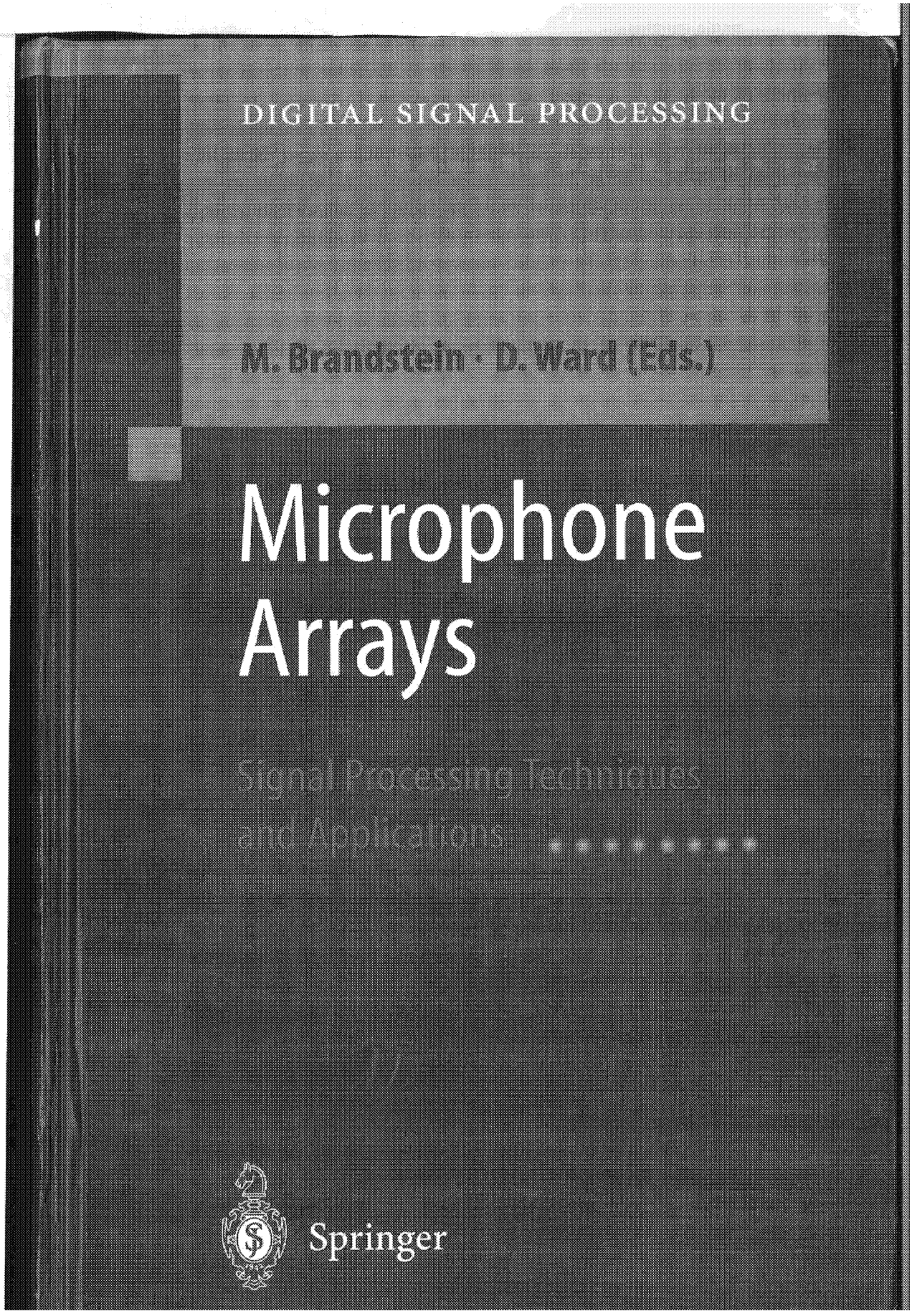


Amazon v. SoundClear
US Patent 9,031,259
Amazon EX-1011

RICHARD STERN

Springer
Berlin
Heidelberg
New York
Barcelona
Hongkong
London
Milan
Paris
Singapore
Tokyo



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 Sciences
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 3-540-41953-5 Springer-Verlag Berlin Heidelberg New York

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. Violations are liable for prosecution act under German Copyright Law.

Springer-Verlag is a part of Springer Science+Business Media
http://www.springer.de

© Springer-Verlag Berlin Heidelberg New York 2001

Printed in Germany

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: 11559276 62/3111 5 4 3 2 1 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

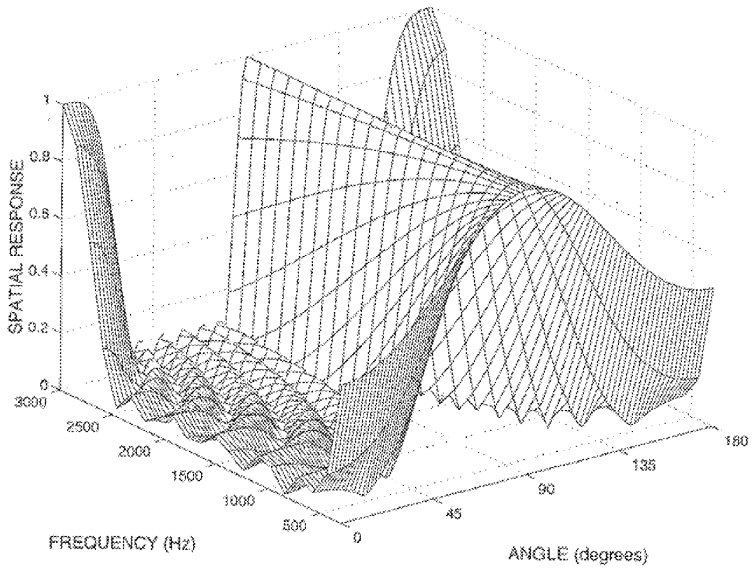


Fig. 1.1. Response of a narrowband array operated over a wide bandwidth.

(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.,

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

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:

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

where p_n is the location of the nth discrete sensor, and $\hat{b}_{\text{FI}}$ 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

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

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

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

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

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

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 $p_{ref} = \max|p_n|$ 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

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

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 $b_{ref}(u)$. Setting $f = \zeta/p_{ref}$, $B(\zeta)$ 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

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

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

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

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

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

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

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.

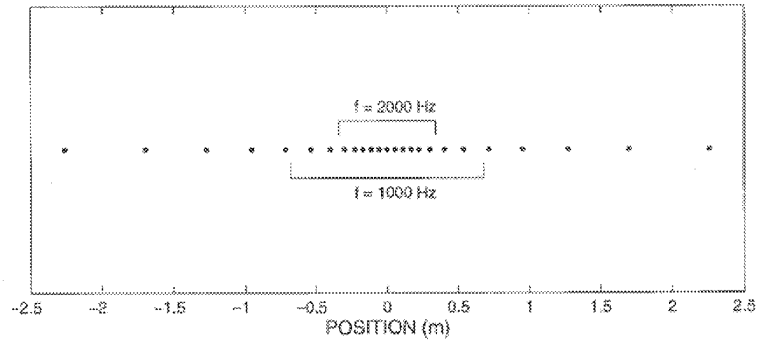


Fig. 1.2. Array geometry used for example CDB.

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-

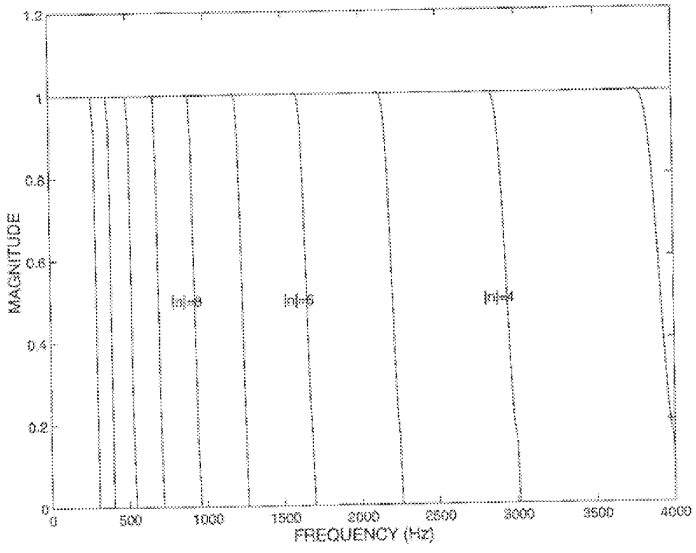


Fig. 1.3. Frequency responses of the weights on each sensor.

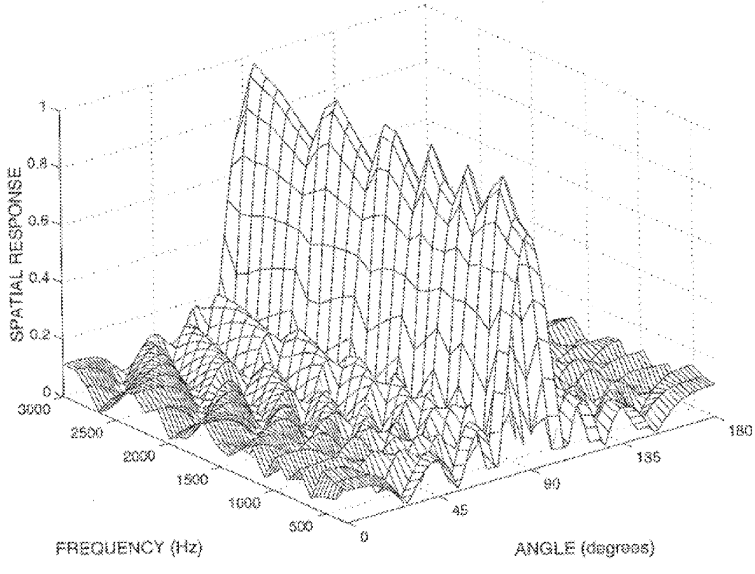


Fig. 1.4. Spatial response of example CDB.

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.

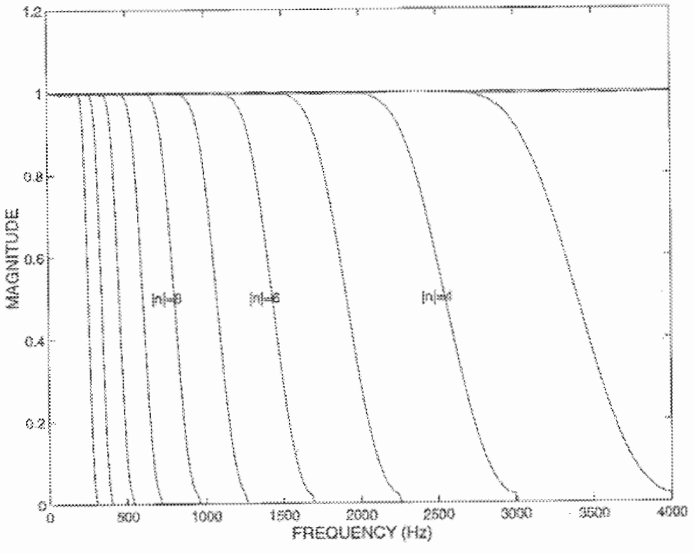


Fig. 1.5. Frequency responses of the weights on each sensor.

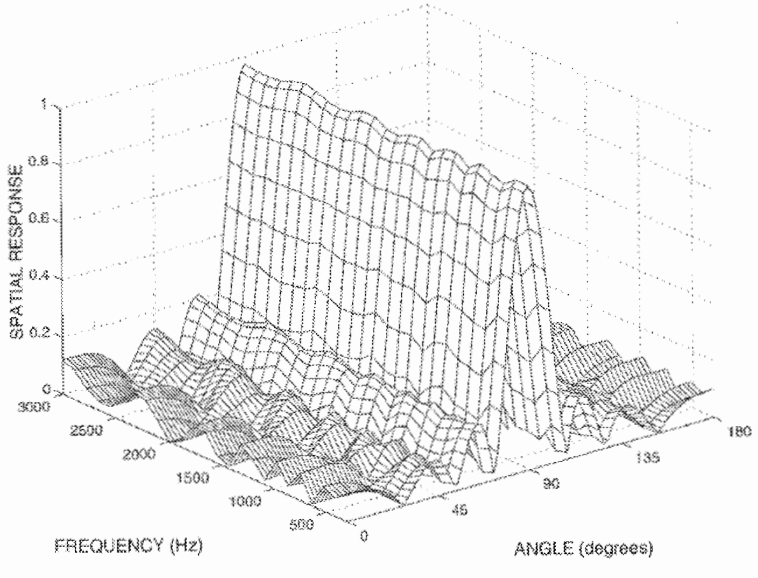


Fig. 1.6. Spatial response of example CDB.

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.

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.

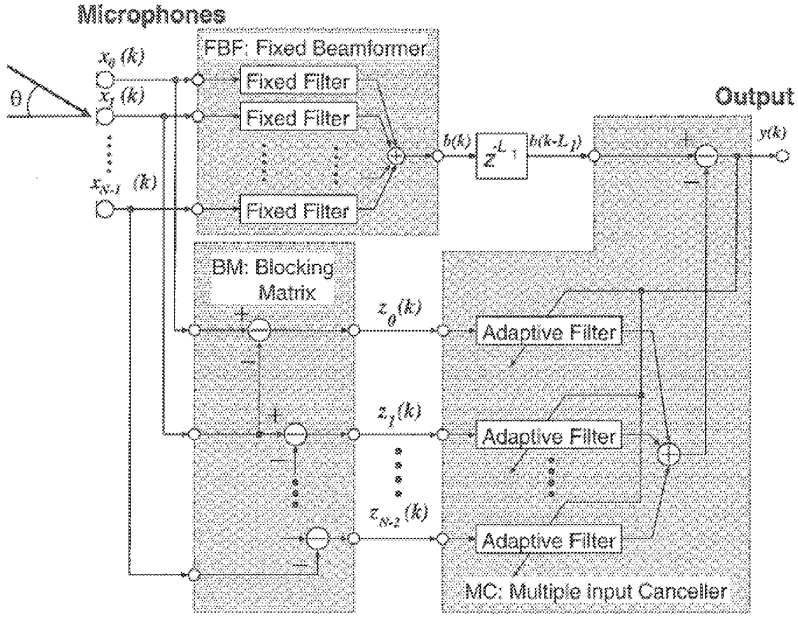


Fig. 5.1. Griffiths-Jim beamformer. It comprises a fixed beamformer (FBF), a multiple-input canceler (MC), and a blocking matrix (BM).

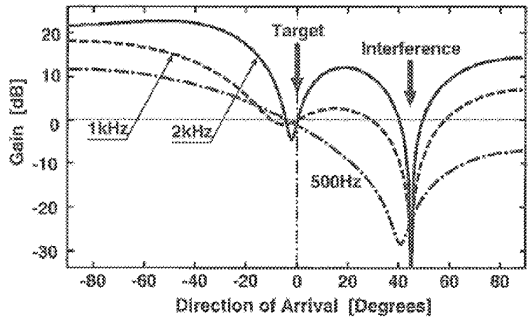


Fig. 5.2. Example directivity pattern of the Griffiths-Jim beamformer.

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

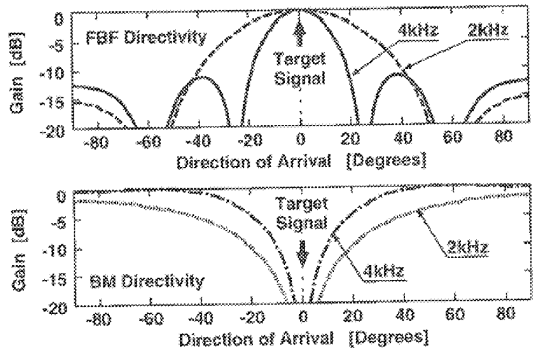


Fig. 5.3. Directivity pattern of a fixed beamformer (FBF) and a blocking matrix (BM).

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.

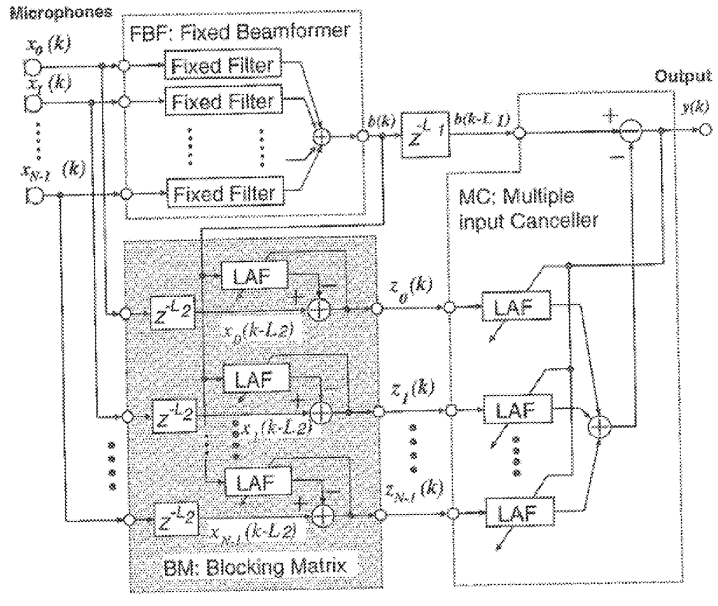


Fig. 5.4. GJBF with a LAF-LAF Structure.

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.

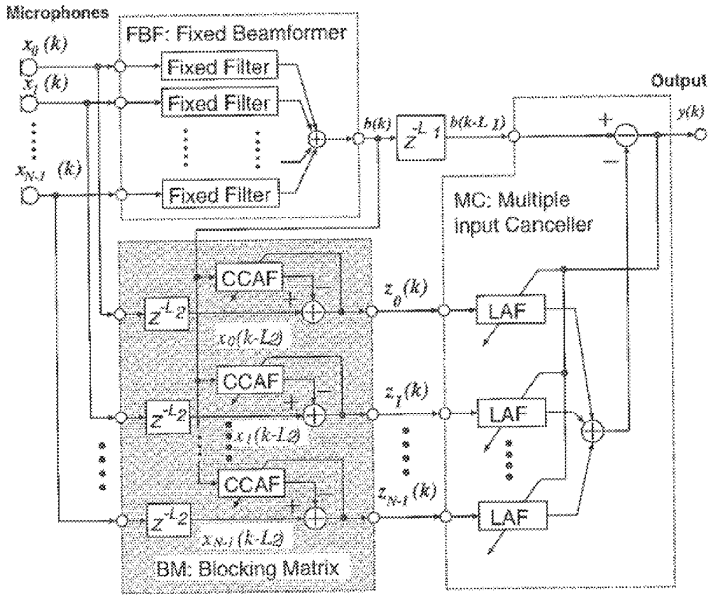


Fig. 5.5. GJBF with a CCAF-LAF Structure.

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

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

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

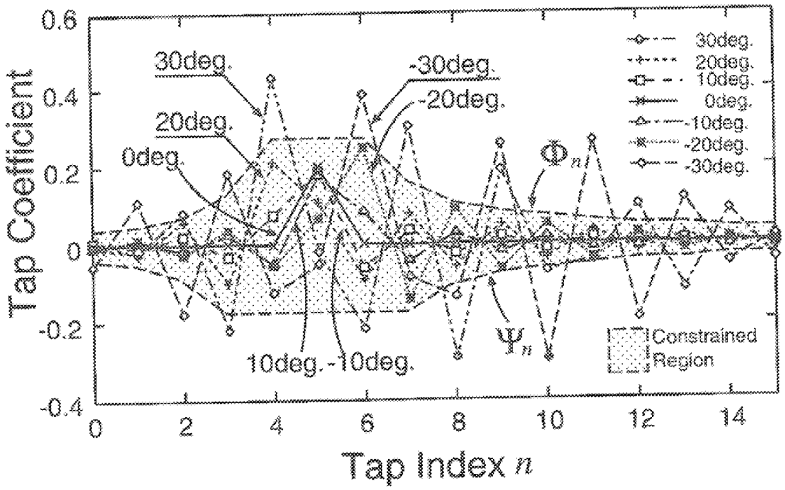


Fig. 5.6. An example of CCAF coefficients to minimize signals from different DOAs and their constraints. When the CCAF coefficients are constrained in the hatched region, up to 20° error in look direction could be allowed.

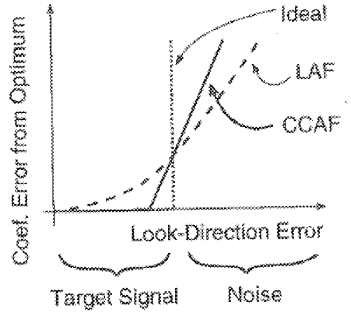


Fig. 5.7. Comparison of selectivity in LAF and CCAF.

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.

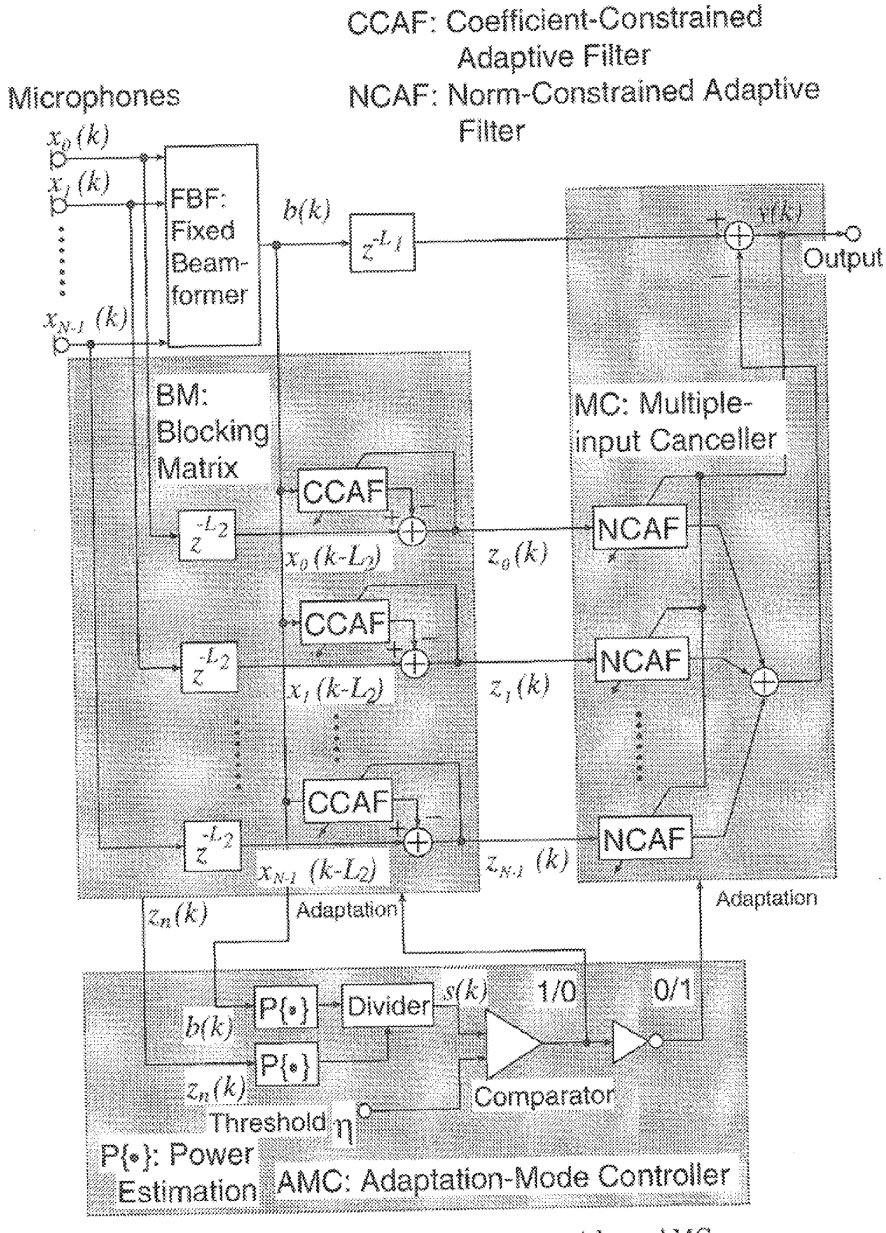


Fig. 5.8. CCAF-NCAF structure with an AMC.

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).

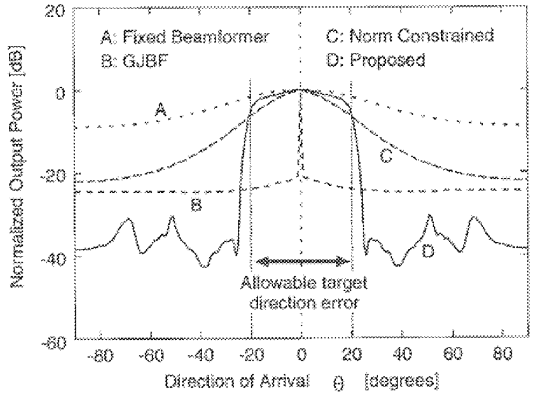


Fig. 5.9. Normalized output power after convergence as a function of DOA.

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.

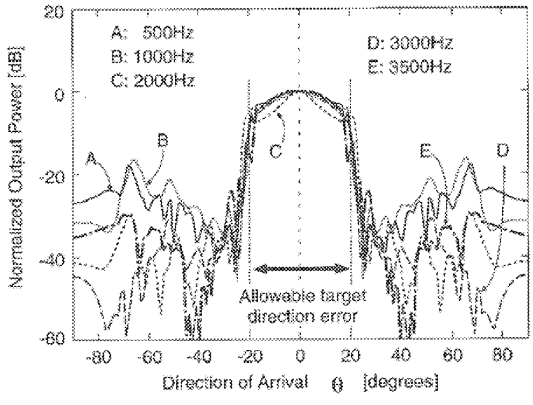


Fig. 5.10. Sensitivities after convergence as a function of DOA at different frequencies.

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.

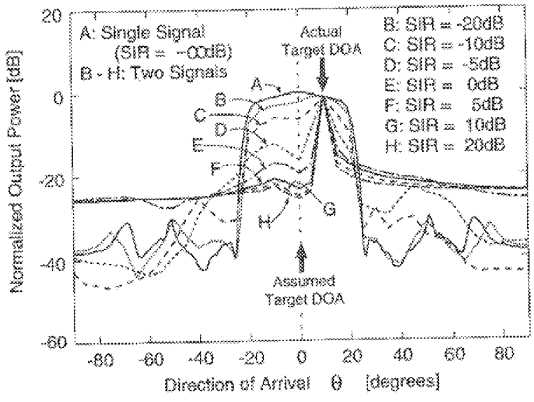


Fig. 5.11. Normalized output power after convergence as a function of DOA with different SIRs.

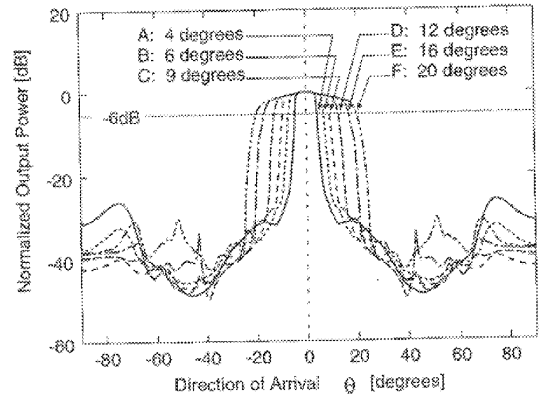


Fig. 5.12. Normalized output power after convergence for different allowable target directions.

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,

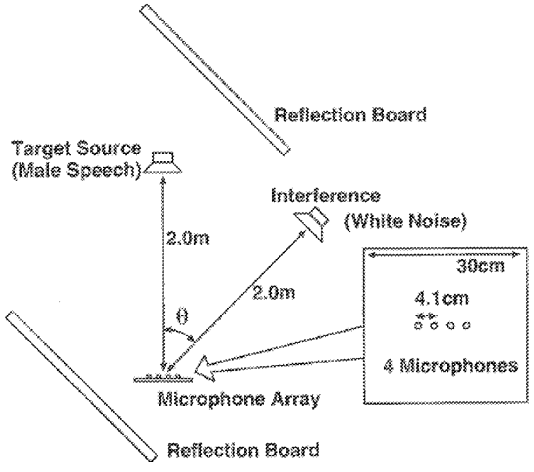


Fig. 5.13. Experimental set-up.

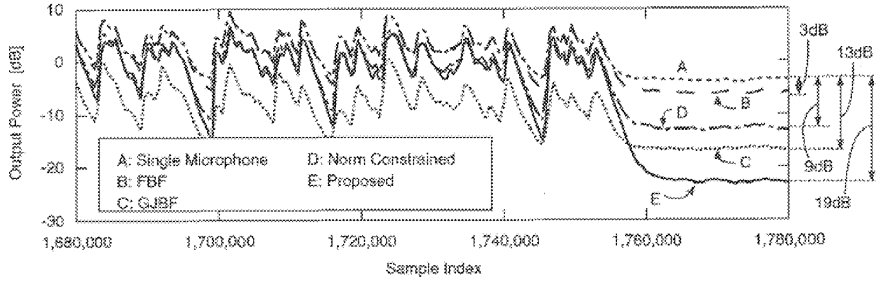


Fig. 5.14. Output Powers for a male speech signal and white noise.

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-

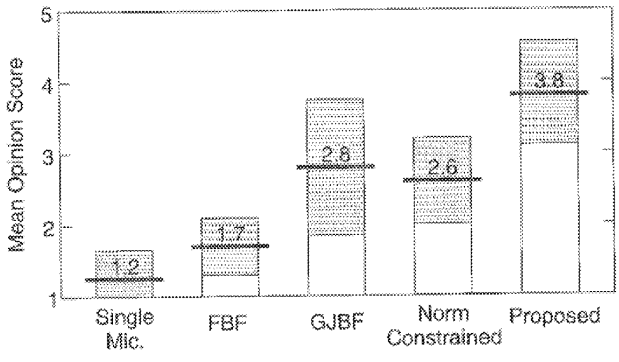


Fig. 5.15. Mean opinion score results.

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.

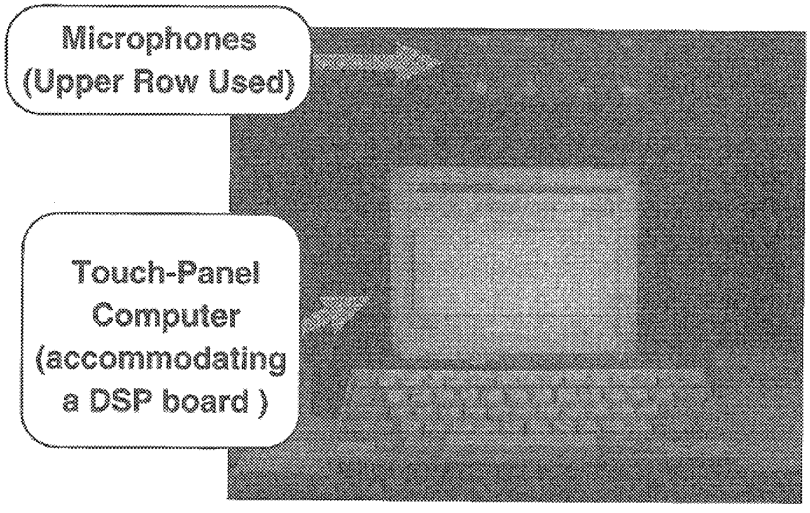


Fig. 5.16. Real-time DSP system.

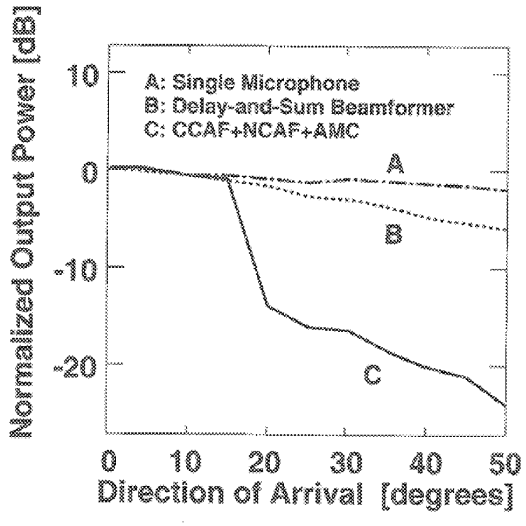


Fig. 5.17. Directivity patterns (i.e., the output powers normalized by the power at the center) measured in 5-degree intervals.

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.

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 n^{th} microphone

and the source, respectively. The received signal at the n^{th} microphone may now be expressed as

$$x_n(t) = s(t) \star h_n(q_s, t) + v_n(t) \tag{8.1}$$

The overall impulse response, $h_n(q_s, t)$, 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 p_n as well as the source location, 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:

$$x_n(t) = \frac{1}{r_n} s(t - \tau_n) \star g_n(q_s, t) + v_n(t) \tag{8.2}$$

where r_n is the source-microphone separation distance, τ_n is the direct path time delay, and $g_n(q_s, t)$ 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.

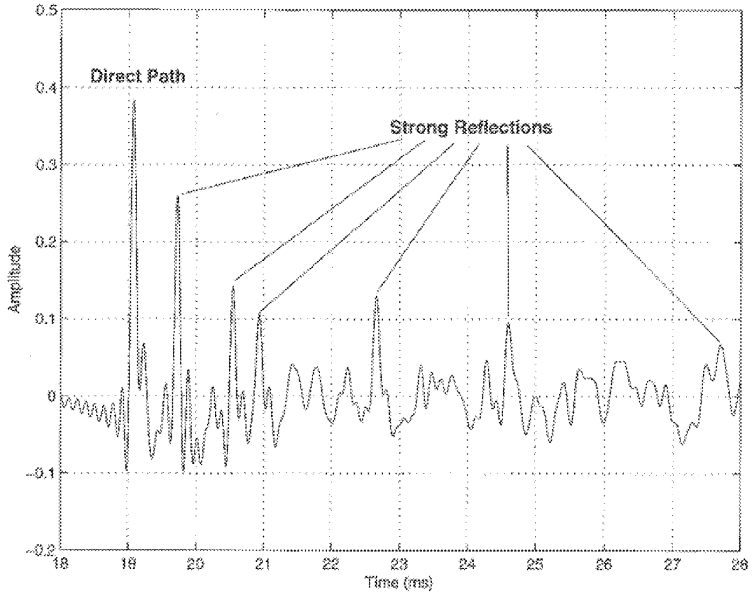


Fig. 8.1. A close-up of a 10-millisecond segment of a room impulse response measured in a typical conference room. The direct-path component and some strong reflected components are highlighted.

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

$$\begin{aligned} x_1(t) &= \frac{1}{\tau_1} s(t - \tau_1) \star g_1(q_s, t) + v_1(t) \\ x_2(t) &= \frac{1}{\tau_2} s(t - \tau_1 - \tau_{12}) \star g_2(q_s, t) + v_2(t). \end{aligned} \tag{8.4}$$

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 i^{th} 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, q_s, the true TDOA relative to the i^{th} sensor pair will be denoted by $T(\{p_{i1}, p_{i2}\}, q_s)$, and is calculated from the expression

$$T(\{p_{i1}, p_{i2}\}, q_s) = \frac{|q_s - p_{i2}| - |q_s - p_{i1}|}{c} \tag{8.10}$$

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, $\hat{\tau}_i \neq T(\{p_{i1}, p_{i2}\}, q_s)$.

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 $p_{i2} - p_{i1}$ 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 i^{th} 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, 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

$$E(q) = \sum_{i=1}^{M} (\hat{\tau}_i - T(\{p_{i1}, p_{i2}\}, q))^2. \tag{8.12}$$

The location estimate is then found from

$$\hat{q}_s = \underset{q}{\operatorname{argmin}}\, E(q). \tag{8.13}$$

The criterion in (8.12) will be referred to as the LS-TDOA error. As stated earlier, the evaluation of $\hat{q}_s$ 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

$$y(t, q_s) = \sum_{n=1}^{N} x_n(t + \Delta_n) \tag{8.14}$$

where Δ_n are the *steering delays* appropriate for focusing the array to the source spatial location, 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 $y(t, q_s)$ 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 n^{th} 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 l^{th} 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,

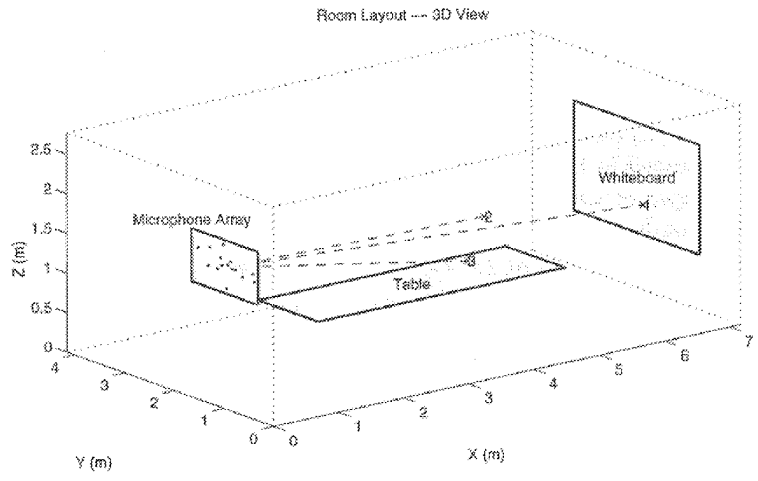


Fig. 8.2. Conference room layout.

$P(q)$ 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

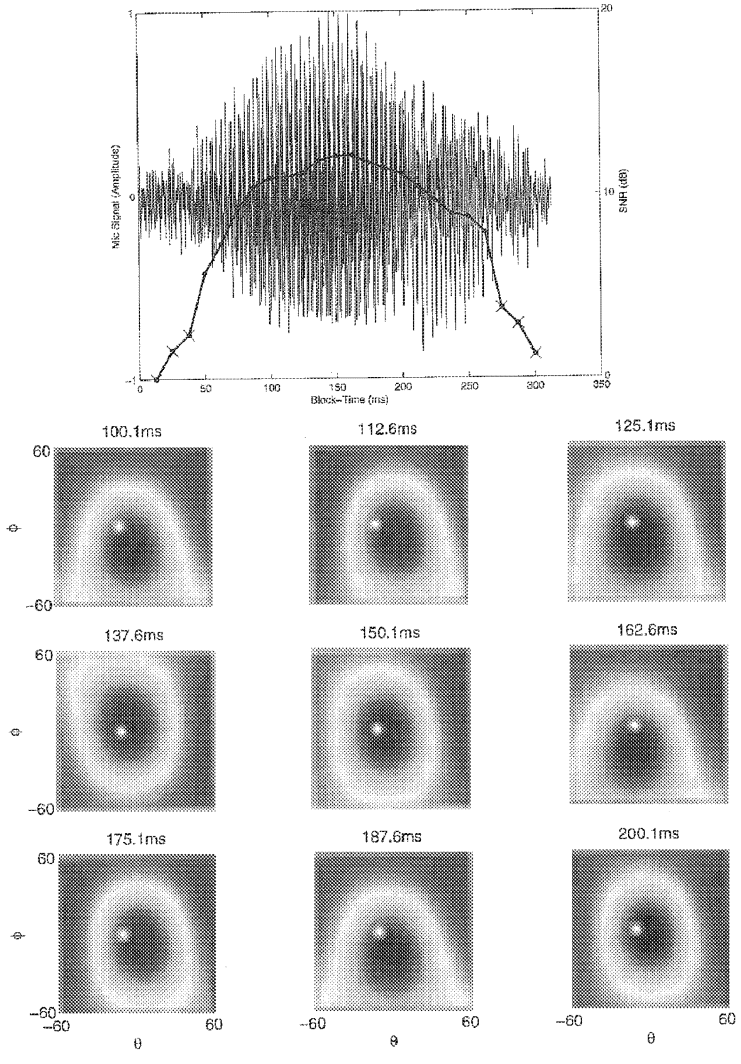


Fig. 8.3. Speech segment (top) with nine frames of the LS-TDOA error surfaces.

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

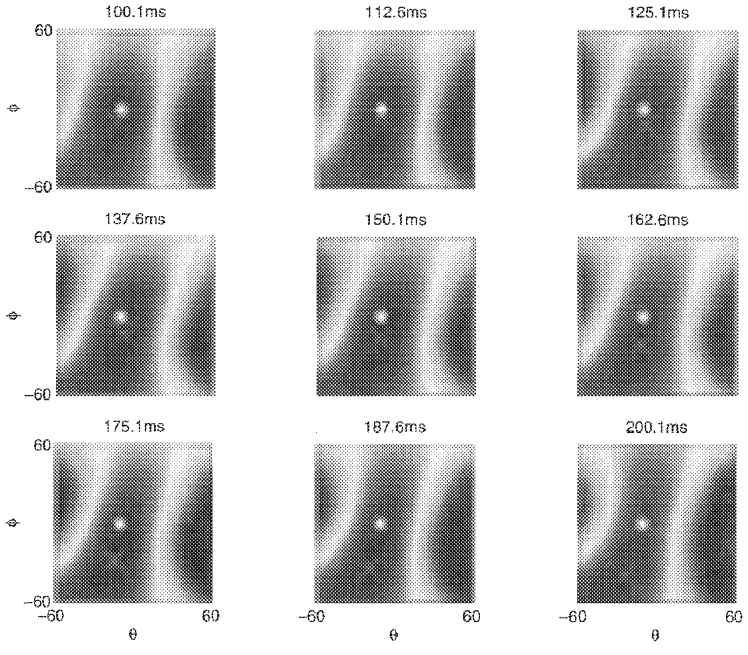


Fig. 8.4. Delay-and-sum beamformer SRP over nine, 25 ms frames.

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

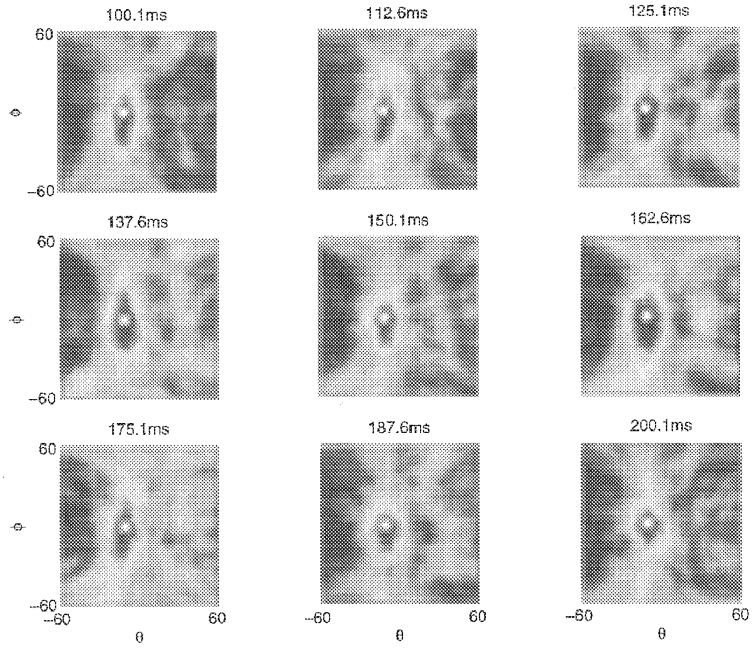


Fig. 8.5. SRP-PHAT response over nine, 25 ms frames.

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°.

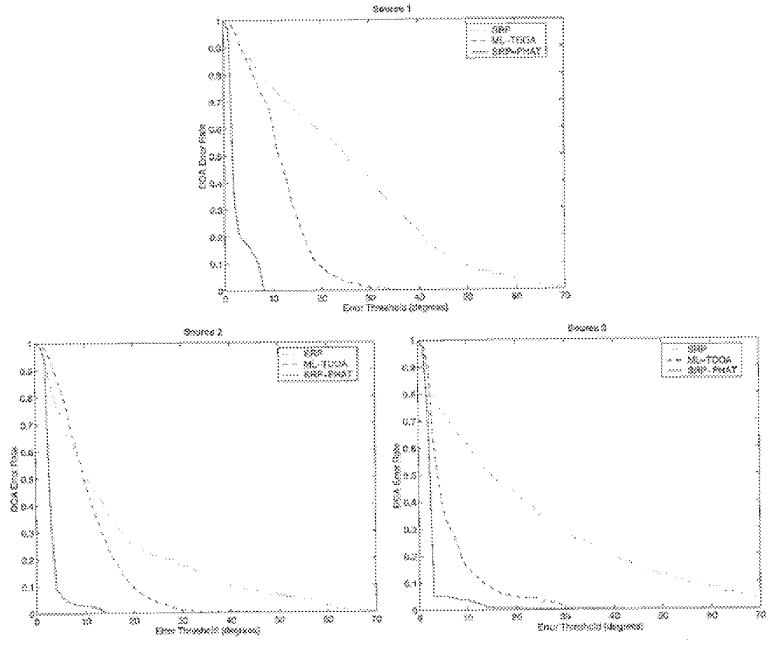


Fig. 8.6. Localizer DOA error rates for three different sources.

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]:

$$\boldsymbol{x}_i[k] = \mathbf{A}(f_i)\boldsymbol{s}_i[k] + \boldsymbol{v}_i[k]\,, \tag{9.1}$$

$$\mathbf{A}(f_i) = \left[\boldsymbol{a}_1(f_i, \theta_1) \cdots \boldsymbol{a}_D(f_i, \theta_D)\right]\,. \tag{9.2}$$

$\mathbf{A}(f_i)$ 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:

$$\mathbf{R}_{xx} = \sum_{i=1}^{J} \alpha_i \mathbf{T}(f_i)\mathbf{R}_{ss}(f_i)\mathbf{T}^H(f_i) \cong \mathbf{B}(f_0)\mathbf{R}_{ss}\mathbf{B}^H(f_0) + \sigma_v \mathbf{R}_{vv} \,. \quad (9.5)$$

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:

$$\hat{\theta}_{MUSIC} = \underset{\theta}{argmin} \left\{ \frac{|\hat{\mathbf{E}}_v^H \mathbf{V}^{-1} b(f_0, \theta)|_2}{|\mathbf{V}^{-1} b(f_0, \theta)|_2} \right\} \,. \quad (9.7)$$

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, $\hat{\theta}_{WB-WSF}$, 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 ($\hat{\cdot}$), to avoid ambiguities.

of the data is simply the sum of its narrowband components. The resulting WB-WSF estimator then assumes the following form:

$$\hat{\theta}_{WB-WSF} = \underset{\theta}{argmin} \left\{ \frac{1}{J} \sum_{i=1}^{J} |\mathbf{A}(f_i, \theta)\hat{\mathbf{C}}_i - \hat{\mathbf{E}}_s(f_i)\mathbf{W}(f_i)|_F^2 \right\}, \qquad (9.8)$$

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]:

$$\mathbf{W}(f_i)_{[m,m]} = \frac{\mu_m(f_i) - \sigma_v}{\sqrt{\mu_m(f_i)\ \sigma_v}}. \qquad (9.9)$$

The matrices $\hat{\mathbf{C}}_i(D \times\ \eta_i)$ in functional (9.8) are linked to selected weighting matrices through the equation $\hat{\mathbf{C}}_i = \mathbf{A}^{\dagger}(f_i, \theta)\hat{\mathbf{E}}_s(f_i)\mathbf{W}(f_i)$ [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

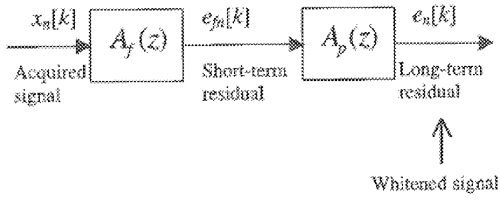


Fig. 9.1. Schematic of the Linear Prediction processing.

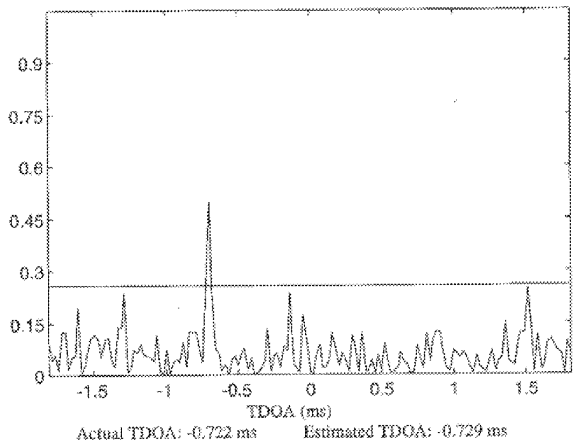


Fig. 9.2. Cross-Power-Spectrum Phase TDOA estimate without pre-whitening.

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.

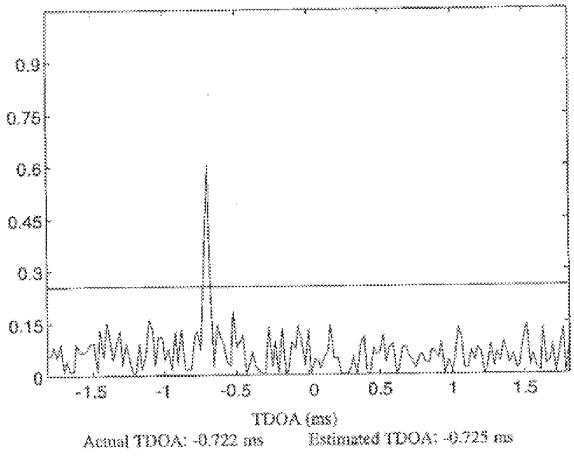


Fig. 9.3. Cross-Power-Spectrum Phase TDOA estimate with pre-whitening.

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 $\epsilon_1[k]$ and $\epsilon_2[k]$ 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 $P_{12}(f)$:

$$P_{12}(f) = \sum_{d=1}^{D} S_d(f) \left(\sum_{k=1}^{K_d} \sum_{l=1}^{K_d} \alpha_{1dk}(f)\alpha_{2dl}^*(f) e^{-j2\pi f(\tau_{1dk}-\tau_{2dl})} \right) , \tag{9.11}$$

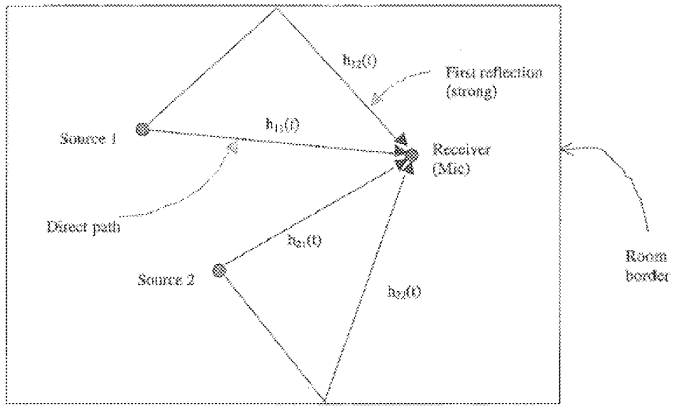


Fig. 9.4. Model of a reverberant room.

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 $\Delta_{dkl} = (\tau_{1\,dk} - \tau_{2\,dl})$ 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 $\hat{r}_l[m]$ and the true one, $r_l[m]$ [6]:

$$J_l[m] = E\left[|\hat{r}_l[m]w_l[m] - r_l[m]|^2\right] . \tag{9.15}$$

The final weights are set proportional to:

$$w_l[m] \propto \begin{cases} 1 - \dfrac{P\hat{\sigma}_p}{|\hat{r}_l[m]|^2} & \text{if } |\hat{r}_l[m]|^2 > P\hat{\sigma}_p \\ 0 & \text{elsewhere} \end{cases} . \tag{9.16}$$

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.

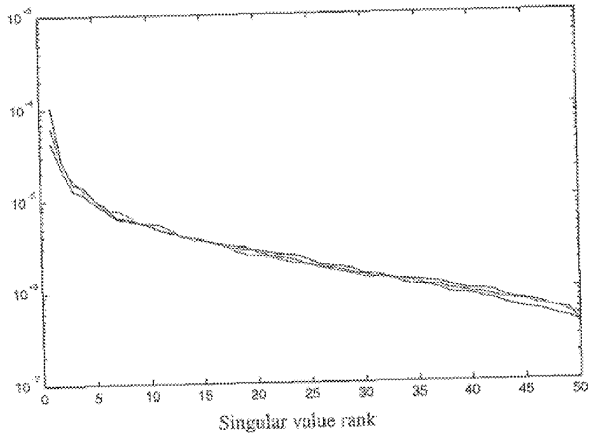


Fig. 9.5. Sample singular value spectrum of $\mathbf{P}[m]$ in three different realizations.

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.

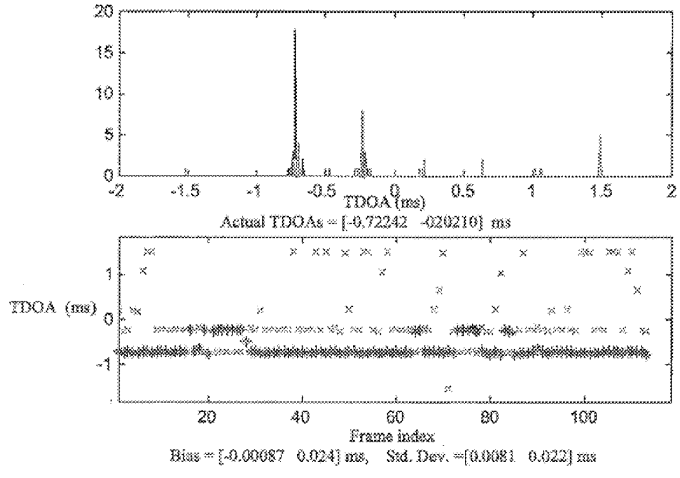


Fig. 9.6. Histogram and plot over time of TDOA estimates in a two-source case.

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

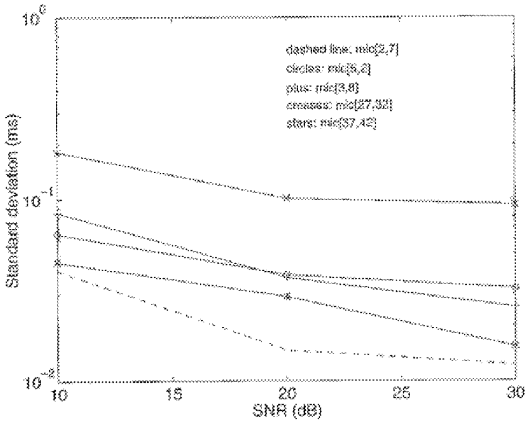


Fig. 9.7. Standard deviations of TDOA estimates vs. SNR for different pairs of microphones, in milliseconds.

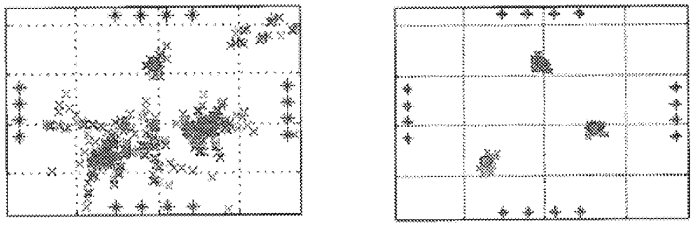


Fig. 9.8. Left: scatter plot of raw locations. Right: scatter plots of centroids.

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.

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.

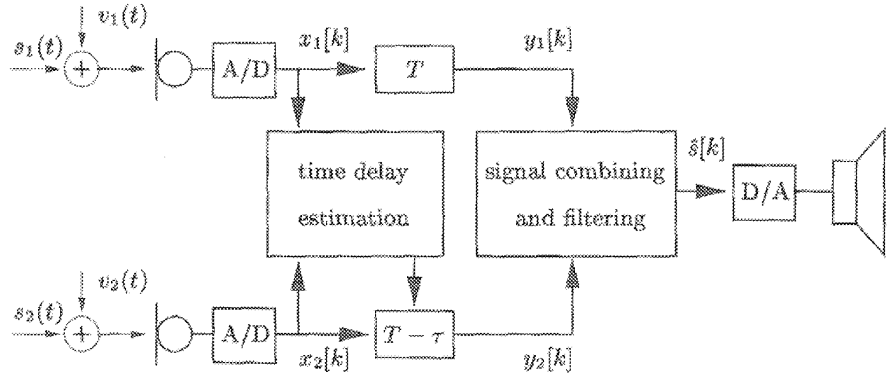


Fig. 12.1. Block diagram of the symmetric two-microphone noise reduction system.

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.

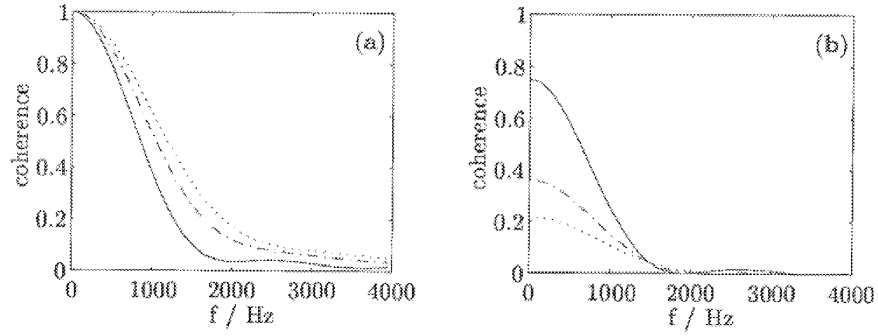


Fig. 12.2. Magnitude squared coherence of two microphone signals in an ideally diffuse noise field [17]. The microphones have a cardioid directivity pattern and different look directions. The microphone distance is 0.1 m. 0° equals broadside orientation.
(a): both microphones are turned clockwise by 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

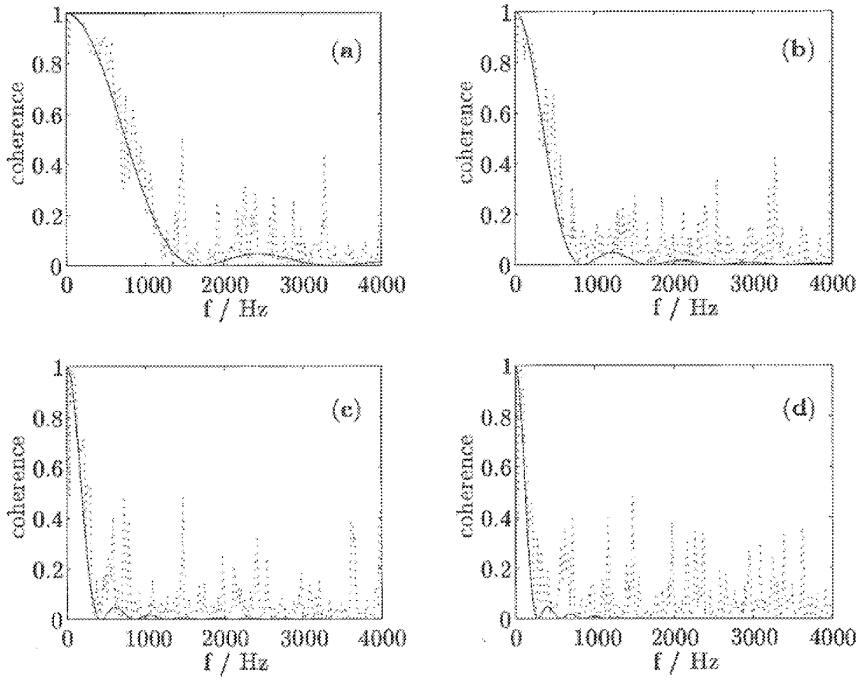


Fig. 12.3. Coherence of the ideally diffuse sound field (solid) and measured coherence (dotted) of office noise for omnidirectional microphones and microphone distances $d_{\text{mic}} = 0.1\,\text{m}$ (a), $d_{\text{mic}} = 0.2\,\text{m}$ (b), $d_{\text{mic}} = 0.4\,\text{m}$ (c), and $d_{\text{mic}} = 0.6\,\text{m}$ (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 ($T_{60} = 0.7\,\text{s}$) 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 $f_s = 8\,\text{kHz}$. 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

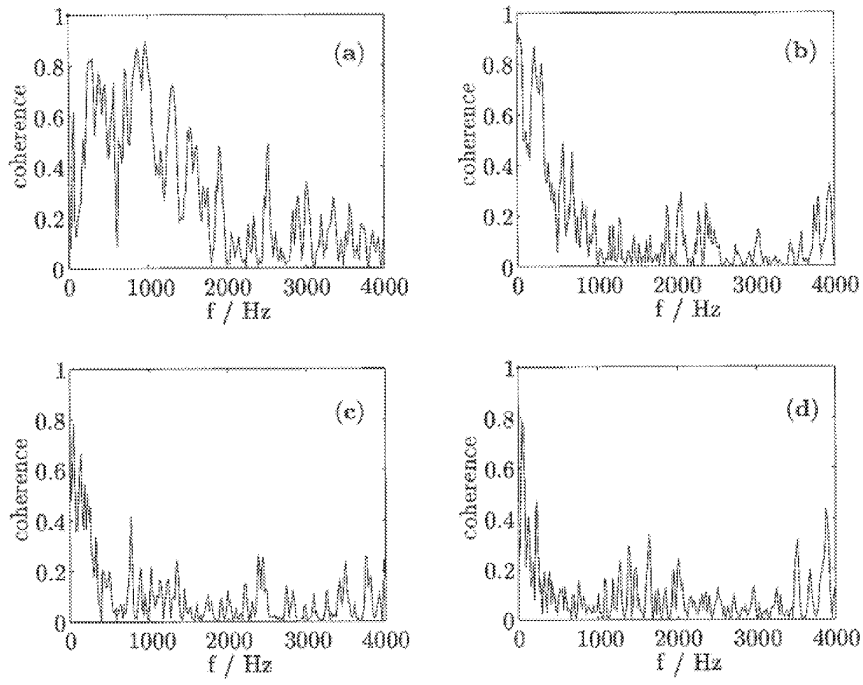


Fig. 12.4. Measured coherence of noise in a car for hypercardioid microphones and microphone distances $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$ (a), $d_{\mathrm{mic}} = 0.2\,\mathrm{m}$ (b), $d_{\mathrm{mic}} = 0.4\,\mathrm{m}$ (c), and $d_{\mathrm{mic}} = 0.6\,\mathrm{m}$ (d).

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,

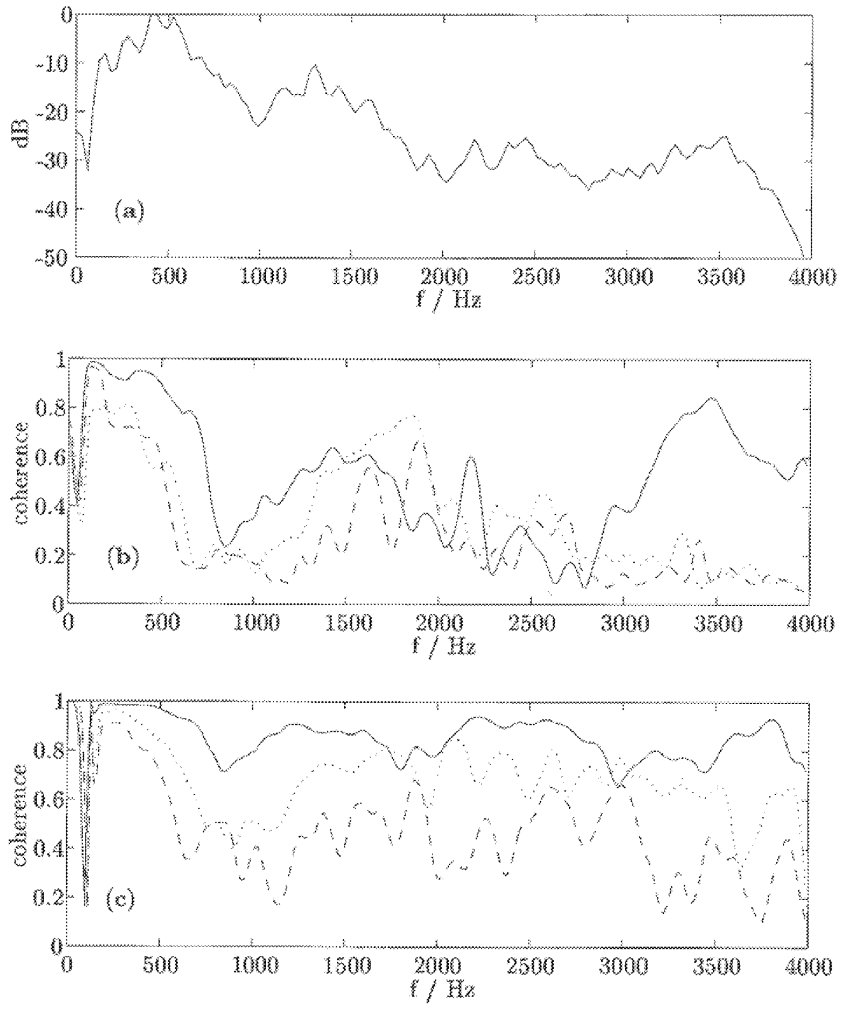


Fig. 12.5. Long term speech spectrum (a) and long term average of the MSC of a speech signal in an office room for omnidirectional (b) and for hypercardioid (c) microphones and various distances ((——): 0.5 m; (· · · · ·): 1 m; (— — —): 2 m) from speaker to microphones. The distance between microphones 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. $H_{s\text{opt}}(\Omega) = P_{s1s2}(\Omega)/P_{x2x2}(\Omega)$ and $H_{v\text{opt}}(\Omega) = P_{v1v2}(\Omega)/P_{x2x2}(\Omega)$ 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

$$E\left\{e_r^2[k]\right\} = E\left\{\left(s_1[k] - \hat{s}[k]\right)^2\right\} \tag{12.7}$$

$$= E\left\{\left(s_1[k] - \hat{\hat{s}}_1[k]\right) s_1[k]\right\} + E\left\{v_1[k]\hat{\hat{v}}_1[k]\right\} \tag{12.8}$$

$$= E\left\{v_1^2[k]\right\} + E\left\{\left(s_1[k] - \hat{\hat{s}}_1[k]\right) s_1[k]\right\} - E\left\{\left(v_1[k] - \hat{\hat{v}}_1[k]\right) v_1[k]\right\} \tag{12.9}$$

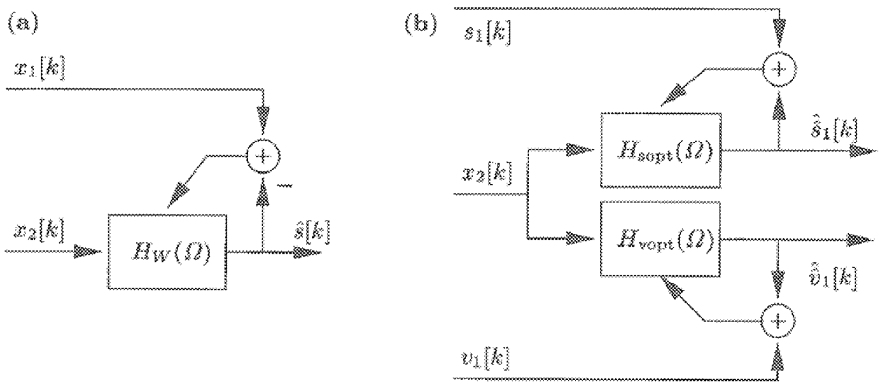


Fig. 12.6. The Wiener filter (a) and its subfilter decomposition (b).

where we used $E\left\{\widehat{v}_1^2[k]\right\} = E\left\{v_1[k]\widehat{v}_1[k]\right\}$ 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]

$$E\left\{\left(v_1[k] - \widehat{v}_1[k]\right) v_1[k]\right\}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1 v_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{\text{vopt}}(\Omega)\, P^*_{v_1 v_2}(\Omega)\, d\Omega \quad (12.10)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1 v_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1 v_2}(\Omega)|^2}{P_{s_2 s_2}(\Omega) + P_{v_2 v_2}(\Omega)} d\Omega \quad (12.11)$$

and

$$E\left\{\left(s_1[k] - \widehat{s}_1[k]\right) s_1[k]\right\}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1 s_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{\text{sopt}}(\Omega)\, P^*_{s_1 s_2}(\Omega)\, d\Omega \quad (12.12)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1 s_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{s_1 s_2}(\Omega)|^2}{P_{s_2 s_2}(\Omega) + P_{v_2 v_2}(\Omega)} d\Omega \quad (12.13)$$

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

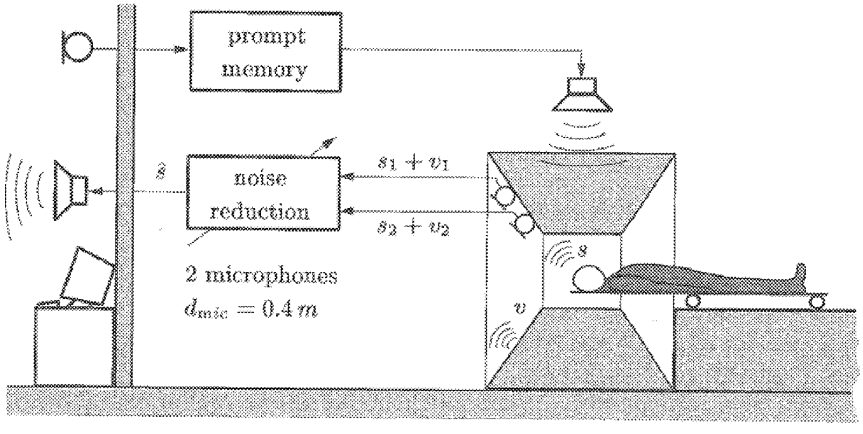


Fig. 12.7. Application of the two-microphone noise reduction system for voice communication in a computed tomography scanner.

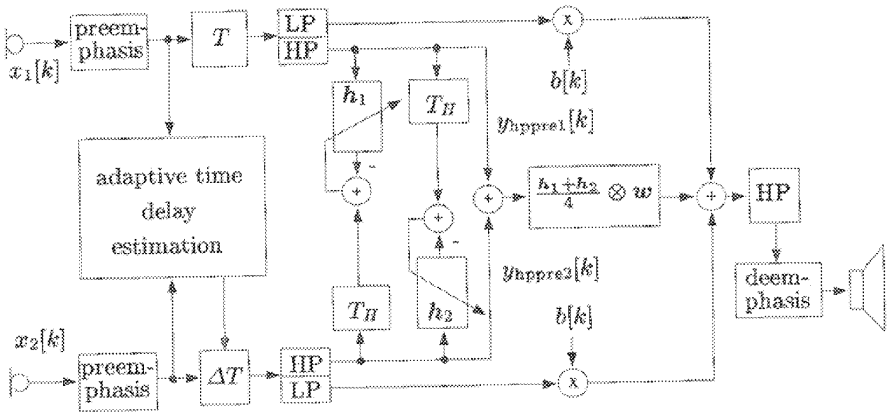


Fig. 12.8. Block diagram of the two-microphone noise reduction system. LP: lowpass filter, HP: highpass filter, $\Delta T = T - \tau$.

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 ($T_H = L_H/2$),

$$\boldsymbol{h}_1[k+1] = \boldsymbol{h}_1[k] + \alpha e_1[k] \frac{\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\mathrm{hppre1}}[k]}{\boldsymbol{y}^T_{\mathrm{hppre1}}[k] \boldsymbol{y}_{\mathrm{hppre1}}[k]} \tag{12.17}$$

$$\boldsymbol{h}_2[k+1] = \boldsymbol{h}_2[k] + \alpha e_2[k] \frac{\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\mathrm{hppre2}}[k]}{\boldsymbol{y}^T_{\mathrm{hppre2}}[k] \boldsymbol{y}_{\mathrm{hppre2}}[k]}, \tag{12.18}$$

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

$$e_1[k] = y_{\mathrm{hppre2}}[k - T_H] - \boldsymbol{y}^T_{\mathrm{hppre1}}[k] \boldsymbol{h}_1[k] \tag{12.20}$$

$$e_2[k] = y_{\mathrm{hppre1}}[k - T_H] - \boldsymbol{y}^T_{\mathrm{hppre2}}[k] \boldsymbol{h}_2[k], \tag{12.21}$$

and

$$\boldsymbol{y}_{\mathrm{hppre1}}[k] = \left(y_{\mathrm{hppre1}}[k],\ \ldots, y_{\mathrm{hppre1}}[k - L_H]\right)^T \tag{12.22}$$

$$\boldsymbol{y}_{\mathrm{hppre2}}[k] = \left(y_{\mathrm{hppre2}}[k],\ \ldots, y_{\mathrm{hppre2}}[k - L_H]\right)^T \tag{12.23}$$

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 $\boldsymbol{w} = (w_0, w_1, \ldots, w_{L_H})^T$

$$\boldsymbol{h}[k] = \frac{\boldsymbol{h}_1[k] + \boldsymbol{h}_2[k]}{2} \otimes \boldsymbol{w}. \tag{12.24}$$

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 $\tilde{s}[k]$ with respect to the unprocessed delayed speech signal $s[k - T_H]$

$$SEGSNR^s_{\tilde{s}-s} = \frac{1}{K}\sum_{m=0}^{K-1} \max\left(SNR^s_{\tilde{s}-s}(m), 0\right) \tag{12.27}$$

with

$$SNR^{s}_{s-s}(m) = 10 \cdot \log_{10} \left(\frac{\sum_{k=mM}^{mM+M-1} s^2[k - T_H]}{\sum_{k=mM}^{mM+M-1} \left(\tilde{s}[k] - s[k - T_H]\right)^2} \right) . \tag{12.28}$$

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.

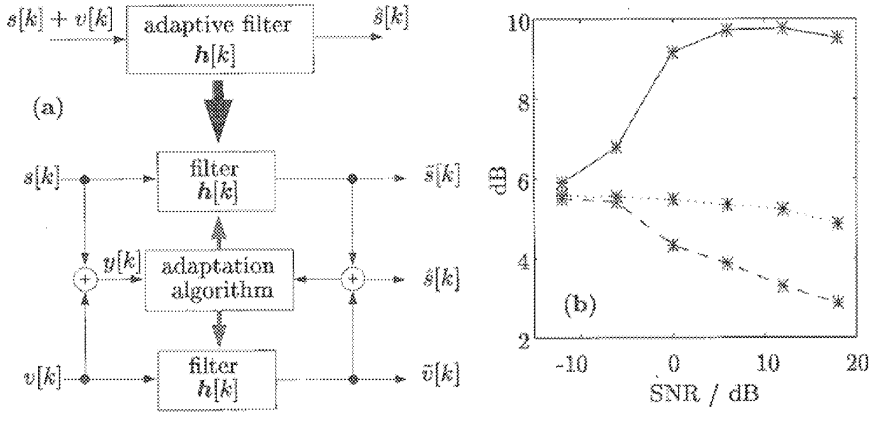


Fig. 12.9. Method for computing objective measures (a). Distortion of speech signal and noise reduction during speech activity and during speech pause vs. input SNR of the adaptive filter in the upper band (b). Step size: $\alpha = 0.1$, filter order $L_H = 64$.
(———): Segmental SNR of speech; (· · · ·): 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 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

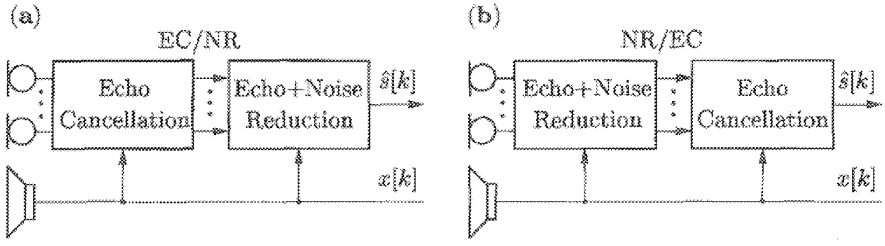


Fig. 12.10. Processing options for combined systems: EC precedes ENR (a), ENR precedes EC (b).

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.

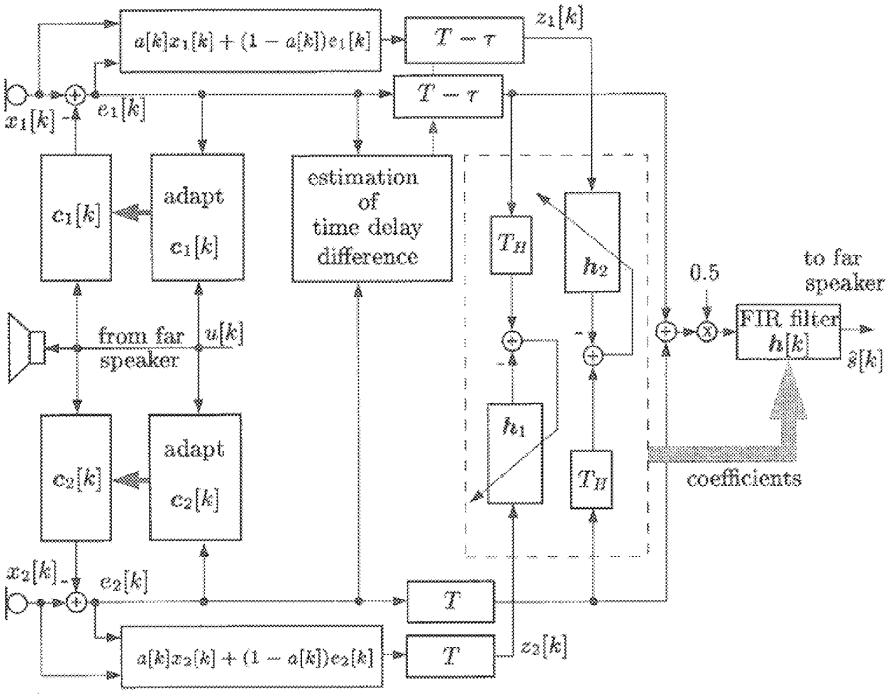


Fig. 12.11. Block diagram of the two microphone combined echo and noise reduction algorithm.

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 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.

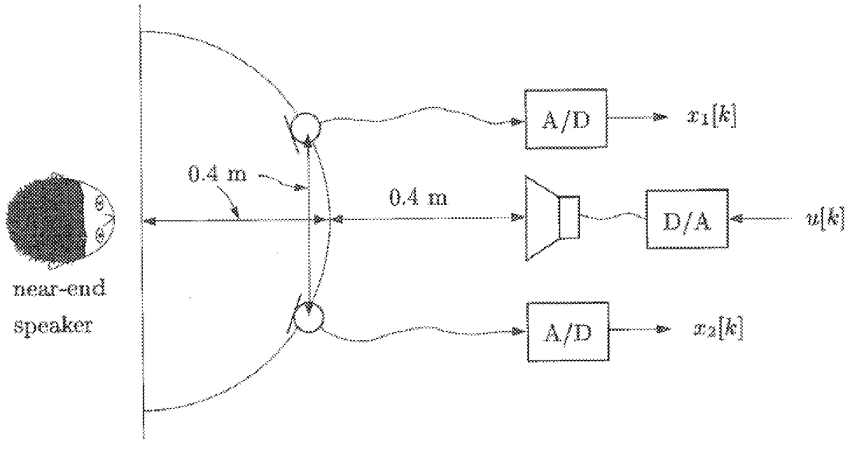


Fig. 12.12. Setup of the near-end speaker, the microphones, and the loudspeaker for conferencing experiments. The microphones have a hypercardioid directivity pattern.

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.

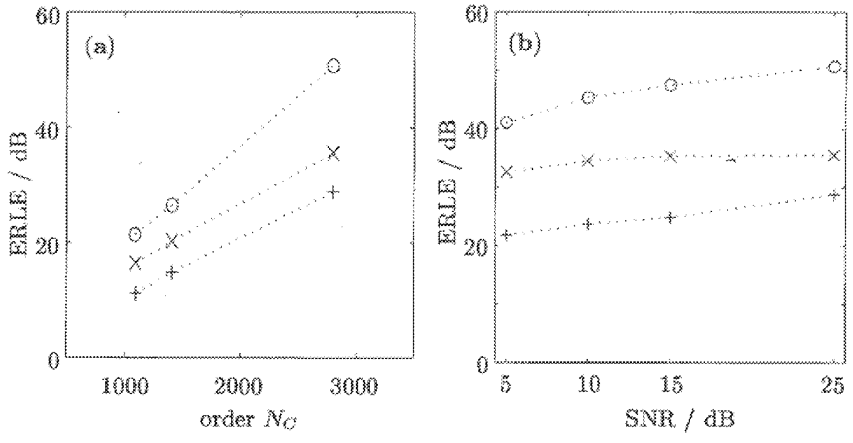


Fig. 12.13. Mean ERLE for single talk vs. the compensator order N_C for SNR ≈ 30 dB (**a**) and vs. the SNR for $N_C = 2800$ (**b**).
+: mean ERLE of echo canceller only;
x: mean ERLE of combined system with $a[k] \equiv 0$;
o: mean ERLE of combined system with $a[k]$ adaptive.

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_0-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.

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.

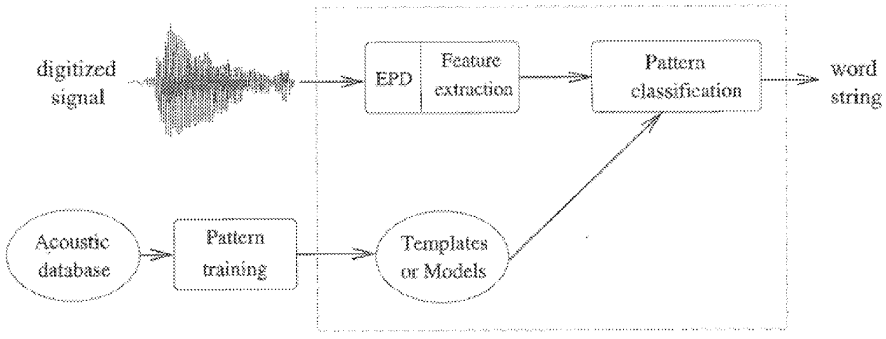


Fig. 15.1. General block diagram of a pattern recognition based ASR system.

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

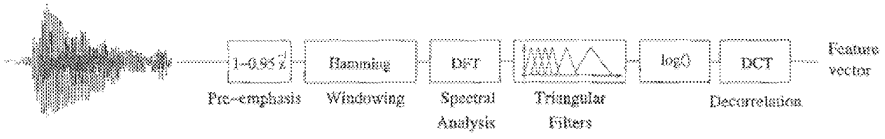


Fig. 15.2. Block diagram of the computation of Mel-based cepstral coefficients.

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

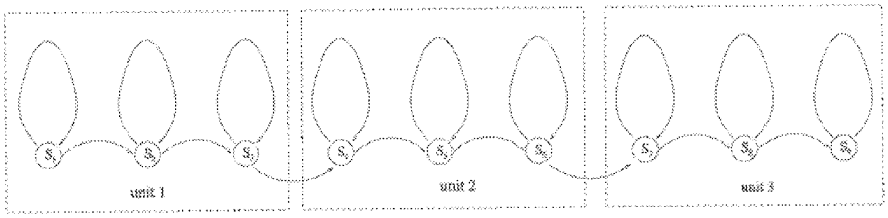


Fig. 15.3. Concatenation of three HMM-based phone models, characterized by a left-to-right three-state topology.

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.

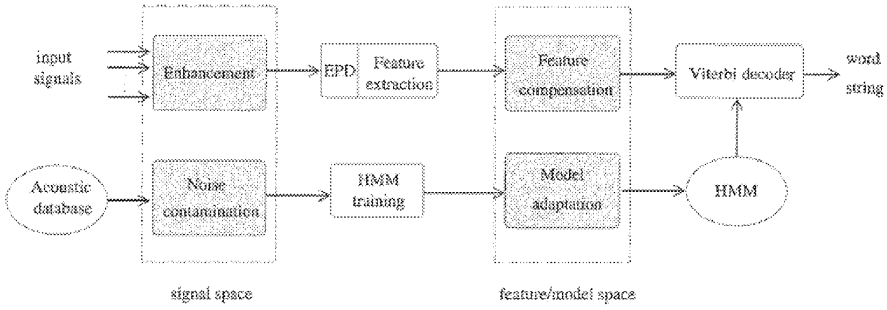


Fig. 15.4. Block diagram of a robust ASR system operating in an adverse environment. Highlighted blocks, operating either in the signal space or in the feature/model space, contribute to reduce the mismatch between noisy data and the acoustic models used by the recognizer.

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,

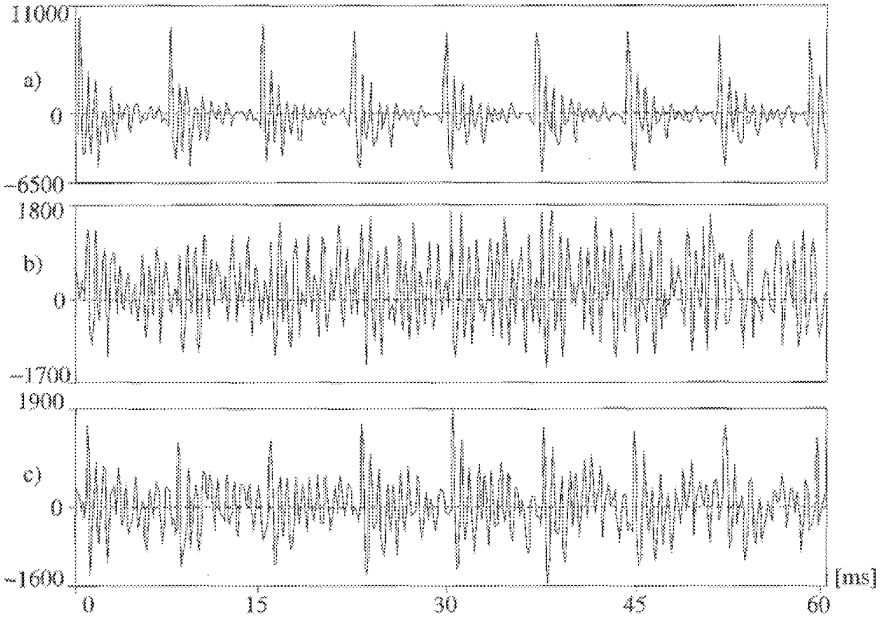


Fig. 15.5. Portion of a vowel /a/ uttered at 4 m distance from a linear array of six microphones. The upper waveform is the close-talker signal, the middle plot represents the reverberated signal acquired by one of the microphones, and the lower plot shows the corresponding delay-and-sum beamformed signal.

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.

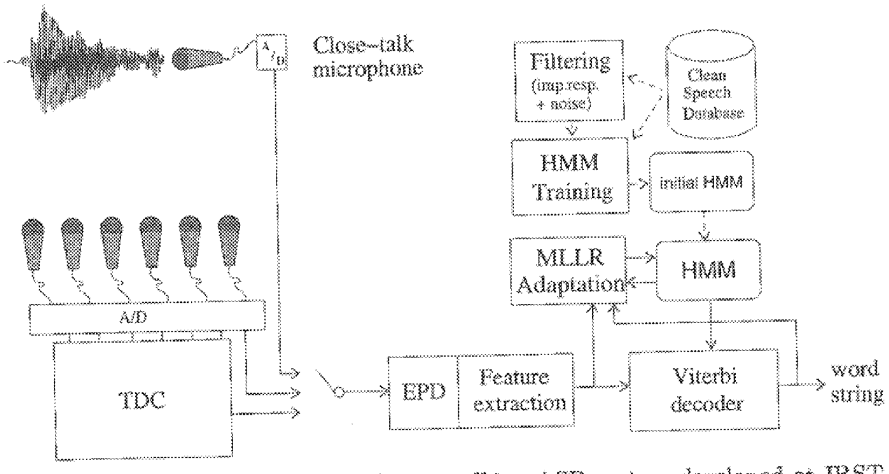


Fig. 15.6. Block diagram of the distant-talking ASR system developed at IRST. It includes signal acquisition, Time Delay Compensation processing, the acoustic front-end, a recognition engine, and the modules for training and adaptation of the acoustic models. Close-talker and single far-field microphones 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.

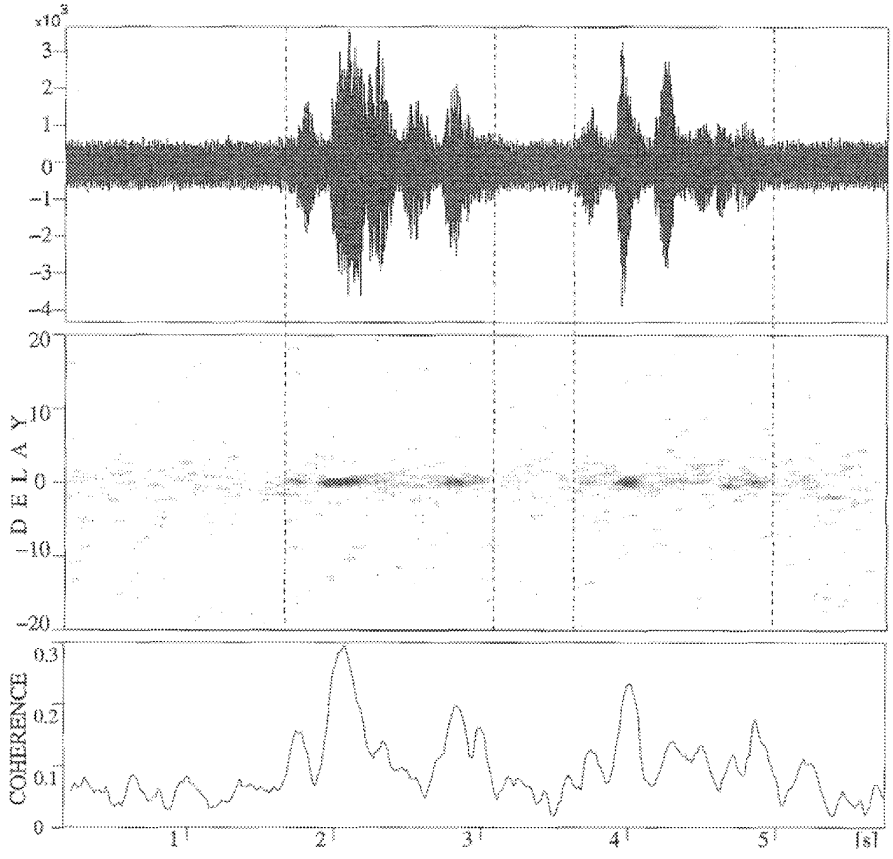


Fig. 15.7. Example of coherence computation for the signals of a microphone pair. It includes one of the noisy speech waveforms acquired by the microphones, a grey level coherence measure representation at various inter-channel delays, and the coherence levels at the correct delay (0 samples).

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.

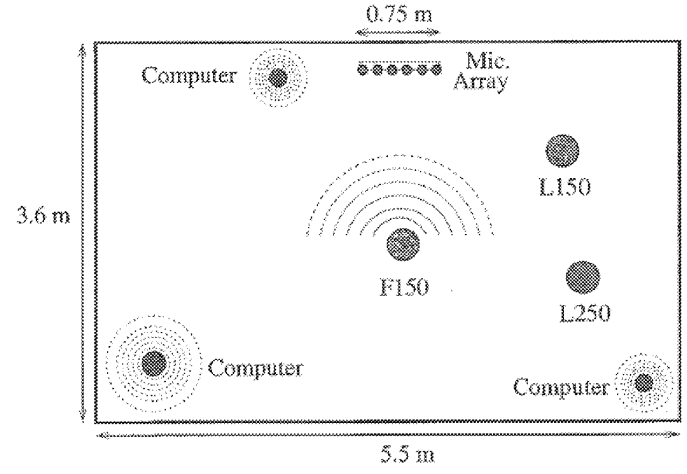


Fig. 15.8. Map of the experimental room showing the positions of the talker, the microphone array, and the computers (noise sources). The label at each position indicates, in a compact form, the orientation (F for frontal, L for lateral) and the distance in cm from the array.

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.

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

Printing: Mercedes-Druck, Berlin
Binding: Stein+Lehmann, Berlin

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. 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. 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.

ISBN 3-540-41953-5



http://www.springer.de