x \cdot \bar{\mathbf{R}}_{X[k-1]} \\ \mathbf{x}^T[k] \cdot \bar{\mathbf{Q}}[k-1] \end{bmatrix} \cdot \bar{\mathbf{Q}}^T[k-1]\,. \tag{6.39}$$

The matrix $\mathbf{V}[k]$ is similarly defined. First, the upper triangular factors are restored by performing QR-updates with the transformed input vectors $\bar{\mathbf{x}}^T[k] = \mathbf{x}^T[k] \cdot \bar{\mathbf{Q}}[k-1]$ or $\bar{\mathbf{v}}^T[k] = \mathbf{v}^T[k] \cdot \bar{\mathbf{Q}}[k-1]$. QR-updating can be performed by using orthogonal Givens transformations, zeroing the elements on the bottom row. Since either $\mathbf{x}[k]$ or $\mathbf{v}[k]$ is $\mathbf{0}$, only one QR-update is required. Assuming that speech is present ($\mathbf{v}[k] = \mathbf{0}$), the QR-update for $\mathbf{X}[k]$ can be written as

$$\mathbf{X}[k] = \underbrace{\begin{bmatrix} \mathbf{U}_{X[k-1]} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 \ldots 0 & 1 \end{bmatrix} \cdot \tilde{\mathbf{Q}}_{X[k]}}_{\tilde{\mathbf{U}}_{X[k]}} \cdot \tilde{\mathbf{R}}_{X[k]} \cdot \bar{\mathbf{Q}}^T[k-1]\,. \tag{6.40}$$

Here $\tilde{\mathbf{Q}}_{X[k]}$ is an $(M+1) \times M$ matrix with orthogonal columns. The matrix $\bar{\mathbf{Q}}[k-1]$ is not altered by the QR-update.

Secondly, the GSVD-procedure is resumed in order to parallelize the rows of $\tilde{\mathbf{R}}_{X[k]}$ and $\tilde{\mathbf{R}}_{V[k]}$ (of course $\tilde{\mathbf{R}}_{V[k]} = \bar{\mathbf{R}}_{V[k]}$ in the described case with $\mathbf{v}[k] = \mathbf{0}$). A fixed number of sweeps (s) is performed, where the pivot index i takes up r consecutive values. Typically one sweep is performed ($s = 1$), where the pivot index takes up all possible values ($r = M - 1$) [23]. The complexity of one GSVD-update is equal to $2.5M^2$ (matrix-vector multiplication) + $3M^2$ (QR-update) + $s \cdot r \cdot 18M$ (GSVD-steps). For $s = 1$ and $r = M - 1$ the total computational complexity amounts to $23.5M^2$.

The computational complexity can be further reduced by using square-root-free implementations for the QR-updates and for the calculation of the

elementary 2×2 SVDs. The calculation of the rotation angles for a QR-update and for an elementary 2×2 SVD respectively requires one and three square roots. In [24] a square-root-free procedure for QR-updating is developed where a one-sided factorization of the upper triangular R-factor is used. However, since the above SVD schemes as such do not lend themselves to square-root-free implementation, alternative schemes based on approximate formulas for the calculation of the rotation angles θ and ϕ have to be considered [21]. When combined with a generalized Gentleman procedure with a two-sided factorization of the upper triangular R-factor, these schemes eventually yield square-root-free SVD-updating algorithms [25], which can be easily extended to square-root-free GSVD-updating algorithms [23].

The computational complexity of one square-root-free GSVD-update is equal to $2.5M^2$ (matrix-vector multiplication) + $2M^2$ (square-root-free QR-update) + $s \cdot r \cdot (14M - r)$ (square-root-free GSVD-steps). For $s = 1$ and $r = M - 1$ the total computational complexity amounts to $17.5M^2$, which is less expensive than 'conventional' (non square-root-free) GSVD-updating.

The optimal filter $\mathbf{W}_{WF}[k]$ at time k is computed as

$$\mathbf{W}_{WF}[k] = \bar{\mathbf{Q}}[k] \cdot \bar{\mathbf{R}}_{X[k]}^{-1} \cdot \mathrm{diag}\{1 - \alpha \cdot \frac{p}{q} \frac{(\bar{\mathbf{R}}_{V[k]}^{ii})^2}{(\bar{\mathbf{R}}_{X[k]}^{ii})^2}\} \cdot \bar{\mathbf{R}}_{X[k]} \cdot \bar{\mathbf{Q}}^T[k] \,. \tag{6.41}$$

The computational complexity for computing one column of $\mathbf{W}_{WF}[k]$ is $4M^2$. If exponential weighting factors λ_x and λ_v are used, the factor p/q in (6.41) has to be replaced by $(1 - \lambda_v^2)/(1 - \lambda_x^2)$.

6.4.3 Downsampling Techniques

For stationary acoustic environments the computational complexity can be further reduced by using downsampling techniques without any loss in performance. In this context downsampling means that the GSVD of $\mathbf{X}[k]$ and $\mathbf{V}[k]$ and the filter $\mathbf{W}_{WF}[k]$ are not updated for every sample, but that the GSVD is updated every d_g samples and that the filter is updated every d_f samples. The drawback of using downsampling is slower convergence towards the optimal filter, implying that downsampling has to be limited in non-stationary acoustic environments.

6.4.4 Simulations

Using simulations we have compared the performance of the conventional and the square-root-free recursive GSVD-update algorithms. For the square-root-free recursive GSVD-update algorithm three different approximations have been used. Figure 6.7 shows the unbiased SNR of the enhanced signal for different values of s (number of sweeps) and r (GSVD-steps). This figure

Table 6.1. Computational complexity for GSVD-based optimal filtering technique

	Non-recursive	Recursive	Square-root-free
	$\frac{\beta}{d_g}(17M^3 + 3qM^2)$	$\frac{23.5M^2}{d_g} + \frac{4M^2}{d_f}$	$\frac{17.5M^2}{d_g} + \frac{4M^2}{d_f}$
$d = 1$	684 Gflops	1408 Mflops	1101 Mflops
$d = 20$	34.2 Gflops	70.4 Mflops	55.0 Mflops

shows that there is almost no difference in performance between the conventional and the square-root-free GSVD-updating algorithm. When performing more than one sweep, the SNR only marginally improves. When performing less than $M - 1$ GSVD-steps, the SNR gradually decreases. Of course, the drawback of using of smaller number of GSVD-steps, is slower convergence towards the optimal filter, implying that the number of GSVD-steps has to be large enough in non-stationary acoustic environments.

6.4.5 Computational Complexity

The total computational complexity of the different algorithms has already been computed in Sects. 6.4.1 and 6.4.2. Table 6.1 shows the complexity in floating point operations per second for the different algorithms with $f_s = 8\,\text{kHz}$, $N = 4$, $L = 20$ in case of no downsampling and in case of downsampling with $d_f = d_g = 20$ ($q = 4000$, $\beta = 1$, $s = 1$, $r = M - 1$). Although the computational complexity of the recursive GSVD-update algorithms is still quite high, it is possible to implement the GSVD-based optimal filtering technique in real-time on a standard Pentium-III 450 MHz PC.

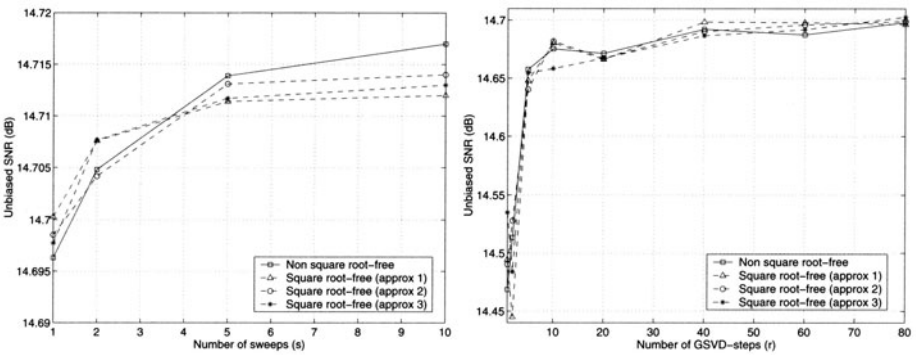


Fig. 6.7. Noise reduction performance of conventional and square-root-free GSVD-updating algorithms in function of (**a**) number of sweeps and (**b**) number of GSVD-steps ($N = 4$, $L = 20$, $T_{60} = 200\,\text{ms}$, SNR= 0 dB, $\lambda_x = \lambda_v = 0.9999$)

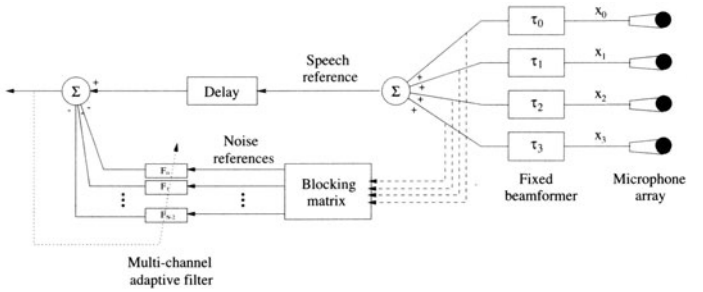


Fig. 6.8. Generalized Sidelobe Canceler

6.5 Combination with ANC Postprocessing Stage

6.5.1 Creation of Speech and Noise References

A well known adaptive beamformer, depicted in Fig. 6.8, is the Generalized Sidelobe Canceler (GSC) [26][27][28]. A GSC uses the output of a fixed delay-and-sum beamformer as a speech reference and creates noise references by combining the delayed microphone signals using a so-called blocking matrix. In the adaptive noise cancellation (ANC) stage a multi-channel adaptive filter (e.g. NLMS, APA, RLS) then removes the correlation between the noise references and the speech reference. For this algorithm to work properly, the noise references therefore should be (highly) correlated with the noise still present in the speech reference. Since in most cases signal leakage into the noise references cannot be completely avoided, the adaptive filter is only allowed to adapt during periods when no speech is present [29][30].

Although the GSVD-based optimal filtering technique in itself significantly increases the signal-to-noise ratio, its noise reduction performance can still be improved by using it in a GSC-like structure, as an alternative for the simple delay-and-sum beamformer (see Fig. 6.9). In the ANC postprocessing stage of the GSVD-based optimal filtering technique, the output of the optimal filter is used as a speech reference. For the creation of a noise reference different possibilities exist.

Define the speech reference $r_{speech}^{l,\Delta}[k]$ as the optimal estimate of the delayed speech component in the lth microphone signal, such that

$$r_{speech}^{l,\Delta}[k] = \hat{s}_l[k - \Delta] \; . \tag{6.42}$$

An obvious choice for creating a noise reference consists of simply subtracting the speech reference $r_{speech}^{l,\Delta}[k]$ from the delayed lth microphone signal,

$$r_{noise}^{1}[k] = x_l[k - \Delta] - r_{speech}^{l,\Delta}[k] \; . \tag{6.43}$$

Indeed, it can be shown that if $\mathbf{W}_{WF}$ is the optimal filter for estimating the speech component $\mathbf{s}[k]$, then $\mathbf{I} - \mathbf{W}_{WF}[k]$ is the optimal filter for estimating

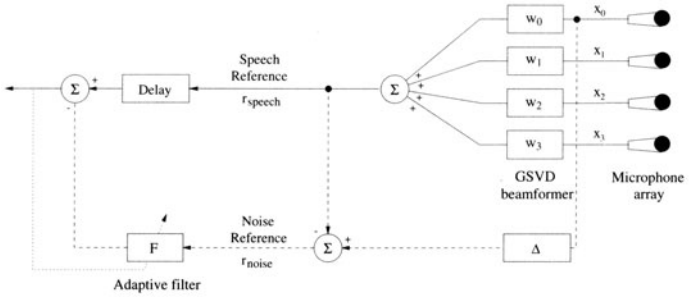


Fig. 6.9. GSVD-based optimal filtering technique with ANC postprocessing stage

the noise component $\mathbf{v}[k]$. In the remainder of this chapter we will use this noise reference and we will take $i = \frac{L}{2}$, such that $l = 0$ and $\Delta = \frac{L}{2} - 1$ (see (6.18) and (6.19)).

However different possibilities for creating noise references exist. Instead of only calculating a noise reference for the lth microphone signal, one can calculate and use the noise references for all the microphone signals, i.e.

$$r^2_{noise}[k] = \begin{bmatrix} x_0[k-\Delta] - r^{0,\Delta}_{speech}[k] \\ x_1[k-\Delta] - r^{1,\Delta}_{speech}[k] \\ \vdots \\ x_{N-1}[k-\Delta] - r^{N-1,\Delta}_{speech}[k] \end{bmatrix}. \tag{6.44}$$

However since different speech references are needed for the calculation of $r^2_{noise}[k]$, the microphone signals have to be filtered with different filters $\mathbf{w}^i_{WF}$, implying increased computational complexity.

6.5.2 Noise Reduction Performance of ANC Postprocessing Stage

In Fig. 6.10 the unbiased SNR of the output signal of the ANC postprocessing stage is plotted as a function of the filter length L and the filter length L_{ANC} of the adaptive filter in the ANC postprocessing stage for two values of the noise overestimation factor α (see (6.22)). This figure shows that the unbiased SNR improves with increasing filter lengths L and L_{ANC}. If an NLMS or fast-RLS adaptive algorithm is used in the ANC postprocessing stage, the total computational complexity is $\mathcal{O}(L^2) + \mathcal{O}(L_{ANC})$, such that it is sometimes better to use a relatively small filter length L for the GSVD-based optimal filter and a relatively large filter length L_{ANC} for the adaptive filter.

This figure also shows that the SNR of the enhanced signal improves with increasing noise overestimation factor α. For higher values of the factor α the SNR of the enhanced signal improves, but also signal distortion is introduced since the MMSE-criterion is not optimized any more. Therefore the factor α has to be limited.

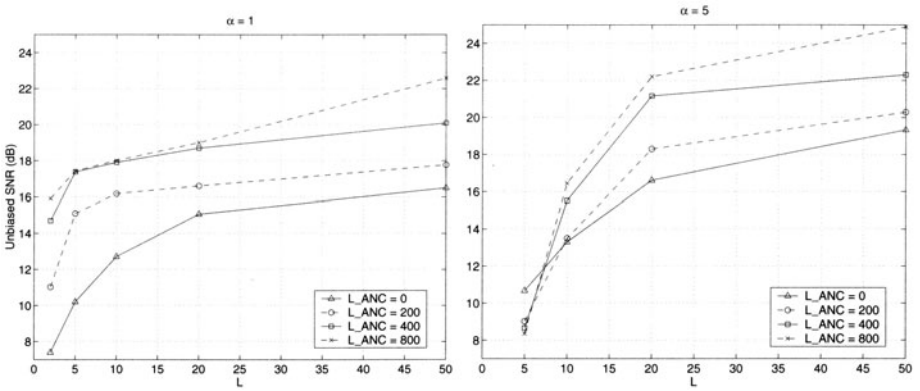


Fig. 6.10. Dependence of unbiased SNR on filter length L, filter length L_{ANC} in ANC postprocessing stage and factor α ($N = 4$, $T_{60} = 200\,\text{ms}$, SNR= 0 dB)

6.5.3 Comparison with Standard Beamforming Techniques

Figure 6.11 compares the performance of the GSVD-based optimal filtering technique (with and without ANC postprocessing stage) with standard beamforming algorithms (delay-and-sum and GSC) for different reverberation times T_{60}. The filter length L_{ANC} of the adaptive filter in the ANC postprocessing stage is 800, for the GSC as well as for the GSVD-based optimal filter.

Figure 6.11 shows that for small T_{60} the GSC performs better than the GSVD-based optimal filtering technique ($L = 20$), while for large T_{60} the GSVD-based optimal filtering technique performs better than the GSC. However, when adding the ANC postprocessing stage, the GSVD-based filtering technique clearly outperforms the GSC for all reverberation times.

6.6 Conclusion

In this chapter a class of multi-microphone signal enhancement techniques has been described, which are based on a generalized singular value decomposition. When applied to multi-microphone acoustic noise reduction, simulations show that the GSVD-based optimal filtering technique has a better noise reduction performance than standard beamforming techniques for all reverberation times and that it is more robust to deviations from the nominal situation. When speech and noise references are created and an adaptive noise cancellation (ANC) postprocessing stage is added, the SNR is further increased. Because of the high computational complexity, recursive and approximate GSVD-updating techniques and downsampling techniques are needed to make this signal enhancement technique amenable to real-time implementation.

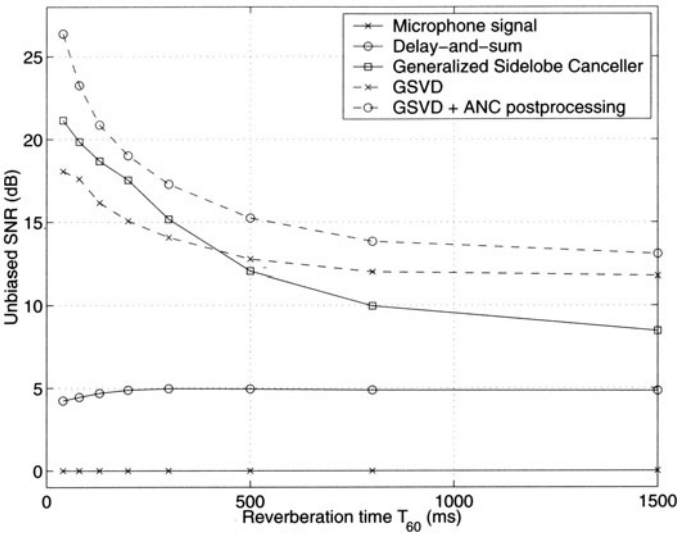


Fig. 6.11. Comparison of unbiased SNR for different signal enhancement algorithms for different reverberation times ($N = 4$, $L = 20$, $L_{ANC} = 800$, $\alpha = 1$, SNR= 0 dB)

References

1. M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise : A regenerative approach," *Speech Communication*, vol. 10, no. 2, pp. 45–57, Feb. 1991.
2. Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.
3. S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of Broad-Band Noise in Speech by Truncated QSVD," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 6, pp. 439–448, Nov. 1995.
4. U. Mittal and N. Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise," *IEEE Trans. Speech, Audio Processing*, vol. 8, no. 2, pp. 159–167, Mar. 2000.
5. F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech Enhancement Based on the Subspace Method," *IEEE Trans. Speech, Audio Processing*, vol. 8, no. 5, pp. 497–507, Sept. 2000.
6. S. Doclo and M. Moonen, "SVD-based optimal filtering with applications to noise reduction in speech signals," in *Proc. of the IEEE Workshop Applicat. Signal Processing to Audio and Acoust. (WASPAA'99)*, New Paltz, NY, USA, Oct. 1999, pp. 143–146.
7. S. Doclo and M. Moonen, "Robustness of SVD-based Optimal Filtering for Noise Reduction in Multi-Microphone Speech Signals," in *Proc. of the 1999 IEEE Int. Workshop Acoust. Echo and Noise Control (IWAENC'99)*, Pocono Manor, PA, USA, Sept. 1999, pp. 80–83.

8. S. Doclo and M. Moonen, "Noise Reduction in Multi-Microphone Speech Signals using Recursive and Approximate GSVD-based Optimal Filtering," in *Proc. IEEE Benelux Signal Processing Symp. (SPS2000)*, Hilvarenbeek, The Netherlands, Mar. 2000.
9. S. Doclo, E. De Clippel, and M. Moonen, "Multi-microphone noise reduction using GSVD-based optimal filtering with ANC postprocessing stage," in *Proc. of DSP2000 Workshop*, Hunt, TX, USA, Oct. 2000.
10. S. Doclo, E. De Clippel, and M. Moonen, "Combined Acoustic Echo and Noise Reduction using GSVD-based Optimal Filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP2000)*, Istanbul, Turkey, June 2000, vol. 2, pp. 1061–1064.
11. S. Van Gerven and F. Xie, "A Comparative Study of Speech Detection Methods," in *Proc. EUROSPEECH*, Rhodos, Greece, Sept. 1997, vol. 3, pp. 1095–1098.
12. S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech, Audio Processing*, vol. 8, no. 4, pp. 478–482, July 2000.
13. F. T. Luk, "A parallel method for computing the generalized singular value decomposition," *Internat. J. Parallel Distr. Comp.*, vol. 2, pp. 250–260, 1985.
14. G. H. Golub and C. F. Van Loan, *Matrix Computations*, MD : John Hopkins University Press, Baltimore, 3rd edition, 1996.
15. P. Butler and A. Cantoni, "Eigenvalues and eigenvectors of symmetric centrosymmetric matrices," *Linear Algebra and its Applications*, vol. 13, pp. 275–288, Mar. 1976.
16. I. Dologlou and G. Carayannis, "Physical Representation of Signal Reconstruction from Reduced Rank Matrices," *IEEE Trans. Signal Processing*, vol. 39, no. 7, pp. 1682–1684, July 1991.
17. S. Doclo and M. Moonen, "SVD-based optimal filtering with applications to noise reduction in speech signals," Tech. Rep. ESAT-SISTA/TR 1999-33, ESAT, K.U.Leuven, Belgium, Apr. 1999.
18. J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.
19. C. C. Paige, "Computing the generalized singular value decomposition," *SIAM J. Sci. Statist. Comput.*, vol. 7, pp. 1126–1146, 1986.
20. C. Van Loan, "Computing the CS and the Generalized Singular Value Decomposition," *Numer. Math.*, , no. 46, pp. 479–491, 1985.
21. J. P. Charlier, M. Vanbegin, and P. Van Dooren, "On efficient implementations of Kogbetliantz's algorithm for computing the singular value decomposition," *Numer. Math.*, vol. 52, pp. 279–300, 1988.
22. M. Moonen, P. Van Dooren, and J. Vandewalle, "A Singular Value Decomposition Updating Algorithm for Subspace Tracking," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 4, pp. 1015–1038, Oct. 1992.
23. M. Moonen, P. Van Dooren, and J. Vandewalle, "A systolic algorithm for QSVD updating," *Signal Processing*, vol. 25, pp. 203–213, 1991.
24. W. M. Gentleman, "Least squares computation by Givens transformations without square roots," *J. Inst. Math. Appl.*, vol. 12, pp. 329–336, 1973.
25. M. Moonen, P. Van Dooren, and J. Vandewalle, "A systolic array for SVD updating," *SIAM Journal on Matrix Analysis and Applications*, vol. 14, no. 2, pp. 353–371, 1993.

26. L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, pp. 27–34, Jan. 1982.
27. K. M. Buckley, "Broad-Band Beamforming and the Generalized Sidelobe Canceller," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, no. 5, pp. 1322–1323, Oct. 1986.
28. J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Theoretical Noise Reduction Limits of the Generalized Sidelobe Canceller (GSC) for Speech Enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Phoenix, AZ, USA, May 1999, vol. 5, pp. 2965–2968.
29. D. Van Compernolle, "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Albuquerque, USA, Apr. 1990, vol. 2, pp. 833–836.
30. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.

7 Explicit Speech Modeling for Microphone Array Applications

Michael Brandstein and Scott Griebel

Harvard University, Cambridge MA, USA

Abstract. In this chapter we address the limitations of current approaches to using microphone arrays for speech acquisition and advocate the development of multi-channel techniques which employ an explicit model of the speech signal. The goal is to combine the advantages of spatial filtering achieved through beamforming with knowledge of the desired time-series attributes. We offer two examples of algorithms which incorporate this principle. The first is a frequency domain approach illustrating the utility of model-based microphone array processing for improved speech spectral analysis. The second is a time domain procedure which highlights the benefits of speech modeling/spatial filtering fusion for the purpose of improving speech quality.

7.1 Introduction

A major thrust of microphone array research has focused on improving the spatial filtering capability of the system in its operating environment. The previous chapters have directly addressed a number of the methods which have proven beneficial in this regard. The goal of this chapter is to present an alternative and complementary strategy which emphasizes the incorporation of explicit speech modeling into the microphone array processing. As will be shown, by combining specific knowledge of the desired time-series attributes with the advantages of multi-channel spatial filtering, this approach can significantly improve the quality of the speech signal and its derived quantities obtained in challenging acoustic environments.

Much of this work is motivated from current research in the field of single-channel speech systems where, due to necessity perhaps, improvements have been obtained by focusing on the underlying content of the signal of interest in addition to its environmental degradations. There is a rich history of work addressing the use of single channel methods for speech enhancement. Summaries of these techniques may be found in texts on the subject, such as [1–3]. While capable of improving perceived quality in restrictive environments (additive noise, no multipath, high to moderate signal-to-noise ratio (SNR), single source), these approaches do not perform well in the face of reverberant distortions, competing sources, and severe noise conditions. In recent years, sophisticated speech models have been applied to the enhancement problem. In addition to utilizing the periodic features of the speech, as in the case of comb filtering, these systems exploit the signal's mixture of harmonic and

stochastic components [4–6]. Such model-based techniques offer an improved performance, both in speech quality and intelligibility. Additionally, these methods, by virtue of their specific parameterization of the speech signal, offer some applicability to the more general acquisition problem. Currently, however, these model-based estimation schemes have been limited to single channel applications.

By employing spatial filtering in addition to temporal processing, microphone arrays offer a distinct performance advantage over single-channel techniques in the presence of additive noise, interfering sources, and distortions due to multipath channel effects. Fixed and adaptive adaptive beamforming techniques generally assume the desired source is slowly varying and at a known location. While dynamic localization schemes, like those of [7], and robust weighting constraints may be incorporated into the adaptation procedure, these methods are very sensitive to steering errors which limit their noise source attenuation performance and frequently distort or cancel the desired signal. Furthermore, these algorithms are oriented solely toward noise reduction and have limited effectiveness at enhancing a desired signal corrupted by reverberations.

Similarly, the class of post-filtering algorithms detailed in Chapter 3 are limited in their assumption that the additive noise is uncorrelated. Once again the emphasis here is additive noise reduction. The performance (and appropriateness) of these post-filters quickly degrades in the presence of multi-path distortions and coherent noise.

Another general approach is based upon attempting to undo the effects of multipath propagation. The channel responses themselves are in general not minimum phase and are thus non-invertible. By beamforming to the direct path and the major images, it is possible to use the multipath reflections constructively to increase SNR's well beyond those achieved with a single beamformer. The result is a matched filtering process [8] which is effective in enhancing the quality of reverberant speech and attenuating noise sources. Unfortunately, this technique has a number of practical shortcomings. The matched filter is derived from the source location-dependent room response and as such is difficult to estimate dynamically. The channel responses obtained in this manner do not address the issue of non-stationary or unknown source locations or changing acoustic environments. These problems are addressed by Affes and Grenier [9] in attempting to adaptively estimate the channel responses and incorporate the results into an adaptive beamforming process.

In general, beamforming research has dealt with algorithms to attenuate undesired sources and noise, track moving sources, and deconvolve channel effects. These approaches, while effective to some degree, are fundamentally limited by the nature of the distant talker environment. Array design methods are overly sensitive to variations in their assumptions regarding source locations and radiation patterns, and are inflexible to the complex and time-

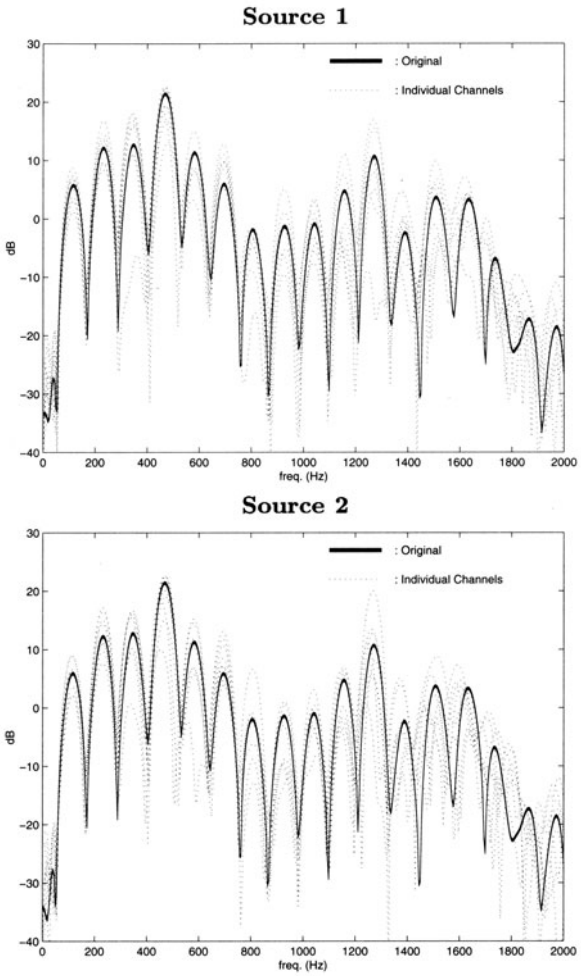


Fig. 7.1. Spectra of a voiced speech segment and its individual channels simulated at source locations spaced 10cm apart.

varying nature of the enclosure's acoustic field. Motion as little as a few centimeters or a talker turning his or her head is frequently sufficient to compromise the optimal behavior of these schemes in practical scenarios [10,11]. Similarly, matched filter processing, while shown to be capable of tracking source motion to a limited degree, requires significant temporal averaging and is not adaptable at rates sufficient to effectively capture the motions of a realistic talker.

To illustrate the acoustic environment's sensitivity to source location variations, a simple example is presented. Two source locations spaced 10cm apart are simulated in the center of a noiseless 4m x 4m x 3m rectangular

room. The enclosure is assumed to have plane reflective surfaces and uniform, frequency-independent reflection coefficients equivalent to a 400ms reverberation time. Room impulse responses are generated for 8 microphones with 25cm spacing positioned along one wall of the enclosure using the Allen and Berkley image model technique [12] with intra-sample interpolation and up to sixth order reflections. Both the microphones and sources are assumed to have cardioid patterns and the sources are oriented toward the center of the array. Figure 7.1 plots the spectra of a voiced speech segment generated at each source location. The bold lines correspond to the spectrum of the original speech while the dotted lines plot the spectra of the data received at each of 8 microphones. The reverberation effects are multiplicative in the frequency domain and vary considerably from channel to channel. The results of this simulation show that there are significant variations in the spectra of the individual channels (up to 10dB for some frequencies) when the source is moved just a few inches. An implication of this example is that any system which attempts to estimate the reverberation effects and apply some means of inverse filtering would have to be adaptable on almost a frame-by-frame basis to be effective. However, the temporal averaging required by these processes prohibits adaptation at such a high rate.

7.2 Model-Based Strategies

Single-channel techniques exploit various features of the speech signal; multi-channel methods focus primarily on improving the quality of the spatial filtering process. In [11] we proposed an alternative to the traditional microphone array methods by explicitly incorporating the Dual Excitation Speech Model [5] into the beamforming process. Our work in [13] extended this idea using a multi-channel version of the Multi-Pulse Linear Predictive Coding (MPLPC) model [14,15] and a nonlinear event-based processing method to discriminate impulses in the received signals due to channel effects from those present in the desired speech. These concepts were then expanded upon in [16,17] by employing the wavelet domain for a multi-resolution signal analysis and reconstruction of the LPC residual. These works illustrated the ability of this approach to suppress the deleterious effects of both reverberations and additive noise without explicitly identifying the channel while being adaptive on a frame by frame basis. This allows the model-based processing paradigm to be applied for effective speech analysis and enhancement under general conditions.

Two examples of this prior work are summarized here. The first is a frequency domain approach illustrating the utility of model-based microphone array processing for improved speech spectral analysis. The second is a time domain procedure which highlights the benefits of the speech modeling/spatial filtering fusion advocated for the purpose of improving speech quality.

7.2.1 Example 1: A Frequency-Domain Model-Based Algorithm

In the single-channel, Dual Excitation (DE) Speech Model [5], a windowed segment of speech, $s[n]$, is represented as the sum of two components: a voiced signal, $v[n]$, and an unvoiced signal, $u[n]$. In the frequency-domain, the relationship may be expressed as:

$$S(\omega) = V(\omega) + U(\omega) \tag{7.1}$$

where $S(\omega)$, $V(\omega)$, and $U(\omega)$ correspond to the Fourier transforms of $s[n]$, $v[n]$, and $u[n]$, respectively. The voiced portion is assumed to be periodic over the time window and may be represented as the sum of the harmonics of a fundamental frequency, ω_0:

$$V(\omega) = \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \tag{7.2}$$

where $W(\omega)$ is the Fourier Transform of the analysis window, A_m is the complex spectral amplitude of the mth harmonic, and M is the total number of harmonics ($M = \lfloor \pi/\omega_0 \rfloor$). Following [18], the fundamental frequency and harmonic amplitudes are estimated through minimization of the mean-squared error criterion:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(\omega) - V(\omega)|^2 d\omega \tag{7.3}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(\omega) - \sum_{m=-M}^{M} A_m W(\omega - m\omega_0)|^2 d\omega. \tag{7.4}$$

This non-linear optimization problem may be decoupled efficiently by noting that for a given fundamental frequency, the harmonic amplitudes which minimize the error are found through the solution of a set of linear equations. The optimal parameter set may then be calculated through global minimization of the error function in (7.4) versus all fundamental frequencies of interest. To effectively model the spectrum of the higher harmonics, a fundamental frequency resolution of less than 1 Hz is typically required. In practice, this exhaustive procedure may be computationally prohibitive. A more efficient approach is to evaluate a coarse, integer pitch estimate via a traditional time-domain pitch estimation procedure and then use the above frequency-domain analysis by synthesis procedure to refine the fundamental frequency estimate.

The estimated unvoiced signal plus noise spectrum, $\hat{U}(\omega)$, is then found from the difference spectrum:

$$\hat{U}(\omega) = S(\omega) - \hat{V}(\omega) \tag{7.5}$$

where $\hat{V}(\omega)$ is the estimated voiced spectrum derived from (7.2) using the estimated values of ω_0 and A_m.

The utility of the DE model for improving speech degraded by background noise lies in its independent enhancement of the voiced and unvoiced components of the speech. Assuming that the degrading noise is independent of the harmonic structure, the voiced spectrum is subjected to only a minor thresholding operation relative to the background noise power. The bulk of the enhancement is achieved by nulling out the unvoiced portions of strongly voiced harmonics and applying a modified Wiener filter to the remaining unvoiced spectral regions.

The Dual Excitation model is extended here for the multi-channel problem to improve its effectiveness for the additive noise case and to address the more general distant-talker scenario involving multipath channels and multiple sources. Consider first the extension of the DE error criterion in (7.4) to include data from N channels:

$$\mathcal{E}_N = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{i=1}^{N} G_i(\omega) S_i(\omega) - \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \right|^2 d\omega \qquad (7.6)$$

where $G_i(\omega)$ is the filter weighting associated with the ith channel and $S_i(\omega)$ is the short-term spectrum of the data received at the ith microphone. Alternatively, for environments where the dominant degradation effect is reverberant, it may be advantageous to recast the above error criterion as the L_2 norm in the log spectrum domain.

The voiced signal estimate, $\hat{V}_N(\omega)$, derived from the parameters minimizing (7.6) would then be used to produce the unvoiced signal plus noise spectrum from:

$$\hat{U}_N(\omega) = \frac{1}{N} \sum_{i=1}^{N} H_i(\omega) \left[G_i(\omega) S_i(\omega) - \hat{V}_N(\omega) \right] \qquad (7.7)$$

The channel weightings, $G_i(\omega)$, could be designed to provide appropriate spatial filtering, addressing issues of noise-reduction, attenuation of interfering sources, and dereverberation. Additionally, the channel-dependent weighting filters, $H_i(\omega)$, could be incorporated as a multi-channel post-processor to exploit known signal characteristics.

In the simplest case of independent additive noise, the extension of the Dual Excitation model to a plurality of channels would stand to improve its enhancement performance by virtue of the data averaging alone. With the inclusion of the spatial filtering afforded through (7.6) and (7.7) it is possible to give the DE model a robustness to channel effects and interfering sources. With regard to multiple sources, the error criterion in (7.6) could be extended explicitly to include L sources and N channels by:

$$\mathcal{E}_{LN} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1}{L} \sum_{j=1}^{L} \left(\sum_{i=1}^{N} G_{ij}(\omega) S_i(\omega) - \sum_{m=-M}^{M} A_{mj} W(\omega - m\omega_{0j}) \right) \right|^2 d\omega$$

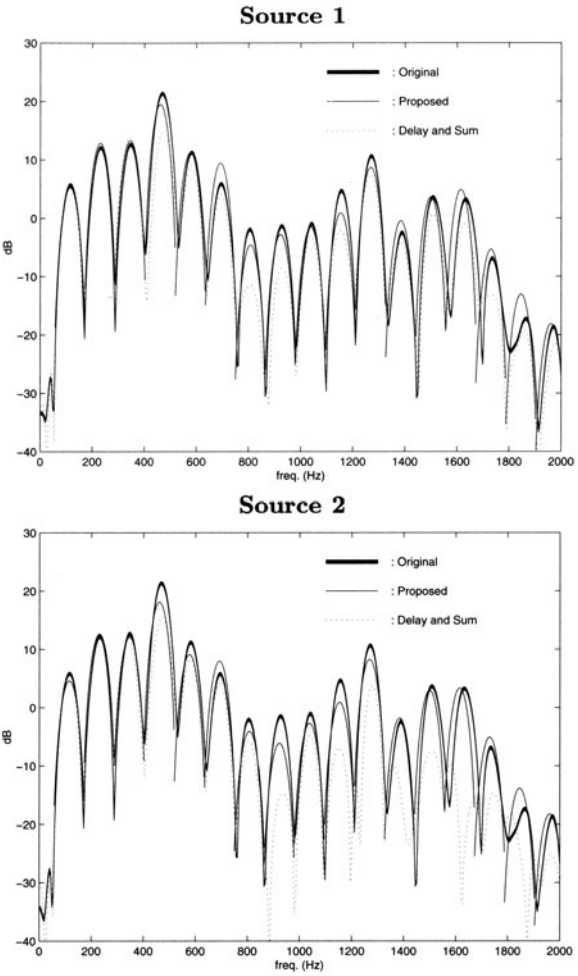


Fig. 7.2. Spectra of the Original, Delay and Sum Beamformer result, and Voiced Multi-Channel Dual Excitation result for the voiced segment in Figure 7.1.

where $G_{ij}(\omega)$ is the spatial filter associated with the ith channel and jth signal source, ω_{0j} is the fundamental frequency of the jth source, and A_{mj} is the amplitude of the mth harmonic associated with the jth source. Using this approach it would be possible to track individual sources through a combination of location and pitch data. Such a multi-channel DE model would have the ability to isolate and enhance a desired source signal by employing both spatial and signal-content information.

To illustrate the potential of such an approach, again consider the example of the voiced speech segment in Fig. 7.1. Figure 7.2 shows the relationship between the Delay and Sum Beamformer and the voiced signal estimate, $\hat{V}_N(\omega)$,

derived from the proposed multi-channel scheme for the two closely-spaced source locations. The pair of results was generated using delays appropriate for the source 1 location. This would correspond to a 10cm mis-aim in the source 2 case. The Delay and Sum method, like any beamforming technique, lacks any signal-dependent constraints on the output produced. As the plots suggest, by exploiting the periodic structure of the desired signal, the Multi-Channel Dual Excitation Model is significantly more robust to the local spectral variations produced by channel reverberations. Unlike the Delay and Sum method, the approach is relatively insensitive to imperfect knowledge of the source location suggesting a robustness to the small, but nominal, variations encountered in a practical operating environment. This result is confirmed by more quantitative methods, such as SNR and log spectral distortion scores.

7.2.2 Example 2: A Time-Domain Model-Based Algorithm

The reverberant speech signal, $x_i[n]$, observed at the ith microphone ($i = 1, 2, \ldots, I$) can be modeled in the time-domain as:

$$x_i[n] = h_i[n] * s[n] + u_i[n] \tag{7.8}$$

where $s[n]$ is the clean speech utterance, $u_i[n]$ is noise, and $h_i[n]$ is the room impulse response between the speech source and the ith microphone. Under all but ideal circumstances, the room impulse response is a complicated function of the environment's acoustical and geometrical properties. The noise term, $u_i[n]$, is assumed to be uncorrelated, both temporally and spatially.

A very general model for speech production approximates the vocal tract as a time-varying all-pole filter [2]. In the case of voiced speech, the filter excitation is modeled as a quasi-periodic impulse train where the average width between consecutive impulses is the pitch period. For unvoiced speech, the excitation signal is approximated by random noise. The proposed algorithm relies on the assumption that the detrimental effects of additive noise and reverberations introduce only zeros into the overall system and will primarily affect only the nature of the speech excitation sequence, not the all-pole filter. It is also assumed that the noise and errant impulses contributed to the excitation sequences are relatively uncorrelated across the individual channels, while the excitation impulses due to the original speech are correlated after performing time-delay compensation. Essentially, the approach will be to identify the clean speech excitation signal from a set of corrupted excitation signals. The enhanced speech is then reconstructed by using the enhanced excitation signal as input to an estimate of the all-pole filter representing the vocal system.

The proposed algorithm offers an effective method for estimating and then reconstructing the excitation signal by employing a class of wavelets to decompose the LPC residual signals. In [19], quadratic spline wavelets

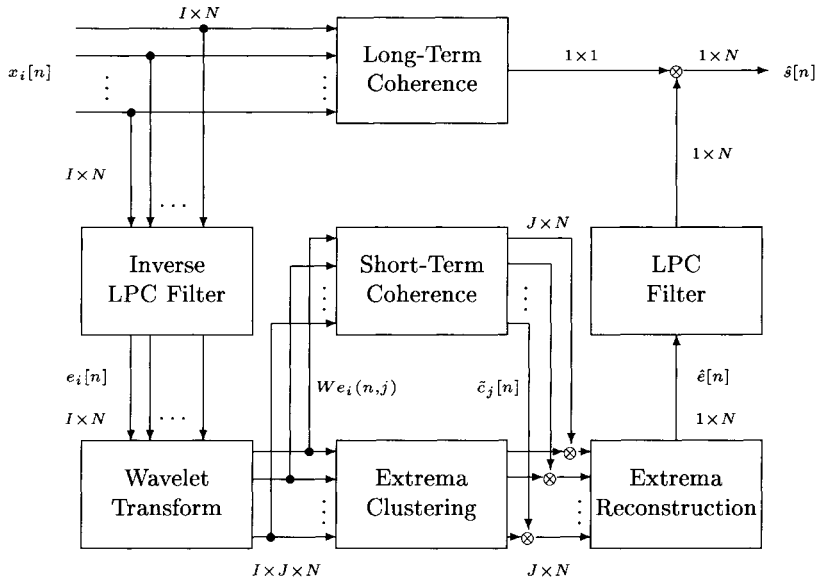


Fig. 7.3. Outline of the proposed algorithm for an N-point frame of speech.

were shown to be effective at detecting singularities in a signal. Because these quadratic spline wavelets are the derivatives of smoothing functions, significant wavelet transform (WVT) coefficients, or wavelet extrema, prove to be appropriate indicators of discontinuities in a signal. The set of extrema at each wavelet scale is an effective representation of the impulses in the excitation signals. By locating the extrema which are well clustered across all microphone channels, we attempt to capture the underlying impulsive structure (for voiced speech) of the original non-reverberant LPC residual signal. The estimated 'clean' multi-scale extrema can be used to reconstruct the desired excitation signal with minimal approximation error [19].

The algorithm is outlined in Fig. 7.3. The received signals, $x_i[n]$, are assumed to have been time-aligned either through *a priori* knowledge of the source and microphone geometry or as the result of applying a relative time-delay estimation procedure to the original channel data. For the following algorithm discussion, the analysis frame is 32ms ($N = 256$ samples at 8kHz sampling rate) and the number of channels is $I = 8$. Details of the simulation procedure are provided below.

A joint LPC filter is derived from N-point windows of the I channels of speech. The linear prediction formalism assumes the speech source signal can be modeled as:

$$s[n] = \sum_{p=1}^{P} a_p s[n-p] + e[n] \tag{7.9}$$

where $e[n]$ is the residual or excitation signal discussed previously. The LPC coefficients, $\{a_1, a_2, \ldots, a_P\}$, are estimated by minimizing the average energy of the residual signal. For this multi-channel setting the criterion is expressed as:

$$\frac{1}{I}\sum_{i=1}^{I}\sum_{n=1}^{N} e_i^2[n] = \frac{1}{I}\sum_{i=1}^{I}\sum_{n=1}^{N}\left\{ s[n] - \sum_{p=1}^{P} a_p s[n-p] \right\}^2 . \tag{7.10}$$

The minimum energy solution is achieved by solving a system of linear equations as in the single-channel case, but involves a multi-channel autocorrelation function, given by:

$$r[k] = \frac{1}{I}\sum_{i=1}^{I}\sum_{n=k}^{N} x_i[n]x_i[n-k]. \tag{7.11}$$

The resulting LPC coefficients can be used to generate a set of residual signals, $e_i[n]$, via inverse filtering. The goal of the following wavelet extrema reconstruction is to use the multiple residual signals to create a single enhanced residual signal, $\hat{e}[n]$. This residual will then be used as input to the previously estimated LPC filter to generate an enhanced speech signal, $\hat{s}[n]$. For the current simulations, the LPC model order is $P = 13$.

The first J scales of the WVT are calculated for each of the I residual signals. Here the WVT is calculated for the first $J = 5$ scales. The WVT of the residual signal of the ith channel, $e_i[n]$, is denoted by $We_i(n, j)$ where j is the scale parameter and n is the temporal index within that scale. We use the non-decimated quadratic spline wavelet discussed in [19]. Because the wavelet is the derivative of a smoothing function, the resulting WVT coefficients will contain local extrema at large changes in the signal. Therefore, impulses in the residual signal of voiced speech segments will be manifested as large WVT coefficients at different scales. This same wavelet has been used on non-reverberant speech for pitch detection [20].

In the following stages, this algorithm is used to reconstruct an enhanced residual signal, where wavelet extrema at different scales correspond to impulses in the voiced excitation. While the significance of the wavelet extrema in the unvoiced case is not as clearly delineated, this approach is effective in reconstructing an appropriate noise-like residual signal.

Based upon the observation that residual impulses due to the original speech tend to be clustered across channels and scales (after time-delay compensation) while those due to reverberation effects are relatively uncorrelated

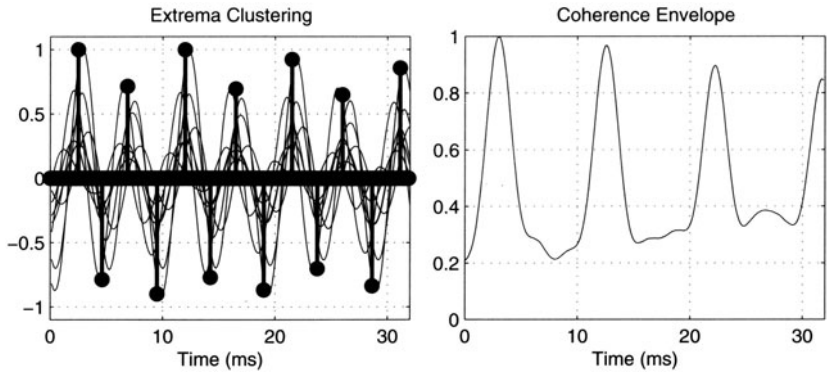


Fig. 7.4. Clustering Results and Coherence Envelope.

in both amplitude and time, a clustering technique is now applied to estimate the 'clean' extrema. In [13] we presented a nonlinear clustering criterion for this application, while in [16] we employed a Gaussian smoothing approach. However, the local extrema of WVT coefficients of the delay-and-sum beamformed speech have been found to perform sufficiently well in this context. Because of its computational advantages and near comparable performance, we will employ this delay and sum strategy for extrema clustering in what follows. The left-hand plot in Figure 7.4 illustrates a single scale's WVT coefficients for 8 channels and the resulting clustered extrema. In general, for voiced speech, extrema due to glottal excitations tend to be more closely clustered than those due to additional impulses from the reverberant channel responses. This causes the less clustered peaks to have a relatively small amplitude. The clustering operation maps the multi-channel WVT extrema data to a single WVT extrema representation suitable for input to the WVT extrema reconstruction algorithm described later.

We now apply the first of two coherence-based scaling functions. The purpose of the amplitude modulation scheme detailed here is to further de-emphasize the clustered peaks due to room reverberations beyond what is achieved by the averaging process associated with the extrema clustering. The motivation is that wavelet extrema due to room effects will be less coherent from channel to channel than those due to the vocal excitation signal.

The following short-term coherence window is multiplied by the wavelet extrema at each scale j:

$$c_j[n] = \frac{2}{I-1} \frac{\sum_{i=1}^{I} \sum_{k=i+1}^{I} \sum_{l=-L}^{L} We_i(n+l,j) We_k(n+l,j)}{\sum_{i=1}^{I} \sum_{l=-L}^{L} We_i(n+l,j)^2} \qquad (7.12)$$

where $2L+1 \ll N$ is the length of a small subwindow within the analysis frame, on the order of 3 to 5ms. With this window, any clustered extrema in areas of low coherence will be de-emphasized, while highly coherent extrema

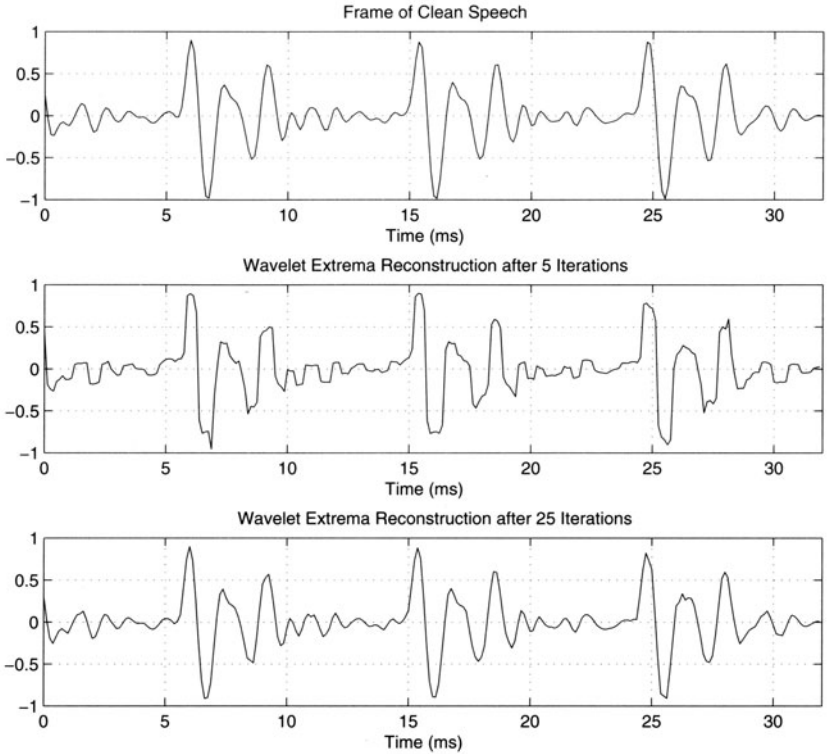


Fig. 7.5. Clean Speech and Wavelet Extrema Reconstructions After 5 and 25 Iterations.

will be relatively unaffected. The coherence function as given above is limited to values between -1 and 1. A thresholding scheme is incorporated to maintain nonnegative scaling terms and to normalize the coherence values within the subwindow. The right-hand plot in Fig. 7.4 shows a typical coherence envelope used for weighting the clustered WVT extrema.

Another advantage of working in the wavelet domain is the ability to approximately reconstruct a signal from the local extrema of its WVT coefficients at each scale [21]. To summarize this process, suppose the wavelet transform of a signal, $f[n]$, with local extrema at $n = n_{j,k}$ in scale j is to be approximated by $\hat{f}[n]$. By minimizing the norm of $\hat{f}[n]$ and imposing the constraint $Wf(n_{j,k}, j) = W\hat{f}(n_{j,k}, j)$, it is possible to reconstruct $\hat{f}[n]$ via a conjugate gradient approach [21]. Figure 7.5 shows a clean voiced speech segment and its wavelet extrema reconstruction after 5 and 25 iterations of the conjugate gradient method. Note that this method tends to approximate a signal to within a very high frequency error term, which, for the speech signals we studied, was inaudible.

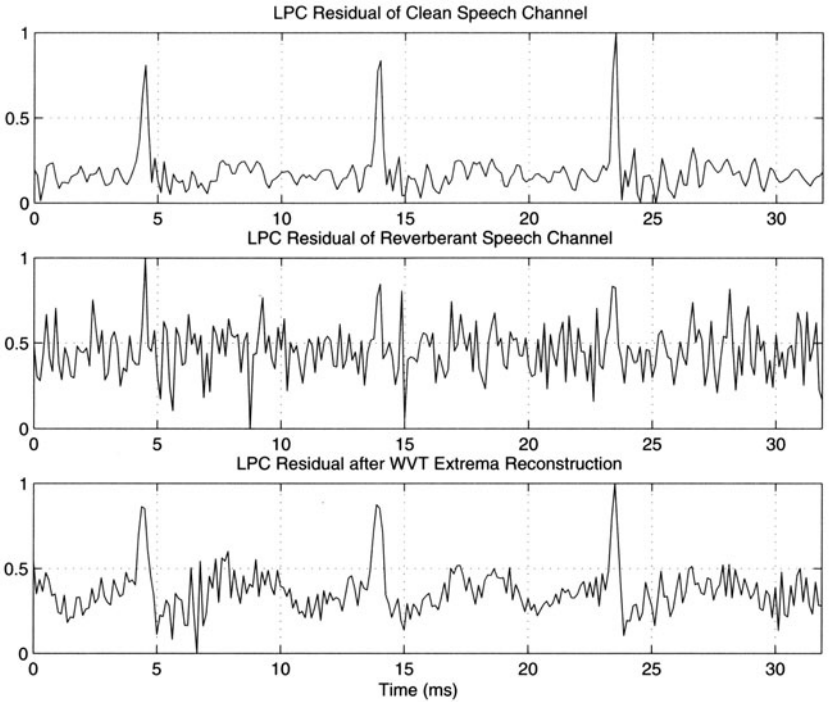


Fig. 7.6. LPC Residual of Clean Speech, After Beamforming, and After Wavelet Clustering Technique.

After multiplying the clustered peaks for each wavelet scale by the appropriate scale-dependent coherence envelope, the enhanced LPC residual is reconstructed using the wavelet extrema reconstruction algorithm just discussed. Listening tests show that the variations in amplitude produced by the extrema synthesis process across consecutive periods of voiced speech can produce audible distortions in the resulting speech signal. To lessen this effect, a quasi-pitch-synchronous smoothing approach is employed at the coarser wavelet scales to ensure that the peaks from consecutive voiced periods are slowly varying. This procedure involves averaging WVT extrema with their corresponding extrema in neighboring pitch periods and implicitly assumes a pitch estimate is available. In [22] a variety of multi-channel pitch estimators are evaluated in the environmental conditions considered here. Several are shown to be effective for reverberation levels of 0 to 1000ms.

Figure 7.6 shows an example of the LPC residual for clean and reverberant speech frames followed by the residual signal resulting from the WVT extrema reconstruction algorithm.

For each analysis frame, the reconstructed residual signal, $\hat{e}[n]$, is applied to the estimated LPC filter to generate an enhanced speech signal. Figures 7.7 and 7.8 offer the results for two different frames of speech. Figure 7.7 displays

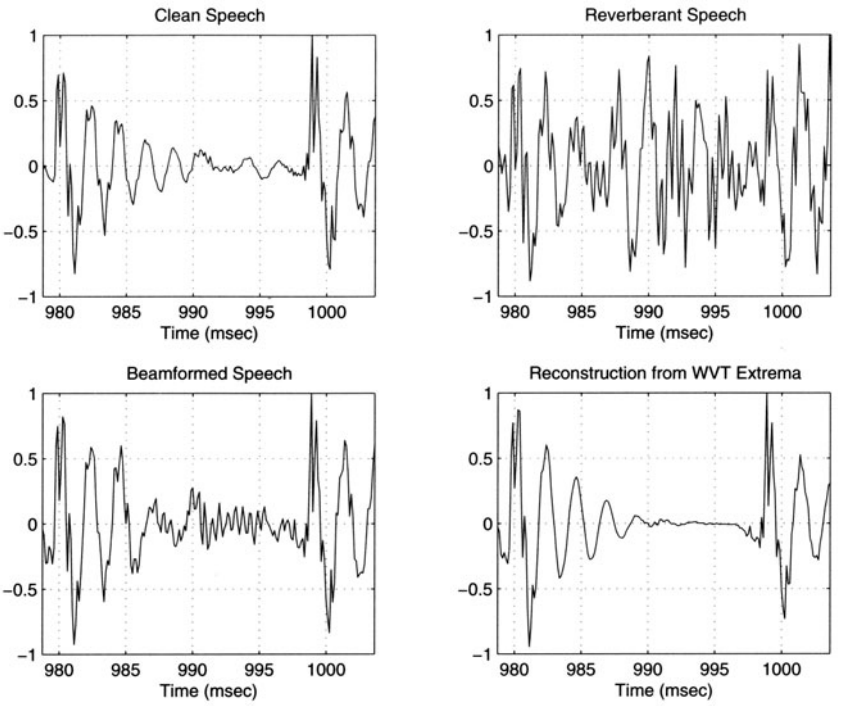


Fig. 7.7. Comparison 1: Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech.

a single analysis frame corresponding to a little over one period of a voiced speech segment. The four plots compare the reconstructed speech signal to the original speech utterance, a single reverberant channel, and a simple delay-and-sum beamforming of the I reverberant speech channels. It is apparent that the primary effect of the multipath distortion is to inflate the signal energy between 987ms and 997ms. The beamformer is ineffective at attenuating these distortions. The proposed algorithm is capable of detecting the excitation energy due to the clean speech and produces a synthesized signal very similar to the original. Figure 7.8 demonstrates a voiced speech segment from early in a vowel segment. At this point, the speech is more affected by additive noise than reverberations, and the enhancement procedure diminishes its presence.

The final step in the enhancement procedure is to scale the reconstructed speech frames by a measure that is based on both the long-term coherence of the speech channels and the energy of the beamformed signal. This second weighting procedure serves two purposes. First, it adjusts the energy of each frame to provide a relatively smoothly varying output. The short-term coherence window can vary substantially from frame to frame. The energy of the resulting speech will vary accordingly, potentially resulting in audible

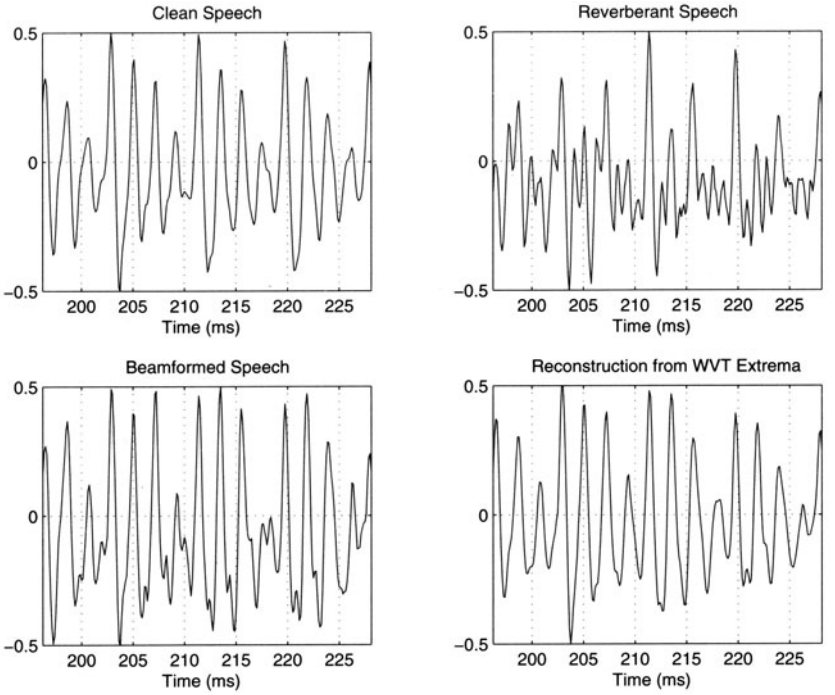


Fig. 7.8. Comparison 2: Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech.

distortions. The long-term coherence window incorporates an energy statistic to address this issue. Second, it allows for the de-emphasis of lengthy signal portions which are due primarily to reverberation effects. This is especially useful at the end of voiced sounds.

Simulations have been performed to evaluate the quality of the enhanced speech derived from this procedure. Again, a 4m × 4m × 3m rectangular enclosure assumed to have plane reflective surfaces and uniform, frequency-independent reflection coefficients was employed. A total of eight microphones were used: one pair on 3 of the walls and another pair on the ceiling, as shown in Figure 7.9. The source was located near the wall containing no microphones. The microphones and sources were all assumed to have cardioid patterns. Room reverberation times, T, ranged from 100ms to 1s. Uncorrelated Gaussian noise was added to each observed signal with SNR of 10dB.

Figures 7.10 and 7.11 illustrate a series of three second speech segments which have been simulated in the reverberant enclosure described above. Each figure plots the original speech, the signal received at a single microphone, the product of delay-and-sum beamforming, and the result achieved using the proposed enhancement algorithm. Figure 7.10 shows the results derived for 400ms reverberant speech while Fig. 7.11 shows the case for 400ms re-

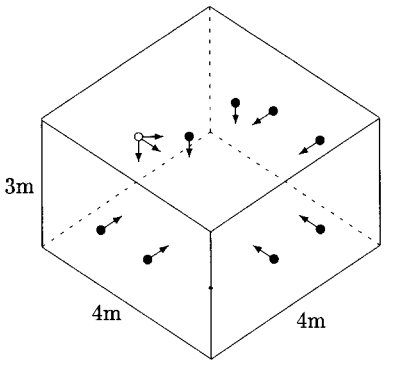


Fig. 7.9. Room Setup - • Represents Microphones, ○ Represents the Speech Source.

verberant speech plus additive noise. In this latter case, uncorrelated white Gaussian noise was added to each channel so that the SNR relative to the original non-reverberant utterance was 10dB. The superior performance of the proposed algorithm relative to delay-and-sum beamforming at removing both long and short term reverberation effects and additive noise is apparent. Listening tests confirm the efficacy of the procedure. The synthesized speech reduces the audible reverberation and noise effects which are clearly present in both the degraded and delay-and-sum beamformed speech.

To quantify the improvements of the proposed algorithm, the average Bark Spectral Distortion (BSD)[23] was calculated over all frames of speech. The BSD is a speech quality assessment method which employs a series of perceptually motivated transformations to the signal spectra prior to computing a spectral distance. The result is an objective distortion measure which correlates well with perceptual quality ratings. Figure 7.12 shows the Bark spectral distance results for reverberation times from 100ms to 1s with additive white noise (SNR = 10dB). The proposed algorithm shows an improvement over delay-and-sum beamforming at all reverberation levels. These figures are consistent with informal listening tests.

7.3 Conclusion

This work has sought to show how the analysis and enhancement of speech acquired in multi-channel environments benefits through the exploitation of known features in the desired speech signal. Rather than focusing solely on the spatial filtering aspects of the problem, we have advocated strategies which incorporate an explicit speech model into the microphone array processing.

Two specific samples of this philosophy were presented. The first was a frequency-domain procedure which employed the Dual Excitation Speech

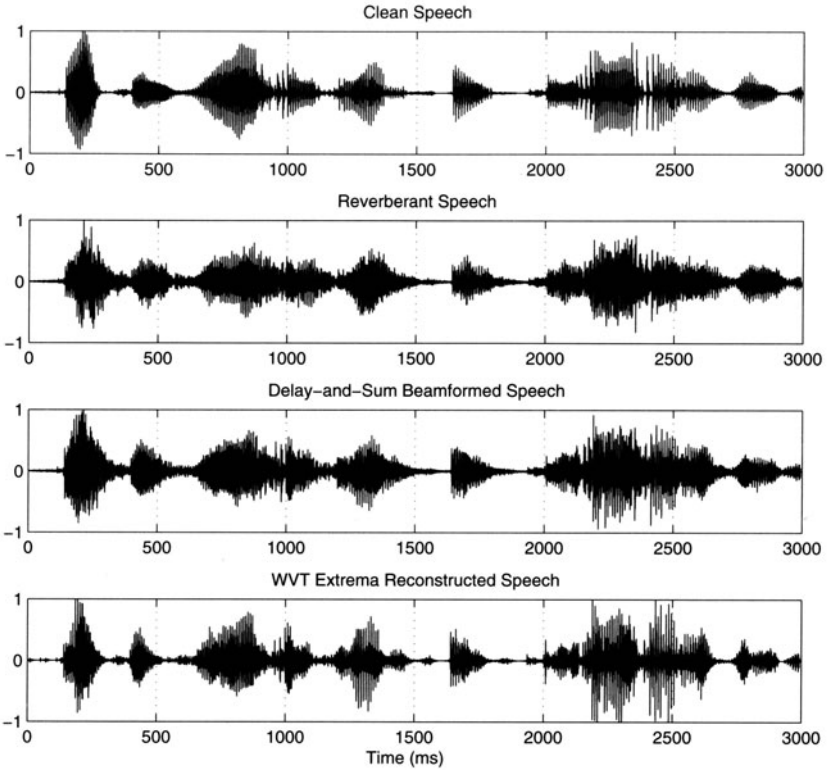


Fig. 7.10. Comparison of Clean, Reverberant, Beamformed, and the Proposed Algorithm (Reverberation-Only Case).

Model to utilize the periodicity of voiced speech. In the context of spectral analysis, we illustrated the improved performance obtained relative to Delay-and-Sum Beamforming in reverberant environments. We then outlined methods to address multi-source scenarios and for fusing this approach with traditional array processing techniques. While not presented here, we have applied this frequency-domain model to the problem of speech time-delay estimation in noisy and reverberant environments [24,25]. By emphasizing these spectral regions exhibiting a periodic structure, it is possible to achieve a time-delay estimator that outperforms standard methods based upon generalized cross-correlation.

The second example focused on time-domain residual onset data to illustrate the effectiveness and practicality of model-based processing for microphone array applications. The proposed wavelet-domain algorithm is capable of discriminating impulses of the excitation residual generated by the desired speech signal from those brought about by multipath echos and uncorrelated noise. The process achieves a significant attenuation of environmental rever-

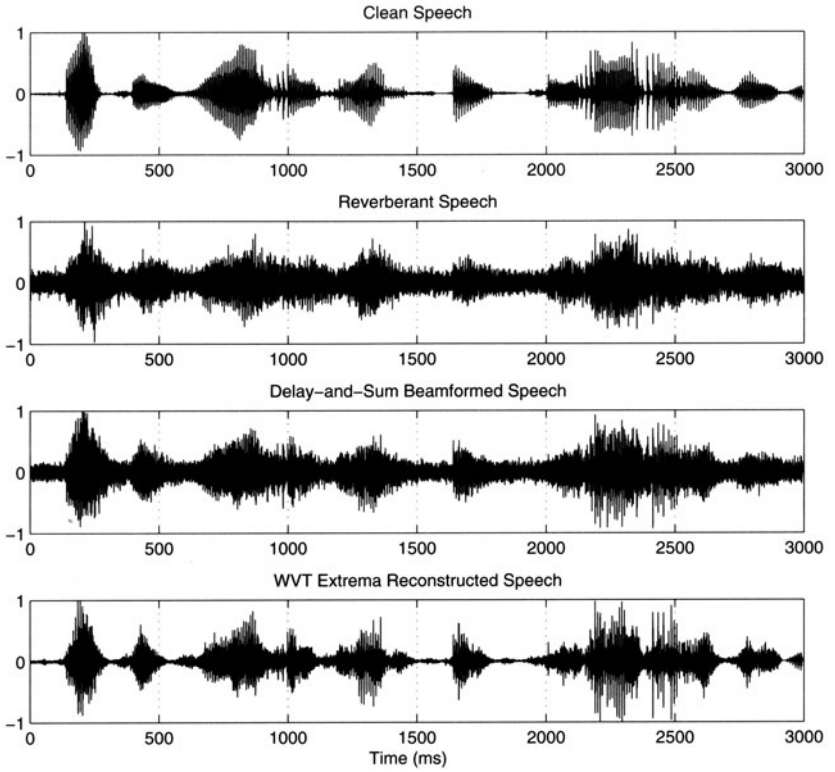


Fig. 7.11. Comparison of Clean, Reverberant, Beamformed, and the Proposed Algorithm (Reverberation plus Noise Case).

berations and additive noise without explicitly requiring estimation of the channel or noise characteristics. These principles may be extended to address the case of interfering sources and unknown source locations. Essentially, the impulse events may be associated with one or more source locations by evaluating their relative delays across the channels given knowledge of the microphone placements. Once this association is performed, the individual speech signals may be independently reconstructed through the synthesis procedure outlined above. This approach to speaker isolation represents a distinct contrast to the spatial filtering paradigm which relies on very specific knowledge and assumptions regarding talkers' locations and radiation patterns to generate appropriate channel weightings. The procedure outlined in this chapter, by virtue of its underlying speech model and exploitation of impulse data alone, has the potential to be much less sensitive to environmental uncertainties (e.g. imperfect source localization, non-ideal radiator effects, and unknown channels).

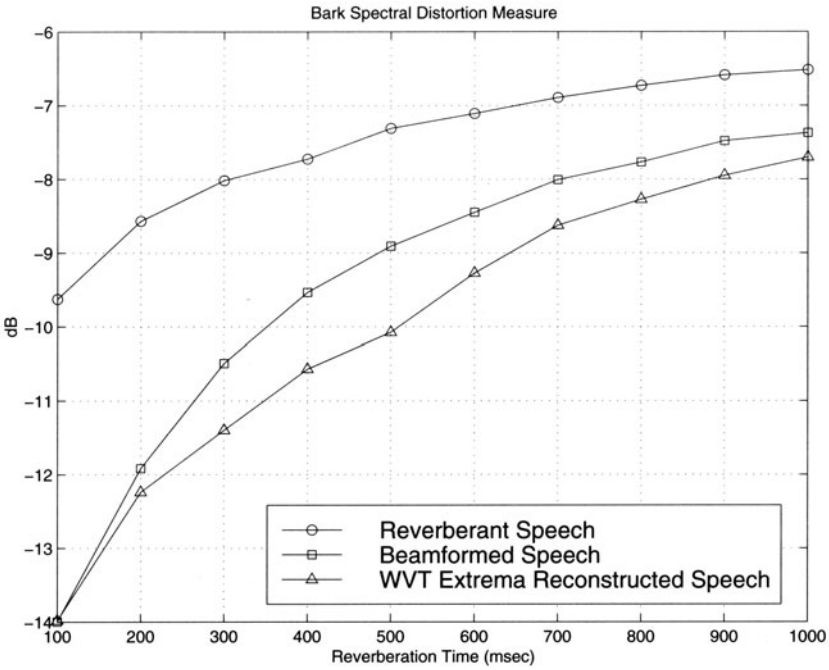


Fig. 7.12. Bark Spectral Distortion Results

There are a variety of speech models available for this application. This work has employed a pair of specific examples to illustrate the utility of the overall process. Similarly, the multi-channel methods based upon these models represent only two of the many possible analysis and enhancement schemes. Future work along these lines will investigate strategies to fuse proven methods in the single and multiple channel enhancement fields and to develop novel and effective algorithms.

References

1. J. Lim, ed., *Speech Enhancement.* Prentice Hall, 1983.
2. J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals.* Prentice Hall, first ed., 1987.
3. S. Furui and M. Sondhi, eds., *Advances in Speech Signal Processing.* Marcel Dekker, first ed., 1992.
4. R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, August 1986.
5. J. Hardwick, *The Dual Excitation Speech Model.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1992.

6. J. Laroche, Y. Stylianou, and E. Moulines, "Hns: Speech modification based on a harmonic + noise model," in *Proceedings of ICASSP93*, pp. II–550–II–553, IEEE, 1993.
7. M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, April 1997.
8. J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, no. 1-2, pp. 207–222, 1993.
9. S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 425–437, September 1997.
10. B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. ICASSP-99*, (Phoenix, AZ), pp. 881–884, March 15-19 1999.
11. M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. ICASSP-98*, (Seattle, WA), pp. 3613–3616, May 12-15 1998.
12. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, April 1979.
13. M. Brandstein, "An event-based method for microphone array speech enhancement," in *Proc. ICASSP-99*, (Phoenix, AZ), pp. 953–956, IEEE, March 15-19 1999.
14. B. S. Atal and J. R. Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *Proc. ICASSP-82*, pp. 614–617, 1982.
15. S. Singhal and B. S. Atal, "Improving performance of multi-pulse lpc coders at low bit rates," in *Proc. ICASSP-84*, pp. I–131–I–134, 1984.
16. S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverbearation," in *IEEE Workshop on Acoustic Echo and Noise Control*, (Pocono Manor, Pennsylvania), September 27-30 1999.
17. M. Brandstein and S. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing for Telecommunication* (S. Gay and J. Benesty, eds.), pp. 261–279, Kluwer, 2000.
18. D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. vol.36, pp. 1223–1235, August 1988.
19. S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 710–732, July 1992.
20. S. Kadambe and G. F. Boudreaux-Bartels, "Applications of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Information Theory*, vol. 38, pp. 917–924, March 1992.
21. S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.
22. S. M. Griebel, "Multi-channel wavelet techniques for reverberant speech analysis and enhancement," Tech. Rep. 5, HIMMEL, Harvard University, Cambridge, MA, February 1999.
23. S. Wang, A. Sekey, and A. Gersho, "An objective measure for predecting subjective quality of speech coders," *IEEE J. Selected Areas in Communications*, vol. 10, pp. 819–829, June 1992.

24. M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.
25. M. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, New York), October 19-22 1997.

Part II

Source Localization

8 Robust Localization in Reverberant Rooms

Joseph H. DiBiase[1], Harvey F. Silverman[1], and Michael S. Brandstein[2]

[1] Brown University, Providence RI, USA
[2] Harvard University, Cambridge MA, USA

Abstract. Talker localization with microphone arrays has received significant attention lately as a means for the automated tracking of individuals in an enclosure and as a necessary component of any general purpose speech capture system. Several algorithmic approaches are available for speech source localization with multi-channel data. This chapter summarizes the current field and comments on the general merits and shortcomings of each genre. A new localization method is then presented in detail. By utilizing key features of existing methods, this new algorithm is shown to be significantly more robust to acoustical conditions, particularly reverberation effects, than the traditional localization techniques in use today.

8.1 Introduction

The primary goal of a speech localization system is accuracy. In general, estimate precision is dependent upon a number of factors. Major issues include (1) the quantity and quality of microphones employed, (2) microphone placement relative to each other and the speech sources to be analyzed, (3) the ambient noise and reverberation levels, and (4) the number of active sources and their spectral content. The performance of localization techniques generally improves with the number of microphones in the array, particularly when adverse acoustic effects are present. This has spawned the research and construction of large array systems (e.g. 512 elements) [1]. However, when acoustic conditions are favorable and the microphones are positioned judiciously, source localization can be performed adequately using a modest number (e.g. 4 elements) of microphones. Performance is clearly affected by the array geometry. The optimal design of the array based on localization criteria is typically dependent on the room layout, speaking scenarios, and the acoustic conditions [2]. In practice, many of these design considerations are very dependent on the specific application conditions, the hardware available, and non-scientific cost criteria. In an effort to make its applicability as general as possible, this chapter will focus primarily on speech localization effectiveness as a function of the acoustic degradations present, namely background noise and reverberations, rather than attempt to address more specific environmental scenarios.

In addition to high accuracy, these location estimates must be updated frequently in order to be useful in practical tracking and beamforming appli-

cations. Consider the problem of beamforming to a moving speech source. It has been shown that for sources in close proximity to the microphones, the array aiming location must be accurate to within a few centimeters to prevent high-frequency rolloff in the received signal [3] and to allow for effective channel equalization [4]. A practical beamformer must therefore be capable of including a continuous and accurate location procedure within its algorithm. This requirement necessitates the use of a location estimator capable of fine resolution at a high update rate. Additionally, any such estimator would have to be computationally non-demanding and possess a short processing latency to make it practical for real-time systems.

These factors place tight constraints on the microphone data requirements. While the computation time required by the algorithm largely determines the latency of the locator, it is the data requirements that define theoretical limits. The work in [5], for example, focuses on reducing the size of the data segments necessary for accurate source localization in realistic room environments.

The goal of this chapter is to detail the issues associated with the problem of speech source localization in reverberant and noisy rooms and to present an effective methodology for its solution. While the focus will be the single-source scenario, the techniques described, in many cases, are applicable to situations where several individuals are conversing. The more general problem of simultaneous, multi-talker localization is addressed further in Chapter 9. The following section contains a summary of the existing genres for speech source localization using microphone arrays and highlights their relative merits. It is followed in Section 8.3 by the development of a speech source localization algorithm designed specifically for reverberant enclosures which combines two of these general approaches. Section 8.4 then offers some experimental results and conclusions.

8.2 Source Localization Strategies

Existing source localization procedures may be loosely divided into three general categories: those based upon maximizing the steered response power (SRP) of a beamformer, techniques adopting high-resolution spectral estimation concepts, and approaches employing time-difference of arrival (TDOA) information. These broad classifications are delineated by their application environment and method of estimation. The first refers to any situation where the location estimate is derived directly from a filtered, weighted, and summed version of the signal data received at the sensors. The second will be used to term any localization scheme relying upon an application of the signal correlation matrix. The last category includes procedures which calculate source locations from a set of delay estimates measured across various combinations of microphones.

8.2.1 Steered-Beamformer-Based Locators

The first categorization applies to passive arrays for which the system input is an acoustic signal produced by the source. The optimal Maximum Likelihood (ML) location estimator in this situation amounts to a focused beamformer which steers the array to various locations and searches for a peak in output power. Termed *focalization*, derivations of the optimality of the procedure and variations thereof are presented in [6–8]. Theoretical and practical variance bounds obtained via focalization are detailed in [6,7,9] and the steered-beamformer approach has been extended to the case of multiple-signal sources in [10].

The simplest type of steered response is obtained using the output of a delay-and-sum beamformer. This is what is most often referred to as a conventional beamformer. Delay-and-sum beamformers apply time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. These signals are time-aligned and summed together to form a single output signal. More sophisticated beamformers apply filters to the array signals as well as this time alignment. The derivation of the filters in these filter-and-sum beamformers is what distinguishes one method from another.

Beamforming has been used extensively in speech-array applications for voice capture. However, due to the efficiency and satisfactory performance of other methods, it has rarely been applied to the talker localization problem. The physical realization of the ML estimator requires the solution of a nonlinear optimization problem. The use of standard iterative optimization methods, such as steepest descent and Newton-Raphson, for this process was addressed by [10]. A shortcoming of each of these approaches is that the objective function to be minimized does not have a strong global peak and frequently contains several local maxima. As a result, this genre of efficient search methods is often inaccurate and extremely sensitive to the initial search location. In [11] an optimization method appropriate for a multimodal objective function, Stochastic Region Contraction (SRC), was applied specifically to the talker localization problem. While improving the robustness of the location estimate, the resulting search method involved an order of magnitude more evaluations of the objective function in comparison to the less robust search techniques. Overall, the computational requirements of the focalization-based ML estimator, namely the complexity of the objective function itself as well as the relative inefficiency of an appropriate optimization procedure, prohibit its use in the majority of practical, real-time source locators.

Furthermore, the steered response of a conventional beamformer is highly dependent on the spectral content of the source signal. Many optimal derivations are based on *a priori* knowledge of the spectral content of the background noise, as well as the source signal [7,8]. In the presence of significant reverberation, the noise and source signals are highly correlated, making ac-

curate estimation of the noise infeasible. Furthermore, in nearly all array-applications, little or nothing is known about the source signal. Hence, such optimal estimators are not very practical in realistic speech-array environments.

The practical shortcomings of applying correlation-based localization estimation techniques without a great deal of intelligent pruning is typified by the system produced in [12]. In this work a sub-optimal version of the ML steered-beamformer estimator was adapted for the talker-location problem. A source localization algorithm based on multi-rate interpolation of the sum of cross-correlations of many microphone pairs was implemented in conjunction with a real-time beamformer. However, because of the computational requirements of the procedure, it was not possible to obtain the accuracy and update rate required for effective beamforming in real-time given the hardware available.

8.2.2 High-Resolution Spectral-Estimation-Based Locators

This second categorization of location estimation techniques includes the modern beamforming methods adapted from the field of high-resolution spectral analysis: autoregressive (AR) modeling, minimum variance (MV) spectral estimation, and the variety of eigenanalysis-based techniques (of which the popular MUSIC algorithm is an example). Detailed summaries of these approaches may be found in [13,14]. While these approaches have successfully found their way into a variety of array processing applications, they all possess certain restrictions that have been found to limit their effectiveness with the speech-source localization problem addressed here.

Each of these high-resolution processes is based upon the spatiospectral correlation matrix derived from the signals received at the sensors. When exact knowledge of this matrix is unknown (which is most always the case), it must be estimated from the observed data. This is done via ensemble averaging of the signals over an interval in which the sources and noise are assumed to be statistically stationary and their estimation parameters (location in this case) are assumed to be fixed. For speech sources, fulfilling these conditions while allowing sufficient averaging can be very problematic in practice.

With regard to the localization problem at hand, these methods were developed in the context of far-field plane waves projecting onto a linear array. While the MV and MUSIC algorithms have been shown to be extendible to the case of general array geometries and near-field sources [15], the AR model and certain eigenanalysis approaches are limited to the far-field, uniform linear array situation.

With regard to the issue of computational expense, a search of the location space is required in each of these scenarios. While the computational complexity at each iteration is not as demanding as the case of the steered-beamformer, the objective space typically consists of sharp peaks. This property precludes the use of iteratively efficient optimization methods. The sit-

uation is compounded if a more complex source model is adopted (incorporating source orientation or head radiator effects, for instance) in an effort to improve algorithm performance. Additionally, it should be noted that these high-resolution methods are all designed for narrowband signals. They can be extended to wideband signals, including speech, either through simple serial application of the narrowband methods or more sophisticated generalizations of these approaches, such as [16–18]. Either of these routes extends the computational requirements considerably.

These algorithms tend to be significantly less robust to source and sensor modeling errors than conventional beamforming methods [19,20]. The incorporated models typically assume ideal source radiators, uniform sensor channel characteristics, and exact knowledge of the sensor positions. Such conditions are impossible to obtain in real-world environments. While the sensitivity of these high-resolution methods to the modeling assumptions may be reduced, it is at the cost of performance. Additionally, signal coherence, such as that created by the reverberation conditions of primary concern here, is detrimental to algorithmic performance, particularly that of the eigenanalysis approaches. This situation may be improved via signal processing resources, but again at the cost of decreased resolution[21]. Primarily for these reasons, localization methods based upon these high-resolution strategies will not considered further in this work. However, this should not exclude their judicious use in other speech localization contexts, particularly multi-source scenarios.

8.2.3 TDOA-Based Locators

With this third localization strategy, a two-step procedure is adopted. Time delay estimation (TDE) of the speech signals relative to pairs of spatially separated microphones is performed. This data along with knowledge of the microphone positions are then used to generate hyperbolic curves which are then intersected in some optimal sense to arrive at a source location estimate. A number of variations on this principle have been developed, [22–28] are examples. They differ considerably in the method of derivation, the extent of their applicability (2-D vs. 3-D, near source vs. distant source, etc.), and their means of solution. Primarily because of their computational practicality and reasonable performance under amicable conditions, the bulk of passive talker localization systems in use today are TDOA-based.

Accurate and robust TDE is the key to the effectiveness of localizers within this genre. The two major sources of signal degradation which complicate this estimation problem are background noise and channel multi-path due to room reverberations. The noise-alone case has been addressed at length and is well understood. Assuming uncorrelated, stationary Gaussian signal and noise sources with known statistics and no multi-path, the ML time-delay estimate is derived from a SNR-weighted version of the Generalized Cross-Correlation (GCC) function [29]. An ML-type weighting appropriate for non-stationary speech sources was presented in [30] and applied successfully to

speech source localization in low-multipath environments [31]. However, once room reverberations rise above minimal levels, these methods begin to exhibit dramatic performance degradations and become unreliable [32,33]. A basic approach to dealing with multi-path channel distortions in this context has been to make the GCC function more robust by deemphasizing the frequency-dependent weightings. The Phase Transform (PHAT) [29] is one extreme of this procedure which has received considerable attention recently as the basis of speech source localization systems [34–36]. By placing equal emphasis on each component of the cross-spectrum phase, the resulting peak in the GCC-PHAT function corresponds to the dominant delay in the reverberated signal. While effective at reducing some of the degradations due to multi-path, the Phase Transform accentuates components of the spectrum with poor SNR and has the potential to provide poor results, particularly under low reverberation, high noise conditions.

Other approaches for TDE of talkers in adverse environments are available. A procedure which utilizes a speech specific criterion in the design of the GCC weighting function is presented in [37]. Cepstral prefiltering [38] has been used to deconvolve the effects of reverberation prior to applying GCC. However, deconvolution requires long data segments since the duration of a typical small-room impulse response is 200-400 ms. It is also very sensitive to the high variability and non-stationarity of speech signals. In fact, the experiments performed in [38] avoided the use of speech as input altogether. Instead, colored Gaussian noise was used as the source signal. While identification of room impulse responses is extremely problematic when the source signal is unknown, the method proposed in [24], which is based on eigenvalue decomposition, efficiently detects the direct paths of the two impulse responses. This method is effective with speech as input, but requires 250 ms of microphone data to converge. A short-time TDE method, which is more complex than GCC, is presented in [33]. It involves the minimization of a weighted least-squares function of the phase data. It was shown to outperform both GCC-ML and GCC-PHAT in reverberant conditions. However, this improvement comes at the cost of a complicated searching algorithm. The marginal improvement over GCC-PHAT may not justify this added cost in computational complexity. Reverberation effects can also be overcome to some degree by classifying TDE's acquired over time and associating them with the direction of arrival (DOA) of the sound waves [39]. This approach, however, is not suitable for short-time TDE. Under extreme acoustic conditions, a large percentage of the TDE's are anomalous, and it takes a considerable period (1-2 s in [39]) to acquire enough estimates for a statistically meaningful classification.

Among the methods summarized above, those that rely on long data segments generally outperform those that do not. This result may be attributed to the ensemble averaging performed under these conditions to improve the quality of the underlying signal statistics. However, the dynamic environ-

ments of many speech array applications require high update rates, which limit the duration of the data segments used for analysis. For example, the automatic camera steering video-conferencing system detailed in [34] utilizes a TDOA-based method with GCC-PHAT TDE applied at update rates of 200-300 ms. With such long data segments, reliable estimates are produced, even in moderately adverse acoustic conditions. However, applications such as adaptive beamforming and the tracking of multiple talkers using a TDOA-based localizer require an appreciably higher estimate rate; source positions must be acquired from independent data segments as short as 20-30 ms. Over such limited durations, the lack of ensemble averaging has a severe impact on the performance of the TDE.

Given a set of TDOA figures with known error statistics, the second step of obtaining the ML location estimate necessitates solving a set of nonlinear equations. The calculation of this result is considerably less computationally expensive than that required for estimators belonging to the two previously discussed genres. There is an extensive class of sub-optimal, closed-form location estimators. designed to approximate the exact solution to the nonlinear problem. These techniques are computationally undemanding and, in many cases, suffer little detriment in performance relative to their more compute-intensive counterparts. [22,25–28,40,41] are typical of these methods. Regardless of the solution method employed, this third class of location estimation techniques possesses a significant computational advantage over the steered-beamformer or high-resolution spectral-estimation based approaches.

TDOA-based locators do present several disadvantages when used as the basis of a general localization scheme. Their primary limitation is the inability to accommodate multi-source scenarios. These algorithms assume a single-source model. While TDOA-based methods with short analysis intervals may be used to track several individuals in a conversational situation [31,42], the presence of multiple simultaneous talkers, excessive ambient noise, or moderate to high reverberation levels in the acoustic field typically results in poor TDOA figures and subsequently, unreliable location fixes. A TDOA-based locator operating in such an environment would require a means for evaluating the validity and accuracy of the delay and location estimates. These shortcomings may be overcome to some degree through judicious use of appropriate detection methods at each stage in the process [31].

While practical, the application of TDOA-based localization procedures is of limited utility in realistic, acoustic environments. Steered-Beamformer strategies are computationally more intensive, but tend to possess a robustness advantage and require a shorter analysis interval. The two-stage process requiring time-delay estimation prior to the actual location evaluation is suboptimal. The intermediate signal parameterization accomplished by the TDOA estimation procedure represents a significant data reduction at the expense of a decrease in theoretical localization performance. However, in

real situations the performance advantage inherent in the optimal steered-beamformer estimator is lessened because of incomplete knowledge of the signal and noise spectral content as well as unrealistic stationarity assumptions.

With these relative advantages and shortcomings in mind, a new localization method, which combines the best features of the steered-beamformer with those of the Phase Transform weighting of the GCC, was introduced in [5]. The goal was to exploit the inherent robustness and short-time analysis characteristics of the steered response power approach with the insensitivity to signal conditions afforded by the Phase Transform. This new algorithm, termed SRP-PHAT, will be detailed in the following section and will be shown to produce highly reliable location estimates in rooms with reverberation times up to 200 ms, using independent 25 ms data segments.

8.3 A Robust Localization Algorithm

Before describing the SRP-PHAT algorithm, it will be necessary to develop further a number of topics addressed in the prior section. Specifically, the following subsections will provide details of the impulse response model, the GCC and its PHAT implementation, ML TDOA-based localization, and the computation of the SRP. These items will then be tied together in the final subsection to motivate and define the SRP-PHAT algorithm.

8.3.1 The Impulse Response Model

It will be assumed that sound waves propagate as predicted by the linear wave equation [43]. With this assumption, the acoustic paths between sound sources and microphones can be modeled as linear systems [44]. This is clearly advantageous to the analysis and modeling of the signals produced by the microphones of an array. Such linear models are valid under the realistic conditions encountered in small-room speech-array environments and are regularly exploited by array-processing techniques [13].

In the presence of sound-reflecting surfaces, the sound waves produced by a single source propagate along multiple acoustic paths. This gives rise to the familiar effects of reverberation; sounds reflect off objects and produce echoes. The walls of most rooms are reflective enough to create significant reverberation. While it is not always noticeable to the occupants, even mild reverberation can severely impact the performance of speech-array systems. Hence, multi-path propagation must be incorporated into the signal-processing model.

The wave field at a particular location inside a reverberant room may be considered to be linearly related to the source signal, $s(t)$. Let the 3-element vectors, $\boldsymbol{p}_n$ and $\boldsymbol{q}_s$, define the Cartesian coordinates of the nth microphone

and the source, respectively. The received signal at the nth microphone may now be expressed as

$$x_n(t) = s(t) \star h_n(\boldsymbol{q}_s, t) + v_n(t) \tag{8.1}$$

The overall impulse response, $h_n(\boldsymbol{q}_s, t)$, is the result of cascading two filters: the room impulse response and the microphone channel response. The former characterizes all acoustic paths between the source and microphone locations, including the direct path. It is a function of $\boldsymbol{p}_n$ as well as the source location, $\boldsymbol{q}_s$, and is highly dependent on these parameters. In general, the room impulse response is affected by environmental conditions, such as temperature and humidity. It also varies with the movement of furniture and individuals inside the room. While such variations are significant, it is reasonable to assume that these factors remain constant over short periods. Hence, a room impulse response may be considered time-invariant for short periods when the source and microphone are spatially fixed. The microphone channel response accounts for the electrical, mechanical and acoustical properties of the microphone system. In general, the microphone's directivity pattern makes its response a function of its orientation as well as its spatial placement relative to the source. The additive term, $v_n(t)$, is the result of channel noise in the microphone system and any propagating ambient noise such as that due to fans or other mechanical equipment. The propagating noise is usually more significant than the channel noise and tends to dominate this term. Generally, $v_n(t)$ is assumed to be uncorrelated with $s(t)$.

Figure 8.1 illustrates a close-up view of the response that was measured in a typical conference room. The direct-path component and some of the strong reflected components are highlighted in this plot. The peaks corresponding to the reflected sound waves are comparable in size to the direct-path peak. These peaks, which occur within 20 ms of the direct-path, are responsible for many of the erroneous results produced by short-time TDE's, which operate on data blocks as small as 25 ms. The large secondary peaks in the room response are highly correlated with the false peaks in the GCC function [5].

The purpose of TDE is to evaluate the temporal disparity between the direct-path components in the two received microphone signals. To this end, it will be useful to rewrite the impulse response specifically in terms of its direct-path component. Equation 8.1 is modified to:

$$x_n(t) = \frac{1}{r_n} s(t - \tau_n) \star g_n(\boldsymbol{q}_s, t) + v_n(t) \tag{8.2}$$

where r_n is the source-microphone separation distance, τ_n is the direct path time delay, and $g_n(\boldsymbol{q}_s, t)$ is the modified impulse response which encompasses the original response minus the direct path component. The microphone signal model is now expressed explicitly in terms of the parameter of interest, namely the time delay, τ_n.

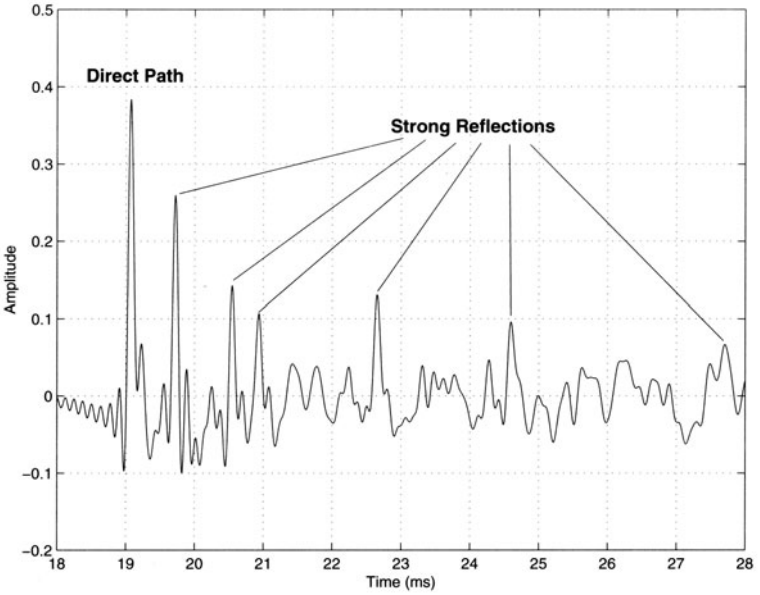


Fig. 8.1. A close-up of a 10-millisecond segment of a room impulse response measured in a typical conference room. The direct-path component and some strong reflected components are highlighted.

8.3.2 The GCC and PHAT Weighting Function

For a pair of microphones, $n = 1, 2$, their associated TDOA, τ_{12}, is defined as

$$\tau_{12} \equiv \tau_2 - \tau_1. \tag{8.3}$$

Applying this definition to their associated received microphone signal models yields

$$\begin{aligned} x_1(t) &= \frac{1}{r_1} s(t - \tau_1) \star g_1(\boldsymbol{q}_s, t) + v_1(t) \\ x_2(t) &= \frac{1}{r_2} s(t - \tau_1 - \tau_{12}) \star g_2(\boldsymbol{q}_s, t) + v_2(t). \end{aligned} \tag{8.4}$$

If the modified impulse responses for the microphone pair are similar, then (8.4) shows that a scaled version of $s(t - \tau_1)$ is present in the signal from microphone 1 and a time-shifted (and scaled) version of $s(t - \tau_1)$ is present in the signal from microphone 2. The cross-correlation of the two signals should show a peak at the time lag where the shifted versions of $s(t)$ align, corresponding to the TDOA, τ_{12}. The cross correlation of signals and is defined as:

$$c_{12}(\tau) = \int_{-\infty}^{+\infty} x_1(t) x_2(t + \tau) dt \tag{8.5}$$

The GCC function, $R_{12}(\tau)$, is defined as the cross correlation of two filtered versions of $x_1(t)$ and $x_2(t)$ [29]. With the Fourier transforms of these filters denoted by $G_1(\omega)$ and $G_2(\omega)$, respectively, the GCC function can be expressed in terms of the Fourier transforms of the microphone signals

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\big(G_1(\omega)X_1(\omega)\big)\big(G_2(\omega)X_2(\omega)\big)^* e^{j\omega\tau}d\omega \tag{8.6}$$

Rearranging the order of the signals and filters and defining the frequency dependent weighting function, $\Psi_{12} \equiv G_1(\omega)G_2(\omega)^*$, the GCC function can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\Psi_{12}(\omega)X_1(\omega)X_2(\omega)^* e^{j\omega\tau}d\omega \tag{8.7}$$

Ideally, $R_{12}(\tau)$ will exhibit an explicit global maximum at the lag value which corresponds to the relative delay. The TDOA estimate is calculated from

$$\hat{\tau}_{12} = \underset{\tau\in D}{\operatorname{argmax}}\, R_{12}(\tau). \tag{8.8}$$

The range of potential TDOA values is restricted to a finite interval, D, which is determined by the physical separation between the microphones. In general, $R_{12}(\tau)$ will have multiple local maxima which may obscure the true TDOA peak and subsequently, produce an incorrect estimate. The amplitudes and corresponding time lags of these erroneous maxima depend on a number of factors, typically ambient noise levels and reverberation conditions.

The goal of the weighting function, Ψ_{12}, is to emphasize the GCC value at the true TDOA value over the undesired local extrema. A number of such functions have been investigated. As previously stated, for realsitic acoustical conditions the PHAT weighting [29] defined by

$$\Psi_{12}(\omega) \equiv \frac{1}{|X_1(\omega)X_2^*(\omega)|} \tag{8.9}$$

has been found to perform considerably better than its counterparts designed to be statistically optimal under specific non-reverberant, noise conditions. The PHAT weighting whitens the microphone signals to equally emphasize all frequencies. The utility of this strategy and its extension to steered-beamforming form the basis of the SRP-PHAT algorithm that follows.

8.3.3 ML TDOA-Based Source Localization

Consider the ith pair of microphones with spatial coordinates denoted by the 3-element vectors, $\boldsymbol{p}_{i1}$ and $\boldsymbol{p}_{i2}$, respectively. For a signal source with known

spatial location, $\boldsymbol{q}_s$, the true TDOA relative to the ith sensor pair will be denoted by $T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s)$, and is calculated from the expression

$$T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s) = \frac{|\boldsymbol{q}_s - \boldsymbol{p}_{i2}| - |\boldsymbol{q}_s - \boldsymbol{p}_{i1}|}{c} \tag{8.10}$$

where c is the speed of sound in air. The estimate of this true TDOA, the result of a TDE procedure involving the signals received at the two microphones, will be given by $\hat{\tau}_i$. In practice, the TDOA estimate is a corrupted version of the true TDOA and in general, $\hat{\tau}_i \neq T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s)$.

For a single microphone pair and its TDOA estimate, the locus of potential source locations in 3-space which satisfy (8.10) corresponds to one-half of a hyperboloid of two sheets. This hyperboloid is centered about the midpoint of the microphones and has $\boldsymbol{p}_{i2} - \boldsymbol{p}_{i1}$ as its axis of symmetry.

For sources with a large source-range to microphone-separation ratio, the hyperboloid may be well-approximated by a cone with a constant direction angle relative to the axis of symmetry. The corresponding estimated direction angle, $\hat{\theta}_i$, for the microphone pair is given by:

$$\hat{\theta}_i = \cos^{-1}\left(\frac{c \cdot \hat{\tau}_i}{|\mathbf{m}_{i1} - \mathbf{m}_{i2}|}\right) \tag{8.11}$$

In this manner each microphone pair and TDOA estimate combination may be associated with a single parameter which specifies the angle of the cone relative to the sensor pair axis. For a given source and TDOA estimate, $\hat{\theta}_i$ is referred to as the DOA relative to the ith pair of microphones.

Given a set of M TDOA estimates derived from the signals received at multiple pairs of microphones, the problem remains as how to best estimate the true source location, $\boldsymbol{q}_s$. Ideally, the estimate will be an element of the intersection of all the potential source loci. In practice, however, for more than two pairs of sensors this intersection is, in general, the empty set. This disparity is due in part to imprecision in the knowledge of system parameters (TDOA estimate and sensor location measurement errors) and in part to unrealistic modeling assumptions (point source radiator, ideal medium, ideal sensor characteristics, etc.). With no ideal solution available, the source location must be estimated as the point in space which best *fits* the sensor-TDOA data or more specifically, minimizes an error criterion that is a function of the given data and a hypothesized source location. If the time-delay estimates at each microphone pair are assumed to be independently corrupted by zero-mean additive white Gaussian noise of equal variance then the ML location estimate can be shown to be the position which minimizes the least squares error criterion

$$E(\boldsymbol{q}) = \sum_{i=1}^{M} (\hat{\tau}_i - T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}))^2. \tag{8.12}$$

The location estimate is then found from

$$\hat{\boldsymbol{q}}_s = \underset{\boldsymbol{q}}{\operatorname{argmin}}\, E(\boldsymbol{q}). \tag{8.13}$$

The criterion in (8.12) will be referred to as the LS-TDOA error. As stated earlier, the evaluation of $\hat{\boldsymbol{q}}_s$ in this manner involves the optimization of a non-linear function and necessitates the use of search methods. Closed-form approximations to this method were given earlier.

8.3.4 SRP-Based Source Localization

The microphone signal model in (8.2) shows that for an array of N microphones in the reception region of a source, a delayed, filtered, and noise corrupted version of the source signal, $s(t)$, is present in each of the received microphone signals. The delay-and-sum beamformer time aligns and sums together the $x_n(t)$, in an effort to preserve unmodified the signal from a given spatial location while attenuating to some degree the noise and convolutional components. It is defined as simply as

$$y(t, \boldsymbol{q}_s) = \sum_{n=1}^{N} x_n(t + \Delta_n) \tag{8.14}$$

where Δ_n are the *steering delays* appropriate for focusing the array to the source spatial location, $\boldsymbol{q}_s$, and compensating for the direct path propagation delay associated with the desired signal at each microphone. In practice, the delays relative to a reference microphone are used instead of the absolute delays. This makes all shifting operations causal, which is a requirement of any practical system, and implies that $y(t, \boldsymbol{q}_s)$ will contain an overall delayed version of the desired signal which in practice is not detrimental. The use of a single reference microphone means that the steering delays may be determined directly from the TDOA's (estimated or theoretical) between each microphone and the reference. This implies that knowledge of the TDOA's alone is sufficient for steering the beamformer without an explicit source location.

In the most ideal case with no additive noise and channel effects, the output of the deal-and-sum beamformer represents a scaled and potentially delayed version of the desired signal. For the limited case of additive, uncorrelated, and uniform variance noise and equal source-microphone distances this simple beamformer is optimal. These are certainly very restrictive conditions. In practice, convolutional channel effects are nontrivial and the additive noise is more complicated. The degree to which these noise and reverberation components of the microphone signals are suppressed by the delay-and-sum beamformer is frequently minimal and difficult to analyze. Other methods have been developed to extend the delay-and-sum concept to the more general filter-and-sum approach, which applies adaptive filtering to the microphone

signals before they are time-aligned and summed. Again, these methods tend to not be robust to non-theoretical conditions, particularly with regard to the channel effects.

The output of an N-element, filter-and-sum beamformer can be defined in the frequency domain as

$$Y(\omega, \boldsymbol{q}) = \sum_{n=1}^{N} G_n(\omega) X_n(\omega) e^{j\omega \Delta_n} \tag{8.15}$$

where $X_n(\omega)$ and $G_n(\omega)$ are the Fourier Transforms of the nth microphone signal and its associated filter, respectively. The microphone signals are phase-aligned by the steering delays appropriate for the source location, $\boldsymbol{q}$. This is equivalent to the time-domain beamformer version. The addition of microphone and frequency-dependent filtering allows for some means to compensate for the environmental and channel effects. Choosing the appropriate filters depends on a number of factors, including the nature of the source signal and the type of noise and reverberations present. As will be seen, the strategy used by the PHAT of weighing each frequency component equally will prove advantageous for practical situations where the ideal filters are unobtainable.

The beamformer may be used as a means for source localization by steering the array to specific spatial points of interest in some fashion and evaluating the output signal, typically its power. When the focus corresponds to the location of the sound source, the SRP should reach a global maximum. In practice, peaks are produced at a number of incorrect locations as well. These may be due to strong reflective sources or merely a byproduct of the array geometry and signal conditions. In some cases, these extraneous maxima in the SRP space may obscure the true location and in any case, complicate the search for the global peak. The SRP for a potential source location can be expressed as the output power of a filter-and-sum beamformer by

$$P(\boldsymbol{q}) = \int_{-\infty}^{+\infty} |Y(\omega)|^2 d\omega \tag{8.16}$$

and location estimate is found from

$$\hat{\boldsymbol{q}}_s = \underset{\boldsymbol{q}}{\operatorname{argmax}}\, P(\boldsymbol{q}). \tag{8.17}$$

8.3.5 The SRP-PHAT Algorithm

Given this background, the SRP-PHAT algorithm may now be defined. With respect to GCC-based TDE, the PHAT weighting has been found to provide an enhanced robustness in low to moderate reverberation conditions. While improving the quality of the underlying delay estimates, it is still not sufficient to render TDOA-based localization effective under more adverse conditions.

The delay-and-sum SRP approach requires shorter analysis intervals and exhibits an elevated insensitivity to environmental conditions, though again, not to a degree that allows for their use under excessive multi-path. The filter-and-sum version of the SRP adds flexibility but the design of the filters is typically geared towards optimizing SNR in noise-only conditions and is excessively dependent on knowledge of the signal and channel content. Originally introduced in [5], the goal of the SRP-PHAT algorithm is to combine the advantages of the steered beamformer for source localization with the signal and condition independent robustness offered by the PHAT weighting.

The SRP of the filter-and-sum beamformer can be expressed as

$$P(\boldsymbol{q}) = \sum_{l=1}^{N}\sum_{k=1}^{N}\int_{-\infty}^{\infty}\Psi_{lk}(\omega)X_l(\omega)X_k^*(\omega)e^{j\omega(\Delta_k-\Delta_l)}d\omega \tag{8.18}$$

where $\Psi_{lk}(\omega) = G_l(\omega)G_k^*(\omega)$ is analogous to the two-channel GCC weighting term in (8.7). The corresponding multi-channel version of the PHAT weighting is given by

$$\Psi_{lk}(\omega) = \frac{1}{|X_l(\omega)X_k^*(\omega)|} \tag{8.19}$$

which in the context of the filter-and-sum beamformer (8.15) is equivalent to the use of the individual channel filters

$$G_n(\omega) = \frac{1}{|X_n(\omega)|}. \tag{8.20}$$

These are the desired SRP-PHAT filters. They may be implemented from the frequency-domain expression above. Alternatively, it may be shown that (8.18) is equivalent to the sum of the GCC's of all possible N-choose-2 microphone pairings. This means that the SRP of a 2-element array is equivalent to the GCC of those two microphones. Hence, as the number of microphones is increased, SRP naturally extends the GCC method from a pairwise to a multi-microphone technique. Denoting $R_{lk}(\tau)$ as the PHAT-weighted GCC of the lth and k^{th} microphone signals, a time-domain version of SRP-PHAT functional can now be expressed as

$$P(\boldsymbol{q}) = 2\pi\sum_{l=1}^{N}\sum_{k=1}^{N}R_{lk}(\Delta_k - \Delta_l). \tag{8.21}$$

This is the sum of all possible pairwise GCC permutations which are time-shifted by the differences in the steering delays. Included in this summation is the sum of the N autocorrelations, which is the GCC evaluated at a lag of zero. These terms contribute only a DC offset to the steered response power since they are independent of the steering delays.

Given either method of computation, SRP-PHAT localization is performed in a manner similar to the standard SRP-based approaches. Namely,

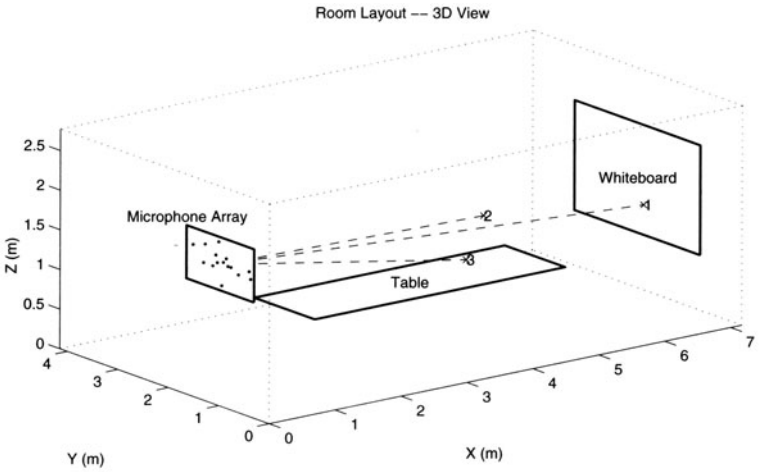


Fig. 8.2. Conference room layout.

$P(\boldsymbol{q})$ is maximized over a region of potential source locations. As will be shown in the next section, relative to the search space indicative of the standard SRP approach, the SRP-PHAT functional significantly deemphasizes extraneous peaks and dramatically sharpens the resolution of the true peak. These desirable features result in a decreased sensitivity to noise and reverberations and more precise location estimates than the existing localization methods offer. Additionally, this is achieved using a very short analysis interval.

8.4 Experimental Comparison

While more extensive results are available in [5], an experiment is offered here to evaluate and compare the relative characteristics and performance of three different source locators: SRP, SRP-PHAT and ML-TDOA. Five second recordings were made for three source locations in a 7 by 4 by 3 m conference room at Brown University using a 15-element microphone array. Figure 8.2 illustrates the room layout. Pre-recorded speech, which was acquired using a close-talking microphone, was played through a loudspeaker while simultaneously recording the signals from the array. The use of the loudspeaker was preferable to an actual talker since the loudspeaker could be precisely located and would be fixed over the duration of the recordings. The talkers were males uttering a unique string of alpha-digits. Source 1 was most distant from the array and was positioned at standing height in front of a white-board. The other two sources were positioned at a seated level

around a conference table, which was located approximately in the center of the room.

The microphone array was composed of eight omni-directional electret condenser microphones, which were randomly distributed on a plane within a .33 by 0.36 m rectangle. The microphones were attached to a rectangular sheet of acoustic foam, which was supported by an aluminum frame. This frame was mounted on a tripod that was placed parallel to the back wall at a distance of 0.9 m. The acoustic foam damps some of the multi-path reflections from this wall and isolates the microphones from vibrations traveling along the mountings.

The loudspeaker faced the array and the volume level was adjusted at each location to maximize SNR conditions. SNR levels at each microphone averaged about 25 dB for the three source locations. Source 3, with its location the closest to the microphone array, had SNRs as high as 36 dB. With such high SNRs, all microphones signals in the conference room dataset have minimal contributions from the background noise, which was primarily produced by the fans inside the computer equipment.

The measured reverberation time of the room was determined to be 200 ms. This qualifies as a mildly reverberant room. However, the near-end peaks in the impulse responses (as in Figure 8.1) combined with a 200 ms reverberation time do, in fact, have a significant impact on localization. This will be demonstrated by the following performance comparisons.

Given the size of the array aperture relative to the source ranges, all three talkers can be considered to lie in the far field of array. Under such conditions, range estimates are ambiguous, and only the azimuth and elevation angles can be estimated reliably. Accordingly, this experiment will focus on DOA measures as opposed to 3-D Cartesian coordinates. Results obtained with more extensive arrays and near-field sources are available in [5].

The recorded data was segmented into 25 ms frames using a half-overlapping Hanning window. SNR-based speech detection was performed for each frame. All frames where any of the eight microphone channels had SNR within 12dB of the background noise were eliminated. Out of the 399 frames per recording, 313, 340, and 297 were retained for sources 1,2, and 3, respectively. The DOA's of the sources were estimated by minimization of the LS-TDOA error and maximization of SRP and SRP-PHAT evaluated over azimuth and elevation relative to the array's origin. The frequency range used to compute both the steered responses and the GCC's was 300 Hz to 8 kHz. These functions were computed over a range of $-60°$ to $+60°$ for both azimuth and elevation with a $0.1°$ resolution.

By taking all possible combinations, 28 microphone pairs were formed using the 8-element array. Hence, for each data frame, 28 TDOA estimates were made for each of the three speech recordings using GCC-PHAT. Figure 8.3 illustrates the LS-TDOA error as a function of azimuth and elevation for a segment of nine successive frames recorded for source 1. The white point in

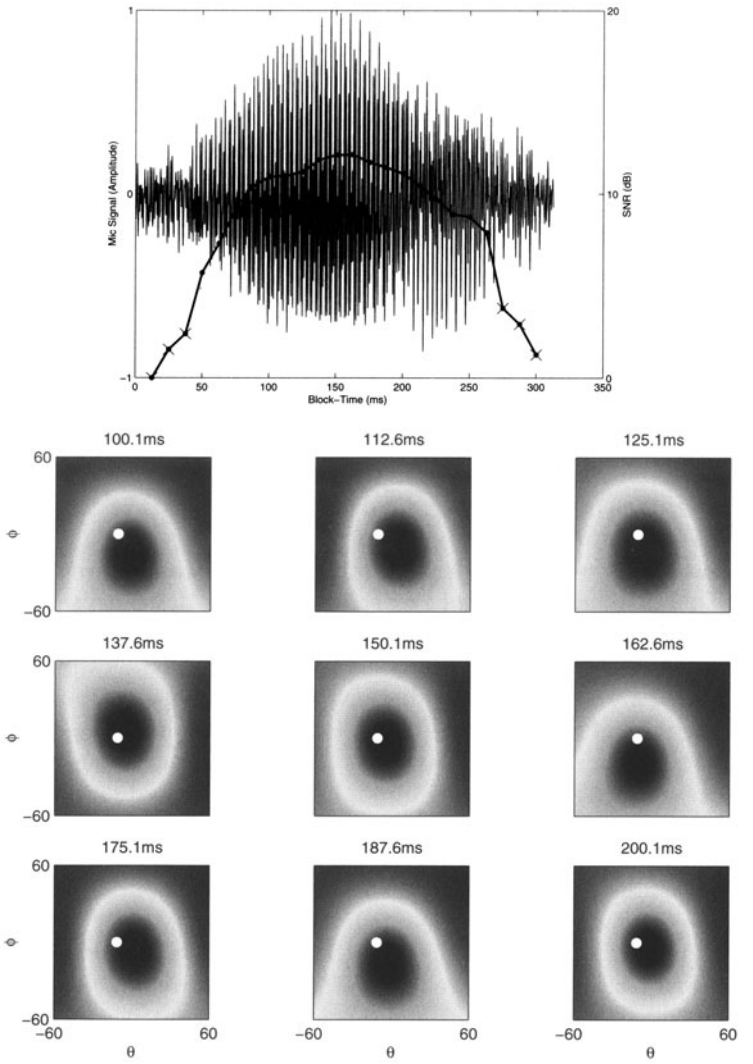


Fig. 8.3. Speech segment (top) with nine frames of the LS-TDOA error surfaces.

each contour plot marks the true DOA. The dark area in the center of the images represents the minima of the LS-TDOA error. At the top of this figure is a plot of the amplitude of the corresponding speech segment, which is the letter "R", spoken as in "Are we there yet?" Superimposed on this speech signal is a curve representing the average power of the signals from the array, with the scale of its vertical axis labeled on the right side of the graph. Each point along this power curve corresponds to the average frame SNR. The three frames at the beginning and end of this speech segment

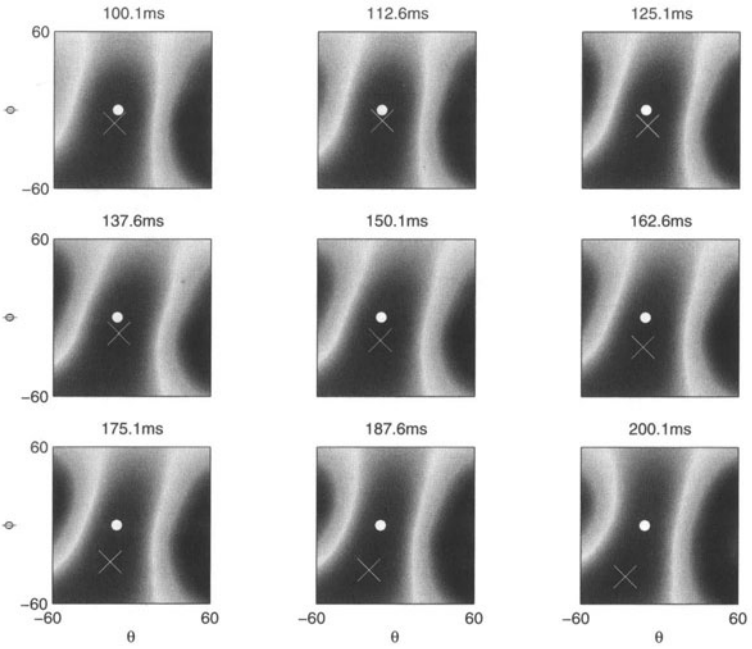


Fig. 8.4. Delay-and-sum beamformer SRP over nine, 25 ms frames.

lacked sufficient SNR to included in the analysis. These plots show that the LS-TDOA error is generally a smooth surface with a global minimum over the angular range of $\pm 60°$. However, from frame to frame the minima vary from the true source location. This inaccuracy is caused by erroneous TDOA estimates. Note also that because of the smooth nature of the error space, the resolution of the DOA estimates is considerably limited.

Figures 8.4 and 8.5 illustrate the error spaces of the SRP and SRP-PHAT as evaluated for the same nine 25 ms frames of speech. Relative to the prior figure the contour images are now inverted in darkness to emphasize the maxima. The plots of the delay-and-sum beamformer SRP in Figure 8.4 bear a noticeable similarity in general shape to their LS-TDOA counterparts. The maximum value in each SRP image, marked by an X, occurs at points distant from the actual DOA, indicated by a white dot. The main beam of the delay-and-sum beamformer is broad and fluctuates considerably over the duration of the speech segment. As a result, many inaccurate location estimates are produced by this method. In contrast to the LS-TDOA and SRP cases, the peaks of SRP-PHAT plots in Figure 8.5 match the actual DOA almost exactly. The main beam of the PHAT beamformer is sharp and consistent over each frame. This produces contour images which appear quite different from the LS-TDOA and SRP versions. The PHAT filters, when applied to the filter-and-sum beamformer, yield an error space that is superior to that of

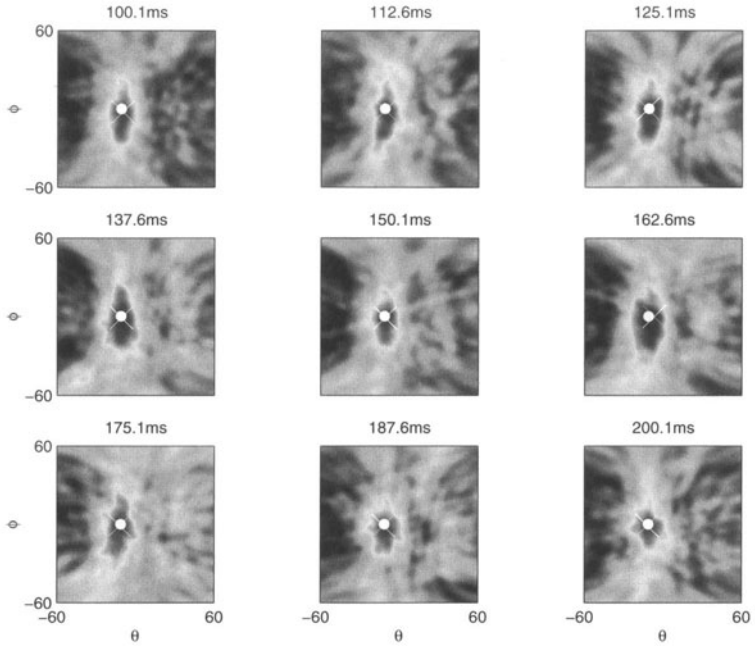


Fig. 8.5. SRP-PHAT response over nine, 25 ms frames.

the delay-and-sum beamformer or the TDOA-based criterion. This qualitative observation will now be corroborated through a numerical performance comparison.

For the DOA estimates produced for each of the three source locations, an RMS DOA error was computed from

$$E_{\text{RMS}}(\hat{\theta}, \hat{\phi}) = \sqrt{(\hat{\theta} - \theta)^2 + (\hat{\phi} - \phi)^2} \tag{8.22}$$

where ϕ and θ are the true azimuth and elevation angles and $\hat{\phi}$ and $\hat{\theta}$ are their estimated counterparts. Figure 8.6 illustrates the results. These plots show the fraction of DOA estimates in each case which exceed a given RMS error threshold. Using this metric, the SRP-PHAT consistently outperforms the other two methods for each of the source locations. The ML-TDOA exhibits definite advantages over the SRP. While the SRP-PHAT's results are nearly identical for all the source locations, including the most distant source 1, the ML-TDOA locator is highly dependent on source location. For example, 60% percent of the estimates from source 1 had error greater than 10° while 50% percent from source 2 and 15% percent from source 3 had error greater 10°. In contrast, nearly all the estimates produced by SRP-PHAT had error less than 10°. About 90% of the estimates from sources 2 and 3, and 80% from source 1 had errors less than 4°.

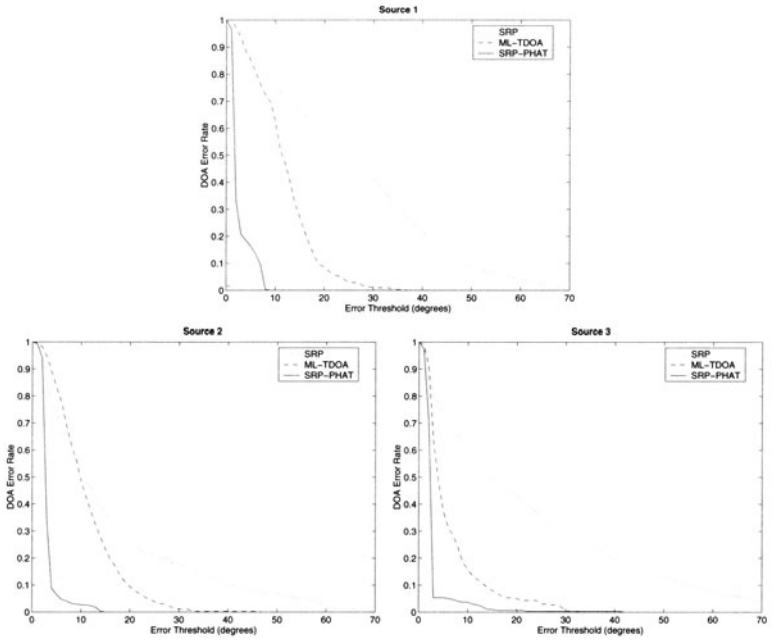


Fig. 8.6. Localizer DOA error rates for three different sources.

The results of this limited experiment illustrate the performance advantages of the SRP-PHAT localizer relative to more traditional approaches for talker localzation with microphone arrays. Other experiments conducted under more general and adverse conditions are consistent with the results here and serve to confirm the utility of combining steered-beamforming and a uniform-magnitude spectral weighting for this purpose.

While the TDOA-based localization method performed satisfactorily for a talker relatively close to the array, it was severely impacted by even the mild reverberation levels encountered when the source was more distant. This result is due to the fact that signal-to-reverberation ratios decrease with increasing source-to-microphone distance. As the reverberation component of the received signal increases relative to the direct path component, the validity of the single-source model inherent in the TDE development is no longer valid. As a result TDOA-based schemes rapidly exhibit poor performance as the talker moves away from the microphones. The SRP-PHAT algorithm is relatively insensitive to this effect. As the results here suggest the proposed algorithm exhibits no marked performance degradation from the near to distant source conditions tested.

The SRP-PHAT algorithm is computationally more demanding than the TDOA-based localization methods. However, its significantly superior performance may easily warrant the additional processing expense. Additionally,

while not discussed here, it is possible to alter the algorithm to dramatically reduce its computational load while maintaining much of its benefit.

References

1. H. Silverman, W. Patterson, J. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 251–254, April 1997.
2. M. Brandstein, J. Adcock, and H. Silverman, "Microphone array localization error estimation with application to sensor placement," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3807–3816, 1996.
3. J. Flanagan and H. Silverman, eds., *International Workshop on Microphone-Array Systems: Theory and Practice*, Brown University, Providence RI, USA, October 1992.
4. B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 881–884, March 1999.
5. J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*, PhD thesis, Brown University, Providence RI, USA, May 2000.
6. W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," in *Signal Processing* (J. Griffiths, P. Stocklin, and C. V. Schooneveld, eds.), pp. 577–590, Academic Press, 1973.
7. G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Am.*, vol. .62, pp. 922–926, October 1977.
8. W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform Theory*, vol. IT-19, pp. 608–614, September 1973.
9. W. Hahn, "Optimum signal processing for passive sonar range and bearing estimation," *J. Acoust. Soc. Am.*, vol. 58, pp. 201–207, July 1975.
10. M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 1210–1217, October 1983.
11. V. M. Alvarado, *Talker Localization and Optimal Placement of Microphones for a Linear MIcrophone Array using Stochastic Region Contraction.* PhD thesis, Brown University, Providence RI, USA, May 1990.
12. H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone-array data," *Computer, Speech, and Language*, vol. 6, pp. 129–152, April 1992.
13. D. Johnson and D. Dudgeon, *Array Signal Processing- Concepts and Techniques*, Prentice Hall, 1993.
14. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, second ed., 1991.
15. R. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*, PhD thesis, Stanford University, Stanford CA, USA, 1981.
16. J. Krolik, "Focussed wide-band array processing for spatial spectral estimation," in *Advances in Spectrum Analysis and Array Processing* (S. Haykin, ed.), vol. 2, pp. 221–261, Prentice Hall, 1991.

17. H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 823–831, August 1985.
18. K. Buckley and L. Griffiths, "Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. .36, pp. 953–964, July 1988.
19. A. Vural, "Effects of pertubations on the performance of optimum/adaptive arrays," *IEEE Trans. Aerosp. Electron.*, vol. AES-15, pp. 76–87, January 1979.
20. R. Compton Jr., *Adaptive Antennas*, Prentice Hall, 1988.
21. T. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation in coherent signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 806–811, August 1985.
22. M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 45–50, January 1997.
23. P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 231–234, April 1997.
24. Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 937–940, March 1999.
25. R. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. Aerosp. Electron.*, vol. AES-8, pp. 821–835, November 1972.
26. J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1661–1669, December 1987.
27. H. Lee, "A novel procedure for accessing the accuracy of hyperbolic multilateration systems," *IEEE Trans. Aerosp. Electron.*, vol. AES-11, pp. 2–15, January 1975.
28. N. Marchand, "Error distributions of best estimate of position from multiple time difference hyperbolic networks," *IEEE Trans. Aerosp. Navigat. Electron.*, vol. .11, pp. 96–100, June 1964.
29. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-24, pp. 320–327, August 1976.
30. M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer, Speech, and Language*, vol. 9, pp. 153–169, April 1995.
31. M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, April 1997.
32. S. Bédard, B. Champagne, and A. Stéphenne, "Effects of room reverberation on time-delay estimation performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-94)*, Adelaide, Australia, pp. II:261–264, April 1994.
33. M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 375–378, April 1997.

34. H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 187–190, April 1997.
35. M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event localization," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 288–292, May 1997.
36. P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 231–234, April 1997.
37. M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.
38. A. Stéphenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 3055–3058, May 1995.
39. N. Strobel and R. Rabenstein, "Classification of time delay estimates in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 3081–3084, March 1999.
40. B. Friedlander, "A passive localization algorithm and its accuracy analysis," *IEEE Jour. Oceanic Engineering*, vol. OE-12, pp. 234–245, January 1987.
41. Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Processing*, vol. 42, pp. 1905–1915, August 1994.
42. D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 371–374, April 1997.
43. L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *Fundamentals of Acoustics*, John Wiley & Sons, third ed., 1982.
44. L. Ziomek, *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*, CRC Press, 1995.

9 Multi-Source Localization Strategies

Elio D. Di Claudio and Raffaele Parisi

INFOCOM Dept., University of Rome "La Sapienza", Rome, Italy

Abstract. Localization of multiple acoustic sources is an important task in many practical applications. However, in most cases adopted models are not fully adequate to describe real scenarios. In particular, in the presence of reverberation, the signal model should explicitly take into account both signals radiated by multiple sources and reflections. In alternative to Generalized Cross-Correlation methods, array processing concepts can be effectively applied to multi-source localization in reverberant environments. In this chapter, main features and limitations of wideband array processing approaches are briefly reviewed. Furthermore, a new integrated localization and classification system, based on a robust frequency-domain Time Delay Of Arrival (TDOA) estimation followed by a spatial clustering of raw location estimates, is presented. The proposed method efficiently incorporates TDOA and array processing concepts in an unified approach. Results on simulated data are supplied.

9.1 Introduction

Many practical applications rely on microphone arrays to localize multiple acoustic sources. Direct examples include automated systems for surveillance, videoconferencing and hands-free speech acquisition [1]. Localization data is also used indirectly for a variety of other tasks: dereverberation of speech, fault prediction and analysis in machinery, cuing and tracking of television cameras, speaker verification, etc.

Ideally, localization methods should exploit knowledge of the particular acoustical transfer functions associated with the sources and sensors. However, over reliance on these functions in the form of excessive calibration to a specific environment may result in limited system flexibility and poor adaptation capabilities to different and varying acoustical conditions. This suggests that localization algorithms should be based on a compact and robust parameterization of the transfer functions.

If reverberation is present, a general signal model should explicitly take into account both the signals radiated by multiple sources and their respective multi-path reflections. The effects of reflective surfaces in closed environments can be studied in a rigorous manner by solving the wave equations with proper boundary and initial conditions. This task is generally too expensive and can be very sensitive to numerical errors and changes in assumptions about the environment. A physically acceptable approximation may be derived from

the empirical observation that the major contributions of reflections are produced by large surfaces. This is equivalent to neglecting edge diffractions and diffusion effects and allows for the use of ray tracing methods. A popular example of this approach is Allen and Berkeley's Image Method [2] which introduces virtual sources to simulate multi-path reflections.

Array processing concepts have been effectively applied to the multi-source localization task in the presence of reverberation. Sensor arrays provide a spatial and temporal sampling of the wave-field and have the ability to resolve multiple *point* sources (e.g. defined by a *finite* set of spatial parameters) even when the radiated signals overlap in time and frequency [3]. Of course, simpler approaches may be adopted if source signals are not simultaneously active or if they can be isolated by filtering. These cases will not be considered in this work.

From a statistical point of view, the array model is well-defined only in the context of narrowband signals, e.g. when the signal bandwidth is a small fraction (less than about 1%) of the central frequency of the sensor pass-band [4]. In particular, methods based on the *signal* and *noise subspaces* have become very popular in narrowband array processing. These methods are essentially based on exploitation of the properties of the Cross-Sensor (spatial) Covariance Matrix (CSCM) of array signal envelopes. The signal subspace idea was studied in depth in [5,6] and led to a family of well-known algorithms for narrowband direction finding, namely MUSIC [4], ESPRIT [7], MIN-NORM [8] and WSF [9] that are statistically consistent within the standard array model.

An alternative to signal subspace methods is the classical Maximum Likelihood (ML) method, which requires an exact statistical characterization of received signals [10]. Since the statistics of acoustic signals significantly depart from the standard i.i.d. Gaussian assumption [11], classical ML estimators can hardly be claimed as optimal for this application. Furthermore, ML (and also WSF [12]) requires a cumbersome multi-dimensional search through the entire array manifold to provide reliable Direction of Arrival (DOA) estimates. Moreover, the effectiveness of the method is limited by the quality of the initial DOA estimates (within a fraction of the array beam-width from the true values) and the required knowledge of the correct number of sources.

In contrast to ML-based methods, subspace algorithms have the intrinsic advantage of being widely distribution-independent, since they rely only on the array signal model and the availability of a consistent estimate of the CSCM itself. Furthermore, many subspace algorithms guarantee a fast global numerical convergence without the need for any prior information about the source locations [13].

The main limitations of subspace-based approaches are twofold. First, they are unable to deal with a quantity of uncorrelated arrivals which exceeds the number of available sensors. Such a condition is quite common in reverberant environments due to the fact that reflected signals are frequently difficult to distinguish from additional sources. Second, while the underlying

array equations have a strong dependence on the narrowband signal assumption, the acoustic sources in question usually have a large bandwidth. Subspace and ML approaches have been extended to wideband signals. In these cases the array model is altered significantly within the sensor pass-band and the sources no longer possess a rank-one signature in the CSCM [14].

Several *frequency-domain* approaches to wideband source localization have been proposed. The incoherent approach splits the array pass-band into several frequency bins, applies a narrowband algorithm to each sub-band and clusters the resulting DOA estimates. This solution is not effective at exploiting the full time-bandwidth product and leads to a higher Signal-to-Noise Ratio (SNR) threshold for the resolution of closely spaced sources [15].

The Coherent Signal Subspace Method (CSSM), proposed by Wang and Kaveh in [15] and further developed in [16–19], estimates spatial covariances at multiple frequency bins, aligns them by proper transformation matrices (called *focusing*) and generates a reduced statistic through averaging which (approximately) shares the same eigenstructure as the narrowband CSCM. However, even though CSSM-based methods do overcome some of the drawbacks of subspace techniques, their use in acoustical applications is hampered by the limited precision of the focusing procedure when the array pass-band exceeds about half an octave. Moreover, CSSM cannot deal with the case of more sources than sensors.

As an alternative to these frequency-domain approaches, *time-domain* processing is appealing since the differential arrival delays among sensors typically exceed the sampling period in a wideband environment and the number of relevant reflections is essentially limited by the frame length [20]. Most time-domain approaches to source localization involve the use of time-difference of arrival (TDOA) estimates between pairs (or *doublets*) of co-located microphones [20–23]. Traditionally, the resulting localization procedure requires joint parameter optimization from signals collected simultaneously by many sensors [24].

TDOA estimation is usually performed by Generalized Cross-Correlation (GCC) methods [22]. GCC is appealing for its simplicity and ease of implementation. However, it assumes a single-source model which limits its utility to the multiple-source, reverberant environment problem considered here. It is desirable to identify *all* direct and reflected paths and to correctly associate each with the appropriate source. The localization problem is frequently further complicated by the absence of *a priori* knowledge about the room geometry and the number of active sources. Moreover, individual signals may be received with sufficient power by only a subset of the sensor doublets. In such a case, the localization algorithm should be robust to missing TDOA's.

In the following, after a short review of wideband array processing techniques, an integrated localization and classification system will be introduced. The algorithm involves robust frequency-domain TDOA estimation followed by *spatial* clustering of raw location estimates [25]. The proposed method

efficiently incorporates TDOA and array processing concepts in an unified approach for the localization of multiple sources in reverberant environments.

9.2 Background

In this section the classical concepts of wideband array processing in the frequency domain are briefly summarized with particular reference to the CSSM and WSF approaches.

9.2.1 Array Signal Model

A common signal model in wideband array processing is based on the sub-band decomposition of the received signals. Suppose that signals radiated by D point sources impinge on an N-element array. Each sensor signal is demodulated, digitized and decomposed by a filterbank (or a DFT) [26] into J complex sub-band components. The bandwidth of each filter is much smaller than its central frequency f_i $(i = 1, 2, ..., J)$, so that each sub-band snapshot vector $\boldsymbol{x}_i[k]$ approximately satisfies the classical narrowband discrete-time equations for $(k = 0, 1, ..., K-1)$ [4]:

$$\boldsymbol{x}_i[k] = \mathbf{A}(f_i)\boldsymbol{s}_i[k] + \boldsymbol{v}_i[k] \,, \tag{9.1}$$

$$\mathbf{A}(f_i) = \left[\boldsymbol{a}_1(f_i, \boldsymbol{\theta}_1) \cdots \boldsymbol{a}_D(f_i, \boldsymbol{\theta}_D)\right] \,. \tag{9.2}$$

$\mathbf{A}(f_i)$ is the $N \times D$ array transfer matrix at frequency f_i, $\boldsymbol{s}_i[k]$ is the vector of the complex source signal envelopes and $\boldsymbol{v}_i[k]$ is the $N \times 1$ vector of additive noise components at frequency f_i and time k. Each column of $\mathbf{A}(f_i)$ is referred to as *steering vector* and represents the array response to an impinging wavefront. The generic steering vector, $\boldsymbol{a}(f, \boldsymbol{\theta})$, is normalized and is a known, continuous and differentiable function of the frequency, f, and of the location parameters contained in vector $\boldsymbol{\theta}$ (azimuth, elevation, range,...). In order to guarantee the non-ambiguity of the model, $\mathbf{A}(f_i)$ must have full rank for any set of $D < N$ source locations [4].

Noise and source waveforms are assumed to be realizations of independent, zero mean, complex, circular random processes, having at least finite second- and fourth-order moments [6]. The background noise is considered temporally white and isotropic so that its CSCM at any frequency is proportional to the identity matrix, $\mathbf{I}$, through its *a priori* unknown variance, σ_v. Colored noise with a known spatio-temporal covariance may be accommodated for through a pre-whitening procedure [4,6].

If the sub-band sampling period is sufficiently large and filter pass-bands do not overlap, snapshots $\boldsymbol{x}_i[k]$ may be considered independent with respect to both time and frequency indices. Under these standard hypotheses, the CSCM at frequency f_i, $\mathbf{R}_{xx}(f_i)$, obeys to the following relationship:

$$\mathbf{R}_{xx}(f_i) = E[\boldsymbol{x}_i[k]\boldsymbol{x}_i^H[k]] = \mathbf{A}(f_i)\mathbf{R}_{ss}(f_i)\mathbf{A}^H(f_i) + \sigma_v\mathbf{I} \,, \tag{9.3}$$

where $E[.]$ indicates the expectation operator and $(.)^H$ is the Hermitian transpose operator. The subspace spanned by the steering vectors contains the span of the eigenvectors corresponding to the $\eta_i \leq D$ largest eigenvalues $\{\mu_1(f_i) \geq \mu_2(f_i) \geq \cdots \geq \mu_{\eta_i}(f_i)\}$ of the CSCM. These eigenvectors are collected in a $N \times \eta_i$ matrix $\mathbf{E}_s(f_i)$ and constitute a basis for the *signal subspace* [4] at frequency f_i. If the wavefronts are not completely coherent, the source covariance matrix, $\mathbf{R}_{ss}(f_i)$, has rank D and the range of $\mathbf{A}(f_i)$ coincides with the signal subspace [4].

The orthogonal complement to the signal subspace is the *noise subspace*, $\mathbf{E}_v(f_i)$, and is constructed from the $N - \eta_i$ eigenvectors associated with the smallest eigenvalues of $\mathbf{R}_{xx}(f_i)$.

9.2.2 Incoherent Approach

The incoherent approach may be considered as a direct derivation of narrowband methods [14,15]. It consists of parallel DOA estimation performed independently on several narrow sub-bands. Estimates obtained in the single sub-bands are clustered in angle to obtain the number of sources (i.e. the number of relevant clusters) and the final DOA (the centroid coordinates).

The main limitations of this method are the high dispersion of DOA estimates and a threshold effect of the CSCM estimator at low SNR [27]. These factors necessitate the use of a high number of snapshots to get reliable location estimates. Such a long-term ensemble analysis may not be valid given the time-varying statistical properties of acoustic sources.

The use of incoherent methods may be advantageous if processing speed is a strict requirement and signals have a high SNR in some frequency bins. This is the case of speech signals, where formant frequencies may be easily identified through a number of methods, e.g. Linear Prediction [14,28]. However, more effective wideband strategies seek to combine the information obtained by sub-band processing to generate a reduced statistic which approximates the narrowband CSCM model. Referred to as *coherent methods*, a few of the more popular approaches along these lines are briefly described below.

9.2.3 Coherent Signal Subspace Method (CSSM)

The CSSM [15] uses $N \times N$ transformation matrices $\mathbf{T}(f_i)$ to map each steering vector $\boldsymbol{a}(f_i, \theta)$ to the corresponding steering vector $\boldsymbol{b}(f_0, \theta)$ of a reference array operating at the focusing frequency f_0. Perfect focusing is obtained when:

$$\mathbf{T}(f_i)\mathbf{A}(f_i) = [\boldsymbol{b}(f_0, \theta_1) \cdots \boldsymbol{b}(f_0, \theta_D)] = \mathbf{B}(f_0) \; . \tag{9.4}$$

The Universal Spatial Covariance Matrix (USCM), $\mathbf{R}_{xx}$, is found from the average of the focused $\mathbf{R}_{xx}(f_i)$ with preselected scalar positive weights

α_i:

$$\mathbf{R}_{xx} = \sum_{i=1}^{J} \alpha_i \mathbf{T}(f_i)\mathbf{R}_{xx}(f_i)\mathbf{T}^H(f_i) \cong \mathbf{B}(f_0)\mathbf{R}_{ss}\mathbf{B}^H(f_0) + \sigma_v \mathbf{R}_{vv} \,. \quad (9.5)$$

The *Universal Signal Subspace*, $\mathbf{E}_s$, and its orthogonal complement, $\mathbf{E}_v$, the *Universal Noise Subspace* are computed from the eigendecomposition of the matrix $\mathbf{V}^{-1}\mathbf{R}_{xx}(\mathbf{V}^{-1})^H$, where $\mathbf{V}$ is the *matrix square root* [29] of the Universal Noise Spatial Covariance, $\mathbf{R}_{vv}$, found according to [19]:

$$\mathbf{R}_{vv} = \mathbf{V}\mathbf{V}^H = \sum_{i=1}^{J} \alpha_i \mathbf{T}(f_i)\mathbf{T}^H(f_i) \,. \quad (9.6)$$

The matrix $\mathbf{R}_{xx}$ must be estimated from $K \gg N$ independent snapshots for each frequency, yielding the sample subspaces[1] $\hat{\mathbf{E}}_s$ and $\hat{\mathbf{E}}_v$ that may be used in any subspace-based algorithm for direction finding. For instance, the classical MUSIC estimate [4] becomes:

$$\hat{\boldsymbol{\theta}}_{MUSIC} = \underset{\boldsymbol{\theta}}{argmin} \left\{ \frac{|\hat{\mathbf{E}}_v^H \mathbf{V}^{-1} \boldsymbol{b}(f_0, \boldsymbol{\theta})|_2}{|\mathbf{V}^{-1} \boldsymbol{b}(f_0, \boldsymbol{\theta})|_2} \right\} \,. \quad (9.7)$$

In room acoustics applications the CSSM suffers from several drawbacks. Reflections are seen as a large number of scarcely correlated sources. Since the number of resolvable sources is bounded by the number of available microphones, the resulting USCM is characterized by a colored background noise which is present only when the sources are active. This fact induces a theoretical loss of resolution and in practice, a strong bias in the DOA estimates as well as errors in the identification of the number of sources.

9.2.4 Wideband Weighted Subspace Fitting (WB-WSF)

The classical CSSM spatially transforms signals and noise. In contrast, the WB-WSF algorithm [12] uses only signal eigenvectors to get asymptotically efficient DOA estimates, $\hat{\boldsymbol{\theta}}_{WB-WSF}$, under a Gaussian assumption for both signals and noise.

The narrowband WSF method attempts to fit in a weighted Least Squares fashion a linear combination of calibrated steering vectors onto the signal subspace eigenvectors. Extension of the WSF criterion to the wideband case is straightforward. If the snapshots can be considered independent among different frequency bins [17], the global and concentrated Gaussian Log-Likelihood

[1] Throughout this work, estimated quantities will be marked by the hat superscript ($\hat{}$), to avoid ambiguities.

of the data is simply the sum of its narrowband components. The resulting WB-WSF estimator then assumes the following form:

$$\hat{\boldsymbol{\theta}}_{WB-WSF} = \underset{\boldsymbol{\theta}}{argmin} \left\{ \frac{1}{J} \sum_{i=1}^{J} |\mathbf{A}(f_i, \boldsymbol{\theta}) \hat{\mathbf{C}}_i - \hat{\mathbf{E}}_s(f_i) \mathbf{W}(f_i)|_F^2 \right\} , \tag{9.8}$$

where the subscript F stands for the Frobenius norm of a matrix.

In the case of Gaussian-distributed signals and noise (and perfectly calibrated arrays) optimal weighting matrices $\mathbf{W}(f_i)$ are diagonal with weights proportional to [9]:

$$\mathbf{W}(f_i)_{[m,m]} = \frac{\mu_m(f_i) - \sigma_v}{\sqrt{\mu_m(f_i)\, \sigma_v}} . \tag{9.9}$$

The matrices $\hat{\mathbf{C}}_i (D \times \eta_i)$ in functional (9.8) are linked to selected weighting matrices through the equation $\hat{\mathbf{C}}_i = \mathbf{A}^{\dagger}(f_i, \boldsymbol{\theta}) \hat{\mathbf{E}}_s(f_i) \mathbf{W}(f_i)$ [9]. Optimality is assured even if consistent estimates of the weights are used instead of their true values.

From (9.8) it may be seen that the WB-WSF method requires the solution of an expensive multi-dimensional optimization problem in the location parameters space. Moreover, in most cases this is an ill-conditioned problem since the error surface is quite flat far from the true DOA's and characterized by many local minima [12]. This property limits the convergence rate of *any* optimization algorithm, making the WB-WSF unsuitable for real-time applications unless an accurate initialization is provided (e.g. tracking of a moving source previously localized by a different algorithm).

Another issue with the WB-WSF approach (and similar covariance fittings) is its very high sensitivity to the appropriateness of the propagation model. Errors in the steering vectors of high SNR components may be greatly amplified by weighting, inducing a loss of convergence or an increased variance of the DOA estimates [30]. This fact makes the use of WB-WSF extremely difficult in reverberant and time-varying environments.

9.3 The Issue of Coherent Multipath in Array Processing

From narrowband array processing it is well known that specular multipath creates several identification problems in subspace methods [4] since the rank of $\mathbf{R}_{ss}(f_i)$ is less than D. In such a situation the signal subspace is strictly contained in the span of $\mathbf{A}(f_i)$ and most subspace algorithms do not guarantee source identification. This problem is particularly serious with incoherent processing, as described in the following example.

Example. Consider for simplicity a single source and a single reflection and suppose that the propagation difference between the direct and the reflected

paths is 3 m. The TDOA is about 10 ms. If a formant frequency of 300 Hz is considered, the sub-band filter should have a pass-band of about 4 Hz. This means that the sub-band Nyquist frequency allows for discrimination of coherent path delays down to about 250 ms using the eigen-analysis of the CSCM (9.3). Below this limit, the MUSIC algorithm may not resolve the two paths and a WSF-type approach is required [9]. However, the initialization will be very difficult [12] and the final estimate very sensitive to calibration errors. The CSSM would be capable of focusing an instantaneous bandwidth to about 150 Hz (0.73 octaves), allowing for the discrimination of a TDOA around 6 ms. By contrast, a time-domain approach, such as GCC, allows for significantly improved resolution. With an 8 KHz sampling frequency, it would be possible to detect a TDOA less than about 250 μs.

This example suggests that time-domain are advantageous in room acoustics applications. However, in the presence of multiple sources and reflections, the problem arises of discriminating and correctly pairing the TDOA's. This task is quite complex if spatial information is not exploited.

9.4 Implementation Issues

Microphone arrays for surveillance or multi-media applications should be relatively cheap and easy to deploy, but capable of functioning in real time. In particular, they should be largely insensitive to environmental changes and calibration should be as simple and automated as possible. These requirements tend to rule out approaches based on joint optimization procedures working with all signals at the same time. In fact, since the microphones may be deployed relatively distant from each other in an effort to achieve a sufficient spatial aperture, synchronization, calibration and simultaneous processing of all sensors may be very expensive and technically difficult. This requires high-speed links and enormous processing power which typically grows with the *square* of the array size in the case of aforementioned signal subspace and ML algorithms.

In the following section, a possible approach to the robust localization of multiple simultaneous speakers in reverberant rooms is described. The proposed method circumvents most of the drawbacks of existing array processing algorithms in terms of the number of sources identifiable, robustness to calibration and estimation errors, and capability of discriminating among direct and reflected paths. Moreover, the algorithm may be computed in parallel on cheap, but powerful, DSP processing boards connected to selected pairs of microphones. This property reduces the physical length of the inter-processor communication paths and simplifies hardware manufacturing, system deployment, and software development. Only a limited number of parameters are collected at *slow data rates* by the main processing unit (e.g. a workstation), which fuses the partial information and furnishes the final interface to end-users (humans or other sub-systems).

9.5 Linear Prediction-ROOT-MUSIC TDOA Estimation

The proposed solution to the multi-source localization problem is based on a frequency-domain time delay estimation procedure which employs a *disturbed harmonics* model and involves a three-stage strategy. The first stage consists of time-varying data pre-whitening through the use of Linear Prediction [28]. In the second stage, the TDOA's for the direct paths and early (strongest) reflections are estimated by a closed-form parametric approach based on the ROOT-MUSIC algorithm [13]. ROOT-MUSIC is applied to a matrix built with samples of the Cross Power Spectrum (CPS) [22] of the pre-processed sensor signals. The Singular Value Decomposition (SVD) of this matrix yields information about the number of sources and main reflections, and their corresponding TDOA's. Each set of TDOA's is then converted into a set of points in 3D-space by classical geometrical methods. Finally, in the third stage the most likely positions of the speakers are found by means of a clustering in space performed among all the estimated locations. The centroids of the most dense clusters are selected as candidate sources, thus eliminating most of the false detections generated by outliers (virtual sources, localization ambiguities, impulsive noise, etc.). In the following sections, these three steps are separately described.

9.5.1 Signal Pre-Whitening

Like most TDOA estimators in the presence of reverberation, the proposed algorithm separately processes signals received by pairs (or *doublets*) of microphones. The underlying hypothesis is that since the microphones in each doublet are physically close to each other, the two room impulse responses will be very similar. Under such circumstances it may be assumed that the two microphones receive time-shifted replicas of each source signal filtered by the same acoustic transfer function.

Since the power spectrum of acoustic signals is typically characterized by a wide dynamic range across frequency bands, signal pre-whitening has been found to be an effective tool for improving the quality of speech enhancement schemes [31] and TDOA estimators in reverberant environments [22,20,23]. Figure 9.1 shows a schematic of the pre-whitening process [32] by Linear Prediction. The short-term prediction filter $A_f(z)$ acts on formant frequencies while the long-term predictor $A_p(z)$ cancels spectral periodicities induced by the pitch and reflections. The presence of multiple sources (speakers and virtual sources) may require filter orders that are slightly higher than those normally used for speech coding applications. Moreover, multiple cascaded long-term predictors may be required.

Linear Prediction removes the common spectral features present in doublet signals (including the pitch) and minimizes spectral fluctuations induced by multi-path effects. These considerations may be extended to a larger class

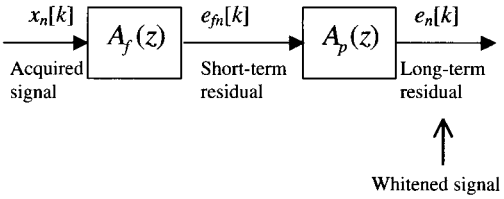


Fig. 9.1. Schematic of the Linear Prediction processing.

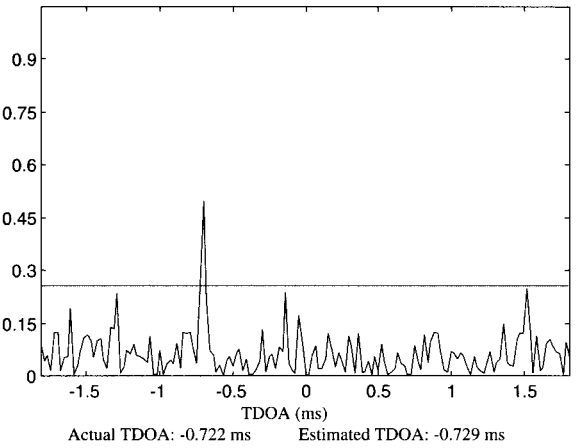


Fig. 9.2. Cross-Power-Spectrum Phase TDOA estimate without pre-whitening.

of acoustic sources besides speech, like air conditioners, computer fans, and rotating machinery in factories. It is important to remark that this procedure is a *linear* process and does not alter either the general signal model or the local signal-to-noise ratio (SNR) in the frequency domain. The pre-whitening produces an approximate concentration of the Likelihood function (under a simplifying Gaussian assumption) [20] and improves the Signal-to-Reverberation ratio, thus reducing the number of TDOA's to be estimated. Figures 9.2 and 9.3 show the improvements obtained when pre-whitening is included in the standard Cross-Power Spectrum Phase (CPSP) TDOA estimator [1].

Optimal processing would require separate filtering of each signal and compensation for the different phase responses before the CPSP estimation. This leads to increased computational cost and higher variance of the TDOA estimates for short data records [22]. Conversely, in the proposed approach short- and long-term predictors are computed on the basis of the *average* autocorrelation of doublet signals. Resulting filters are smoother in frequency and lead to a negligible loss in performance.

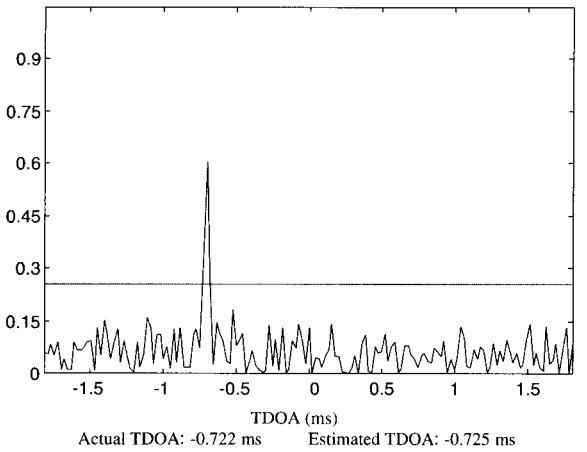


Fig. 9.3. Cross-Power-Spectrum Phase TDOA estimate with pre-whitening.

Among Linear Prediction algorithms, those assuring a smooth and statistically stable estimate should be preferred. In particular, Levinson and Schur-type algorithms (such as Yule-Walker, Burg, Makhoul) compute the minimal set of parameters (reflection coefficients) of a Toeplitz correlation matrix, typical of stationary signals [33]. The Toeplitz constraint, coupled with a *biased* correlation estimate, avoids the presence of deep nulls in the prediction filters which may cancel the speech formants and ensures the stability of the AR model assumed by Linear Prediction [33].

9.5.2 An Approximate Model for Multiple Sources in Reverberant Environments

The proposed solution is an effective alternative to existing Generalized Cross-Correlation methods. GCC is not able to cope with multiple sources that are simultaneously active. Array processing concepts may be effectively introduced for this task. The TDOA search may be recast as a *disturbed harmonics retrieval* problem [34,35] from the CPS, $P_{12}(f)$, of a generic doublet.

Indicating by $e_1[k]$ and $e_2[k]$ the sampled signals acquired from the doublet (after Linear Prediction whitening) and with $\mathcal{E}_1(f)$ and $\mathcal{E}_2(f)$ as their corresponding Fourier transforms, $P_{12}(f)$ is defined by:

$$P_{12}(f) = E\left[\mathcal{E}_1(f)\mathcal{E}_2^*(f)\right] . \tag{9.10}$$

Under mild hypotheses, it may be shown that the following holds for $P_{12}(f)$:

$$P_{12}(f) = \sum_{d=1}^{D} S_d(f) \left(\sum_{k=1}^{K_d} \sum_{l=1}^{K_d} \alpha_{1dk}(f)\alpha_{2dl}^*(f) e^{-j2\pi f(\tau_{1dk}-\tau_{2dl})} \right) , \tag{9.11}$$

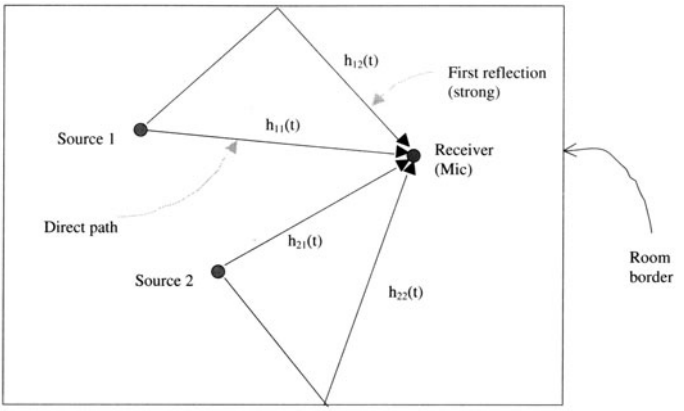


Fig. 9.4. Model of a reverberant room.

where D is the number of sources, $S_d(f)$ is the spectrum of the d-th whitened signal, K_d is the number of significant arrivals from a single source (direct path *and* reflections), $\alpha(f)$ are slowly-varying transfer functions due to mismatches between the microphone impulse responses, and $\Delta_{dkl} = (\tau_{1dk} - \tau_{2dl})$ are the TDOA's to be estimated (i.e. the time differences of all possible combinations in pairs of direct paths and main reflections, as illustrated in Figure 9.4). The model in (9.11) assumes that $P_{12}(f)$ is the sum of a few sinusoids modulated by slowly-varying envelopes and embedded in a nearly white noise [34].

It is important to note that in case of multiple sources the Linear Prediction-based whitening filter has a square frequency response which is roughly proportional to the reciprocal of the sum of the signal spectra. As a consequence, if the spectrum of the d-th source exhibits a peak at some frequency, the whitened spectrum $S_d(f)$ will have a broader and lower peak while the other sources will experience a strong fading at the same frequency. The corresponding sinusoids in (9.11) will present localized envelope fluctuations that can be regarded as noise. However, since the power spectra are positive, it can be verified that the CPS, $P_{12}(f)$, of sampled signals has a first-order periodicity [35] at each TDOA that can be detected through autocorrelation analysis [36].

9.5.3 Robust TDOA Estimation via ROOT-MUSIC

The cross power spectrum (9.11) is evaluated at discrete equi-spaced frequencies, f_i, from consecutive frames of Linear Prediction residuals by means of an unwindowed FFT and time averaging. Many algorithms for the estimation of the delays Δ_{dkl} are available in literature. Among the possible solutions,

techniques based on eigenanalysis are preferred for the detection of disturbed and ensemble harmonics [36] like those present in (9.11). In particular, an *ad hoc* version of the ROOT-MUSIC algorithm [13] known to be very robust to envelope fluctuations has been developed.

Specifically, ROOT-MUSIC is applied to the covariance of a Hankel data matrix, formed by the estimates of $P_{12}(f_i)$ at each doublet [34]. Let:

$$\boldsymbol{e}_n[m] = \left[e_n[(m-1)L], \ldots, e_n[mL]\right] , \tag{9.12}$$

be the m-th frame of Linear Prediction residuals at the n-th sensor. L is the frame length. The running CPS estimate at frequency f_i is computed by exponential windowing with a forgetting factor γ $(0 < \gamma < 1)$ according to:

$$\hat{P}_{12}[m,i] = (1-\gamma)\, \hat{P}_{12}[m-1,i] + \gamma\, \mathcal{E}_1[m,i]\, \mathcal{E}_2^*[m,i] , \tag{9.13}$$

where $\mathcal{E}_n[m,i]$ is the i-th sample of the FFT of $\boldsymbol{e}_n[m]$.

The estimates in (9.13) are used to build a CPS matrix $\mathbf{P}[m]$, which has the following structure:

$$\begin{aligned}\mathbf{P}[m] &= \begin{pmatrix} \hat{P}_{12}[m,i_1] & \cdots & \hat{P}_{12}[m,i_1+P-1] \\ \hat{P}_{12}[m,i_1+1] & \cdots & \hat{P}_{12}[m,i_1+P] \\ \vdots & \ddots & \vdots \\ \hat{P}_{12}[m,i_2-P+1] & \cdots & \hat{P}_{12}[m,i_2] \end{pmatrix} = \\ &= \begin{pmatrix} \hat{\boldsymbol{r}}_1[m] \\ \hat{\boldsymbol{r}}_2[m] \\ \vdots \\ \hat{\boldsymbol{r}}_{i_2-i_1-P+2}[m] \end{pmatrix} , \end{aligned} \tag{9.14}$$

where i_1 and i_2 are the respective indices of the lowest and highest frequency bins considered, and P is the "prediction" order.

As shown in (9.13), the matrix $\mathbf{P}[m]$ is asymptotically (i.e. when $m \to \infty$ and $\gamma \to 0$) low-rank if all the power spectra are flat [34]. However, it must be noted that spectral fluctuations generate an *irreducible* noise, independent of the observation time. This noise limits the improvement in the variance of the TDOA estimate with the SNR and the observation time. Experiments show that the variance tends to a finite limit with the increase in the SNR; this is typical of array processing in the presence of random model mismatches [30].

In addition, each CPS sample $\hat{P}_{12}[m,i]$ is affected by an *estimation* error which can be shown to be nearly uncorrelated with respect to the bin index, i, if the signals have been whitened. The variance, $\sigma_p[m,i]$, of this error is proportional to $E\left[|\mathcal{E}_1[m,i]|^2\right] \times E\left[|\mathcal{E}_2[m,i]|^2\right]$ since the FFT samples approach a Gaussian distribution [11] and can be considered constant relative to the frequency of the Linear Prediction residuals.

The Singular Value Decomposition [29] is now applied to the matrix, $\mathbf{P}[m]$, to obtain an estimate of σ_p and the numerical rank, η_p, required by

the standard ROOT-MUSIC [13]. The rows of $\mathbf{P}[m]$ may be conveniently re-weighted if the SNR in the generic FFT bin changes through subsequent frames. A possible choice for the weights, $w_l[m]$, comes from the minimization of the Mean Square Error (MSE), $J_l[m]$, between the measured $\hat{\boldsymbol{r}}_l[m]$ and the true one, $\boldsymbol{r}_l[m]$ [6]:

$$J_l[m] = E\left[|\hat{\boldsymbol{r}}_l[m]w_l[m] - \boldsymbol{r}_l[m]|^2\right] . \tag{9.15}$$

The final weights are set proportional to:

$$w_l[m] \propto \begin{cases} 1 - \dfrac{P\hat{\sigma}_p}{|\hat{\boldsymbol{r}}_l[m]|^2} & \text{if } |\hat{\boldsymbol{r}}_l[m]|^2 > P\hat{\sigma}_p \\ 0 & \text{elsewhere} \end{cases} . \tag{9.16}$$

It is convenient to normalize the weights so that $\sum w_l^2[m] = 1$.

The SVD of the re-weighted matrix $\mathbf{P}_w[m]$ is computed and the rank η_p is estimated again. The $P - \eta_p$ right singular vectors corresponding to the smallest singular values, $\{\hat{\nu}_{\eta_p+1} \geq \hat{\nu}_{\eta_p+2} \geq \ldots \geq \hat{\nu}_P\}$, define the noise subspace $\hat{\mathbf{E}}_p$ of the matrix $\mathbf{P}_w[m]$.

At this point, the sinusoidal frequencies in (9.11) are estimated from the zeros of the following rational function [13]:

$$\mathrm{H}(z) = \boldsymbol{d}^T(z^{-1})\, \hat{\mathbf{E}}_p \begin{pmatrix} \hat{\nu}_{\eta_p+1}^{-2} & 0 & \cdots \\ 0 & \ddots & 0 \\ \vdots & 0 & \hat{\nu}_P^{-2} \end{pmatrix} \hat{\mathbf{E}}_p^H\, \boldsymbol{d}(z) , \tag{9.17}$$

where $\boldsymbol{d}(z) = \left[\, 1 \; z \; z^2 \; \ldots \; z^{P-1} \,\right]^T$. The weighting of the noise subspace $\hat{\mathbf{E}}_p$ in (9.17) helps to compensate for deviations of the noise covariance matrix from the identity matrix [5]. Finally, a Least Squares fitting of the sample CPS using (9.11) is recommended to validate the detected sources according to their relative power [4].

9.5.4 Estimation of the Number of Relevant Reflections

Before the ROOT-MUSIC algorithm may be applied, it is first necessary to estimate the number of significant sinusoids in (9.11). The effective rank of $\mathbf{P}[m]$ may be calculated and used for this task since it is linked through (9.11) to the number of relevant arrivals.

Rank estimation in array processing is often performed by setting a proper threshold on the eigenvalues of the CSCM. A number of possible criteria are available for this purpose [37]. In the present approach, the rank η_p of the matrix $\mathbf{P}[m]$ (or $\mathbf{P}_w[m]$) is estimated via a simple and robust algorithm based on the SVD.

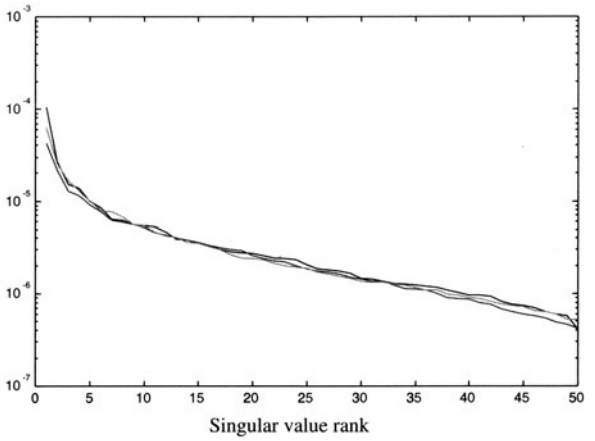


Fig. 9.5. Sample singular value spectrum of $\mathbf{P}[m]$ in three different realizations.

It is assumed that the squares of noise-induced singular values of $\mathbf{P}[m]$ (i.e. the eigenvalues of the data correlation matrix $\mathbf{P}^H[m]\mathbf{P}[m]$ [29]) are clustered around the noise power, σ_p, scaled by the number of observations and are approximately Gaussian or χ^2 distributed. In contrast, the largest (or *signal*) eigenvalues are assumed to be drawn from a different and *unknown* distribution [37]. These may be considered as *outliers* when observed on the eigenvalue spectrum.

This particular interpretation allows for the robust estimation of the noise eigenvalue distribution in order to set an appropriate threshold [38]. In this work, the sample median and the absolute median deviation of the eigenvalues have been successfully used to estimate, respectively, σ_p and the dispersion. As an example, Figure 9.5 shows the spectra of the singular values of $\mathbf{P}[m]$ for three different realizations.

The use of robust estimation requires a sufficient amount of data. For example, with median-based estimators it is known that the order, P, of the matrix $\mathbf{P}[m]$ should exceed twice the number of significant reflections [38]. Information theoretic criteria like MDL or AIC [37] do not possess this limitation, but do have the drawback of being highly sensitive to modeling errors and approximations always present when dealing with multiple reverberated signals.

9.5.5 Source Clustering

Each doublet generates a set of TDOA's. By pairing the TDOA's of nearby doublets, a set of candidate source positions may be generated by means of efficient geometric algorithms available in the literature. In particular, the Linear Intersection (LI) method [24] has been found to produce reliable

results during computer simulations. In this case each source position is given by the point of closest intersection between only *two* bearing lines. Points that lie outside of the room borders are immediately discarded.

Location estimates are clustered in space using a *fast* fuzzy algorithm [39]. The clustering procedure's processing speed is of paramount importance during real-time operations since this is the only task that cannot be parallelized onto the local DSP processors associated with each doublet.

The maximum number of estimates that can be attributed to a single source, Q_d, can be easily computed from the array geometry [24]. For example, in the case of azimuth-only (2-D) estimation, it is:

$$Q_d \propto \binom{N}{2} . \tag{9.18}$$

Most of the remaining estimates are generated by pairing TDOA's that refer to different acoustic sources and have no physical meaning. Other incorrect locations arise from geometrical ambiguities in the array itself, reflections from furniture (whose virtual sources are internal to the room), or diffraction effects.

Dense clusters possessing an approximately elliptical shape are formed around the locations of actual sources while other estimates usually generate disperse clusters containing only a few points. The centroids of clusters gathering a number of elements close to Q_d are finally selected as source locations. This clustering procedure may be extended over consecutive time frames in an effort increase performance.

9.5.6 Experimental Results

The ROOT-MUSIC TDOA algorithm was compared with the Cross-Power Spectrum Phase approach [23]). Sources were simulated as spherical clusters of radius 15 mm formed by highly coherent point sources radiating speech signals previously recorded in the near-field and open space. Reverberations were simulated using the Image Method [2] applied to a room of size ($6 \times 7 \times 4$ m) with reflection coefficients equal to 0.8 on the walls and 0.6 on the ceiling and the floor. Spatially distributed air absorption with an attenuation coefficient of 1 dB/m at 20 KHz was also employed. Microphones arranged in four square arrays of 16 elements each were selectively used. The frame length was 1024 samples with a sampling frequency of 44.1 KHz. At each step, the matrix $\mathbf{P}[m]$ was formed from a *single* 1024-points FFT. Clustering of location estimates was performed over ten consecutive frames with an overlap of 50%.

Pre-whitening was found to improve the performance of both algorithms. Table 9.1 presents the sample TDOA statistics obtained from a single doublet in the presence of a single speech source for varying SNR values. The superior performance of the proposed approach is clearly visible.

Estimator		SNR = 10 dB	SNR = 20 dB	SNR = 30 dB
CPSP	Bias:	0.0278	0.0153	0.015
	Std. dev.:	0.165	0.0264	0.025
ROOT-MUSIC	Bias:	0.0126	0.0087	0.0065
	Std. dev.:	0.0472	0.0183	0.0151

Table 9.1. TDOA estimate of CPSP and ROOT-MUSIC: sample bias and standard deviation, in milliseconds.

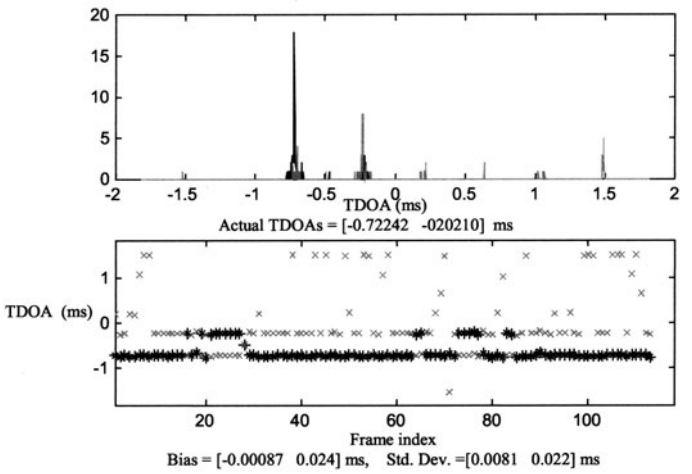


Fig. 9.6. Histogram and plot over time of TDOA estimates in a two-source case.

Figure 9.6 illustrates the properties of the ROOT-MUSIC TDOA estimate obtained from a single doublet in the presence of two simultaneous speakers. The upper plot shows the histogram of the TDOA's, corresponding to the two sources and to the main reflections. The lower plot tracks the polynomial roots of (9.17) [13]. The effects of most of the reflections are visible only in sporadic frames.

Figure 9.7 shows the standard deviations of TDOA estimates derived for various doublets and varying SNR values. It can be seen that at high SNR levels the standard deviations depart from the ideal linear behaviour expected in classical array processing techniques with perfectly calibrated arrays [9]. This effect is due to the aforementioned model approximations and, in particular, to the envelope fluctuations of the CPS.

Figure 9.8 shows an example of clustering over time in the presence of three simultaneous speakers. The plot on the left shows the scatter plots of

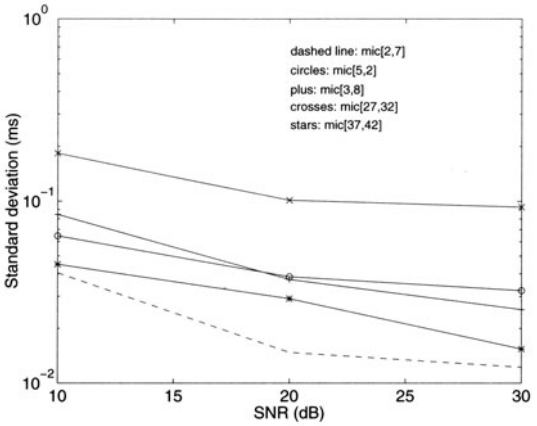


Fig. 9.7. Standard deviations of TDOA estimates vs. SNR for different pairs of microphones, in milliseconds.

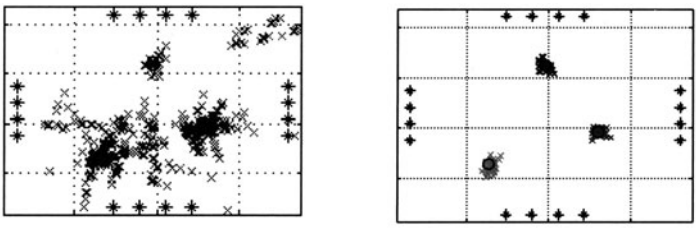


Fig. 9.8. Left: scatter plot of raw locations. Right: scatter plots of centroids.

raw locations obtained by application of the LI method [24] to the ROOT-MUSIC TDOA estimates. All locations obtained in ten consecutive frames and passed to the clustering algorithm are indicated. The actual source positions are also indicated by circles. The plot on the right depicts the scatter plots of the *centroids* of the main detected clusters associated with the sources.

Finally, Table 9.2 details the sample statistics of the centroids depicted in Figure 9.8 drawn out of 50 outcomes. In particular, the number of locations attributed to each source is shown as a fraction of the theoretical Q_d. This number can be assumed as a measure of the compactness of the clusters and of the quality of the overall algorithm. The very low bias of the estimated source positions should be noted.

References

1. M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 75–95, Aug. 1998.

Source No.	Percentage	Bias [x,y,z]	Standard deviation [x,y,z]
1	0.87	[-10.9, 17.1, 1.6]	[80.9, 63.8, 15.6]
2	0.80	[13.2, 9.3, 3.1]	[69.2, 59.1, 14.0]
3	0.60	[12.4, 8.6, 5.4]	[73.9, 48.2, 28.8]

Table 9.2. Clustering performance: percentage of location estimates inside the cluster, sample bias and standard deviation of the centroids, in millimeters (SNR=20 dB).

2. J. B. Allen, and A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
3. H. Krim, and M. Viberg, "Two decades of array signal processing research," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, July 1996.
4. R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
5. G. Bienvenu, and L. Kopp, "Optimality of high-resolution array processing using the eigensystem approach," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 31, no. 5, pp. 1235–1248, Oct. 1983.
6. B. De Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. on Signal Processing*, vol. 41, no. 9, pp. 2826–2838, Sept. 1993.
7. R. Roy, and K. Kailath, "ESPRIT - Estimation of Signal Parameter via Rotational Invariance Techniques," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.
8. R. Kumaresan, and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. on Aerosp. Electro. Syst.*, vol. AES-19, pp. 134–139, Jan. 1983.
9. M. Viberg, B. Ottersten, and T. Kailath, "Detection and estimation in sensor arrays using weighted subspace fitting," *IEEE Trans. on Signal Processing*, vol. 39, no. 11, pp. 2436–2449, Nov. 1991.
10. P. Stoica, and A. Nehorai, "Music, maximum likelihood and Cramer-Rao bound: further results and comparisons," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 12, pp. 2140–2150, Dec. 1990.
11. D. B. Williams, and D. H. Johnson, "Robust estimation of structured covariance matrices," *IEEE Trans. on Signal Processing*, vol. 41, no. 9, pp. 2891–2906, Sept. 1993.
12. J. A. Cadzow, "Multiple source location - The signal subspace approach," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 7, pp. 1110–1125, July 1990.
13. B. D. Rao, and K. V. S. Hari, "Performance analysis of Root-MUSIC," *IEEE Trans. on Signal Processing*, vol. 37, no. 12, pp. 1939-1949, Dec. 1989.
14. G. Su, and M. Morf, "Signal subspace approach for multiple wide-band emitter location," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 31, no. 12, pp. 1502–1522, Dec. 1983.

15. H. Wang, and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 4, pp. 823–831, Aug. 1985.
16. H. Hung, and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol 36, no. 8, pp. 1272–1281, Aug. 1988.
17. M. A. Doron, and A. J. Weiss, "On focusing matrices for wide-band array processing," IEEE Trans. on Signal Processing, vol. 40, no. 6, pp. 1295–1302, June 1992.
18. B. Friedlander, and A. J. Weiss, "Performance analysis of wideband direction finding using interpolated arrays," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP92*, vol. 4, pp. 457-460, 1992.
19. B. Friedlander, and A. J. Weiss, "Direction finding for wide-band signals using an interpolated array," *IEEE Trans. on Signal Processing*, vol. 41, no. 4, pp. 1618–1634, Apr. 1993.
20. G. Clifford Carter, *Coherence and Time Delay Estimation.* IEEE Press, 1993.
21. B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, Mar. 1996.
22. C. H. Knapp, and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
23. M. Omologo, and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, May 1997.
24. M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form estimator for use with room environment microphone arrays," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, Jan. 1997.
25. E. D. Di Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by ROOT-MUSIC and clustering," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, June 2000.
26. S. Haykin, ed., *Array Signal Processing*, Prentice Hall, 1984.
27. J. Krolik, and D. N. Swingler, "Multiple wide-band source location using steered covariance matrices," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1481–1494, Oct. 1989.
28. B. S. Atal, and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2, pp. 637–655, 1971.
29. G. H. Golub, and C. F. Van Loan, *Matrix computations.* Johns Hopkins University Press, 3rd edn., 1996.
30. A.L. Swindlehurst, and T. Kailath, "A performance analysis of subspace-based methods in the presence of model errors: part II- multidimensional algorithms," *IEEE Trans. on Signal Processing*, vol. 41, no. 9, pp. 2882–2890, Sept. 1993.
31. B. Yegnanarayana, C. Avendano, H. Hermansky, and P. Satyanarayana Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25–42, 1999.

32. R. P. Ramachandran, and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 4, pp. 467–478, Apr. 1989.
33. S. Lawrence Marple, *Digital spectral analysis with applications.* Prentice Hall, 1987.
34. M. A. Hasan, M. R. Azimi-Sadjadi, and G. J. Dobeck, "Separation of multiple time delays using new spectral estimation schemes," *IEEE Trans. on Signal Processing*, vol. 46, no. 6, pp. 1580–1590, June 1998.
35. W. A. Gardner, *Statistical spectral analysis, a nonprobabilistic theory*, Prentice Hall, 1988.
36. V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophys. Journal of Royal Astron. Soc.*, vol. 33, pp. 347–366, 1973.
37. M. Wax, and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp 387–392, Apr. 1985.
38. P. J. Huber, *Robust statistics.* John Wiley, 1981.
39. S. Chiu, "Fuzzy model identification based on cluster estimation," *Journ. of Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267–278, 1994.

10 Joint Audio-Video Signal Processing for Object Localization and Tracking

Norbert Strobel[1], Sascha Spors[2], and Rudolf Rabenstein[2]

[1] Siemens Medical Solutions, Erlangen, Germany
[2] University Erlangen-Nuremberg, Erlangen, Germany

Abstract. Applications such as videoconferencing, automatic scene analysis, or security surveillance involving acoustic sources can benefit from object localization within a complex scene. Many single-sensor techniques already exist for this purpose. They are, e.g., based on microphone arrays, video cameras, or range sensors. Since all of these sensors have their specific strengths and weaknesses, it is often advantageous to combine information from various sensor modalities to arrive at more robust position estimates.

This chapter presents a joint audio-video signal processing methodology for object localizing and tracking. The approach is based on a decentralized Kalman filter structure modified such that different sensor measurement models can be incorporated. Such a situation is typical for combined audio-video sensing, since different coordinate systems are usually used for the camera system and the microphone array.

At first, the decentralized estimation algorithm is presented. Then a speaker localization example is discussed. Finally, some estimation results are shown.

10.1 Introduction

Microphone arrays are becoming increasingly important for applications related to the recording of speech or music for subsequent processing, transmission or storage. In most of these cases, the goal is to extract information by listening. In object localization and tracking, however, things are different. There, the information conveyed by acoustic waves is not related to the pattern of the signal they carry but to the position of the sound emitting source.

Of course, sound waves are not the only physical quantity that can be used for source localization. Other sensors may, for example, measure the range of the source or detect its retinal image. Since all sensors have their specific strengths and weaknesses, there is no single modality for object localization that always outperforms all others. Therefore, it is desirable to integrate the information of various sensor modalities to exploit different physical phenomena.

A practical example for such a multi-sensor system may comprise a microphone array and a video camera, for example. Either sensor can be used to estimate a (single-sensor) source position. Taken together, they can be viewed as a joint audio-video object localization system. Unfortunately, such

a multi-sensor architecture is of limited use unless it is known how to combine the single-sensor position estimates. This fusion problem is addressed here after a brief overview of object localization based on audio and video sensors.

Acoustic Object Localization Using audio signals, one can estimate the object position from time-differences of arrival (TDOA's) of sound waves recorded at a microphone array. There are *direct* and *indirect* acoustic source localization methods. The direct approach is based on summing the systematically delayed microphone signals and observing the power of the overall output signal. This strategy is usually implemented using a steered filter-and-sum beamformer. Steered beamformers have been extensively studied in the past, and a significant amount of work focused, for example, on sonar target localization [1–5]. More recent applications either concentrate on acoustic applications such as improved sound reception or on speaker localization [6–9].

Indirect techniques, on the other hand, require two distinct processing steps. A set of time differences of arrival (TDOA's) of an acoustic wavefront recorded at separate microphone sensors is computed first [10]. They can even be smoothed using a Kalman filter [11]. Then geometrical properties are used to infer the source position [12,13]. Another method is to apply a non-directed gradient descent search over all possible locations to find the best match for a source location given a set of TDOA's between multiple microphones [14].

Visual Object Localization Acoustic object localization only works if sound waves can be reliably received at the microphone array. When faced with low signal-to-noise ratios or strong room reverberations, acoustic object localization methods may fail to deliver accurate source position estimates. Then it may be beneficial to use a complementary sensor system to obtain another source position estimate. Often a video camera is a practical choice for such a complementary sensor system.

To find and identify an object in a video sequence, visual object characteristics must be used. Since the visual appearance of various objects can vary widely, it becomes necessary to define a problem domain. A popular area of research is face localization and tracking [15–20].

Knowing the position of a face in a camera reference frame is only a first step. To actually estimate the position of a person's face in a world coordinate system requires a calibrated camera [21]. Such an object recognition and tracking problem is very common in three-dimensional computer vision. A geometry-based approach covering examples from recognition-based tracking and navigation can be found in [22], for example. Other application areas are recognition-based vehicle tracking [23], traffic sign localization [24], or remote video surveillance [25].

Object Localization involving Audio and Video Sensors Up to now, the single-sensor audio or video object localization problem has received significantly more research attention than joint audio-video approaches. Those have started to emerge only recently, and researchers have mainly focused on "intermediate" applications so far. Intermediate applications integrate audio and video sensors, but they do not directly fuse the single-sensor position estimates.

There are basically two ways to integrate audio and video sensors. One can either visually localize an acoustic source first and then use a beamformer to improve the sound reception [26,27]. Or one can use a microphone array first to steer a camera towards an acoustic source. Once the source appears within the camera field of view, it can then be observed with greater accuracy [28–31].

An approach to object localization based on combining audio and video position estimates can be found in [32]. Assuming a stationary speaker, it was shown how two independent source position estimates produced by an audio and a video sensor can be combined using sensor fusion techniques [33]. In more recent work, joint audio-video object localization based on recursive state estimation was investigated [34,35].

The objective of this chapter is to detail how classical tracking techniques based on the decentralized Kalman filter can be applied to joint audio-video object localization and tracking. To this end, a review of the linear and the extended Kalman filter for tracking a single maneuvering target is first presented. One may assume that all the observations (position estimates, in our case) can be fed into a single *centralized* Kalman filter. Unfortunately, such a centralized architecture cannot be used for joint audio-video object localization due to different measurement models at the audio and video object localizers. As a solution to this problem, a decentralized Kalman filter architecture is introduced. Such a structure can be obtained by decomposing the linear Kalman filter into two autonomous local processors and one global coordinator. Since the local (linear) Kalman filters operate independently, it is possible to replace one local linear Kalman filter by a local extended Kalman filter taking into account the nonlinear measurement model at the microphone array. Although the resulting parallel system is no longer theoretically optimal, it will be shown that it still offers properties that are often very desirable such as an increased robustness against sensor failure. Finally, some open problems are pointed out.

10.2 Recursive State Estimation

First, the traditional linear Kalman filter is discussed. Since it cannot be used for measurements which are nonlinearly related to the system state, the extended Kalman filter is also introduced. It is then shown how to construct a decentralized version of the linear Kalman filter. Finally, integrating the

parallel state estimates of linear and extended Kalman filters into a linear sensor fusion framework is discussed.

10.2.1 Linear Kalman Filter

The Kalman filter is based on an internal model of the system dynamics and a comparison of incoming measurements with ongoing estimates. It recursively produces estimates of the system state. Important factors such as measurement accuracy and object motion are taken into account. They affect the gains (weights) applied to the input data. If the models are accurate, the resulting state estimates are optimal in the mean square sense.

Measurement Equation The input to the Kalman filter is a n-dimensional measurement vector, $\mathbf{y}[k]$, recorded at time instant (or step) k. In the linear case, it can be modeled as

$$\mathbf{y}[k] = \mathbf{H}[k]\,\mathbf{x}[k] + \mathbf{n}[k], \tag{10.1}$$

where $\mathbf{H}[k]$ denotes the observation matrix, $\mathbf{x}[k]$ represents the system state vector at time instant k, and $\mathbf{n}[k]$ is additive, white Gaussian noise. To apply the Kalman filter to object localization and tracking, within a two-dimensional plane, a state vector based on a Cartesian coordinate system may be used (superscript T denotes transposition)

$$\mathbf{x}[k] = \left[x[k]\ \dot{x}[k]\ y[k]\ \dot{y}[k]\right]^T. \tag{10.2}$$

In (10.2), $x[k]$ and $y[k]$ denote the horizontal and vertical coordinates of the object position, while $\dot{x}[k]$ and $\dot{y}[k]$ refer to their associated velocities. Augmenting the state vector by the velocities, $\dot{x}[k]$ and $\dot{y}[k]$, may improve the performance of the state estimates provided a motion model that properly represents the dynamic behavior of the object may be found. Given only observations of the object position (no velocity data), the measurement equation becomes

$$\underbrace{\begin{bmatrix} x[k] \\ y[k] \end{bmatrix}}_{\mathbf{y}[k]} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{H}[k]} \underbrace{\begin{bmatrix} x[k] \\ \dot{x}[k] \\ y[k] \\ \dot{y}[k] \end{bmatrix}}_{\mathbf{x}[k]} + \mathbf{n}[k]. \tag{10.3}$$

For simplicity, only two spatial dimensions are considered. It is conceptually straightforward to extend the derivation to three dimensions.

State Equation The state behavior is modeled as

$$\mathbf{x}[k+1] = \mathbf{A}[k]\,\mathbf{x}[k] + \mathbf{v}[k], \tag{10.4}$$

where $\mathbf{A}[k]$ is the quadratic nonsingular state matrix. The initial mean and covariance of the state are assumed to be known.

Both measurement noise, $\mathbf{n}[k]$, and process noise, $\mathbf{v}[k]$, are assumed to be normal, zero-mean, and white. They are further required to be independent of each other and also independent of the initial state, $\mathbf{x}[0]$. The associated measurement noise and process noise covariance matrices are called $\mathbf{C_{nn}}[k]$ and $\mathbf{C_{vv}}[k]$, respectively.

Although finding a good state model for a moving object seems to be a challenging problem, a linear model often serves as a good starting point. Such a model assumes constant object speed. Starting with a motion model in continuous time, a state model for the discrete Kalman filter can be found by transforming the continuous model into discrete time as shown, e.g., in [36]. This results in

$$\underbrace{\begin{bmatrix} x[k+1] \\ \dot{x}[k+1] \\ y[k+1] \\ \dot{y}[k+1] \end{bmatrix}}_{\mathbf{x}[k+1]} = \underbrace{\left[\begin{array}{cc|cc} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{array}\right]}_{\mathbf{A}[k]} \underbrace{\begin{bmatrix} x[k] \\ \dot{x}[k] \\ y[k] \\ \dot{y}[k] \end{bmatrix}}_{\mathbf{x}[k]} + \mathbf{v}[k], \tag{10.5}$$

where T is the time between two subsequent state estimations.

Recursive Kalman Equations A Kalman filter can be characterized as a (state) model-based predictor followed by an observation-dependent corrector. Assuming the existence of a dynamics equation and of an *a posteriori* state estimate, $\hat{\mathbf{x}}[k|k]$, based on a set of observations taken up to and including time k, the *a priori* state estimate at time $k+1$ is predicted from

$$\hat{\mathbf{x}}[k+1|k] = \mathbf{A}[k]\,\hat{\mathbf{x}}[k|k]. \tag{10.6}$$

The initial state is assumed to be a Gaussian random variable with known mean and covariance matrix.

Given the previous *a priori* state estimate, $\hat{\mathbf{x}}[k|k-1]$ at time k, the *a priori* estimation error is defined as

$$\mathbf{e}[k|k-1] = \mathbf{x}[k] - \hat{\mathbf{x}}[k|k-1]. \tag{10.7}$$

Substituting (10.4) and (10.6) into (10.7), gives

$$\mathbf{e}[k|k-1] = \mathbf{A}[k-1]\left(\mathbf{x}[k-1] - \hat{\mathbf{x}}[k-1|k-1]\right) + \mathbf{v}[k]. \tag{10.8}$$

The associated error covariance matrix is defined as

$$\mathbf{P}[k|k-1] = E\{\mathbf{e}[k|k-1]\,\mathbf{e}[k|k-1]^T\}, \tag{10.9}$$

where $E\{\}$ denotes the expected value. Using (10.8), the *a priori* error covariance matrix can be written as

$$\mathbf{P}[k|k-1] = \mathbf{A}[k-1]\,\mathbf{P}[k-1|k-1]\,\mathbf{A}^T[k-1] + \mathbf{C}_{\mathbf{v}\,\mathbf{v}}[k-1], \qquad (10.10)$$

where $\mathbf{P}[k-1|k-1]$ refers to the *a posteriori* estimation error covariance matrix further explained below in (10.15). Based on $\hat{\mathbf{x}}[k|k-1]$, the associated observations can be predicted by

$$\hat{\mathbf{y}}[k|k-1] = \mathbf{H}[k]\,\hat{\mathbf{x}}[k|k-1]. \qquad (10.11)$$

The *a priori* state estimate $\hat{\mathbf{x}}[k|k-1]$ can be improved (corrected) by adding the difference between the new observation at time k, $\mathbf{y}[k]$, and its predicted value, $\hat{\mathbf{y}}[k|k-1]$. The resulting *a posteriori* state estimate then becomes

$$\hat{\mathbf{x}}[k|k] = \hat{\mathbf{x}}[k|k-1] + \mathbf{K}[k]\left(\mathbf{y}[k] - \hat{\mathbf{y}}[k|k-1]\right). \qquad (10.12)$$

The matrix $\mathbf{K}[k]$ is commonly referred to as Kalman gain, and the difference $\mathbf{y}[k] - \hat{\mathbf{y}}[k|k-1]$ is known as the innovation sequence, $\boldsymbol{\nu}[k|k-1]$. The Kalman gain is chosen such that the Kalman filter recursively computes a minimum mean-square linear state estimate [37,38]. Its covariance formulation can be expressed as

$$\mathbf{K}[k] = \mathbf{P}[k|k]\,\mathbf{H}^T[k]\left[\mathbf{H}[k]\,\mathbf{P}[k|k-1]\,\mathbf{H}^T[k] + \mathbf{C}_{\mathbf{n}\,\mathbf{n}}[k]\right]^{-1}. \qquad (10.13)$$

An alternative means of computing the Kalman gain is [38]

$$\mathbf{K}[k] = \mathbf{P}[k|k]\,\mathbf{H}^T[k]\,\mathbf{C}_{\mathbf{n}\,\mathbf{n}}^{-1}[k]. \qquad (10.14)$$

As mentioned earlier, $\mathbf{P}[k|k]$ is the *a posteriori* estimation error covariance matrix defined as

$$\mathbf{P}[k|k] = E\{\mathbf{e}[k|k]\,\mathbf{e}[k|k]^T\}, \qquad (10.15)$$

while

$$\mathbf{e}[k|k] = \mathbf{x}[k] - \hat{\mathbf{x}}[k|k] \qquad (10.16)$$

represents the *a posteriori* estimation error. The *a posteriori* estimation error covariance matrix may be computed from

$$\mathbf{P}^{-1}[k|k] = \mathbf{P}^{-1}[k|k-1] + \mathbf{H}^T[k]\,\mathbf{C}_{\mathbf{n}\,\mathbf{n}}^{-1}[k]\,\mathbf{H}[k]. \qquad (10.17)$$

From (10.14), it may be seen that the Kalman gain is small when the measurement noise is large. In this case, (10.12) shows that the new state estimate follows the predicted (*a priori*) state estimate more closely. On the other hand, the Kalman gain increases with $\mathbf{P}[k|k]$. This implies that more

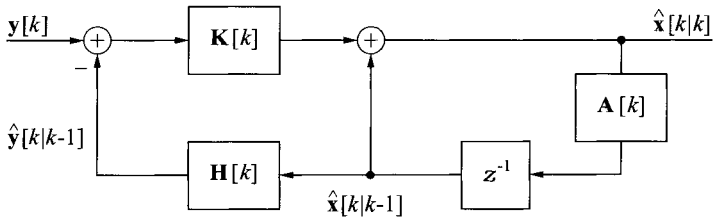


Fig. 10.1. The Kalman filter loop representing a predictor-corrector structure: the weighted difference between the current observation, $\mathbf{y}[k]$, and its predicted value, $\hat{\mathbf{y}}[k|k-1]$, is added to the (predicted) *a priori* state estimate, $\hat{\mathbf{x}}[k|k-1]$, to obtain the *a posteriori* state estimate, $\hat{\mathbf{x}}[k|k]$.

weight is given to the latest measurements when the *a posteriori* state estimation error is large, i.e. when the state cannot be accurately predicted. This usually happens when an object trajectory is undergoing rapid changes.

Equations (10.14) and (10.17) together with (10.10) and (10.12) establish the *information formulation* of the Kalman filter.

Equations (10.6), (10.10), (10.11), (10.12), (10.14), and (10.17) comprise the recursive Kalman equations. First, forward projections are made by computing the *a priori* state estimate, $\hat{\mathbf{x}}[k|k-1]$, and the associated error covariance matrix, $\mathbf{P}[k|k-1]$, using (10.6) and (10.10). Next, the *a posteriori* error covariance matrix is updated using (10.17). Then, the Kalman gain is found, for example, via (10.14). In the following step, the difference between the current measurement, $\mathbf{y}[k]$, and its predicted counterpart, $\hat{\mathbf{y}}[k|k-1]$, as shown in (10.11) is taken. The resulting innovation term is subsequently used to update the *a posteriori* estimate as shown in (10.12). This recursion is summarized in Figure 10.1. Once the loop is entered, it can be continued *ad infinitum*.

Alternative Form of the Kalman Equations An alternative equation for the *a posteriori* state estimate is now presented. It is needed to derive the decentralized Kalman filter. To this end, (10.11) is used to replace $\hat{\mathbf{y}}[k+1|k]$ in (10.12) arriving at

$$\hat{\mathbf{x}}[k|k] = \hat{\mathbf{x}}[k|k-1] + \mathbf{K}[k]\,\mathbf{y}[k] - \mathbf{K}[k]\,\mathbf{H}[k]\,\hat{\mathbf{x}}[k|k-1]. \tag{10.18}$$

Next, $\mathbf{K}[k]$ is expanded using (10.14) and replacing $\mathbf{H}^T[k]\,\mathbf{C}_{\mathbf{nn}}^{-1}[k]\,\mathbf{H}[k]$ by $\mathbf{P}^{-1}[k|k]-\mathbf{P}^{-1}[k|k-1]$ according to (10.17). Rearranging some terms, another *a posteriori* state update equation is obtained

$$\hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\big(\mathbf{P}^{-1}[k|k-1]\,\hat{\mathbf{x}}[k|k-1] + \mathbf{H}^T[k]\mathbf{C}_{\mathbf{nn}}^{-1}[k]\,\mathbf{y}[k]\big). \tag{10.19}$$

Implementation of this alternative Kalman filter loop is illustrated in Figure 10.2.

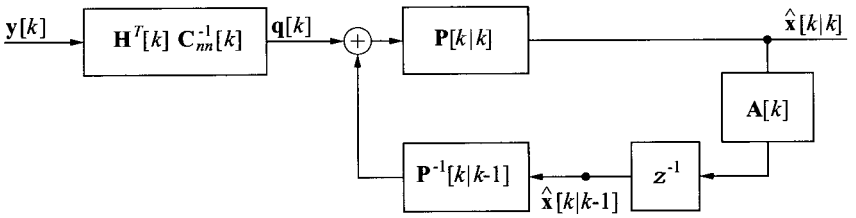


Fig. 10.2. An alternative Kalman filter loop.

For later reference, the following quantity is introduced

$$\mathbf{q}[k] = \mathbf{H}^T[k]\mathbf{C}_{\mathbf{nn}}^{-1}[k]\,\mathbf{y}[k]. \tag{10.20}$$

It may be interpreted as a weighted observation vector. Substituting (10.20) into (10.19), produces

$$\hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\big(\mathbf{P}^{-1}[k|k-1]\,\hat{\mathbf{x}}[k|k-1] + \mathbf{q}[k]\big)\ . \tag{10.21}$$

Note that there is another way of writing $\mathbf{q}[k]$. Solving (10.14) for $\mathbf{H}^T[k]\mathbf{C}_{\mathbf{nn}}^{-1}[k]$ and substituting the result into (10.20), results in

$$\mathbf{q}[k] = \mathbf{P}^{-1}[k|k]\,\mathbf{K}[k]\,\mathbf{y}[k]. \tag{10.22}$$

At this point, the Kalman filter equations have been manipulated into a form that makes it straightforward to derive the decentralized Kalman filter. Nevertheless, the recursion presented above may not be numerically optimal. Alternative forms of the Kalman filter can be found in [38,39].

10.2.2 Extended Kalman Filter due to a Measurement Nonlinearity

Despite the Kalman filter's importance, it is limited to problems in which the state and measurement equations are linear functions of the current state. However, for a microphone array yielding estimates of the azimuth, $\theta[k]$, and the range, $r[k]$, a linear Kalman filter cannot be used when the state vector is expressed in Cartesian coordinates. This is due to the measurement vector of the audio localizer represented as

$$\mathbf{y}[k] = \begin{bmatrix}\theta[k]\\ r[k]\end{bmatrix} + \mathbf{n}[k]. \tag{10.23}$$

The observation, $\mathbf{y}[k]$, is nonlinearly related to the state vector, $\mathbf{x}[k]$, as defined in (10.2), since

$$\theta[k] = \arctan\left(\frac{y[k]}{x[k]}\right), \text{ and} \tag{10.24}$$

$$r[k] = \sqrt{x^2[k] + y^2[k]}. \tag{10.25}$$

Despite the nonlinear measurement model, it is still possible to apply Kalman filtering ideas provided the measurement nonlinearities are carefully linearized. One approach is to use a (first-order) extended Kalman filter (EKF). The EKF is closely related to the linear Kalman filter. The only difference is that the nonlinear measurement equation is relinearized at each iteration.

To explain operation of the EKF, assume a linear state-space model as expressed in (10.4). Then the *a priori* state estimate at time $k+1$ can be predicted using (10.6). The associated *a priori* and *a posteriori* error covariance matrices follow from (10.10) and (10.17), respectively. As to the issue of predicting the next measurement, make the assumption that the measurement equation can be expressed as

$$\mathbf{y}[k] = \mathbf{h}(\mathbf{x}[k]) + \mathbf{n}[k], \tag{10.26}$$

where $\mathbf{h}(.)$ represents a nonlinear measurement function and $\mathbf{n}[k]$ refers to additive, white Gaussian noise with zero mean as shown in (10.23). Using a first-order extended Kalman filter, it is possible to predict the measurements via

$$\hat{\mathbf{y}}[k|k-1] = \mathbf{h}(\hat{\mathbf{x}}[k|k-1]). \tag{10.27}$$

Assuming that the measurement model $\mathbf{h}(.)$ can be linearized about the predicted state vector $\hat{\mathbf{x}}[k|k-1]$, the Jacobian (observation) matrix may be introduced

$$\hat{\mathbf{H}}[k] = \left.\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x} = \hat{\mathbf{x}}[k|k-1]}. \tag{10.28}$$

For the measurement model of the audio localizer from (10.23) to (10.25), the Jacobian matrix takes the form

$$\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} -\dfrac{y}{x^2+y^2} & 0 & \dfrac{x}{x^2+y^2} & 0 \\ \dfrac{x}{\sqrt{x^2+y^2}} & 0 & \dfrac{y}{\sqrt{x^2+y^2}} & 0 \end{bmatrix}. \tag{10.29}$$

At this point, the Kalman gain can be defined as [40,38]

$$\mathbf{K}[k] = \mathbf{P}[k|k]\,\hat{\mathbf{H}}^T[k]\left[\hat{\mathbf{H}}[k]\,\mathbf{P}[k|k]\,\hat{\mathbf{H}}^T[k] + \mathbf{C}_{\mathbf{nn}}[k]\right]^{-1}. \tag{10.30}$$

Finally, the correction step is computed from

$$\hat{\mathbf{x}}[k|k] = \hat{\mathbf{x}}[k|k-1] + \mathbf{K}[k]\,(\mathbf{y}[k] - \hat{\mathbf{y}}[k|k-1]). \tag{10.31}$$

This is basically the same operation as in the linear case. The only real difference is that a linearized measurement function is used at each iteration instead of a truly linear measurement matrix.

Although the extended Kalman filter is widely used, it must be kept in mind that it is a first-order approximation to the optimal estimator. It can be expected that the EKF will be effective if $\mathbf{h}(.)$ is "close" to linear in the state, and if the state estimate errors are not too large. However, if $\mathbf{h}(.)$ is highly nonlinear, the EKF may diverge. Similarly, if the errors in the state estimate are large, the effect of the nonlinearity becomes more severe, and divergence is likely [40,39]. No convergence results are known for the EKF, and it must be viewed as an *ad hoc* filter. Better results can be obtained using an iterated EKF [41].

10.2.3 Decentralized Kalman Filter

The constraint of a single sensor type can be a severe restriction. Consider, for example, the scenario of a video conferencing participant moving in front of the camera. The individual may disappear if the camera does not track sufficiently fast. To increase the robustness of any given video tracking algorithm, it may be beneficial to use a microphone array together with the camera.

The parallelized or decentralized (linear) Kalman filter (DKF) is a multisensor Kalman filter that has been divided up into modules, each one associated with a particular sensor system. The structure of this modular system is derived in the following paragraphs. First, the observation vector of the linear Kalman filter depicted in Figure 10.2 is partitioned into two components, one for each modality (audio and video). Next, a somewhat artificial construction, the so-called inverse Kalman filter, is introduced. It facilitates the parallelization of the estimation algorithm. Finally, the linear Kalman filter shown in Fig. 10.1, the inverse Kalman filter, and a Kalman filter whose observation vector has been partitioned into two components are used to construct the decentralized Kalman filter.

Partitioning the Observation Vector For a two-sensor system, the observation vector at step k, $\mathbf{y}[k]$, can be partitioned into two components, i.e.,

$$\mathbf{y}[k] = \begin{bmatrix} \mathbf{y}_1[k] \\ \mathbf{y}_2[k] \end{bmatrix}. \tag{10.32}$$

Treating the zero-mean measurement noise component, $\mathbf{n}[k]$, similarly, the audio (i=1) and video measurement equations (i=2) can be expressed as

$$\mathbf{y}_i[k] = \mathbf{H}_i\,\mathbf{x}[k] + \mathbf{n}_i[k], \quad i = 1, 2, \tag{10.33}$$

provided both (local) sensors observe the same state. If the state, $\mathbf{x}[k]$, and the measurement noise components, $\mathbf{n}_1[k]$ and $\mathbf{n}_2[k]$, are independent, then the centralized Kalman filter can be parallelized.

Rewriting (10.19) to make use of the fact that the audio and video measurement noises are independent produces

$$\mathbf{H}^T[k]\,\mathbf{C}_{\mathbf{nn}}^{-1}[k]\,\mathbf{y}[k] = \mathbf{H}_1^T[k]\,\mathbf{C}_{\mathbf{n}_1\mathbf{n}_1}^{-1}[k]\,\mathbf{y}_1[k] + \mathbf{H}_2^T[k]\,\mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k]\,\mathbf{y}_2[k], \tag{10.34}$$

where

$$\mathbf{H}[k] = \begin{bmatrix} \mathbf{H}_1[k] \\ \mathbf{H}_2[k] \end{bmatrix} \tag{10.35}$$

and

$$\mathbf{C}_{\mathbf{nn}}^{-1}[k] = \begin{bmatrix} \mathbf{C}_{\mathbf{n}_1\mathbf{n}_1}^{-1}[k] & 0 \\ 0 & \mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k] \end{bmatrix}. \tag{10.36}$$

Substituting (10.34) into (10.19) obtains

$$\begin{aligned} \hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\big(&\mathbf{P}^{-1}[k|k-1]\,\hat{\mathbf{x}}[k|k-1] + \\ &\mathbf{P}_1^{-1}[k|k]\,\mathbf{K}_1[k]\,\mathbf{y}_1[k] + \mathbf{P}_2^{-1}[k|k]\,\mathbf{K}_2[k]\,\mathbf{y}_2[k]\big). \end{aligned} \tag{10.37}$$

Note that (10.14) was used replace the terms $\mathbf{H}_i^T[k]\,\mathbf{C}_{\mathbf{n}_i\mathbf{n}_i}[k]$ found in (10.34) with $\mathbf{P}_i^{-1}[k|k]\,\mathbf{K}_i[k]$, $i = 1, 2$. The matrices $\mathbf{P}[k|k-1]$ and $\mathbf{P}[k|k]$ denote the global *a priori* and *a posteriori* error covariances, while $\mathbf{P}_i[k|k-1]$ and $\mathbf{P}_i[k|k]$, $i = 1, 2$, are their counterparts at the two local processors, respectively. The vector $\hat{\mathbf{x}}[k|k-1]$ is the global *a priori* state estimate based on the previous measurements up to time step k.

Similar to (10.20), the weighted observation vectors at the two local Kalman filters ($i = 1, 2$) are introduced as

$$\mathbf{q}_i[k] = \mathbf{P}_i^{-1}[k|k]\mathbf{K}_i[k]\,\mathbf{y}_i[k]. \tag{10.38}$$

Substituting into (10.37) yields

$$\hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\big(\mathbf{P}^{-1}[k|k-1]\,\hat{\mathbf{x}}[k|k-1] + \mathbf{q}_1[k] + \mathbf{q}_2[k]\big). \tag{10.39}$$

The structure of the associated Kalman filter is shown in Figure 10.3. It is closely related to Figure 10.2. The only difference is that there are now two inputs instead of one.

Inverse Kalman Filters The normal operation of a Kalman filter is to read observations from the input and to produce state estimates at the output. Formally, this mapping may also be inverted in the form of an *inverse* Kalman filter. It reads an estimate and produces the corresponding observation. It is not attempted to realize such a structure or to defend its practical use. The only justification for its existence is the fact that the equations for a Kalman filter may also be solved for the observations given the state estimates. Arranging a Kalman filter and its corresponding inverse Kalman filter in series formally reproduces the initial observations at the output.

The description of the inverse Kalman filter follows by rearranging terms in (10.21). Although this expression is valid for a single Kalman filter, it is only a matter of notation to extend it to two individual Kalman filters indexed with $i = 1$ and $i = 2$, respectively. That is,

$$\mathbf{q}_i[k] = \mathbf{P}_i^{-1}[k|k]\,\hat{\mathbf{x}}_i[k|k] - \mathbf{P}_i^{-1}[k|k-1]\,\hat{\mathbf{x}}_i[k|k-1], \quad i = 1, 2. \tag{10.40}$$

The structure of the inverse Kalman filter is shown in Figure 10.4. The name inverse Kalman filter was chosen to emphasize that this structure (re)produces the weighted observation vectors, $\mathbf{q}_i[k]$, from state estimates, $\hat{\mathbf{x}}_i[k|k]$.

Fusion Center Substituting the local *inverse* Kalman filters from (10.40) into (10.39), produces the joint *a posteriori* state estimate

$$\begin{aligned}\hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\big(&\mathbf{P}^{-1}[k|k-1]\,\hat{\mathbf{x}}[k|k-1] + \\ &\mathbf{P}_1^{-1}[k|k]\,\hat{\mathbf{x}}_1[k|k] - \mathbf{P}_1^{-1}[k|k-1]\,\hat{\mathbf{x}}_1[k|k-1] + \\ &\mathbf{P}_2^{-1}[k|k]\,\hat{\mathbf{x}}_2[k|k] - \mathbf{P}_2^{-1}[k|k-1]\,\hat{\mathbf{x}}_2[k|k-1]\big).\end{aligned} \tag{10.41}$$

The variables $\hat{\mathbf{x}}_i[k|k-1]$ and $\hat{\mathbf{x}}_i[k|k]$, $i = 1, 2$, refer to the local *a priori* and local *a posteriori* state estimates, respectively. The differences between the weighted local *a posteriori* and *a priori* estimates on the right hand side of (10.41) can be viewed as local *state error information* vectors. The joint *a posteriori* state estimate is the global state estimate.

Equation (10.41) is central for the fusion of the local audio and video state estimates. It constitutes the so-called fusion center which receives the local *a priori* and *a posteriori* state estimates together with their associated error covariance matrices, and it finally produces the global state estimate, $\hat{\mathbf{x}}[k|k]$. To this end, the global *a priori* error covariance matrix, $\mathbf{P}[k|k-1]$, and the global *a posteriori* error covariance matrix, $\mathbf{P}[k|k]$, are needed at each time step. The global *a priori* error covariance matrix follows from (10.10). The global *a posteriori* error covariance matrix, on the other hand, can be derived by substituting (10.35) and (10.36) into the *a posteriori* error covariance update equation of the centralized Kalman filter as shown in (10.17). This

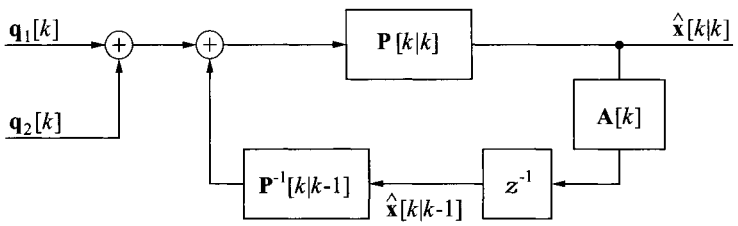


Fig. 10.3. Two-input Kalman filter

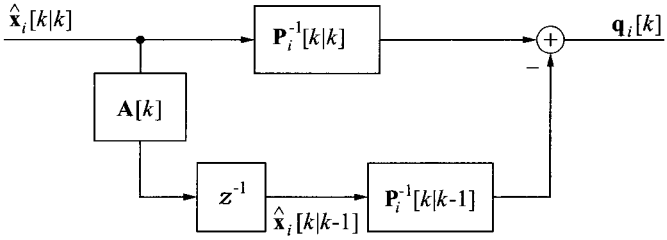


Fig. 10.4. Structure of the inverse Kalman filter.

yields

$$\mathbf{P}^{-1}[k|k] = \mathbf{P}^{-1}[k|k-1] + \mathbf{H}_1^T[k]\,\mathbf{C}_{\mathbf{n}_1\mathbf{n}_1}^{-1}[k]\,\mathbf{H}_1[k] + \mathbf{H}_2^T[k]\,\mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k]\,\mathbf{H}_2[k]. \tag{10.42}$$

When applied to either the audio (i=1) or video (i=2) systems, (10.17) can be rearranged to

$$\mathbf{H}_i[k]^T\mathbf{C}_{\mathbf{n}_i\mathbf{n}_i}^{-1}[k]\,\mathbf{H}_i[k] = \mathbf{P}_i^{-1}[k|k] - \mathbf{P}_i^{-1}[k|k-1]. \tag{10.43}$$

Taking $i = 1$ and $i = 2$ in (10.43) and substituting these terms back into (10.42), allows for the computation of the global *a posteriori* error covariance matrix via

$$\mathbf{P}^{-1}[k|k] = \mathbf{P}^{-1}[k|k-1] + \sum_{i=1}^{2}\left(\mathbf{P}_i^{-1}[k|k] - \mathbf{P}_i^{-1}[k|k-1]\right). \tag{10.44}$$

The overall sum comprising the global *a priori* error covariance matrix, $\mathbf{P}^{-1}[k|k-1]$, and the differences between the local *a posteriori* and *a priori* error covariance matrices is the global *a posteriori* error covariance matrix.

Structure of the decentralized Kalman filter. At this point, the decentralized linear Kalman filter for a two-sensor system has been introduced. Its simplified block diagram is shown in Figure 10.5. An audio localizer ($i = 1$) and a video localizer ($i = 2$) provide independent source position estimates (observations), $\mathbf{y}_i[k]$, which serve as inputs to the local Kalman filters, KF_1 and KF_2. The underlying structure of these filters can be found in Figure 10.1 or 10.2. The local Kalman filters compute their associated full-order *a priori* and *a posteriori* state estimates, $\hat{\mathbf{x}}_i[k|k-1]$ and $\hat{\mathbf{x}}_i[k|k]$, respectively. They also keep track of the associated *a priori* and *a posteriori* error covariance matrices, $\mathbf{P}_i[k|k-1]$ and $\mathbf{P}_i[k|k]$. Then, the local estimators pass their respective *a priori* and *a posteriori* state estimates together with the associated error covariance matrices on to the fusion center. The fusion center finally derives a global *a posteriori* state estimate, $\hat{\mathbf{x}}[k|k]$, by recursively processing (10.41).

In the linear case, the estimate generated by the central processor coincides with the global centralized estimate. Thus, there is no performance loss

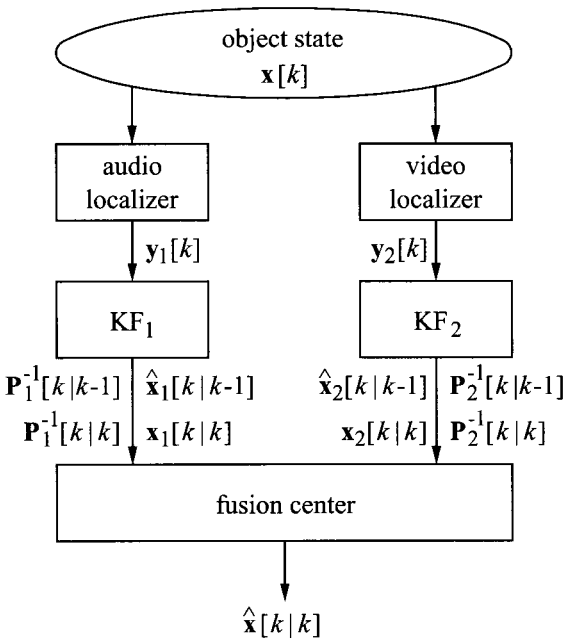


Fig. 10.5. Structure of the decentralized Kalman filter.

for a decentralized linear system. However, the algorithm does assume that the local processors are synchronized to the same speed. In general, this may not be the case. A solution to the problem of asynchronous operation can be found in [42].

Examining Sensor Failure It is illuminating to investigate what happens when one sensor fails. For example, suppose the microphone sensor ($i = 1$) provides no audio signal and subsequently, there is no localization information provided through this modality. Intuitively, it is expected that in this case the global state estimate equals the local video state estimate once all transient effects have vanished. This is indeed the case as will be shown below.

The absence of audio information is modeled by setting $\mathbf{H}_1[k] = \mathbf{0}$. Equation (10.37) can now be written as

$$\hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\big(\mathbf{P}^{-1}[k|k-1]\,\hat{\mathbf{x}}[k|k-1] + \mathbf{H}_2^T[k]\,\mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k]\,\mathbf{y}_2[k]\big). \quad (10.45)$$

Using (10.43) and (10.44), it may be seen that for $\mathbf{H}_1[k] = \mathbf{0}$ the global *a posteriori* error covariance matrix simplifies to

$$\mathbf{P}^{-1}[k|k] = \mathbf{P}^{-1}[k|k-1] + \mathbf{H}_2[k]^T\,\mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k]\,\mathbf{H}_2[k]. \quad (10.46)$$

Solving (10.46) for $\mathbf{P}^{-1}[k|k-1]$, substituting the result into (10.45), and rearranging some terms, gives

$$\hat{\mathbf{x}}[k|k] = \hat{\mathbf{x}}[k|k-1] + \mathbf{P}[k|k]\,\mathbf{H}_2^T[k]\,\mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k]\boldsymbol{\nu}[k|k-1] \tag{10.47}$$

where

$$\boldsymbol{\nu}[k|k-1] = \mathbf{y}_2[k] - \mathbf{H}_2[k]\,\hat{\mathbf{x}}[k|k-1]. \tag{10.48}$$

Note that (10.47) is basically the same state update equation as found at the local Kalman filter for the video sensor. Referring to (10.18), for example, finds that

$$\hat{\mathbf{x}}_2[k|k] = \hat{\mathbf{x}}_2[k|k-1] + \mathbf{P}_2[k|k]\,\mathbf{H}_2^T[k]\,\mathbf{C}_{\mathbf{n}_2\mathbf{n}_2}^{-1}[k]\boldsymbol{\nu}_2[k|k-1], \tag{10.49}$$

where

$$\boldsymbol{\nu}_2[k|k-1] = \mathbf{y}_2[k] - \mathbf{H}_2[k]\,\hat{\mathbf{x}}_2[k|k-1]. \tag{10.50}$$

A simple comparison of (10.47) and (10.49) shows that

$$\hat{\mathbf{x}}[k|k] = \hat{\mathbf{x}}_2[k|k], \tag{10.51}$$

$$\hat{\mathbf{x}}[k|k-1] = \hat{\mathbf{x}}_2[k|k-1], \text{ and} \tag{10.52}$$

$$\mathbf{P}[k|k] = \mathbf{P}_2[k|k]. \tag{10.53}$$

As promised, the parameters for the global recursive state estimate are indeed equivalent to those of the (local) video processor.

Combining Linear Kalman Filter and Extended Kalman Filter for Joint Object Localization The local Kalman filter at the microphone array, the local Kalman filter at the video camera, and the global Kalman filter are the three main components needed to recursively calculate a joint object position estimate. All the Kalman filters use the same dynamical model.

The video camera may be adjusted to return the object position directly in Cartesian coordinates, i.e.,

$$\mathbf{y}_2[k] = \begin{bmatrix} x_2[k] \\ y_2[k] \end{bmatrix} + \mathbf{n}_2[k]. \tag{10.54}$$

Its associated local state is

$$\mathbf{x}_2[k] = \left[x_2[k]\,\dot{x}_2[k]\,y_2[k]\,\dot{y}_2[k]\right]^T. \tag{10.55}$$

The microphone array, on the other hand, observes the source position in terms of azimuth and range, i.e., it provides measurements

$$\mathbf{y}_1[k] = \begin{bmatrix} \theta[k] \\ r[k] \end{bmatrix} + \mathbf{n}_1[k], \tag{10.56}$$

where $\theta[k]$ and $r[k]$ are given by (10.24) and (10.25), respectively. Since, $\theta[k]$ and $r[k]$ are nonlinearly related to the local state

$$\mathbf{x}_1[k] = \left[x_1[k]\, \dot{x}_1[k]\, y_1[k]\, \dot{y}_1[k]\right]^T \tag{10.57}$$

at the associated Kalman filter, a linear Kalman filter may no longer be used. An extended Kalman filter must be applied instead.

At this point, a linear and an extended Kalman filter are available for use at the local sensors, but a rigorous mathematical framework has only been developed for the linear case. Fortunately, the decentralized Kalman filter shown in Figure 10.5 is composed of autonomous components. This makes it possible to replace the local, initially linear audio Kalman filter, KF_1, with an extended Kalman filter. The linear Kalman filter for video localization, KF_2, and the fusion center remain unchanged.

10.3 Implementation

10.3.1 System description

The implementation of a joint audio-video processing system based on the theory discussed so far is illustrated below. There are many ways to implement the decentralized state estimator shown in Figure 10.5. These depend on the objects observed, the types of sensors available, and the requirements for localization and tracking. A system intended to track a single person in an audio-visual environment is presented here. It consists of a microphone array for audio localization and a video camera for tracking of human faces. Both sensors are combined using the recursive estimation scheme shown in Figure 10.5.

The structure of the audio localization system is shown in Figure 10.6. Its main elements are a microphone array, means to compute the cross correlations between pairs of microphone signals, and a summed correlator beamformer. The functionality of the summed correlator beamformer is basically the same as the classical steered beamformer, however, no continuous variation of delay elements is necessary. To inhibit erroneous estimates when no speech signal is present, a speech pause detector is employed. More details about the audio localizer can be found in [8,9].

The video localization system is a real-time face tracker with the following main elements: foreground-background segmentation, detection of skin-color regions, and detection of eye-like regions. A sample result can be found in Figure 10.7. The large rectangle in the foreground contains the detected skin color region and the two small rectangles indicate the eye positions. The face position is placed at the center of the bounding rectangle. This combination of skin-color and eye detection improves the robustness of the face tracker.

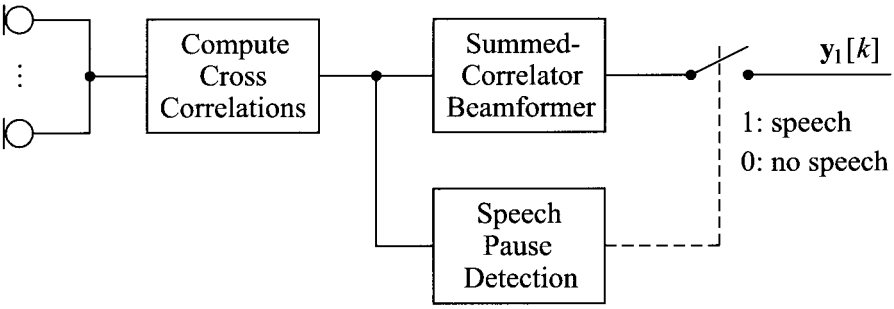


Fig. 10.6. Audio localization system: Summed correlator implementation of a steered beamformer.

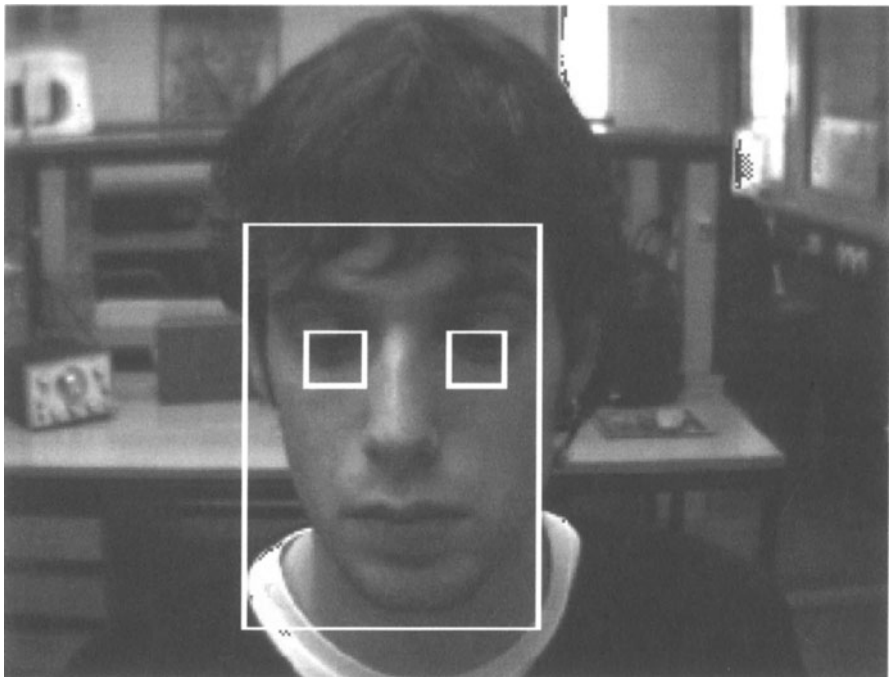


Fig. 10.7. Sample result of video localization.

10.3.2 Results

Tracking of a human speaker in an audio-visual environment is a very interesting application. Unfortunately, it does not easily facilitate a quantitative analysis since the true speaker position cannot be determined accurately by other means. To demonstrate the robustness and accuracy of joint audio-video tracking, this work will resort to an alternative setup: tracking of a model railway along an oval track in a plane. Knowledge of the fixed railway

track contour together with continuous measurements of the engine's exact position along the track provided the ground truth against which the audio-video tracking results can be compared. Figure 10.8 shows an example. To demonstrate the increased robustness of joint audio-video processing against sensor failure, it has been assumed that both modalities suffer from poor localization conditions at different times.

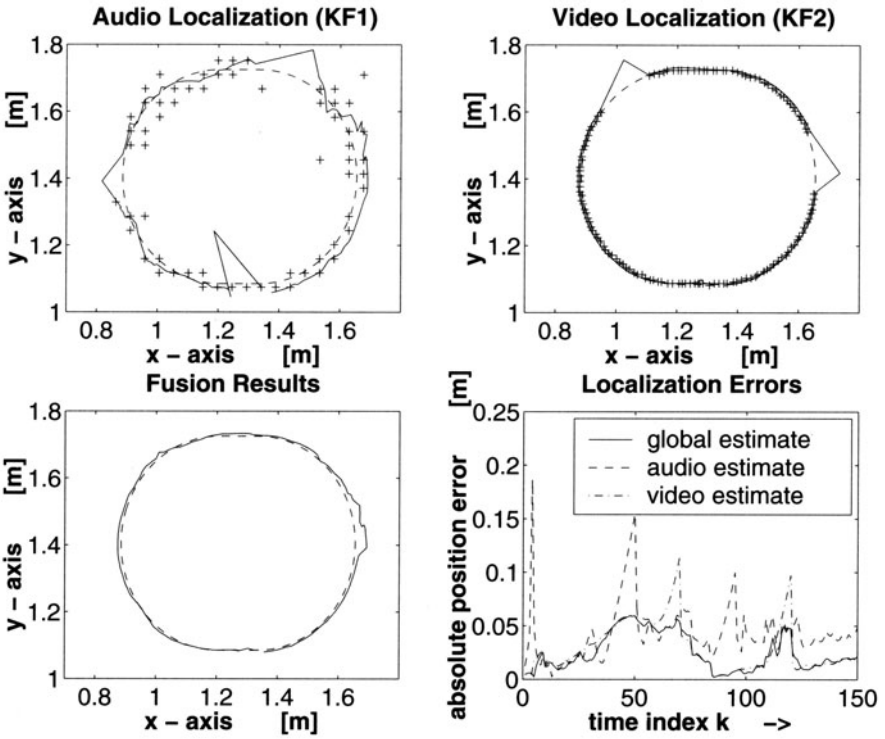


Fig. 10.8. Quantitative analysis of tracking errors: audio localization (top left), video localization (top right), joint audio-video localization (bottom left), localization errors (bottom right).

The audio localization results are shown on the left in the top row. The dashed line is the railway track. The sequence of position estimates from the summed correlator beamformer is indicated by crosses (+). They represent the input data, $\mathbf{y}_1[k]$, to the local extended Kalman filter, KF_1. The estimation result computed by the Kalman filter is depicted as a solid line. Since the state estimation process started at the bottom part of the track, the initial error during the first steps of the Kalman recursion is clearly visible. Furthermore, there are two instances in the sequence of position estimates where the raw position estimates (observations) were dropped to mimic a silent acoustic

source. In both cases, the Kalman filter extrapolated the position estimates based on the linear motion model of the local Kalman filter. When new input data became available, the position estimates resumed their proper course.

The situation is similar for video localization shown on the right of the top row. Since the camera usually has a much higher spatial resolution than the microphone array, the video position estimates are significantly more accurate in general. Again, two instances with missing video observations were simulated. As in the case of the audio localizer, the associated video position estimates were linearly extrapolated since the associated video Kalman filter, KF_2, uses the same motion model as the audio Kalman filter, KF_1. The fusion result is shown on the left of the bottom row. It may be seen that the joint estimation algorithm successfully removes deviations due to unreliable audio or video observations.

Finally, the plot on the right of the bottom row shows how the audio, video, and joint audio-video position estimates differ from the true object positions. The absolute position errors of the audio and video position estimates peak at the startup of the audio estimator and when there are failures related to missing mono-modal sensor observations. Since these deviations do not coincide in time, the joint estimate relies on the more accurate single localizer estimate in these cases. This example shows that joint audio-video object localization provides more robust results than either of the two mono-modal methods employed independently.

10.4 Discussion and Conclusions

The decentralized Kalman filter recursively combines local audio and video state estimates into a more reliable global state and, thus, position estimate. To this end, a common model of the system dynamics and a common coordinate system is needed. For a video sensor system operating in Cartesian coordinates, a linear local Kalman filter is appropriate. However, since the audio position estimates at the microphone array are computed with respect to a polar coordinate system, an extended Kalman filter is required. For a dynamical model based on Cartesian coordinates, the global coordinator fusing the local state estimates can apply a linear Kalman filter as well.

Although audio position estimates are often less accurate than the results obtained with a video localizer, they can still provide useful input for a joint audio-video object localization system. Taking a look at the absolute position error shown on the lower right in Figure 10.8 it may be seen that the global position estimate is always close to the best single-sensor output, but it is not necessarily more accurate. Nevertheless, by introducing a joint audio-video processor, a localizer that yields more reliable results than either one of the single-sensor systems is obtained.

The parallel Kalman filter structure which was applied only provides a starting point for joint audio-video object localization. There are many other

challenging problems along these lines which were not discussed here. For example, one issue centers around techniques used for modeling uncertainty in the integration and fusion process. In this case, the lack of precise measurement models is one factor that limits the performance of the fusion algorithm. Just knowing a position estimate obtained either using an audio or a video localizer is not really sufficient. To successfully fuse these estimates, it is necessary to know how accurate these observations are. For microphone arrays a method for predicting the error region associated with a particular source position estimate can be found in [43].

Another difficulty arises when tracking is based only on direction or angle measurements. This happens when an acoustic source leaves the near-field of a microphone array. One intuitively appealing and conceptually simple approach in this case is to estimate range or position based on two or more angle measurements from different directions. This can be obtained using distributed audio arrays [44,36] or multiple cameras [42]. A mathematical analysis of the associated triangulation error statistics can be found in [40].

One important ingredient of the Kalman tracking algorithm is the one-step predictor. It assumes that the source moves along a predictable trajectory with a deterministic velocity. As a result, such a tracking scheme is bound to be sensitive to sudden source maneuvers. One frequently used approach to take maneuvers into account employs multiple Kalman trackers operating in parallel [45,36,39,34]. To determine which state model is most consistent with the observations, it is necessary to evaluate the likelihood of the different state models. Once known, there are two ways to obtain the source's position: either use the result produced by the most likely tracker or weigh each model's position estimate by its probability and sum up the individual weighted position estimates.

For multi-sensor, multi-target applications involving many acoustic sources, a key issue is to determine which sensor observations are associated with which target. This issue is commonly referred to as the *data association* problem [46]. One way to distinguish between the tracks of various sources is to rely on source motion models [47]. A different approach is to use the interacting multiple model (IMM) estimator [48]. It automatically adapts itself to each object's motion mode. Current research in this area focuses on techniques such as multiple hypothesis tracking, probabilistic data association methods [49,50], and multiple criteria optimization theory [51].

References

1. W. Bangs and P. Schultheiss, "Space-time processing for optimal parameter estimation," in *Signal Processing* (J. Griffiths, P. Stocklin, and C. Schooneveld, eds.), pp. 577–591, Academic Press, 1973.
2. W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. on Information Theory*, vol. 19, no. 5, pp. 608–614, 1973.

3. W. Hahn, "Optimum signal processing for passive sonar range and bearing estimation," *Journal of the Acoustical Society of America*, vol. 58, no. 1, pp. 201–207, 1975.
4. G. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 463–470, 1981.
5. N. Owsley and G. Swope, "Time delay estimation in a sensor array," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 519–523, 1981.
6. W. Kellermann, "A self-steering digital microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-91)*, Toronto, Canada, pp. 3581–3584, June 1991.
7. H. Silverman and S. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Computer Speech and Language*, vol. 6, no. 2, pp. 129–152, 1992.
8. N. Strobel, T. Meier, and R. Rabenstein, "Speaker localization using steered filtered-and-sum beamformers," in *Proceedings Vision, Modeling, and Visualization '99* (B. Girod, H. Niemann, and H.-P. Seidel, eds.), (Erlangen), pp. 195–202, 1999.
9. N. Strobel and R. Rabenstein, "Robust speaker localization using a microphone array," in *Proceedings of the X European Signal Processing Conference*, vol. III, 2000.
10. M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*,vol. 11, no. 2, pp. 91–126, April 1997.
11. D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 371–374, April 1997.
12. Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. on Signal Processing*, vol. 42, no. 8, pp. 1905–1915, 1994.
13. H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 187–90, April 1997.
14. D. Rabinkin, R. Renomeron, A. Dahl, *et al.*, "A DSP implementation of source location using microphone arrays," in *SPIE Proceedings '96*, vol. 2846, pp. 88–98, 1996.
15. R. Chellappa, C. Wilson, and A. Sirohey, "Human and machine recognition of faces: A survey," *IEEE Proceedings*, vol. 83, no. 5, pp. 705–740, 1995.
16. A. Eleftheriadis and A. Jacquin, "Automatic face location, detection and tracking for model-assisted coding of video teleconferencing sequences at low bitrates," *Signal Processing: Image Communication*, vol. 7, no. 3, pp. 231–248, 1995.
17. L. Bala, K. Talmi, and J. Liu, "Automatic detection and tracking of faces and facial features in video sequences," in *Proceedings of the 1997 Picture Coding Symposium*, no. 143 in ITG-Fachberichte, pp. 251–256, 1997.
18. P. Fieguth and D. Terzopoulos, "Color-based tracking of heads and other mobile objects at video frame rates," in *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 21–27, 1997.

19. J. Crowley and P. Berard, "Multi-modal tracking of faces for video communications," in *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 640–645, 1997.
20. R. Quian, M. Sezan, and K. Matthews, "A robust real-time face tracking algorithm," in *Proceedings of the 1998 IEEE International Conference on Image Processing*, vol. 1, pp. 131–135, 1998.
21. R. Tsai, "A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE Trans. Robot. Autom.*, vol. 3, pp. 323–344, 1987.
22. O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.
23. D. Koller, D. Daniilidis, and H. Nagel, "Model-based object tracking in monocular image sequences and road traffic scenes," *International Journal of Computer Vision*, vol. 10, no. 3, pp. 257–281, 1993.
24. P. Arnoul, M. Viala, J. Guerin, and M. Mergy, "Traffic signs localisation for highway inventory from a video camera on board a moving collecting van," in *Proceedings of the 1996 IEEE Intelligent Vehicle Symposium*, pp. 682–687, 1996.
25. G. L. Foresti, "Object recognition and tracking for remote video surveillance," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1045–1062, 1999.
26. U. Bub, M. Hunke, and A. Waibel, "Knowing who to listen to in speech recognition: Visually guided beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 848–851, May 1995.
27. M. Collobert, R. Freauc, G. Tourneur, *et al.*, "LISTEN: a system for locating and tracking individual speakers," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 283–288, 1996.
28. G. Pingali, "Integrated audio-visual processing for object localization and tracking," in *Proceedings of the SPIE*, vol. 3310, pp. 206–213, 1997.
29. C. Wang and M. Brandstein, "A hybrid real-time face tracking system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 3737–3740, May 1998.
30. C. Wang and M. Brandstein, "Multi-source face tracking with audio and visual data," in *IEEE Int. Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, pp. 169–174, September 1999.
31. Y. Huang, J. Benesty, and G. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-2000)*, Instanbul, Turkey, pp. 1384–1387, June 2000.
32. N. Strobel and R. Rabenstein, "Fusion of multisensor data," in *Principles of 3D Image Analysis and Synthesis* (B. Girod, G. Greiner, and H. Niemann, eds.), pp. 309–322, Kluwer, 2000.
33. J. Richardson and K. Marsh, "Fusion of multisensor data," *International Journal of Robotics Research*, vol. 7, no. 6, pp. 78–96, 1988.
34. N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization using a recursive multi-state, multi-sensor estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-2000)*, Istanbul, Turkey, pp. 3781–3784, June 2000.
35. N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, Jan. 2001.

36. D. H. Johnson and D. E. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.
37. L. Scharf, *Statistical Signal Processing–Detection, Estimation, and Time Series Analysis*, Addison-Wesley, 1991.
38. J. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice Hall, 1995.
39. R. G. Brown and P. Y. Hwang, *Introduction to random signals and applied Kalman filtering*, Wiley, 1997.
40. T. Broida, "Kinematic and statistical models for data fusion using Kalman filtering," in *Data Fusion in Robotics and Machine Intelligence* (Abidi and Gonzales, eds.), pp. 311–365, Academic Press, 1992.
41. A. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.
42. B. Rao, H. Durrant-Whyte, and J. Sheen, "A fully decentralized multi-sensor system for tracking and surveillance," *International Journal of Robotics Research*, vol. 12, no. 1, pp. 20–44, 1993.
43. M. Brandstein, J. Adcock, and H. Silverman, "Microphone array localization error estimation with application to sensor placement," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3807–3816, 1996.
44. M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, 1997.
45. Y. Bar-Shalom and T. Fortman, *Tracking and Data Association*, Academic Press, 1988.
46. S. Blackman, "Association and fusion of multiple sensor data," in *Multitarget-Multisensor Tracking: Advanced Applications* (Y. Bar-Shalom, ed.), pp. 187–218, Artech House, 1990.
47. Y. Bar-Shalom and X. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*, Univ. of Conneticutt, 1995.
48. M. Yeddanapudi, Y. Bar-Shalom, and K. Pittipati, "IMM estimation for multitarget-multisensor air traffic surveillance," *IEEE Proceedings*, vol. 85, no. 1, pp. 80–94, 1997.
49. R. Mahler, "A unified foundation for data fusion," in *Seventh Joint Service Data Fusion Symposium*, 1994.
50. I. Goodman, "A general theory for the fusion of data," in *Tri-Service Data Fusion Symposium*, 1987.
51. A. Poore, "Multi-dimensional assignment formulation of data association problems arising from multi-target and multi-sensor tracking," *Computational Optimization Applicat.*, vol. 3, pp. 27–57, 1994.
52. G. Wang, R. Rabenstein, N. Strobel, and S. Spors, "Object localization by joint audio-video signal processing," in *Proceedings Vision, Modeling, and Visualization 2000* (B. Girod, G. Greiner, H. Niemann, and H.-P. Seidel, eds.), Saarbrücken, Germany, pp. 97–104, Nov. 2000.

Part III

Applications

11 Microphone-Array Hearing Aids

Julie E. Greenberg[1] and Patrick M. Zurek[2]

[1] Massachusetts Institute of Technology, Cambridge MA, USA
[2] Sensimetrics Corporation, Somerville MA, USA

Abstract. Microphone-array hearing aids provide a promising solution to the problems encountered by hearing-impaired persons when listening to speech in the presence of background noise. This chapter first discusses implementation issues and performance metrics specific to the hearing-aid application. A review of previous work on microphone-array hearing aids includes systems with directional microphones, fixed beamformers, adaptive beamformers, physiologically-motivated processing, and binaural outputs. Recent simulation results of one promising adaptive beamforming system are presented. The performance of microphone-array hearing aids depends heavily on the acoustic environments in which they are used. Additional information about the level of reverberation, number of interferers, and relative levels of interferers encountered by hearing-aid users in everyday situations is required to quantify the benefit of microphone-array hearing aids and to select the optimal processing strategy.

11.1 Introduction

Hearing impairments affect 10–16% of the total adult population and at least one-third of the elderly population [1–5]. Recent surveys in the United States indicate that roughly 20% of the hearing-impaired population (5–6 million people) use hearing aids, but only 60% of users are satisfied with their devices [6].

Users' overall satisfaction with their hearing aids is highly dependent on the situations in which they are used. While 90% of hearing aid users indicated that they were satisfied with their hearing aid for one-on-one listening, 62% were satisfied in small group settings, 45% in workplaces and restaurants and only 25% in large groups [7]. In an analysis of why hearing aid owners do not use their hearing aids, the second most prevalent reason, given by 25% of respondents, was that the hearing aids "did not work" in background noise [8].

These complaints are not surprising given the greater susceptibility of hearing-impaired listeners to interference from background noise. Normal-hearing listeners typically exhibit speech reception thresholds[1] (SRT's) of

[1] The speech reception threshold is the signal-to-interference ratio at which speech is 50% intelligible.

−5 dB when listening monaurally. Hearing-impaired persons, however, typically require a speech-to-interference ratio (SIR) that is 5–10 dB higher than a normal-hearing person to achieve the same level of speech understanding [9].

The problem of poor performance in background noise with conventional hearing aids has motivated the use of microphone arrays to create directionally sensitive hearing aids that selectively amplify or attenuate sounds based on their direction of arrival. The goal of microphone-array hearing aids is to improve the SIR when the interference arises from a different direction than the desired speech signal. Because hearing aid users still require additional processing to compensate for frequency-dependent and level-dependent characteristics of their hearing loss, the microphone array can be considered as a preprocessor, followed by conventional hearing aid processing.

Profoundly hearing-impaired individuals who are not helped by conventional hearing aids are candidates for cochlear implants. Cochlear implant systems typically include an ear-level microphone, a body-worn speech processor, and an array of electrodes implanted in the inner ear. Since the vast majority of cochlear implants are unilateral, and since the speech processor encodes only a subset of the acoustic information in the speech signal, cochlear implant users are even more sensitive than hearing aid users to interference from background noise. A study of phoneme recognition found that relative to normal hearing listeners, cochlear implant users require an additional 11 dB SIR to achieve the same performance [10]. As in the hearing aid application, a microphone array can be considered as a preprocessor to the cochlear implant speech processor.

11.2 Implications for Design and Evaluation

11.2.1 Assumptions Regarding Sound Sources

Considering the microphone array as a preprocessor to a hearing aid or cochlear implant, the problem is to design a system to improve the desired speech-to-interference ratio. Since the interference may also be speech, it is first necessary to distinguish between desired speech and interference based on the direction of arrival.

It is assumed that the desired signal arises from a small range of angles, while signals arriving from all other angles outside this *look direction* are interferers.[2] This assumption will be violated by any coherent reflections of the desired speech that arrive from outside the look direction. Although later reflections do not contribute to speech intelligibility and may appropriately be considered as interference, early reflections (arriving within up to 50–95 ms of the direct wave) do contribute to intelligibility and should be considered part of the desired speech signal [12]. But since these early reflections most likely

[2] The look direction should include a range of angles spanning at least 20° azimuth to account for minor steering errors and normal head movements [11].

arrive from outside the look direction, they may be treated as interference by the microphone array. Furthermore, these reflections constitute a source of correlated interference, and their presence may lead to degrading performance for many conventional adaptive beamforming algorithms.

It is also usually assumed that the user physically steers the array's look direction toward the desired speech source. For head-worn arrays, this is accomplished by head movements to face the desired source; for hand-held systems the device can be pointed at the desired sound source. Automatic steering (electronic control of the look direction) is also possible but is not considered here.

11.2.2 Implementation Issues

The hearing aid application presents particular challenges for the design of microphone arrays with regard to configuration, size, weight, and power consumption. The popularity of 'invisible' in-the-canal hearing aids indicates that many hearing aid users consider cosmetic acceptability to be extremely important. Because cochlear implant users already wear a body-worn processor connected to ear-level components with wires, this population may be less concerned with cosmetic acceptability.

Functionally, a microphone-array hearing aid consists of three components: the microphone array itself, the processing unit, and the receivers (speakers) that deliver sound to the ears. Ideally, these components should be interconnected via wireless communication, although that is not generally possible with current technology. In the absence of wireless communication, the ideal processing unit would be small enough to be incorporated into either the mechanical structure of the microphone array or the receivers, in order to avoid cumbersome wiring between components. If a separate body-worn processor is necessary, it must be relatively lightweight (comparable to a personal portable stereo system) and have reasonable battery life. Although the availability of specialized, low-power electronic components makes advanced signal processing algorithms feasible, computational complexity remains an important issue in determining which algorithms can be implemented in practical devices.

The array size and number of microphones is limited to that which can be comfortably mounted on eyeglass frames or a headband, or worn around the neck ($< 20\,\text{cm}$). In order to be comfortable and not unwieldy, hand-held devices face a similar size limit. Simpler and more cosmetically-acceptable array designs may be obtained by using one microphone (or two closely-spaced microphones) at each ear as in conventional binaural hearing aids, although this leads to spatial undersampling and requires (currently-unavailable) wireless or (cumbersome) wired signal transmission between components at the two ears. Increasing the number of microphones beyond two provides diminishing improvements. For arrays spanning up to 20 cm with realistic levels of

sensor noise, the incremental improvement in directionality due to additional microphones is negligible beyond four to six microphones [13].

11.2.3 Assessing Performance

A number of physical and behavioral metrics can be used to assess the performance of microphone-array hearing aids. The *directivity* of a microphone array is the ratio of the array's power gain in the look direction to the array's power gain averaged over all directions. The directivity of an omnidirectional microphone in free space is 0 dB. The directivity of a microphone array varies with frequency. For applications involving broadband signals such as speech, the directivity values are commonly averaged across the frequency range of interest. The *intelligibility-weighted directivity* [14] can be computed by applying frequency-dependent weights to the directivity values for individual frequencies. These weights, derived from Articulation Index theory [15], reflect the relative importance of each frequency band to speech intelligibility.

For the special case of isotropic noise, the source is attenuated by an amount equal to the directivity of the array. For a non-isotropic source, the amount of attenuation depends on its spatial and frequency characteristics and may be more or less than the directivity. The attenuation of interference alone is an indicator of array performance, but does not take into account algorithmic effects on, or due to, the desired signal. The *improvement in SIR*, or gain, of an array (using linear processing) is the difference between the attenuation of the interference and the attenuation of the desired source. Although ideally the desired source arrives from the look direction and is not attenuated at all, in practice attenuation may result from sensor mismatch, mis-steering, or reflections of the desired signal arriving from outside the look direction. Like the directivity, broadband measures of SIR improvement may be obtained by averaging across frequency, and a measure of *intelligibility-weighted gain* is obtained by incorporating frequency-dependent weights.

Measures of directivity and improvement in SIR are useful as preliminary metrics to quantify the physical performance of a microphone array. However, the final assessment of any array-processing hearing aid must be based on intelligibility tests with hearing-impaired subjects. Percent correct scores on a particular listening test can be used to show differences among algorithms. However, because percent correct scores saturate in extremely favorable and unfavorable listening situations, and because the relationship between intelligibility and SIR varies with different speech materials, it is difficult to translate percent correct scores to more general measures of benefit or to make comparisons across studies.

The SRT is generally preferable to percent correct scores, because SRT scores can be interpreted directly with respect to an algorithm's effect on SIR. However, SRT scores do have an important limitation. Because they are typically measured via an adaptive procedure which varies the interference level to obtain 50% intelligibility, a single SRT score reflects algorithm performance

at a range of SIR's, and SRT scores for different subjects performing the same task may reflect algorithm performance at widely varying SIR's. This is not a problem for algorithms that provide a fixed directional pattern regardless of SIR. However, for algorithms whose directional responses vary with relative signal strength, the SRT will not reveal variations in performance at different SIR's. In particular, the SRT does not measure performance at the two extremes of interference alone and speech in quiet. For adaptive algorithms, SRT scores should be augmented by another measure to verify that the algorithm does not have detrimental effects on desired speech in quiet.

Meaningful evaluations of more complex microphone-array hearing aids should be based on comparison to the performance of simpler microphone-array systems, rather than an omnidirectional microphone. For example, the reference condition for evaluating fixed beamformers should be either a single directional microphone or a pair of directional microphones worn in binaural hearing aids, depending on the comparison desired. The reference condition for evaluating adaptive beamformers should be the *underlying fixed beamformer* based on a simple delay-and-sum combination of microphone signals for the same physical array. In this way, it is possible to assess the incremental benefits due to each increase in the size and computational complexity of microphone-array hearing aids.

11.3 Hearing Aids with Directional Microphones

The goal of microphone-array hearing aids – to selectively attenuate acoustic signals based on the direction of arrival – can to some extent be accomplished simply by using a directional microphone in a conventional hearing aid. Because a directional microphone samples the acoustic signal at two points in space, it can be considered a simple acoustic array.

Traditional directional microphones use an acoustic network to delay and combine acoustic signals from the two ports of the device, followed by electrical transduction of the resulting acoustic signal. Some newer directional microphones instead transduce the acoustic signals at both ports and subsequently delay and combine the two signals electronically. It is possible to use either technique to achieve desired directional specifications. However, the directional characteristics of devices using an acoustic network are static, while devices that combine the signals electronically are more flexible, permitting the directional characteristics to be altered in real time.

Hearing aids with directional microphones have been available for over 25 years, and their advantages for listening to speech in background noise are well established [16,17]. Even so, devices with directional microphones have traditionally accounted for a relatively small fraction of hearing aids sales, perhaps due to the inflexibility of their directional patterns. The recent introduction of commercial devices with multiple directional patterns has spurred renewed interest, increasing the popularity of hearing aids with directional

microphones. These devices include Phonak's AudioZoom™ [18], which provides a user-controlled switch between directional and omnidirectional modes, and GN Resound's Digital 5000 [19], which creates omnidirectional and three different directional patterns. Phonak's Adaptive digital AudioZoom™ attempts to suppress the loudest interferer by automatically switching among different directional patterns.

A single directional microphone improves the directivity by 4–6 dB relative to an omnidirectional microphone [18–20] and has been shown to increase SRT's by up to 8 dB in favorable acoustic environments [21], although SRT improvements of 3–5 dB are more typical [17,19,22]. One study suggests that the benefit may be reduced to 1–2 dB in strong reverberation [23], presumably because the directional pattern reduces reflections of the desired signal as well as the interference, an effect which is discussed in Section 11.8.

11.4 Fixed-Beamforming Hearing Aids

Fixed beamformers provide an attractive processing option for microphone-array hearing aids. The required processing is relatively simple and can be implemented with small, low-power analog or digital electronics to meet the constraints discussed in Section 11.2.2. Unlike adaptive beamformers discussed in Section 11.5, fixed beamformers are relatively robust to minor steering errors and reflections correlated with the desired signal. Well-designed fixed beamformers provide optimal or near-optimal interference reduction in isotropic noise and in extremely reverberant environments. However, because they do not alter their directivity patterns to provide nulls in particular directions, fixed beamformers provide suboptimal performance when directional interferers are present in less reverberant environments.

Stadler and Rabinowitz [20] analyzed the effects of array orientation, number of microphones, and directionality of microphones for several types of fixed array processing, including classic delay-and-sum beamforming, maximum directivity at 4000 Hz, and superdirective arrays with constraints on noise sensitivity. For 14-cm broadside arrays, they found intelligibility-weighted directivity values of 7–8 dB for arrays of two directional microphones with uniform weights (that is, classic delay-and-sum beamforming). For broadside arrays, there was little or no benefit from adding more than two microphones or using the more computationally demanding sensitivity-constrained array processing. In contrast, endfire arrays showed a clear advantage for additional microphones and more complicated processing. Five-microphone, 11-cm endfire arrays with sensitivity-constrained processing provided 9–11 dB of intelligibility-weighted directivity. However, using a practical method for implementing the superdirective sensitivity-constrained processing and a performance metric designed to distinguish the beneficial early reflections of the desired signal from detrimental late reflections, Liu and

Sideman [24] found the intelligibility-weighted gain of this endfire array to be only 5 dB in extreme reverberation.

Kates and Weiss [25] evaluated three fixed processing strategies using a 10-cm endfire array with five omnidirectional microphones in two moderately reverberant rooms. The optimal superdirective array outperformed classic delay-and-sum beamforming and an oversteered array, providing 7–11 dB of gain weighted for speech intelligibility. In a follow-up study with 16 hearing-impaired subjects [26], the superdirective array provided an intelligibility benefit of 5–6 dB relative to a single omnidirectional microphone, but only 2–3 dB relative to a single cardioid directional microphone.

Soede et al. [27,28] designed, implemented, and evaluated five-microphone fixed beamforming hearing aids based on 14-cm broadside and 10-cm endfire arrays with relatively simple processing designed to obtain maximum directivity at 4000 Hz. Initial analysis and measurements revealed directivity values ranging from 5 dB at 500 Hz to 11–13 dB at 4000 Hz, corresponding to intelligibility-weighted directivities of 8–9 dB [20]. Evaluations with 45 hearing-impaired subjects in an artificial diffuse noise field found an average of 7 dB improvement in SRT for both the broadside and endfire arrays relative to the subjects' own hearing aids with omnidirectional microphones.

Etymotic Research's ArrayMic^{TM} is a commercial endfire array that implements a fixed beamforming preprocessor for hearing aids. This device utilizes relatively simple processing designed to obtain maximum directivity in the range 1–2 kHz. It is available in a 6-cm, three-microphone version and an 11-cm, five-microphone version with intelligibility-weighted directivities of 7 dB and 10 dB, respectively. The benefit of these devices should be compared to 4 dB for D-Mic^{TM}, Etymotic Research's directional hearing aid microphone, which is small enough to fit in an in-the-ear hearing aid.

Another commercial fixed-processing device is the Radiant Beam Array from Starkey Laboratories. It includes six microphones mounted in a 21-cm wide housing worn around the neck and interfaced to an ear-level hearing aid via a telecoil. Pairs of microphone signals are bandpass filtered and summed to maximize directivity for individual frequency ranges based on intermicrophone spacings [29]. The device provides directivities ranging from 8–11 dB as a function of frequency. Part of this directivity presumably results from the acoustic shadow cast by the body for rearward sources.

11.5 Adaptive-Beamforming Hearing Aids

Adaptive beamformers alter their directional patterns in response to changes in the acoustic environment, potentially providing better performance than fixed beamformers under certain conditions. Those conditions include acoustic environments with low-to-moderate reverberation and relatively few interfering sources. However, adaptive beamformers are more sensitive than fixed beamformers to array errors such as sensor mismatch and mis-steering, and

to correlated reflections arriving from outside the look direction. A serious problem with some adaptive beamforming algorithms is that they may degrade the desired signal when such situations occur, especially when the SIR is high.

Because the performance of adaptive beamformers is heavily influenced by the acoustic environments encountered by the user, and because laboratory studies can only attempt to approximate a range of real-world listening situations, the benefit of adaptive-beamforming over fixed-beamforming hearing aids will ultimately be determined by field trials. Even so, initial assessment of adaptive beamforming algorithms necessarily takes place in simulations and laboratory environments. Because different studies evaluating adaptive-beamforming algorithms generally use different acoustic environments, and because the performance of these adaptive algorithms is heavily dependent on the environment, it is virtually impossible to make meaningful comparisons of improvement in SIR across studies. Furthermore, these studies have used a variety of speech materials and performance metrics for intelligibility tests, further confounding comparisons. As a result, this section will summarize promising adaptive algorithms that have been evaluated in the context of the hearing aid application, without attempting across-study comparisons of benefits.

11.5.1 Generalized Sidelobe Canceler with Modifications

One promising approach that has received much attention is the generalized sidelobe canceler (GSC) [30] (Figure 11.1). This system includes a blocking matrix that removes signals arriving from the look direction to produce reference signals free of the desired signal. The reference signals are input to adaptive filters performing unconstrained minimization, and the adaptive filter outputs are then subtracted from a delayed primary signal consisting of desired signal plus interference. The unconstrained minimization can be accomplished in the time-domain using the LMS algorithm, which is attractive because of its low computational complexity. Because misadjustment of the adaptive weights is proportional to desired signal strength even in the ideal case, and because some desired signal leaks into the reference signal under realistic conditions (due to sensor mismatch, mis-steering, or reverberation), the traditional GSC performs poorly at high SIR, degrading the desired signal.

A number of modifications and adjustments to the GSC effectively overcome this problem. First, cancellation of the desired signal based on its reflections can be completely prevented by appropriate selection of the primary signal delay [31]. In particular, the delay must be shorter than the interval between the arrival of the direct wave and the first reflection at the microphones [32]. This limits the non-causal portion of the adaptive filter's response that can cancel the direct desired signal based on its reflections received at the array. Second, the sum method [33] alters the normalization

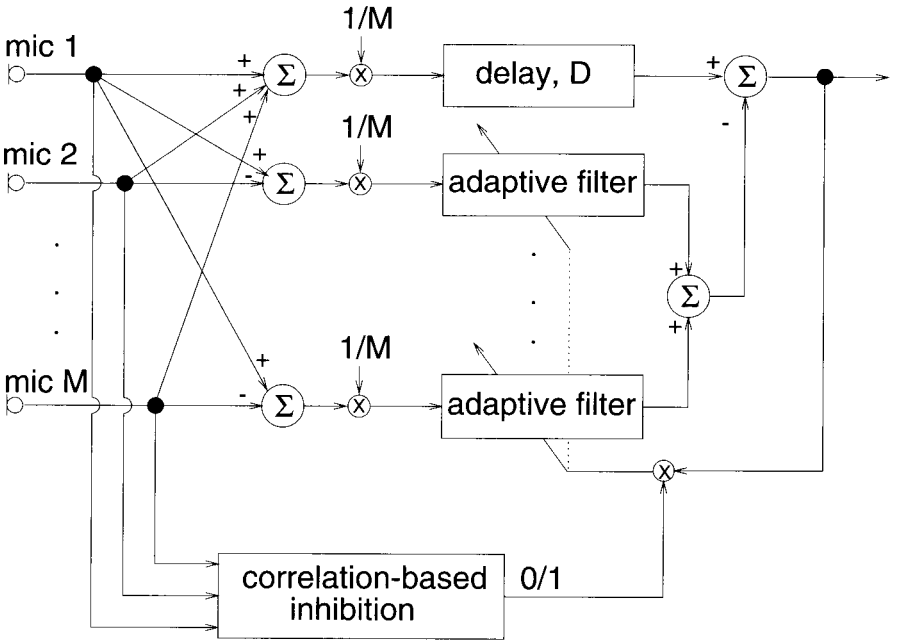


Fig. 11.1. Block diagram of the modified generalized sidelobe canceler for arrays with an arbitrary number of microphones. The primary signal is formed from the weighted sum of all microphone signal, and the blocking matrix is implemented by pairwise differences between microphone signals.

of the LMS step size, reducing misadjustment due to strong desired signals. Finally, correlation-based inhibition [34] (Figure 11.2) freezes the adaptive weights when the SIR is estimated to be high, further mitigating the effects of strong desired signals.

Effect of Modifications Simulations of a modified GSC with linear broadside arrays of two and five omnidirectional microphones under a variety of conditions verified that proper selection of the primary delay eliminates desired signal cancellation due to reverberation and that the modifications mitigate the problems of strong desired signals [34,35]. Figure 11.3 shows the steady-state[3] intelligibility-weighted gain, G_I [14], with and without the sum method and correlation-based inhibition for two SIRs and three levels of reverberation. Negative values of G_I illustrate the detrimental effects of the unmodified GSC caused by both high excess mean-square error from strong desired signals and desired signal cancellation due to mis-steering of the array. These detrimental effects are prevented by the modifications, which also improve performance in less critical conditions. These results clearly show the

[3] A discussion of the transient behavior is given in [35].

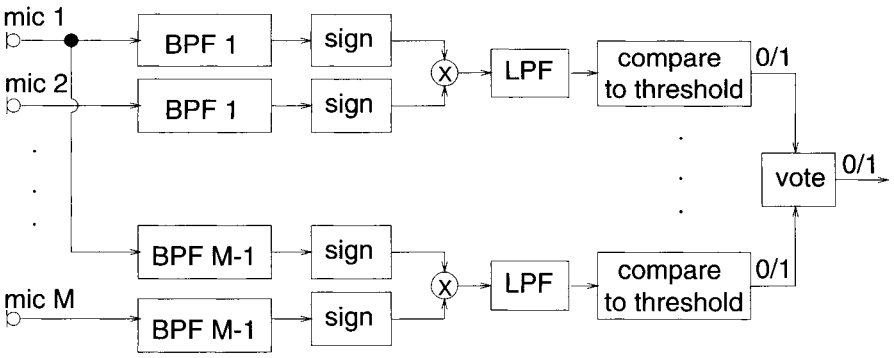


Fig. 11.2. Block diagram of correlation-based inhibition for arrays with an arbitrary number of microphones. Each pair of microphone signals is bandpass filtered (BPF). The center frequencies of the bandpass filters are selected according to $f = \frac{c}{2d}$, where c is the speed of sound and d is the intermicrophone spacing, in order to minimize the expected correlation from sources outside the look direction. For the 7-cm, two-microphone array considered here, the passband was 1650–3300 Hz. For the 16-cm, five-microphone array, the passbands were 3000–5000 Hz for 4 cm intermicrophone spacing, 1700–3000 Hz for 8 cm, 1250–1700 Hz for 12 cm, and 860–1250 Hz for 16 cm. The bandpass filtered signals are hard limited to ± 1 by taking the sign and then multiplied together. This instantaneous correlation is then smoothed by a first-order recursive lowpass filter (LPF) with a time constant of 10 ms, selected to track power fluctuations corresponding to the duration of individual phonemes. The lowpass filtered correlation values from each microphone pair is compared to a threshold of zero to make a binary decision about the range of SIR in that frequency band. Finally, an overall binary decision about the range of SIR is reached based on the majority of frequency bands voting for high or low SIR.

efficacy of both modifications, as the worst performance for each condition is generally obtained with neither modification, while the best performance is always obtained with both modifications.

Number of Microphones and Adaptive Filter Length Figure 11.4 shows simulation results with both modifications and the same four combinations of number of microphones and adaptive filter length as above, operating at three levels of SIR and in three levels of reverberation. In addition, performance of the underlying fixed beamformers is provided as a reference to assess the benefit due to the adaptive systems.

Performance is generally better for longer filters than shorter ones. The magnitude of this difference is greater in reverberation, because longer filters are better able to cancel not only the direct interference but also some of its reflections. The improvement in the anechoic environment results from the longer filter providing a better approximation of the summed signal in the primary channel based on the difference signals in the reference channels.

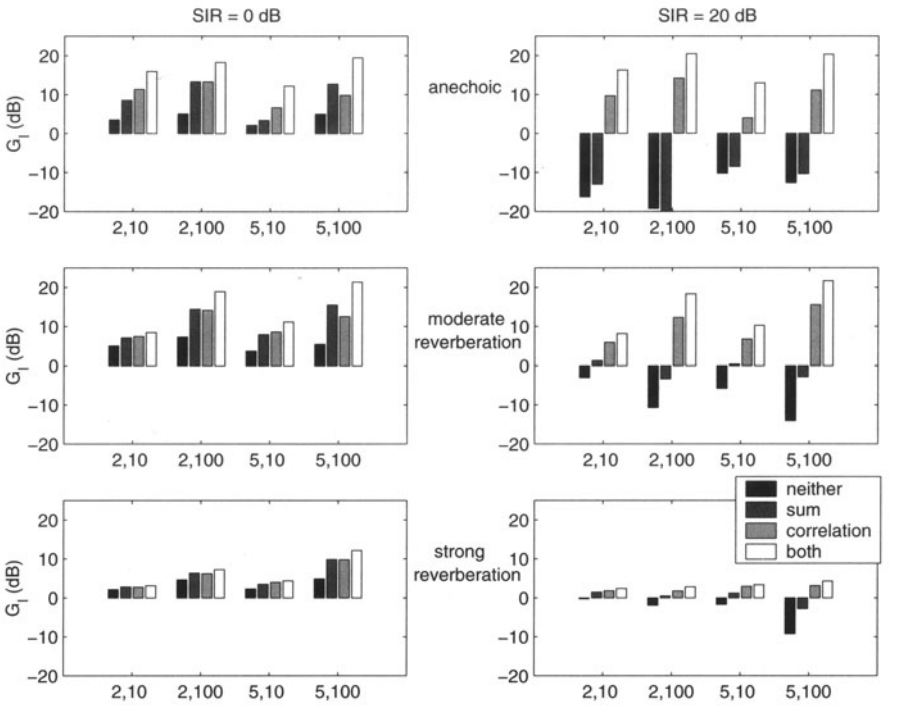


Fig. 11.3. Effect of modifications discussed in the text on the intelligibility-weighted gain, G_I. The pairs of numbers labelling the abscissa indicate the number of microphones (2 or 5) and the adaptive filter length (10 or 100 ms). The primary delay was $D = 5$ ms. The speech and interference (babble) sources were located at 10° and 55° azimuth relative to array broadside, respectively. Signal levels were varied to produce SIRs of 0 and +20 dB (two columns). A room simulation [36] incorporating a free-space linear array (either 7-cm two-microphone array or 16-cm five-microphone array) was used with different wall absorption values to produce three acoustic environments (three rows): anechoic, moderately reverberant (+6 dB direct-to-reverberant ratio, 150 ms reverberation time), and strongly reverberant (−2 dB direct-to-reverberant ratio, 620 ms reverberation time). Full processing details and parameters are given in [35].

In reverberation, increasing the number of microphones provides improvements, although small, because the system can direct additional nulls toward the strongest reflections of the interferer. However, these simulations only include a single interferer, and it is expected that the benefit of additional microphones would be more pronounced in the presence of multiple independent sources of interference.

In the anechoic environment, increasing the number of microphones degrades performance in some cases due to attenuation of the desired signal. Although the positive values of G_I seen in Figure 11.4 are primarily the

result of interference suppression, there is often a secondary, negative component due to attenuation of the desired signal. In most cases, this effect is small (typically 3 dB or less in reverberant conditions). However, for the five-microphone array in an anechoic environment, this effect causes 10-20 dB of desired signal attenuation.[4] This occurs because the five-microphone array can steer four independent nulls; with only one interferer to cancel and the array slightly mis-steered, it directs a null towards the desired source, which it finds during intervals when SIR is low enough to allow adaptation but which still contain the desired speech. Because the processing improves overall SIR by suppressing the interference more than the desired signal, it might be possible to counteract this effect by simply increasing the broadband gain. However, the desired signal attenuation is not uniform across frequency, so any such compensation would need to be rather sophisticated in order to compensate for frequency-specific effects. Overall, this result indicates that for arrays with more than two microphones, the modifications mitigate, but do not completely eliminate, desired signal cancellation.

Effect of Reverberation Another important trend in Figure 11.4 is the decrease in performance of the adaptive system with increasing reverberation. This is consistent with previous evaluations of the modified GSC [34], as well as with work on other adaptive beamformers discussed in the remainder of this section. Increasing reverberation results in decreasing interference cancellation because it increases the effective length of the impulse response relating the primary and reference signals, so that less of that response can be matched by a fixed-length filter. As a result, performance of any adaptive system declines with increasing reverberation for constant filter length. Conversely, cancellation performance improves with increasing filter length for a fixed level of reverberation.

The foregoing analysis assumes that the only effect of increasing reverberation is to increase the length of the temporal response needed to characterize the relationship between primary and reference signals. It also assumes that this temporal response can be adequately matched by an adaptive filter of sufficient length. In real-world applications, however, both the stationarity of the reverberant environment and the practical adaptation time of the filter will limit the temporal range over which increasing reverberation can be counteracted by increasing the length of the adaptive filter. In extreme reverberation the interference sound field is approximately diffuse and no adaptive benefit is possible; interference reduction is limited to the directivity provided by the underlying fixed processing.

[4] The most extreme case of this occurs for the five-microphone array with 10-ms adaptive filter in the anechoic space at 0 dB SIR, where the G_I value of 12 dB results from 31 dB of interference suppression together with 19 dB of desired signal attenuation.

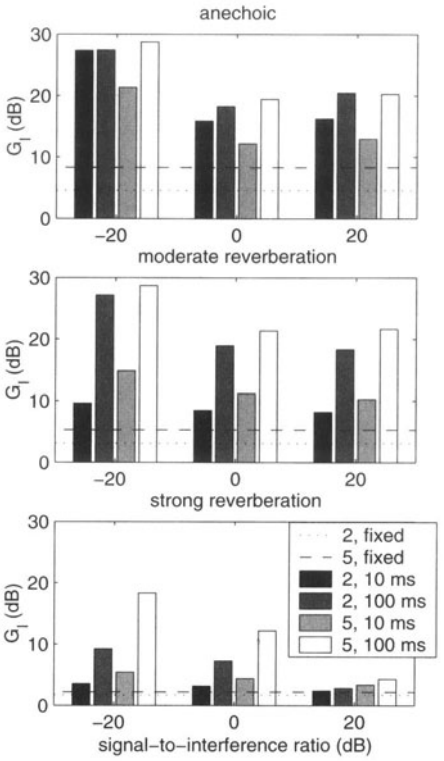


Fig. 11.4. Effect of design parameters. Intelligibility-weighted gain, G_I, for a generalized sidelobe canceler with both modifications discussed in the text, for three SIRs and three levels of reverberation. Values in each cluster of bars represent results for two- and five-microphone arrays with adaptive filter lengths of 10 ms and 100 ms. The lines on each plot indicate the performance of the underlying fixed beamformer. Other parameters and conditions are as in Figure 11.3

Other GSC Modifications and Evaluations Kompis and Dillier designed and evaluated adaptive systems with two microphones placed at the ears, using processing similar to the modified GSC described above. They incorporated directional microphones [37,38] and explored variations of correlation-based inhibition [39,40]. Objective intelligibility tests with normal-hearing and hearing-impaired subjects verified that the modified adaptive beamformer with directional microphones always outperformed either the underlying fixed beamformer with directional microphones or the same adaptive beamformer with omnidirectional microphones. Although they did not measure SRTs directly, their results indicate that the adaptive beamformer improved SRTs by more than 5 dB compared to the fixed beamformer (both using directional microphones).

Vanden Berghe and Wouters [41] studied an adaptive GSC-based system for a small endfire array consisting of two directional microphones spaced by 3 cm on a behind-the-ear (BTE) hearing aid shell. They showed a clear benefit (5 dB in SRT) of adaptive processing in moderate reverberation relative to a single directional microphone, but it is not clear if similar benefit could have been obtained using the underlying fixed beamformer.

Van Hoesel and Clark [42] evaluated a modified GSC as a preprocessor for a cochlear implant system. They used two ear-level cardioid microphones and an alternate inhibition method to freeze the adaptive weights when the SIR is high. Physical measurements showed approximately 10 dB of interference reduction, compared to 3 dB for the underlying fixed beamformer. Intelligibility tests with four experienced cochlear implant users in a moderately reverberant environment showed virtually no effect of the adaptive beamformer on speech in quiet (indicating no degrading effect of the beamforming) and an improvement of 35 percentage points in sentence recognition at 0 dB SIR, relative to the underlying fixed beamformer.

Hamacher et al. [43] also evaluated cochlear implant preprocessors based on the modified GSC with and without spectral subtraction to suppress diffuse noise and reverberation. For a single directional interferer in moderate and strong reverberation, the modified GSC improved SRTs by 6 and 2 dB, respectively, relative to the underlying fixed beamformer, while the modified GSC with spectral subtraction improved SRTs by 3 and 2 dB in those same conditions. However, in a moderately reverberant space with a large number of independent interferers, the addition of spectral subtraction improved SRTs by 4 dB, compared to only 1 dB without spectral subtraction.

11.5.2 Scaled Projection Algorithm

Other researchers have attempted to design robust adaptive beamformers for hearing aids using alternate techniques. Hoffman and Buckley [44] implemented a robust processor based on a combination of linear, quadratic, and eigenvector constraints using the scaled projection algorithm [45] in conjunction with a GSC. They evaluated head-sized arrays with three, five, and seven microphones and measured intelligibility-weighted gains and speech reception thresholds. Results show that this approach provides 6–8 dB improvement over the underlying fixed beamformer in moderate reverberation and approximately 2 dB in strong reverberation. These improvements are robust for any number of microphones, even in the presence of strong desired signals [31]. As in other work, performance is limited by reverberation, and in extreme reverberation the performance of the adaptive beamformer approaches that of the underlying fixed beamformer.

Kates and Weiss [25] also used the scaled projection algorithm, with and without a composite correlation matrix intended to improve performance in the presence of strong desired signals and correlated reflections. In reverberation, the scaled projection algorithm alone was superior to delay-and-sum beamforming, but showed little improvement over the fixed superdirective array discussed in Section 11.4. Performance of the scaled projection algorithm with the composite correlation matrix was inferior to that of the scaled projection algorithm alone.

11.5.3 Direction of Arrival Estimation

Wang et al. [46] and Korompis et al. [47] investigated a two-stage approach consisting of direction-of-arrival (DOA) estimation followed by maximum energy array processing designed to preserve the signal arriving from a range of angles surrounding the look direction while directing nulls towards known interference locations. If accurate estimates of source location are available, then the maximum energy approach effectively nulls directional interferers. However, these results do not demonstrate effective methods for performing the DOA estimation adaptively and robustly in the presence of reverberation and strong desired signals.

The location-estimation null-steering (LENS) algorithm [48] takes a similar approach. First, it uses a narrowband technique to estimate optimal null locations for each frequency to minimize interference power. These estimates are subject to robustness control processing that limits desired signal cancellation and then are used to formulate adaptive filters that steer the nulls accordingly. Simulations demonstrate that the LENS algorithm provides rapid convergence and robust performance in the presence of strong desired signals and steering errors, for multiple interferers. As with the adaptive algorithms discussed previously, performance decreases with increasing levels of reverberation.

11.5.4 Other Adaptive Approaches and Devices

Other attempts to design adaptive beamforming hearing aids have encountered similar problems and limitations with respect to strong desired signals and reverberation. Shields and Campbell [49,50] evaluated a system that performs adaptive beamforming independently in a number of subbands, with adaptation strategies that vary from band-to-band based on the coherence between bands. Tests with normal-hearing and hearing-impaired listeners demonstrated that the system improves intelligibility for SIRs up to 3 dB in moderate reverberation, but it has little effect in strong reverberation and was not evaluated at high SIRs.

McKinney and DeBrunner [51] and DeBrunner and McKinney [52] have investigated two systems: an array of two hypercardioid microphones with a multirate filter bank plus a modified narrowband minimum mean-square error algorithm, and a four-microphone array consisting of two hypercardiod microphones at each ear, separated by 1.25 cm in the horizontal plane. Both of these arrays exhibited promising performance in anechoic environments, but should be evaluated further in the presence of reverberation. In addition, the four-element array exhibited substantial reductions in performance at high SIR.

Investigators at Planning Systems Inc. have developed a prototype device based on a combination of fixed and adaptive beamforming. A 9-cm, four-microphone endfire array was attached to the side of the user's eyeglass frames

for testing. Evaluations with hearing-impaired listeners in a variety of acoustic environments indicate that the array is more effective (up to 5 dB in SRT) than a commercially-available directional-microphone hearing aid [53].

AVR Communications, Ltd., markets a hand-held, 15-cm four-microphone adaptive beamforming array that uses FM transmission to interface to an ear-level hearing aid. The device is advertised to reduce interferers by 15 dB or more.

11.6 Physiologically-Motivated Algorithms

The normal human auditory system does a remarkable job of extracting a desired speech signal from interference. For spatially distinct interferers, the typical SRT of −5 dB for normal-hearing persons listening monaurally is augmented by an additional 5 dB from binaural hearing. This advantage is comprised of equal contributions from head shadow and binaural interaction [54].

Some researchers have proposed microphone-array hearing aids utilizing two microphones placed at the ears, with processing based on the binaural characteristics of the human auditory system [55–57]. Frequency-domain analysis of the microphone signals is followed by computation of intermicrophone phase and level differences in each frequency band. A frequency-dependent weight vector is set to unity for bands with intermicrophone phase/level differences near values expected for sound sources arriving from straight ahead; lower weight values are applied to attenuate frequency bands with values deviating from the desired range. The weight function can also be modified based on estimates of the diffuseness of the sound field derived from the intermicrophone coherence [55], or based on voice detection and pitch estimates [58].

Evaluations of these algorithms revealed no benefit for normal hearing subjects listening at relatively low SIRs, but some improvements for hearing-impaired listeners operating at moderate SIRs. In a variety of acoustic environments, hearing-impaired listeners show modest improvements in speech intelligibility and listener satisfaction at SIRs ranging from −3 dB to +3 dB [55,57,59]. The Audallion BEAMformer^{TM} is a commercial device using this type of algorithm. It was developed at AudioLogic Inc. and is available from Cochlear Ltd. as a preprocessor to their cochlear implant system. Evaluation of this device in a relatively nonreverberant environment showed that eight out of nine implant users tested at +5 or +10 dB SIR received substantial improvements in intelligibility scores, while two implant users tested at +20 dB SIR showed no improvement [60]. In another study, four cochlear implant users showed an average SRT improvement of 2.5 dB in two moderately reverberant environments, but no improvement in stronger reverberation, relative to the underlying fixed beamformer [61].

11.7 Beamformers with Binaural Outputs

The above discussion largely assumes that the array processor uses inputs from multiple microphones to produce a single output that is presented monaurally or diotically[5]. However, some hearing-impaired listeners retain significant binaural hearing abilities and benefit from wearing two hearing aids [17,62]. This suggests designing microphone-array hearing aids that have two outputs to preserve binaural cues, allowing the wearer to use any residual binaural capacities both for noise reduction and for maintaining a sense of auditory space.

A straightforward way to obtain binaural outputs is by using two independent arrays, one on each side of the head, and presenting the output of each array to the ipsilateral ear. Soede et al. [28] compared monaural and binaural presentation of signals from endfire arrays worn on eyeglass temples and found that the SRT of hearing-impaired listeners improved by 2.5 dB with the addition of the second array; this improvement is comparable to that of binaural over monaural hearing aids (using omnidirectional microphones).

Another approach is to design a single array with two outputs, where processing is a compromise between improved directionality and preservation of binaural cues. Merks et al. [63] suggested a head-worn system based on a five-element broadside array, with two sets of fixed weights selected to direct the main lobes 10° to the left or right. Desloge et al. [64] designed and evaluated a variety of techniques for merging array processing with the preservation of binaural cues. One simple design is the lowpass/highpass system, where signals are split into two frequency bands. The low-frequency components obtained from microphones placed near the ears are presented binaurally to preserve directional cues, while high-frequency components are array-processed to improve directionality. This design takes advantage of the fact that low-frequency interaural delay cues are particularly important for sound localization [65], while most array processing techniques are more effective at high frequencies.

Using normal hearing listeners, Welker et al. [66] evaluated a lowpass/highpass system with adaptive beamforming in the high-frequency channel and found that the cutoff frequency between the low- and high-frequency bands effectively controlled the tradeoff between localization ability due to binaural cues and intelligibility improvement due to noise reduction. However, a subsequent study [67] suggests that hearing impaired listeners do not benefit as much as expected from improved SIR in the high frequency band, and may require that the beamformer operate over the entire frequency range in order to obtain practical benefits. This result is likely due to limitations on the ability of hearing-impaired listeners to extract useful information from audible speech cues in the high frequency region [68,69] and emphasizes the need for evaluations including hearing-impaired subjects. Furthermore,

[5] same signal at both ears

it suggests inclusion of a user-controlled switch that selects between different operating modes for interference reduction and binaural cue preservation, similar to recent commercial use of switching between directional and omnidirectional microphones.

11.8 Discussion

Recent technological advances and new commercial devices hold promise for overcoming the difficulties experienced by hearing aid users listening to speech in background noise. In some acoustic environments, hearing aids with a single directional microphone improve SRTs by 3–4 dB, a modest intelligibility benefit. While directional microphones *alone* provide modest benefits, those benefits are generally additive, supporting the use of directional microphones as elements of larger arrays.

Results of studies with fixed beamformers indicate that arrays spanning 10–15 cm provide directivity values up to 10 dB, which is expected to provide substantial benefits for understanding speech in background noise. Although simple two-microphone broadside arrays with uniform weights provide a few dB improvement over a single directional microphone, the best performance is provided by five-microphone endfire arrays using directional microphones. Such an array must be hand held or body worn and is likely to be conspicuous, making it cosmetically unacceptable to a significant segment of the population.

Two-element adaptive arrays with ear-level microphones offer a cosmetically acceptable option, particularly if wireless communication can be achieved between elements at the ears. The performance of two-microphone adaptive systems varies with the acoustic environment, but several studies suggest that modified generalized sidelobe cancelers provide 5–7 dB improvement over comparable fixed beamformers in moderate reverberation, for a total improvement of 10–12 dB over a single omnidirectional microphone. This additional benefit of the adaptive system is significant in reaching the 5–10 dB improvement in SIR typically required by hearing aid users to obtain intelligibility comparable to normal-hearing listeners. These adaptive systems generally provide the same benefit as the underlying fixed beamformer in extreme reverberation, with the notable exception of the adaptive system augmented with spectral subtraction [43], which warrants further evaluation.

The benefit of adaptive beamformers with more than two microphones remains to be established. These systems have the cosmetic disadvantages of five-microphone fixed beamformers, as well as high computational requirements. Even so, such systems may appeal to a portion of the hearing-impaired population if they substantially outperform other microphone arrays. With more than two microphones, the modified generalized sidelobe canceler still suffers some desired signal cancellation, making the LENS algorithm [48] and Hoffman's robust processing [44] the most promising candidates.

Although computer simulations and tests with normal-hearing listeners play an important role in algorithm selection and refinement, future evaluations of fixed- and adaptive-beamforming hearing aids must include hearing-impaired listeners. Physical metrics and tests with normal hearing listeners often overestimate the benefit to hearing-impaired listeners [26,28]. Presumably this is because intelligibility-weighted performance metrics are appropriate for normal-hearing listeners, but underestimate the importance of low frequencies for hearing impaired listeners [26,68,69]. This is a serious challenge in the design of microphone-array hearing aids, because it reduces the expected benefit of array processing, which generally provides greater directivity at higher frequencies.

Field trials are also necessary to evaluate algorithms under realistic acoustic conditions, since laboratory experiments cannot consider the full range of acoustic environments encountered by hearing aid users in their daily lives. The most significant acoustic variable is the degree of reverberation. Reverberation has two main effects on array performance, related to the interference and the desired signal. Because these effects are further complicated by some aspects of adaptive beamformers, the following discussion first considers the performance of arrays with fixed directional patterns (directional microphones and fixed beamformers).

Because a fixed system always presents the same directional pattern, the amount by which it attenuates the direct wave and reflections of a reverberant source is determined by the source's frequency content and angles of incidence on the array. The direct waves of the desired signal and interference will be treated differently, because the desired signal arrives from the look direction and is not attenuated, while the interference arrives from outside the look direction and will be attenuated according to the angle of incidence. Reflections of both the desired signal and the interference will arrive from all directions and will be attenuated approximately equally. In low-to-moderate reverberation, the direct wave effect will dominate, and the directional pattern will improve SIR. In extreme reverberation, the reflected energy effect will dominate, and the fixed directional pattern will not improve SIR.

Rather than considering the reverberation time, for these purposes the level of reverberation is more accurately quantified by the direct-to-reverberant ratio. The direct-to-reverberant ratio may differ for different sources in the same acoustic environment, depending on their proximity to the array. In any level of reverberation, the benefit of a fixed system can be predicted from the directional pattern, the source locations, and the direct-to-reverberant ratios of the desired signal and interference. In low-to-moderate reverberation, SIR improvements primarily result from the array's effect on the direct waves, discussed above. In strong reverberation, SIR improvements are primarily related to the amount by which the direct-to-reverberant ratio of the desired signal exceeds the direct-to-reverberant ratio of the interference.

In contrast to fixed systems, adaptive beamformers attenuate interference primarily by directing nulls toward the direct wave of the interference; the magnitude of this cancellation can be estimated from the direct-to-reverberant ratio of the interference. In moderate reverberation, long adaptive filters may provide secondary benefits by nulling strong reflections and/or modeling the pattern of reflections. In extreme reverberation, the interference resembles isotropic noise, and optimal performance is provided by a fixed beamformer designed to maximize directivity. As a result, in extreme reverberation, the performance of well-designed adaptive beamformers will be comparable to that of the underlying fixed beamformer, with SIR improvements determined by the difference in direct-to-reverberant ratios, as discussed above.

An alternate way to describe the above effects of reverberation is by considering improvement in SIR as a function of level of reverberation. (See Figure 10 of [34] and Figure 7 of [24] for examples.) For interference alone, the performance of a fixed beamformer exhibits moderate, monotonic variation with direct-to-reverberant ratio, while adaptive beamformer performance greatly exceeds fixed beamformer performance in low reverberation and asymptotically approaches the fixed performance in extreme reverberation. Including the desired signal[6] reduces the fixed beamformer's overall performance by an amount related to the desired signal's direct-to-reverberant ratio. This of course also affects the asymptotic performance of adaptive beamformers.

The benefits of microphone-array hearing aids will ultimately depend strongly on the nature of the acoustic environments in which they are used. If hearing aid users frequently experience difficulties due to interferers in mild-to-moderate reverberation, then directional microphones, fixed beamformers, and adaptive beamformers will each provide incremental benefits. In strong reverberation with the desired source in close proximity to the array (thereby increasing the direct-to-reverberant ratio), directional microphones and fixed beamformers will continue to provide practical benefits. If interference sources are in close proximity, the additional complexity of adaptive systems will be worthwhile.

Increasing reverberation requires longer adaptive filters to reach the same level of interference reduction, and for many adaptive algorithms, the convergence time is proportional to filter length. Real-world constraints on filter length, convergence time, and stationarity of the user and the acoustic environment will limit the additional improvements provided by adaptive systems.

Other acoustic variables of interest are the number and relative levels of interferers. In theory, the main benefit of adaptive arrays with more than two microphones (relative to adaptive arrays with two microphones) will be

[6] either by considering the entire signal as favorable to intelligibility or distinguishing between early and late reflections, as in [24]

manifest in environments with multiple independent, directional interferers. Survey data on the frequency with which hearing-aid users encounter such interferers would be useful for specifying the number of microphones required in adaptive systems.

Acknowledgements

This work was supported by Grant No. NIH-5-R01-DC00117, Grant No. NIH-5-R01-DC00270, and Contract No. N01-DC-5-2017 from the National Institute on Deafness and Other Communication Disorders.

References

1. E. Kramarow, H. Lentzner, R. Rooks, J. Weeks, and S. Saydah, *Health and Aging Chartbook*, National Center for Health Statistics, 1999.
2. A.C. Davis, "The prevalence of hearing impairments and reported hearing disability among adults in Great Britian," *Int. J. Epidemiol.*, vol. 18, pp. 911–917, 1989.
3. D.H. Wilson, P.G. Walsh, L. Sanchez, A.C. Davis, A.W. Taylor, G. Tucker, and I. Meagher, "The epidemiology of hearing impairment in an Australian adult population," *Int. J. Epidemiol.*, vol. 28, pp. 247–252, 1999.
4. B. Karlsmose, T. Lauritzen, and A. Parvin, "Prevalence of hearing impairment and subjective hearing problems in a rural Danish population ages 31–50 years," *Br. J. Audiol.*, vol. 33, pp. 395–402, 1999.
5. S. Uimonen, K. Huttunen, K. Jounio-Ervasti, and M. Sorri, "Do we know the real need for hearing rehabilitation at the population level? Hearing impairments in the 5- to 75-year-old cross-sectional Finnish population," *Br. J. Audiol.*, vol. 33, pp. 53–59, 1999.
6. S. Kochkin, " 'Baby Boomers' spur growth in potential market, but penetration rate declines," *Hearing J.*, vol. 52, no. 1, pp. 33–48, 1999.
7. S. Kochkin, "MarkeTrak V: Consumer satisfaction revisited," *Hearing J.*, vol. 53, no. 1, pp. 38–55, 2000.
8. S. Kochkin, "MarkeTrak V: 'Why my hearing aids are in the drawer': The consumers' perspective," *Hearing J.*, vol. 53, no. 2, pp. 34–42, 2000.
9. R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," *J. Speech Hear. Res.*, vol. 29, pp. 146–154, 1986.
10. I. Hochberg, A. Boothroyd, M. Weiss, and S. Hellman, "Effects of noise and noise suppression on speech perception by cochlear implant users," *Ear Hear.*, vol. 13, pp. 263–270, 1992.
11. T.J. Schwander and H. Levitt, "Effect of two-microphone noise reduction on speech recognition by normal-hearing listeners," *J. Rehabil. Res. Dev.*, vol. 24, pp. 87–92, 1987.
12. L. Cremer and H.A. Müller, *Principles and Applications of Room Acoustics*, (T.J. Schultz, translator), Applied Science Publishers, 1982.
13. P.M. Peterson, *Adaptive array processing for multiple microphone hearing aids*, PhD Thesis, Massachusetts Institute of Technology, Cambridge MA, USA, 1989.

14. J.E. Greenberg, P.M. Peterson, and P.M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Am.*, vol. 94, pp. 3009–3010, 1993.
15. ANSI, *American National Standard Methods for the Calculation of the Articulation Index, ANSI S3.5-1969*, American National Standards Institute, 1969.
16. H.B. Nielsen and C. Ludvigsen, "Effect of hearing aids with directional microphones in different acoustic environments," *Scand. Audiol.*, vol. 7, pp.217–224, 1978.
17. D.B. Hawkins and W.S. Yacullo, "Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation," *J. Speech Hear. Disord.*, vol. 49, pp. 278–286, 1984.
18. H. Bächler and A. Vonlanthen, "Audio Zoom – Signal processing for improved communication in noise," *Phonak Focus*, vol. 18, 1995.
19. GN Resound: personal communication.
20. R.W. Stadler and W.M. Rabinowitz, "On the potential of fixed arrays for hearing aids," *J. Acoust. Soc. Am.*, vol. 94, pp. 1332–1342, 1993.
21. M. Valente, D.A. Fabry, and L.G. Potts, "Recognition of speech in noise with hearing aids using dual microphones," *J. Am. Acad. Audiol.*, vol. 6, pp. 203–210, 1995.
22. M. Valente, G. Schuchman, L.G. Potts, and L.B. Beck, "Performance of dual-microphone in-the-ear hearing aids," *J. Am. Acad. Audiol.*, vol. 11, pp. 181-189, 2000.
23. A.R. Leeuw and W.A. Dreschler, "Advantages of directional hearing aid microphones related to room acoustics," *Audiol.*, vol. 30, pp. 330-344, 1991.
24. C. Liu and S. Sideman, "Simulations of fixed microphone arrays for directional hearing aids," *J. Acoust. Soc. Am.*, vol. 100, pp. 848–856, 1996.
25. J.M. Kates and M.R. Weiss, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Am.*, vol. 99, pp. 3138–3148, 1996.
26. G. Saunders and J. Kates, "Speech intelligibility enhancement using hearing-aid array processing," *J. Acoust. Soc. Am.*, vol. 102, pp. 1827–1837, 1997.
27. W. Soede, A.J. Berkhout, and F.A. Bilsen, "Development of a directional hearing instrument based on array technology," *J. Acoust. Soc. Am.*, vol. 94, pp. 785–798, 1993.
28. W. Soede, F.A. Bilsen, and A.J. Berkhout, "Assessment of a directional microphone array for hearing impaired listeners," *J. Acoust. Soc. Am.*, vol. 94, pp. 799–808, 1993.
29. B. Widrow and M.N. Brearley, "Directional hearing aid," U.S. Patent no. 4,751,738, 1988.
30. L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. 30, pp. 27–34, 1982.
31. M.W. Hoffman, T.D. Trine, K.M. Buckley, and D.J. Van Tasell, "Robust adaptive microphone array processing for hearing aids: Realistic speech enhancement," *J. Acoust. Soc. Am.*, vol. 96, pp. 759–770, 1994.
32. J.E. Greenberg and P.M. Zurek, "Preventing reverberation-induced target cancellation in adaptive-array hearing aids," *J. Acoust. Soc. Am.*, vol. 95, pp. 2990–2991, 1994.
33. J.E. Greenberg, "Modified LMS algorithms for speech processing with an adaptive noise canceller," *IEEE Trans. Speech Audio Proc.*, vol. 6, pp. 338–351, 1998.

34. J.E. Greenberg and P.M. Zurek, "Evaluation of an adaptive-beamforming method for hearing aids," *J. Acoust. Soc. Am.*, vol. 91, pp. 1662–1676, 1992.
35. J.E. Greenberg, *Improved design of microphone-array hearing aids*, PhD Thesis, Harvard-MIT Division of Health Sciences and Technology, Cambridge MA, USA, 1994.
36. P.M. Peterson, "Simulation of the impulse response between a single source and multiple, closely-spaced receivers in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, pp. 1527–1529, 1986.
37. M. Kompis and N. Dillier, "Noise reduction for hearing aids: Combining directional microphones with an adaptive beamformer," *J. Acoust. Soc. Am.*, vol. 96, pp. 1910–1913, 1994.
38. M. Kompis, P. Feuz, G. Valentini, and M. Pelizzone, "A combined fixed/adaptive beamforming noise-reduction system for hearing aids," in *Proc. 20th Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 3136–3139, 1998.
39. N. Dillier, T. Frölich, M. Kompis, H. Bögli, and W.K. Lai, "Digital signal processing (DSP) applications for multiband loudness correction digital hearing aids and cochlear implants," *J. Rehabil. Res. Dev.*, vol. 30, pp. 95–109, 1993.
40. M. Kompis, N. Diller, J. Francois, J. Tinembart, and R. Häusler, "New target-signal-detection schemes for multi-microphone noise-reduction systems for hearing aids," in *Proc. 19th Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 1990–1993, 1997.
41. J. Vanden Berghe and J. Wouters, "An adaptive noise canceller for hearing aids using two nearby microphones," *J. Acoust. Soc. Am.*, vol. 103, pp. 3621–3626, 1998.
42. R.M. vanHoesel and G.M. Clark, "Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients," *J. Acoust. Soc. Am.*, vol. 97, pp. 2498–2503, 1995.
43. V. Hamacher, W.H. Döering, G. Mauer, H. Fleischmann, and J. Hennecke, "Evaluation of noise reduction systems for cochlear implant users in different acoustic environments," *Am. J. Otol.*, vol. 18, pp. S45–S49, 1997.
44. M.W. Hoffman and K.M. Buckley, "Robust time-domain processing of broadband microphone array data," *IEEE Trans. Speech Audio Proc.*, vol. 3, pp. 193–203, 1995.
45. H. Cox, R.M. Zeskind, and M.M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 35, pp. 1365–1376, 1987.
46. A. Wang, K. Yao, R.E. Hudson, D. Korompis, F. Lorenzellii, and S. Gao: "Microphone array for hearing aid and speech enhancement applications," in *Proc. IEEE Int. Conf. Application Specific Systems, Architectures and Processors*, pp. 231–239, 1996.
47. D. Korompis, A. Wang, and K. Yao, "Comparison of microphone array designs for hearing aid," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 2739–2742, May 1995.
48. J.G. Desloge, *The location-estimating null-steering (LENS) algorithm for adaptive microphone-array processing*, PhD Thesis, Massachusetts Institute of Technology, Cambridge MA, USA, 1998.
49. P.W. Shields and D.R. Campbell, "Multi-microphone sub-band adaptive signal processing for improvement of hearing aid performance: preliminary results using normal hearing volunteers," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 415–418, 1997.

50. P.W. Shields and D.R. Campbell, "Multi-microphone noise cancellation for improvement of hearing aid performance," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 3633–3636, 1998.
51. E.D. McKinney and V.E. DeBrunner, "A two-microphone adaptive broadband array for hearing aids," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-96)*, Atlanta GA, USA, pp. 933–936, 1996.
52. V.E. DeBrunner, E.D. McKinney, "A directional adaptive least-mean- square acoustic array for hearing aid enhancement," *J. Acoust. Soc. Am.*, vol. 98, pp. 437–444, 1995.
53. G.L. Gibian, W. Koroljow, A. LaRow, S. Shaw, P. Nelson, and L. Sherlock, "Clinical trials of a hybrid adaptive beamformer (HAB) for improved speech understanding in noise," in *Collected Papers, 137th meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association, Forum Acusticum, integrating the 25th German Acoustics DAGA Conference*, Berlin, Germany, Mar. 1999.
54. P.M. Zurek, "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, (G.A. Studebaker and I. Hochberg, eds.), Allyn and Bacon, pp. 255–275, 1993.
55. T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W.S. Woods, and B. Kollmeier, "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction," *Acustica*, vol. 83, pp. 684–699, 1997.
56. B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, vol. 95, pp. 1593–1602, 1994.
57. H. Schweitzer, A.M. Terry, and M.A. Grim, "Three experimental measures of a digital beamforming signal processing algorithm," *J. Am. Acad. Audiol.*, vol. 7, pp. 230–239, 1996.
58. E. Lindemann and J.L. Melanson, "Noise reduction system for binaural hearing aid," U.S. Patent no. 5,651,071, 1997.
59. B. Kollmeier, J. Pessig, and V. Hohmann, "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," *J. Rehabil. Res. Dev.*, vol. 30, pp. 82–94, 1993.
60. V. Margo, C. Schweitzer, and G. Feinman, "Comparisons of Spectra 22 performance in noise with and without an additional noise reduction preprocessor," *Seminars in Hearing*, vol. 18, pp. 405–415, 1997.
61. V. Hamacher, H. Fleischmann, G. Mauer, and W.H. Döring, "Evaluation of noise reduction systems for cochlear implant users in different acoustic environment," presented at *Third European Symposium on Paediatric Cochlear Implantation*, Hannover, Germany, June 1996.
62. D. Byrne and W. Noble, "Optimizing sound localization with hearing aids," *Trends in Amplification*, vol. 3, pp. 51–73, 1998.
63. I.L.D.M. Merks, M.M. Boone, and A.J. Berkhout, "Design of a broadside array for a binaural hearing aid," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk NY, USA, Oct., 1997.
64. J.G. Desloge, W.M. Rabinowitz, and P.M. Zurek, "Microphone-array hearing aids with binaural output. I. Fixed-processing systems," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 529–542, 1997.
65. F.L. Wightman and D.J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, pp. 1648–1661, 1991.

66. D.P. Welker, J.E. Greenberg, J.G. Desloge, and P.M. Zurek, 'Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 543–551, 1997.
67. P.M. Zurek and J.E. Greenberg, "Two-microphone adaptive array hearing aids with monaural and binaural outputs," in *Proc. Ninth IEEE DSP Workshop*, Hunt TX, USA, Oct. 2000.
68. T.Y.C. Ching, H. Dillon, and D. Byrne, "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," *J. Acoust. Soc. Am.*, vol. 103, pp. 1128–1140, 1998.
69. C.A. Hogan and C.W. Turner, "High-frequency audibility: Benefits for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 104, pp. 432–441, 1998.

12 Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction

Rainer Martin

Institute of Communication Systems and Data Processing
Aachen University of Technology, Aachen, Germany

Abstract. This chapter presents arrays with two microphones and a postfilter for noise reduction and acoustic echo cancellation. The postfilter algorithm exploits the spatial coherence of the microphone signals. In contrast to single-microphone enhancement algorithms, it does not need an explicit noise power spectral density estimate. An analysis of the mean square error reveals that the coherence properties of the microphone signals are of paramount importance for the performance of the postfilter. Coherence measurements of signals in various acoustic environments are presented. The influence of the directivity and orientation of the microphones on the measured coherence is discussed and rules for the design of the acoustic interface are given. Finally, applications of this approach are presented. The two-microphone algorithm is employed to reduce the non-stationary noise in the voice intercom of a computed tomography scanner. It is also combined with echo cancellers to be used in a robust desktop conferencing device.

12.1 Introduction

Hands-free operation of voice communication terminals presents a challenging signal processing task. The relatively large distance between speaker and microphones, the feedback of acoustic echoes, and the "anywhere and anytime" paradigm of mobile communications can all contribute to considerably disturbed speech signals. To achieve a reasonable communication quality the hands-free terminal must therefore reduce disturbing environmental acoustic noise as well as acoustic echoes in received speech signals.

Because of their ease of implementation and use, single-microphone speech enhancement systems are favored in many applications. However, multi-microphone systems have considerable advantages over single-microphone systems when the noise is non-stationary or the speech is reverberated. Multi-microphone systems take the spatial correlation of sound fields into account. The spatial correlation can be exploited to dereverberate the desired speech signal and to reduce noise and acoustic echoes. The simplest system which takes advantage of some of these benefits is the two-microphone array.

In contrast to larger arrays, the two-microphone approach relies less on the beamforming gain of the array but more on the noise and echo suppression of a postfilter. The postfilter combines and processes the two microphone signals in order to compute an estimate of the clean speech signal. The array can be easily implemented using widely available stereo A/D converters.

Figure 12.1 depicts the basic components of the two-microphone speech enhancement system. The microphone signals are assumed to be a linear combination of clean speech signals $s_n(t)$ and noise signals $v_n(t)$, where $n \in \{1, 2\}$ denotes the microphone index. The system is symmetric, i.e., we assume that both microphones pick up speech and noise alike. This is quite different from the noise cancellation algorithm [1], where the primary microphone picks up both speech and noise and the secondary microphone serves as a noise reference only. It has been demonstrated that the noise cancellation approach does not work well in reverberant environments [2]. According to Fig. 12.1, the sampled microphone signals $x_1[k]$ and $x_2[k]$ (sampling frequency f_s) are adjusted for possible time delay differences in the range of $-T \leq \tau \leq T$, where T denotes the maximum delay difference. The signals are then combined and filtered. The filtering takes the spatial correlation (coherence) of the microphone signals into account and constructs an estimate $\hat{s}[k]$ of the clean speech for instance by minimizing a mean square error criterion. In contrast to most single-microphone algorithms, the coherence based approach does not rely on an explicit noise power spectral density estimate. Its performance, however, depends to a large extent on the acoustics of the environment.

Two-microphone arrays with coherence based postfilters were pioneered in [3,4] and later improved in [5–8]. Systems with more microphones and with significantly higher array gains were investigated in e.g. [9,10] and in [11]. The latter approach also makes use of subarrays and subband processing.

In this contribution we will first present the magnitude squared coherence function (MSC) as a tool for the analysis of the spatial correlation of the microphone signals. Using the Wiener filter as an example, we will then motivate in Section 12.3, why the spatial coherence of speech and noise signals is important for the performance of the two-microphone postfilter. Finally, we will describe applications of the two-microphone postfilter in the voice intercom of a computed tomography scanner and in desktop conferencing experiments.

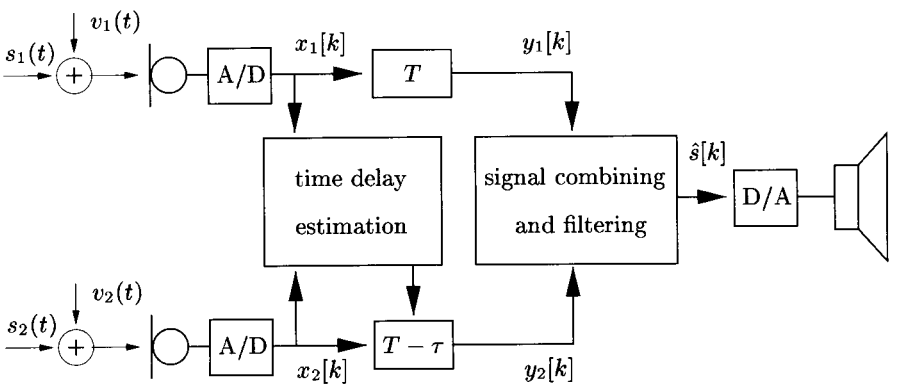


Fig. 12.1. Block diagram of the symmetric two-microphone noise reduction system.

12.2 Coherence of Speech and Noise

The MSC, $C_{x_1x_2}(\Omega)$, is a frequency domain measure of correlation between two signals. As it turns out, it is also a powerful tool for analyzing the potential and the performance of multi-microphone noise reduction systems. In this section we first define the MSC and the *reverberation distance*, another measure which helps to characterize the acoustic environment. We then present coherence measurements for various noise types and for speech. We will find that the directivity of the microphone and its orientation towards the speaker play an important role. The aim of this section is to derive rules for the design of the acoustic interface of the two-microphone noise reduction system.

12.2.1 The Magnitude Squared Coherence

For stationary input signals, $x_1[k]$ and $x_2[k]$, the MSC is defined as the ratio of the magnitude squared cross power spectral density, $P_{x_1x_2}(\Omega)$, to the power spectral densities, $P_{x_1x_1}(\Omega)$ and $P_{x_2x_2}(\Omega)$, of the input signals [12,13]

$$C_{x_1x_2}(\Omega) = \frac{|P_{x_1x_2}(\Omega)|^2}{P_{x_1x_1}(\Omega)\, P_{x_2x_2}(\Omega)}, \tag{12.1}$$

where Ω denotes a normalized frequency variable, $\Omega = 2\pi f/f_s$. The MSC takes on values between zero and one, $0 \leq C_{x_1x_2}(\Omega) \leq 1$.

It is well known that the coherence of two bandlimited and sampled signals recorded with omnidirectional microphones in an ideally diffuse (isotropic) sound field is given by [14,15]

$$C_{\text{diffuse}}(\Omega) = \frac{\sin^2\left(\Omega\, f_s\, d_{\text{mic}}\, c^{-1}\right)}{\left(\Omega\, f_s\, d_{\text{mic}}\, c^{-1}\right)^2}, \tag{12.2}$$

where d_{mic} denotes the distance between the microphones and c the speed of sound. $C_{\text{diffuse}}(\Omega)$ attains its first zero at frequency $f_0(d_{\text{mic}}) = c/(2d_{\text{mic}})$. The sound field is highly correlated for frequencies below $f_0(d_{\text{mic}})$ while the correlation is low for frequencies above $f_0(d_{\text{mic}})$. Equation (12.2) is a necessary but not a sufficient condition for a sound field to be ideally diffuse. Hence, it is possible to construct sound fields which have an MSC according to (12.2) and which are not ideally diffuse [16].

Besides the spatial distribution of the sound sources and the room acoustics the directivity of the microphones also has an impact on the measured coherence. For omnidirectional microphones the coherence of the ideally diffuse sound field is given by (12.2) independent of the microphone orientation. For directional microphones the coherence depends on the orientation of the microphones to each other. Figure 12.2 plots the coherence of the ideally diffuse sound field for microphones with a cardioid directivity pattern.

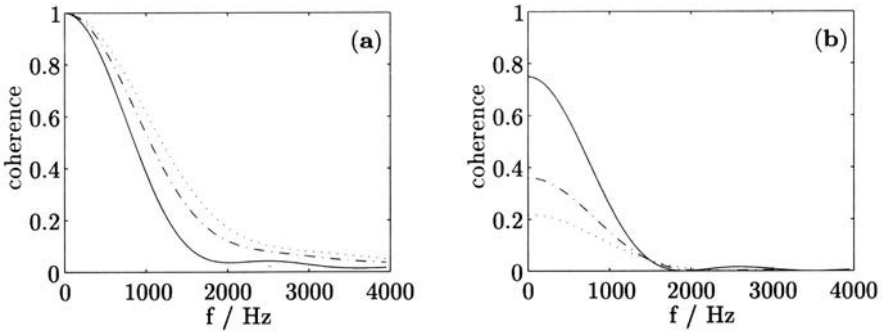


Fig. 12.2. Magnitude squared coherence of two microphone signals in an ideally diffuse noise field [17]. The microphones have a cardioid directivity pattern and different look directions. The microphone distance is 0.1 m. 0° equals broadside orientation.
(**a**): both microphones are turned clockwise by the same angle;
(**b**): one microphone is turned clockwise, the other counterclockwise.
(——): 30° turn, (– · –): 60° turn, (· · · · ·): 90° turn.

In Figure 12.2a the microphones were turned from a broadside orientation (0°) clockwise by the same angle. In Figure 12.2b one microphone is turned clockwise, the other is turned counterclockwise by the same angle. In the 90° position the directions of maximal sensitivity face each other. When the microphones look in different directions, the MSC at low frequencies is significantly reduced [17].

12.2.2 The Reverberation Distance

The coherence of the microphone signals depends on the amount of spatially uncorrelated sound energy and thus also on the amount of reverberated sound within these signals. The *reverberation distance* [14], r_H, can be used to characterize the ratio of direct sound energy to reverberated sound energy. When a sound source radiates sound equally under free field conditions in all spatial directions, the energy density at the distance

$$r_H = \sqrt{\frac{\overline{\alpha}_{Ab}\, A}{16\pi}} \approx 0.1\,\mathrm{m}\sqrt{\frac{V/\mathrm{m}^3}{\pi\, T_{60}/\mathrm{s}}} \tag{12.3}$$

has the same magnitude as the steady-state energy density which is obtained when the same sound power is radiated in a reverberant enclosure. $\overline{\alpha}_{Ab}$ denotes the absorption coefficient averaged over all walls of the enclosure. A is the area of all of these walls, V is the volume of the enclosure, and T_{60} is the reverberation time. The approximation on the right hand side of (12.3) is obtained when the reverberation time is computed using *Sabine's* equation [14], which holds when $\overline{\alpha}_{Ab}$ is small compared to unity. For an office

room with $V = 100\,\mathrm{m}^3$ and a reverberation time of 0.7 s the reverberation distance is about $r_H \approx 0.67\,\mathrm{m}$. The direct sound energy outweighs the reverberated sound energy when the receiver is within a sphere with radius r_H. The portion of direct sound energy in the microphone signals is significantly increased if the sound source and the sound receiver have a pronounced directivity. For microphones with a hypercardioid sensitivity pattern the effective reverberation distance is approximately twice as large as the reverberation distance of omnidirectional receivers.

12.2.3 Coherence of Noise and Speech in Reverberant Enclosures

In this section we present and discuss coherence measurements for noise and speech signals. As we will see in Section 12.3, the two-microphone postfilter approach relies on a low coherence of noise signals and a high coherence of the desired speech signal.

Coherence of Office Noise Figure 12.3 shows the coherence of the ideally diffuse sound field (solid) and the measured coherence of noise in a reverberant office room (dotted). The noise in this room is generated by computer fans and hard disk drives. The microphones have an omnidirectional directivity pattern. We find that the width of the main maximum is well modeled by the coherence of the ideally diffuse sound field. To avoid coherent noise within the telephone bandwidth of $300 \leq f \leq 3400\,\mathrm{Hz}$, the microphone distance must be larger than $0.4\,m$.

Coherence of Car Noise Figure 12.4 plots the coherence of noise recorded in a car. In this case the microphones have a hypercardioid directivity pattern and were turned towards the driver by about 15 degrees. In accordance with Fig. 12.2 the application of directional microphones results in a significant reduction of the MSC at low frequencies.

Coherence of Speech In contrast to the spatially distributed noise sources of the typical, noisy environment (office or car), the near-end speaker can be modeled by a point source provided the microphones are located sufficiently far from the speaker's mouth. The transmission of the speech signal from the mouth of the speaker to the microphones can be then described by linear transfer functions.

The MSC as defined in (12.1) is invariant under linear transformations of the input signals [12]. The MSC of a single speaker in a noise-free, reverberant enclosure should be therefore close to one regardless of where the speaker is situated with respect to the microphones and regardless of the reverberation distance. However, in a practical application where the coherence must be estimated from finite signal segments the estimated coherence might

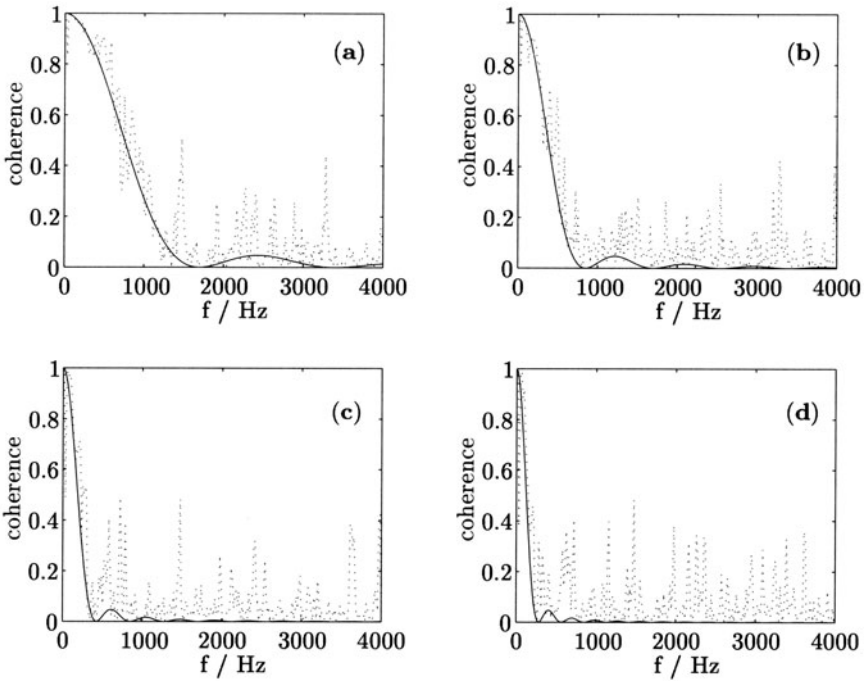


Fig. 12.3. Coherence of the ideally diffuse sound field (solid) and measured coherence (dotted) of office noise for omnidirectional microphones and microphone distances $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$ (**a**), $d_{\mathrm{mic}} = 0.2\,\mathrm{m}$ (**b**), $d_{\mathrm{mic}} = 0.4\,\mathrm{m}$ (**c**), and $d_{\mathrm{mic}} = 0.6\,\mathrm{m}$ (**d**).

be severely biased. If, for instance, the coherence is estimated by averaging magnitude squared DFT frames (periodograms) the coherence estimate of reverberated speech is biased towards zero. The bias depends on the ratio of the block length of the DFT to the length of the impulse response of the acoustic path, and on the distribution of energy in the impulse response [17].

Figure 12.5 plots the estimated coherence of a speech signal uttered by a speaker in a reverberant room ($T_{60} = 0.7\,\mathrm{s}$) for omnidirectional and for hypercardioid microphones and three distances from speaker to microphones. To estimate the coherence, the speech signals were segmented into frames of 128 signal samples at a sampling rate of $f_s = 8$ kHz. To improve the rendering of the coherence plots, the signal frames were zero-padded to a DFT frame length of 512 samples. A short term coherence estimate was then computed on the basis of short time averaged periodograms and the final coherence estimate by long term averaging the short term coherence estimates. Taking the directivity of a human speaker into account, the effective reverberation distance of this setup is about 0.9 m for omnidirectional microphones. For

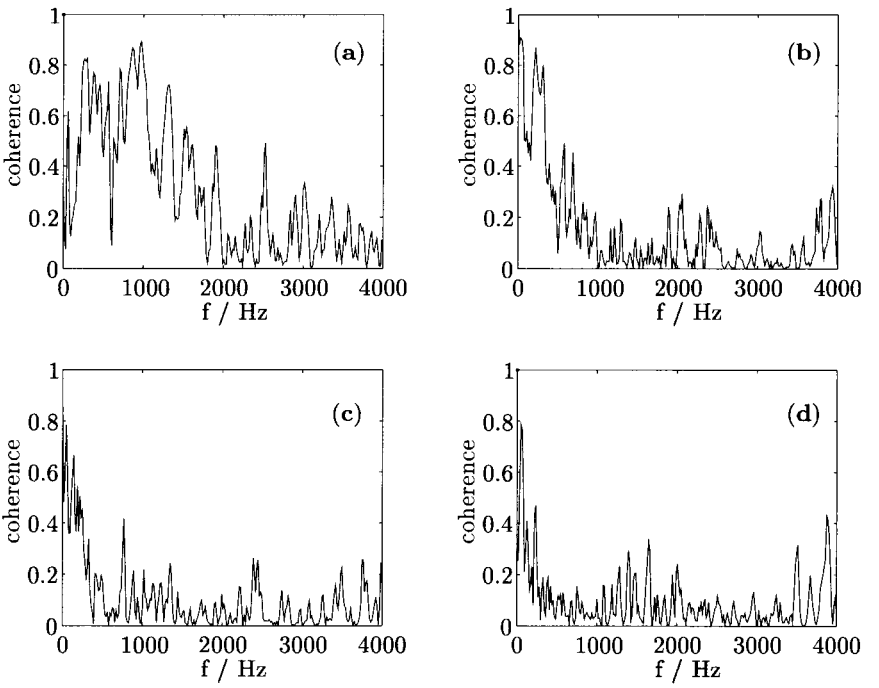


Fig. 12.4. Measured coherence of noise in a car for hypercardioid microphones and microphone distances $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$ (**a**), $d_{\mathrm{mic}} = 0.2\,\mathrm{m}$ (**b**), $d_{\mathrm{mic}} = 0.4\,\mathrm{m}$ (**c**), and $d_{\mathrm{mic}} = 0.6\,\mathrm{m}$ (**d**).

the hypercardioid microphones the effective reverberation distance is about 1.6 m. We find that even when the speaker is well within the reverberation distance the coherence is significantly below unity. This is a result of the relatively large reverberation time and the small frame size of the coherence estimation procedure. The estimated coherence is less biased in environments with shorter reverberation times, e.g., in a car. Nevertheless, it is important that the speaker is located well within the reverberation distance since the coherence will be additionally reduced by incoherent ambient noise.

Similar results are obtained when a speech signal is radiated from a (small) loudspeaker of a hands-free conferencing terminal into the acoustic environment. Since the feedback of speech echoes via the noise reduction system to the far-end side is not desired, the coherence of these speech echoes should be low. They can be then treated in the same way as ambient noise. The coherence of the microphone signals in the presence of speech echoes depends to a large extent on the placement of the loudspeaker with respect to the microphones, on the directional pattern of the microphones, and the room acoustics. To increase the robustness of a hands-free conferencing terminal,

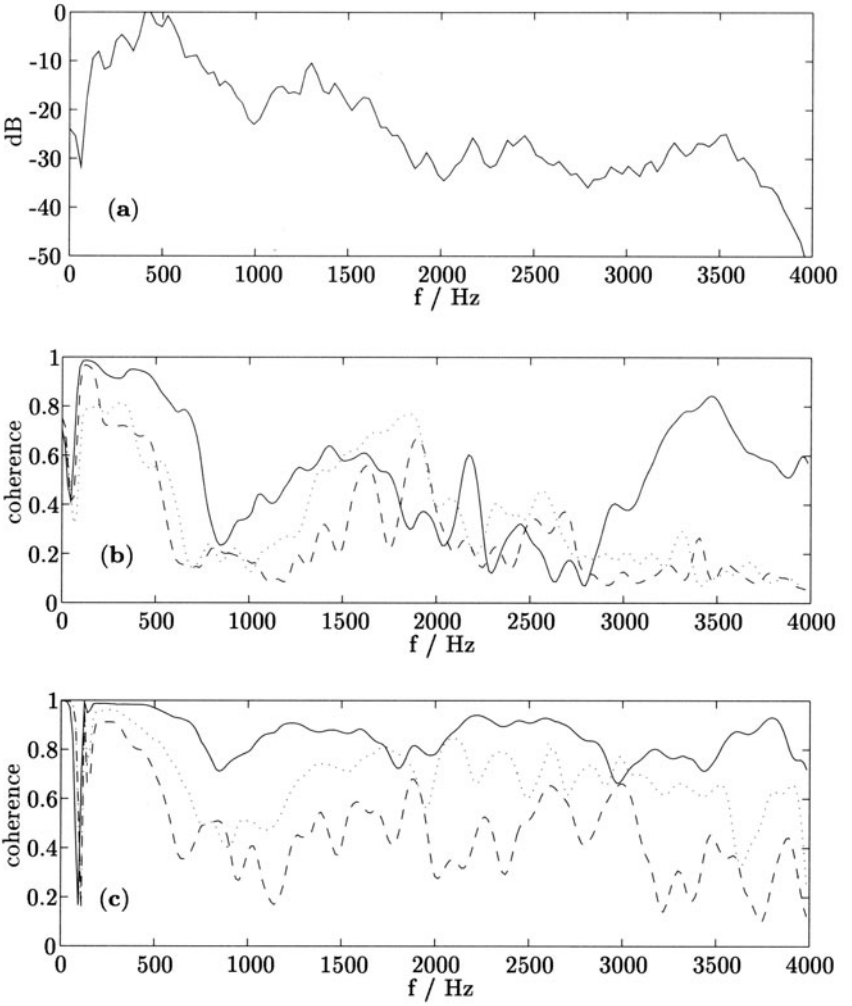


Fig. 12.5. Long term speech spectrum (**a**) and long term average of the MSC of a speech signal in an office room for omnidirectional (**b**) and for hypercardioid (**c**) microphones and various distances ((——): 0.5 m; (·····): 1 m; (– – –): 2 m) from speaker to microphones. The distance between microphones is 0.4 m.

the coupling between the loudspeaker and the microphones and the coherence of the speech echoes must be minimized. This can be, for instance, achieved by using directional microphones and by placing the loudspeaker in the direction of minimum microphone sensitivity.

12.3 Analysis of the Wiener Filter with Symmetric Input Signals

In this section we compute the mean square error of the two-microphone adaptive algorithm with symmetric input signals when the filter is adapted by means of the unconstrained Wiener filter [18]. In contrast to other Wiener filter optimization approaches, we do not use the undisturbed desired speech signal as a reference in the derivation of the optimal filter. We show that the performance of the system can be characterized using the magnitude squared coherence (MSC) function of the microphone signals. The analysis supplements the coherence measurements of the previous section.

For the computation of the linear MMSE filter and its mean square error we consider an IIR filter as shown in Figure 12.6a and assume that the input signals $x_n[k] = s_n[k] + v_n[k]$, $n \in \{1, 2\}$, are the sum of clean speech signals $s_n[k]$ and noise signals $v_n[k]$. In Figure 12.6a, $x_2[k] = s_2[k]+v_2[k]$ is the input signal of the adaptive filter while $x_1[k] = s_1[k] + v_1[k]$ serves as a reference signal. Since both microphones pick up speech and noise alike, the reference signal contains not only (reverberated) speech but also noise.

The Wiener filter in Figure 12.6a minimizes the mean square error $E\{(x_1[k] - \hat{s}[k])^2\}$, where $\hat{s}[k]$ is computed using the non-causal IIR filter. When the speech and the noise signals are statistically independent, the frequency response of the Wiener solution is given by [18]

$$H_W(\Omega) = \frac{P_{x1x2}(\Omega)}{P_{x2x2}(\Omega)} = \frac{P_{s1s2}(\Omega) + P_{v1v2}(\Omega)}{P_{s2s2}(\Omega) + P_{v2v2(\Omega)}} \tag{12.4}$$

$$= \frac{P_{s1s2}(\Omega)}{P_{s2s2}(\Omega) + P_{v2v2}(\Omega)} + \frac{P_{v1v2}(\Omega)}{P_{s2s2}(\Omega) + P_{v2v2}(\Omega)}, \tag{12.5}$$

which is recognized as a linear combination of two optimal subfilters, see Figure 12.6b. $H_{\text{sopt}}(\Omega) = P_{s1s2}(\Omega)/P_{x2x2}(\Omega)$ and $H_{\text{vopt}}(\Omega) = P_{v1v2}(\Omega)/P_{x2x2}(\Omega)$ are linear MMSE estimators for the speech component $s_1[k]$ and the (undesired) noise component $v_1[k]$ of the reference signal $x_1[k]$, respectively. The estimation error between the desired speech signal $s_1[k]$ and the output of the Wiener filter $\hat{s}[k]$ can therefore be written in terms of the estimation errors of the subfilters

$$e_r[k] = s_1[k] - \hat{s}[k] = \left(s_1[k] - \hat{\hat{s}}_1[k]\right) - \hat{\hat{v}}_1[k]\,. \tag{12.6}$$

The overall minimum mean square error can be then decomposed into

$$E\left\{e_r^2[k]\right\} = E\left\{\left(s_1[k] - \hat{s}[k]\right)^2\right\} \tag{12.7}$$

$$= E\left\{\left(s_1[k] - \hat{\hat{s}}_1[k]\right) s_1[k]\right\} + E\left\{v_1[k]\hat{\hat{v}}_1[k]\right\} \tag{12.8}$$

$$= E\left\{v_1^2[k]\right\} + E\left\{\left(s_1[k] - \hat{\hat{s}}_1[k]\right) s_1[k]\right\} - E\left\{\left(v_1[k] - \hat{\hat{v}}_1[k]\right) v_1[k]\right\} \tag{12.9}$$

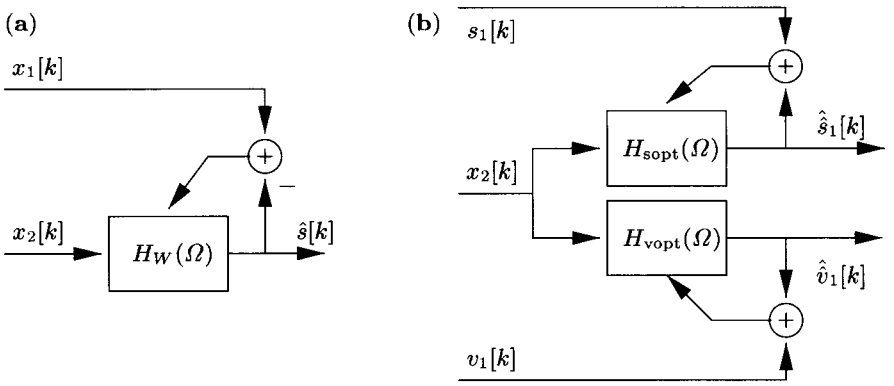


Fig. 12.6. The Wiener filter (**a**) and its subfilter decomposition (**b**).

where we used $E\left\{\widehat{\widehat{v}}_1^2[k]\right\} = E\left\{v_1[k]\widehat{\widehat{v}}_1[k]\right\}$ which holds for the MMSE filter.

Using Parceval's relation, the minimum mean square errors of the unconstrained subfilters $H_{\text{sopt}}(\Omega)$ and $H_{\text{vopt}}(\Omega)$ can be rewritten in the frequency domain as [18]

$$E\left\{\left(v_1[k] - \widehat{\widehat{v}}_1[k]\right) v_1[k]\right\}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1v_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{\text{vopt}}(\Omega)\, P^*_{v_1v_2}(\Omega)\, d\Omega \quad (12.10)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1v_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1v_2}(\Omega)|^2}{P_{s_2s_2}(\Omega) + P_{v_2v_2}(\Omega)} d\Omega \quad (12.11)$$

and

$$E\left\{\left(s_1[k] - \widehat{\widehat{s}}_1[k]\right) s_1[k]\right\}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1s_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{\text{sopt}}(\Omega)\, P^*_{s_1s_2}(\Omega)\, d\Omega \quad (12.12)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1s_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{s_1s_2}(\Omega)|^2}{P_{s_2s_2}(\Omega) + P_{v_2v_2}(\Omega)} d\Omega \quad (12.13)$$

where $P_{xy}(\Omega)$ denotes the (cross) power spectral density of the signals in the subscript.

Applying these results to (12.9), we obtain for the overall MMSE

$$E\left\{e_r^2[k]\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1 s_1}(\Omega)\, d\Omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1 v_2}(\Omega)|^2 - |P_{s_1 s_2}(\Omega)|^2}{P_{s_2 s_2}(\Omega) + P_{v_2 v_2}(\Omega)} d\Omega\,. \tag{12.14}$$

The representation of the MMSE in the frequency domain shows that a successful application (i.e. a small MMSE) of the two-microphone Wiener filter to the speech enhancement task requires a high correlation of the speech components $s_1[k]$ and $s_2[k]$ and a low correlation of the noise components $v_1[k]$ and $v_2[k]$. In the following sections we consider two special cases.

12.3.1 No Near End Speech

During speech pause, $s_1[k] \equiv s_2[k] \equiv 0$, the MMSE $E\left\{e_r^2[k]\right\}$ is given by

$$E\left\{e_r^2[k]\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1 v_2}(\Omega)|^2}{P_{v_2 v_2}(\Omega)} d\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1 v_1}(\Omega)\, C_{v_1 v_2}(\Omega)\, d\Omega\,. \tag{12.15}$$

The residual noise at the output of the optimal filter depends on the power spectral density $P_{v_1 v_1}(\Omega)$ of the noise and the coherence $C_{v_1 v_2}(\Omega)$ of the noise components.

12.3.2 High Signal to Noise Ratio

If the microphone signals have a high SNR the MMSE can be written as a function of the power spectral density of the speech signal and the coherence $C_{s_1 s_2}(\Omega)$ of the speech components. The approximations $P_{s_2 s_2}(\Omega) \gg P_{v_2 v_2}(\Omega)$ and $P_{s_2 s_2}(\Omega) \gg |P_{v_1 v_2}(\Omega)|^2$ lead to

$$E\left\{e_r^2[k]\right\} \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1 s_1}(\Omega)\,(1 - C_{s_1 s_2}(\Omega))\, d\Omega\,. \tag{12.16}$$

A prerequisite for high speech quality is thus a coherence of the speech components which is close to one. Incoherent speech components generated, for instance, by reverberation will be attenuated. Whether this attenuation constitutes an improvement or a reduction of the perceived speech quality depends on the ratio of coherent and incoherent speech sounds and the noise level. If the speech components are less coherent than the ambient noise the SNR will not be improved. A sufficient amount of coherent speech is therefore of paramount importance for a good performance of the coherence based two-microphone speech enhancement system. The coherence of speech signals can be improved by using directional microphones.

12.4 A Noise Reduction Application

In this section we describe a noise reduction algorithm which is based on the Wiener filter as discussed in Section 12.3. The algorithm uses a time domain implementation of the Wiener filter and was developed for the voice intercom of a computed tomography scanner.

The voice communication between a patient in a computed tomography (CT) scanner and the operator at the control desk is disturbed by acoustic noise which originates from the CT scanner. The acoustic noise in the gantry tunnel is due to numerous cooling fans and to the rotating x-ray imaging system. To reduce the fatigue of the operator, it is very desirable to reduce the level of noise transmitted from the gantry of the scanner to the control desk. Since the noise is highly non-stationary, single microphone speech enhancement methods do not perform well in this environment.

Figure 12.7 illustrates the application of the two-microphone noise reduction algorithm in the computed tomography scanner. The microphones which pick up the patient's speech are mounted inside the gantry tunnel at a distance of 0.4 m. The microphone signals are sampled and processed on a DSP. The enhanced signal $\hat{s}[k]$ is then played back on a loudspeaker at the control desk.

12.4.1 An Implementation Based on the NLMS Algorithm

Figure 12.8 shows a block diagram of the time domain implementation. The microphone signals are bandlimited to 3600 Hz and sampled with $f_s = 8000$ Hz. Preemphasis filters whiten the input signals and thus improve the convergence of the adaptive filter. They also help to improve the

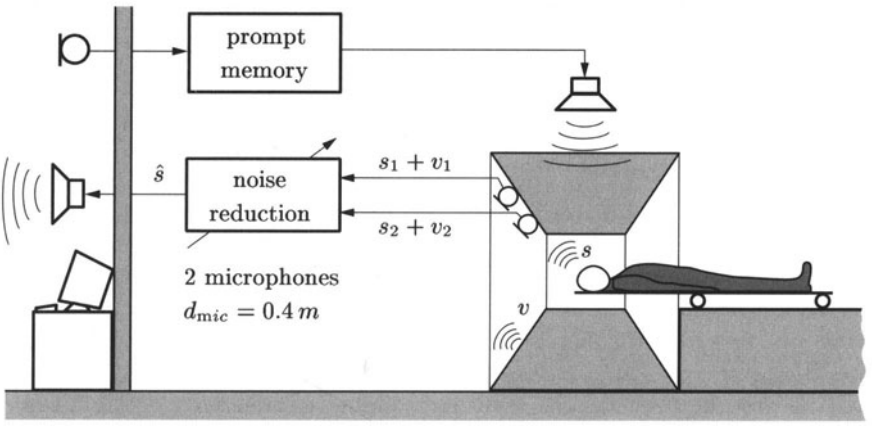


Fig. 12.7. Application of the two-microphone noise reduction system for voice communication in a computed tomography scanner.

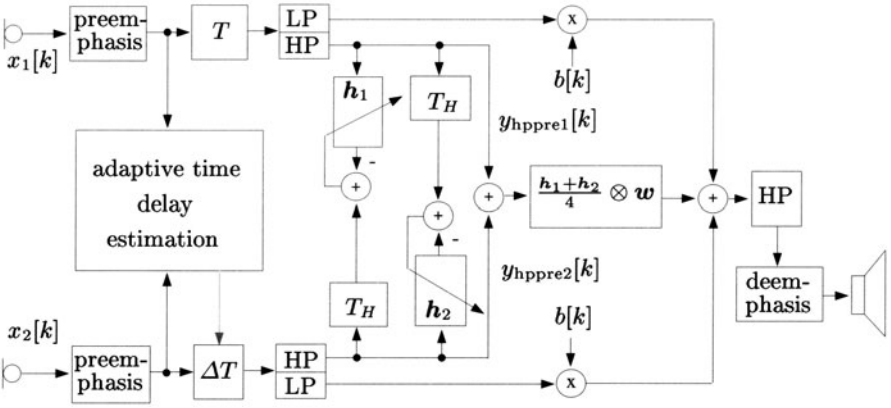


Fig. 12.8. Block diagram of the two-microphone noise reduction system. LP: lowpass filter, HP: highpass filter, $\Delta T = T - \tau$.

reproduction of high frequency speech components in a fixed-point implementation.

An adaptive time delay estimation algorithm compensates time delays between the two input signals. The time delay compensation is based on the correlation of the two microphone signals and on the SNR. When a high SNR is detected, a recursively smoothed correlation function is searched for maxima and the time delay is determined [17]. To avoid noticeable lowpass or comb-filtering effects, the magnitude of the delay error after delay compensation should be smaller than $1/(4f_s)$.

The spectral density and the coherence properties of the microphone signals suggest processing the speech signal in two frequency bands. Above 800 Hz, a linear phase adaptive Wiener filter is used which suppresses incoherent signal components (noise and reverberated speech) and passes highly coherent speech signals. Processing the frequency band below 800 Hz with this adaptive filter would result in noticeable fluctuations of the residual noise. These fluctuations are caused by the correlation of the noise signals at low frequencies. The noise in the band between 240 and 800 Hz is therefore suppressed by an adaptive scalar factor $b[k]$. This factor is controlled by the speech activity of the person inside the gantry tunnel. The speech activity is determined by the SNR estimator which is also used to increase the robustness of the time delay estimation algorithm [19]. The frequency band below 240 Hz is attenuated by 20 dB by means of a second order recursive highpass filter. The deemphasis filter at the output of the speech enhancement system restores the spectral characteristics of the speech signal. This noise reduction system is currently used in the Siemens SOMATOM PLUS 4 CT scanner[1].

[1] Siemens and SOMATOM PLUS 4 are registered trademarks of Siemens AG, Germany.

The computational complexity of the algorithm is about 10 MIPS on a 24 bit fixed-point DSP. In what follows we briefly describe the algorithm.

12.4.2 Processing in the 800 – 3600 Hz Band

The Wiener filter is approximated by two antiparallel linear phase NLMS adapted FIR filters (see Figure 12.8). The adaptive filters with the coefficient vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ of order $L_H = 64$ are updated using a linear phase version [20] of the NLMS algorithm ($T_H = L_H/2$),

$$\boldsymbol{h}_1[k+1] = \boldsymbol{h}_1[k] + \alpha e_1[k] \frac{\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\mathrm{hppre1}}[k]}{\boldsymbol{y}_{\mathrm{hppre1}}^T[k] \boldsymbol{y}_{\mathrm{hppre1}}[k]} \tag{12.17}$$

$$\boldsymbol{h}_2[k+1] = \boldsymbol{h}_2[k] + \alpha e_2[k] \frac{\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\mathrm{hppre2}}[k]}{\boldsymbol{y}_{\mathrm{hppre2}}^T[k] \boldsymbol{y}_{\mathrm{hppre2}}[k]}, \tag{12.18}$$

where $\boldsymbol{I}$ denotes the identity matrix and

$$\boldsymbol{I}^R = \begin{pmatrix} 0 & 0 & 0 & .. & & 1 \\ 0 & 0 & .. & & .. & \\ 0 & .. & & 1 & & \\ .. & & 0 & & .. & \\ & 1 & & & .. & 0 \\ & .. & & & .. & 0 \; 0 \\ 1 & & & .. & 0 \; 0 \; 0 & \end{pmatrix} \tag{12.19}$$

denotes a modified reflection matrix. The error signals $e_1[k]$ and $e_2[k]$ are given by

$$e_1[k] = y_{\mathrm{hppre2}}[k - T_H] - \boldsymbol{y}_{\mathrm{hppre1}}^T[k] \boldsymbol{h}_1[k] \tag{12.20}$$

$$e_2[k] = y_{\mathrm{hppre1}}[k - T_H] - \boldsymbol{y}_{\mathrm{hppre2}}^T[k] \boldsymbol{h}_2[k], \tag{12.21}$$

and

$$\boldsymbol{y}_{\mathrm{hppre1}}[k] = \left(y_{\mathrm{hppre1}}[k], \; ..., y_{\mathrm{hppre1}}[k - L_H]\right)^T \tag{12.22}$$

$$\boldsymbol{y}_{\mathrm{hppre2}}[k] = \left(y_{\mathrm{hppre2}}[k], \; ..., y_{\mathrm{hppre2}}[k - L_H]\right)^T \tag{12.23}$$

denote the data vectors of the filter input signals. $\alpha \approx 0.1$ is the stepsize parameter of the NLMS algorithm. Because of the symmetry of the coefficient vector updates $\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\mathrm{hppre1}}[k]$ and $\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\mathrm{hppre2}}[k]$ and a symmetric initialization, the coefficient vectors $\boldsymbol{h}_1[k+1]$ and $\boldsymbol{h}_2[k+1]$ are symmetric for all k. Therefore, only the first half of the vectors need to be adapted.

To filter the combined input signals of the upper band, $(y_{\mathrm{hppre1}}[k] + y_{\mathrm{hppre2}}[k])/2$, we use the mean of the two adaptive coefficient vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ and an additional smoothing window $\boldsymbol{w} = (w_0, w_1, \; ..., w_{L_H})^T$

$$\boldsymbol{h}[k] = \frac{\boldsymbol{h}_1[k] + \boldsymbol{h}_2[k]}{2} \otimes \boldsymbol{w}. \tag{12.24}$$

The symbol $\otimes$ denotes the pointwise multiplication of two vectors. The window function is used to smooth the frequency response of the adaptive filter. A Kaiser window [21] with a shape parameter β_{Kaiser} in the range of $3 \leq \beta_{Kaiser} \leq 5$ results in good speech quality and increased noise reduction. Since the coefficient vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ represent linear phase filters, the averaging of the vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ in (12.24) yields an average of the amplitude spectra of these filters without errors due to mismatching phase spectra.

12.4.3 Processing in the 240 – 800 Hz Band

The attenuation factor $b[k]$ ($b_{min} \leq b[k] \leq b_{max}$) is controlled by a speech activity detector and adjusted according to the estimated SNR. Whenever the estimated SNR is below a preselected threshold the attenuation is slowly and successively increased until a maximum attenuation of 40 dB (corresponding to $b_{min} = 0.01$) is reached. Whenever the estimated SNR is above the threshold the attenuation is rapidly decreased to a minimum value of 3 dB ($b_{max} = 0.5$). The SNR threshold is set to 3 dB. Thus, the attenuation factor $b[k]$ is computed using the recursive system

$$b[k+1] = b[k]\beta_1 + b_{max}\,(1-\beta_1)\,,\ SNR > \text{threshold} \tag{12.25}$$

$$b[k+1] = b[k]\beta_2 + b_{min}\,(1-\beta_2)\,,\ SNR \leq \text{threshold}\,. \tag{12.26}$$

The smoothing constants β_1 and β_2 are set to $\beta_1 = 0.9996$ and $\beta_2 = 0.99999$.

12.4.4 Evaluation

We assess the performance of the noise reduction algorithm in terms of

- distortion of the speech signal;
- noise reduction during speech activity;
- noise reduction during speech pause.

These criteria can be measured during simulation of the speech enhancement system. For the purpose of measuring the above properties, the adaptive filter is duplicated such that the undisturbed speech signal and the noise signal can be processed independently [22]. Figure 12.9a outlines this approach. It requires separate recordings of the noise and the speech signals.

The speech signal distortion can then be measured as the segmental SNR of the filtered signal $\widetilde{s}[k]$ with respect to the unprocessed delayed speech signal $s[k - T_H]$

$$SEGSNR^{s}_{\widetilde{s}-s} = \frac{1}{K}\sum_{m=0}^{K-1} \max\left(SNR^{s}_{\widetilde{s}-s}\,(m)\,,0\right) \tag{12.27}$$

with

$$SNR^{s}_{\tilde{s}-s}(m) = 10 \cdot \log_{10} \left(\frac{\sum_{k=mM}^{mM+M-1} s^2[k - T_H]}{\sum_{k=mM}^{mM+M-1} (\tilde{s}[k] - s[k - T_H])^2} \right) . \tag{12.28}$$

M denotes the segment length and K the number of segments. To reduce the influence of speech pauses, we average the SNR of only those speech signal frames which exhibit an SNR larger than 0 dB. Since we use a linear phase filter, the segmental SNR measures the amplitude distortion of the speech signal and is therefore well correlated with perceived distortions.

The attenuation of the noise signals during speech activity NR_{active} and during speech pause NR_{pause} is measured as the power ratio of the noise signals before and after the adaptive filter

$$NR_{active} = 10 \log_{10} \left(\overline{P_v[k - T_H]} / \overline{P_{\tilde{v}}[k]} \right) , \ \overline{P_s}[k] \neq 0 \tag{12.29}$$

$$NR_{pause} = 10 \log_{10} \left(\overline{P_v[k - T_H]} / \overline{P_{\tilde{v}}[k]} \right) , \ \overline{P_s}[k] = 0 , \tag{12.30}$$

where $\overline{P_v[k]}$ and $\overline{P_{\tilde{v}}[k]}$ denote the average power of the unprocessed and the processed noise signal, respectively. $\overline{P_s}[k]$ denotes the short term power of the speech signal.

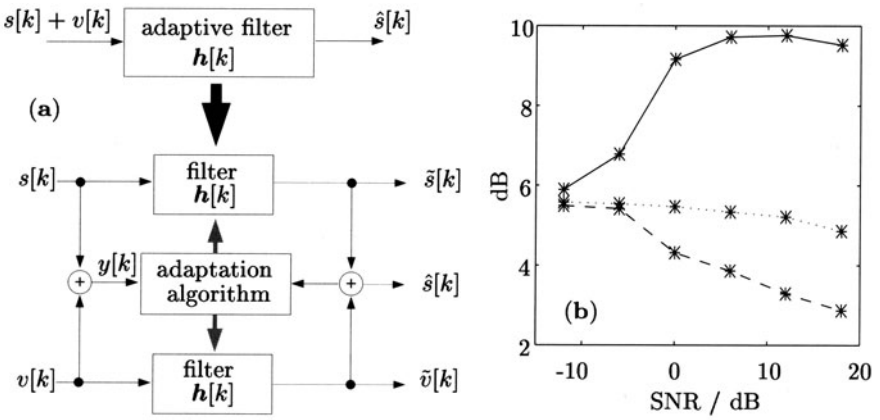


Fig. 12.9. Method for computing objective measures (**a**). Distortion of speech signal and noise reduction during speech activity and during speech pause vs. input SNR of the adaptive filter in the upper band (**b**). Step size: $\alpha = 0.1$, filter order $L_H = 64$.
(——): Segmental SNR of speech; (· · · · ·): noise attenuation during speech pause; (– – –): noise reduction during speech activity.

Figure 12.9b plots the objective measures as a function of the input SNR. For this experiments, adaptive filters of order $L_H = 64$ and a rectangular window $\boldsymbol{w}$ were used (equivalent to $\beta_{Kaiser} = 0$). It can be seen that the speech signal distortion increases as the input SNR decreases. When the incoherent noise becomes dominant at about 0 dB, the speech SNR significantly degrades. The noise reduction during speech pause is about 5 – 6 dB. It can be improved by 2 – 3 dB if a tapered smoothing window is used. The overall noise reduction during speech pause including the adaptive scalar weighting in the lower band and the highpass filter is about 14 dB during speech pause.

12.4.5 Alternative Implementations of the Coherence Based Postfilter

The speech enhancement system as outlined above exploits the coherence properties of the microphone signals. Coherence based noise reduction systems can be also implemented in the frequency domain [3,9,5,6]. Since the coherence based approach does not rely on an explicit noise power spectral density estimate, the performance of these systems is limited by the coherence of the noise and the speech signals. To improve the noise reduction especially for low frequencies, the combination with spectral weighting techniques has been proposed. In [5] the coherence function is also employed to detect speech pauses and to enable noise power spectral estimation during speech pauses. In [23] the cross power spectral density of the microphone signals is used to derive an explicit noise power spectral density estimate, which is then used in a two-channel spectral subtraction. Combined with small superdirective endfire arrays, this system led to significant improvements of speech intelligibility in conjunction with cochlear implants [24].

12.5 Combined Noise and Acoustic Echo Reduction

A hands-free conferencing system has to cope not only with ambient noise but also with acoustic echoes. The feedback of the far-end speech signal via the loudspeaker, the room, and the microphone (the "LRM system") back to the far-end side necessitates an echo suppression device to guarantee the stability of the electro-acoustic loop and to supply sufficient echo reduction. While the stability of the electro-acoustic loop can be treated as a control problem, the echo reduction aims at making the echo imperceptible. The echo reduction problem is therefore closely linked to psychoacoustics, especially to masking effects in the human auditory system.

The noise and the echo reduction problems were addressed independently for many years (see e.g. [25–28] and [29,30] for reviews of these methods). To achieve optimal performance it has been recognized, however, that the echo control and noise reduction problem should be tackled in a combined approach [31,8,32–34]. The combined treatment yields algorithms which deliver better performance at less computational costs than systems based on

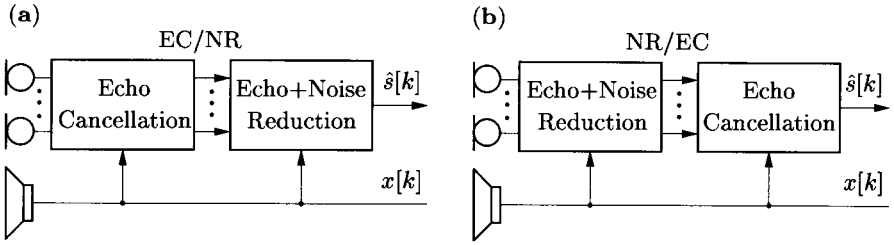


Fig. 12.10. Processing options for combined systems: EC precedes ENR (**a**), ENR precedes EC (**b**).

separate algorithms [8,35]. In this section we will outline one solution to the combined echo and noise reduction problem and show how the coherence based noise reduction postfilter can be extended to achieve a high level of echo reduction. This requires the integration of echo cancellers into the two-microphone speech enhancement system.

When acoustic echo cancellation is combined with an echo and noise reduction postfilter it must be asked in which order these two processing operations should be performed [36]. Figures 12.10a and 12.10b depict two principal cases: the configuration EC/NR where the acoustic echo cancellation (EC) precedes a speech enhancement postfilter (NR), and vice versa, the configuration NR/EC.

Although the echo canceller can benefit from the noise reduction in the NR/EC configuration, there are good reasons why the configuration of Figure 12.10a, where the echo cancellation precedes the noise reduction, is preferable. The main advantage of the EC/NR configuration is that the noise reduction postfilter is not presented with the disturbing and possibly highly coherent echo that is found in the microphone signals and that there is no time varying noise reduction filter in the echo path. Besides that, if the echo canceller does not deliver sufficient echo attenuation, the residual echo can be treated similar to the background noise signal and can be further attenuated by the postfilter. This idea is successfully exploited in a frequency selective echo reduction technique, called *echo shaping* [37], which does not require complete cancellation of the echo by the echo canceller. Instead, the total echo attenuation is split between the echo cancellers and the postfilter. A disadvantage of the EC/NR approach is, however, that it requires cancellation of the microphone signals of the array by individual cancellers.

The combination of acoustic echo cancellation with an adaptive microphone array is a challenging task by itself [38–40]. If we provide a canceller for each microphone channel, the echo cancellers converge as well as in the single microphone case. However, the computational load may be too large. If a single echo canceller is placed after the summation point of the array, the adaptation of the echo canceller might be severely disturbed when the look direction of the array is adapted to the speaker position.

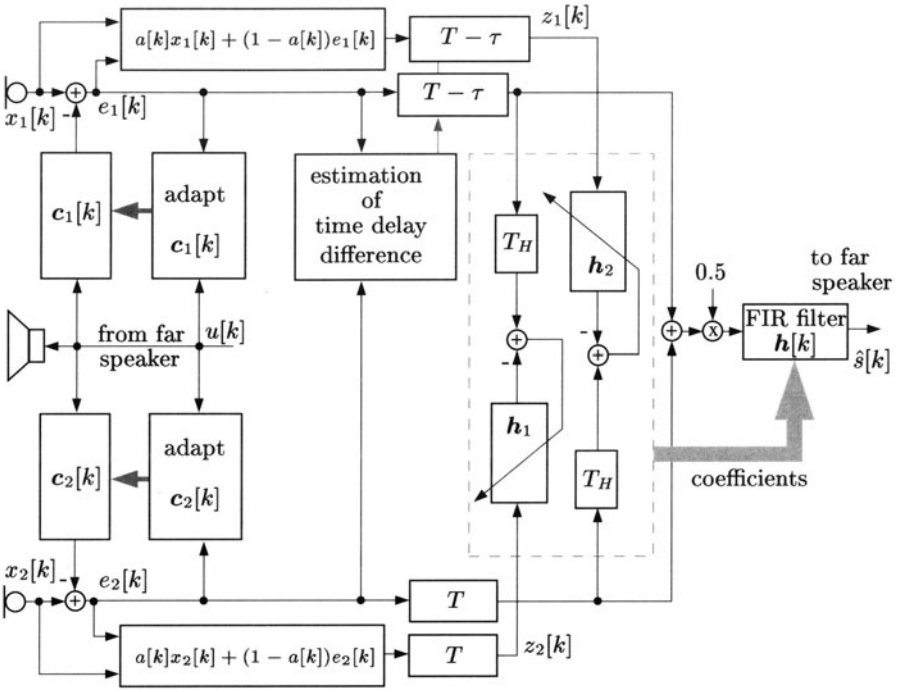


Fig. 12.11. Block diagram of the two microphone combined echo and noise reduction algorithm.

To avoid the adaptation and the complexity problems, we adopt a strategy as follows: The echo cancellers are placed at the microphone inputs but they are equipped with a reduced number of filter taps. The echo attenuation of the cancellers will be reduced but since they have fewer filter taps their speed of convergence will be improved. The reduced echo attenuation of the echo cancellers is compensated for by the array gain [41] and the noise reduction postfilter which will also reduce acoustic echoes, especially the incoherent late reverberation portion. To improve the echo suppression, we apply additional echo attenuation in the postfilter by using the *echo shaping* technique for the postfilter adaptation. The *echo shaping* technique attenuates only those frequencies for which the echo dominates the near-end signal. It leads to significantly increased echo attenuation, to acceptable near-end signal distortions during double talk, and to modest computational demands as compared to single-microphone speech enhancement systems.

Figure 12.11 depicts a block diagram of the combined echo and noise reduction algorithm. The cancellers at the microphone inputs are NLMS adapted FIR filters with adaptive stepsize control [42]. The echo cancelled signals are time delay compensated. Because some of the echo is incoherent, the noise reduction postfilter will also reduce acoustic echoes. To amplify this

effect and to reduce coherent echoes, the amount of echo in the inputs of the adaptive filters is deliberately increased by using a linear combination of the microphone signal $x_n[k]$ and the echo compensated signal $e_n[k]$ as an input to the adaptive filters. The reference signal of the adaptive filters is the echo compensated signal of the other microphone channel. Neglecting the time delay compensation (or letting $\tau = T$), the input to the adaptive filter $\boldsymbol{h}_2$ is given by

$$z_1[k] = a[k]x_1[k] + (1 - a[k])e_1[k] \tag{12.31}$$

$$= s_1[k] + v_1[k] + d_1[k] - (1 - a[k])\hat{d}_1[k] \,, \tag{12.32}$$

where $d_1[k]$ denotes the echo signal and $\hat{d}_1[k]$ the echo estimate of the echo canceller. The linear combination is controlled by the time varying factor $a[k]$. For $a[k] > 0$ there will be more echo in the input signal of the adaptive filter than in the reference signal since the reference signal is the echo compensated signal. To match the echo level of the reference signal, the adaptive filters will attenuate the echo if it dominates the near-end signals. Therefore, this mechanism provides for additional echo attenuation whenever the echo is disturbing. An algorithm for the adaptation of the "mixing factor" $a[k]$ is outlined in [17].

12.5.1 Experimental Results

The two-microphone algorithm as explained above was evaluated in a desktop conferencing experiment. Since we aim for low coherence of the acoustic echoes and high inherent robustness, the acoustic interface was designed such that there is little coupling between the loudspeaker and the microphones. Also, to increase the performance of the system, directional microphones or small superdirective endfire arrays [43] should be used. Figure 12.12 explains the experimental setup. The near-end speaker and the microphones are placed at a table in accordance with ITU-T recommendation P.34 [44]. The microphones have a hypercardioid directivity pattern and the loudspeaker is placed in a direction of low sensitivity of the microphones. The specific arrangement of the microphones and the loudspeaker combined with the gains of the loudspeaker and the microphone amplifiers resulted in an echo return loss of 9 dB. A similar setup (with similar results) was also used in a car. In this case the two microphones were mounted at the sun visor and the loudspeaker was attached to the dashboard.

Figure 12.13a plots the echo return loss enhancement (ERLE) for *single talk* and a stationary LRM system as a function of the number of compensator taps. For $a[k] \equiv 0$, the postfilter reduces noise and incoherent echo components only. The additional echo reduction delivered by the postfilter is then about 6 dB, independent of the compensator order. A significant increase of the echo attenuation is achieved when the *echo shaping* algorithm is turned on ($a[k]$ adaptive). It can be shown with the noise-free case that

for $a[k] \equiv 1$ the postfilter delivers the same amount of echo suppression as the echo canceller alone [17]. Hence, the slope of the ERLE vs. compensator order plot is about twice as steep as the slope of the plot for the echo canceller only. Figure 12.13b plots the ERLE as a function of the SNR. Again we find a significant increase of the echo attenuation for the *echo shaping* technique. For *double talk* most of the echo reduction is delivered by the echo canceller since the coherent near-end speech is passed by the filter. Only for frequencies where the residual echo dominates the near-end signal is significant echo reduction applied. If the far-end speaker is not active (no echo), the combined system behaves exactly like the two-microphone noise reduction systems of Section 12.4.1. For the office situation the average noise reduction is about 10 dB.

12.6 Conclusions

This chapter has presented a two-microphone postfilter approach to noise reduction and to combined echo and noise reduction. It was shown that the performance of these systems is closely linked to the spatial coherence of the speech, the noise, and the echo signals. The coherence based processing is useful only above a cutoff frequency which depends on the microphone distance. However, as a high coherence of the speech signal is also of great importance, the microphone distance cannot be made arbitrarily large. Best results are therefore obtained when microphones or small (endfire) arrays with a high directivity are used in conjunction with the proposed postfilter.

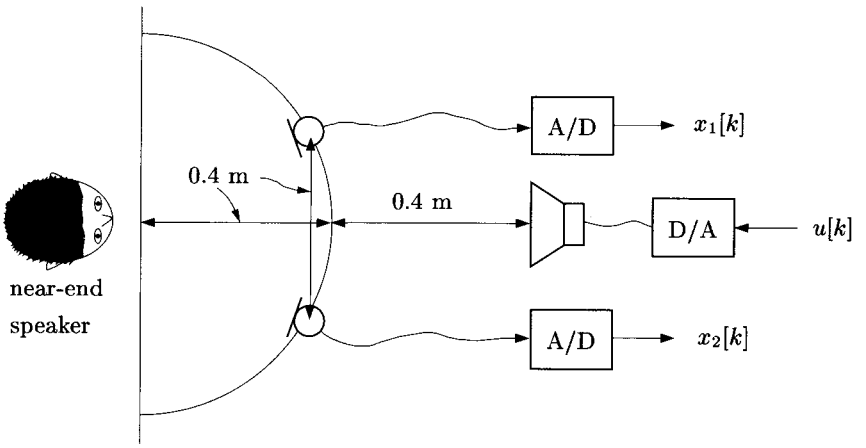


Fig. 12.12. Setup of the near-end speaker, the microphones, and the loudspeaker for conferencing experiments. The microphones have a hypercardioid directivity pattern.

The array and postfilter approach has been successfully deployed in the voice intercom of a computed tomography scanner.

Acknowledgements

The author is grateful to M. Dörbecker (Ericsson Eurolab Nuremberg), H. Hagena and T. Lotter (Aachen University of Technology) for reviewing the manuscript.

References

1. B. Widrow, J.R. Glover, J.M. McCool, et al., "Adaptive Noise Cancelling: Principles and Applications," in *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.
2. W. Armbrüster, R. Czarnach, and P. Vary, "Adaptive Noise Cancellation with Reference Input - Possible Applications and Theoretical Limits," in *Signal Processing III, Theories and Applications*, pp. 391–394, Elsevier, 1986.
3. J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.
4. E.R. Ferrara and B. Widrow, "Multichannel Adaptive Filtering for Signal Enhancement," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-29, no. 3, pp. 766–770, 1981.
5. R. Le Bouquin and G. Faucon, "On Using the Coherence Function for Noise Reduction," in *Signal Processing V: Theories and Applications*, pp. 1103–1106, Elsevier, 1990.

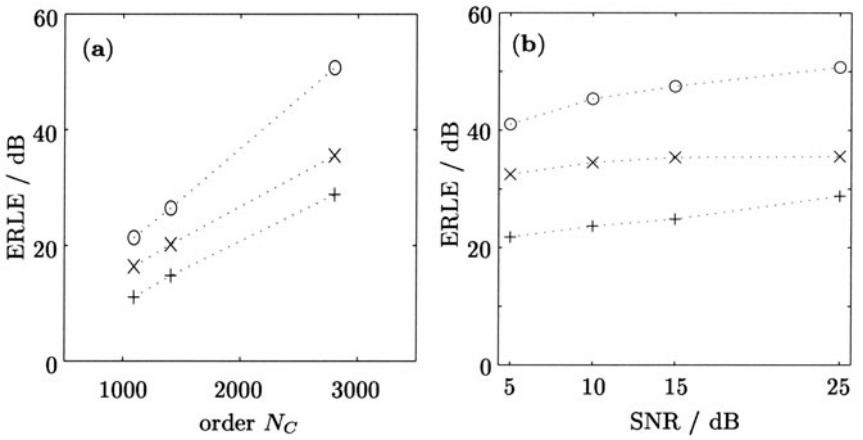


Fig. 12.13. Mean ERLE for single talk vs. the compensator order N_C for SNR ≈ 30 dB (**a**) and vs. the SNR for $N_C = 2800$ (**b**).
+: mean ERLE of echo canceller only;
x: mean ERLE of combined system with $a[k] \equiv 0$;
o: mean ERLE of combined system with $a[k]$ adaptive.

6. R. Le Bouquin and G. Faucon, "Study of a noise cancellation system based on the coherence function," in *Signal Processing VI: Theories and Applications*, pp. 1633–1636, Elsevier, 1992.
7. R. Martin and P. Vary, "A Symmetric Two Microphone Speech Enhancement System – Theoretical Limits and Application in a Car Environment," in *Proc. Fifth IEEE Signal Processing Workshop*, pp. 4.5.1–4.5.2, 1992.
8. R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation, and Noise Reduction: A Two Microphone Approach," in *Proc. Third Int. Workshop on Acoustic Echo Control*, Lannion, France, pp. 125–132, 1993.
9. R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-88)*, New York NY, USA, pp. 2578–2581, Apr. 1988.
10. R. Zelinski, "Noise Reduction Based on Microphone Array with LMS Adaptive Post-Filtering," *Elect. Lett.*, vol. 26, no. 24, pp. 2036–2037, 1990.
11. C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. SAP*, vol. 6, no. 3, pp. 240–259, 1998.
12. J.S. Bendat and A.G. Piersol, *Measurement and Analysis of Random Data*, Wiley, 1966.
13. G.C. Carter, "Coherence and Time Delay Estimation," in *Proc. IEEE*, vol. 75, no. 2, pp. 236–255, 1987.
14. H. Kuttruff, *Room Acoustics*, Elsevier Science, 3rd edn., 1990.
15. B.F. Cron and C.H. Sherman, "Spatial-Correlation Functions for Various Noise Models," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1732–1736, 1962.
16. P. Dämmig, "Zur Messung der Diffusität von Schallfeldern durch Korrelation," *Acustica*, vol. 7, pp. 387, 1957.
17. R. Martin, "Hands-free Telephones based on Multi-Microphone Echo Cancellation and Noise Reduction," PhD Thesis (in German), Institute of Communication Systems and Data Processing, Aachen University of Technology, Aachen, Germany, 1995.
18. T. Kailath, *Lectures on Wiener and Kalman Filtering*, CISM Courses and Lectures No. 140, Springer Verlag, 1981.
19. R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals," in *Proc. EUROSPEECH*, Berlin, Germany, pp. 1093–1096, 1993.
20. C.F.N. Cowan and P.M. Grant, *Adaptive Filters*, Prentice Hall, 1985.
21. J.F. Kaiser, "Nonrecursive Digital Filter Design Using the I_o-sinh Window Function," in *Proc. IEEE Int. Symp. on Circuits and Syst.*, pp. 20–23, 1974.
22. R. Le Bouquin, G. Faucon, and A. Akbari Azirani, "Proposal of a Composite Measure for the Evaluation of Noise Cancelling Methods in Speech Processing," in *Proc. EUROSPEECH*, Berlin, Germany, pp. 227–230, 1993.
23. M. Dörbecker and S. Ernst, "Combination of Two-Channel Spectral Subtraction and Adaptive Wiener Post-Filtering for Noise Reduction and Dereverberation," in *Proc. EUSIPCO*, Trieste, Italy, pp. 995–998, 1996.
24. M. Dörbecker, *Multi-channel Algorithms for the Enhancement of Noisy Speech for Hearing Aids*, PhD Thesis (in German), Institute of Communication Systems and Data Processing, Aachen University of Technology, Aachen, Germany, 1998.
25. C. Breining, P. Dreiseitel, E. Hänsler, et al., "Acoustic Echo Control. An application of very high order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, 1999.

26. S.L. Gay, J. Benesty (eds.),*Acoustic Signal Processing for Telecommunication*, Kluwer, 2000.
27. E. Hänsler, "The Hands-Free Telephone Problem - An Annotated Bibliography," *Signal Processing*, vol. 27, pp. 259–271, 1992.
28. E. Hänsler, "The Hands-Free Telephone Problem - An Annotate Bibliography Update," *Annales des Télécommunication*, vol. 49, no. 7-8, pp. 360–367, 1994.
29. J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993.
30. Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.
31. Y. Grenier, M. Xu, J. Prado, and D. Liebenguth, "Real-Time Implementation of an Acoustic Antenna for Audioconferencing," in *1st Intl. Workshop on Acoustic Echo Control*, Berlin, Germany, 1989.
32. R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation, and Noise Reduction: A Two Microphone Approach," *Annales des Télécommunications*, vol. 49, no. 7–8, pp. 429–438, 1994.
33. B. Ayad and G. Faucon, "Acoustic Echo and Noise Cancelling for Hands-Free Communication Systems," in *Proc. Fourth Int. Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 91–94, 1995.
34. G. Faucon and R. Le Bouquin Jeannes, "Joint System for Acoustic Echo Cancellation and Noise Reduction," in *Proc. EUROSPEECH*, Madrid, Spain, pp. 1525–1528, 1995.
35. R. Martin and J. Altenhöner, "Coupled Adaptive Filters for Acoustic Echo Control and Noise Reduction," in *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 3043–3046, May 1995.
36. R. Martin and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands-Free Telephony - State of the Art and Perspectives," in *Proc. EUSIPCO*, Trieste, Italy, pp. 1107–1110, 1996.
37. R. Martin and S. Gustafsson, "The Echo Shaping Approach to Acoustic Echo Control," *Speech Communication*, vol. 20, pp. 181–190, 1996.
38. W. Herbordt and W. Kellermann, "GSAEC - Acoustic Echo Cancellation Embedded into the Generalized Sidelobe Canceller," in *Proc. EUSIPCO*, Tampere, Finland, 2000.
39. W. Kellermann, "Strategies for Combining Acoustic Echo Cancellation and Adaptive Beamforming Microphone Arrays," in *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 219–222, 1997.
40. R. Martin, S. Gustafsson, and M. Moser, "Acoustic Echo Cancellation for Microphone Arrays Using Switched Coefficient Vectors," in *Proc. 5th Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, London, England, pp. 85–88, 1997.
41. W. Kellermann, "Some Properties of Echo Path Impulse Responses of Microphone Arrays and Consequences for Acoustic Echo Cancellation," in *Proc. Fourth Int. Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 39–43, 1995.
42. C. Antweiler, *Orthogonalizing Algorithms for Digital Compensation of Acoustic Echoes*. PhD Thesis (in German), Institute of Communication Systems and Data Processing, Aachen University of Technology, Aachen, Germany, 1995.
43. M. Dörbecker, "Small Microphone Arrays with Optimized Directivity for Speech Enhancement," in *Proc. EUROSPEECH*, Rhodes, Greece, pp. 327–330, 1997.

44. ITU-T Recommendation P.34, *Transmission Characteristics of Hands-Free Telephones*, Melbourne, Australia, 1988.

13 Acoustic Echo Cancellation for Beamforming Microphone Arrays

Walter L. Kellermann

University Erlangen – Nürnberg, Germany

Abstract. Acoustic feedback from loudspeakers to microphones constitutes a major challenge for digital signal processing in interfaces for natural, full-duplex human–machine speech interaction. Two techniques, each one successful on its own, are combined here to jointly achieve maximum echo cancellation in real environments: For one, acoustic echo cancellation (AEC), which has matured for single-microphone signal acquisition, and, secondly, beamforming microphone arrays, which aim at dereverberation of desired local signals and suppression of local interferers, including acoustic echoes. Structural analysis shows that straightforward combinations of the two techniques either multiply the considerable computational cost of AEC by the number of array microphones or sacrifice algorithmic performance if the beamforming is time-varying. Striving for increased computational efficiency without performance loss, the integration of AEC into time-varying beamforming is examined for two broad classes of beamforming structures. Finally, the combination of AEC and beamforming is discussed for multi-channel recording and multi-channel reproduction schemes.

13.1 Introduction

For natural human–machine interaction, acoustic interfaces are desirable that support seamless full-duplex communication without requiring the user to wear or hold special devices. For that, the general scenario of Figure 13.1 foresees several loudspeakers for multi-channel sound reproduction and a microphone array for acquisition of desired signals in the local acoustic environment. Acoustic signal processing is employed to support services such as speech transmission, speech recognition, or sound field synthesis offered by communication networks or autonomous interactive systems. Such hands-free acoustic interfaces may be tailored for incorporation into a wide variety of communication terminals, including teleconferencing equipment, mobile phones and computers, car information systems, and home entertainment equipment.

For signal acquisition, microphone arrays allow spatial filtering of arriving signals and, thus, desired signals can be enhanced and interferers can be suppressed. With full-duplex communication, echoes of the loudspeaker signals will join local interferers to corrupt the desired source signals. Beamforming, however, does not exploit the available loudspeaker signals as reference information for suppressing the acoustic echoes. This is accomplished by acoustic

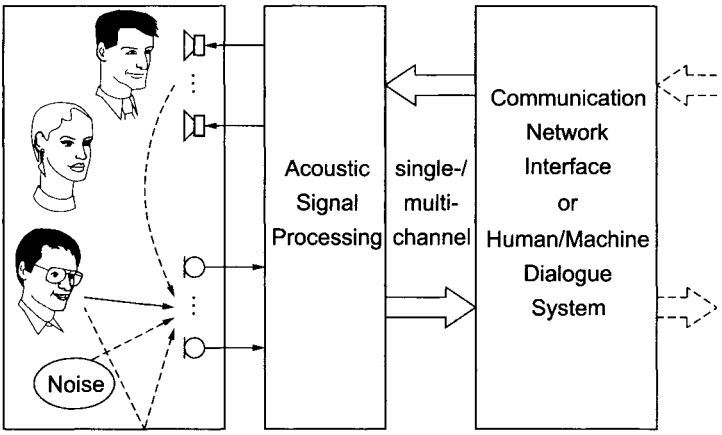


Fig. 13.1. Acoustic interface for natural human–machine communication.

echo cancellation (AEC) algorithms [1–3]. For discussing the combination of AEC with microphone arrays, the concept of AEC is first reviewed in Section 13.2 and beamforming methods are categorized in Section 13.3 with respect to the properties determining the interaction with AEC. Then, generic concepts for the combination of AEC and beamforming are discussed in Section 13.4. Structures for integrating AEC into beamforming are investigated in Section 13.5. Finally, the extension from single-channel reproduction to the case of multiple reproduction channels is outlined.

13.2 Acoustic Echo Cancellation

The concept of AEC is first considered for the case of a single loudspeaker and a single microphone according to Figure 13.2. To remove the echo from the microphone signal $x(n)$ (with n denoting discrete time), AEC aims at generating a replica $\widehat{v}(n)$ for the signal $v(n)$, which is an echoed version of the loudspeaker signal $u(n)$. Aside from the echo $v(n)$, $x(n)$ contains components originating from local desired sources and local interferers, $s(n)$ and $r(n)$, respectively. Introducing the residual echo

$$e(n) = v(n) - \widehat{v}(n), \tag{13.1}$$

the estimate for the desired signal $\widehat{s}(n)$ can be written as:

$$\widehat{s}(n) = x(n) - \widehat{v}(n) = s(n) + e(n) + r(n). \tag{13.2}$$

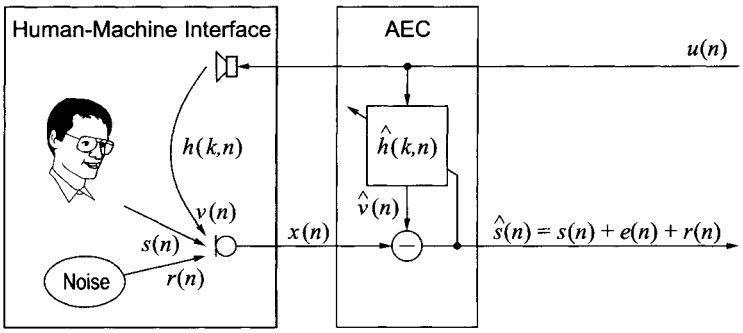


Fig. 13.2. Basic structure for single-channel AEC.

The amount of echo attenuation achieved by AEC is expressed by the *echo return loss enhancement* ($ERLE$)[1]:

$$ERLE_{log}(n) = 10 \cdot \log \frac{\mathcal{E}\{v^2(n)\}}{\mathcal{E}\{e^2(n)\}} \quad [\text{dB}], \tag{13.3}$$

with $\mathcal{E}\{\cdot\}$ denoting the expectation operator. As long as potential nonlinearities of the loudspeaker system can be neglected [4], the loudspeaker-enclosure-microphone(LEM) system is completely characterized by its generally time-varying impulse response $h(k,n)$. Indeed, the impulse response may vary drastically and unpredictably over time, as a slight change in position of any object can alter many coefficients significantly [2]. The number of impulse response samples that must be modeled for an $ERLE_{log}$ value of x dB is estimated by [2,5]

$$L_{AEC} \approx \frac{x}{60} \cdot f_s \cdot T_{60}, \tag{13.4}$$

where f_s denotes the sampling frequency, and T_{60} is the reverberation time[2]. Based on this estimate, more than $L_{AEC} = 1000$ impulse response coefficients must be perfectly matched to assure 20 dB of $ERLE_{log}$ for a typical office with $T_{60} = 400$ ms and an echo canceller operating at $f_s = 8$ kHz.

As a model for the LEM system, a digital FIR filter structure with a time-varying impulse response $\hat{h}(k,n)$ of length L_{AEC} is employed, so that the estimated echo $\hat{v}(n)$ is given by

$$\hat{v}(n) = \hat{\mathbf{h}}^T(n) \cdot \mathbf{u}(n) \tag{13.5}$$

[1] As $v(n)$ and $e(n)$ are not accessible in practical situations, $ERLE$ must be estimated from $\hat{s}(n)$ and $x(n)$ [2].

[2] As characteristic parameter of an enclosure, the reverberation time T_{60} is the time until the sound energy decays by 60dB after switching off the source.

where ^{T} denotes transposition and

$$\widehat{\mathbf{h}}(n) = \left[\widehat{h}(0,n), \widehat{h}(1,n), \ldots, \widehat{h}(L_{AEC}-1,n)\right]^T, \tag{13.6}$$

$$\mathbf{u}(n) = [u(n), u(n-1), \ldots, u(n-L_{AEC}+1)]^T. \tag{13.7}$$

The misalignment between the FIR model $\widehat{\mathbf{h}}(n)$ and the LEM system $\mathbf{h}(n)$ is described by the logarithmic system error norm $D_{log}(n)$:

$$D_{log}(n) = 10 \cdot \log \frac{||\mathbf{h}(n) - \widehat{\mathbf{h}}(n)||_2^2}{||\mathbf{h}(n)||_2^2}, \tag{13.8}$$

with $||\cdot||_2$ denoting the l_2 norm[3].

13.2.1 Adaptation algorithms

For identifying the time-varying impulse response $h(k,n)$, adaptive filtering algorithms derive an optimum vector $\widehat{\mathbf{h}}_{opt}(n)$ by minimizing a mean square error criterion based on the input $u(n)$ and the estimation error $e(n)$ (assuming here, for simplicity, $s(n) = r(n) = 0$). Three fundamental algorithms are introduced below for the general case of complex signals (for a comprehensive treatment of adaptive FIR filtering see, e.g., [6,7]). Adaptation control in the context of AEC is addressed and frequency domain implementations are outlined briefly.

Fundamental algorithms. Minimizing the mean squared error $E\{|e(n)|^2\}$ for (at least) wide-sense stationary signals and a time-invariant echo path $h(k,n) = h(k)$ leads to the Wiener-Hopf equation for the optimum echo canceller $\widehat{\mathbf{h}}_{opt}$ [7]

$$\widehat{\mathbf{h}}_{opt} = \mathbf{R}_{\mathbf{uu}}^{-1} \cdot \mathbf{r}_{\mathbf{uv}} \tag{13.9}$$

with the time-invariant correlation matrix $\mathbf{R}_{\mathbf{uu}}$ and the crosscorrelation vector $\mathbf{r}_{\mathbf{uv}}$ given by

$$\mathbf{R}_{\mathbf{uu}} = E\{\mathbf{u}(n)\mathbf{u}^H(n)\}, \tag{13.10}$$

$$\mathbf{r}_{\mathbf{uv}} = E\{\mathbf{u}(n)v^*(n)\}, \tag{13.11}$$

respectively. (* denotes complex conjugation and ^{H} conjugate complex transposition.) For nonstationary environments, iterative or recursive algorithms are required to approach the Wiener solution in (13.9). As the most popular adaptation algorithm, the **NLMS**(*Normalized Least Mean Square*) algorithm [6,7] updates the filter according to

$$\widehat{\mathbf{h}}(n+1) = \widehat{\mathbf{h}}(n) + \alpha \frac{\mathbf{u}(n)}{\mathbf{u}^H(n)\mathbf{u}(n)} e^*(n) \tag{13.12}$$

[3] If the length of $\mathbf{h}(n)$ is greater than L_{AEC}, then $\widehat{\mathbf{h}}(n)$ must be complemented with zeros accordingly.

and may be understood as a stochastic approximation of the steepest descent algorithm, with $\mathbf{u}(n)$ approximating the negative gradient vector, and a step-size parameter α, $0 < \alpha < 2$. Obviously, for correlated signals such as speech, $\mathbf{u}(n)$ will not cover uniformly the L_{AEC}-dimensional vector space, which implies that the convergence to minimum system error $D_{log}(n)$ in (13.8) is slow [7]. The popularity of the NLMS is based on its robust convergence behavior [2] and its low computational complexity (about $2L_{AEC}$ multiplications per sampling interval T (MUL's per T) are needed for implementing (13.1), (13.5), and (13.12)).

To improve the convergence for speech signals, the *Affine Projection Algorithm* (**APA**) uses P previous input vectors

$$\mathbf{U}(n) = [\mathbf{u}(n), \mathbf{u}(n-1), \ldots, \mathbf{u}(n-P+1)] \tag{13.13}$$

to compute an error vector

$$\mathbf{e}(n) = \mathbf{v}(n) - \mathbf{U}^T(n) \cdot \widehat{\mathbf{h}}^*(n), \tag{13.14}$$

where

$$\mathbf{e}(n) = [e(n), e(n-1), \ldots, e(n-P+1)], \tag{13.15}$$

$$\mathbf{v}(n) = [v(n), v(n-1), \ldots, v(n-P+1)]. \tag{13.16}$$

The filter coefficients are then updated according to

$$\widehat{\mathbf{h}}(n+1) = \widehat{\mathbf{h}}(n) + \alpha \mathbf{U}(n) \left[\mathbf{U}^H(n)\mathbf{U}(n) - \delta \mathbf{I}\right]^{-1} \mathbf{e}^*(n), \tag{13.17}$$

with the regularization parameter δ ($\delta \geq 0$) and $\mathbf{I}$ denoting the identity matrix. Thus, the APA can be interpreted as a generalization of the NLMS algorithm, which in turn corresponds to an APA with $P = 1, \delta = 0$. The gradient estimate for the APA is equal to the projection of the system misalignment vector $\mathbf{h}(n) - \widehat{\mathbf{h}}(n)$ onto the P-dimensional subspace spanned by $\mathbf{U}(n)$. Thus, the complementary orthogonal component of the misalignment vector becomes smaller with increasing P. The computational complexity of the APA amounts to approximately $(P+1) \cdot L_{AEC} + O(P^3)$ MUL's per T, where, typically, $P = 2, \ldots, 32$, and L_{AEC} is given by (13.4). Fast versions of the APA reduce the computational load to $2L_{AEC} + 20P$, but require additional measures to assure numerical stability [2,6].

As the most powerful and computationally demanding adaptation method, the **RLS**(*Recursive Least Squares*) algorithm directly minimizes a weighted sum of previous error samples

$$J(\widehat{\mathbf{h}}, n) = \sum_{k=1}^{n} \beta(k) |e(k)|^2, \quad \text{with} \quad 0 < \beta \leq 1. \tag{13.18}$$

The solution has the form of (13.9), however with time-dependent estimates for $\mathbf{R}_{\mathbf{uu}}(n)$, $\mathbf{r}_{\mathbf{u}v}(n)$ given by

$$\widehat{\mathbf{R}}_{\mathbf{uu}}(n) = \sum_{k=1}^{n} \beta(k)\mathbf{u}(k)\mathbf{u}^H(k), \tag{13.19}$$

$$\widehat{\mathbf{r}}_{\mathbf{u}v}(n) = \sum_{k=1}^{n} \beta(k)\mathbf{u}(k)v^*(k). \tag{13.20}$$

The update equation reads here

$$\widehat{\mathbf{h}}(n+1) = \widehat{\mathbf{h}}(n) + \widehat{\mathbf{R}}_{\mathbf{uu}}^{-1}(n)\mathbf{u}(n)e^*(n). \tag{13.21}$$

If an exponential window $\beta(k) = \lambda^{n-k}$ with the forgetting factor $0 < \lambda < 1$ is used, the inversion of $\widehat{\mathbf{R}}_{\mathbf{uu}}(n)$ is avoided by exploiting the matrix inversion lemma that allows recursive update of the inverse [7]. Then, the complexity of the RLS algorithm is on the order of L_{AEC}^2 MUL's per T [6]. Similarly to the APA, fast versions for the RLS algorithm have been proposed which reduce computational complexity to $7L_{AEC}$ MUL's per T. However, the large filter order L_{AEC} and the nonpersistent excitation $u(n)$ require extra efforts to assure stable convergence [6]. A simplified version of fast RLS algorithms is the *Fast Newton Algorithm* [6], which reduces the complexity to $L_{AEC} \cdot P$ MUL's per T, with P being a predictor order that should be matched to the correlation properties of the input $u(n)$. (For speech signals, $P \approx 10$ is a typical value at $f_s = 8$kHz.)

Adaptation control. Adaptation control has to satisfy two contradicting requirements. On one hand, changes in the echo path $h(k,n)$ should be tracked as fast as possible. This requires a large stepsize, α, for the NLMS and APA algorithms in (13.12) and (13.17)), and a rapidly decaying β for the RLS algorithm in (13.21), respectively. On the other hand, the adaptation must be robust to interfering local sources $s(n)$ and noise $r(n)$, which requires a small stepsize, α, and a slowly decaying β, respectively [2,7]. When a local talker is active, adaptation should be stalled immediately to avoid divergence of $\widehat{\mathbf{h}}(n)$. Therefore, a fast and reliable detection of local source activity and estimation of background noise levels is decisive for efficient AEC operation. Correspondingly, a significant amount of computational complexity is invested in monitoring parameters and signals which support adaptation control [2]. With properly tuned adaptation control, acoustic echoes are attenuated by, typically, about 25 dB of $ERLE_{log}$ during steady state using the above adaptation algorithms.

Frequency subband and transform domain structures. To reduce computational load and to speed up convergence of adaptation algorithms

which do not inherently decorrelate $u(n)$ (e.g., the NLMS algorithm), frequency subband and transform domain structures have been developed [1,8]. Subband structures decompose the fullband signals $u(n)$ and $x(n)$ into M subbands which are usually downsampled by $R < M$ [3,9]. The adaptive subband filters operate at a reduced sampling rate and require fewer coefficients which leads to overall computational savings by a factor of close to R^2/M compared to fullband adaptive filtering. After subtraction, the subband signals are synthesized to yield again a fullband signal $\widehat{s}(n)$. While the additional complexity for the analysis/synthesis filter banks is relatively small for large L_{AEC}, the introduced signal delay for $\widehat{s}(n)$ is objectionable in some applications [2,31].

Transform-domain structures draw their computational advantage over direct time-domain implementations from the fast Fourier transform (FFT) and its use for fast convolution [1,6,8]. Block-exact adaptation algorithms, which behave exactly like their time-domain counterparts, have been proposed for all the fundamental algorithms above. For the long impulse responses at issue, the system model $\widehat{h}(k,n)$ is often partitioned into shorter subsystems to reduce the signal delay [2].

13.2.2 AEC for multi-channel sound reproduction

Considering a multi-channel reproduction unit (see Figure 13.1) broadcasting K different sound channels $\mathbf{u}_\kappa(n)$ $(\kappa = 0, \ldots, K-1)$ with usually time-varying mutual correlation, any microphone records the sum of K echo signals produced by different echo paths $h_\kappa(k,n)$,

$$v(n) = \sum_{\kappa=0}^{K-1} \mathbf{h}_\kappa(n)^T \cdot \mathbf{u}_\kappa(n), \tag{13.22}$$

with $\mathbf{h}_\kappa(n), \mathbf{u}_\kappa(n)$ being defined according to (13.6) and (13.7). Correspondingly, K echo cancellers, $\widehat{\mathbf{h}}_\kappa(n)$, are needed to model the respective echo paths. As only one error signal, $e(n)$, is available, the K inputs, $u_\kappa(n)$, must be mutually decorrelated without perceptible distortion to allow identification of the individual $\widehat{\mathbf{h}}_\kappa(n)$. This difference to single-channel AEC defines an even more challenging system identification problem, which has been considered only for the stereo case ($K = 2$) so far [1,10–12]. Current adaptation schemes still exhibit slower convergence and multiply computational load by more than K compared to their single-channel AEC counterparts.

13.2.3 AEC for multi-channel acquisition

A straightforward extension of the single-loudspeaker/ single-microphone scenario to an N-microphone acquisition system essentially multiplies the number of adaptive filters by N. The N-channel echo cancellation is captured by

extending the signals in (13.2) to N-dimensional column vectors,

$$\widehat{\mathbf{s}}(n) = \mathbf{x}(n) - \widehat{\mathbf{v}}(n) = \mathbf{s}(n) + \mathbf{r}(n) + \mathbf{e}(n) \tag{13.23}$$

$$= \mathbf{s}(n) + \mathbf{r}(n) + \mathbf{v}(n) - \widehat{\mathbf{H}}^T(n)\mathbf{u}(n) \tag{13.24}$$

with $\mathbf{u}(n)$ according to (13.7), with $\mathbf{e}(n), \mathbf{r}(n), \widehat{\mathbf{s}}(n), \mathbf{s}(n), \mathbf{v}(n), \widehat{\mathbf{v}}(n), \mathbf{x}(n)$ as column vectors of the form

$$\mathbf{x}(n) = [x_0(n), \ldots, x_{N-1}(n)]^T, \tag{13.25}$$

and with $\widehat{\mathbf{H}}(n)$ as a matrix containing the impulse responses $\widehat{\mathbf{h}}_\nu(n)$ as columns according to

$$\widehat{\mathbf{H}}(n) = \left[\widehat{\mathbf{h}}_0(n), \ldots, \widehat{\mathbf{h}}_\nu(n), \ldots, \widehat{\mathbf{h}}_{N-1}(n)\right]. \tag{13.26}$$

While this implies a corresponding multiplication of the computational cost for filtering, the cost for adaptation and its control is not necessarily multiplied by N. All operations depending only on the input data, $u(n)$, have to be carried out only once for all N channels, which would include the matrix inversion in the APA or RLS algorithms, (13.17) and (13.21), respectively. However, some fast versions draw their efficiency from interweaving matrix inversion and update equations [6] and, therefore, do not completely support this separation. Frequency subband and transform domain algorithms [1,6,8,9] support this separation at least by requiring the analysis transform of $u(n)$ only once for all channels.

13.3 Beamforming

This section only aims at categorizing beamforming algorithms with respect to their interaction with AEC. For a comprehensive treatment of fundamental techniques see, e.g., [13,14], while the current state of beamforming technology with microphone arrays is covered in several other chapters of this book.

13.3.1 General structure

Consider a microphone array capturing N real-valued sensor signals, $x_\nu(n)$, which are filtered by linear time-varying systems with impulse responses $g_\nu(k,n)$ and then summed up (Figure 13.3). The resulting beamformer output, $y(n)$, can be written as

$$y(n) = \mathbf{G}^T(n) \cdot \mathbf{X}(n) = \mathbf{G}^T(n) \cdot [\mathbf{S}(n) + \mathbf{R}(n) + \mathbf{V}(n)], \tag{13.27}$$

with the column vector $\mathbf{G}(n)$ representing the concatenated impulse response vectors $\mathbf{g}_\nu(n)$

$$\mathbf{G}(n) = \left[\mathbf{g}_0^T(n), \ldots, \mathbf{g}_{N-1}^T(n)\right]^T, \tag{13.28}$$

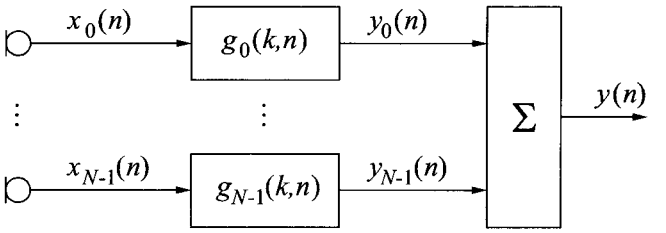


Fig. 13.3. General structure for a beamforming microphone array

where all $\mathbf{g}_\nu(n)$ are of length L_{BF}:

$$\mathbf{g}_\nu(n) = [g_\nu(0,n), \ldots, g_\nu(L_{BF}-1,n)]^T . \tag{13.29}$$

The column vector $\mathbf{X}(n)$ (and, equally, $\mathbf{R}(n), \mathbf{S}(n), \mathbf{V}(n)$) contains the latest L_{BF} signal samples of each microphone signal

$$\mathbf{X}(n) = \left[\mathbf{x}_0^T(n), \ldots, \mathbf{x}_{N-1}^T(n)\right]^T \tag{13.30}$$

with

$$\mathbf{x}_\nu(n) = [x_\nu(n), \ldots, x_\nu(n-L_{BF}+1)]^T . \tag{13.31}$$

In the scenario of Figure 13.1, beamforming aims at spatial filtering to dereverberate the components $\mathbf{s}(n)$ originating from the desired source(s) and to suppress interfering signals $\mathbf{r}(n)$ and echoes $\mathbf{v}(n)$.

For ideal dereverberation of a single source, the desired signal as it is emitted by the source, $s^{(0)}(n)$, should be retrieved except for some delay $n_0 > 0$:

$$\mathbf{G}^T(n) \cdot \mathbf{S}(n) = s^{(0)}(n-n_0). \tag{13.32}$$

Assuming that delayed versions of $s^{(0)}(n)$ are contained in $\mathbf{s}_\nu(n)$ defined by (13.31), the filters $g_\nu(k,n)$ have to equalize the corresponding delays and the sum of the filters has to provide a flat frequency response for all signals arriving from the source direction. Obviously, delay equalization requires knowledge about the location of the desired source(s). For the following, it is assumed that the source location is given by *a priori* knowledge or separately determined by some source localization algorithm (see, e.g., Chapters 8-10). For an anechoic environment and with the desired signal components being delay-equalized by the array geometry, the total impulse response, $g(k,n)$, of the beamformer to the desired source $s^{(0)}(n)$ should ideally fulfill

$$g(k,n) = \sum_{\nu=1}^{N} g_\nu(k,n) \stackrel{!}{=} \delta(k-k_0) \tag{13.33}$$

to assure a constant frequency response with unity gain and constant group delay k_0.

For interference suppression, the beamformer should minimize its response to all undesired signal components, which include here local interferers and loudspeaker echoes. Using, the mean squared error (MSE) as optimization criterion, this reads:

$$\mathcal{E}\left\{\left(\mathbf{G}^T(n)\cdot[\mathbf{R}(n)+\mathbf{V}(n)]\right)^2\right\}\stackrel{!}{=}\min. \tag{13.34}$$

Based on this general concept and with AEC in mind, basic methods for time-invariant or time-varying beamforming are outlined below.

13.3.2 Time-invariant beamforming

Time-invariant beamforming, i.e., $\mathbf{G}(n)=\mathbf{G}$, $\mathbf{g}_\nu(n)=\mathbf{g}_\nu$, is used for applications where the beamformer does not have to change the 'look direction' and where the potential nonstationarity of the involved signals, $\mathbf{s}(n),\mathbf{r}(n),\mathbf{v}(n)$, is not accounted for.

As the most basic beamforming method, the delay-and-sum beamformer (DSB) realizes in its simplest form a tapped delay line with a single non-zero coefficient for each filter $\mathbf{g}_\nu(n)$ [13,14]. If the required delays for the desired 'look direction' do not coincide with integer multiples of the sampling period, interpolation filters are required for realizing fractional delays [15–17]. Accounting for the wideband nature of speech and audio signals, nested arrays are often employed using different sets of sensors for different frequency bands to approximate a constant ratio between aperture width and signal wavelength [17–19]. As a generalization of DSB, filter-and-sum beamforming (FSB) aims for a frequency-independent spatial selectivity within each frequency band as detailed in Chapter 1 and [20]. Both beamforming concepts, DSB and FSB, were first developed on the basis of the far-field assumption [18], but may also be extended to near-field beamforming as described in Chapter 1. Time-invariant DSB and FSB are mostly signal-independent, i.e., no attention is paid to the power spectral densities of the signals $\mathbf{s}(n)$, $\mathbf{r}(n)$, $\mathbf{v}(n)$ and the direction of arrival (DOA) of interferers.

Such 'beamsteering' techniques are obviously appropriate for human–machine interfaces in reverberant environments with a restricted range of movement for a single desired source and where, due to reverberation, unwanted signal components of comparable level must be expected from all directions.

Nevertheless, time-invariant beamforming can incorporate additional spatial information to suppress dominant interferers [21,22]. Moreover, knowledge about long-term statistics of the noise field can be accounted for [23] and may lead to statistically optimum beamformers with superdirective behaviour for low frequencies as described in Chapter 2 and [24].

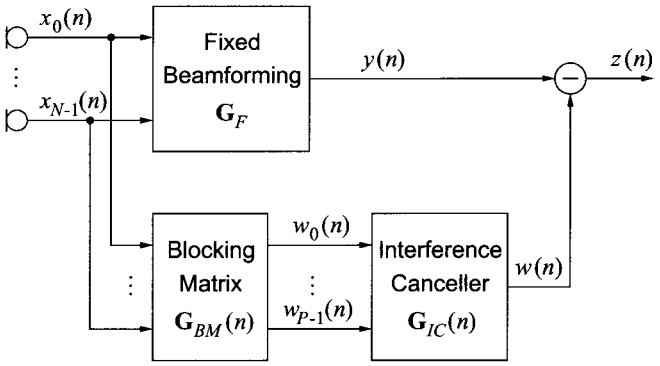


Fig. 13.4. Generalized sidelobe canceller structure for adaptive beamforming.

13.3.3 Time-varying beamforming

For nonstationary environments with both nonstationary signal characteristics and potentially moving sources, the beamformer should be able to track the time-variance of the signal characteristics and the spatial arrangement of the interfering sources. For that purpose, adaptive beamforming methods design filters $g_\nu(k,n)$ which minimize a statistical error criterion based on the array output, $y(n)$, with constraints for the DOA of a desired source (or 'target') such as formulated in (13.33) and (13.34) [13,14,25–27]. See also Chapter 5.

Generalized Sidelobe Canceller (GSC). As an example for an efficient implementation of adaptive beamformers that minimize a mean square error (MSE) criterion subject to a linear constraint, the generalized sidelobe canceller structure [13,25] is considered (Figure 13.4). Here, the adaptive beamforming is separated into two parallel paths: The upper path is a time-invariant, signal-independent beamformer, $\mathbf{G}_F$, steered toward the desired source. In the lower path, the first stage implements a blocking-matrix, $\mathbf{G}_{BM}(n)$, which, ideally, completely suppresses the components of the desired source, $\mathbf{s}(n)$, by a linear combination of the microphone channels [13] or filtering [28]. This topic is also detailed in Chapter 5. The $P \leq N$ outputs, $w_i(n), i = 0, \ldots, P-1$, are then used by the adaptive interference canceller, $\mathbf{G}_{IC}(n)$, to form an estimate for the interference component in $y(n)$. Optimization of $\mathbf{G}_{IC}(n)$ becomes an unconstrained Wiener filtering problem when the MSE criterion of (13.9) is used, and ideally leads to removal of all components in $y(n)$ which are correlated to $w_i(n)$. For identifying the optimum $\mathbf{G}_{IC}(n)$, the same adaptation algorithms as for echo cancellation can be used, i.e., (13.12),(13.17),(13.21), with gradient-type algorithms like the NLMS algorithm being most common.

13.3.4 Computational complexity

For both time-invariant and time-varying beamforming, the computational load is essentially proportional to the number of sensors N. The FIR filter lengths typically do not exceed $L_{BF} = 128$ [17,20,29,30]. With increasing filter length, computational savings are obtained by frequency-domain implementations of the filtering [20,29]. As with AEC, for adaptive beamforming implementations a significant share of computational complexity is dedicated to fast and reliable source activity detection which forms the basis of adaptation control.

13.4 Generic structures for combining AEC with beamforming

First, the combination of AEC with beamforming is motivated by comparing practical requirements with typical performance of AEC and beamforming. Then, the main properties of two generic options for a combination are discussed in some detail.

13.4.1 Motivation

Although AEC and beamforming are two distinct signal processing concepts, their goals meet with regard to acoustic echoes. While AEC subtracts from $\mathbf{x}(n)$ an echo estimate, $\widehat{\mathbf{v}}(n)$, based on $u(n)$ as reference information, beamforming suppresses echoes within $\mathbf{x}(n)$ as undesired interference by its spatial filtering capability. With beamforming being undisputed for its effectiveness in suppressing local noise and reverberance of local desired sources, the need for a complementary AEC unit for acoustic echo suppression is discussed in the following.

As a guideline for desired echo suppression for telecommunication, [31] requires $ERLE_{log} \geq 45$ dB during single-talk and at least 30 dB during double-talk, assuming a 'natural' echo attenuation of up to 6 dB between the loudspeaker signal, $u(n)$, and the microphone signal, $x(n)$. Echo suppression methods other than AEC, e.g., noise reduction, loss insertion, or nonlinear devices, impair full-duplex communication and, thus, are only acceptable as supplementary measures [2]. For full-duplex speech dialogue systems employing automatic speech recognition, the echo attenuation requirements are not as well-defined and will depend on the desired recognition rate as well as on the robustness of the speech recognizer with respect to speech-like interference. In view of these requirements, the echo attenuation provided by microphone arrays and the echo path gain for a microphone array are examined below.

Array gain. The echo attenuation provided by a microphone array is usually identified with the array gain for the desired sources relative to echoes as interference. For signal-independent time-invariant beamforming, the directivity index quantifying the array gain of the desired direction over the average of all other directions [26] does typically not exceed 20 dB over a wide frequency range, and is much smaller at low frequencies (< 500 Hz) due to usual geometrical aperture constraints [19,26]. This contrasts with the fact that acoustic echoes usually exhibit their maximum energy at low frequencies [2]. As a remedy, differential beamforming realizes superdirective array gains at low frequencies and allows for a directivity index of up to 12 dB in practical implementations [1,27]. On the other hand, for adaptive beamforming, interference suppression is usually also limited to about 20 dB for reverberant environments if distortion of the desired source signal $s^{(0)}(n)$ should be precluded. See Chapters 2 and 5 as well as [19,32].

Echo path gain. For microphone array applications, the echo path gain between $u(n)$ and the beamformer output, $y(n)$, will often be higher than for single-microphone systems (-6 dB), because the sum of the distances from the loudspeaker to the listener, and from the desired source to the microphone array, will usually be greater (e.g. in teleconferencing). The user will typically increase the gains for the loudspeaker signal and the microphone array correspondingly to compensate for the decay of the sound level ($\approx$ 6 dB per doubling of distance in the far-field). If the microphone array and loudspeaker are relatively close, then the required echo attenuation will be increased accordingly.

13.4.2 Basic options

Restricting the scenario to a single reproduction channel, $u(n)$, and a single acquisition channel, $\widehat{s}(n)$, a combination of AEC and beamforming is obviously conceivable in two fundamentally different ways as shown in Figure 13.5. Here, 'AEC first' realizes one adaptive filter for each microphone in $\widehat{\mathbf{H}}^{(I)}(n)$ of (13.26), whereas 'Beamforming first' uses a single-channel AEC, $\widehat{\mathbf{h}}^{(II)}(n)$, which obviously has to include the beamformer, $\mathbf{G}(n)$, into its echo path model.

13.4.3 'AEC first'

This structure suggests that $\widehat{\mathbf{H}}^{(I)}(n)$ may operate without any repercussions from the beamforming so that the AEC problem corresponds to that described by (13.23). On the other hand, with perfect echo cancellation, the beamforming will be undisturbed by acoustic echoes and will concentrate on suppressing local interferers and reverberation.

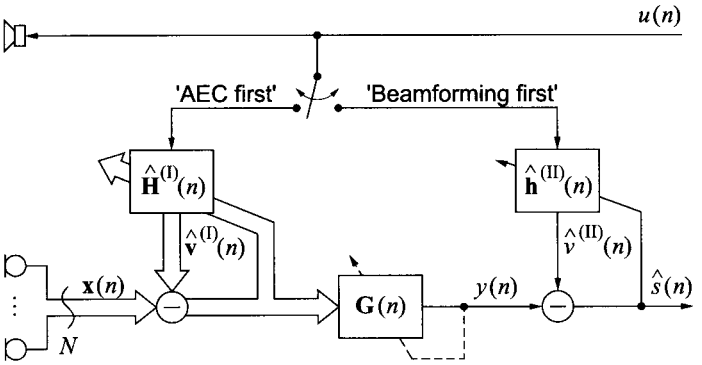


Fig. 13.5. Generic structures for combining AEC with beamforming.

AEC properties. Although AEC could operate independently from the beamforming, synergies with beamforming should be exploited with regard to detection of local source activity and computational complexity.

Local source activity detection. As noted above, the adaptation of $\widehat{\mathbf{H}}^{(I)}(n)$, requires a fast and reliable detection of local source activity to avoid divergence. With single-channel AEC, the detection is based on comparing estimates for

$$Q_\nu(n) = \frac{E\left\{v_\nu^2(n)\right\}}{E\left\{(r_\nu(n) + s_\nu(n))^2\right\}} \tag{13.35}$$

to a given threshold. With subsequent beamforming, this decision can be derived from estimates of

$$Q(n) = \frac{E\left\{\left(\mathbf{G}^T(n)\mathbf{V}(n)\right)^2\right\}}{E\left\{\left(\mathbf{G}^T(n)\left[\mathbf{R}(n) + \mathbf{S}(n)\right]\right)^2\right\}} \tag{13.36}$$

which reflect local source activity much clearer than $Q_\nu(n)$ as $\mathbf{r}_\nu(n), \mathbf{v}_\nu(n)$ are suppressed relative to $\mathbf{s}(n)$ by beamforming. Thus, $Q(n)$ reduces uncertainty in local source activity detection and allows adaptation during time intervals where adaptation might have been stalled if its control was based on $Q_\nu(n)$.

Computational complexity. At least the filtering and the filter coefficient update of the AEC adaptation will require N-fold computational cost compared to a single-channel AEC. Even with continuing growth of the performance-cost ratio of signal processing hardware, this computational load will remain prohibitive in the near future for many cost-sensitive or very large systems employing $N = 5, \ldots, 512$ sensors [17,19,26,30,33,34]. One option to alleviate the computational burden is to reduce the length L_{AEC} in (13.4) of the FIR filter models, $\widehat{\mathbf{h}}_\nu$, and to rely on the beamformer for suppressing the residual

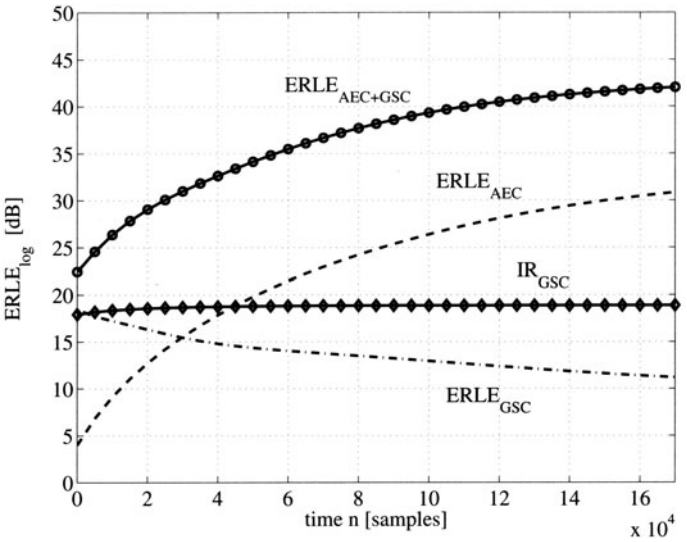


Fig. 13.6. Example for convergence of $ERLE_{log}$ components and local interference suppression(IR) for 'AEC first' structure ($N = 8$, $T_{60} \approx 300$ ms, $f_s = 12$ kHz, $L_{AEC} = 2500$, $L_{BM} = 16$, $L_{IC} = 50$).

echoes, $\mathbf{e}(n)$. Shortening $\widehat{\mathbf{h}}_\nu$ implies however, that the adaptation of the AEC is disturbed by an increased noise component, which is due to the unmodeled tail of the true echo path impulse response, $\mathbf{h}_\nu(n)$ [2].

Beamforming performance. For a signal-independent beamformer, the presence and performance of the AEC has no impact on the beamforming. The signal-independent spatial filtering will increase echo suppression according to its directivity while suppression of local interferers remains unaffected.

Signal-dependent beamformers use $\mathbf{w}(n) = \mathbf{x}(n) - \widehat{\mathbf{v}}^{(I)}(n)$ for optimizing the beamforming filters $\mathbf{G}(n)$. Thereby, at the cost of local interference suppression, the beamformer will concentrate on suppressing echo components, $\mathbf{e}(n)$, if their levels exceed that of local interferers, $\mathbf{r}(n)$, and it will further suppress residual echoes as long as they are not negligibly small compared to the local interferers. For illustration, the typical convergence behaviour for 'AEC first' using a GSC beamformer is shown in Figure 13.6 for $\mathbf{r}(n)$, $\mathbf{s}(n)$, $u(n)$ being white noise signals, and for alternating adaptation of $\mathbf{G}_{IC}(n)$, and $\widehat{\mathbf{H}}^{(I)}(n)$ (see also [32]). Due to its short filters, the beamformer converges almost instantaneously to about $ERLE_{GSC} = 18$ dB, and thereby provides a significant amount of $ERLE_{log}$ long before $\widehat{\mathbf{H}}^{(I)}(n)$ has converged. At the same time, suppression of local interference, IR_{GSC}, remains essentially con-

stant over time, as it converges very rapidly to almost 20 dB and is not allowed to converge much further to preclude distortion of the desired signal.

13.4.4 'Beamforming first'

In this structure, the beamformer is essentially independent from the AEC so that the beamforming performance agrees with Section 13.3 for acoustic echoes being perceived as another source of interference. AEC is realized by a single adaptive filter $\widehat{\mathbf{h}}^{(II)}(n)$ as in Figure 13.5 which is attractive with regard to computational complexity. However, the system identification problem faced by $\widehat{\mathbf{h}}^{(II)}(n)$ requires closer examination.

Echo path for AEC. Incorporating the beamformer, $\mathbf{G}(n)$, into the echo path model means that, ideally, the adaptive filter, $\widehat{\mathbf{h}}^{(II)}(n)$, models the sum of N echo paths from the loudspeaker input, $u(n)$, to the beamformer output, $y(n)$, (see Figure 13.3)

$$\widehat{\mathbf{h}}_{opt}^{(II)}(n) = \mathbf{f}(n) = \sum_{\nu=1}^{N} \mathbf{f}_\nu(n), \tag{13.37}$$

with the impulse responses, $\mathbf{f}(n)$, given by ($\star$ denotes linear convolution):

$$\mathbf{f}_\nu(n) = [f_\nu(0,n), \ldots, f_\nu(L_{AEC+BF}-1, n)]^T, \tag{13.38}$$

$$f_\nu(k,n) = h_\nu(k,n) \star g_\nu(k,n). \tag{13.39}$$

Thus, the impulse response length of $\widehat{\mathbf{h}}^{(II)}(n)$ depends on the beamforming, and, if any $g_\nu(k,n)$ is time-varying, $\widehat{\mathbf{h}}^{(II)}(n)$ has to track this time-variance as well[4]. The required length, L_{AEC+BF}, for $\widehat{\mathbf{h}}^{(II)}(n)$ is essentially the sum of the length L_{BF} and the necessary length for the acoustic path (including loudspeaker and microphone), L_{AEC}:

$$L_{AEC+BF} = L_{AEC} + L_{BF} - 1. \tag{13.40}$$

Note that for a given desired $ERLE_{log}$, L_{AEC} can be chosen smaller than given by (13.4) depending on the expected contribution of beamforming to $ERLE_{log}$ (see also [35]).

Signal-independent, time-invariant beamformers. Due to the time-invariance of $g_\nu(k,n)$, the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ only has to track the time-variance of $\mathbf{h}_\nu(n)$ and, thus, the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ is identical to the adaptation of one of the N filters $\widehat{\mathbf{h}}_\nu^{(I)}(n)$ in the 'AEC first' structure except for the different filter length L_{AEC+BF}.

[4] Note that the time-varying components $h_\nu(k,n)$ cannot be identified separately, although $g_\nu(k,n)$ is known ('knapsack problem').

Signal-dependent, time-varying beamformers. Here, the main problem is that the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ has to track the time-variance of $\mathbf{G}(n)$. As for the adaptation algorithms discussed in Section 13.2.1 an increasing filter order involves a reduced tracking capability [7], the high-order filter, $\widehat{\mathbf{h}}^{(II)}(n)$, cannot follow the time-variance of the low-order filters of $\mathbf{G}(n)$ ($L_{AEC+BF} \gg L_{BF}$). Therefore, $\widehat{\mathbf{h}}^{(II)}(n)$ can find a useful echo path model only when $\mathbf{G}(n)$ remains time-invariant for a sufficiently long time. In Figure 13.7, the adaptation behaviour of the 'beamforming first' structure is analyzed for a speech conversation with a GSC as adaptive beamformer [28,32]. Inspecting the time domain signals $u(n)$ and $s(n)$ in Figures 13.7a and 13.7b shows that a 'double-talk' period occurs for time $n = 3.5 \ldots 4.0 \cdot 10^5$. Figure 13.7c illustrates which component is adapted at a given time. To track slight movements of the desired local source, the blocking matrix, $\mathbf{G}_{BM}(n)$, is adapted if only the local source and noise are present [28,32]. The system error of (13.8) depicted in Figure 13.7d converges monotonically when $\widehat{\mathbf{h}}^{(II)}(n)$ is adapted. When the interference canceller, $\mathbf{G}_{IC}(n)$, or the blocking matrix, $\mathbf{G}_{BM}(n)$, are adapted the system error rises instantaneously ($n = 2 \ldots 3.5 \cdot 10^5$). This is not critical as long as $u(n) = 0$, however, during double-talk ($n = 3.5 \ldots 4.0 \cdot 10^5$), a large residual error, $e(n)$, arises (Figure 13.7e,f) as $\widehat{\mathbf{h}}^{(II)}(n)$ cannot reconverge. Consequently, with the 'beamforming first' structure, the benefits of AEC are missing when they are desired most, i.e., during double-talk and during transitions from far-end activity to local activity and vice-versa (at other times primitive echo suppression methods, such as loss insertion [2], are less objectionable).

13.5 Integration of AEC into time-varying beamforming

As time-varying beamforming, $\mathbf{G}(n)$, cannot be tracked satisfactorily by the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$, a compromise is desirable for AEC which avoids the computational burden of $\widehat{\mathbf{H}}^{(I)}(n)$ for large N and provides improved echo cancellation compared to $\widehat{\mathbf{h}}^{(II)}(n)$. For this, the beamformer is decomposed into a time-invariant and a time-varying part in the sequel, with AEC acting only on the output of the time-invariant part. Two options for arranging the time-invariant and the time-varying stage are examined: First, a cascade with the time-invariant stage followed by the time-varying stage, and second, a parallel arrangement of the two stages.

13.5.1 Cascading time-invariant and time-varying beamforming

As illustrated in Figure 13.8, instead of a single beamformer output, $y(n)$, (see Figure 13.3), $M < \ldots \ll N$ beamformer output signals $\mathbf{y}(n) =$

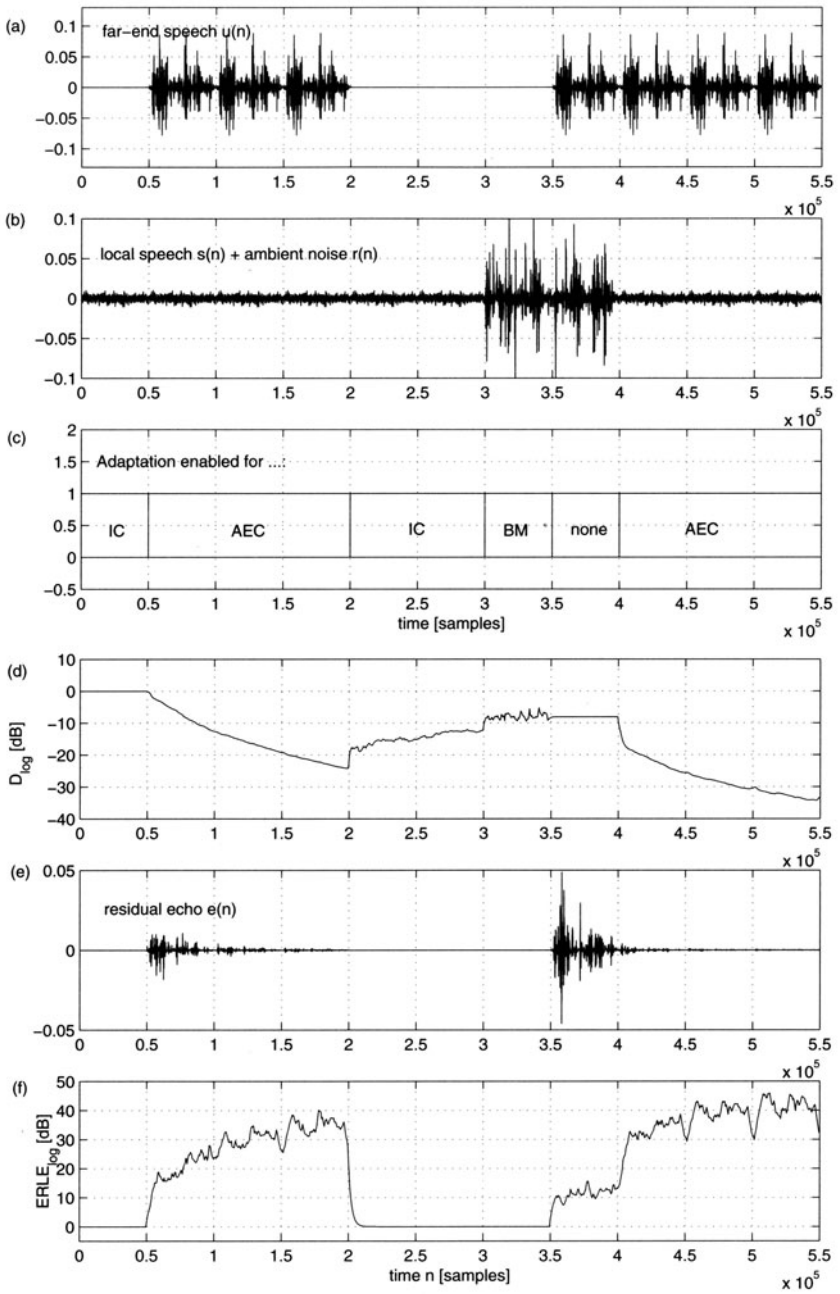


Fig. 13.7. Adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ in 'beamforming first' structure ($N = 8$, $T_{60} \approx 50$ ms, $f_s = 12$ kHz, $L_{AEC+BF} = 300$, $L_{BM} = 16$, $L_{IC} = 50$, adaptation by NLMS algorithm)

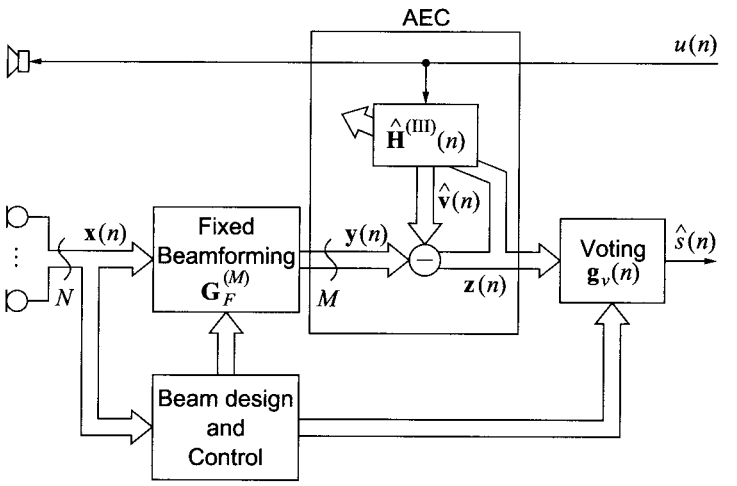


Fig. 13.8. AEC integrated into cascaded beamforming.

$[y_0(n), \ldots y_{M-1}(n)]^T$ are produced by M sets of fixed beamforming filters $\mathbf{G}_F^{(M)}$ according to

$$\mathbf{y}(n) = \mathbf{G}_F^{(M)^T} \cdot \mathbf{X}(n), \tag{13.41}$$

where $\mathbf{X}(n)$ is given by (13.30) and

$$\mathbf{G}_F^{(M)} = [\mathbf{G}_{F,0}^T, \ldots, \mathbf{G}_{F,\mu}^T, \ldots, \mathbf{G}_{F,M-1}^T] \tag{13.42}$$

with $\mathbf{G}_{F,\mu}$ according to (13.28). For AEC, $\widehat{\mathbf{H}}^{(III)}(n)$ realizes M adaptive echo cancellers $\widehat{\mathbf{h}}_\mu(n), \mu = 0, \ldots, M-1$, which exhibit the same performance as $\widehat{\mathbf{h}}^{(II)}(n)$ with time-invariant $\mathbf{G}(n)$ (see Section 13.4.4). Thus, if $M < N$ and $L_{AEC+BF} \approx L_{AEC}$, AEC operates at a reduced computational cost compared to $\widehat{\mathbf{H}}^{(I)}(n)$ (see Section 13.4.3). The time-varying part of the beamforming implements a weighted sum ('voting') using time-varying weights, $g_{v,\mu}(n)$:

$$\widehat{s}(n) = \mathbf{g}_v^T(n) \cdot \mathbf{z}(n) \tag{13.43}$$

with

$$\mathbf{g}_v(n) = [g_{v,0}(n), \ldots, g_{v,\mu}(n), \ldots, g_{v,M-1}(n)]^T, \tag{13.44}$$

$$\mathbf{z}(n) = [z_0(n), \ldots, z_\mu(n), \ldots, z_{M-1}(n)]^T. \tag{13.45}$$

Fixed beamformer design. The fixed beamformers of $\mathbf{G}_F^{(M)}$ may be designed to account for various situations, for instance, different beamformers could be employed for the presence or absence of echo, $\mathbf{v}(n)$, and of certain

local interferers, $\mathbf{r}(n)$. This concept is easily extended to cover several desired sources or moving desired sources, which is especially attractive for teleconferencing [5,17,18,22,26]. For the actual design of $\mathbf{G}_{F,\mu}$, techniques based on both time-invariant or time-varying beamforming can be applied. Updating may be attractive to allow for long-term flexibility.

$\mathbf{G}_F^{(M)}$ *based on time-invariant beamforming.* As a straightforward approach, $M_0 > M$ signal-independent fixed beams may be formed to cover several possible interference scenarios and/or all possible desired source positions. The output of these M_0 beamformers is monitored and a subset of M beamformers is used for $\mathbf{G}_F^{(M)}(n)$ to produce potentially desired signals $\mathbf{y}(n)$. As an example, in a teleconferencing studio with $M_0 = 7$ seats and three local participants being present, only $M = 3$ beams should produce desired signals (for examples see [17,18,22,26]).

$\mathbf{G}_F^{(M)}$ *based on adaptive beamforming.* Signal-dependent adaptive beamforming can be used to identify fixed beamformers for typical interference scenarios. To this end, an adaptive beamformer operates at a normal adaptation rate with its filter coefficients acting as a training sequence for finding M representative fixed beamformers. *A priori* knowledge of the desired source location(s) for incorporating constraints is necessary as well as initial training [5].

Initializing and updating $\mathbf{G}_F^{(M)}$. The monitoring of M_0 fixed beams, or the learning of optimum beamformers for deciding upon $\mathbf{G}_F^{(M)}$ can be carried out during an initial training phase only, or continuously. Continuous monitoring is recommended when changes are expected that demand the updating of $\mathbf{G}_F^{(M)}$. Monitoring of M_0 beams helps also to establish reliable estimates for background noise levels and supports detection of local talker activity so that convergence speed and robustness of AEC adaptation can be improved. Generally, as long as updating of $\mathbf{G}_F^{(M)}$ occurs less frequently than significant changes in the acoustic path, the model of time-invariant beamforming is justified with respect to AEC behavior. Aiming at minimum computational complexity for AEC, more frequent updates of $\mathbf{G}_F^{(M)}$ may be accepted for reduced M. The update should preferably occur at the beginning of 'far-end speech only' periods, as then, the AEC $\hat{\mathbf{H}}^{(III)}(n)$ can immediately adapt to the new echo path.

Voting. The time-varying weights, $g_{v,\mu}(n)$, in (13.44) must be chosen to allow for a fast reaction ($\leq$ 20 ms) to newly active local sources or changing interference scenarios, while at the same time avoiding the perception of switching, e.g., by applying a sigmoïd-like gain increase over time. For maximum spatial selectivity, only one beam signal should have a nonzero weight,

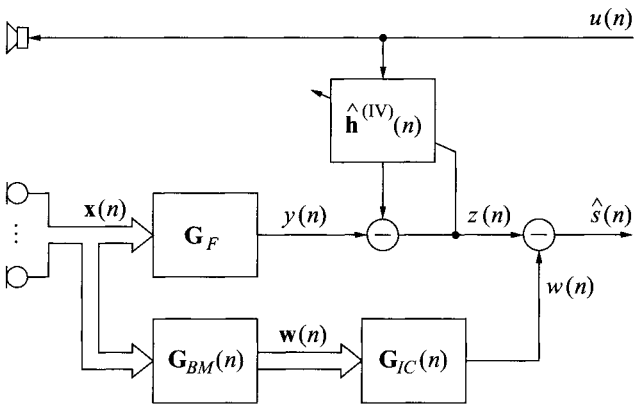


Fig. 13.9. GSC with embedded AEC.

$g_{v,\mu}(n)$, in the stationary case. Frequent toggling between beams is subjectively objectionable and should be prevented by hysteresis mechanisms (see also [17,26]).

13.5.2 AEC with GSC-type beamforming structures

As a popular representative of adaptive beamformers, the GSC (see Section 13.3.3) is also an example for a parallel arrangement of time-varying and time-invariant beamforming. If an integrated AEC should only see time-invariant beamforming in the echo path, it has to act on the output of the fixed beamformer, $y(n)$, as depicted in Figure 13.9 [32]. Obviously, only a single adaptive filter, $\widehat{\mathbf{h}}^{(IV)}(n)$, is necessary which faces the same system identification task as $\widehat{\mathbf{h}}^{(II)}(n)$ for time-invariant beamforming (see Section 13.4.4), which in turn is essentially identical to the plain single-microphone AEC problem. However, residuals of acoustic echoes, $\mathbf{v}(n)$, will also be contained in $w(n)$ unless they are eliminated by $\mathbf{G}_{BM}(n)$ or $\mathbf{G}_{IC}(n)$. Here, leaving echo suppression to the interference canceller, $\mathbf{G}_{IC}(n)$, seems to be the obvious solution. Recall that $\mathbf{G}_{IC}(n)$ minimizes the quadratic norm of $\widehat{s}(n)$ to remove all components from $z(n)$ that are correlated with $\mathbf{w}(n)$. If $\widehat{\mathbf{h}}^{(IV)}(n)$ is perfectly adjusted, no echo components remain in $z(n)$ and the echo estimate within $w(n)$ should be zero. On the other hand, local interference components in $\mathbf{w}(n)$ should be linearly combined using nonzero filter coefficients, so that $w(n)$ can remove interference residuals from $z(n)$. Clearly, a conflict in the design of $\mathbf{G}_{IC}(n)$ arises [32].

For illustration, consider a stationary situation for a given frequency, ω_0, in a 2-D plane containing a linear beamforming array with time-invariant $\mathbf{G}_F$, $\mathbf{G}_{BM}$, and $\mathbf{G}_{IC}$. A local interferer, $\mathbf{r}(n)$, arrives as a planar wave from θ_0 and passes the blocking matrix which is transparent for $\mathbf{r}(n)$ ($\mathbf{G}_{BM}^T \cdot \mathbf{r}(n) = \mathbf{r}(n)$).

Then, for perfect interference cancellation, $\mathbf{G}_{IC}(n)$ has to model the response of the fixed beamformer, $\mathcal{F}\{\mathbf{G}_{IC}\}(\theta_0,\omega_0) = \mathcal{F}\{\mathbf{G}_F\}(\theta_0,\omega_0)$, with $\mathcal{F}\{\mathbf{G}_{(\cdot)}\}(\theta,\omega)$ denoting the frequency response for a plane wave of frequency ω arriving from θ. If, on the other hand, an acoustic echo arrives from the same direction, θ_0, with nonzero spectral support at ω_0, this should be perfectly suppressed if $z(n)$ contains no echo, which means $\mathcal{F}\{\mathbf{G}_{IC}\}(\theta_0,\omega_0) = 0$. Obviously, this conflict requires a compromise at the cost of either local interference suppression or echo attenuation. Here, adaptation algorithms will automatically favor the dominant signal component in $\mathbf{w}(n)$. Even if echo and local interference do not arrive from the same direction, the finite number of degrees of freedom limits the ability of $\mathbf{G}_{IC}$ to suppress echo and local interference simultaneously. This is especially true for reverberant environments which possess a very large (if not infinite) number of DOAs for both echoes and local interference.

To avoid the conflict of interests within $\mathbf{G}_{IC}$, a suppression of the acoustic echoes, $\mathbf{v}(n)$, using $\mathbf{G}_{BM}(n)$ seems attractive. Considering the options, it is obvious that the output, $\mathbf{w}(n)$, should be freed from $\mathbf{v}(n)$ without suppressing $\mathbf{r}(n)$ or impairing the suppression of $\mathbf{s}(n)$. This means that no additional filtering on $\mathbf{x}(n)$ is allowed. As an alternative, estimates for the echoes, $\mathbf{v}(n)$, could be subtracted from $\mathbf{w}(n)$, which requires one adaptive filter for each of the $P \leq N$ channels and is similar to the generic concept of Section 13.4.3.

13.6 Combined AEC and beamforming for multi-channel recording and multi-channel reproduction

Multi-channel recording means that the output of the acquisition part of the acoustic interface in Figure 13.1 consists of several ($L > 1$) channels which, e.g., are necessary to convey spatial information for remote multi-channel sound reproduction, but may also support local signal processing. In Figures 13.5, 13.6, and 13.7 this translates to an L-dimensional output vector $\widehat{\mathbf{s}}(n)$. With respect to the beamforming, it means a duplication of the filtering and adaptation for each channel using the techniques outlined in Section 13.3. Both, time-invariant and adaptive beamforming will typically use L different 'look directions.' Regarding the generic methods to combine AEC with beamforming (Section 13.4), this means that for the 'AEC first' structure, the AEC part, $\widehat{\mathbf{H}}^{(I)}(n)$, remains unchanged while only the beamforming has to be duplicated. For the 'beamforming first' structure, the AEC realized by $\widehat{\mathbf{h}}^{(II)}(n)$ has to be duplicated as well.

When AEC is integrated into cascaded beamforming (see Section 13.5.1) the extension to the multi-channel recording case is included in the concept. The number of parallel fixed beams simply must equate or exceed the number of recorded channels, $M \geq L$, and the voting must be chosen accordingly. The AEC structure, $\widehat{\mathbf{H}}^{(III)}(n)$, remains unchanged. If the AEC is embedded into

a GSC-like structure, both the beamforming, $\mathbf{G}(n)$, and the AEC structure, $\widehat{\mathbf{h}}^{(IV)}(n)$, have to be implemented L times. However, removal of the acoustic echoes in the blocking matrix is necessary only once if performed directly on the microphone signals, $\mathbf{x}(n)$.

Multi-channel reproduction introduces a K-channel AEC problem (as described in Section 13.2.2), wherever a single echo cancellation filter is employed for single-channel reproduction, regardless of whether echo is to be removed from a microphone output or from a beamformer output. Essentially, this deteriorates convergence behavior and increases computational complexity for all structures discussed in Sections 13.4 and 13.5, accordingly.

Finally, for a system with both multi-channel reproduction and multi-channel recording as suggested in Figure 13.1, the complexity for combined AEC and beamforming obeys the superposition principle with respect to filtering and filter adaptation. Synergies are obtained by the common use of control information for several channels. The nature of the problems, however, does not change compared to the basic scenarios studied in Sections 13.2.2, 13.4, and 13.5 so that the corresponding results remain meaningful.

13.7 Conclusions

Beamforming and acoustic echo cancellation have been shown to jointly contribute to the suppression of acoustic feedback occurring in hands-free acoustic man-machine interfaces. While for time-invariant beamforming a single adaptive AEC filter suffices in the case of single-channel reproduction and single-channel recording, time-varying beamformers demand multiple adaptive filters if echo cancellation performance is not to degrade severely. However, realizing a time-varying beamformer as a cascade of time-invariant beamforming and time-varying voting requires only a few adaptive echo cancellers even for microphone arrays with many sensors. Implementing a combination of AEC and beamforming for a multi-channel reproduction and multi-channel recording system involves a corresponding increase in computational complexity. Signal processing performance, however, is still determined by the solutions for the elementary problems.

Acknowledgement

The author wishes to thank Wolfgang Herbordt for providing the simulation results and Susanne Koschny for preparing the illustrations.

References

1. S.L. Gay and J. Benesty, eds., *Acoustic Signal Processing for Telecommunication*, Kluwer, 2000.

2. C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, July 1999.
3. M.M. Sondhi and W. Kellermann, "Echo cancellation for speech signals," in *Advances in Speech Signal Processing*, (S. Furui and M.M. Sondhi, eds.), Marcel Dekker, 1991.
4. A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, pp. 1747–1760, 2000.
5. W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.219–222, Apr. 1997.
6. G.-O. Glentis, K. Berberidis, and S. Theodoridis. "Efficient least squares adaptive algorithms for FIR transversal filtering," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 13–41, July 1999.
7. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.
8. J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
9. W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-88)*, New York NY, USA, pp.2570–2573, Apr. 1988.
10. M.M. Sondhi, D.R. Morgan, and J.L. Hall, "Stereophonic echo cancellation: An overview of the fundamental problem," *IEEE Signal Processing Letters*,vol. 2, no. 8, pp. 148–151, Aug. 1995.
11. S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp.3059–3062, May 1995.
12. J.Benesty, D.R. Morgan, and M.M. Sondhi, "A hybrid mono/stereo acoustic echo canceler," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 468–475, Sept. 1998.
13. B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
14. D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1993.
15. R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice Hall, 1983.
16. R.G. Pridham and R.A. Mucci, "Digital interpolation beamforming for low-pass and bandpass signals," *Proceedings of the IEEE*, vol. 67, no. 6, pp. 904–919, June 1979.
17. W. Kellermann, "A self-steering digital microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-91)*, Toronto, Canada, pp.3581–3584, May 1991.
18. J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985.
19. C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

20. T. Chou, "Frequency-independent beamformer with low response error," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp.2995–2998, May 1995.
21. Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, Dec. 1986.
22. P. Chu, "Desktop mic array for teleconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp.2999–3002, May 1995.
23. M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.239–242, Apr. 1997.
24. P. Chu, "Superdirective microphone array for a set-top videoconferencing system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.235–238, Apr. 1997.
25. L.J. Griffiths and C.W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
26. J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, 1991.
27. G. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, pp. 229–240, 1996.
28. O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix unsing constrained adaptive filters, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-96)*, Atlanta GA, USA, pp.925–928, May 1996.
29. I. Claesson, S.E. Nordholm, B.A. Bengtsson, and P. Eriksson, "A multi-DSP implementation of a broad-band adptive beamformer for use in a hands-free mobile radio telephone," *IEEE Trans. on Vehicular Technology*, vol. 40, no. 1, pp. 194–202, Feb. 1991.
30. S. Oh, V. Viswanathan, and P. Papamichalis, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-92)*, San Francisco CA, USA, pp. I:281–I:284, Mar. 1992.
31. ITU-T recommendation G.167, *Acoustic echo controllers*, Mar. 1993.
32. W. Herbordt and W. Kellermann, "GSAEC - Acoustic echo cancellation embedded into the generalized sidelobe canceller," in *Signal Processing X: Theories and Applications (Proceedings of EUSIPCO-2000)*, Tampere, Finland, vol.3, pp.1843–1846, Tampere, Finland, Sept. 2000.
33. S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson, "Noise reduction using an adaptive microphone array in a car – a speech recognition evaluation," in *Conference Recordings of the ASSP Workshop on Application of Digital Signal Processing to Audio and Acoustics*, New Paltz NY, USA, Oct. 1993.
34. H. Silverman, W. R. Patterson, J.L. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.251–254, Apr. 1997.

35. W. Kellermann, "Some properties of echo path impulse responses of microphone arrays and consequences for acoustic echo cancellation," in *Conf. Rec. of the Fourth International Workshop on Acoustic Echo Control*, Røros, Norway, June 1995.

14 Optimal and Adaptive Microphone Arrays for Speech Input in Automobiles

Sven Nordholm[1], Ingvar Claesson[2], and Nedelko Grbić[2]

[1] Curtin University of Technology, Perth, Australia
[2] Blekinge Institute of Technology, Ronneby, Sweden

Abstract. In this chapter, speech enhancement and echo cancellation for hands-free mobile telephony are discussed. A number of microphone array methods have been tested and results from car measurements are given. Traditional methods such as linearly constrained beamforming and generalized sidelobe cancelers are discussed as well as array gain optimization methods. An *in situ* calibrated method which gives an overall improved performance is also presented. Algorithms such as Least-Squares (LS) and Normalized-Least-Mean-Squares (NLMS) are used to find optimal weights. Improved performance using an LS-method is shown, but at the cost of increased numerical complexity limiting its implementation in real-time applications. By introducing subband processing this issue can be avoided. The results show a noise suppression of approximately 18 dB and hands-free loudspeaker suppression of the same order.

14.1 Introduction: Hands-Free Telephony in Cars

The increased use of mobile telephones in cars has created a greater demand for hands-free, in-car installations. The advantages of hands-free telephones are safety and convenience. In many countries and regions hand-held telephony in cars is prohibited by legislation. The car manufacturers also prohibit such use since it will interact with other electronic devices in the car such as air bags, navigation equipment, etc. This means that a mobile telephone should be properly installed and an external antenna should be used. However, by installing the microphone far away from the user, a number of disadvantages, such as poor sound quality and acoustic feedback from the far-end side, are introduced. This means that some form of filtering is required in order to obtain sound quality comparable to that of hand-held telephony. This filtering operation must suppress the loudspeaker, as well as background noise and room reverberation, without causing severe speech distortion. A number of potential methods will be presented to address this problem.

For automobile applications, there has also been the desire to replace many hand-controlled functions with voice controls. The signal degradations in this context have many similarities to those encountered in distant-talker speech recognition applications. A study of recognition in car environments was presented in [1,2]. However this topic is beyond the primary goal here and is the specific subject of Chapter 15.

Hands-free car installations result in noisy near-end speech as well as an acoustic feedback of the far-end speech. The near-end disturbances, resulting in substantial speech distortion, are mainly room reverberation and car noise. Furthermore, acoustic feedback, generated at the near-end side, is a problem for the far-end side talker, who will hear his voice echoed with 100-200 ms delay, making speech conversation substantially more difficult. Three major tasks must be addressed in order to improve the quality of hands-free mobile telephones: noise suppression, room reverberation suppression, and acoustic feedback suppression of the hands-free loudspeaker. Because of the cabin conditions, room reverberation suppression is not a critical issue in most standard automobiles. In trucks, buses, people movers and 4-WD with their larger interiors, it may need to be considered. The measurements presented here are from a normal-sized station wagon. The acoustic channel is non-minimum phase and thus quite hard to deconvolve [3]. Matched filtering approaches which do not require explicit channel deconvolution [3,4] and several other methods detailed earlier in this text are available for reverberation suppression under more adverse conditions.

Speech enhancement in hands-free mobile telephony can be performed using spectral subtraction [5–8] or temporal filtering such as Wiener filtering, noise cancellation and multi-microphone methods using a variety of different array techniques [9–11]. Room reverberation in this context is most effectively handled using array techniques or by proper microphone design and placement. Acoustic feedback for hands-free mobile telephony is usually addressed by conventional echo cancellation techniques [12–15] although subband echo cancellation has been popular lately, see for instance [15,16].

A broad-band microphone array is able to perform all the given tasks, i.e. speech enhancement, echo cancellation and reverberation suppression, in a concise and effective manner. This is due to the fact that the spatial domain is utilized as well as the time domain. An effective combination of spatial and temporal processing will lead to a very efficient solution. Hence, improved speech enhancement performance is achieved compared to single microphone solutions. The microphone array technique also handles the acoustic feedback in an efficient way. The hands-free loudspeaker represents a single source despite having been filtered by the channel associated with the car's interior. Similarly, the main talker (driver) represents an additional single source within the cabin. These two sources will have different locations. The echo-suppression level and speech distortion will depend on how "well apart" these two sources are placed [17].

The outline of this chapter is as follows:

- Section 2, Optimum and Adaptive Beamforming topics are reviewed from a hands-free mobile telephone perspective, specifically:
  1. Signal Model
  2. Constrained Minimum Variance Beamforming and Generalized Sidelobe Canceler (GSC)

3. *In situ* Calibrated Microphone Array
4. Time-Domain Minimum-Mean-Square-Error Beamformer
5. Frequency-Domain Minimum-Mean-Square-Error Beamformer
6. Optimal Near-field Signal-to-Noise plus Interference Beamformer (SNIB)

- Section 3, Subband Implementation of the Microphone Array
  1. Description of LS-Subband Beamforming
- Section 4, Multi-resolution Time-Frequency Adaptive Beamforming
  1. Complexity Comparisons
- Section 5, Evaluation and examples
- Section 6, Summary and conclusions

14.2 Optimum and Adaptive Beamforming

The hands-free mobile telephony problem in an automobile is well suited for optimum or adaptive beamforming. The user is in a fixed and known location relative to the array and enclosure. The geometrical array configuration is known. It is further possible to place the loudspeaker in a position that is advantageous from a beamforming perspective. Early approaches to this task involved the direct adoption of adaptive antenna array methods in use since the 1960's [18]. However, this proved not to be a straightforward task and required the development of approaches specific to the application environment, e.g. [11,19,10,20].

The most common means of applying adaptive beamforming concepts is to treat the problem as one of constrained optimization. These methods rely on geometrical constraints, where the location of the source is known either perfectly or with some accuracy. They require sensor calibration and stable hardware (e.g. to avoid low temperature drift). The algorithms may be extended using different robustness constraints [21–23] and noise sub-space constraints [3]. The problem may also be viewed as several multi-dimensional filtering problems. These filters are combined with an interference cancelling structure [10,24,25].

14.2.1 Common Signal Modeling

In order to provide a consistent description it will be useful to introduce a simple signal model which is general in the sense that microphone elements and sources with any spectral content can be placed arbitrarily. The D different point signal sources $s_d(t)$, $d = 1, 2, \ldots, D$, with spectral densities $R_{s_d s_d}(\omega)$ are assumed to be mutually uncorrelated, i.e. the cross power spectral density $R_{s_d s_e}(\omega)$ is zero if $d \neq e$. All sources impinge on an array of N microphone elements, corrupted with non-directional independent diffuse additive noise $v(t)$. The transfer function between source d and array element n is denoted $G_{d,n}(\omega)$ and is either measured or modeled. In this model, a spherical source

in a free-field and homogeneous medium is assumed. In a real world situation with measured data such an assumption may be suspect. The signal received at the nth microphone element, $x_n(t)$, is

$$x_n(t) = \sum_{d=1}^{D} s_d(t) \star g_{d,n}(t) + v(t) \quad n = 1, 2, \ldots, N, \tag{14.1}$$

where $\star$ denotes convolution. Each source signal is treated as a point source filtered by an LTI system. An implication of these assumptions is that variations in the acoustic channel are slow relative to the filter update rate. In the sequel all signals are assumed to be bandlimited and sampled with a discrete-time index k.

14.2.2 Constrained Minimum Variance Beamforming and the Generalized Sidelobe Canceler

In minimum variance beamforming, the objective is to minimize the output of a (broadband) array while maintaining a constant gain constraint towards the desired source, in this case the talker of interest.

The output of the beamformer is given as

$$y[k] = \sum_{n=1}^{N} \mathbf{w}_n^H \mathbf{x}_n[k] \tag{14.2}$$

where the weight vector and input data vector both are of length L.

The expression to be minimized is the "variance" of the assumed zero-mean output, $E(|y[k]|^2)$, with respect to the filter weights given by

$$r_{yy}[0] = E(y[k]y^H[k]) = \textstyle\sum_{n=1}^{N} \mathbf{w}_n^H E(\mathbf{x}_n[k]\mathbf{x}_n^H[k])\mathbf{w}_n. \tag{14.3}$$

If it is assumed, without loss of generality, that point source one is the talker of interest, then the major task is to find the constraint on the weight vector such that $y[k] = s_1(t)|_{t=kT}$, i.e. the output is distortion free. A natural way to do this is to express the minimization in the frequency domain and include matched filtering [3]. For this process, there is a strict requirement of accurate signal modeling or robust constraints, otherwise super-resolution will cancel the source [23,10].

The Generalized Sidelobe Canceler can be viewed as a constrained beamformer which has been converted to a non-constrained beamformer by means of a blocking matrix. Thus, the problem is separated into two tasks: the design of a fixed beamformer to determine the response for the desired source and a matrix filter that blocks the desired source from entering. In the simplest case of a free-field, far-field source and a perfectly calibrated array this blocking matrix will amount to a point constraint [26,27]. For the near field situation and a reverberant enclosure, special measures must be taken. The

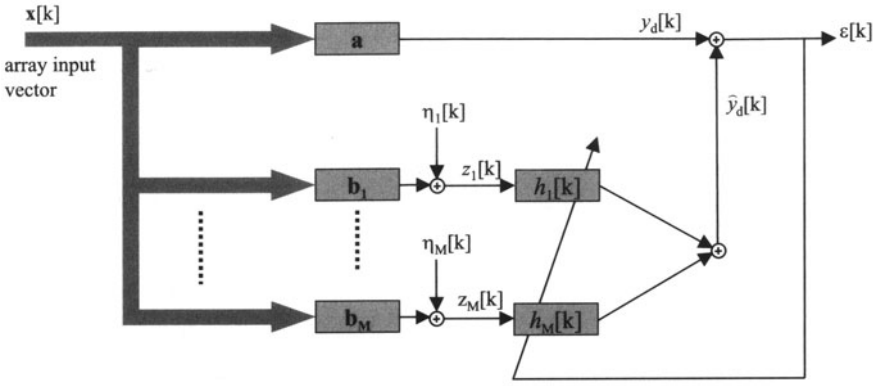


Fig. 14.1. Structure of the Generalized Sidelobe Canceler.

original form of the GSC only provides for a point constraint and is excessively sensitive to calibration and direction errors [23,22]. A number of methods have been proposed which are more suitable for microphone array applications [27,28,21,10,29]. Details of this appropriate GSC implementation will now be presented.

The GSC structure shown in Figure 14.1 consists of two main parts: an upper fixed beamformer and a blocking matrix with subsequent interference cancelers. In order to avoid attenuation of the desired signal it is critical that the input to these interference cancelers contain only the undesired signals.

The input signal vector, $\mathbf{x}[k]$, is filtered by the upper beamformer, $\mathbf{a}$, steering towards the talker of interest and creating an output $y_d[k]$,

$$y_d[k] = \mathbf{a}^H \mathbf{x}[k]. \tag{14.4}$$

This beamforming filter in its simplest form consists of a vector of ones. More generally, it consists of FIR filters forming a multi-dimensional filter. The blocking matrix should form signals that are orthogonal to the desired signal. Thus, the input to the interference cancelers should contain only undesired signals, (and some injected noise)

$$z_m[k] = \mathbf{b}_m^T \mathbf{x}[k] \quad (+\eta_m[k]). \tag{14.5}$$

When designing the lower beamformers, $\mathbf{b}_m$, which implement the signal blocking matrix, the requirement is that the desired signal should be blocked totally. This is not practically feasible. To do so would require knowledge of the transfer function from the desired source to the input of the lower beamformers with extremely high precision. By choosing to relax this requirement and viewing the problem from a filter design perspective where the desired signal should be suppressed below a certain level determined by an artificial injected noise level, $\eta_m[k]$, one may overcome these limitations [10]. This injected noise is not actually present, it is only included in the filter weight

updating algorithm used in the adaptive implementation of the interference canceler. The desired signal will not be picked up and attenuated by the interference canceler as long as the injected noise dominates over the desired signal. This approach is also valid for the background noise free case [25].

Another approach to this constrained optimization problem is the use of subspace techniques such as that suggested in [3]. This method requires several adaptive steps and also a Voice Activity Detector (VAD). The speech distortion is related to how well the transfer functions from the desired source to each microphone element, $G_{1i}(\omega)$, are identified. The upper beamformer is then created as a matched spatial temporal filter and the blocking matrix is created as a projection matrix that is orthogonal to the transfer function vector $(G_{11}(\omega), G_{12}(\omega), \ldots, G_{1N}(\omega))$. This implies that, as long as this orthogonality constraint is valid, no target signal will leak through. All of this assumes that a good estimate of the transfer function vector is used, the talker can be represented by a point source, and the conditions are time invariant.

Experience using the GSC has shown that it provides a very good suppression of background noise, but that control of the signal distortion and calibrating for a combined array are problematic [10]. Another observation reached from implementation experience is the importance of using a precise VAD. The interference canceler is very effective at exploiting correlations with the target and adjusting its weights to suppress or heavily distort the desired signal. A combination of VAD and leaky LMS was used in the implementation [10] to give a reasonable result. Still it was difficult to obtain satisfactory results with long term tests in a car, i.e. over a few days of measurements using an initial calibration. This suggests the need to have a means for very simple *in situ* calibration.

14.2.3 *In Situ* Calibrated Microphone Array (ICMA)

The basic idea when developing this scheme was to find a robust yet effective strategy to an undistorted version of the desired signal with significant suppression of background noise and unwanted sources. A primary goal was to overcome the environmental sensitivity inherent in the constrained optimization strategies outlined above. A way to achieve this is to record calibration sequences through the actual system in a real situation with all of its imperfections. These calibration sequences contain information regarding the statistical properties of the speaker, from both a spatial and temporal point of view. All calibration signals are gathered from the correct position and with the actual hardware. The adaptive system, as such, is then designed not to suppress signals close to the calibration point, i.e it should have low sensitivity to perturbation errors and avoid super-resolution. This can be achieved by moving the source (spatial dithering) around the nominal point during calibration or exploiting temporal dithering in the A/D converters. Calibration sequences are recorded from both the jammer position(s) and the target position when *no car noise is present.* These signals are stored in memory for

later use as training signals in an adaptive phase. As will be shown it is only necessary to save the second order statistics of the calibration signals in the implementation phase. This approach gives an inherent calibration where it is possible to average and weigh interesting frequency bands, microphones, and spatial points. The methodology does not rely on any geometric *a priori* information or array element similarities with accurate positioning. The result is a system that is tailored for the actual situation. The system has been studied from a theoretical [17] and implementation perspectives [30–32]. The ICMA uses a Minimum-Mean-Square-Error (MMSE) optimization that is approximated by either an NLMS implementation [30,31] or an LS solution [33]. An LS or Recursive-Least-Squares (RLS) solution becomes practical when using a subband implementation. The ICMA design can be viewed as an MMSE beamformer where there is separate access to the undesired noise and desired speech signal.

14.2.4 Time-Domain Minimum-Mean-Square-Error Solution

Assume that the input to the beamformer consists of a sum of known calibration sequence observations, $s_n[k]$, $n = 1 \ldots N$, sent out from the position of interest, and noise-plus-interference signals, $x_n[k]$, $n = 1 \ldots N$, consisting of the actual environment signal observations. The time-domain objective can be formulated as

$$\mathbf{w}_{opt} = \arg\min_{\mathbf{w}} \quad E\left[(y[k] - s_r[k])^2\right] \tag{14.6}$$

where the output, $y[k]$, from the beamformer is given by

$$y[k] = \sum_{n=1}^{N} \mathbf{w}_n^H \left(\mathbf{x}_n[k] + \mathbf{s}_n[k]\right). \tag{14.7}$$

The desired signal, $s_r[k]$, is chosen from a single calibration array sensor observation, $s_n[k]$, or a separate reference microphone signal chosen as the reference sensor. In theory the true source signal would be desirable to use instead of a sensor observation, but the true source signal is simply not accessible in a noise-filled car. The optimal weights which minimizes the mean square error between the output and the reference signal are found by [34]

$$\mathbf{w}_{opt} = \left[\mathbf{R_{ss}} + \mathbf{R_{vv}}\right]^{-1} \mathbf{r}_s. \tag{14.8}$$

where $\mathbf{R_{ss}}$ is defined as

$$\mathbf{R_{ss}} = \begin{pmatrix} \mathbf{R}_{s_1 s_1} & \mathbf{R}_{s_1 s_2} & \cdots & \mathbf{R}_{s_1 s_N} \\ \mathbf{R}_{s_2 s_1} & \mathbf{R}_{s_2 s_2} & \cdots & \mathbf{R}_{s_2 s_N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{s_N s_1} & \mathbf{R}_{s_N s_2} & \cdots & \mathbf{R}_{s_N s_N} \end{pmatrix} \tag{14.9}$$

and

$$\mathbf{R}_{s_m s_n} = \begin{pmatrix} r_{s_m s_n}[0] & r_{s_m s_n}[1] & \dots & r_{s_m s_n}[L-1] \\ r^*_{s_m s_n}[1] & r_{s_m s_n}[0] & \dots & r_{s_m s_n}[L-2] \\ \vdots & \vdots & \ddots & \vdots \\ r^*_{s_m s_n}[L-1] & r^*_{s_m s_n}[L-2] & \dots & r_{s_m s_n}[0] \end{pmatrix} \tag{14.10}$$

with

$$r_{s_m s_n}[l] = E\{s_m[k]s_n[k+l]\} \qquad l = 0, 1, \cdots, L-1 \tag{14.11}$$

The noise correlation matrix, $\mathbf{R_{vv}}$, is defined similarly and consists of the correlation estimates of all noise plus interference signals. The filter weights, $\mathbf{w}$, are arranged according to

$$\mathbf{w}^T = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ \dots \ \mathbf{w}_N^T] \tag{14.12}$$

where

$$\mathbf{w}_n^T = [w_n[0] \ w_n[1] \ \dots \ w_n[L-1]] \qquad n = 1, 2, \cdots, N. \tag{14.13}$$

The cross correlation vector, $\mathbf{r_s}$, is defined as

$$\mathbf{r_s} = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \dots \quad \mathbf{r}_N] \tag{14.14}$$

with

$$\mathbf{r}_n = [r_n[0] \quad r_n[1] \quad \dots \quad r_n[L-1]] \qquad n = 1, 2, \cdots, N \tag{14.15}$$

and each element as

$$r_n[l] = E[s_n[k]s_r[k+l]] \tag{14.16}$$

$$n = 1, 2, \cdots, N, \quad r \in [1, 2, \cdots, N], \quad l = 0, 1, \cdots, L-1.$$

14.2.5 Frequency-Domain Minimum-Mean-Square-Error Solution

The formulation of the MMSE beamformer can be expressed in terms of individual frequency bands. The optimal beamformer consists of the frequency-dependent weights that minimize the mean square error across the individual frequency bands. This is provided that the different frequency bands are essentially independent and that the fullband signal can be represented accurately via this subband decomposition. The frequency domain design criterion

can be formulated in a fashion similar to that of the time domain. For each subband with center frequency, f, the criterion will be

$$\mathbf{w}_{opt}^{(f)} = \arg\min_{\mathbf{w}^{(f)}} \quad E\left[|y^{(f)}[k] - s_r^{(f)}[k]|^2\right] \tag{14.17}$$

where the output, $y^{(f)}[k]$, from the beamformer is given by

$$y^{(f)}[k] = \sum_{n=1}^{N} w_n^{(f)} \left[x_n^{(f)}[k] + s_n^{(f)}[k]\right] \tag{14.18}$$

where $x_n^{(f)}[k]$, $s_n^{(f)}[k]$ and, $y^{(f)}[k]$ are narrow-band signals containing essentially only components of frequency f. The single sensor observation, $s_r[k]^{(f)}$, is again one of the microphone observations chosen as the reference sensor. The optimal weights, which minimize the mean square error between the output and the reference signal for each frequency band, are found by

$$\mathbf{w}_{opt}^{(f)} = \left[\mathbf{R}_{ss}^{(f)} + \mathbf{R}_{vv}^{(f)}\right]^{-1} \mathbf{r}_s^{(f)}. \tag{14.19}$$

where

$$\mathbf{R}_{ss}^{(f)} = E\{\mathbf{s}^{(f)}[k]\mathbf{s}^{(f)}[k]^H\} \tag{14.20}$$

where

$$\mathbf{s}^{(f)}[k] = \left[s_1^{(f)}[k] \quad s_2^{(f)}[k] \quad \cdots \quad s_N^{(f)}[k]\right]^T. \tag{14.21}$$

Here each signal, $s_n^{(f)}[k], \quad n = 1, 2, \cdots, N$, is the narrow-band observation when only the source signal of interest is active. The correlation matrix $\mathbf{R}_{vv}^{(f)}$ is defined similarly where each microphone observation consists of only the noise and interference signals. The cross correlation vector, $\mathbf{r}_s^{(f)}$, is defined as

$$\mathbf{r}_s^{(f)} = [r_1^{(f)} \quad r_2^{(f)} \quad \ldots \quad r_N^{(f)}] \tag{14.22}$$

with each element as

$$r_n^{(f)} = E[s_n^{(f)}[k]s_r^{(f)}[k]] \tag{14.23}$$

$$n = 1, 2, \cdots, N, \quad r \in [1, 2, \cdots, N].$$

The frequency dependent weights, $\mathbf{w}^{(f)}$, are defined as complex valued vectors of dimension N.

14.2.6 Optimal Near-Field Signal-to-Noise plus Interference Beamformer

The output signal-to-noise plus interference power ratio (SNIR) is defined as

$$Q = \frac{\text{average signal output power}}{\text{average noise-plus-interference output power}} \tag{14.24}$$

and the beamformer which maximizes the ratio, Q, is the optimal SNIB. Expressing the mean signal output power as a function of the filter weights in the beamformer and finding the optimal weights which maximize Q is done below.

Time-Domain Formulation The beamformer output power when only the signal of interest, $s[k]$, is active, is found from the zero lag of the autocorrelation function, $r_{y_s y_s}[0]$, as

$$r_{y_s y_s}[0] = \mathbf{w}^H \mathbf{R_{ss}} \mathbf{w} \tag{14.25}$$

The matrix, $\mathbf{R_{ss}}$, is defined in (14.9). The weights, $\mathbf{w}$, are arranged as in (14.12) and (14.13).

An expression for the noise-plus-interference power, $r_{y_v y_v}[0]$, is found from

$$r_{y_v y_v}[0] = \mathbf{w}^H \mathbf{R_{vv}} \mathbf{w} \tag{14.26}$$

when all the surrounding noise sources are active and the source signal of interest is inactive.

Now, the optimal weights are found by maximizing the ratio of two quadratic forms, according to

$$\mathbf{w}_{opt} = \arg\max_{\mathbf{w}} \left\{ \frac{\mathbf{w}^H \mathbf{R_{ss}} \mathbf{w}}{\mathbf{w}^H \mathbf{R_{vv}} \mathbf{w}} \right\}. \tag{14.27}$$

Frequency-Domain Formulation The formulation of the optimal signal-to-noise plus interference beamformer may be derived for individual frequency subbands. The weights that maximize the quadratic ratios at individual frequencies constitute the optimal beamformer that maximizes the total output power ratio, provided the subband signals are independent.

For frequency, f, the quadratic ratio between the output signal power and the output noise-plus-interference power is

$$\mathbf{w}_{opt}^{(f)} = \arg\max_{\mathbf{w}^{(f)}} \left\{ \frac{\mathbf{w}^{(f)H} \mathbf{R}_{ss}^{(f)} \mathbf{w}^{(f)}}{\mathbf{w}^{(f)H} \mathbf{R}_{vv}^{(f)} \mathbf{w}^{(f)}} \right\} \tag{14.28}$$

where the matrices, $\mathbf{R}_{ss}^{(f)}$ and $\mathbf{R}_{vv}^{(f)}$, are defined as in (14.20). The weights, $\mathbf{w}^{(f)}$, are defined as the complex valued vectors of dimension N.

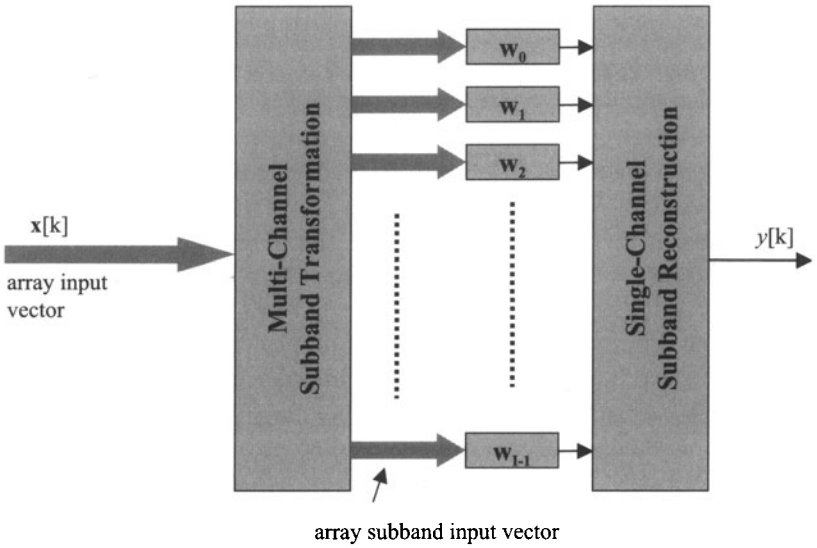


Fig. 14.2. Subband Beamforming Structure.

14.3 Subband Implementation of the Microphone Array

Noise and echo suppression exhibit significant gains when an LS solution is used in place of the NLMS algorithm [33]. However, computational considerations make use of the LS criterion impractical for the wide-band problem. Subband frequency transformations as shown in Figure 14.2 provide efficient means of allowing for the use of second order methods (such as RLS), while keeping computational complexity low. The frequency-domain algorithms have a least-squares objective function, as described in (14.29).

An uniform DFT analysis-synthesis filterbank [35] will be employed here. The filterbank is used to decompose the full-rate sampled signals, $x_n[k]$, into I subband signals [36]. The subband signals are essentially generated from a common bandpass filter with varying center frequency, $\frac{2\pi i}{I}$, and cover the entire frequency range. As a special case, when the number of subbands equals the length of the prototype filter, the subband decomposition will equal the overlapped Short-Time Fourier Transform (STFT) and the prototype filter is chosen as a simple, uniform moving average. The subband signals are decimated to a lower sampling rate allowing for a polyphase implementation. This provides an analysis-synthesis structure with approximately the same computational complexity as the STFT [35].

14.3.1 Description of LS-Subband Beamforming

The MMSE beamforming scheme formulated in (14.17) may be reexpressed in the time domain using an LS formulation as subband number

$$\mathbf{w}_{ls,opt}^{(i)}(N) = \arg\min_{\mathbf{w}^{(i)}} \left\{ \sum_{k=0}^{K-1} \left[|y^{(i)}[k] - s_r^{(i)}[k]|^2 \right] \right\} \tag{14.29}$$

where i indicates the subband index, K is the number of data points considered, and where $y^{(i)}[k]$ is given by (14.18) with $f = 2\pi i/I$. The reference source signal, $s_r^{(i)}[k]$, is not directly available, but a calibration sequence gathered in a quiet environment can be used in its place. This calibration signal contains the source's temporal and spatial information. Since $s_r^{(i)}[k]$ is independent of the actual data $x_n^{(i)}[k]$, at least for large K, the LS problem can be expressed as the sum of two components by

$$\mathbf{w}_{ls,opt}^{(i)}(K) = \arg\min_{\mathbf{w}^{(i)}} \left\{ \sum_{k=0}^{K-1} \left[|\mathbf{w}^{(i)H}\mathbf{s}^{(i)}[k] - s_r^{(i)}[k]|^2 + |\mathbf{w}^{(i)H}\mathbf{x}^{(i)}[k]|^2 \right] \right\}. \tag{14.30}$$

The equation may be rewritten as

$$\mathbf{w}_{ls,opt}^{(i)}(K) = \arg\min_{\mathbf{w}^{(i)}} \left\{ \mathbf{w}^{(i)H} \left[\hat{\mathbf{R}}_{ss}^{(i)}(K) + \hat{\mathbf{R}}_{xx}^{(i)}(K) \right] \mathbf{w}^{(i)} - \mathbf{w}^{(i)H}\hat{\mathbf{r}}_s^{(i)}(K) - \hat{\mathbf{r}}_s^{(i)H}(K)\mathbf{w}^{(i)} + \hat{r}_{s_r}^{(i)} \right\} \tag{14.31}$$

where the source correlation estimates can be precalculated in the calibration phase from

$$\begin{aligned} \hat{\mathbf{R}}_{ss}^{(i)}(K) &= \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{s}^{(i)}[k]\mathbf{s}^{(i)H}[k] \\ \hat{\mathbf{r}}_s^{(i)}(K) &= \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{s}^{(i)}[k]s_r^{(i)H}[k] \end{aligned} \tag{14.32}$$

where

$$\mathbf{s}^{(i)}(K) = [s_1^{(i)}[k], \quad s_2^{(i)}[k], \quad \cdots \quad s_N^{(i)}[k]]^T$$

is the microphone data when the source signal alone is active. The least-squares minimization of (14.31) is found by

$$\mathbf{w}_{ls,opt}^{(i)}(K) = \left[\hat{\mathbf{R}}_{ss}^{(i)}(K) + \hat{\mathbf{R}}_{xx}^{(i)}(K) \right]^{-1} \hat{\mathbf{r}}_s^{(i)}(K) \tag{14.33}$$

where

$$\hat{\mathbf{R}}_{xx}^{(i)}(K) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{x}^{(i)}[k]\mathbf{x}^{(i)H}[k] \tag{14.34}$$

is the observed data correlation matrix estimate. This implies that an estimate of the calibration data may be used as part of the solution.

14.4 Multi-Resolution Time-Frequency Adaptive Beamforming

The performance of the algorithm stated in previous section requires that the number of subbands is large enough for the frequency-domain representation to be accurate. The number of subbands is proportional to the length of the equivalent time-domain filters, and the more subbands chosen, the more degrees of freedom inherent in the beamformer. The number of subbands is also related to the delay caused by the frequency transformations since a large number of subbands necessitates a longer prototype filter, which in its turn will cause an increased delay.

The algorithm is easily extended to a combination of time- and frequency-domain representations. Each subband signal can be seen as a time-domain signal sampled at a reduced sampling rate and containing only frequencies in that particular subband. By applying a time-domain algorithm in each subband, the degrees of freedom for the filters are increased while the number of subbands can be held constant. The lengths of the corresponding filters may differ across the subbands to produce a multi-resolution framework.

14.4.1 Memory Saving and Improvements

The proposed beamformer consists of a source signal information gathering phase followed by the operation phase. Information about the source signal is represented through the frequency-dependent, source-only correlation matrix estimates, $\hat{\mathbf{R}}_{ss}^{(i)}$. These estimates are calculated and stored for each of the I subbands. When there are known unwanted sources, such as hands-free loudspeakers, which have a fixed location in relation to the microphones and the enclosure, correlation estimates from these signals are also estimated and saved. Estimates of the frequency-dependent cross-correlation vectors, $\hat{\mathbf{r}}_s^{(i)}$, are also maintained. The number of elements, P, required in memory to store the fullband time-domain solution is:

$$P^{time} = [NL(NL+2)]^2$$

where N is the number of microphone channels and L is the fullband FIR filter length. For the frequency-domain representation the number of storage elements needed is

$$P^{freq} = I[N\frac{L}{I}(N\frac{L}{I}+2)]^2$$

where index I is the number of subbands. As an example, Figure (14.3), shows the ratio of the number of storage elements required for the time- and the frequency-domain implementations as a function of the fullband time-domain filter length and subbands values of $I = 16, 32, 64, 128, 256, 512$. The number of channels is $N = 6$. Even for moderate filter lengths, the size of the mem-

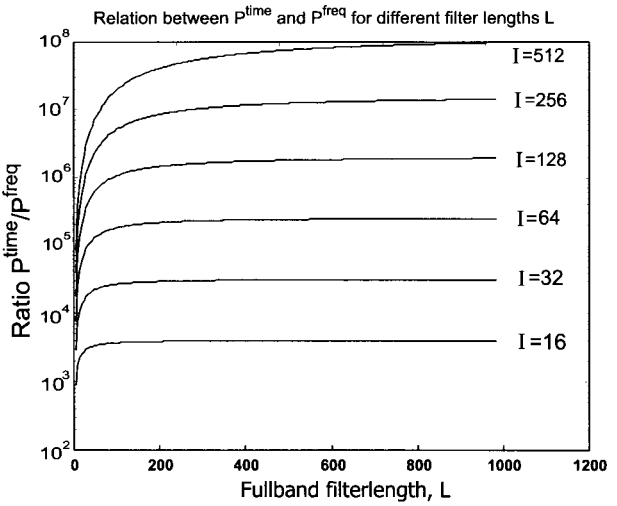


Fig. 14.3. The ratio P^{time}/P^{freq} for filter lengths L varying from 4 to 1024, and the number of subbands, I, is varying from 16 to 512. The number of channels is $N = 6$.

ory is reduced substantially with the frequency-domain implementation. The number of multiplications is proportional to the number of stored elements, and the relationship between computational costs for the time-domain and frequency-domain implementations is the same as in Figure 14.3.

14.5 Evaluation and Examples

14.5.1 Car Environment

A performance evaluation of the beamformer was made in a hands-free situation with a six-element microphone array mounted on the passenger-side visor of a Volvo station wagon. Data was gathered on a multi-channel DAT-recorder with a 12 kHz sampling rate and a 300-3400 Hz bandwidth. In order to facilitate simultaneous driving and recording, an artificial talker was mounted in the passenger seat to simulate a real person engaging in a conversation. Initially, a white noise sequence within the bandwidth was emitted from the artificial talker, in a non-moving car with the engine turned off. This sequence served as the desired sound source calibration signal in all of the following simulations. Interference signals were recorded by emitting an independent sequence of bandlimited, white noise from the hands-free loudspeaker. This recording functioned as the point-source interference calibration signal and was referred to as the echo signal. In order to gather background noise estimates, the car was driven at a speed of 110 km/h over a paved road. The car cabin noise environment consisted of a number of unwanted sound sources,

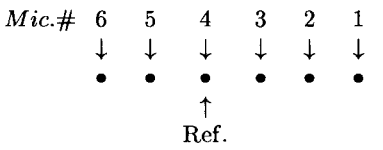


Fig. 14.4. Geometry of the six-element, linear array with an adjacent microphone spacing of 5 cm.

mostly with broad spectral content, e.g. wind and tire noise. Recordings with real speech signals, from both the artificial talker and the hands-free loudspeaker, were recorded both individually and while driving. These recordings served as the beamformer evaluation signals. All of the performance measures presented in Section 14.5.5 were based on these real speech recordings.

14.5.2 Microphone Configurations

The sensors used in this evaluation were high-quality Sennheiser microphones mounted flat on the visor. The speaker was positioned 35 cm from the center of the array and oriented perpendicular to the its axis. The mounting of the six-element linear array is given in Figure 14.4. The spacing between adjacent elements in the array was 5 cm.

14.5.3 Performance Measures

There are two objectives for the beamformer: minimize the distortion caused by the beamforming filters (measured by the deviation between the beamformer output and the source signal) and maximize noise and interference suppression. In order to measure the performance the normalized distortion quantity, D, is introduced as

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} |C_d \hat{P}_{y_s}(w) - \hat{P}_{x_s}(w)| dw \tag{14.35}$$

where $w = 2\pi f$ and f is the normalized frequency. The constant, C_d, is defined as

$$C_d = \frac{\int_{-\pi}^{\pi} \hat{P}_{x_s}(w) dw}{\int_{-\pi}^{\pi} \hat{P}_{y_s}(w) dw} \tag{14.36}$$

where $\hat{P}_{x_s}(w)$ is a spectral power estimate of a single sensor observation and $\hat{P}_{y_s}(w)$ is the spectral power estimate of the beamformer output when the source signal alone is active. The constant C_d normalizes the mean output spectral power to that of the single sensor spectral power. The single sensor

observation is chosen as the reference microphone observation, i.e. microphone 4 in the array. This distortion measure is essentially an estimate of the mean output spectral power deviation from the observed single sensor spectral power, and under ideal circumstances will be zero.

To measure normalized noise suppression, the quantity, S_N, is computed from

$$S_N = C_s \frac{\int_{-\pi}^{\pi} \hat{P}_{y_N}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{x_N}(w)dw} \tag{14.37}$$

where

$$C_s = \frac{1}{C_d} \tag{14.38}$$

and $\hat{P}_{y_N}(w)$ and $\hat{P}_{x_N}(w)$ are the spectral power estimates of the beamformer output and the reference sensor observation, respectively, when the surrounding noise alone is active.

Similarly, the normalized interference suppression quantity, S_I, is given by

$$S_I = C_s \frac{\int_{-\pi}^{\pi} \hat{P}_{y_I}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{x_I}(w)dw} \tag{14.39}$$

with $\hat{P}_{y_I}(w)$ and $\hat{P}_{x_I}(w)$ being the spectral power estimates when the interference and desired signals, respectively, are active alone. Both of these suppression measures are normalized to the amplification (or attenuation) caused by the beamformer relative to the reference sensor observation when the source signal is active alone. Accordingly, when the beamformer scales the source signal by a specific amount, the noise and interference suppression quantities are adjusted appropriately.

14.5.4 Spectral Performance Measures

In order to evaluate the performance within individual subbands, the above definitions may be made frequency-dependent by omitting the integration operations, i.e.

$$S_N(w) = \frac{C_s \hat{P}_{y_N}(w)}{\hat{P}_{x_N}(w)} \tag{14.40}$$

and

$$S_I(w) = \frac{C_s \hat{P}_{y_I}(w)}{\hat{P}_{x_I}(w)} \tag{14.41}$$

where the definition of C_s and the power spectral estimates are the same as above.

In practice, the above measures were calculated using Welch's averaged periodogram spectral estimation method with non-overlapping Hanning windows of length 256. The integrals were approximated by discrete summation over the periodograms. All measures were calculated from the time-domain signals, which implies that any distortions created by the frequency transformations were also taken into account.

14.5.5 Evaluation on car data

In this evaluation, 8 s white noise calibration signals were used. These were emitted individually from the artificial talker and the hands-free loudspeaker, as the source and interference calibration signals, respectively. The calibration input sequence used to generate all the optimal beamformer weights consisted of these signals along with the car cabin noise signals, gathered at a specific time instant, t.

In order to evaluate the optimal beamformers, input signals were created by emitting independent speech signals from the artificial talker and the hands-free loudspeaker and recording the microphone observations with car cabin noise taken at time instant $t + 8$ s. The beamformer output was generated by filtering the inputs with the fixed filter weights found from the calibration sequences.

In the time-domain implementations, the FIR filter length was chosen as $L = 256$. For the frequency-domain implementations, the total number of subbands was set to $I = 64$. By setting the prototype filter length in the filterbank to 256, the same filter order as for the corresponding time-domain filters was obtained. This comes from the fact that the number of time-domain lags used in the frequency transformation equals the prototype filter length.

14.5.6 Evaluation Results

Performance measures in dB of the distortion, noise, and interference quantities, as described in section 14.5.3, are presented in Table 14.1. In general, the optimal SNIB beamformers have better suppression levels for both noise and interference when compared to the LS beamformers. However, the LS beamformers have much lower distortions values. Additionally, the subband-LS beamformer has performance comparable to the fullband-LS solution as the number of subband weights is increased.

Evaluation plots are now presented for the least-squares beamformer. Figure 14.5 illustrates the short-time (20 ms) power estimates in dB derived from an 8 s sequence of the single-reference microphone observation without any processing, followed by 8 s of the beamformer output signal acquired using the time-domain least-squares beamformer. Source speech, hands-free interference and car cabin noise are all active simultaneously. The near-end

Table 14.1. Distortion, noise, and interference performance measures of the beamformer output.

Performance [dB]	D	S_N	S_I
Time domain			
SNIB	-19.4	18.1	30.7
NLMS	-24.9	4.04	3.78
LS	-30.6	15.2	17.2
Frequency domain			
SNIB	-19.8	18.0	23.7
NLMS 1-tap	-21.1	8.68	5.00
NLMS 2-tap	-20.9	7.95	5.55
NLMS 3-tap	-20.8	7.45	4.96
NLMS 4-tap	-20.7	7.19	4.68
NLMS 5-tap	-20.7	7.11	4.54
NLMS 6-tap	-24.8	7.05	4.45
LS 1-tap	-28.6	12.9	13.6
LS 2-tap	-28.8	13.4	14.4
LS 3-tap	-30.0	13.8	15.2
LS 4-tap	-30.4	14.2	15.4
LS 5-tap	-30.5	14.3	15.7
LS 6-tap	-30.7	14.3	15.8

speech, coming from the location of interest is denoted in the plot as "Speech Male/Female" while the far-end speech echo, i.e. the interfering hands-free loudspeaker, is denoted by "Echo Male/Female".

Figures 14.6 and 14.7 show the spectral power estimates in dB of the reference microphone observation and the normalized spectral estimate of the least-squares beamformer outputs when the noise and the interference signals are active individually. These plots correspond to the numerator and the denominator of (14.40) and (14.41), respectively.

14.6 Summary and Conclusions

A number of optimal, time- and frequency-domain beamformers based on different error criteria were presented. The beamformers were evaluated in a real environment, a car hands-free telephony situation. Simulations with real speech signals acquired by a linear microphone array show that noise reduction of 18 dB and echo suppression of 30 dB can be achieved, simultaneously. This was accomplished by the time-domain version of the optimal signal-to-noise plus interference beamformer. With the time-domain least-squares implementation, noise suppression of 15 dB and hands-free suppression of 17 dB were found. The least-squares implementation yields ten times less distortion, as compared to the optimal signal-to-noise plus interference beamformer.

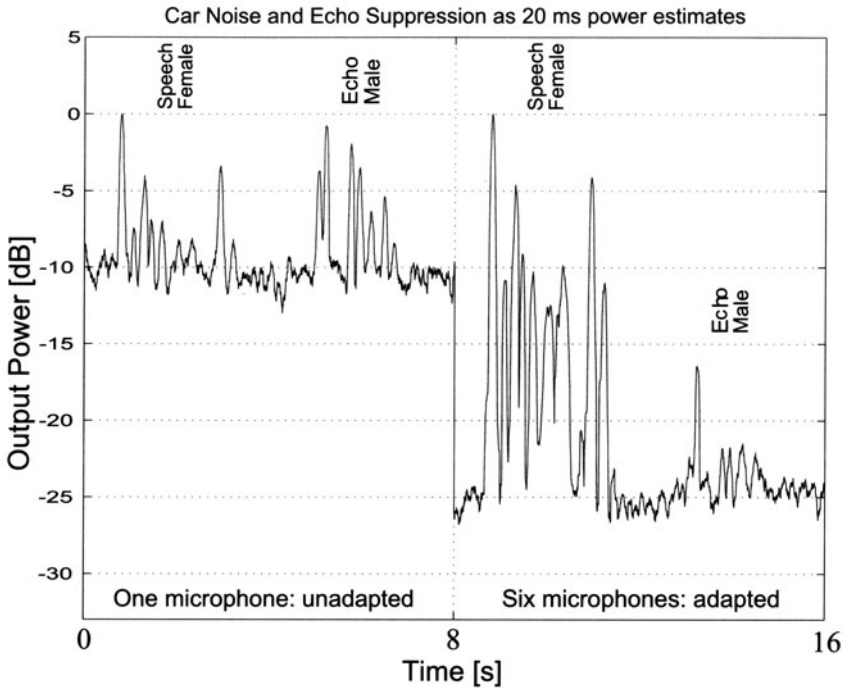


Fig. 14.5. Short-time (20 ms) power estimates of an unprocessed single microphone observation followed by the time-domain least-squares beamformer output signal.

The frequency-domain implementations show a similar relation between the optimal beamfomers. Better suppression is achieved with the optimal signal-to-noise plus interference beamformer, but the distortion is much higher than that for the least-squares implementations.

The subband least-squares beamformer evaluation showed that the performance on the real speech recordings is very close to that of the optimal time-domain least-squares beamformer. The noise and echo suppression were 14 dB and 16 dB, respectively, while the computational complexity was substantially reduced, thereby making it amenable to real-time processors. The distortion caused by the proposed method is the same as with the optimal time-domain least-squares beamformer.

Further research includes blind speech source extraction where the desired cross-correlation vector may be interchanged with a nonlinear function of the averaged beamformer output, for each frequency. The performance relies on the difference between the probability density functions of the source speech and the background noise. Implementations at an early stage show encouraging results. Source tracking is implicitly possible since a calibration sequence is unnecessary, and the objective function is made invariant to source movements.

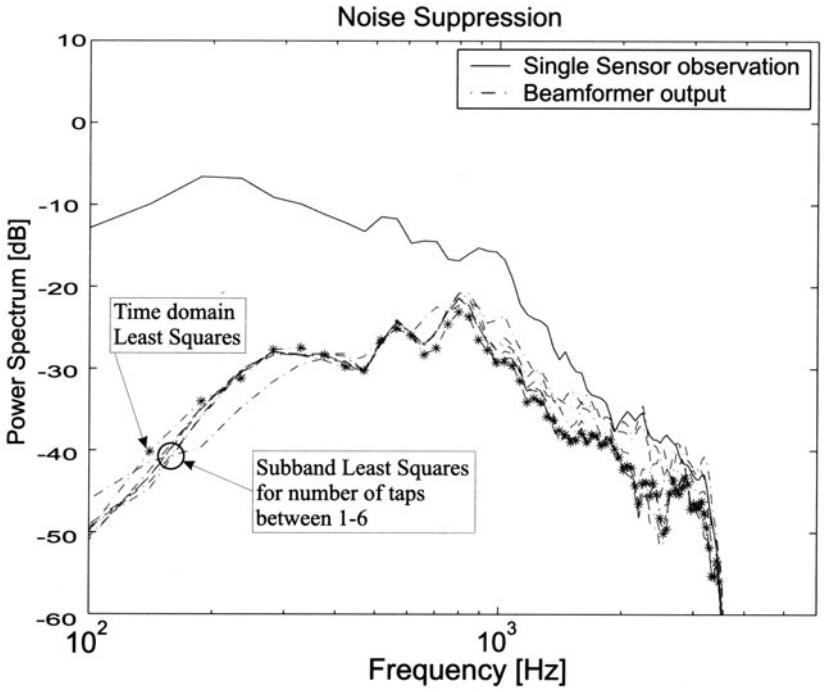


Fig. 14.6. Power spectrum of unprocessed single microphone observation (solid line) and power spectrum of the least-squares beamformer output signals (dashed-dotted lines) when only car cabin noise is present. The time-domain least-squares beamformer is marked by dashed-dots with stars.

References

1. S. Nordebo, B. Bengtsson, I. Claesson, S. Nordholm, A Roxström, M. Blomberg, and K. Elenius, "Noise reduction using an adaptive microphone array for speech recognition in a car," in *Proc. RVK93, Radio Vetenskaplig Konferens*, Lund, Sweden, Apr. 1993.
2. S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson, "Noise reduction using an adaptive microphone array in a car– a speech recognition evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz NY, USA Oct. 1993.
3. S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 425–437, Sept. 1997.
4. J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, Oct. 1993.
5. J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Macmillan, 1993.

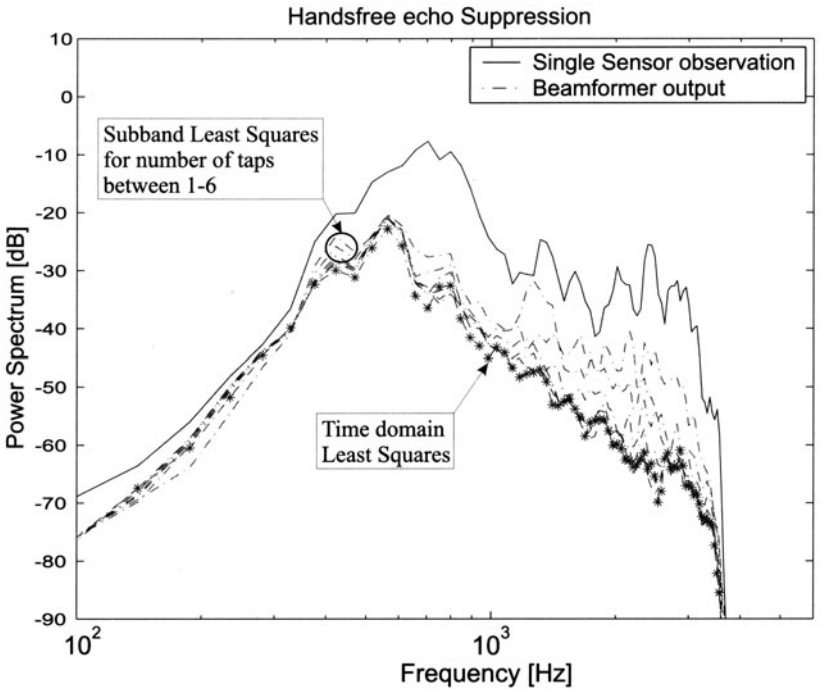


Fig. 14.7. Power spectrum of unprocessed single microphone observation (solid line) and power spectrum of the least-squares beamformer output signals (dashed-dotted lines) when only hands-free speech interference is present. The time-domain least-squares beamformer is marked by dashed-dots with stars.

6. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP–27, no. 2, pp. 113–120, April 1979.
7. J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, Minneapolis MN, USA, vol. II, pp. 363–366, April 1993.
8. H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction, truly linear convolution and a spectrum dependant adaptive averaging method," *Submitted for publication in IEEE Trans. Speech Audio Processing*, June 1999.
9. Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-34, no. 6, pp. 1391–1400, Dec. 1986.
10. S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Vehicular Tech.*, vol. 42, no. 4, pp. 514–518, Nov. 1993.
11. Y. Grenier and M. Xu, "An adaptive array for speech input in cars," in *Proc. Int. Symp. Automotive Technology and Automation (ISATA)*, 1990.
12. M. Sondhi and D.A. Berkley, "Silencing echoes in the telephone network," *Proc. IEEE*, vol. 68, pp. 948–963.

13. D.G. Messerschmidt, "Echo cancellation in speech and data Transmission," *IEEE J. Sel. Areas Commun.*, vol. SAC-2, pp. 283–297, Mar. 1982.
14. M. Sondhi and W. Kellermann, "Adaptive echo cancellation for speech signals," in *Advances in speech signal processing*, (S. Furui and M. Sondhi, eds.), ch. 11, Dekker, 1992.
15. C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt and J. Tilp, "Acoustic echo control, an application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, pp. 42–69, July 1999.
16. S. L. Gay and J. Benesty, eds., *Acoustic signal processing for telecommunication*, Kluwer, 2000.
17. S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals. an analytical evaluation," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 241–252, May 1999.
18. B. Widrow and S. D. Stearns, *Adaptive signal processing*, Prentice Hall, 1985.
19. I. E. Claesson, S. E. Nordholm, B. A. Bengtsson, and P. F. Eriksson, "A multi-DSP implementation of a broad-band adaptive beamformer for use in a hands-free mobile radio telephone," *IEEE Trans. Vehicular Tech.*, vol. 40, no. 1, pt. 2, pp. 194–202, Feb. 1991.
20. S. Nordebo, I. Claesson, and S. Nordholm, "An adaptive microphone array Employing calibration signals recorded on-site" in *Proc. ICSPAT94*, Dallas TX, USA, Oct. 1994.
21. M. H. Er, "A robust formulation for an optimum beamformer subject to amplitude and phase perturbations," *Signal Processing*, vol. 19, no. 1, pp. 17–26, Jan. 1990.
22. H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-35, pp. 1365–1376, Oct. 1987.
23. E. Walach, "On superresolution effects in maximum likelihood adaptive antenna arrays," *IEEE Trans. Antennas Propagat.*, vol. 32, pp. 259–263, March 1984.
24. I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. 40, no. 9, pp. 1093–1096, Sept. 1992.
25. S. Nordholm, I. Claesson, and S. Nordebo, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE J. Oceanic Eng.*, vol. 19, no. 4, pp. 583–590, Oct. 1994.
26. L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. 30, pp. 27–34, Jan. 1982.
27. M. H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 31, no. 6, pp. 1378–1393, Dec. 1983.
28. M. H. Er and A. Cantoni, "A new set of linear constraints for broad-band time domain element space Processors," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 2, pp. 320–329, Mar. 1986.
29. O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

30. M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with a microphone array," *IEEE Trans. Vehicular Tech.*, vol. 48, no. 5, pp. 1518–1526, Sep. 1999.
31. M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, vol. 1, pp. 239–242, 1997.
32. N. Grbić, M. Dahl, and I. Claesson, "Neural network based adaptive microphone array system for speech enhancement," *IEEE World Congress on Computational Intelligence*, Anchorage AK, USA, vol. 3, pp. 2180–2183, May 1998.
33. N. Grbić, *Speech signal extraction - a multichannel approach*, University of Karlskrona/Ronneby, Sweden, Nov. 1999.
34. S. Haykin, *Adaptive filter theory*, Prentice Hall, 3rd edition, 1996.
35. P.P. Vaidyanathan, *Multirate systems and filter-banks*, Prentice Hall, 1993.
36. S.K. Mitra, *Digital signal processing*, McGraw-Hill, 1998.

15 Speech Recognition with Microphone Arrays

Maurizio Omologo, Marco Matassoni, and Piergiorgio Svaizer

ITC-IRST (Istituto per la Ricerca Scientifica e Tecnologica), Povo (Trento), Italy

Abstract. Microphone arrays can be advantageously employed in Automatic Speech Recognition (ASR) systems to allow distant-talking interaction. Their beamforming capabilities are used to enhance the speech message, while attenuating the undesired contribution of environmental noise and reverberation. In the first part of this chapter the state of the art of ASR systems is briefly reviewed, with a particular concern about robustness in distant-talker applications. The objective is the reduction of the mismatch between the real noisy data and the acoustic models used by the recognizer. Beamforming, speech enhancement, feature compensation, and model adaptation are the techniques adopted to this end. The second part of the chapter is dedicated to the description of a microphone-array based speech recognition system developed at ITC-IRST. It includes a linear array beamformer, an acoustic front-end for speech activity detection and feature extraction, a recognition engine based on Hidden Markov Models and the modules for training and adaptation of the acoustic models. Finally the performance of this system on a typical recognition task is reported.

15.1 Introduction

During the last decade research on ASR technology has made significant advances. As a result, high performance systems are now available for situations where there is a good match between testing and training conditions [1,2]. However, these same systems tend to suffer from a limited robustness to variability in their operating environment [3–8].

One of the most attractive potential features of ASR technology is the flexibility afforded through hands-free interaction. Not being encumbered by a hand-held or head-mounted microphone may be of considerable utility to the user. Of particular concern is the distant-talker case where the user is beyond the normal acquisition range of the system microphone[1] (e.g. at more than one meter in the case of an omnidirectional microphone). For ASR applications of moderate/high language complexity this represents a very ambitious task.

The development of distant-talker ASR will allow for the expansion of voice activated technology into a number of areas where it has until now

[1] The distinctions among different types of microphones will not be treated here. However, the choices available and their relative characteristics should be specifically addressed in the design of the application.

been ineffective (e.g. noisy offices or factory floors) and will improve its functionality for applications where it has already seen some use (e.g. computer dictation or the home). In the first example, voice messages of a varying nature would have to be recognized as isolated word commands immersed in a background of multi-talker speech and noise. The second example involves a rather clean environment, but a large vocabulary of words to identify.

In the real-world applications it would also be necessary to account for various factors related to the means of interaction. The talker's position may be unknown and time-varying in an unpredictable fashion. Because of sound attenuation and talker radiation effects, the quality of the input signal may be influenced by even subtle head movements. Moreover, environmental noise and room acoustics play an important role, especially in the case of highly reverberant conditions and unstationary noise sources. In the most adverse noisy conditions, a talker will tend to speak more loudly than usual and thereby modify the underlying characteristics of the speech signal produced relative to normal speaking conditions. This is known as the Lombard effect [9]. Additionally, when the language/dialogue model becomes more complex, the variability in talking style may increase and one can expect that the talker will often speak in spontaneous mode.

For these reasons and others, there are many challenging and as yet unsolved problems in this field. In the last few years, some work has been devoted to the application of multi-microphone based processing for distant-talker speech recognition. Compared to the number of labs working on improving the robustness of single-channel ASR systems, this effort is relatively small. This fact may be due to the incipient nature of microphone array technology and the increase in hardware complexity that is required for a multi-channel front-end. However, judging by the advancements in ASR performance that may be attributed to improvements in input signal quality brought about by microphone array processing, this work is well justified.

The remainder of this chapter is organized into two sections. The following section summarizes the current state of research activity in the field of ASR, particularly with regard to the distant-talker situation. The final section details a specific microphone-array based recognition system, namely an Italian language recognizer developed at ITC-IRST.

15.2 State of the Art

15.2.1 Automatic Speech Recognition

Automatic speech recognition can be viewed as a problem of conversion from speech into text by a decoding process that involves several processing stages. The characteristics and the difficulty of an ASR application differ substantially based upon various features. These include vocabulary size and confusability, speaker independence, language complexity, and input speech quality.

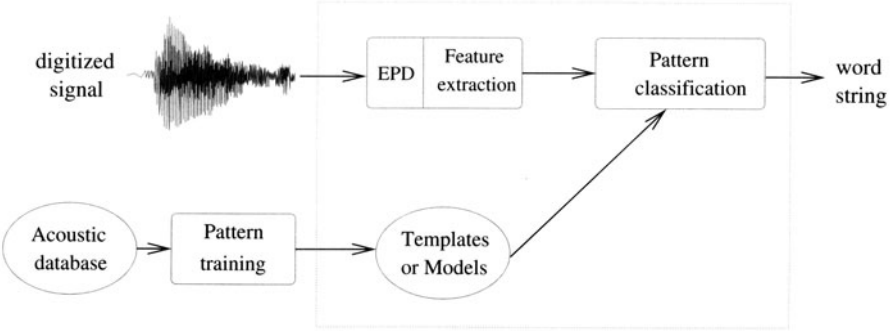


Fig. 15.1. General block diagram of a pattern recognition based ASR system.

Research on ASR has been conducted for more than four decades. The related literature is very large; good overviews of the most significant achievements can be found in [1,2]. Many different techniques for ASR have been investigated. Currently, the most widely used approaches are based on some form of statistical pattern recognition. Thanks to these modern methods, the growth in hardware capability, and the availability of very large speech corpora for training, the last decade has witnessed high level performance achieved on recognition tasks of progressively increasing complexity.

The purpose of this section is to give a very brief introduction to the basic problem and to the most common solutions which have been adopted, with specific reference to the distant talker problem.

Pattern recognition based ASR A general block diagram of an ASR system based on the pattern recognition approach is shown in Figure 15.1. It is assumed that the speech message has been transduced by a microphone into an electrical signal and then converted into an equivalent digital representation with an adequate sampling rate and quantization level (e.g. 16 kHz and 16 bit, respectively). In general, at a preprocessing level ASR systems include a pre-emphasis step in the form of a single tap high-pass filter. The goal of which is to emphasize high-frequency formants which typically have a reduced magnitude due to a negative spectral tilt in the speech signal, particularly voiced sounds.

A speech activity detection process, also called End-Point Detection (EPD), is employed to isolate speech events from other segments and background noise. Several techniques are available for EPD, e.g. [10–14]. These are usually based on criteria such as short-term energy and zero-crossing rate. However, they may also rely on the same acoustic features used during the recognition process.

The objective of Feature Extraction (FE) is to convert the input signal into some form of compressed parametric representation. The most common examples of FE are based on short-time spectral analysis. Speech can be

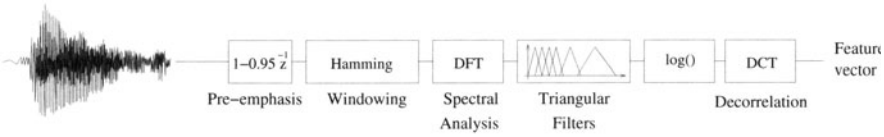


Fig. 15.2. Block diagram of the computation of Mel-based cepstral coefficients.

considered statistically stationary over short periods of time (a few tens of milliseconds). As a consequence, the analysis frame size employed for FE is generally over 15 ms with a step size in the range of 5-30 ms. The FE process produces a sequence of vectors of dimensionality normally ranging between 8 and 12. These vectors are frequently augmented by including data to characterize the first and second order time derivatives of the given features.

A number feature sets have been investigated for ASR. Popular ones include Mel-scaled Cepstral Coefficients (MCC's) (see Figure 15.2), LPC coefficients, PLP coefficients [5,15]. Currently, MCC's are the most widely used acoustic features for ASR. Figure 15.2 outlines their method of production. Basically, a triangular filter-bank is applied to the output of a short-term spectral analysis. The logarithmic-like Mel scale models the frequency resolution of the human ear and for this reason is preferred to a linear scale in the filter-bank. A Discrete Cosine Transform (DCT) is then applied to decorrelate the log-filter-bank output. The resulting MCC's may then be statistically modeled through Gaussians with diagonal covariance matrices. This property is useful in the case of HMM-based recognition discussed below.

Note that non-linearities and approximations are included in the processing that derives the acoustic features from the signal or its power spectrum representation. For instance, the output of the filter-bank used for MCC computation provides a rough approximation to the FFT-based spectral analysis method. As a consequence, when addressing the impact of microphone arrays to distant-talker ASR, it is possible that improvements in signal quality produced by the array processing may be rendered ineffective during successive stages of the recognition chain because of these approximations.

In the pattern recognition approach to ASR, the acoustic feature vector sequence derived from the unknown speech is compared to the feature sequence of reference speech. Among the various ways to perform this comparison, three methods have been primarily utilized: Hidden Markov Model (HMM), Dynamic Time Warping (DTW), and Artificial Neural Networks (ANN). A detailed discussion of these techniques, the functional relationships between them, and the hybrid solutions which have been studied (e.g. ANN/HMM), goes beyond the scope of this chapter and can be found in [1,2].

The literature reports that for simple tasks (e.g. connected digit recognition) in controlled and matched environments (i.e. the user interacts with a close-talking microphone and the system has been trained on clean speech), satisfactory recognition performance can be obtained using any of the above

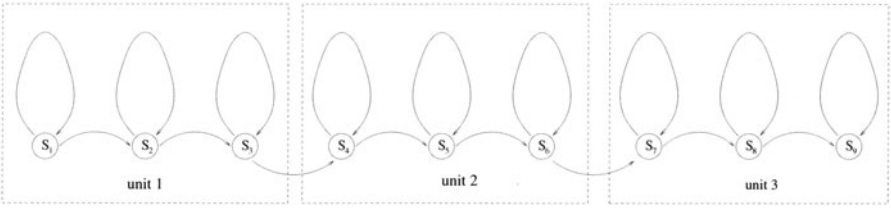


Fig. 15.3. Concatenation of three HMM-based phone models, characterized by a left-to-right three-state topology.

methods. However, HMM's are currently considered to be the most effective and stable framework for speech modeling in a general context. This is particularly the case for large vocabulary tasks and when statistical language modeling as well as integration of the recognizer with a dialogue manager are required by the application. As a consequence, the remainder of this chapter will assume the use of HMM's as the means for pattern classification. Accordingly, it will be necessary to present a modicum of detail regarding the procedure.

HMM framework In the statistical ASR paradigm, a generic utterance consists of a sequence of unknown words and the recognizer finds the most probable word string, given the observed feature vector sequence provided by the front-end processing. HMM's serve as the statistical model used to classify the utterances and quantify their observation probabilities.

Bayes' rule is used to decompose the required probability into two components: the *a priori* probability of observing the sequence of words (the "language model") and the probability of observing the feature vector sequence given that word string (the "acoustic model").

Each word is represented by a chain of basic sounds called phones. An HMM is adopted for each phone. In practice, the model consists of a number of states with the sequencing through them determined by a set of transition probabilities. Each state produces observations which are characterized by a set of state-dependent observation statistics. These are frequently modeled as mixtures of Gaussian densities. Figure 15.3 shows an example of a three-state, Markov model. In this case the non-zero transition probabilities are constrained to produce a left-to-right topology which is very common in ASR applications.

The *Training Problem* involves learning the appropriate HMM parameters given a reference ensemble of feature sequences associated with the desired word. An efficient procedure known as the Baum-Welch algorithm is available for this purpose. For what concerns the *Scoring Problem* during recognition, once the most likely state sequence is selected, the related recognized text is provided. The Viterbi algorithm is commonly used to efficiently evaluate word string likelihoods.

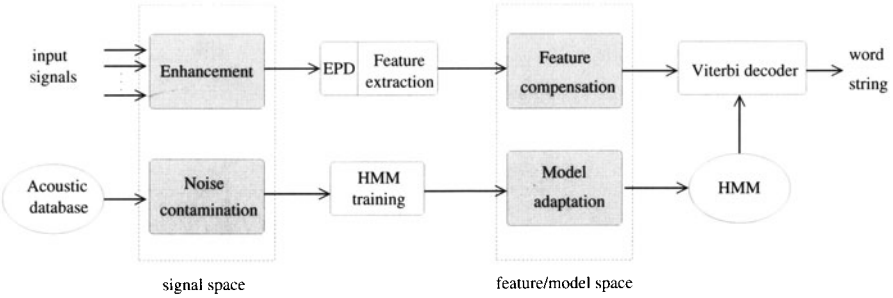


Fig. 15.4. Block diagram of a robust ASR system operating in an adverse environment. Highlighted blocks, operating either in the signal space or in the feature/model space, contribute to reduce the mismatch between noisy data and the acoustic models used by the recognizer.

For a more thorough development of HMM theory and practice, see [16]. As to the main concerns of this chapter, note that the theory of HMM and its application to ASR are based on a number of signal assumptions, among which is that of statistical independence of the observations over time. This is not satisfied in many cases, particularly for speech acquired in the presence of noise and reverberations.

15.2.2 Robustness in ASR

The crucial aspect for most ASR techniques is robustness. In practice, performance very often degrades rapidly when these systems are used with speech input taken from a noisy environment or, in general, with speech input having characteristics which differ from those observed during the training phase. Mismatch between training and test conditions can be caused by many factors. Some examples are background noise, transducers, channel noise, inter-speaker variability, and spontaneous-speech phenomena. Hence, flexibility and robustness with respect to these sources of variability is one of the main objectives of current ASR research [3–8]. Generally, the training of speech recognizers is accomplished by using large speech corpora. In principle, for each noisy environmental condition and, in this case, for each microphone and talker position, a specific corpus should be used. This solution being impractical, a fundamental task for researchers and technology developers becomes that of exploiting as much as possible from existing corpora, tools, techniques, and *a priori* knowledge, in order to build robust recognizers.

Current approaches to improving robustness of noisy speech recognizers can be classified into a number of categories as reported in [4,7,17]. Four of these general approaches (signal enhancement, feature compensation, model adaptation, and noise contamination) are summarized below. Figure 15.4 shows how these methods may be introduced as modules into a generic HMM-based architecture to improve system robustness.

Enhancement techniques are used to increase the quality of the signal provided to the recognizer [18–21]. Their impact on ASR is not obvious. There is no direct relationship between the SNR (or perceptual quality) of the resulting signal and speech recognition performance, even when the recognizer is trained using the same preprocessing. In this regard, a further critical issue is that of end-point detection. Several algorithms (generally based on energy-thresholding techniques) have been proposed, which can be applied successfully with a SNR as low as 10 dB. However, most of these algorithms generally become unusable with lower SNR conditions [5].

Many techniques have been proposed which address the parametric representation of the signal. These aim at constructing a compact and robust feature set and processing it to compensate for the mismatch between acoustic spaces of the clean and noisy speech. As an example, a very simple feature normalization technique, often combined with MCC's, is Cepstral Mean Subtraction (CMS). It consists of removing from each cepstral coefficient sequence the mean evaluated across the whole utterance (or over an extended interval). CMS aids in reducing the influence of slow variations in the acoustic feature vectors, like those related to convolutional channel effects (e.g. change of microphone) and to speaker-dependent biases. Other relevant compensation techniques, operating either in the feature space or in the model space, are signal bias removal [22,23], stochastic matching [24,17], noise modeling and masking [25–27], Parallel Model Combination [28,29].

In the recent years increasing attention has been devoted to acoustic model adaptation. When the application requirements allow, these techniques attempt to adapt system parameters to the speaker and environment by exploiting data samples representative of the actual acoustic conditions. For systems based on continuous-density HMM's, most popular adaptation schemes rely on maximum *a posteriori* (MAP) estimation [30] or maximum likelihood linear regression (MLLR) [31–33] of the model parameters.

Finally, it is possible to adopt an approach known in the literature as *training data contamination* [4,34,35], which provides a way of training acoustic models which are more robust and representative of the given real noisy environment than those derived through training on the corresponding clean speech. In practice, training data are produced by injecting real or artificial noise into the clean speech material. Clearly, this approach is time-consuming and may become impractical when the size of training set grows large. However, it does offer some advantage. For instance, it is free from the negative spectrum problem typical of noise subtraction schemes [36].

15.2.3 Microphone Arrays and Related Processing for ASR

The utility of a microphone array as input to a speech recognition system lies in its ability to acquire a higher quality signal than that provided by a single far-field microphone. The signal enhancement is obtained by emphasizing the talker's speech as well as by reducing noise and reverberation components.

These methods are the specific topics of other chapters in this work. The relevant approaches will only be summarized very briefly here.

In order to reduce mismatch effects in the recognizer, a first requirement is that of having uniform improvement levels across the complete speech spectrum. Additionally, it would be desirable to have these spectral enhancements be independent of the talker's position. Unfortunately, microphone array frequency responses are characterized by significant variations across source angles and distances. Consequently, it is necessary to reduce as much as possible this variability, which may introduce significant discrepancies with respect to the training conditions.

In practice, an enhanced output can be derived from a microphone array by the application of beamforming techniques. The simplest and most commonly used approach is the delay-and-sum beamformer, which reduces the output power for directions other than that of the steering location by means of destructive interference. Figure 15.5 illustrates a typical result from this procedure. The delay-and-sum beamformer may be used passively to realign the signals given a set of delay estimates, or actively by aiming the array towards a specific direction. In the former case, delay estimates are derived from Time Delay Estimation (TDE) techniques [37], as shown in the case of talker location in [38], where a Cross-Power Spectrum Phase analysis was adopted.

Applying Time Delay Compensation (TDC) processing represents a good solution for the case of an isotropic (diffuse) noise field, as no spatial coherence is exploitable to suppress undesired components. Very much dependent on the number of array elements and their geometry relative to the source position, delay-and-sum beamforming provides only moderate directivity gains. Additional drawbacks [39–43,8], are grating lobes in the directivity pattern and a low-pass effect due to both the beam narrowness at high frequencies and to imperfect steering caused by imprecise inter-channel delay estimates.

In the case of coherent noise sources linearly constrained adaptive beamformers, such as those proposed by Frost [44] or the Generalized Sidelobe Canceler [45], have the specific objective of eliminating noise contributions in directions outside the directivity lobe. The main limitation of these schemes in a reverberant environment is the issue of signal cancellation. Since the degradations are correlated with the desired signal, the suppression process introduces distortions to the desired signal. Superdirective beamformers have also been proposed [46] to suppress interfering signals effectively.

A technique specifically developed to address the reverberation phenomenon of enclosures is the Matched Filter Array . This technique utilizes the acoustic impulse responses of the environment to create constructive interference between direct and reflected components of the speech signal [47,48].

Beamforming techniques may also be combined with adaptive postfilters [46] (e.g. based on Wiener theory) for further noise reduction. However,

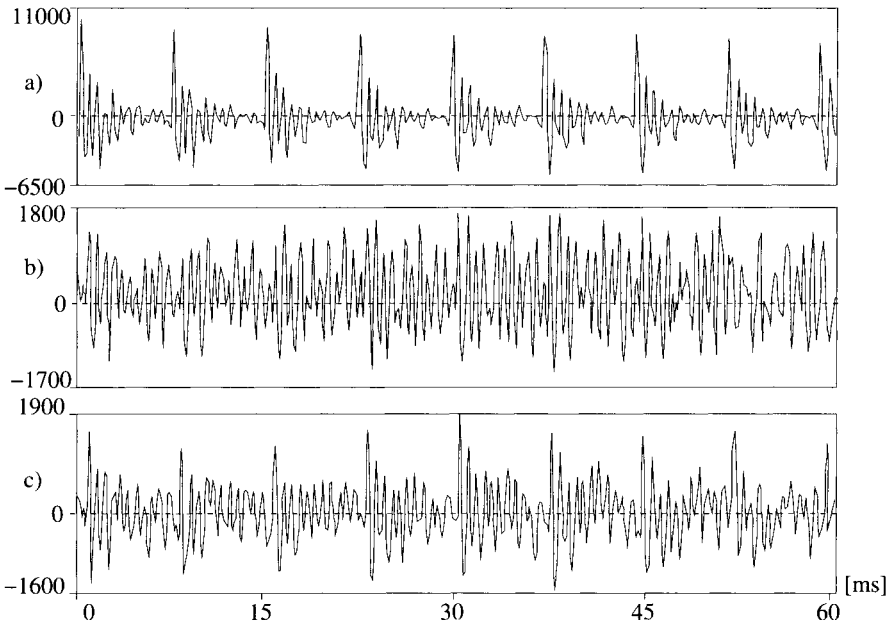


Fig. 15.5. Portion of a vowel /a/ uttered at 4 m distance from a linear array of six microphones. The upper waveform is the close-talker signal, the middle plot represents the reverberated signal acquired by one of the microphones, and the lower plot shows the corresponding delay-and-sum beamformed signal.

the use of post-filtering, like its above counterparts, can introduce artifacts into the reconstructed signal, particularly for the case of reverberant environments [49], and may consequently limit recognition improvements.

15.2.4 Distant-Talker Speech Recognition

A sizeable body of work on distant-talker ASR have been produced in recent years [50–56]. The use of either single microphones or multi-microphone systems has focused primarily on experimental contexts (e.g. car environment) for tasks generally characterized by a small-size vocabulary and by a low complexity language. Simple multi-microphone products are available commercially and have replaced traditional input devices in some ASR applications. However, these devices are of limited practical utility and are typically only effective for talkers no more than one meter away from the array, at a fixed and a quite narrow range of angles, and in a rather quiet environment.

This section provides a brief overview of the literature relating to various research topics related to distant-talker ASR. The intention is not to give a thorough description of all the techniques which have been investigated, but rather to indicate the general issues and approaches. Section 15.3 will explore in detail the system developed at IRST labs.

Array Geometry This first topic concerns the array characteristics as well as their influence on recognition performance. Clearly, the number of microphones represents an important factor. In an effort to limit hardware complexity, most investigators utilize 16 or fewer microphones. Optimal array design techniques have been addressed by other chapters of this book and, with specific reference to ASR, in [57]. For practical ASR systems, the bulk of array geometries investigated have either been of the linear equi-spaced or harmonic nesting varieties. The latter is the most common in practice, despite the fact that a real advantage in application to hands-free ASR is not evident [58,59]. In general, demonstrating a potential advantage inherent in a specific geometry is difficult. The speech quality improvement due to the array configuration is counterbalanced by several approximations (e.g. in the generation of artificial signals, when simulation is used, and in the acoustic feature extraction), and by effects related to the application of compensation/adaptation techniques applied to these features or to the acoustic modeling.

Beamforming The literature reports on the use of various beamforming techniques for distant-talker ASR. The delay-and-sum beamformer is the most commonly used, despite its limited speech enhancement capabilities. The joint use of delay-and-sum beamforming and a talker localization module is investigated in [60]. The use of adaptive beamforming techniques (e.g. GSC), generally under the assumptions of fixed talker and noise source positions, is also common. Some examples are shown in [61,51,50,62]. As expected, in the presence of coherent noise sources, adaptive beamformers yield more robust recognition performance than delay-and-sum beamformers. However, many authors report that the improvements are lower than what would be expected on the basis of the SNR or of the reconstructed signal's perceived quality. This observation seems to be more common for data acquired in real environments.

End-Point Detection This topic is very critical, even in moderately noisy and reverberant environments. Most of the experiments reported in the literature are based on the use of a "push-to-talk" speech acquisition method. In the past, a few works addressed the impact of EPD on distant-talker recognition performance. In [50] adaptive energy thresholds were applied to the output of the delay-and-sum beamformer to identify speech boundaries. In [38], a CSP-based coherence measure between two input channels was used to detect a generic acoustic event. Its effectiveness in speech activity detection will be confirmed in the next section. The application of a similar coherence measure to speech/noise classification is also documented in [63].

Acoustic Features In ASR research and applications, there is the tendency today to adopt a standard acoustic feature set (e.g. Mel or LPC cepstral co-

efficients). This is also the case for distant-talker ASR research. Most of the systems described in the literature are based on the use of MCC's or LPC cepstral coefficients together with their first/second order derivatives and cepstral mean subtraction. However, some other acoustic features more robust to noise, such as PLP or short-time modified coherence (SMC) [64,46], have been investigated. In the interest of acoustic feature robustness to reverberation effects, possible future approaches may be inspired by techniques which selectively process the linear prediction residual [65] or incorporate speech production modeling into the given multi-channel framework [66,67].

Recognition Engine The majority of recognizers investigated in the literature are based on the use of traditional HMM-based solutions. Training HMM's on artificially contaminated speech may lead to robust solutions when a large noisy database is not available, as shown in [68] and in [69]. In some cases, MAP, MLLR or ANN-based adaptation techniques have been adopted [70,71,54,53] to further reduce the mismatch between training and test conditions which remains after the microphone array based processing. In [56], MLLR is also compared favorably to Parallel Model Combination. Broadly speaking, these approaches allow the system to learn more about the speaker characteristics, the environmental noise, and the "artifacts" (e.g. low-pass effects typical of delay-and-sum beamforming) introduced by the multi-channel processing. Overall, the use of adaptation techniques has had a significant positive impact on hands-free ASR system performance.

An alternative approach that deserves to be mentioned is proposed in [61]. Here a new dimension is added to the search space used by the Viterbi algorithm to account for the different directions from which the talker may be speaking. In this way, changes in source position should be implicitly detected on the basis of a maximum likelihood criteria. The resulting system is more flexible, but possesses a considerable complexity increase and requires a consistent HMM training to be performed initially.

Another alternative approach is reported in [72], where an ANN is used to perform a transformation/normalization of the acoustic features extracted from the delay-and-sum beamformed signal. During training, the ANN learns information related to the talker position. The influence of the talker location is addressed in [71], where the effectiveness of a location independent ANN is demonstrated.

Speech Corpora and Tasks A variety of recognition tasks have been investigated in the literature. The most common is connected-digits. However, the choice of the experimental task is probably not as relevant as the way this task is created. In order to derive speech material for training and especially for test experiments, three main approaches have been adopted:

- Speech data is collected from a sampling of *real* talkers using multi-channel recording hardware. This method requires much more effort than

its alternatives, but it represents the most reliable means for investigation in this field and for comparing results to theory.

- Speech data (e.g. extracted from a given clean database) is played back through a loudspeaker. Again, multi-channel hardware is used to record the signals. With this method there is the advantage of repeatability of the same utterances under different conditions, array geometries, etc. However, the experiments may be influenced by artifacts inherent in the recording process, such as the dependency on loudspeaker response, the different radiation modeling, and other variabilities in the environmental conditions. It is worth mentioning the work done in Bell Laboratory's VarEchoic Chamber [73], by which any reverberant condition can be investigated without the risk of changing other environmental characteristics across recording sessions.
- Speech data is reproduced by simulation, typically based on a simplified additive/convolutive channel modeling. In this case, the reverberation effects on the various input channels are generally recreated by convolving the close-talker signal with real impulse responses measured using a time-stretched pulse [74,75] or with artificial impulse responses derived by applying the Image Method [76]. A simulation-based experiment has a clear limitation due to the fact that many phenomena occurring in a real environment may be neglected. Moreover, the use of simulation both in training and in test may provide misleading results due to biases in the artificial data generation.

15.3 A Microphone Array-Based ASR System

This section describes the distant-talking recognition system being developed and experimented with at IRST labs [70,60,59,77,68,78,79]. Figure 15.6 shows a block diagram of the system, consisting of: a microphone array and the related TDC processing, an acoustic front-end for speech activity detection and acoustic feature extraction, a recognition engine module (Viterbi decoding) and related modules for HMM adaptation. Each module of the system as well as the experimental framework will be described below together with the most relevant recognition results so far obtained.

15.3.1 System Description

Speech Acquisition Distant-talker speech signals were acquired by a linear array of six equi-spaced (at 15 cm) omnidirectional microphones. Each channel was synchronously sampled at a 16 kHz rate with 16-bit accuracy. Delays estimated between the channels (through CSP-based time delay estimation) were used to align the signals in a delay and sum beamformer.

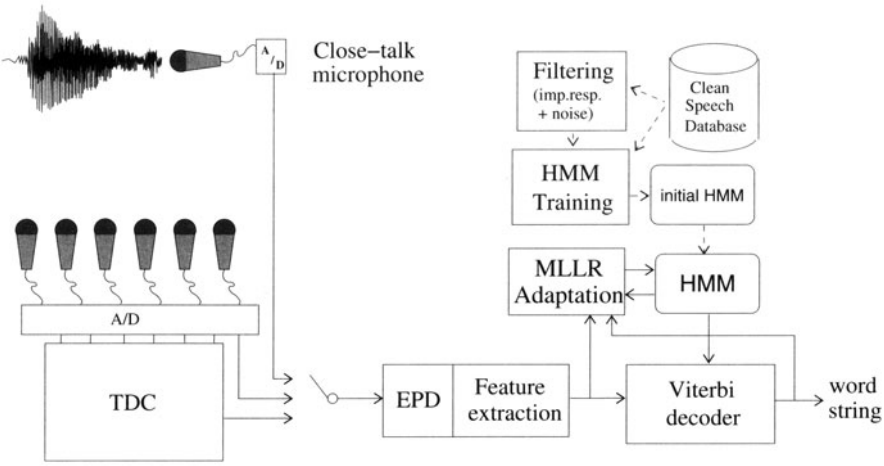


Fig. 15.6. Block diagram of the distant-talking ASR system developed at IRST. It includes signal acquisition, Time Delay Compensation processing, the acoustic front-end, a recognition engine, and the modules for training and adaptation of the acoustic models. Close-talker and single far-field microphones are used for reference purposes in addition to the microphone array.

Speech-Activity Detection The use of a microphone array adds a spatial dimension to the domain of the time/frequency analysis of conventional single input systems. Besides source localization and selective acquisition by beamforming, an additional benefit of multi-microphone systems is the capability of discerning between coherent directive sources (e.g. a talker facing the microphones) and spatially diffuse, low coherence disturbances. The discriminating feature is a coherence measure between the signals of different microphones, such as the phase correlation [37,38]. Coherence measure computation is here extended to several microphone pairs to provide a more robust speech activity function. This function is effective for low SNR and reverberant signals, where an energy-based approach would not be.

Figure 15.7 illustrates an example of this procedure. The upper plot depicts the noisy speech signal acquired by a single microphone in the array. The middle plot represents the corresponding phase correlation between two channels of the array as a function of time (horizontal axis) and mutual delay in samples between the channels (vertical axis). A darker gray level denotes higher coherence. The lower plot shows the coherence measure at the true delay versus time.

In practice, the EPD technique proposed here is based on a preliminary selection of an inter-channel delay for each microphone pair. Given the inter-channel delays associated with the various microphone pairs, the appropriate coherence functions are summed to derive a speech activity function. Adaptive thresholds are then applied in order to determine speech boundaries.

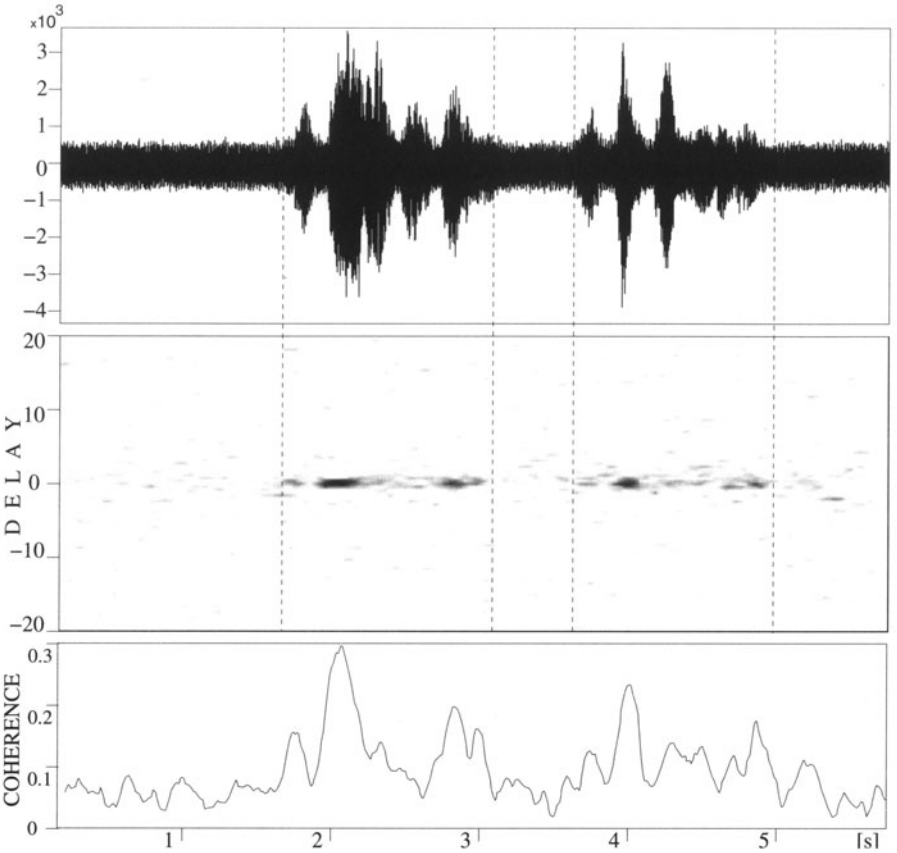


Fig. 15.7. Example of coherence computation for the signals of a microphone pair. It includes one of the noisy speech waveforms acquired by the microphones, a grey level coherence measure representation at various inter-channel delays, and the coherence levels at the correct delay (0 samples).

Acoustic Feature Extraction In the experimental set-up described here, the input to the feature extractor (see Figure 15.6) is either the output of the TDC processing derived from the microphone array data, the signal acquired by a single microphone within the array, or that acquired by the close-talker microphone.

In any case, the input signal is pre-emphasized and blocked into 20 ms, half-overlapping frames. For each frame, 8 MCC's and the log-energy are extracted. CMS is then applied to each feature sequence. The resulting normalized MCC's and log-energy, together with their first and second order time derivatives, are arranged into a single observation vector of 27 components.

HMM Recognizer Acoustic modeling is based on a set of 34 phone units. Each unit is modeled with a three state left-to-right continuous-density HMM, with output probability distributions represented by the means of mixtures having 16 Gaussian components with diagonal covariance matrices. Phone HMM's are trained either using a clean speech database or a noisy version, obtained by simulation as described below. Once the HMM's have been trained, the resulting models are adapted to the real environment by applying a MLLR adaptation technique [31–33].

15.3.2 Speech Corpora and Task

Various speech corpora have either been collected or produced in order to perform the experiments to be discussed.

Clean Speech Corpus The initial step of HMM training is accomplished through the standard Baum-Welch procedure. For this purpose, phonetically rich sentences representing a portion of APASCI [80] were used. This corpus was acquired in a quiet room (SNR $\geq$ 40 dB) using a high quality close-talker microphone. The training set consisted of 2100 utterances collected from a total of 100 speakers (50 males and 50 females).

Multi-Channel Real Noisy Corpus The multi-channel noisy corpus consists of speech material collected in an office of size (5.5 m $\times$ 3.6 m $\times$ 3.5 m) characterized by a moderate amount of reverberation ($T_{60} \simeq 0.3$ s) as well as by the presence of coherent noise due to secondary sources (e.g. computers, air conditioning, etc). Multi-channel recording of each utterance was accomplished by using both a close-talker (CT) directional microphone and the linear array described above.

Speech material was collected from 8 speakers (4 males and 4 females) during a series of recording sessions with variable environmental noise conditions. Each speaker uttered 50 connected digit strings (400 digit occurrences), both at frontal position F150 (1.5 m distance from the array) and at lateral position L250 (2.5 m distance, left of the array). Four of the individuals also uttered the same string set at position L150 (1.5 m distance, 60° right of the array).

Utterances were recorded with background noise segments of varying length at the beginning and end of each digit sequence. SNR, expressed as the ratio of the average speech to noise energy measured at the array microphones, was 12.6 dB mean with 3 dB standard deviation for the frontal recordings, and 10.7 dB mean and 2.8 dB standard deviation for the lateral recordings. As reference, SNR estimated on the CT microphone signals possessed a 28 dB mean and 3.8 dB standard deviation.

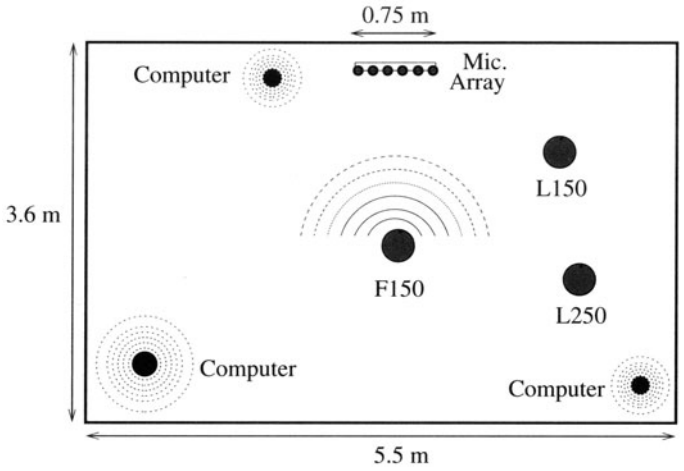


Fig. 15.8. Map of the experimental room showing the positions of the talker, the microphone array, and the computers (noise sources). The label at each position indicates, in a compact form, the orientation (F for frontal, L for lateral) and the distance in cm from the array.

Contaminated Speech Corpora A set of training databases consisting of acoustically realistic signals was artificially recreated using the APASCI clean corpus along with knowledge (e.g. room impulse responses and background noise signals) of the real operating environment. For this purpose, a simplified additive/convolutive model was adopted as follows:

$$s_{co}(t) = h_r(t) \star s_{cl}(t) + k \cdot n(t) \tag{15.1}$$

where $h_r(t)$ is an impulse response of the room, k is a scaling factor, $n(t)$ is background noise acquired in the room, s_{cl} is the clean speech, s_{co} is the contaminated speech, and $\star$ denotes convolution. The effect of background noise is accounted for by scaling the noise recorded inside the room using an appropriate amplitude to reproduce different SNR's (ranging from 0 to 21 dB) and then adding the result to reverberant speech. The reverberation effects of a room can be simulated in several ways. In this case, it was achieved by convolving the close-talker signal with impulse responses measured using a time-stretched pulse.

15.3.3 Experiments and Results

Experimental results involving connected-digit recognition are reported below. These are expressed in terms of Word Recognition Rate (WRR), computed as the average performance obtained by testing on material obtained from all the speakers and positions. As a result, each test experiment consists of the recognition of 8000 digits.

Table 15.1. Word recognition rates obtained on a connected-digit recognition task using different phone models, input devices, and end-point detection methods.

Models | Input	FarMic	Arr_{CSP_EPD}	Arr_{ID_EPD}
Clean	33.7	57.1	61.2
Rob	91.5	95.2	96.3
AdaRob	95.6	98.3	98.6

Experiments were conducted using either the microphone array (Arr) or a single microphone of the array (FarMic). For comparison purposes, results obtained testing on material acquired with the close-talker microphone (and using clean models) was approximately 99%. This reference result represents a sort of upper bound of any experiment conducted.

As shown in Table 15.1, training with filtered clean speech (Rob) improves recognition performance tangibly, even in the case of a single far microphone input. This result is consistent with other work [68,78,79]. The results confirm that the use of the microphone array, in combination with the TDC module ensures superior recognition performance relative to a single microphone. However, the advantage of using the array is more relevant in the case of robust models. In this case, more than 40% relative improvement was obtained (from 91.5% to 95.2%).

A second issue investigated was the impact of the speech activity detection method on the recognition performance. In addition to difficulties due to the distance between the talker and microphones, the system is prone to insertions in this experimental framework. This is due to the adoption of a digit-loop grammar with no information about string length. As a consequence, pauses inside a digit sequence and long noise segments, preceding and following the speech utterance, can cause many errors because of mismatched acoustic modeling. The right hand column in Table 15.1 (Arr_{ID_EPD}) shows the results obtained using an "ideal" end point detector. These were acquired using utterance boundaries identified manually and leads to a relative performance improvement of about 20% compared to results obtained using the coherence-based EPD method described earlier (Arr_{CSP_EPD}). Resolving this performance disparity is a goal for the future development of a more accurate EPD algorithm.

The results show the further improvement provided by adopting on-line incremental HMM adaptation (AdaRob). On-line adaptation is more suitable for real-time applications where environmental conditions, talker position, etc. may vary substantially with time. The Table shows that in the best case, that is starting from robust models and exploiting manually segmented speech boundaries, 98.6% WRR was obtained, not far from the close-talker reference performance. In previous work [68,79] it was shown that, when starting from

clean models, both batch and on-line adaptation techniques do not achieve this performance level. Finally, note that the adaptation produces a score of 95.6% WRR in the case of a single far-microphone.

15.4 Discussion and Future Trends

Hands-free interaction represents the most natural form of human communication. Research on hands-free speech recognition is drawing scientists together to form an important discipline with numerous potential applications. In particular, various multi-modal/multi-media interaction scenarios have be conceived thanks to the enhanced functionality added to traditional ASR systems.

Because of growing research and prototype development, the field of distant-talker speech recognition using microphone arrays has developed dramatically. As seen in this chapter, the introduction of a microphone array into an ASR system has the potential to improve performance significantly. However, this is at the cost of hardware and software complexity. Additional improvements are possible through the use of adaptation/compensation techniques and specific methods for acoustic model training. Through these approaches, performance increases can be achieved even using a single microphone. Hence, it seems reasonable that future research will focus on the use of arrays consisting of few microphones and the joint application of effective techniques for an on-line reduction of the mismatch between the operating conditions and those under which the system was trained.

Given the current state of the art, future research is needed in all the directions highlighted in the previous sections, from microphone array processing, to speech activity detection, to robust acoustic features, to adaptation of the recognizer to the real environmental conditions. Furthermore, new approaches will have to be envisaged to deal with the various environmental uncertainties which characterize distant-talker speech recognition applications.

Distributed multi-microphone systems [69], with instantaneous selection of the most reliable microphone input, may represent a promising approach. Along these lines, a specific sub-band recognizer or full-band recognizer may be associated with each of the given microphones. This would be with the purpose of realizing a competitive parallel recognition framework, where the recognized word string is selected among different hypotheses.

Another approach that deserves future study is that of incorporating speech production modeling into the multi-channel system and applying nonlinear analysis techniques as those proposed in [66,67] and detailed in Chapter 7. In this way, the system may be made less sensitive to the influence of variabilities related to reverberation effects, imperfect talker location, or a talker's head movements and may better focus attention on the speech propagating in the environment.

Finally, experimental tasks and activities are important aspects to highlight. Because of relevant differences in the experimental frameworks and the type of speech material (and languages) which are adopted, results obtained by the various research teams are often not comparable to one another. Moreover, results are often provided only on the basis of simulation experiments, while real world experiments are always needed to confirm a given theory. In the past, the most relevant and widely known activities for the development of basic speech recognition technology were carried out under the ARPA-CSR program. This produced the development of common speech material and standard evaluation criteria. Hence, the creation of a common framework for all the research centers operating in this field, may allow for significant advances in this exciting discipline.

References

1. L.R. Rabiner, B.H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.
2. R. De Mori, *Spoken dialogues with computers*, Academic Press, 1998.
3. A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, Kluwer, 1992.
4. Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.
5. J.C. Junqua and J.P. Haton, *Robustness in automatic speech recognition.* Kluwer, 1996.
6. C.H. Lee, F.K. Soong, and K.K. Paliwal, *Automatic speech and speaker recognition.* Kluwer, 1996.
7. S. Furui, "Recent advances in robust speech recognition," in *Proc. of ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 11–20, 1997.
8. M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, pp. 75–95, 1998.
9. J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.
10. L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals* Prentice Hall, 1978.
11. L.R. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Sys. Tech. Journal*, vol. 54, no. 2, pp. 297–315, 1975.
12. L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, pp. 777–785, 1981.
13. H. Ney, "An optimization algorithm for determining the endpoints of isolated utterances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-81)*, Atlanta GA, USA, pp. 720–723, 1981.
14. J.C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, pp. 406–412, 1994.

15. D. O'Shaughnessy, *Speech Communications*, IEEE Press, 2000.
16. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
17. C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, vol. 25, pp. 29–47, 1998.
18. J.S. Lim, *Speech Enhancement*, Prentice Hall, 1983.
19. Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. on Signal Processing*, vol. 40, pp. 1303–1316, 1992.
20. S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction* Wiley and Teubner, 1996.
21. S. Boll, "Speech enhancement in the 1980s, Noise suppression with pattern matching," in *Advances in Speech Signal Processing*, (S. Furui and M.M. Sondhi, eds.), pp.309–325, Marcel Dakker, 1992.
22. M. Rahim and B.H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 19–30, 1996.
23. C. Lawrence and M. Rahim, "Integrated bias removal techniques for robust speech recognition," *Computer Speech and Language*, vol. 13, pp. 283–298, 1999.
24. A. Sankar and C.H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 190–202, 1996.
25. A. Nadas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.
26. I. Sanches, "Noise-compensated hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5 , pp. 533–540, 2000.
27. Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 255–266, 2000.
28. M.J.F. Gales, *Model-based techniques for noise robust speech recognition*, PhD thesis, Cambridge University, Cambridge, England, 1995.
29. M. J. F. Gales and S. J. Young, "Robust speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 352-359, 1996.
30. J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
31. C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.
32. M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.
33. M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
34. S. Das, R. Bakis, A. Nadas, D. Nahamoo, and M. Picheny, "Influence of background noise and microphone on the performance of the ibm tangora speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, Minneapolis MN, USA, pp. 95–98, Apr. 1993.

35. B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 31, no. 4, pp. 793–897, 1983.
36. S. Furui, "Robust speech recognition under adverse conditions," in *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 31–42, 1992.
37. C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
38. M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.
39. Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, 1986.
40. R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-88)*, New York NY, USA, pp. 2578–2581, Apr. 1988.
41. S. Haykin, ed., *Advances in spectrum analysis and array processing.* Prentice Hall, 1995.
42. M. W. Hoffman and K. M. Buckley, "Robust time-domain processing of broadband microphone array data," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 3, pp. 193–203, 1995.
43. S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, no. 3-4, pp. 215-27, 1996.
44. O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. of IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
45. L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1 , pp. 27–34, 1982.
46. J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Multi-microphone noise reduction techniques for hands-free speech recognition - a comparative study," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, pp. 171–174, 1999.
47. J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, 1993.
48. E.E. Jan, P. Svaizer, and J.L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. of IEEE ISCAS*, pp. 1460–1463, 1995.
49. C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech and Audio Proc.*, vol. 6, no. 3, pp. 240–259, 1998.
50. D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-90)*, Albuquerque NM, USA, pp. 833–836, Apr. 1990.
51. Y. Grenier, "A microphone array for car environments," *Speech Communication*, vol. 12, pp. 25–39, 1993.

52. T.M. Sullivan and R.M. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, Minneapolis MN, USA, pp. 91–94, Apr. 1993.
53. P. Raghavan, R.J. Renomeron, C. Che, D.S. Yuk, and J.L. Flanagan, "Speech recognition in a reverberant environment using matched filter array (MFA) processing and linguistic-tree maximum likelihood linear regression (LT-MLLR) adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 777–780, Mar. 1999.
54. T.B. Hughes, H.S. Kim, J.H. DiBiase, and H.F. Silverman, "Performance of an HMM Speech Recognizer using a real-time tracking microphone array as input," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 3 , pp. 346–349, 1999.
55. T. Takiguchi, S. Nakamura, and K. Shikano, "Speech recognition for a distant moving speaker based on HMM composition and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 1403–1406, June 2000.
56. J. Kleban and Y. Gong, "HMM adaptation and microphone array processing for distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 1411–1414, June 2000.
57. D. V. Rabinkin, R. J.Renomeron, J. C. French, and J. L. Flanagan, "Optimum microphone placement for array sound capture," *Proc. of the SPIE*, vol. 3162, pp. 227-39, 1997.
58. M. Inoue, S. Nakamura, T. Yamada, and K. Shikano, "Microphone array design measures for hands-free speech recognition," in *Proc. of EUROSPEECH*, pp. 331–334, 1997.
59. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of different microphone array configurations for hands-free speech recognition in noisy and reverberant environments," in *Proc. of EUROSPEECH*, pp. 347–350, 1997.
60. M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 227–230, Apr. 1997.
61. S. Nakamura, T. Yamada, P. Heracleous, and K. Shikano, "Recognition of distant-talking speech based on 3-D trellis search using a microphone array and adaptive beamforming," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, pp. 219–222, 1999.
62. S. Oh, and V. Viswanathan, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-92)*, San Francisco CA, USA, pp. 281–284, Mar. 1992.
63. R. Le Bouquin, "Enhancement of noisy speech signals, application to mobile radio communications," *Speech Communication*, vol. 18, pp. 3–19, 1996.
64. D. Mansour and B.H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 6, pp. 795–804, 1989.
65. B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendano, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 405–408, May 1998.
66. M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA pp. 3613–3616, May 1998.

67. M. Brandstein, "An event-based method for microphone array speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 953–956, Mar. 1999.
68. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 449–452, Mar. 1999.
69. Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 197–200, June 2000.
70. D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation," in *Proc. of ICSLP*, pp. 1329–1332, 1996.
71. Q. Lin, C.W. Che, D.S. Yuk, L. Jin, B. de Vries, J. Pearson, and J.L. Flanagan, "Robust distant-talking speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-96)*, Atlanta GA, USA, pp. 21-24, May 1996.
72. C. Che, Q. Lin, J. Pearson, B. de Vries, and J.L. Flanagan, "Microphone arrays and neural networks for robust speech recognition," in *Proc. ARPA Human Language Technology (HLT)*, pp. 342–348, 1994.
73. W. Ward, G. Elko, R. Kubli, and W. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. of Wallance Clement Sabine Centennial Symposium*, pp. 343–346, 1994.
74. N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1484–1488, 1981.
75. Y. Suzuki, F. Asano, H. Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1995.
76. J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
77. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Experiments of HMM adaptation for hands-free connected digit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 473–476, May 1998.
78. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of filtered clean speech for robust HMM training," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp. 99–102, 1999.
79. M. Matassoni, M. Omologo, and D. Giuliani, "Hands-free speech recognition using a filtered clean corpus and incremental HMM adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 1407–1410, June 2000.
80. B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," in *Proc. of ICSLP*, pp. 1391–1394, 1994.

16 Blind Separation of Acoustic Signals

Scott C. Douglas

Southern Methodist University, Dallas TX, USA

Abstract. This chapter presents an overview of criteria and algorithms for the blind separation of linearly mixed acoustic signals. Particular attention is paid to the underlying statistical formulations of various approaches to the convolutive blind signal separation task, and comparisons to other blind inverse problems are made. Several algorithms are described, including a novel algorithm that largely maintains the spectral content of the original mixtures in the extracted source signals. Numerical experiments are also provided to explore the behaviors of the algorithms in real-world blind signal separation tasks.

16.1 Introduction

16.1.1 The Cocktail Party Effect

Consider for a moment the last time you had a conversation with a friend at a social gathering. Numerous conversations were occurring around you, and yet you had the capability of focusing your attention on your friend's phrases. This so-called "cocktail party effect" [1] is well-known and taken for granted by almost everyone. What are the system elements at work in this situation?

- *An acoustic sensor array.* In our case, two sensors (our ears) collect acoustic information, although there is nothing precluding the use of more than two sensors in the task from a system design standpoint.
- *A computational processing system.* Most auditory models allow for several layers of signal processing, from lower-level frequency-dependent feature extraction within the inner ear to higher-level pattern recognition within the brain [2].

The current boom in electronic devices, while creating a revolution in human lifestyles, also creates the need for sensing and translating acoustical information into useful electronic forms. Since we use our ears as information sensors in conversations with others, it seems only natural to develop similar listening capabilities in information-processing devices. This quest remains a challenging problem, despite several decades of research in microphone arrays, voice recognition, and natural language processing. Moreover, while speech can be accurately collected using close-to-the-mouth microphones, there are applications in which such placements are impractical or undesirable, such as in audio teleconferencing systems and automobile speakerphones. Clearly,

there is a need for systems that separate multiple audio signals for improved intelligibility and fidelity. Such efforts, if successful, could transform the way we interact with devices around us. In addition, elements of these methods would be useful in a wide range of other applications:

- *Machine monitoring.* Almost all mechanical systems require periodic service and maintenance to function over long periods. When such systems are close to failing, they often emit sounds that indicate their worn-out state. Signal separation could be used as a first step in identifying the eminent failure of mechanical systems by extracting useful acoustical features from multi-channel acoustic and/or vibration signals.
- *Medical diagnosis.* Many non-invasive medical sensing technologies, such as acoustic ultrasound, characterize bodily processes through multi-channel recordings that can be difficult to decipher due to the complicated propagation properties of human body tissue. Signal separation offers the potential of extracting coherent and identifiable signal features that can be more-easily tied to specific bodily functions or ailments.
- *Musical performance.* When recording musical performances, it may be undesirable or impractical to acoustically isolate individual instruments within an ensemble. Signal separation could be used to separate musical instrument sounds in multi-channel audio recordings. Such processing methods could be an initial step in the analysis of musical performances or the removal of individual musical performances (e.g., for creating karaoke-type recordings).

16.1.2 Chapter Overview

This chapter discusses the problem of *blind signal separation (BSS)* of acoustic mixtures as recorded by a microphone array. This task is one scenario within the larger field of BSS. In BSS, multiple independent streams of information are extracted from linear mixtures of these signal streams without specific knowledge of the source signals, the mixing conditions, or the sensor array configuration. BSS has applications in a number of fields, particularly in digital communications where the separation of multiple transmitted signals in a wireless communications system is desired. BSS has also garnered interest in the neural network research community as a canonical problem for biologically inspired computational systems. The aim of this chapter is to describe the statistical underpinnings of existing acoustic BSS methods and to provide a technical foundation for further research into the field. As these methods continue to be the focus of much research, it is too early to deem the usefulness of most existing approaches; such judgments are best left for the reader to ponder.

The organization of the chapter is as follows. We first outline the structure of the acoustic BSS task in Section 16.2 and relate it to the cocktail party effect. Criteria for BSS of acoustic signals are considered in Section 16.3, along

with the corresponding assumptions on the acoustic signals. Section 16.4 discusses several algorithms for BSS of acoustic signal mixtures, along with their features and limitations. Section 16.5 presents simulation results of selected algorithms as applied to the acoustic BSS task. Conclusions and a discussion of open issues are provided in Section 16.6.

In terms of mathematical notation, all quantities are real-valued unless denoted otherwise, and discrete-time signals and systems are assumed. Scalar, vector, and matrix quantities shall be indicated by lowercase italic, lowercase bold, and uppercase bold letters, respectively. Because of the statistical formulations used within the algorithm derivations, all signals are assumed to be realizations of sampled discrete-time random processes with time index k. In practice, the procedures developed in this chapter can be used for "real world" signals that do not have simple statistical models. In such cases, deterministic averages replace statistical expectations within the formulations.

16.2 Blind Signal Separation of Convolutive Mixtures

16.2.1 Problem Structure

We first outline the structure of the BSS task and relate it to the cocktail party effect. Figure 16.1 illustrates this problem in block diagram form. On the left is the source signal vector sequence

$$\mathbf{s}(k) = [s_1(k)\ s_2(k)\ \cdots\ s_m(k)]^T, \tag{16.1}$$

where m is the number of sources and $s_i(k)$ is the ith signal source. In the cocktail party problem, each $s_i(k)$ corresponds to a sampled version of an acoustic signal as measured at its source position[1]. These source signals pass through an $(m \times n)$ linear, time-invariant system with matrix impulse response $\mathbf{A}_i$, $0 < i < \infty$, to produce the measured signal vector sequence

$$\mathbf{x}(k) = [x_1(k)\ \cdots\ \mathbf{x}_n(k)]^T = \sum_{i=0}^{\infty} \mathbf{A}_i \mathbf{s}(k-i). \tag{16.2}$$

The entries of each $(n \times m)$ matrix $\mathbf{A}_i$ within (16.2) are determined by the source locations, the sensor array locations, the acoustical properties of the physical environment, and the capabilities (e.g., directivity, frequency response) of each sensor. We lump all of these properties into the matrix sequence $\{\mathbf{A}_i\}$ for descriptive simplicity. This mixing model is termed *convolutive* because (16.2) describes a multi-channel discrete-time convolutive process. Causality of the system model is assumed, although most BSS formulations do not use this feature. Equation (16.2) assumes that no broadband

[1] The definition of a source's position is somewhat arbitrary in that the individual temporal properties of each $s_i(k)$ are unspecified. Generally, the source position is defined as the spatial center of the source's acoustic power, although any related position (e.g., the mouth of each talker in speech separation) is reasonable.

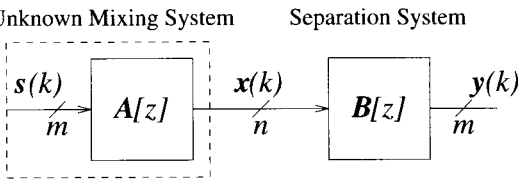


Fig. 16.1. Block diagram of the convolutive BSS task.

sensor noise is present, although many BSS techniques perform adequately for low levels of sensor noise.

The term "blind" in convolutive BSS refers to the lack of knowledge about the source signals and mixing system other than its generic linear form. In most formulations, little knowledge of the sources that generate the signals $\{s_i(k)\}$, such as their physical locations, their exact temporal characteristics, or their exact statistical properties, is available. As a consequence, the sensors need not possess a certain geometry. Thus, convolutive BSS becomes an interesting alternative to other array processing methodologies when precise knowledge of the source and/or sensor locations is either too difficult or time-consuming to obtain.

As for the numbers of sources m and sensors n in the convolutive BSS task, most formulations described in the scientific literature assume that $n \geq m$. In other words, the number of sensors is at least as great as the number of sources. Thus, each source $s_i(k)$ can be uniquely extracted from the sensor measurements $\{x_i(k)\}$ using the $(n \times m)$ multi-channel linear system[2]

$$\mathbf{y}(k) = \sum_{l=0}^{\infty} \mathbf{B}_l \mathbf{x}(k-l). \tag{16.3}$$

In (16.3), the sequence of $(m \times n)$ matrices $\mathbf{B}_l$ describe the separation system, and the output vector sequence $\mathbf{y}(k) = [y_1(k) \cdots y_m(k)]^T$ contains the estimates of the individual source signals. Again, we have assumed a causal separation model (i.e., $\mathbf{B}_l = \mathbf{0}$ for $l < 0$) for practical reasons. When implementing convolutive BSS procedures, the separation system must have a finite number of parameters, as addressed in Section 16.4.

In contrast to the physical system in (16.2), the separation system in (16.3) is an electronic signal processing device whose parameters $\{\mathbf{B}_l\}$ are computed. Equation (16.3) is analogous to the signal processing performed by the human hearing system to extract useful acoustic information. There

[2] The requirement that $n \geq m$ allows for, but does not guarantee the existence of, a multi-channel linear system as a separation solution for the convolutive BSS task. In addition, the multi-channel transfer function $\mathbf{A}[e^{\iota\omega}]$ must have a rank equal to m at all frequencies $|\omega| \leq \pi$ for which there is source signal content. This condition is similar to a pseudo-invertibility condition on the unknown system, although convolutive BSS does not require a system inverse to be calculated.

are, however, several differences between the cocktail party problem and the convolutive BSS task:

- The processing model in (16.3) is linear in form, whereas it is not clear whether human binaural processing is linear in form [2].
- The number of sources m in convolutive BSS is no greater than the number of sensors n, whereas it is often the case in human binaural processing that the number of sources m far outnumber the two ears used to collect acoustical information.
- For the most part, the content of the source signals is not leveraged in convolutive BSS solutions, whereas such acoustical content clearly plays a role in human binaural processing, e.g., in deciphering speech.

One could argue that existing approaches to convolutive BSS "miss the mark" by not utilizing nonlinear models or content recognition. Indeed, the "as-many-sensors-as-sources" assumption is a by-product of both the linear demixing model and the absence of any knowledge about the source signal content. By restricting the form of the convolutive BSS task, however, several practical and promising signal processing methods have been developed largely due to the simplicity of the problem formulation.

16.2.2 Goal of Convolutive BSS

The overall goal of the convolutive BSS task is straightforward:

> **Goal:** *Adjust the impulse response of the demixing system such that each output signal $y_i(k)$ contains one filtered version of each source signal $s_j(k)$ without replacement and without any loss of information.*

Mathematically, we can describe this solution as

$$y_i(k) = \sum_{l=0}^{\infty} d_{ijl} s_j(k-l), \quad 1 \leq \{i,j\} \leq m \tag{16.4}$$

where the one-to-one mapping $j \rightarrow i$ is arbitrary and the sequences d_{ijl}, $0 \leq l < \infty$ satisfy the mild frequency response condition

$$\sum_{l=0}^{\infty} d_{ijl} e^{\iota \omega l} \neq 0, \quad |\omega| \leq \pi \tag{16.5}$$

for every valid pair $\{i,j\}$ in the mapping, where $\iota = \sqrt{-1}$. We can only recover a filtered version of each source signal because we assume nothing about the temporal characteristics of each source signal. Hence, there is no obvious temporal structure that we can impose on the extracted signals.

The convolutive BSS task will be largely driven by a single underlying collective assumption about the source signals:

Main Assumption: *Each $s_i(k)$ is statistically independent of each $s_j(l)$ for all $i \neq j$, all k, and all l.*

This assumption implies that, for any two samples $s_1 = s_i(k)$ and $s_2 = s_j(l)$ from any two different source signals within the mixture, the joint probability density function (p.d.f.) of s_1 and s_2 can be factored into the product of their marginal p.d.f.'s as

$$p_{s_1 s_2}(s_1, s_2) = p_{s_1}(s_1) \cdot p_{s_2}(s_2). \tag{16.6}$$

Statistical independence of each source signal from every other is the main requirement in most formulations to the convolutive BSS task[3]. Statistical independence of the sources, however, is necessary but not sufficient to guarantee separability; all methods leverage some additional side information about the sources, as discussed in Section 16.3.

16.2.3 Relationship to Other Problems

The convolutive BSS task is similar in structure and goal to two other signal processing problems. We discuss these similarities here.

Instantaneous Blind Signal Separation. Interest in the blind signal separation task first surfaced in various research communities in a much simpler form than that depicted in Figure 16.1 [5,6]. In instantaneous blind signal separation, the measured vector model is given by

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k), \tag{16.7}$$

where $\mathbf{A}$ is an unknown $(n \times m)$ mixing matrix and $\mathbf{s}(k)$ is the unknown m-dimensional signal vector containing independent components. This model is a special case of that in (16.2), in which $\mathbf{A}_j = \mathbf{A}\delta(j)$, where $\delta(j)$ is the Kronnecker unit impulse function. Here, mixing is termed *instantaneous* in that no dispersive effects or time delays are present. As a result, the $(m \times n)$ separation system can be instantaneous in form as well, as given by

$$\mathbf{y}(k) = \mathbf{B}\mathbf{x}(k). \tag{16.8}$$

The goal of instantaneous BSS is to adjust the entries in $\mathbf{B}$ such that

$$\mathbf{B}\mathbf{A} = \mathbf{\Phi}\mathbf{D}, \tag{16.9}$$

where $\mathbf{\Phi}$ is an $(m \times m)$ permutation matrix and $\mathbf{D}$ is a diagonal matrix of non-zero entries d_j. If such is the case, then

$$y_i(k) = d_j s_j(k), \quad 1 \leq \{i, j\} \leq m \tag{16.10}$$

[3] Some convolutive BSS formulations only require that the sources be spatially and temporally uncorrelated from each other as opposed to being statistically independent. Additional features about the source signals and/or mixing conditions must then be imposed to solve the convolutive BSS task [3,4]

where again the one-to-one mapping $j \rightarrow i$ that is defined by the non-zero entries of $\boldsymbol{\Phi}$ is arbitrary.

Instantaneous BSS can be viewed as a special case of convolutive BSS, although the two problems differ somewhat in the ways they are solved. In fact, designing algorithms and solutions for instantaneous BSS is easier than developing corresponding solutions to convolutive BSS simply due to implementation aspects, as the latter methods require systems with many more adjustable parameters. When the source signals are distinctly narrowband, convolutive BSS reduced to the instantaneous BSS task due to the linear time-invariant nature of the mixing model. Applications of instantaneous BSS can be found in digital communications and medical diagnosis, among other applications. The recently published edited monograph [7] is devoted to the subject, and good overviews of the problems and issues in instantaneous BSS can be found in [8] and [9].

In this chapter, we shall often introduce important concepts and criteria in the context of the instantaneous BSS task, as this approach provides insight into the structure of all BSS problems in a simplified notation.

Multi-channel Blind Deconvolution. Multi-channel blind deconvolution is similar to convolutive BSS both in structure and goal. Both tasks involve identical mixing models and demixing system structures as shown in Figure 16.1. Whereas convolutive BSS attempts to make the extracted output signals $\{y_i(k)\}$ spatially independent, however, multi-channel blind deconvolution attempts to enforce both *spatial and temporal* independence of the extracted output signals. This additional requirement only makes sense if the source signals $\{s_j(k)\}$ are both spatially and temporally independent sequences, such that $s_i(k)$ is independent of $s_j(k+l)$ for all $i \neq j$ and any l or for all $\{i,j\}$ and any $l \neq 0$. In such cases, ideal multi-channel blind deconvolution yields

$$y_i(k) = d_{ij\Delta_j} s_j(k - \Delta_j) \tag{16.11}$$

where $j \rightarrow i$ is again a one-to-one mapping and Δ_j is an arbitrary delay.

At first glance, multi-channel blind deconvolution appears to be a special instance of convolutive BSS in which a particular temporal structure is imposed on the extracted output signals. Such a correspondence has led researchers to use multi-channel blind deconvolution algorithms to solve convolutive BSS tasks [10–13]. As we shall show, however, multi-channel blind deconvolution methods can alter the temporal structure of acoustic source signals within the extracted outputs in undesirable ways. Thus, such deconvolution methods are most appropriate when the source signals are temporally independent. One such case is in wideband antenna arrays for wireless communications. See [14,15] for recent discussions of multi-channel blind deconvolution methods.

Multi-channel blind deconvolution can also be viewed as the multi-channel extension of the single-channel blind deconvolution task. Blind deconvolution

is a well-studied topic, and the two edited monographs [16] and [17] provide a good introduction to the field. In digital communications, blind deconvolution is termed *blind equalization* due to the fact that a deconvolved signal has a nearly flat power spectrum.

16.3 Criteria for Blind Signal Separation

16.3.1 Overview of BSS Criteria

The capabilities of any convolutive BSS technique mainly depend on the separation criterion that is employed. These criteria in turn depend on the underlying assumptions about the source signals. Almost all BSS methods leverage the fact that the source signals are statistically independent of one another; however, all methods make use of additional signal features to achieve separation. We discuss these assumptions in connection with the criteria employed.

Convolutive BSS criteria can be categorized into one of three groups: (i) density modeling criteria, (ii) contrast functions, and (iii) correlation-based criteria. We discuss each of these criteria in turn, introducing basic concepts in the context of the instantaneous BSS problem. Extensions to the convolutive BSS task differ largely in algorithm implementation, as shown in Section 16.4.

16.3.2 Density Modeling Criteria

Density-based BSS methods rely heavily on concepts in *information theory,* a field with applications in communications, economics, neuroscience, and physics [18]. Information theory is useful for BSS because with it one can characterize the amount of shared information in a set of signals. Intuitively, separation is obtained when there is no common information among any two output signal subsets. Such methods are density-based because they effectively model the joint p.d.f. of the output signal vector sequence $\mathbf{y}(k)$ as they adjust the demixing system to produce nearly independent signal sequences. For simplicity, we restrict our study to instantaneously mixed sources such that the models in (16.7) and (16.8) are appropriate.

Although many information-theoretic formulations to instantaneous BSS can be developed [10,19], all of these formulations can be unified using the Kullback-Leibler divergence measure [8]

$$D(p_{\mathbf{y}}||\widehat{p}_{\mathbf{y}}) = \int p_{\mathbf{y}}(\mathbf{y}) \log \left(\frac{p_{\mathbf{y}}(\mathbf{y})}{\widehat{p}_{\mathbf{y}}(\mathbf{y})} \right) d\mathbf{y}, \tag{16.12}$$

where $p_{\mathbf{y}}(\mathbf{y})$ and $\widehat{p}_{\mathbf{y}}(\mathbf{y})$ are the actual and model distributions, respectively, of $\mathbf{y}(k)$. Equation (16.12) measures the "distance" between $p_{\mathbf{y}}(\mathbf{y})$ and $\widehat{p}_{\mathbf{y}}(\mathbf{y})$, although this measure is asymmetric. We can rewrite (16.12) using the expectation operator $E\{\cdot\}$ as

$$D(p_{\mathbf{y}}||\widehat{p}_{\mathbf{y}}) = E\left\{ \log \left(\frac{p_{\mathbf{y}}(\mathbf{y})}{\widehat{p}_{\mathbf{y}}(\mathbf{y})} \right) \right\}. \tag{16.13}$$

Replacing ensemble averages with sample averages, one obtains a criterion that depends on output signal measurements directly. Further simplifications are possible and shall be discussed shortly.

The choice of $\widehat{p}_{\mathbf{y}}(\mathbf{y})$ is governed by the assumptions on and *a priori* knowledge of $\mathbf{s}(k)$. If all $s_i(k)$ are identically distributed, a reasonable choice is

$$\widehat{p}_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{m} p_s(y_i), \tag{16.14}$$

yielding a maximum-likelihood estimate of the demixing matrix $\mathbf{B}$. Alternatively, by estimating the marginal p.d.f.'s $p_{y_i}(y_i)$ for the current $\mathbf{B}$ and setting $\widehat{p}_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{m} p_{y_i}(y_i)$, one obtains a minimum mutual information approach to BSS [8]. In either case, the criterion in (16.12) attempts to measure the degree of independence of the individually extracted source signals, because $\widehat{p}_{\mathbf{y}}(\mathbf{y})$ describes a signal vector with statistically independent components.

The main advantage of (16.12) as a cost function is its statistical efficiency. With a "good" choice of model $\widehat{p}_{\mathbf{y}}(\mathbf{y})$, one can make good use of the source signal characteristics, e.g., to obtain fast convergence for a given data set size. The main drawback to (16.12) is the practical need to *choose* a model. The most general choice of criterion, minimum mutual information, is computationally demanding; whereas the most-restrictive criterion choice, maximum likelihood, can actually fail to achieve separation if the marginal p.d.f.'s within $\widehat{p}_{\mathbf{y}}(\mathbf{y})$ differ too much from the actual source distributions. See [20] for the stability conditions that the marginal p.d.f.'s within $\widehat{p}_{\mathbf{y}}(\mathbf{y})$ must satisfy for (16.12) to be locally stable about a separating solution in instantaneous BSS tasks.

Equation (16.12) has not been proven to possess only separating minima. As such, separation is not guaranteed with this cost function. Even so, numerous simulations have indicated that ill-convergence appears to be rare for even gross marginal approximations within $\widehat{p}_{\mathbf{y}}(\mathbf{y})$. For example, suppose each source signal is a sampled speech signal. The instantaneous distribution of speech is approximately Laplacian-distributed with p.d.f [21]

$$p_s(s) = \frac{1}{\sigma\sqrt{2}} \exp\left(-\frac{|s|\sqrt{2}}{\sigma}\right), \tag{16.15}$$

where σ^2 is the variance of the speech signal. Employing (16.15) within (16.13), one obtains a criterion that consistently yields separated speech signals in instantaneous BSS tasks using simple gradient-type optimization procedures. Moreover, the variance σ^2 can be arbitrarily set within the criterion (e.g., $\sigma^2 = 1$), as the scales of the sources are effectively absorbed into the mixing conditions within the model. In situations where the source signals come from a select set of distributions, one can develop procedures for choosing a "good" distribution model for each extracted source from the family of models [22]. Alternatively, one can develop parametric density models and successively estimate the model parameters from the extracted sources [23].

Although it might not be apparent within (16.13), one cannot employ a Gaussian marginal p.d.f. $p_s(y_i) = (2\pi)^{-1/2}\exp(-|y_i|^2/2)$ to obtain separation; such a choice only *decorrelates* the corresponding output signals [8]. While statistical independence implies uncorrelatedness, the converse is not true. Consider the two signal vectors $\mathbf{s}(k)$ and $\boldsymbol{\Gamma}\mathbf{s}(k)$, where the elements of $\mathbf{s}(k)$ are statistically independent with unit variance and $\boldsymbol{\Gamma}$ is an orthonormal matrix such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \mathbf{I}$. Both $\mathbf{s}(k)$ and $\boldsymbol{\Gamma}\mathbf{s}(k)$ have the same autocorrelation matrix $E\{\mathbf{s}(k)\mathbf{s}^T(k)\} = \mathbf{I}$, i.e., uncorrelated elements. However, $\boldsymbol{\Gamma}\mathbf{s}(k)$ clearly contains signal mixtures. As a correlary, separation methods using density modeling require non-Gaussian-distributed sources in order to uniquely identify a separation system. This issue is not a problem in practice, because most interesting acoustic sources have non-Gaussian p.d.f.'s.

In convolutive BSS, the cost function in (16.13) depends on the p.d.f. of the extracted sources in $\mathbf{y}(k)$, which in turn depends on the impulse response $\{\mathbf{B}_l\}$ of the separation system. Can this expression be simplified further? This issue has been addressed in the context of dispersive systems by Pham [24], who has determined expressions for the entropy of a multi-dimensional random process when filtered by a multi-channel linear system. A cost function $E\{\mathcal{J}(\{\mathbf{B}_l\})\}$ for the impulse response $\{\mathbf{B}_l\}$ of the multi-channel separation system in (16.3) can then be developed that is identical (up to a fixed constant) to that in (16.13). This cost function is

$$E\{\mathcal{J}(\{\mathbf{B}_l\})\} = \oint \log|\det \mathbf{B}[z]|z^{-1}dz - E\{\log \widehat{p}_{\mathbf{y}}(\mathbf{y}(k))\}, \qquad (16.16)$$

where $\mathbf{B}[z]$ is the z-transform of the impulse response sequence $\{\mathbf{B}_l\}$. The first term on the right-hand side of (16.16) is a contour integral about the origin in the complex z-plane and represents the effects of a linear transformation on the joint entropy of the extracted output sequences. The second term measures the degree of independence of the extracted outputs when (16.14) is used. Again, we can replace expectations by sample averages to obtain a criterion that depends entirely on data alone. In later sections, we shall describe algorithms that minimize (16.16) in an iterative fashion.

16.3.3 Contrast Functions

Contrast functions, as introduced in the context of instantaneous BSS by Comon [5], are an alternative to density-based criteria. The concept of a contrast function is similar in spirit to that of an indicator light on an electrical or mechanical device. Such a light indicates a certain operating condition about the device. Similarly, a contrast function identifies when one output $y_i(k)$ of a separation system contains elements of only one source signal $s_j(k)$. The key issue is to determine such a function that depends only on $y_i(k)$ and not on the mixing conditions. Again, we shall introduce the concepts behind contrast functions using instantaneously mixed source signals.

Consider the instantaneous mixture and separation models in (16.7) and (16.8), respectively, and define the combined system matrix $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{BA}. \tag{16.17}$$

Then, we can express the ith extracted output signal in terms of the elements $\{c_{ij}\}$ of $\mathbf{C}$ as

$$y_i(k) = \sum_{j=1}^{m} c_{ij} s_j(k). \tag{16.18}$$

A contrast function is a cost function $\phi[y_i(k)]$ for which a local maximum over all elements of c_{ij}, $1 \leq j \leq m$ corresponds to the separated solution

$$c_{ij} = \begin{cases} d_l, \text{ for a single value of } l,\ 1 \leq l \leq m \\ 0, \text{ otherwise.} \end{cases} \tag{16.19}$$

In practice, this cost function is expressed in terms of the elements of the separation matrix $\mathbf{B}$ for which optimization takes place.

In contrast-based BSS, the contrast function criterion must be designed for the particular separation task. This design is driven by two main goals:

- The contrast function must be simple to evaluate.
- The contrast function must identify a separated result for the given source signal statistics through its maxima.

As an illustrative example, consider the *normalized kurtosis,* which is a measure of the shape of a zero-mean univariate p.d.f. as normalized by its variance. The normalized kurtosis of the random variable y is given by

$$\kappa_y = \frac{E\{|y|^4\}}{E^2\{|y|^2\}} - 3. \tag{16.20}$$

It can be shown that the contrast function

$$\phi[y_i(k)] = |\kappa_{y_i(k)}| \tag{16.21}$$

correctly identifies a separated solution for instantaneous BSS as defined in (16.18) through its local maxima, as long as all source signals $s_j(k)$ have nonzero kurtosis [5]. The most well-known zero-kurtosis distribution is the Gaussian p.d.f., such that contrast-based methods are generally designed for non-Gaussian sources.

To extract all of the sources from a signal mixture, m contrast functions are used to calculate the separation parameters for the m system outputs. How does one guarantee that different sources are extracted at different outputs of the system in this approach? Details regarding this issue are considered in Section 16.4.

Like density matching, BSS methods that use contrast functions rely on the spatial independence and non-Gaussianity of the source signals to perform separation. Unlike density based methods, however, contrast-based BSS

methods do not require significant knowledge about the source signal p.d.f.'s. This feature is particularly important in practical applications, because a "blind" system should perform separation regardless of the source signal distributions. In some situations, however, using certain moment-based criteria such as the normalized kurtosis can yield inefficient estimates of the separation system. Such is the case when extracting speech signals with the normalized kurtosis contrast in (16.21), for example, due to the impulsive nature of typical speech p.d.f.'s.

The above discussion carries over to the case of convolutive BSS with appropriate modifications. It has been shown, for example, that the normalized kurtosis contrast function is valid for convolutive BSS [15,25,26]. Contrast functions have strong connections to *cumulants,* statistical quantities that arise from the characteristic function of a multivariate p.d.f. through a Taylor series expansion [8]. Additional cumulant-based separation criteria can also be developed [26].

16.3.4 Correlation-Based Criteria

Both density modeling and contrast function BSS approaches employ nonquadratic criteria. One must resort to iterative search procedures to optimize the separation system parameters using these criteria. Convergence speed is often an issue in such cases, especially in acoustic BSS tasks where thousands of filter parameters are often involved. In this regard, it is desirable to develop more direct approaches to BSS employing structured computations.

In [27], an approach for blindly separating mixtures of instantaneously mixed sources is presented that employs correlations of the measured signals $\mathbf{x}(k)$ at different time instants. This method assumes that the source signals are statistically independent and stationary but temporally correlated, such that the correlation matrix

$$\mathbf{R}_{\mathbf{xx}}(k,l) = E\{\mathbf{x}(k)\mathbf{x}^T(k-l)\} \tag{16.22}$$

exhibits a unique eigenvalue structure for at least two different time lags $l = l_1$ and $l = l_2$. Note that

$$\mathbf{R}_{\mathbf{xx}}(k,l) = \mathbf{A}E\{\mathbf{s}(k)\mathbf{s}^T(k-l)\}\mathbf{A}^T, \tag{16.23}$$

where the matrix $E\{\mathbf{s}(k)\mathbf{s}^T(k-l)\}$ is diagonal due to the independence of the source signals. Define the normalized matrix

$$\begin{aligned}\overline{\mathbf{R}}(l_1,l_2) &= \mathbf{R}_{\mathbf{xx}}(k,l_1)[\mathbf{R}_{\mathbf{xx}}(k,l_2)]^{-1} \\ &= \mathbf{A}E\{\mathbf{s}(k)\mathbf{s}^T(k-l_1)\}\mathbf{A}^T[\mathbf{A}E\{\mathbf{s}(k)\mathbf{s}^T(k-l_2)\}\mathbf{A}^T]^{-1} \\ &= \mathbf{A}\overline{\boldsymbol{\Lambda}}(l_1,l_2)\mathbf{A}^{-1},\end{aligned} \tag{16.24}$$

where we have defined $\overline{\boldsymbol{\Lambda}}(l_1,l_2) = E\{\mathbf{s}(k)\mathbf{s}^T(k-l_1)\}[E\{\mathbf{s}(k)\mathbf{s}^T(k-l_2)\}]^{-1}$. Since $\overline{\boldsymbol{\Lambda}}(l_1,l_2)$ is diagonal, (16.24) is in the form of an eigenvalue decomposition of $\overline{\mathbf{R}}(l_1,l_2)$, where the mixing matrix $\mathbf{A}$ contains the eigenvectors

of this matrix. Thus, we can determine $\mathbf{A}$ from $\overline{\mathbf{R}}(l_1, l_2)$ using well-known eigenvalue procedures, from which the separation matrix $\mathbf{B}$ can be found by inverting $\mathbf{A}$. By defining a normalized matrix $[\mathbf{R_{xx}}(k, l_1)]^{-1}\mathbf{R_{xx}}(k, l_2)$ whose eigenvectors are $\mathbf{B} = \mathbf{A}^{-1}$, this final inversion step can be avoided.

Matrix inversions are challenging to compute; as such, the above procedures are not particularly useful. Fortunately, it is possible to estimate $\mathbf{B}$ using *joint diagonalization,* in which $\mathbf{B}$ is computed so that

$$\mathbf{B}\mathbf{R_{xx}}(k, l)\mathbf{B}^T = \mathbf{\Lambda}(k, l) \tag{16.25}$$

for two time instants $l = \{l_1, l_2\}$, where $\mathbf{\Lambda}(k, l)$ is diagonal for all k and $l = \{l_1, l_2\}$. In this case, both $\mathbf{B}$ and $\mathbf{\Lambda}(k, l)$ can be estimated jointly, e.g., though a nonlinear optimization on a suitable least-squares cost function [28].

The above separation method relies on data correlation properties that may be difficult to verify in a practical setting. Fortunately, one can employ an identical procedure using assumptions that are quite plausible in many acoustic settings. Suppose that the sources are *nonstationary*, such that

$$\mathbf{R_{xx}}(k) = \mathbf{A}E\{\mathbf{s}(k)\mathbf{s}^T(k)\}\mathbf{A}^T \tag{16.26}$$

exhibits a unique eigenvalue structure for two different time instants $k = k_1$ and $k = k_2$. Then, one can define the normalized matrices $\mathbf{R_{xx}}(k_1)[\mathbf{R_{xx}}(k_2)]^{-1}$ or $[\mathbf{R_{xx}}(k_1)]^{-1}\mathbf{R_{xx}}(k_2)$ for use within the eigenvalue approach to blind system identification or signal separation. More importantly, one can use joint diagonalization to estimate $\mathbf{B}$ directly. Nonstationarity of the sources is clearly reasonable for speech signals. Several issues govern the success of this approach, as discussed in the next section.

16.4 Structures and Algorithms for Blind Signal Separation

We now describe filter structures and algorithms for convolutive blind signal separation. These methods employ the criteria described previously and use iterative gradient optimization procedures for simplicity.

16.4.1 Filter Structures

Because convolutive mixing is a linear process, a multi-channel linear system is sufficient to perform separation. How should this system be implemented? Several issues affect the implementation of acoustic signal separation systems:

- *Amount of room reverberation.* When separating acoustic signals in a room environment, room reverberation plays a significant role in choosing a separation model. To obtain good separation, the impulse response of the BSS system must account for the majority of the room's multipath effects.

- *Stability of the separation system.* Most convolutive BSS methods use adaptive procedures for adjusting the system's parameters. This system must remain bounded-input, bounded-output stable during adaptation.
- *Computational complexity.* Because of the dispersive effects of most room environments, the BSS system's impulse response must be *thousands of taps long* in order to get good separation quality at typical audio sampling rates and for typical rooms.

Of all possible implementations, finite impulse response (FIR) filters represent ideal candidates due to their simplicity and guaranteed stability. Moreover, they can be implemented efficiently in a block fashion using fast convolution techniques, such that their complexity scales as $\mathcal{O}(\log_2 L)$ where L is the system's filter length [29]. In such block implementations, a bulk delay is introduced into the input-output path of the system, and the coefficient adaptation method must inherently be block-based as well. The first issue does not represent a problem in most cases, and the second issue often has a negligible effect on the performance of most convolutive BSS methods.

For the remainder of the chapter, we shall replace (16.3) by

$$\mathbf{y}(k) = \sum_{l=0}^{L} \mathbf{B}_l(k)\mathbf{x}(k-l), \tag{16.27}$$

where $\mathbf{B}_l(k)$, $0 \leq l \leq L$, are the multi-channel FIR filter parameters. See [30] for a discussion of other system structures for convolutive BSS tasks.

16.4.2 Density Matching BSS Using Natural Gradient Adaptation

In this section, we review a stochastic gradient procedure for minimizing the cost function in (16.16) and discuss some of its properties. We then describe a modified version of the algorithm that is ideally suited to acoustic BSS tasks.

A convolutive BSS algorithm based on the criterion in (16.16) can be derived in a standard fashion by (i) calculating the gradient of this cost function, (ii) establishing differential updates for the filter parameters based on gradient minimization, and (iii) discretizing the differential updates. Such an approach leads to a difficulty: the gradient of the first term on the right-hand side of (16.16) requires the *pseudo-inversion* of the multi-channel separation system $\mathbf{B}[z]$, a challenging proposition. Fortunately, it is possible to transform this update into a *natural gradient* adaptation procedure. Natural gradient adaptation is a modified gradient descent procedure in which the structure of the parameter space is employed to adjust the search direction for more efficient adaptation. Unlike Newton-based methods, natural gradient adaptation does not assume a locally quadratic cost function, and no spurious algorithm stationary points are created. Moreover, in the case of inverse problems such as BSS and blind deconvolution, natural gradient adaptation possesses an inherent statistical efficiency that overcomes any ill-conditioning introduced by

the mixing system. Most importantly, the natural gradient update for minimizing (16.16) avoids inverting $\mathbf{B}[z]$. For more details on natural gradient adaptation, including applications to statistical estimation, blind equalization, and subspace tracking, see [31].

Due to space limitations, we avoid deriving the natural gradient algorithm and and instead simply present its time-domain form [32], although frequency-domain variations of the method have been developed by several researchers [10,12,13]. This algorithm is given by

$$\mathbf{B}_l(k+1) = \mathbf{B}_l(k) + \mathbf{M}(k)\left[\mathbf{B}_l(k) - E\{\mathbf{f}(\mathbf{y}(k-L))\mathbf{u}^T(k-l)\}\right], \quad (16.28)$$

where

$$\mathbf{u}(k) = \sum_{q=0}^{L} \mathbf{B}_{L-q}^T(k)\mathbf{y}(k-q) \quad (16.29)$$

$$\mathbf{f}(\mathbf{y}(k)) = [f(y_1(k))\ f(y_2(k))\ \cdots\ f(y_m(k))]^T \quad (16.30)$$

$$f(y) = -\frac{\partial \log p_s(y)}{\partial y}, \quad (16.31)$$

$\mathbf{M}(k)$ is a diagonal matrix of step sizes $\mu_i(k)$, and $E\{\cdot\}$ is an averaging operation. This algorithm assumes that each source signal has the approximate marginal density $p_s(y)$. If Laplacian-distributed sources are assumed, then

$$f(y) = \mathrm{sgn}(y) = \begin{cases} 1 \text{ if } y > 0 \\ 0 \text{ if } y = 0 \\ -1 \text{ if } y < 0. \end{cases} \quad (16.32)$$

This algorithm only employs simple operations and makes extensive use of multiply/adds. In its time-domain form, the algorithm requires approximately four multiply/adds per filter coefficient at each time instant k. Moreover, when block-based updates are employed, the correlation term $E\{\mathbf{f}(\mathbf{y}(k-L))\mathbf{u}^T(k-l)\}$ can be computed using fast convolution methods, making the entire system of $\mathcal{O}(\log_2 L)$ complexity.

The density-matching algorithm in (16.28)–(16.31) was originally derived for the multi-channel blind deconvolution task; hence, it imposes certain temporal constraints that are undesirable for convolutive BSS problems. It can be shown that the stationary points of (16.28) must satisfy

$$E\{f(y_i(k))y_j(k-l)\} = \delta_{ij}\delta(l),\ 1 \le \{i,j\} \le m, \text{ all } l. \quad (16.33)$$

For multi-channel blind deconvolution tasks, these conditions are desirable, as they hold if the elements of $\mathbf{y}(k)$ are temporally independent. For convolutive BSS, however, the conditions in (16.33) are undesirable and lead to extracted signals that have nearly flat frequency spectra. We now propose a modification to (16.28) that avoids temporal constraints on the individually extracted source signals. The proposed algorithm is

$$\begin{aligned}\mathbf{B}_l(k+1) = \mathbf{B}_l(k) + \mathbf{M}(k)\big[&\mathrm{diag}[E\{\mathbf{f}(\mathbf{y}(k-L))\mathbf{y}^T(k-L-l)\}] \\ &- E\{\mathbf{f}(\mathbf{y}(k-L))\mathbf{y}^T(k-L-l)\}\big] \star \mathbf{B}_l(k),\end{aligned} \quad (16.34)$$

where $\star$ denotes discrete-time convolution of matrix sequences over l. Simplifying (16.34) assuming slow coefficient adaptation speeds produces

$$\mathbf{B}_l(k+1) = \mathbf{B}_l(k) + \mathbf{M}(k)\left[E\{\mathbf{V}_l(k)\} - E\{\mathbf{f}(\mathbf{y}(k-L))\mathbf{u}^T(k-l)\}\right] \tag{16.35}$$

where the (i,j)th entry of $\mathbf{V}_l(k)$ is

$$v_{ijl}(k) = f(y_i(k-L))u_{ij}(k-l) \tag{16.36}$$

$$u_{ij}(k) = \sum_{q=0}^{L} w_{ij(L-q)}(k)y_i(k-q). \tag{16.37}$$

The complexity of this algorithm is approximately the same as that of (16.28), because $u_{ij}(k)$ is computed as part of $\mathbf{u}(k) = [u_1(k) \cdots u_n(k)]^T$ through $u_j(k) = \sum_{i=1}^{m} u_{ij}(k)$.

It is a straightforward exercise to determine the conditions on the output sequence $\mathbf{y}(k)$ corresponding to $\mathbf{B}_l(k+1) = \mathbf{B}_l(k)$ for (16.35), i.e., a stationary point of the update. These conditions are

$$E\{f(y_i(k))y_j(k-l)\} = 0 \quad 1 \le i \ne j \le m, \text{ all } l. \tag{16.38}$$

Unlike (16.33), Equation (16.38) implies *no particular temporal structure* on the extracted output signals. Thus, the extracted signals are not deconvolved or decorrelated at a separating solution.

Our convolutive BSS algorithm is a nonholonomic version of the multichannel blind deconvolution algorithm in (16.28). For a discussion of nonholonomic constraints in instantaneous BSS tasks, see [33].

16.4.3 Contrast-Based BSS Under Prewhitening Constraints

In this section, we discuss the general structure of contrast-based methods for convolutive BSS tasks. We then describe one particular BSS method that uses prewhitening as a constraint within a simple contrast optimization procedure.

As discussed in Section 16.3, a contrast function identifies an independent source when it is extracted from a linear mixture. To extract m sources in parallel, the joint maximization criterion

$$\mathcal{J}_C(\mathbf{y}(k)) = \sum_{i=1}^{m} \phi[y_i(k)] \tag{16.39}$$

could be used. Unfortunately, such a procedure does not guarantee that $y_i(k)$ and $y_j(k)$ correspond to different source signals for $i \ne j$. To extract all of the sources, a contrast-based convolutive BSS method must either

- construct the BSS system so that unique parametric solutions for each system output are obtained (i.e., parameter-dependent constraints); or

- impose certain properties on the joint statistics of the extracted signals (i.e., signal-dependent constraints).

As an example of parameter-dependent constraints, consider a factorization of the separation system $\mathbf{B}[z]$ into two multi-channel systems as [14]

$$\mathbf{B}[z] = \mathbf{W}[z]\P[z]. \tag{16.40}$$

The n-input, m-output *decorrelation system* $\P[z]$ filters the signal mixtures $\mathbf{x}(k)$ so that the outputs of this system are both spatially and temporally uncorrelated with unit variances. One way to achieve this result is through *multi-channel linear prediction*, whereby the coefficients of $\P[z]$ are computed using the multi-channel Levinson algorithm and the autocorrelation statistics of $\mathbf{x}(k)$. The $(m \times m)$ *separation system* $\mathbf{W}[z]$ can then be constrained to be *paraunitary*, such that its impulse response $\{\mathbf{W}_l\}$ satisfies

$$\sum_{j=-\infty}^{\infty} \mathbf{W}_j \mathbf{W}_{j+l}^T \approx \delta(l)\mathbf{I} \tag{16.41}$$

for all l. The system constraints in (16.41) guarantee that each output signal $y_i(k)$ corresponds to a different source signal when contrast function optimization is performed over the impulse response $\{\mathbf{W}_l\}$ of the separation system. These constraints can be imposed explicitly through a structured parametrization [34], or they can be imposed implicitly using gradient-based adaptive procedures [35]. The latter methods are preferable in acoustic BSS tasks because they allow direct-form FIR filters to be used.

Alternatively, signal-dependent constraints can be employed. Here, we describe one such procedure [36]. The coefficient updates for the multi-channel FIR filter in (16.3) are given by

$$\begin{aligned}\mathbf{B}_l(k+1) = \mathbf{B}_l(k) &+ \mu(k)\left[\mathbf{B}_l(k) - E\{\mathbf{y}(k-L)\mathbf{u}^T(k-l)\}\right] \\ &+ \mathbf{D}_\beta(k)\left[E\{\mathbf{g}(\mathbf{y}(k-L))\mathbf{u}^T(k-l)\} - E\{\mathbf{N}(k-L)\}\mathbf{B}_l(k)\right],\end{aligned} \tag{16.42}$$

where $\mathbf{D}_\beta(k)$ is a diagonal matrix of step sizes $\beta_i(k)$, $\mathbf{N}(k)$ is a diagonal matrix whose diagonal elements are $y_i^4(k)$, $\mathbf{g}(\mathbf{y}(k))$ is a vector of cubic nonlinearities $g(y_i) = y_i^3$, and the other signal quantities are defined previously. This algorithm is similar to the EASI algorithm for instantaneous BSS tasks [8] in that it uses a two-term update. The first bracketed term on the right-hand side of (16.42) guarantees that the output signals $\mathbf{y}(k)$ are spatially and temporally decorrelated, and the second bracketed term is associated with the maximization of the normalized kurtosis contrast function in (16.21) for all $1 \leq i \leq m$. Although this algorithm appears to be similar to the multi-channel blind deconvolution algorithm in (16.28), it differs in one important respect: it is locally stable at a separating solution if

$$\beta_i(k)\kappa_{y_i(k)} > 0 \tag{16.43}$$

for suitably small $|\beta_i(k)|$ and all values of $\kappa_{y_i(k)} \neq 0$. In other words, (16.42) can potentially separate arbitrary signal mixtures, without specific knowledge of the source signal distributions. The chief drawback to (16.42) with respect to the natural gradient procedure in (16.28) is that the former algorithm requires six multiply/adds per FIR filter coefficient versus four multiply/adds for (16.42). Both algorithms can be implemented using fast convolution techniques.

In addition to the above methods, one can use sequential procedures for extracting sources via contrast functions, in which individual sources are estimated and then removed from the mixture in an m-stage procedure [26].

16.4.4 Temporal Decorrelation BSS for Nonstationary Sources

In this section, we present an algorithm that employs the nonstationarity of acoustic source signals to perform convolutive BSS. More details about this technique can be found in [28].

One technique for translating instantaneous BSS techniques to convolutive BSS tasks is to use frequency-domain processing methods. In such translations, each channel of the current input data block of size T samples is first processed using a T-dimensional discrete Fourier transform (DFT). Then, the resulting transformed components are grouped by bin value to form T different spatial mixtures, at which point each mixture is treated independently as an instantaneous BSS procedure with a complex-valued separation matrix. Once separation is achieved at each bin value, the separated frequency components are processed by an inverse DFT to produce the reconstructed signals. While intuitive, this approach can fail to provide good separation quality due to the possible *order permutation* of the individually extracted frequency components. In other words, the separation procedures do not guarantee that the same source will be extracted at the same output at every frequency bin within the separation structure because the various instantaneous BSS procedures are uncoupled within the system. One potential solution to this difficulty is the use of a truncated separation filter relative to the data block size, such that $L \ll T$ is maintained. This choice introduces coupling between the frequency bins during the optimization task and appears to minimize the artificially introduced permutation problem. Reasonable separation results have been reported with such an approach, although choosing actual values of L and T remains a heuristic procedure.

With the translation technique in place, the correlation-based separation method in Section 16.3 can easily be extended to the convolutive BSS task. Define the complex-valued DFT of the data frame $[\mathbf{x}(k), \ldots, \mathbf{x}(k+T-1)]$ as

$$\mathbf{x}(\omega, k) = \sum_{l=0}^{T-1} \mathbf{x}(k+l) e^{-\iota\omega l}, \quad \omega = \{0, \pi/T, \ldots, 2\pi(T-1)/T\}. \tag{16.44}$$

Similarly, define the DFT of the separation system coefficients at time k as

$$\mathbf{B}_k[e^{\iota\omega}] = \sum_{l=0}^{L} \mathbf{B}_l(k) e^{-\iota\omega l}, \quad \omega = \{0, \pi/T, \ldots, 2\pi(T-1)/T\}. \quad (16.45)$$

Define the autocorrelation matrix of $\mathbf{x}(\omega, k)$ as

$$\mathbf{R}_k(\omega) = \frac{1}{N} \sum_{q=0}^{N-1} \mathbf{x}(\omega, k+qT)\mathbf{x}^H(\omega, k+qT), \quad (16.46)$$

where ^{H} denotes Hermitian (conjugate) transpose and N is the averaging interval. Then, the goal of the nonstationary BSS task is to determine a sequence of complex-valued matrices $\mathbf{B}_k[e^{\iota\omega}]$ for all valid ω such that

$$\mathbf{B}_k[e^{\iota\omega}]\mathbf{R}_{k_\tau}(\omega)\mathbf{B}_k^H[e^{\iota\omega}] = \mathbf{\Lambda}_{k_\tau}(\omega) \quad (16.47)$$

for K non-overlapping time intervals indexed by k_τ, $1 \leq \tau \leq K$, where $\mathbf{\Lambda}_k(\omega)$ is a diagonal matrix of unknown signal powers for a given frequency bin value and time k_τ. The choice of K depends on the number of channels of data being processed, and the intervals k_τ are chosen so that the averaging windows are non-overlapping. See [28] for further details on these issues as well as an evaluation of the procedure on real-world speech data.

16.5 Numerical Evaluations

This section presents the separation results of algorithms in two acoustic BSS tasks. The real-world signal mixtures used for these evaluations are

- a two-channel recording of two male speakers reading different parts of a newspaper article in a conference room [13], and
- a two-channel recording of a male-female *a cappella* duet taken from an audio compact disc [37].

Each signal set presents specific features and challenges. The f_s = 16kHz-sampled, 16-bit, t = 18.75s speech-speech recording contains a moderate level of reverberation as well as a nominal level of fan noise due to room ventilation. The f_s = 44.1kHz-sampled, 16-bit, t = 139.3s *a cappella* duet recording contains close harmonies with significant spectral overlap of the individual vocalists' voices, exhibits both natural and artificial reverberation effects, and is processed at a high sampling rate.

We compare the performances of the density-based multi-channel blind deconvolution (MBD) algorithm in (16.28) as well as the density-based convolutive BSS (ConvBSS) algorithm in (16.42), in which both updates have been implemented in block form using FFT-based fast convolution methods. Each block-based update calculates the averaged quantites denoted by $E\{\cdot\}$ using L-sample sums, so that all filter coefficients are updated every

L time instants. Since impulsive signals are being extracted, we have chosen $f(y) = \text{sgn}(y)$ corresponding to the Laplacian p.d.f. source model in (16.15). The diagonal entries of $\mathbf{M}(k)$ are chosen as

$$\mu_i(k) = \frac{\mu_0}{\beta + \sum_{p=L+1}^{2L} y_i(k-p) f(y_i(k-p))} \tag{16.48}$$

where the constant $\beta > 0$ is used to avoid a divide-by-zero condition. This "quasi-normalized" step size strategy makes the updates scale-independent and generally improves algorithm robustness. For the speech-speech separation task, the parameter values chosen for both algorithms were, $L = 4096$, $\mu_0 = 10/L$, and $\beta_0 = 0.01$, and ten passes through the recording were used to adjust the system coefficients. For the *a cappella* duet recording, the parameter settings were $L = 4096$, $\mu_0 = 1.5/L$, and $\beta = 0.01$, and the systems were trained using two complete passes through the recording.

To evaluate the performance of each algorithm, we have employed a data-only strategy based on identifiable signal content. First, temporal portions of the recordings are identified that only contain a single source. Then, the variances of these portions are computed for each channel, and a signal-to-interference (SIR) ratio is formed by taking the ratio of the signal variances across the same time interval for each channel. For recordings that contain background noise, the variance of the background noise for each channel is estimated from portions of the recording that contain no other signals, and these noise variances are subtracted from the signal-plus-noise variances calculated previously prior to forming the SIR ratios. In addition, we compute the power spectral densities (PSDs) of the original and extracted signals to gauge the temporal effects that each algorithm imposes on the extracted signals.

Table 16.1 lists the SIRs and signal-to-noise ratios (SNRs) for the original and separated signals for the speech-speech separation task. The average SIR of the original signal mixtures is 3.6 dB, and both algorithms are able to improve the level of separation by a moderate amount. In this case, the average SIR of the extracted sources is 10.6 dB for the MBD algorithm and 12.7 dB for the ConvBSS algorithm, indicating that the latter technique provides better separation performance in this example. Comparing the SNRs of the various signals, the MBD algorithm is less sensitive to noise in this example, although both algorithms maintain a noise level that is significantly below that of the residual interference in the extracted signals.

The spectral content of the extracted signals show another distinct advantage of the ConvBSS algorithm over the MBD algorithm, as shown in Figure 16.2. The MBD algorithm nearly equalizes the PSDs of the original signal mixtures in the extracted sources, whereas the ConvBSS algorithm maintains similar spectral content as that in the original mixtures. Comparing the audio quality of the extracted outputs, the ConvBSS method produces separated

Table 16.1. Signal-to-interference and signal-to-noise ratios for the mixed and separated signals in the speech-speech separation example.

	Original Mixture	Separated Outputs MBD Algorithm	Separated Outputs ConvBSS Algorithm
SIR, Left Ch.	5.1 dB	14.8 dB (+9.7 dB)	13.5 dB (+8.4 dB)
SIR, Right Ch.	2.0 dB	6.3 dB (+4.3 dB)	11.9 dB (+9.9 dB)
SNR, Left Ch.	21.3 dB	21.1 dB (-0.2 dB)	17.9 dB (-3.4 dB)
SNR, Right Ch.	18.2 dB	16.9 dB (-1.3 dB)	13.8 dB (-4.4 dB)

signals that sound more natural and listenable than those produced by the MBD algorithm. This feature alone makes the ConvBSS algorithm preferable over multi-channel blind deconvolution approaches to convolutive BSS, as it is challenging to undo the undesirable temporal effects of the latter approaches in a truly blind setting. Note that both algorithms have similar complexities and do not require any signal preprocessing or post-processing.

The performance of the algorithms in the *a cappella* duet example are similar. The average SIR of the original mixture is 3.5 dB, and the average SIRs of the extracted signals is 9.8 dB for the MBD algorithm and 11.0 dB for the ConvBSS algorithm, respectively. The spectrograms in Figure 16.3 illustrate the temporal effects of the algorithms. Figure 16.3(a) and (b) show the spectrograms of the original left and right channels before processing over the interval $113 \leq t \leq 114.2$ sec. In these segments, the male singer is singing "I'm gon-na tell her now that" at different pitches, whereas the female singer is holding the "short-i" vowel sound of the word "him" at a single pitch. The female singer's voice clearly provides the dominant features in both spectrograms. Figure 16.3(c) and (d) show the spectrograms of the outputs of the MBD algorithm over the same time interval. As can be seen, the male's voice is enhanced in the left output channel, whereas the female's voice is enhanced in the right output channel. The signal spectra, however, are much flatter than those of the original mixture signals. Shown in Figure 16.3(e) and (f) are the spectrograms of the ConvBSS algorithm outputs, in which the original input spectra are largely preserved. The resulting extracted outputs sound much more natural in comparison to those produced by the MBD algorithm due to their spectral character.

16.6 Conclusions and Open Issues

In this chapter, an overview of blind signal separation for acoustic signal mixtures has been presented. Both theoretical and practical aspects have been

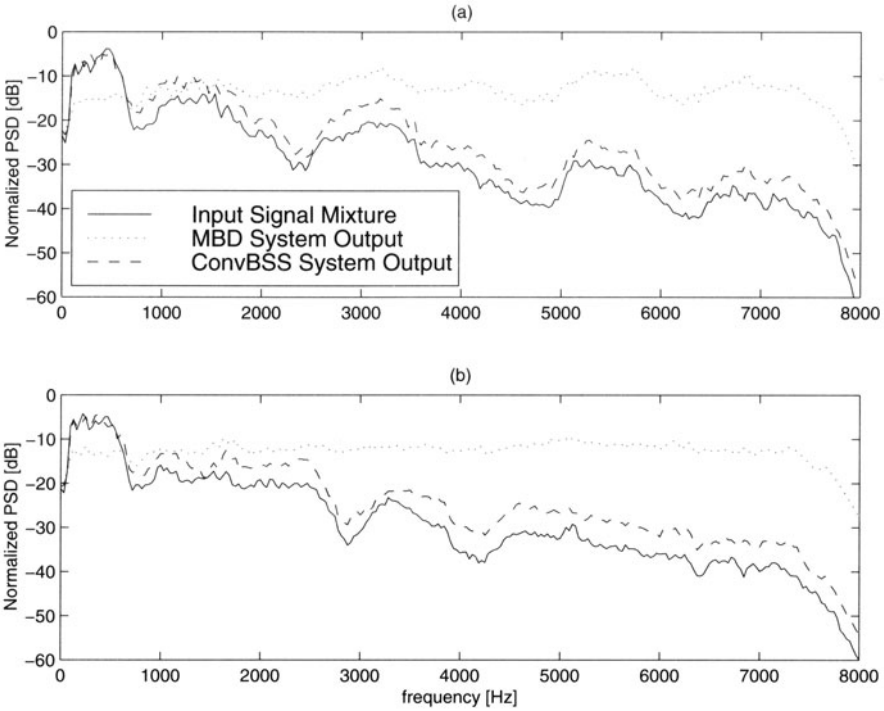


Fig. 16.2. Normalized power spectral densities (PSDs) for the speech-speech separation example: (a) left channel, (b) right channel.

discussed so that one can gain an intuitive understanding of the capabilities and limitations of such methods. Algorithms that are well-matched to the separation of acoustic mixtures have been given. Numerical evaluations indicate that the procedures provide a moderate level of separation quality when applied to real-world data sets.

While state-of-the-art acoustic signal separation methods have achieved some signficant milestones, numerous issues and challenges remain:

- *Time-varying acoustical environments.* In most practical situations, the acoustical channel is ever-changing due to movements of sources and/or objects, changes in air temperature, and the like. What are the typical tracking capabilities of blind signal separation methods? The design of separation algorithms with good tracking capabilities is at its infancy.
- *Changes in source number.* All of the methods in this chapter assume that the number of sources m is known. How can one translate these procedures to more realistic scenarios in which the number of sources is unknown or changes with time? Such is the case in most speech separation tasks due to the intermittency of human conversations.

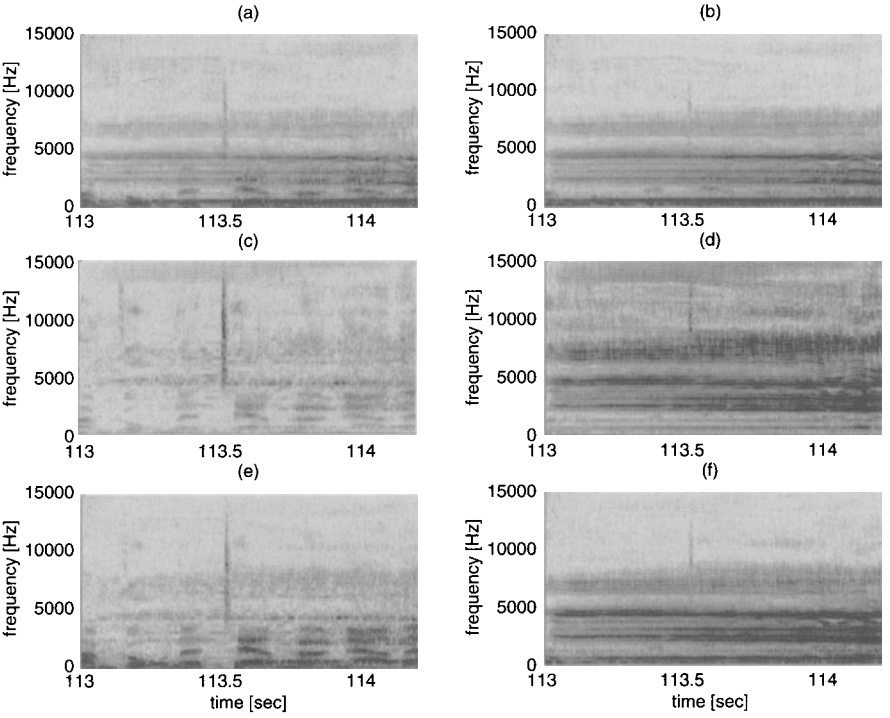


Fig. 16.3. Spectrograms for the *a cappella* duet separation example: (a) original left channel, (b) original right channel, (c) left channel output of MBD algorithm, (d) right channel output of MBD algorithm, (e) left channel output of ConvBSS algorithm, (f) right channel output of ConvBSS algorithm.

- *Robustness issues.* While many signal separation criteria and procedures correctly identify extracted source signals under ideal conditions (e.g., no sensor noise, known source statistics, and well-chosen structures and parameters), these criteria and methods might fail to separate under more-realistic situations. How robust is a given blind signal separation criterion or algorithm to these real-world variations? Clearly, one would like to design criteria and methods that are robust to the non-ideal conditions of practical acoustic settings.
- *"Separating" mixtures containing more sources than sensors.* When the number of sensors is fewer than the number of sources, signal separation as described in this chapter is no longer well-defined. How does one design criteria and algorithms for blindly extracting independent signals in this case? Solutions to this problem may prove more useful than those for traditional signal separation due to the number of practical applications that fit this mixing model.
- *Statistical efficiency of separation methods.* All separation methods employ data-dependent processing. For a given data record length, to what

fundamental accuracy can blind signal separation be performed, and how does the statistical efficiency of a given approach compare to this fundamental limit? Answers to such issues shall determine how well these methods perform in future applications given the ever increasing capabilities of computational hardware.

Despite these challenges, blind signal separation continues to be a source of innovative procedures, criteria, and concepts. The references that follow provide good starting points for future research efforts.

References

1. C. Cherry, "Some experiments on the recognition of speech with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–981, 1953.
2. W. Yost, *Fundamentals of Hearing*, 3rd ed., Academic, 1994.
3. S. Van Gerven and D. Van Compernolle, "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness," *IEEE. Trans. Signal Processing*, vol. 43, pp. 1602–1612, July 1995.
4. E. Weinstein, M. Feder, and A. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE. Trans. Speech Audio Processing*, vol. 1, pp. 405–413, Oct. 1993.
5. P. Comon, "Independent component analysis: A new concept?", *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
6. C. Jutten and J. Herault, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.
7. S. Haykin, ed., *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, Wiley, 2000.
8. J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 90, pp. 2009–2026, 1998.
9. S. Douglas, "Blind Signal Separation and Blind Deconvolution," in *Handbook of Neural Networks for Signal Processing*, (J.-N. Hwang and Y.-H. Hu, eds.), CRC Press, 2001.
10. A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.
11. S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach," in *Proc. 11th IFAC Symp. Syst. Ident.*, Kitakyushu City, Japan, pp. 1057–1062, July 1997.
12. R. Lambert and A. Bell, "Blind separation of multiple speakers in a multipath environment," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 423–426, Apr. 1997.
13. T.-W. Lee *et al.*, "Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 1089–1092, May 1998.
14. Y. Inouye and T. Sato, "Iterative algorithms based on multistage criteria for multichannel blind deconvolution," *IEEE Trans. Signal Processing*, vol. 47, no. 6, pp. 1759-1764, 1999.

15. J. Tugnait, "Identification and deconvolution of multichannel linear non-Gaussian processes using higher-order statistics and inverse filter criteria," *IEEE Trans. Signal Processing*, vol. 45, pp. 658–672, Mar. 1997.
16. S. Haykin, ed., *Blind Deconvolution*, Wiley, 1994.
17. S. Haykin, ed., *Unsupervised Adaptive Filtering, Vol. II: Blind Deconvolution*, Wiley, 2000.
18. T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.
19. S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Proc. Sys. Adv. Neural Inform.*, pp. 757–763, MIT Press, 1996.
20. S. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, Nov. 1997.
21. W. Davenport, Jr., "A study of speech probability distributions," Tech. Rep. 148, MIT Research Laboratory of Electronics, Cambridge, MA, 1950.
22. S. Douglas, A. Cichocki, and S. Amari, "Multichannel blind separation and deconvolution of sources with arbitrary distributions," in *Proc. IEEE Int. Workshop Neural Networks Signal Processing*, Amelia Island FL, USA, pp. 436–445, Sept. 1997.
23. D.-T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, pp. 1712–1725, July 1997.
24. D.-T. Pham, "Mutual information approach to blind separation of stationary sources," in *Proc. Workshop Indep. Compon. Anal. Signal Sep.*, Aussois, France, pp. 215–220, Jan. 1999.
25. C. Simon *et al.*, "Separation of a class of convolutive mixtures: a contrast function approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Phoenix AZ, USA, pp. 1429–1432, Apr. 1997.
26. J. Tugnait, "On blind separation of convolutive mixtures of independent linear signals in unknown additive noise," *IEEE Trans. Signal Processing*, vol. 46, pp. 3117–3123, Nov. 1998.
27. L. Molgedey and H. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, pp. 3634–3637, June 1994.
28. L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320–327, May 2000.
29. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Mag.*, vol. 9, pp. 14–37, Jan. 1992.
30. K. Torkkola, "Blind Separation of Delayed and Convolved Sources," in *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, (S. Haykin, ed.), pp. 321–375, Wiley, 2000.
31. S. Douglas and S. Amari, "Natural Gradient Adaptation," in *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, (S. Haykin, ed.), pp. 13–61, Wiley, 2000.
32. S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. Signal Processing Adv. Wireless Commun.*, Paris, France, pp. 101–104, Apr. 1997.
33. S. Amari, T.-P. Chen, and A. Cichocki, "Nonholonomic orthogonal learning algorithms for blind source separation," *Neural Comput.*, vol. 12, pp. 1463-1484, 2000.

34. P. Regalia and P. Loubaton, "Rational subspace estimation using adaptive lossless filters," *IEEE Trans. Signal Processing*, vol. 40, pp. 2392–2405, Oct. 1992.
35. S. Douglas, S. Amari, and S.-Y. Kung, "Adaptive paraunitary filter banks for spatio-temporal principal and minor subspace analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 1089–1092, Mar. 1999.
36. X. Sun and S. Douglas, "Multichannel blind deconvolution of arbitrary signals: Adaptive algorithms and stability analyses," in *Proc. 34th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove CA, USA, pp. 1412–1416, Oct. 2000.
37. The Bobs, "Boy around the corner," from *Songs For Tomorrow Morning,* [audio recording], Rhino Records, 1988.

Part IV

Open Problems and Future Directions

17 Future Directions for Microphone Arrays

Gary W. Elko

Media Signal Processing Research, Agere Systems, Murray Hill NJ, USA

17.1 Introduction

The telephone, invented more than 140 years ago, has changed very little with respect to acoustic design in the last 40 years. The most recent technological improvements were the replacement of the carbon microphone with an electret microphone and the development of the speakerphone for hands-free telephony in the early sixties. It might seem quite surprising that in spite of the rapid changes in communication systems: digital switching, digital wireless, packet-based data communications and the like, that very little has changed to the acoustic front-end in telephony systems. For instance, with the general availability of broadband communication networks, why have we not significantly increased the audio communication bandwidth? Why don't we see stereo or multichannel in telephony and communication systems? Why don't we see more ubiquitous adoption of voice recognition in consumer products? Why can't we tailor the audio response of our communication systems the same way we do with our entertainment systems? Why do our cell phones sound so tinny, distorted, and noisy? Why is it so difficult to understand people who call us from noisy environments or call us in noisy environments? These are good questions when one ponders the explosion of inexpensive digital signal processing integrated circuits combined with the current bandwidth of communication systems and those that are about to be introduced. Clearly front-end acoustic signal processing will play a large role in addressing these questions.

In our increasingly mobile society we find that hands-free operation of communication systems is becoming more and more the norm, whether we talk to other humans or to machines. There are significant challenges in enabling high quality speech communication in difficult acoustic environments. Microphone arrays are an enabling technology for hands-free communication since these systems offer directional gain to improve the signal-to-noise ratio. This chapter will address some new possibilities as well as outstanding research issues that remain.

17.2 Hands-Free Communication

The hands-free speakerphone was first released by AT&T's Western Electric in 1954. The speakerphone allowed many users gathered in one location to

Microphone Distance (m)	Gain required for 2 cm SNR (dB)	Minimum number of omnidirectional microphones
0.02	0	1
0.1	14	5
1.0	34	50
2.0	40	100

Table 17.1. Table showing minimum number of microphones in an optimal array to attain the equivalent 2 cm SNR (Signal-to-Noise- Ratio) for various distances from a point source in an isotropic noise field.

participate in a telephone conference. Recently the speakerphone has become a standard option on consumer phones. Speakerphones brought two major problems that did not exist in standard handset telephony; acoustic echo as well as, high sensitivity to room reverberation and background noise. The problem that we are concerned with in this book is that of using microphone arrays to minimize noise and reverberation. A simple example shows how this problem is actually an extremely difficult one to solve.

To begin, let us assume that a telephone handset microphone location is 2 cm from a talkers mouth in an acoustic environment that has a diffuse noise field. For simplicity, we assume the speech source is a point source. Now, let us look at the increase in directional gain that would be required for a directional microphone array to maintain the same signal-to-noise ratio as the handset for different distances. The results of the required gain are shown in Table 17.1. The minimum number of omnidirectional microphone elements required to realize this gain are also given in Table 17.1. This minimum number can be calculated by using the maximum gain attainable for an array in an isotropic noise field ($20 \log N$, where N is the number of elements [1]). Note that we have also assumed that the directional array is pointed at the desired source. As can be seen in Table 17.1, the number of microphones required to attain the same SNR as a handset omnidirectional microphone placed at 2 cm from the talker can be very large. The problem is actually even worse. We have assumed that the beamforming gain of the microphone array is optimum. For an isotropic noise field this gain is practically impossible to realize for arrays greater than three elements. For optimum gain the array would have to be differential at low frequencies and the white noise gain would be such that microphone self-noise would dominate the microphone output [1]. For nondiffuse sound fields, higher array gains can be realized, but these fields are specialized cases that are not typically found in a real acoustic background.

Fortunately, humans have the ability to understand speech signals in conditions where the SNR is as small as 0 dB. Thus, humans can deal with speech signals that are not as clean as handset speech. It is well known that SNR increases as small as a few dB can be critical in low SNR cases [2], and thus, any directional gain provided by a microphone array can be a benefit.

Hence, microphone array beamforming can be an effective tool in improving the intelligibility and quality of acoustic communication signals.

17.3 The "Future" of Microphone Array Processing

Humans naturally communicate using speech. It is therefore evident that the future of man-machine communication will focus on speech as a main interface. Also, as devices become smaller and more personal, a speech interface becomes essential. It is well known that present speech recognition algorithms are not very robust to hands-free speech. The significant loss in recognition performance is due to acoustic reverberation and SNR loss that are typical in hands-free applications. It is clear from this observation that we need to look at microphone array designs to also increase automatic speech recognition in the hands-free mode. The main design goal in present microphone array designs is to maximize the SNR. It is not clear, however, that this metric is optimal for speech recognition algorithms. Speech recognition algorithms work on input data that is condensed into feature vectors. Typically these vectors are based on LPC (Linear Predictive Coding) and cepstral analysis of speech. Obvious attempts to increase the accuracy of hands-free speech recognition optimize the beamformer design to minimize the perturbation of the measured feature vectors in reverberation and noise. However, it is not likely that this is equivalent to maximizing the input SNR.

Since microphone arrays have the ability to generate multiple outputs of spatially filtered input signals, it would seem apparent that speech recognition algorithms in the future will incorporate the idea of multiple acoustic inputs, even to the point that the speech recognizer is part of the beamforming processing. One obvious application of this idea would be to have the beamformer give many outputs: one for the desired signal plus noise, and the others as signals that are essentially representative of the spatial noise field. The recognizer would then have an estimate of the background noise field to utilize for robust speech end-point detection and to update the word models and statistics. Similar synergistic possibilities exist in the design of speech and audio coders for operation in noisy environments. Finally, the use of a spatially segmented acoustic field by beamforming could increase the cancellation depth and bandwidth of active noise cancellation hearing systems.

Another area of interest is the design of nearfield microphone arrays. Close-talking microphones have been in use now for more than fifty years. These microphones utilize a beamformer that exploits the nearfield of a desired acoustic source. One of the inherent problems in close-talking microphones is the sensitivity of the microphone to distance and orientation to the source. Microphone systems that account for position variations of close-talking microphones and modify the output signal to compensate for this variation would improve hand-held terminal acoustic performance [4]. Also,

systems that can automatically vary their mode of operation (nearfield or farfield) depending on the estimated position of the microphone relative to the desired source will certainly find favor in the future.

The hands-free communication problem does not solely concern the microphone. There is also the problem of acoustic coupling between the loudspeaker playing the far-end signal and the microphone used to transmit the local talker signal. The common solution to this problem is to identify the acoustic coupling path and subtract out the modeled echo from the near-end microphone signal. This operation is typically referred to as acoustic echo cancellation. It is obvious that time variation of the beamformer to minimize noise or follow the near-end talker could significantly change the coupling path between the loudspeaker to the microphone. Clearly a good design would be to do both beamforming and echo cancellation in one combined operation. Possible approaches to this problem have been identified in [5], and are reviewed in Chapter 13.

As stated earlier, past efforts in microphone array design have utilized standard engineering optimization criteria to derive the "optimum" array response. Although this approach has led to reasonable solutions, the ultimate judgment of the microphone array performance is what a human listener would perceive. Very little work has been published on designing arrays that utilize some perceptual metric in the design procedure. One exception to this is the idea of "intelligibility weighted gain" [6]. Although this work is a reasonable starting point, there is much fertile ground for research in this area.

As people become more connected via communication systems that allow "anytime, anywhere" communication it seems natural that microphone array systems will become very personal—to the point of a wearable acoustic communication devices. These devices will be extremely lightweight and small and have been euphemistically referred to as "acoustic jewelry". There are many new challenges that come with the design of wearable acoustic devices. One obvious challenge is the design of an array that will optimize the pickup of the host's speech as well as that of the desired surrounding signals. Clearly we would like to place the host microphone as close as possible to the talker's mouth, without encumbering the wearer. Thus the design will probably position the microphone array somewhere in the nearfield-to-farfield transition region. Another opportunity presented in wearable microphone systems is to use bone and skin conducted vibration to improve the wearer's speech signal. For farfield signals (similar to the hearing aid scenario), we will have to resort to superdirectional beamformer designs.

Finally, another new area for microphone arrays is in transducer systems based on Micro-Electro-Mechanical Systems (MEMS). The possibility of constructing the microphone on the same piece of silicon that has the signal processing is an intriguing new development. One area where MEMS might be effective is in robustness to turbulent air flow over the microphone.

Since turbulent length scales are orders of magnitude lower than acoustic propagating wavelengths, there is the possibility of using adaptive frequency-wavenumber filtering to remove high wavenumber turbulent sensitivity from the microphone array. MEMS technology also allows one to easily build complex combinations of pressure, velocity, and acceleration signals. Combinations of these different types of sensors with integral signal processing opens up a whole new area for investigation.

17.4 Conclusions

Microphone array beamforming for human-to-human and human-to-machine communication is now a viable and cost effective solution. Microphone arrays will play an essential role in the solution to the difficult problem of hands-free communication. The biggest challenge that remains is that it will be essentially impossible to obtain an SNR that is as high as a microphone mounted close to the desired talker (a telephone handset for instance). Linear combinations of microphones can only offer a maximum gain of $20 \log N$, where N is the number of microphone elements, and typical realizable gains are much less than this limit. Thus one is led to conclude that we need to search for new parametric and nonlinear processing algorithms in hopes of obtaining directional gains that significantly exceed the gains for linear processing. This is the challenge that has to be met if we ever hope to attain handset quality speech in hands-free conditions. One possible solution to this quality gap is to resort to wearable acoustic communication devices that can be thought of as "acoustic jewelry".

References

1. G. W. Elko, "Superdirectional microphone arrays", in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, eds., Kluwer Academic Publishers, chapter 10, pp. 181-237, 2000.
2. K. D. Kryter, "Validation of articulation index", *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1698-1702, Nov. 1962.
3. W. Kellermann, "A self steering digital microphone array", in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, pp. 3581-3584, 1991.
4. G. W. Elko, R. A. Kubli, D. R. Morgan, and J. E. West, *Adjustable filter for differential microphones*, U.S. Patent US05303307, 1994.
5. W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays", in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, pp. 219-222, 1997.
6. P. M. Zurek, J. E. Greenberg, and P. M. Peterson, "Sensitivity to design parameters in an adaptive-beamforming hearing aid", in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, pp. 1129-1132, 1990.

18 Future Directions in Microphone Array Processing

Dirk Van Compernolle

Katholieke Universiteit Leuven, Leuven, Belgium

18.1 Lessons From the Past

Antenna array processing has had long-standing impact on phased array radars, sonars and radio astronomy for several decades. The gigantic antenna arrays that were constructed for deep space observation must stand out as some of the most impressive engineering achievements of any discipline. Success in these related fields of signal processing have without any doubt stimulated interest in microphone array processing. And these successes did not only generate interest, they did much more—they created high expectations. Another interest generating stimulus came from a very different field, i.e. the one of anatomy and physiology. Nature has endowed virtually all species with two ears. One good reason, of course, is that there is always a second as backup when one of the two fails. But at the same time we all know that our sense for orientation is helped considerably by the use of two ears instead of one and that it helps us understand each other in the midst of a noisy crowd.

After 20 years of active research, however, we cannot claim that microphone array processing has had the success many of us hoped for, and many will wonder when the great breakthrough in microphone array processing will finally come, if ever. Nevertheless, progress in computer technology has helped us in a big way. In the early days only analog schemes of limited signal processing complexity were possible. This was followed by early years of high cost DSP computations, where computational cost seemed to impede widespread use of the technology. Today we have affordable DSPs that allow us to implement all but the most complex schemes cheaply in digital signal processing technology in real-time. But this in itself was not enough. Apart from breaking through the computing bottleneck, our understanding of the problems at hand has significantly progressed, as witnessed in this book. Most of the results presented are from recent years and give new insight into both the potential and the limitations of microphone array processing. However, too often the same problems that were considered too hard ten or twenty years ago are still set apart for 'future research'. Admitted weaknesses to proposed solutions are similar to the ones that we have been struggling with for a long time. Generally speaking we may say that many proposed solutions add to our understanding but lack robustness in order to make a bright future for themselves.

So should we not ask ourselves if there is a fundamental issue with microphone array processing? And my answer is 'nothing is fundamentally wrong'. Microphone array processing has only proven to be quite a bit harder than other previously successful array processing applications. We have known the problems from the beginning, but have underestimated the impact of some of them in real-life situations.

The basic problems fall into a small number of categories: (i) the speech signal is broadband; (ii) in many practical situations the desired source is in a reverberant space in the near or mid field, is moving, and cannot be assumed to be a point source; and (iii) the speech signal changes rapidly, it is intermittent, and shares many characteristics with the competing and interfering signals.

It is very difficult to tackle all these issues at once. It is especially difficult to come up with tractable mathematical models for this complex environment. The result of this complex situation is that a lot of research effort has gone into, and continues to go into the search for optimal beamforming strategies that rely on extra assumptions and constraints. Sadly enough, this all too often leads to solutions that lack robustness when evaluated in a variety of real-life situations. It may be that a far field assumption is required, it may be that less reverberation would be sufficient, or it may be that a perfect predictive speech detector will bring the breakthrough. Surely these mathematical developments are relevant and give us a better understanding of broadband beamforming in general. Simultaneously we should admit to ourselves that robustness has been, and still is today, one of the main issues.

The drive to achieve (mathematically) optimal solutions is a natural underpinning of our science and engineering nature. But is microphone array processing not too complex to be solved with optimal approaches? Should we not expect real breakthroughs to come from so-called robust solutions that are clearly sub-optimal for any given circumstance, but applicable in a relatively wide range of situations? Also is it not obvious that there will not be a single solution, but that we need quite different solutions depending on the target application(s)? These observations go hand in hand with one of the major problems that has faced microphone arrays since their debut: size and cost. A large size always seemed a must from the requirement of uniform broadband beamforming. Some of the first microphone arrays, especially the one constructed in the auditorium at Bell Labs, were magically impressive by their shear size and number of microphones. They were great fun as a research project. Also they resulted in functional solutions. At the same time the price of such systems seems exorbitant. Later on we saw many arrays on the order of, say, 1 meter. Any such design is still only applicable to a very limited number of applications such as conference rooms. In the majority of potential applications such a bulky design has no place. No industry has screamed more for tiny and low cost solutions than the hearing aid industry. Here spacing of a few cm are the maximum and processing power is an or-

der of magnitude less than in desktop applications. In all these situations we should not be surprised about the small size and limited number of sensors (two) in human hearing. It is far from optimal, but it works.

18.2 A Future Focused on Applications

If we ask ourselves what will the future bring for microphone array processing, we must envisage a range of widely differing solutions of different sizes and costs.

In the sequel I analyze the potential of the most important market segments and go looking for killer applications. If commercialization has not yet started, the question is of course what hampered commercial introduction and when, if ever, will we see usage of microphone arrays in each of these application fields.

18.2.1 Automotive

If any 'killer application' exists for microphone arrays, then it should be voice input in the car. It has all the right ingredients. Mobile telephony and speech recognition scream for hands-free voice input in a noisy environment. Signal-to-noise ratios obtained by single microphones are just not sufficient. Thus microphone arrays seem the logical solution. There are extra features that should help. The speakers inside a car are not mobile and their position is reasonably constant from session to session. Ultimately, a market potential of tens of millions of units per year should be commercially convincing. All of this should be sufficient for successful uptake of microphone array technology, but is it?

Today, penetration of microphone arrays in cars is minimal, except for a few top brands that are not all that noisy by themselves, and therefore have the least need for it. The major concern of car equipment manufacturers and car manufacturers alike is cost: multiple microphones, multiple wires, extra DSP power required, etc. Every cent in every component counts when putting a car together and microphone arrays have been judged as too expensive. Also, at least for the foreseeable future one should envisage that most microphones are mounted into existing cars, further complicating the story for arrays.

Given this large cost concern I do not believe that large arrays spanning the entire car will ever be viable. On the contrary, the car is an ideal environment for a microphone array that behaves like a traditional directional microphone but with a slightly steerable beam. Such an embedded array can be mounted as any regular microphone by not so specialized technicians. I do believe that the development of better microphones for usage inside the car will be a point of focus for microphone developers in the coming years.

18.2.2 Desktop

Cost has been the stumbling block as well for desktop microphone arrays in conjunction with speech recognition in the PC environment. People just do not like it if you tell them that the accuracy of a $50 speech recognition software package will drastically improve if they buy a $150 microphone array to go with it. It is still unclear if we will ever overcome the cost hurdle in this case. It may just be a question of whether large enough volumes will ever be reached such that current prices can be lowered drastically.

Microphone arrays for the desktop have just started to appear on the market. The reviews so far are ambiguous. In *quiet* environments they work as well as any headset worn. So if you do not want to be physically hooked up to your computer, this is the way to go the reviews say. At the same time the reviews will warn you that the existing commercial array microphones do not work well in considerable noise, and that one should not move around. Current reviews unanimously advise a wireless headset if one needs to move around a lot.

It seems therefore that current commercial implementations only solve a small part of the problem. All of the designs rely primarily on fixed beamforming, most often with limited directionality adjustment. On top of this, some additional noise suppression may be used. The 'speech seeking' part seems to be insufficient in all of the produced arrays. Also the quality and speed of tracking is substandard. It just shows how great the robustness issue really is when bringing microphone array technology to consumer products.

All in all there is reason for optimism, however. Desktop arrays are very pragmatic in their designs. These microphones are built for applications that use a PC screen or monitor, and they sit perfectly well on top of a monitor or attach to the front of it. Overall size is limited to about 20 cm, all computing is done inside the array, and the array connects to other equipment just as any other microphone would do. We have come a very long way to bring prices down enough such that a single enclosure with multiple elements, A/D converters and a DSP can be made at prices competing with traditional high end microphones. And let us not forget that these are first generation devices and that volumes are still very small.

Given some more time, I believe that there is hope that microphone arrays will capture a part of this market. Who knows, 5 years from now microphone arrays may be standard equipment on laptop and desktop computers. There is also a chicken and egg situation here. A wider usage of speech recognition would put more pressure on hardware manufactures to include higher end microphones, including arrays. On the other hand, one of the main hurdles in improving performance and subsequent acceptance of speech recognition is the low quality audio input on most systems today.

18.2.3 Hearing Aids

Hearing aids form a market by themselves. Restrictions on size and computational power are an order of magnitude more stringent than in other areas, leading to substantially different designs. Array sizes of 5 to 20 cm have been used in experiments with hearing aids, but have overall been met with disapproval. Nevertheless, here we have also seen the introduction of a range of new multi-microphone based products in the last couple of years. Many of these products do not use classical arrays, but a combination of microphones with different characteristics, used as inputs to a noise suppression stage. Perhaps even more obvious than in the automotive or desktop case, the evolution is towards an adaptive speech seeking and noise suppressing microphone. The distinction here between microphone technology and array technology is not entirely clear (but that does not really matter).

18.2.4 Teleconferencing

Teleconferencing was for some time seem as one of the potential killer applications. But I think that this is no longer true. On the one hand, the expansion of the teleconferencing market seems to have come down to slow growth and we see nothing of the explosion that some had hoped for. Therefore, the hope for a massive market does not seem justified. Acoustic echo cancellation is the crucial issue and it can not be solved by array processing. When using arrays, as with any multi-microphone input, the problem becomes significantly worse. Special microphone designs, including radial arrays, have been constructed and will continue to play a role in this market. Large wall mounted microphone arrays, however, are unlikely to find their way into teleconferencing rooms in any big way.

18.2.5 Very Large Arrays

Teleconferencing was one of the potential markets for large arrays. Another one is the virtual microphone in large auditoria. However, this can not be considered a booming market either. Design and manufacturing of these arrays is costly and a large degree of optimization may be required from site to site, making the picture even worse. Hence large microphone arrays are doomed to remain a niche market. They will certainly survive in high profile demonstration projects, and as a research topic they will carry on for many years to come. Another (quite niche) market for very large arrays exists in the acoustic monitoring industry.

18.2.6 The Signal Subspace Approach - An Alternative to Spatial Filtering ?

Finally, we should ask ourselves the question if we should not look for alternative solutions to plain spatial filtering. We may think in two directions:

blind signal separation and signal subspace approaches. These techniques do not require sensitive geometric information about the array layout but work with any configuration.

These techniques should result in higher configuration robustness. But at the same time they are computationally very demanding and, while making fewer assumptions about the layout, they make in general more assumptions about the signals. Practical implementations have not appeared so far, but demonstration results are often impressive. So we should keep an eye open for these techniques. It is unlikely we will find them in products in the coming years, but in later generation array processing techniques, they may become the standard way to go.

18.3 Final Remarks

The near-term trend is in one direction: small arrays with few microphones and a high degree of robustness that behave as speech seeking, directional, and noise canceling microphones. Depending on the target application designs may vary from less than a 1 cm in diameter for the hearing aid market, over 5 cm for the car, to a maximum of 20 cm for desktop. After all, human hearing does very well with two ears spaced about 20 cm apart. These designs will not reach maximal noise suppression in any theoretical sense. Their goal is clear: a few dB gain in signal-to-noise ratio across the board at a cost which is only marginally above that of other microphones. A market of several million units for such medium cost devices is realistic and therefore economically viable. Economic potential for large arrays is much more limited and will therefore remain a niche market.

Index

acoustic echo cancellation, 272, 281, 308, 386
acoustic model, 335, 337
adaptation control, 286
adaptive beamformer, 88, 163, 235, 309, 338
affine projection algorithm, 285
AIC, 195
array calibration, 310
array gain, 21, 46, 293
array geometry, 340
articulation index theory, 232
artificial neural network, 334
automotive microphone array, 309, 391
autoregressive modeling, 160, 191

Bark spectral distortion, 148
Baum-Welch algorithm, 335, 345
beamforming, 3, 87, 288
beampattern, 6, 22
blind signal separation, 325, 355
blocking matrix, 34, 88, 291

cardioid microphone, 69, 257
causal separation model, 358
cepstral mean subtraction, 337
cepstral prefiltering, 162
close-talking microphone, 345, 355, 385
clustering, 195
cochlear implant, 230
cocktail party effect, 355
coefficient-constrained adaptive filters, 94
coherence, 22, 63, 143, 161, 244
coherence matrix, 22, 45, 47
coherence of car noise, 259
coherence of office noise, 259
coherence of speech, 259
coherence-based post-filter, 271
coherent signal subspace method, 183
computed tomography scanner, 256, 266
constant directivity beamformer, 4
constrained minimum variance beamformer, 310
constrained optimization, 309
contrast functions, 364
correlation matrix, 114, 160, 314
correlation-based criteria, 366
cross-correlation, 40, 166
cross-correlation matrix, 114
cross-power spectrum, 189, 309
cross-power spectrum phase, 190, 338
cumulant-based separation criteria, 366
cylindrically isotropic noise field, 30, 73

data association problem, 222
decentralized Kalman filter, 212
decorrelation system, 371
delay-and-sum beamformer, 26, 159, 169, 234, 290, 338
density modeling criteria, 362
desktop microphone array, 392
differential microphone, 67
diffuse noise field, 23, 309, 338
diffuse sound field, 49, 257
dipole microphone, 69
direction of arrival, 162, 182, 243, 290, 302
directional gain, 384
directivity index, 23, 44, 293
distributed multi-microphone, 348
disturbed harmonics model, 189
double-talk, 65
dual excitation speech model, 137
dynamic time warping, 334

EASI algorithm, 371
echo path gain, 293
echo return loss enhancement, 274, 283
echo shaping, 272
end-point detection, 333, 340
endfire array, 234, 275
entropy, 364
ESPRIT, 182
extended Kalman filter, 210
Eyring's formula, 118

face tracking, 204, 218
farfield, 25
farfield assumption, 3, 290
fast Newton algorithm, 286
feature extraction, 333, 340, 344
filter-and-sum beamformer, 5, 159, 171, 204, 290
filterbank, 45, 116, 184, 317
fixed beamformer, 234, 310
focused beamformer, 159
frequency-domain beamforming, 5
front-to-back ratio, 24, 30

generalized cross-correlation, 161, 183, 191
generalized sidelobe canceler, 35, 40, 88, 113, 127, 236, 291, 310, 338
generalized singular value decomposition, 115
gradient microphones, 35

hands-free speech acquisition, 181, 255, 307
harmonic nesting, 5, 340
hearing aid, 393
hearing impairment, 229
hidden Markov model, 334
high-resolution spectral estimation, 158
hypercardioid microphone, 36, 243, 259, 274

image method, 52, 118, 182, 342
in situ calibrated microphone array, 312
injected noise, 311
instantaneous blind signal separation, 360
intelligibility-weighted directivity, 232
intelligibility-weighted gain, 232
interacting multiple model, 222
interference suppression, 240
inverse Kalman filter, 213

Jacobian matrix, 211
joint diagonalization, 115, 367

Kalman filter, 206
Kullback-Leibler divergence measure, 362

language model, 332, 335
leaky adaptive filters, 92
leaky LMS, 312
linear intersection method, 195
linear prediction, 141, 185, 334, 371
LMS algorithm, 236
location-estimation null-steering algorithm, 243
Lombard effect, 332
loudspeaker-enclosure-microphone system, 283

magnitude squared coherence function, 257
matched filter, 134, 308, 310, 338
matrix inversion lemma, 43
maximum likelihood linear regression, 337
MDL, 195
mean square error, 114, 290
mel-scaled cepstral coefficients, 334
MEMS, 386
microphone channel response, 165
microphone directivity, 232, 257, 261, 357
microphone orientation, 257
microphone placement, 157, 261, 308
MIN-NORM, 182
minimum mutual information, 363
minimum variance spectral estimation, 160
MMSE beamformer, 313, 314
MMSE filter, 263
multi-channel blind deconvolution, 361
multi-resolution time-frequency adaptive beamformer, 319
MUSIC, 160, 182, 186
MVDR beamformer, 25, 39

narrowband assumption, 3
natural gradient adaptation, 368
nearfield, 25, 290, 385
nearfield superdirective array, 44
nested array, 290
NLMS algorithm, 266, 268, 284, 313
noise canceler, 34, 89
noise cancellation, 61, 256, 308
noise subspace, 182, 309
nonholonomic constraints, 370
nonlinear optimization, 367
normalized interference suppression, 322
normalized kurtosis, 365
normalized noise suppression, 322

omnidirectional microphone, 232, 257, 331
order permutation, 372

parallel model combination, 337
phase transform, 162, 167
phone, 335
phoneme recognition, 230
PLP coefficients, 334
post-filter, 39, 255, 338
power spectral density matrix, 21
pseudo-inversion, 368

QR-decomposition, 122

recursive state estimation, 205
reverberation distance, 258
reverberation time, 118, 164, 173, 283
RLS algorithm, 285, 313
robust estimation, 195
robustness, 88, 121, 336, 377
room impulse response, 162, 367
ROOT-MUSIC, 189

Sabine's equation, 258
scaled projection algorithm, 29, 242
scoring problem, 335
segmental SNR, 269
sensor calibration, 309
sensor fusion, 205
sensor mismatch, 235
short-time Fourier transform, 317
short-time modified coherence, 341
signal bias removal, 337
signal cancellation, 90, 338
signal subspace, 182
signal-to-interference ratio, 374
signal-to-noise plus interference power ratio, 316
single-channel speech enhancement, 136, 255
singular value decomposition, 189, 193
source activity detection, 267, 294
spatial aliasing, 4, 26
spatial clustering, 183
spatial correlation function, 61
spatial dithering, 312
speakerphone, 383
spectral subtraction, 40, 242, 308
speech corpora, 333, 341
speech intelligibility, 230
speech quality measures, 51
speech reception threshold, 229
speech recognition, 281, 292, 307, 331, 385
speech-to-interference ratio, 230
spherically isotropic noise field, 26, 65
state error information vector, 214
statistical pattern recognition, 333
steered response power, 158
steering vector, 169, 184
stochastic matching, 337
stochastic region contraction, 159
subband decomposition, 314
subspace tracking, 369
super-resolution, 310
superdirective array with Wiener post-filter, 44
superdirectivity, 19, 234, 290, 338
supergain, 19

teleconferencing, 281, 300, 355, 393
time delay estimation, 40, 161, 267, 338
time-difference of arrival, 158, 168, 183, 204
time-domain beamforming, 5
time-invariant beamforming, 290
time-varying beamforming, 291
Toeplitz matrix, 117, 191
training data contamination, 337
training problem, 335
transform-domain structure, 287

uncorrelated noise field, 24
universal noise subspace, 186
universal signal subspace, 186
universal spatial covariance matrix, 185
unvoiced speech, 137, 140

varechoic chamber, 342
videoconferencing, 163, 181, 203
Viterbi algorithm, 335
voice activity detector, 312, 343, 347
voiced speech, 137, 140

wavelet transform, 141
white noise gain, 24, 47
whitening filter, 190, 266
wideband weighted subspace fitting, 186
Wiener filter, 39, 41, 62, 114, 263, 291, 308, 338
Wiener solution, 35
Wiener-Hopf equation, 42, 284
word recognition rate, 346
WSF, 182