Digital Signal Processing

Michael Brandstein · Darren Ward
Microphone Arrays

Springer-Verlag Berlin Heidelberg GmbH

Michael Brandstein · Darren Ward (Eds.)

Microphone Arrays

Signal Processing Techniques and Applications

With 149 Figures

Springer

Series Editors

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

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

Editors

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

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

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

Cip data applied for

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

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

Preface

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

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

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

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

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

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

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

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

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

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

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

Cambridge MA, USA — *Michael Brandstein*
London, UK — *Darren Ward*
January 2001

Contents

List of Contributors

Joerg Bitzer
Houpert Digital Audio
Bremen, Germany

Michael S. Brandstein
Harvard Universtiy
Cambridge MA, USA

Ingvar Claesson
Blekinge Inst. of Technology
Ronneby, Sweden

Joseph H. DiBiase
Brown Universtiy
Providence RI, USA

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

Simon Doclo
Katholieke Universiteit Leuven
Leuven, Belgium

Scott C. Douglas
Southern Methodist University
Dallas TX, USA

Gary W. Elko
Agere Systems
Murray Hill NJ, USA

Nedelko Grbić
Blekinge Inst. of Technology
Ronneby, Sweden

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

Scott M. Griebel
Harvard Universtiy
Cambridge MA, USA

Osamu Hoshuyama
NEC Media Research Labs
Kawasaki, Japan

Walter L. Kellermann
University Erlangen-Nuremberg
Erlangen, Germany

Rodney A. Kennedy
The Australian National University
Canberra, Australia

Claude Marro
France Télécom R&D
Lannion, France

Rainer Martin
Aachen University of Technology
Aachen, Germany

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

Marc Moonen
Katholieke Universiteit Leuven
Leuven, Belgium

Sven Nordholm
Curtin University of Technology
Perth, Australia

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

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

Rudolf Rabenstein
University Erlangen-Nuremberg
Erlangen, Germany

Harvey F. Silverman
Brown Universtiy
Providence RI, USA

K. Uwe Simmer
Aureca GmbH
Bremen, Germany

Sascha Spors
University Erlangen-Nuremberg
Erlangen, Germany

Norbert Strobel
Siemens Medical Solutions
Erlangen, Germany

Akihiko Sugiyama
NEC Media Research Labs
Kawasaki, Japan

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

Dirk Van Compernolle
Katholieke Universiteit Leuven
Leuven, Belgium

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

Robert C. Williamson
The Australian National University
Canberra, Australia

Patrick M. Zurek
Sensimetrics Corporation
Somerville MA, USA

Part I

Speech Enhancement

1 Constant Directivity Beamforming

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

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

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

1.1 Introduction

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

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

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

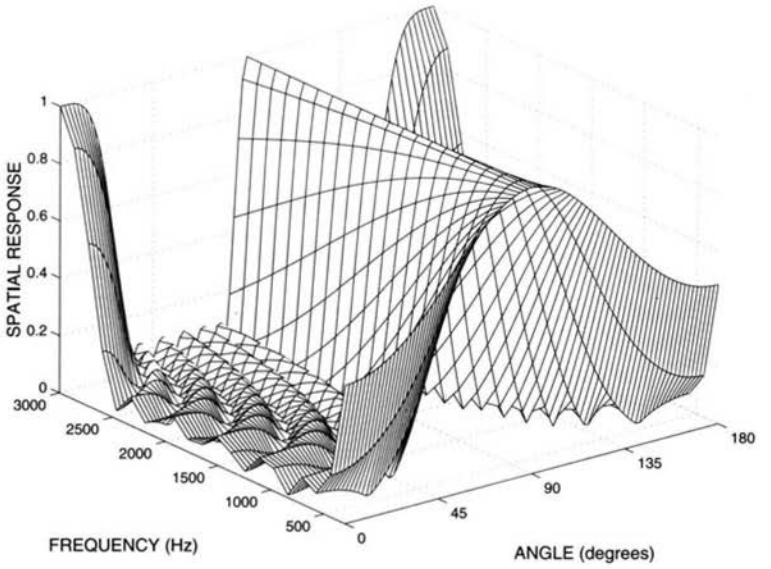

Fig. 1.1. Response of a narrowband array operated over a wide bandwidth.

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

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

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

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

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

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

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

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

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

Time-domain versus frequency-domain beamforming

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

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

1.2 Problem Formulation

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

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

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

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

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

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

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

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

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

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

The spatial response of the beamformer is given by

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

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

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

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

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

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

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

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

1.3 Theoretical Solution

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

1.3.1 Continuous sensor

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

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

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

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

and the spatial response for a source at angle θ is

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

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

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

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

Let the sensor weighting function be given by

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

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

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

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

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

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

1.3.2 Beam-shaping function

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

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

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

1.4 Practical Implementation

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

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

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

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

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

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

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

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

1.4.1 Dimension-reducing parameterization

Define the *reference beam-shaping filter response* as

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

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

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

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

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

where

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

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

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

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

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

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

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

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

where

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

is a L-dimensional *transformation vector*.

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

1.4.2 Reference beam-shaping filter

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

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

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

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

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

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

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

1.4.3 Sensor placement

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

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

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

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

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

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

1.4.4 Summary of implementation

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

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

where

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

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

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

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

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

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

1.5 Examples

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

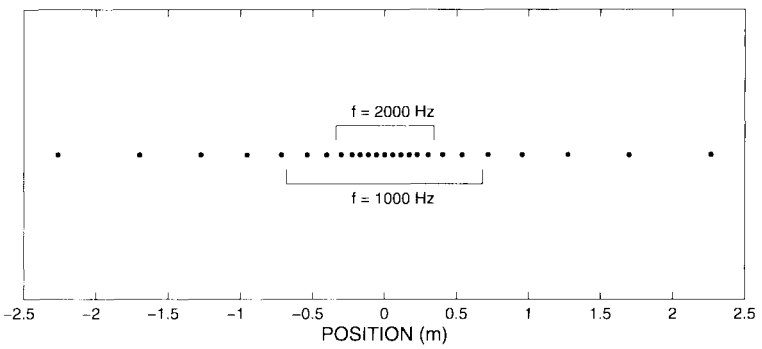

Fig. 1.2. Array geometry used for example CDB.

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

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

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

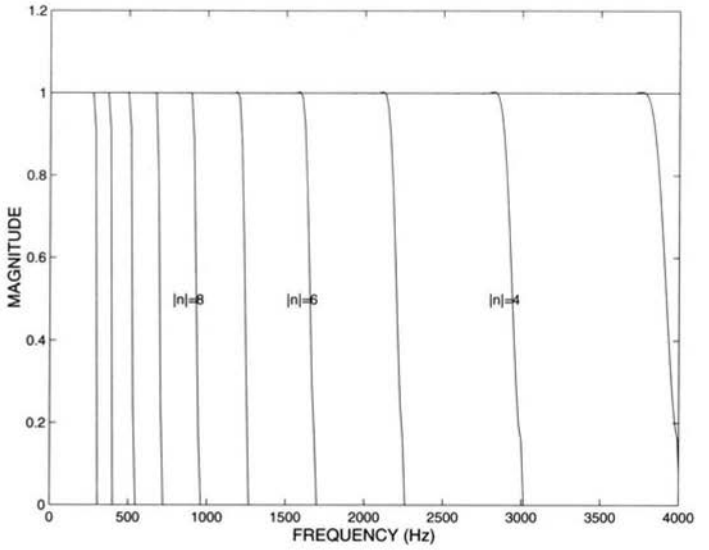

Fig. 1.3. Frequency responses of the weights on each sensor.

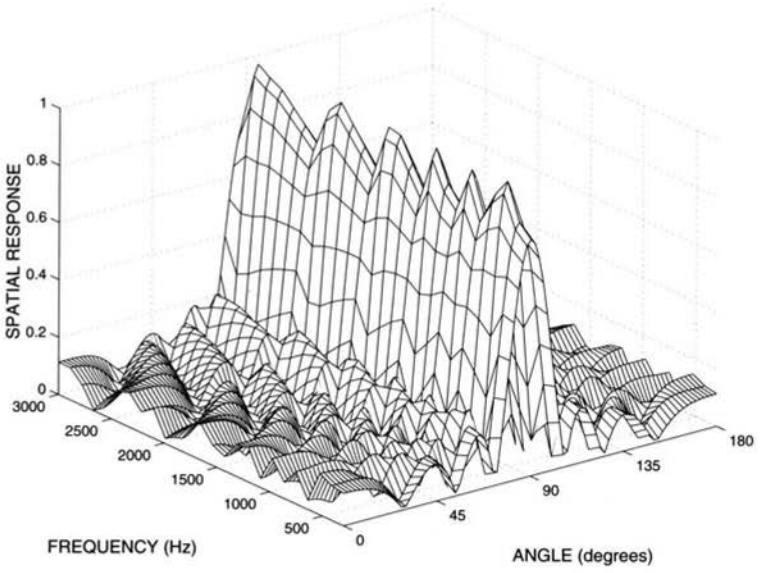

Fig. 1.4. Spatial response of example CDB.

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

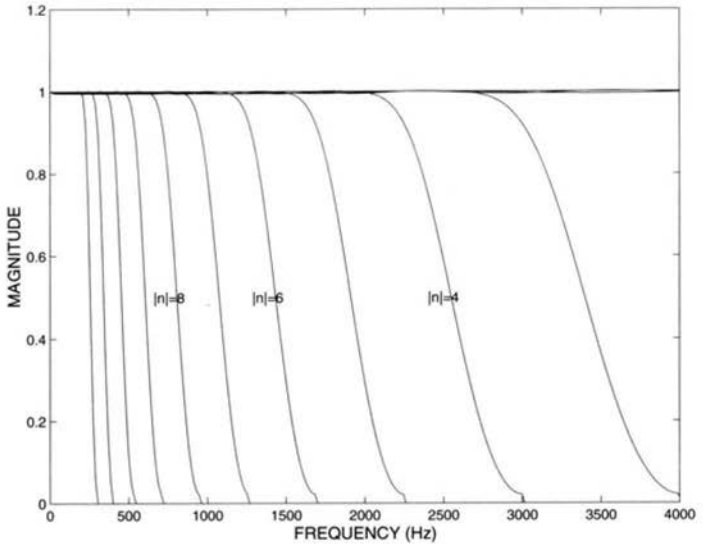

Fig. 1.5. Frequency responses of the weights on each sensor.

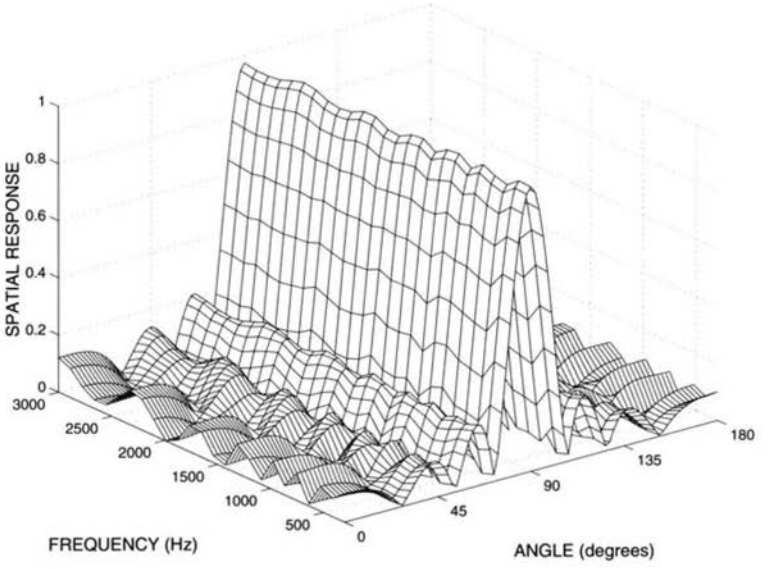

Fig. 1.6. Spatial response of example CDB.

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

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

1.6 Conclusions

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

References

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

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

2 Superdirective Microphone Arrays

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

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

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

2.1 Introduction

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

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

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

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

2.2 Evaluation of Beamformers

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

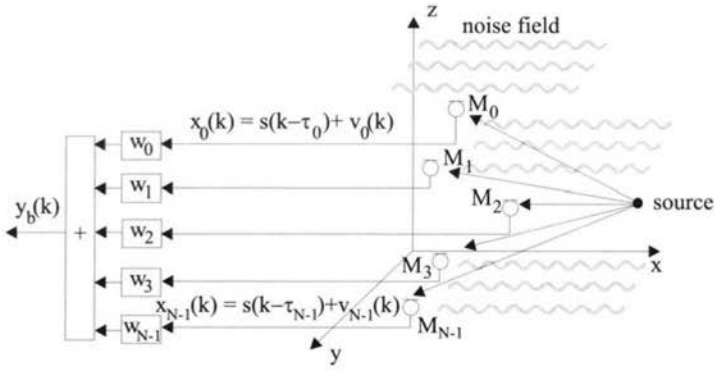

Fig. 2.1. Signal model consisting of noise field and desired source signal

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

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

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

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

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

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

direction of the source signal.

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

Finally, the output signal

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

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

2.2.1 Array-Gain

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

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

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

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

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

where

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

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

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

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

and for the noise-only case the output is

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

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

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

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

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

where

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

is the coherence function [4].

Thus,

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

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

2.2.2 Beampattern

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

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

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

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

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

where

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

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

2.2.3 Directivity

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

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

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

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

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

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

2.2.4 Front-to-Back Ratio

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

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

2.2.5 White Noise Gain

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

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

into (2.13) results in the white noise gain:

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

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

2.3 Design of Superdirective Beamformers

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

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

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

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

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

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

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

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

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

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

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

Examples for desired signal models are:

- Standard farfield model for linear arrays with equidistant sensors:

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

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

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

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

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

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

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

2.3.1 Delay-and-Sum Beamformer

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

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

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

2.3.2 Design for spherical isotropic noise

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

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

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

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

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

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

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

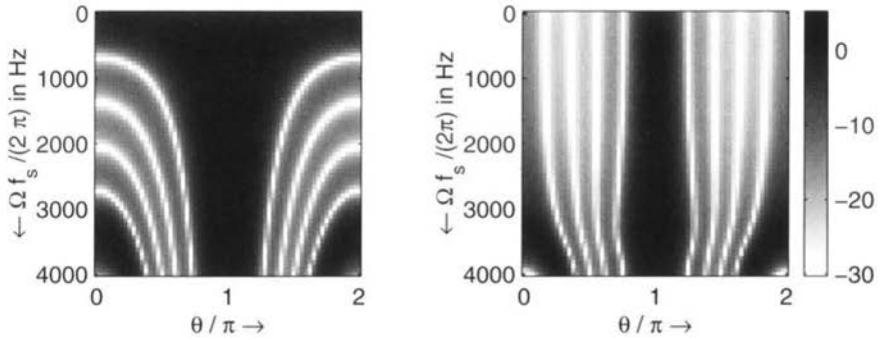

Fig. 2.2. Left: beampattern of a delay-and-sum beamformer. Right: beampattern of an optimal array for isotropic noise (superdirective beamformer). (l = 5 cm, $N = 5$, endfire steering direction)

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

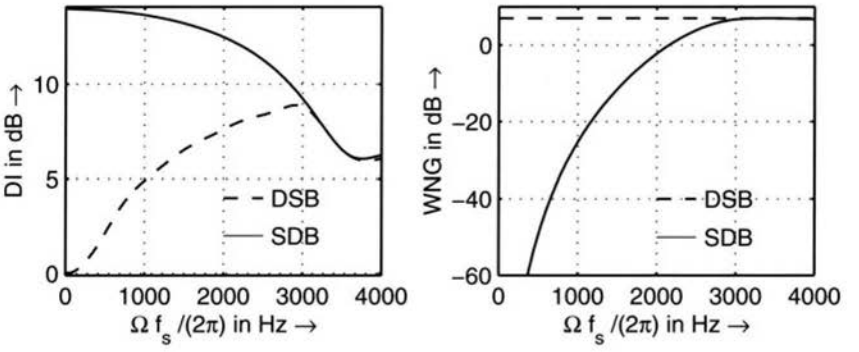

Fig. 2.3. Left: Directivity index (DI) for delay-and-sum beamformer and superdirective beamformer. Right: White noise gain (WNG) for delay-and-sum beamformer and superdirective beamformer. (l = 5 cm, $N = 5$, endfire steering direction)

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

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

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

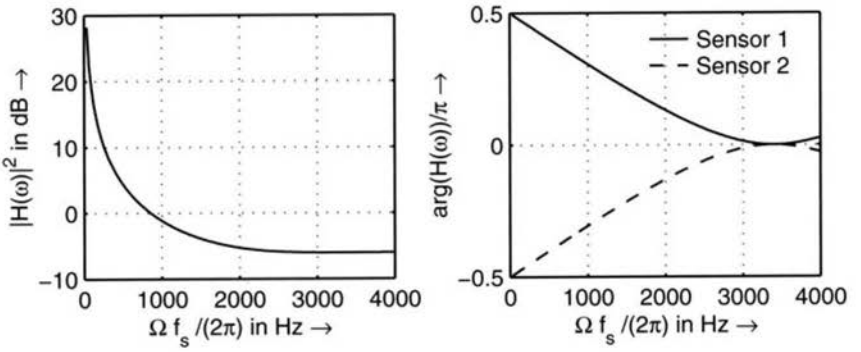

Fig. 2.4. Coefficients of a two channel SDB, left: Magnitude, right: Phase ($l = 5$ cm, $N = 2$, endfire steering direction)

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

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

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

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

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

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

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

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

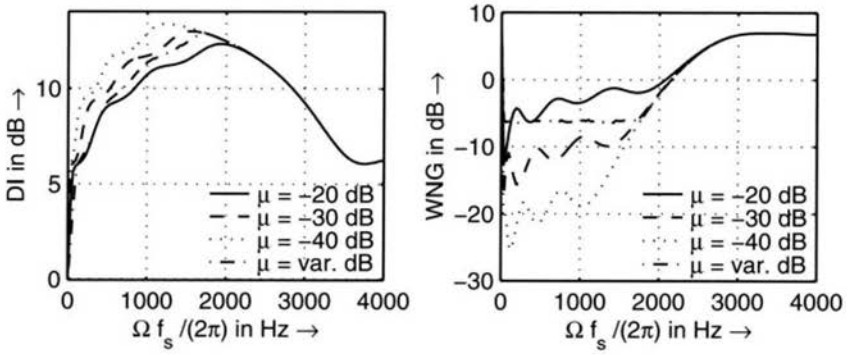

Fig. 2.5. Left: Directivity index (DI) for different constrained designs. Right: White noise gain (WNG) for different constrained designs. ($l = 5$ cm, $N = 5$, endfire steering direction)

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

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

2.3.3 Design for Cylindrical Isotropic Noise

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

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

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

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

2.3.4 Design for an Optimal Front-to-Back Ratio

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

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

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

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

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

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

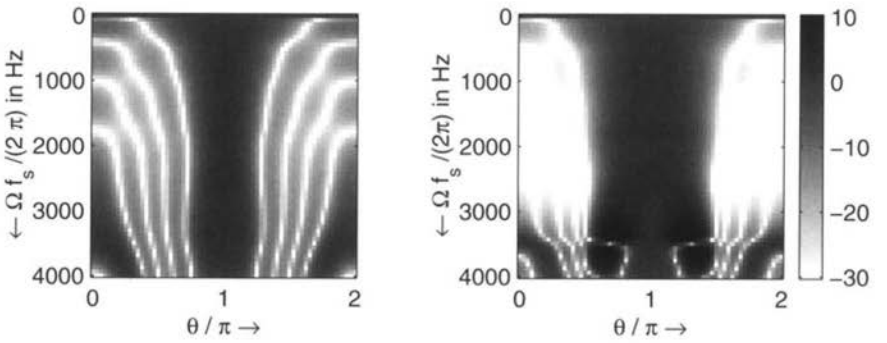

Fig. 2.6. Left: beampattern of a constrained superdirective beamformer. Right: beampattern of a constrained beamformer, designed with (2.36). ($l = 5$ cm, $N = 5$, $\mu = 0.01$, endfire steering direction)

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

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

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

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

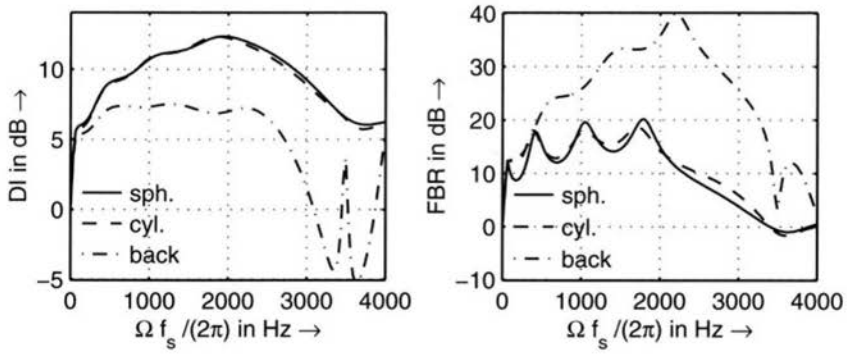

Fig. 2.7. Left: Directivity index (DI) for three optimal designs. Right: Front-to-back ratio (FBR) for three optimal designs. ($l = 5$ cm, $N = 5$, $\mu = 0.01$, endfire steering direction)

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

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

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

2.3.5 Design for Measured Noise Fields

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

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

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

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

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

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

2.4 Extensions and Details

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

2.4.1 Alternative Form

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

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

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

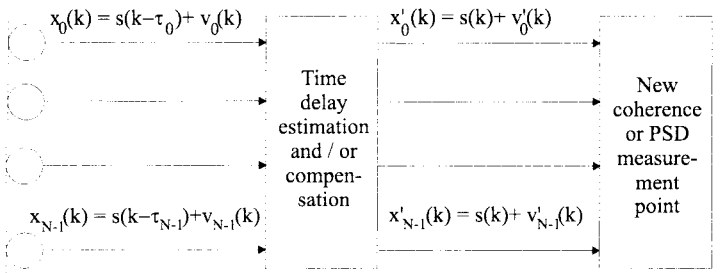

Fig. 2.8. Signal model after time delay compensation

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

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

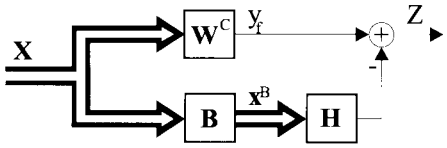

Fig. 2.9. Schematic description of the decomposition of the optimal weight vector into two orthogonal parts

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

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

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

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

An example for $N = 4$ is given by

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

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

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

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

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

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

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

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

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

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

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

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

2.4.2 Comparison with Gradient Microphones

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

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

The output of such systems is given by

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

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

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

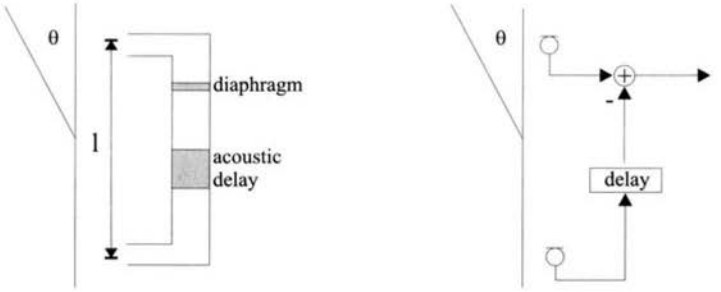

Fig. 2.10. Schematic description of a first order gradient microphone

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

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

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

2.5 Conclusion

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

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

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

References

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

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

3 Post-Filtering Techniques

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

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

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

3.1 Introduction

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

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

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

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

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

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

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

3.2 Multi-channel Wiener Filtering in Subbands

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

3.2.1 Derivation of the Optimum Solution

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

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

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

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

Using the definitions for the power of a complex signal

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

the cross-correlation vector

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

and the correlation matrix

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

the squared error at time k may be written as

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

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

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

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

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

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

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

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

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

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

3.2.2 Factorization of the Wiener Solution

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

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

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

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

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

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

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

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

and the correlation matrix may be expressed as

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

Consequently, the optimum weight vector may be written as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2.3 Interpretation

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

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

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

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

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

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

3.3 Algorithms for Post-Filter Estimation

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

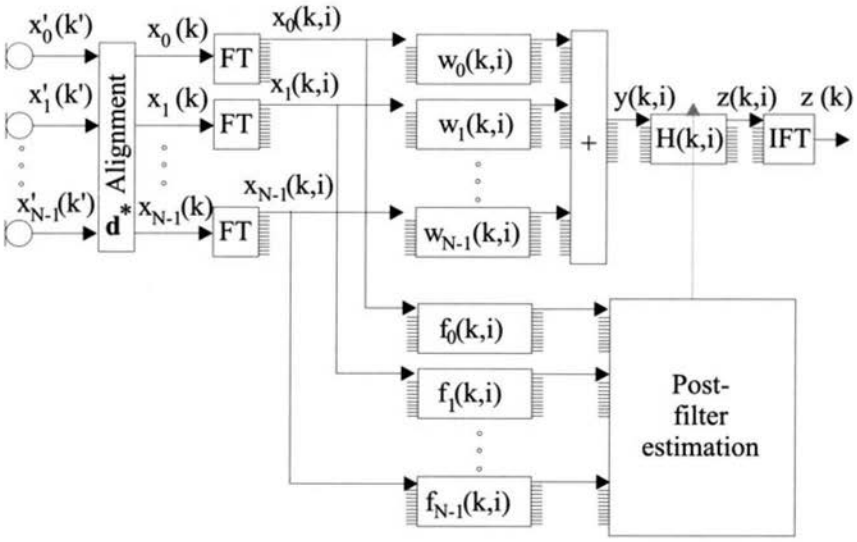

Fig. 3.1. General block diagram of the examined post-filters.

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

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

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

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

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

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

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

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

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

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

3.3.1 Analysis of Post-Filter Algorithms

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3.2 Properties of Post-Filter Algorithms

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

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

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

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

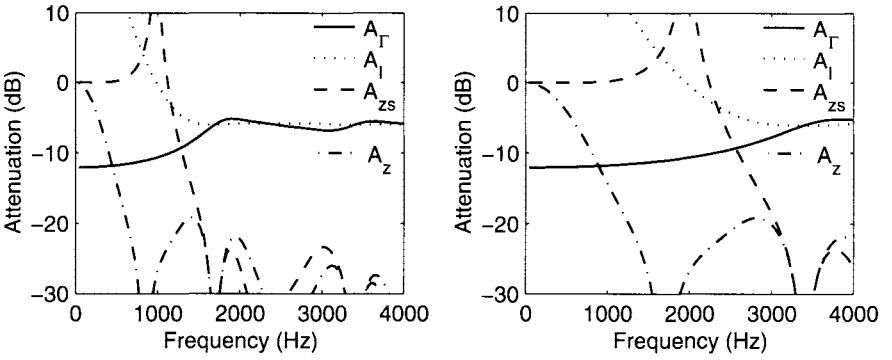

Fig. 3.2. Theoretical noise attenuation of an end-fire array for a diffuse noise field. Left: delay and sum beamformer coefficients. Right: superdirective coefficients.

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

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

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

3.3.3 A New Post-Filter Algorithm

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

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

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

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

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

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

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

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

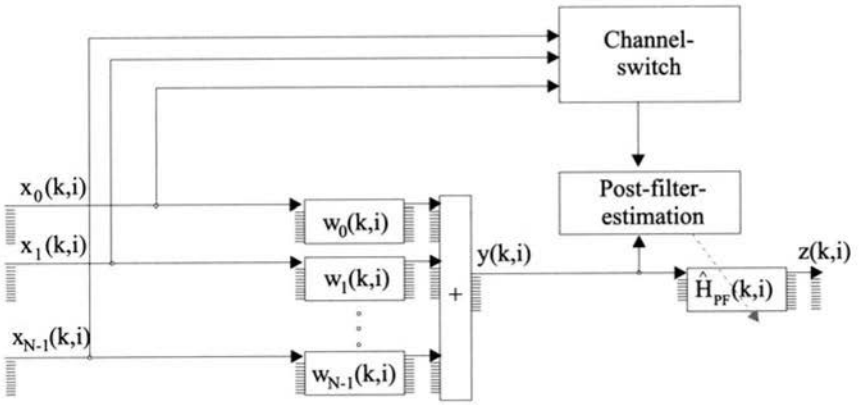

Fig. 3.3. Block diagram of the adaptive post-filter for an arbitrary beamformer (APAB).

3.4 Performance Evaluation

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

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

3.4.1 Simulation System

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

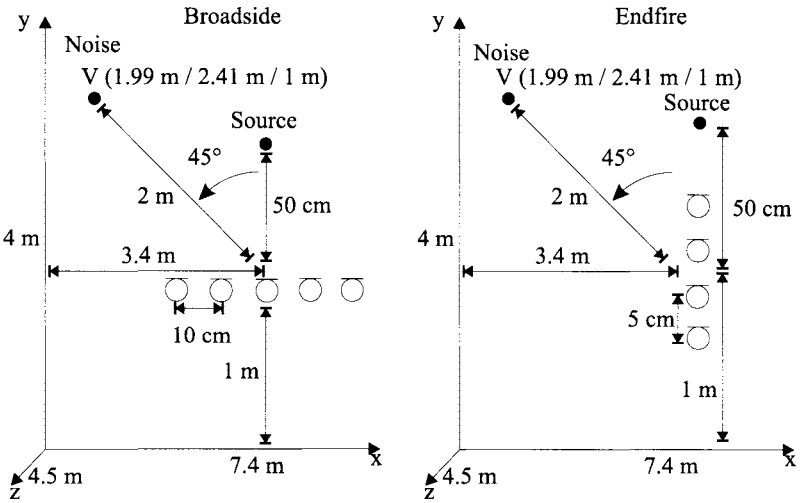

Fig. 3.4. Configuration of the simulated room.

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

3.4.2 Objective Measures

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

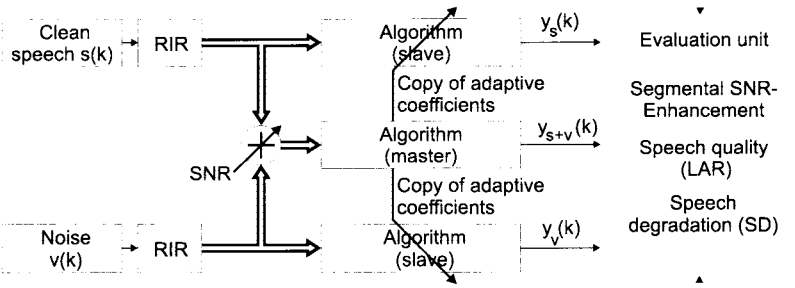

Fig. 3.5. Graphical description of the complete simulation system.

hancement (SNRE):

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

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

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

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

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

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

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

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

3. Compute the LAR of block l

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

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

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

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

3.4.3 Simulation Results

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

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

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

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

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

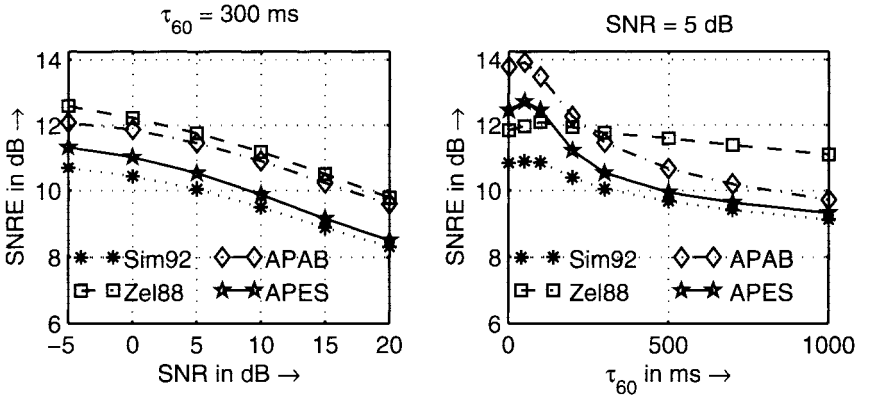

Fig. 3.6. Left: SNRE vs. input-SNR. Right: SNRE vs. reverberation time τ_{60} (Broadside).

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

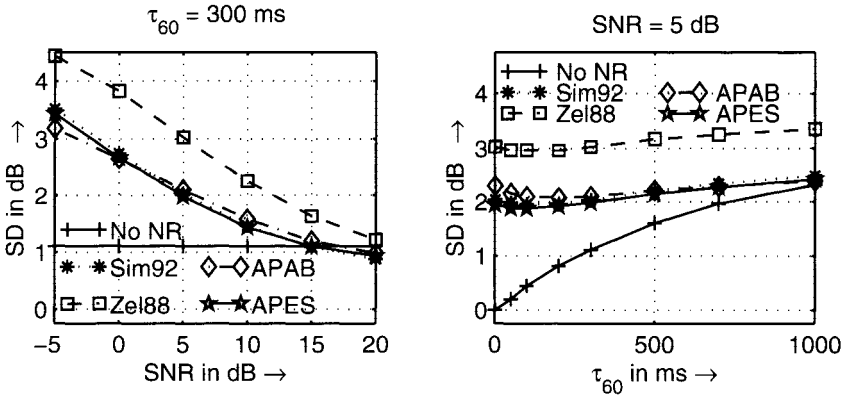

Fig. 3.7. Left: SD vs. input-SNR. Right: SD vs. reverberation time τ_{60} (Broadside).

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

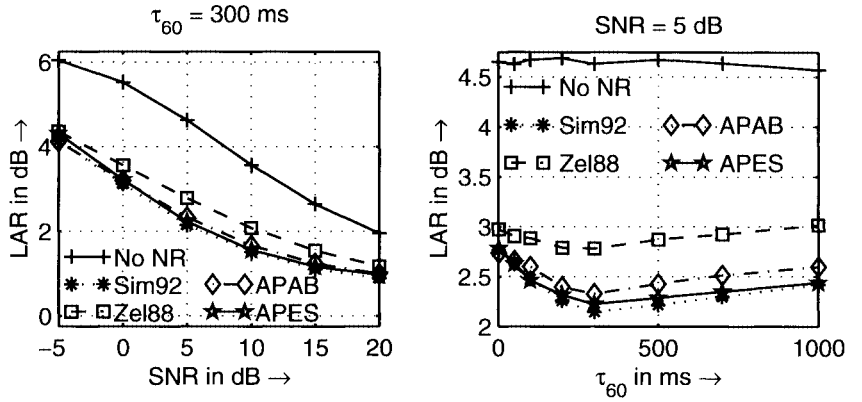

Fig. 3.8. Left: LAR vs. input-SNR. Right: LAR vs. reverberation time τ_{60} (Broadside).

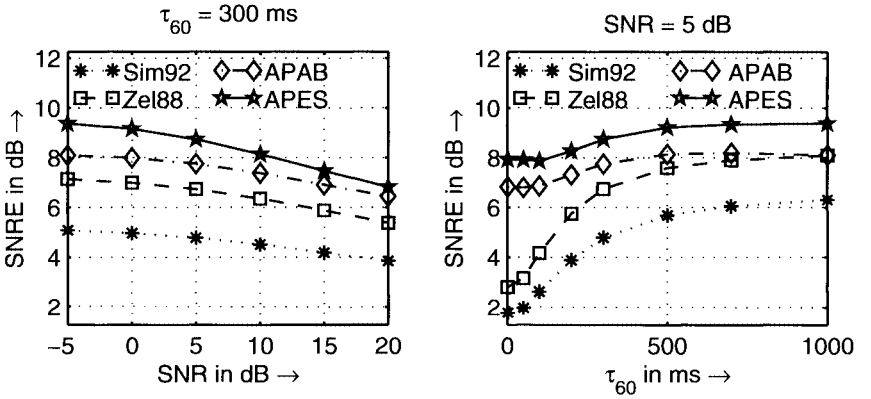

Fig. 3.9. Left: SNRE vs. input-SNR. Right: SNRE vs. reverberation time τ_{60} (Endfire).

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

[1] Smaller values indicate better quality.

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

3.5 Conclusion

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

References

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

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

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

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

4 Spatial Coherence Functions for Differential Microphones in Isotropic Noise Fields

Gary W. Elko

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

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

4.1 Introduction

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

4.2 Adaptive Noise Cancellation

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

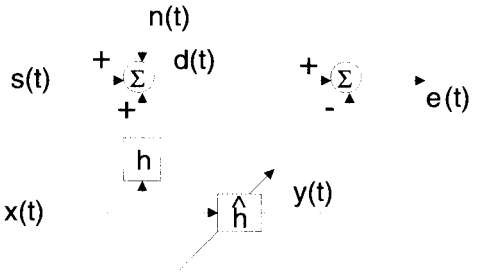

Fig. 4.1. Schematic model of adaptive noise cancellation system.

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

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

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

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

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

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

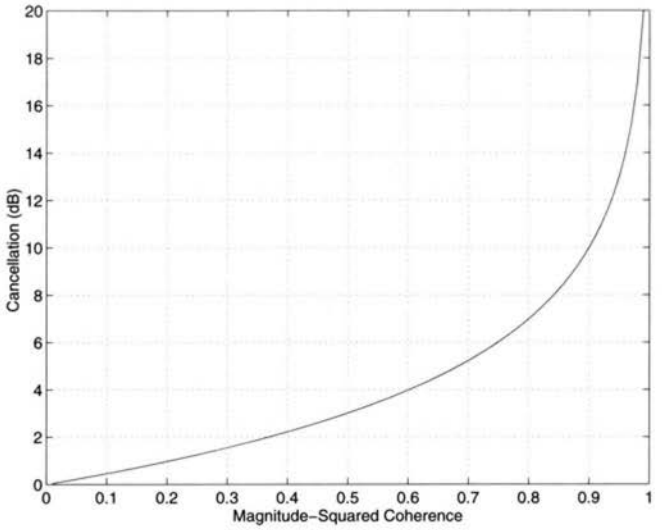

Fig. 4.2. Adaptive cancellation in dB versus the mean-square coherence between the noise signal $x(t)$ and the signal $d(t)$ as defined in Fig 4.1.

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

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

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

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

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

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

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

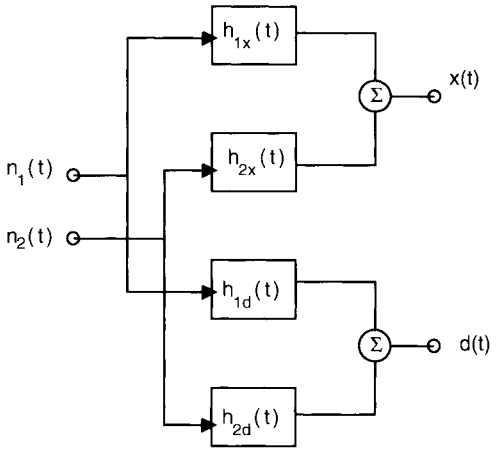

Fig. 4.3. Schematic model of two independent sources n_1 and n_2 combining through filters to form signals x and d.

For this case the autospectral densities are,

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

and

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

The cross-spectral density is,

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

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

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

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

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

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

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

4.3 Spherically Isotropic Coherence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

and

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

the solution for the directivity factor is

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

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

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

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

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

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

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

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

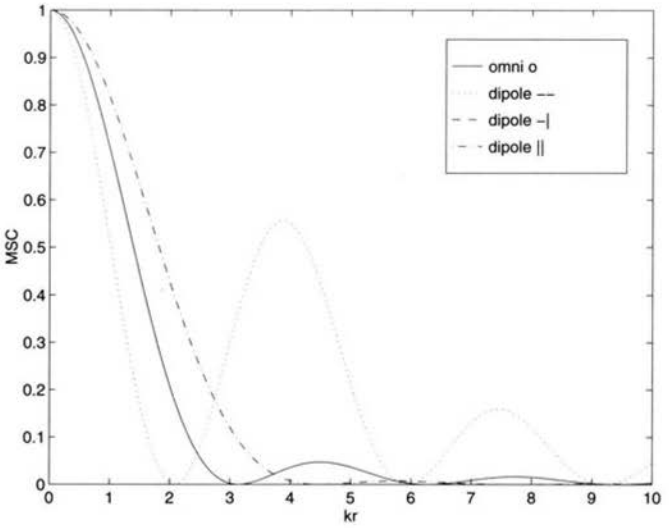

Fig. 4.4. Magnitude-squared coherence (MSC) for omnidirectional and dipole microphones in a spherically isotropic noise field. Note that the curve for the orthogonal dipoles lies along the abscissa.

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

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

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

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

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

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

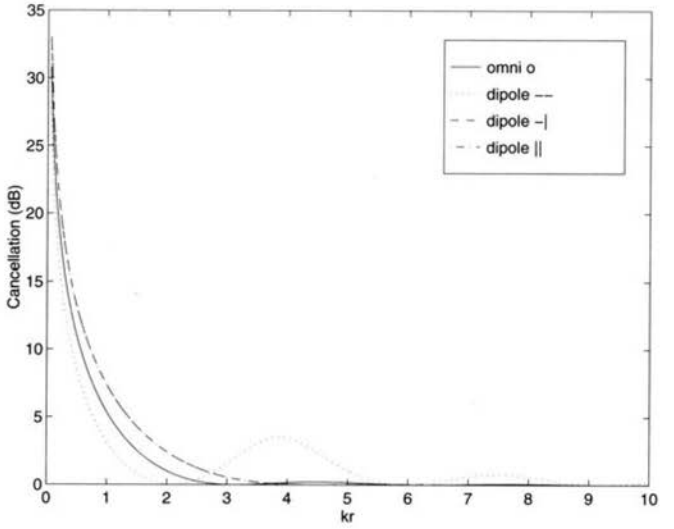

Fig. 4.5. Maximum cancellation (dB) for omnidirectional and of dipole microphones for spherically isotropic noise fields. Note that the curve for the orthogonal dipoles lies along the abscissa.

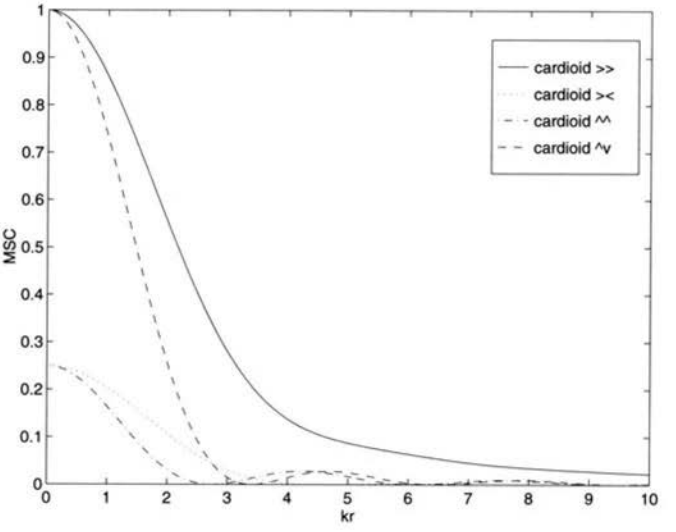

Fig. 4.6. Magnitude-squared coherence (MSC) for various orientations of cardioid microphones in a spherically isotropic noise field.

From Appendix A, the result is

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

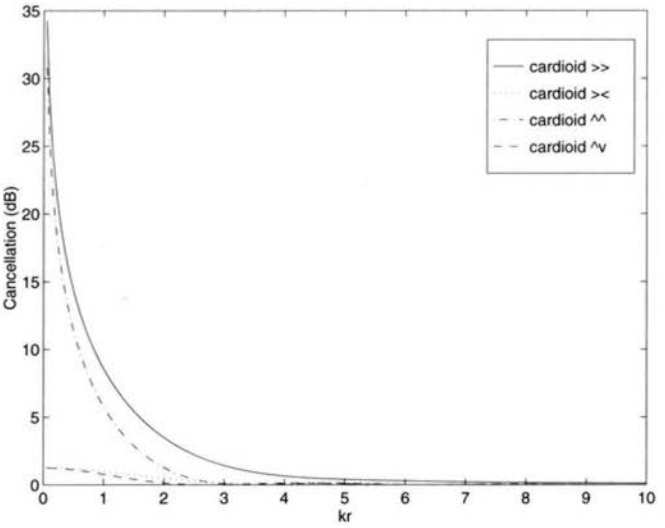

Fig. 4.7. Maximum cancellation (dB) for various orientations of cardioid microphones for spherically isotropic noise fields.

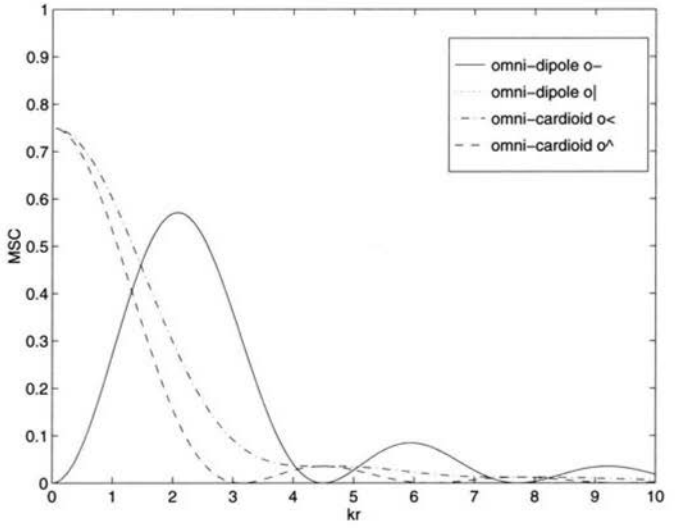

Fig. 4.8. Magnitude-squared coherence (MSC) for various orientations of omnidirectional and dipole and cardioid microphones in a spherically isotropic noise field.

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

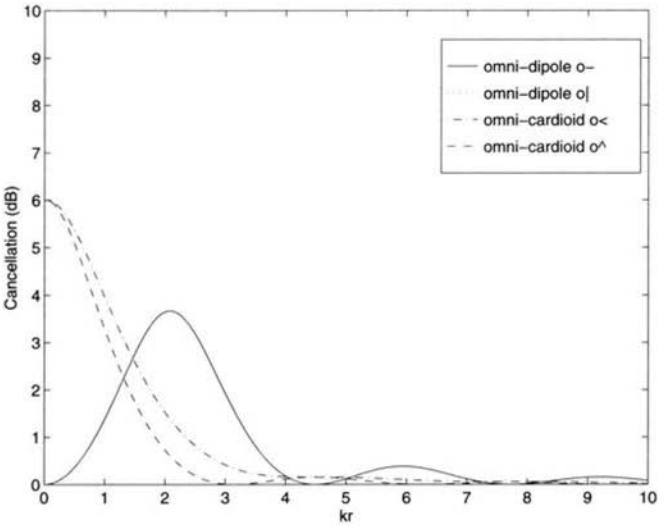

Fig. 4.9. Maximum cancellation (dB) for various orientations of omnidirectional and dipole and cardioid microphones for spherically isotropic noise fields.

combination of collinear differential arrays is

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

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

4.4 Cylindrically Isotropic Fields

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

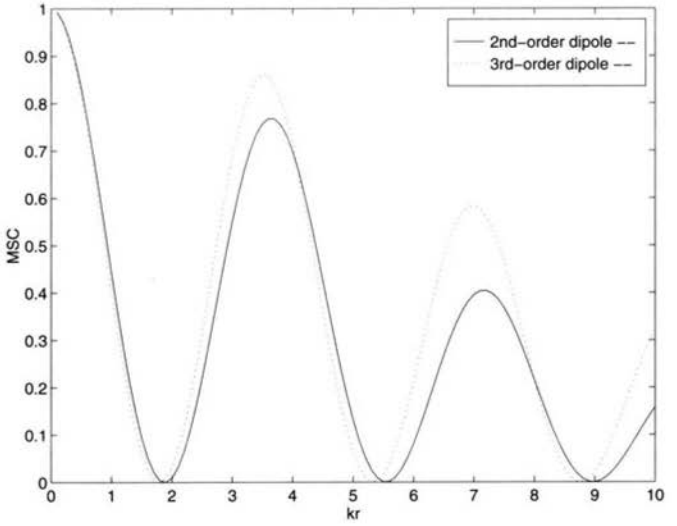

Fig. 4.10. Magnitude-squared coherence for second and third-order collinear dipoles in a spherically isotropic noise field.

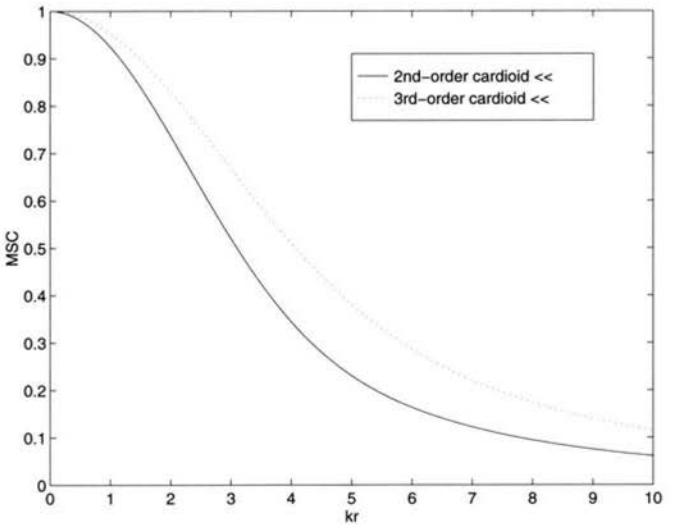

Fig. 4.11. Magnitude-squared coherence for second and third-order collinear cardioids in a spherically isotropic noise field.

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

ϕ_1 and ϕ_2 are

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

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

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

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

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

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

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

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

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

From Appendix C, I_n is,

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

where ε_m is defined as,

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

and the function C is the binomial coefficient [7]

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

The numerator of the coherence function is

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

where the coefficients d_n are components of the vector

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.5 Conclusions

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

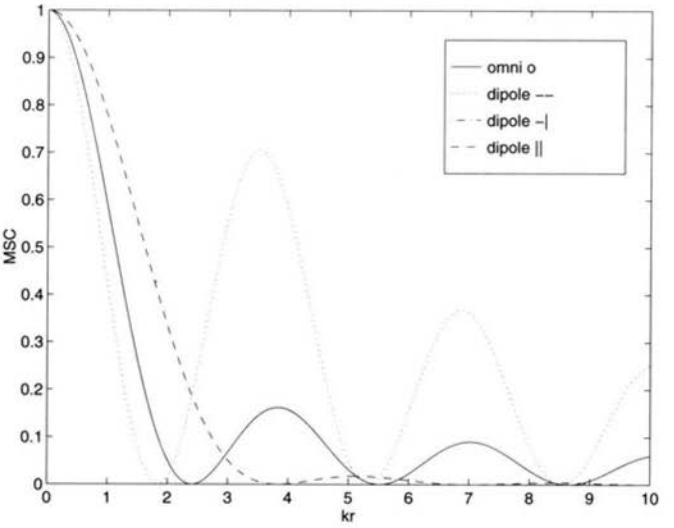

Fig. 4.12. Magnitude-squared coherence (MSC) for omnidirectional and various orientations of dipole microphones in a cylindrically isotropic noise field.

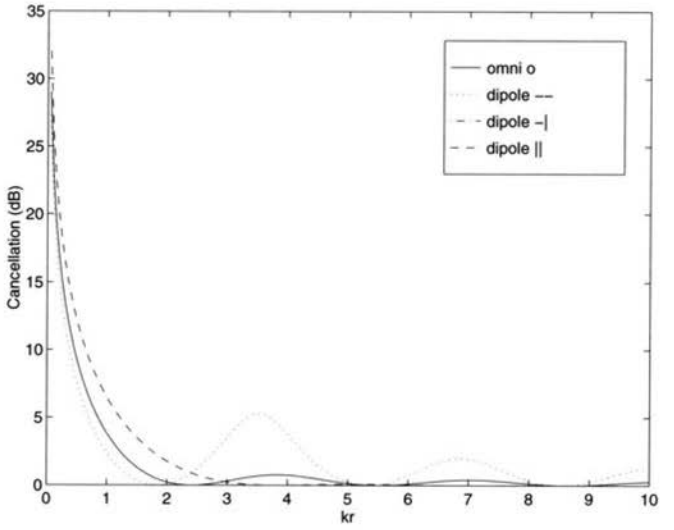

Fig. 4.13. Maximum cancellation (dB) for omnidirectional and various orientations of dipole microphones for cylindrically isotropic fields.

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

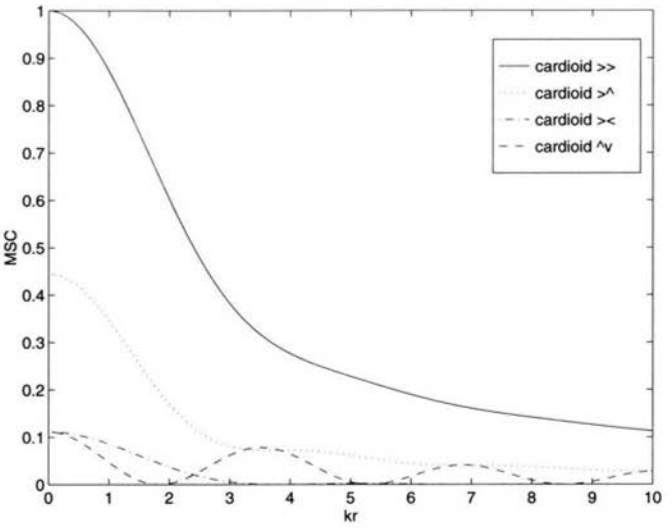

Fig. 4.14. Magnitude-squared coherence (MSC) for various orientations of cardioid microphones in a cylindrically isotropic noise field.

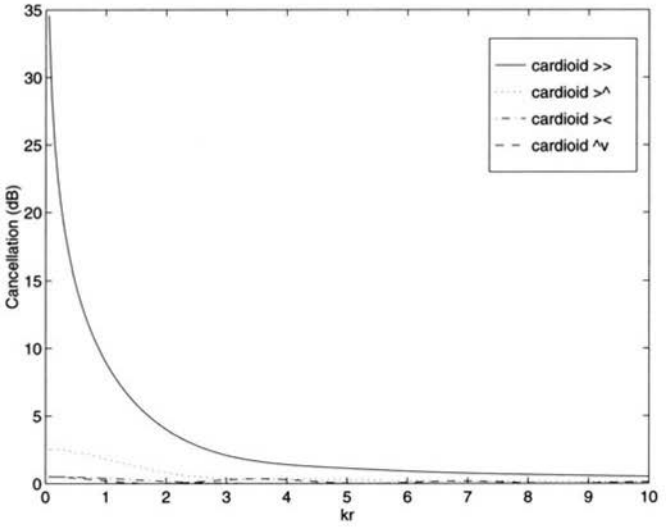

Fig. 4.15. Maximum cancellation (dB) for various orientations of cardioid microphones for cylindrically isotropic fields.

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

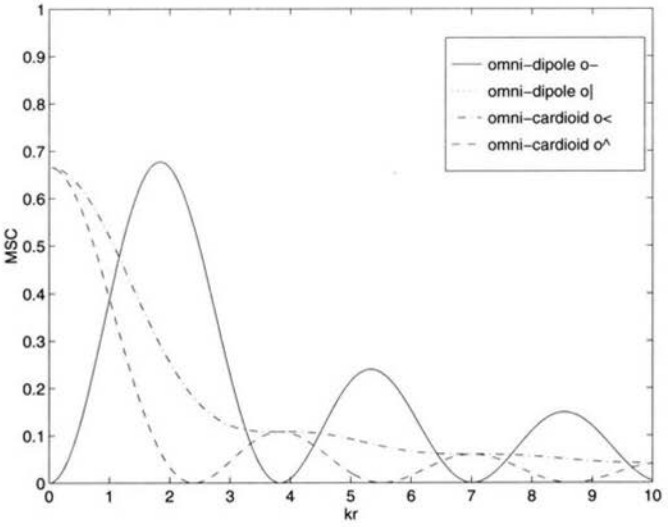

Fig. 4.16. Magnitude-squared coherence (MSC) for various orientations of omnidirectional and dipole and cardioid microphones in a cylindrically isotropic noise field.

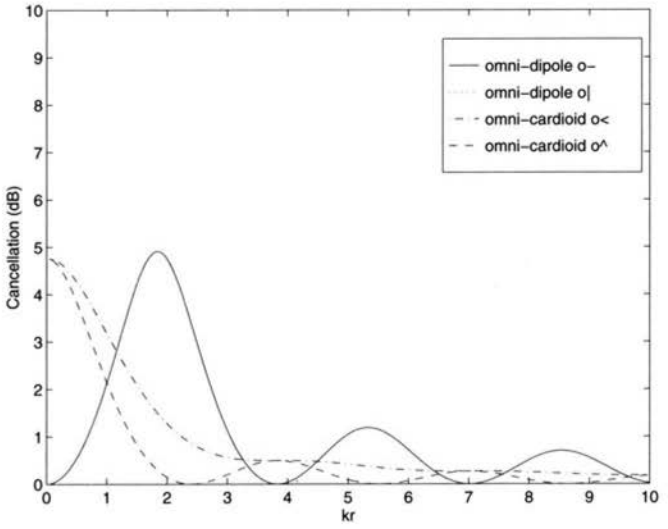

Fig. 4.17. Maximum cancellation (dB) for various orientations of omnidirectional and dipole and cardioid microphones for cylindrically isotropic fields.

Appendix A

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

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

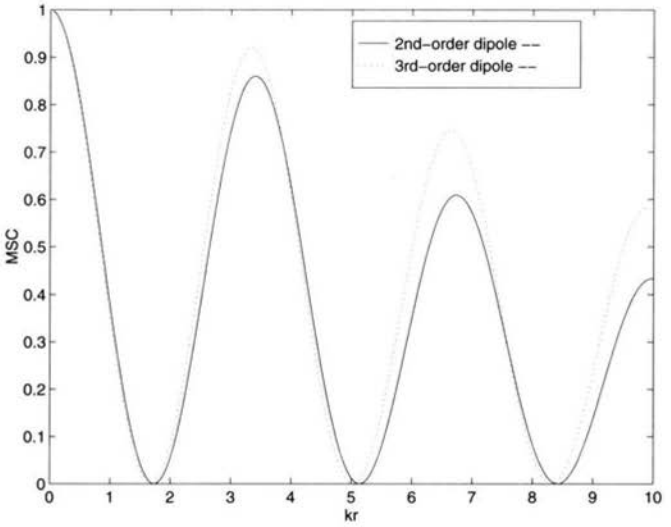

Fig. 4.18. Magnitude-squared coherence for second and third-order collinear dipoles in a cylindrically isotropic noise field.

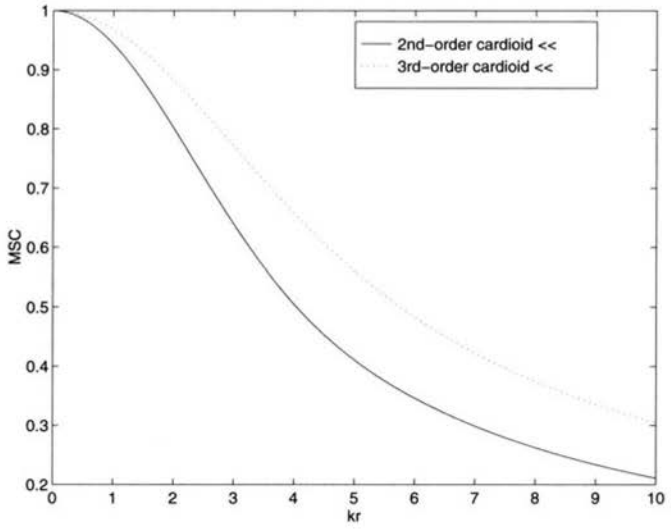

Fig. 4.19. Magnitude-squared coherence for second and third-order collinear cardioids in a cylindrically isotropic noise field.

This integral can be simplified by a change of variables,

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

Therefore,

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

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

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

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

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

Therefore the integral I_n is

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

Appendix B

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

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

Expanding the integrand yields

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

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

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

The next term involves

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

The next term involves

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

The final non-zero term involves

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

The resulting numerator for the spatial coherence function is therefore

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

Appendix C

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

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

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

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

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

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

Using the binomial theorem,

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

where the function C is the binomial coefficient [7]

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

Combining terms and invoking (4.60) yields

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

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

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

References

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

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

5 Robust Adaptive Beamforming

Osamu Hoshuyama and Akihiko Sugiyama

NEC Media Research Labs, Kawasaki, Japan

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

5.1 Introduction

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

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

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

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

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

5.2 Adaptive Beamformers

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

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

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

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

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

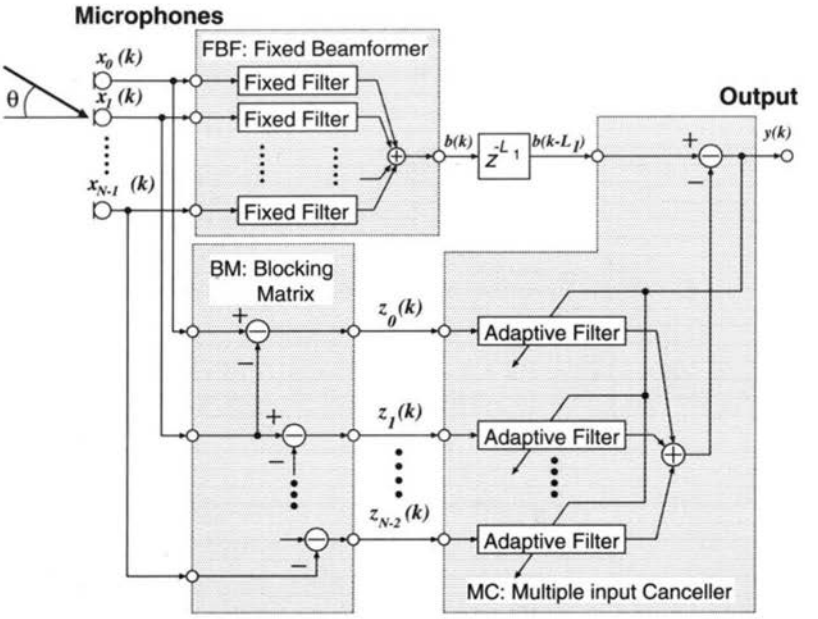

Fig. 5.1. Griffiths-Jim beamformer. It comprises a fixed beamformer (FBF), a multiple-input canceler (MC), and a blocking matrix (BM).

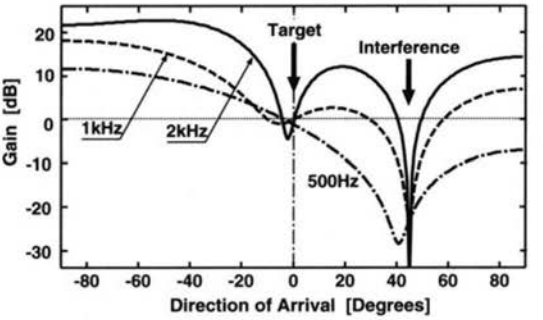

Fig. 5.2. Example directivity pattern of the Griffiths-Jim beamformer.

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

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

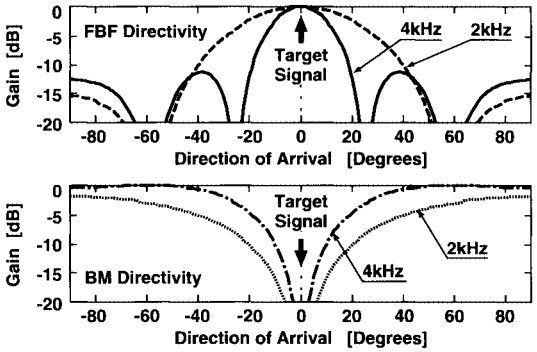

Fig. 5.3. Directivity pattern of a fixed beamformer (FBF) and a blocking matrix (BM).

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

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

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

5.3 Robustness Problem in the GJBF

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

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

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

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

Techniques to reduce target-signal leakage include:

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

Techniques that attempt to restraint excess coefficient growth include:

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

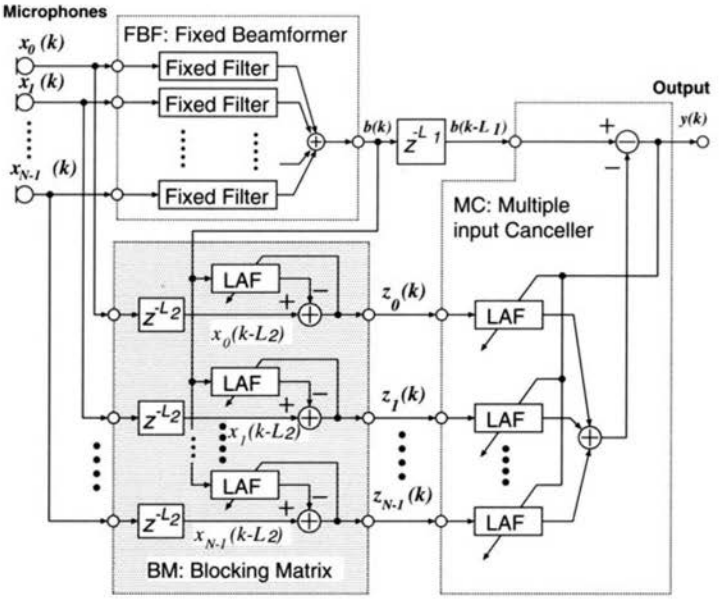

Fig. 5.4. GJBF with a LAF-LAF Structure.

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

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

5.4.1 LAF-LAF Structure

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

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

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

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

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

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

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

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

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

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

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

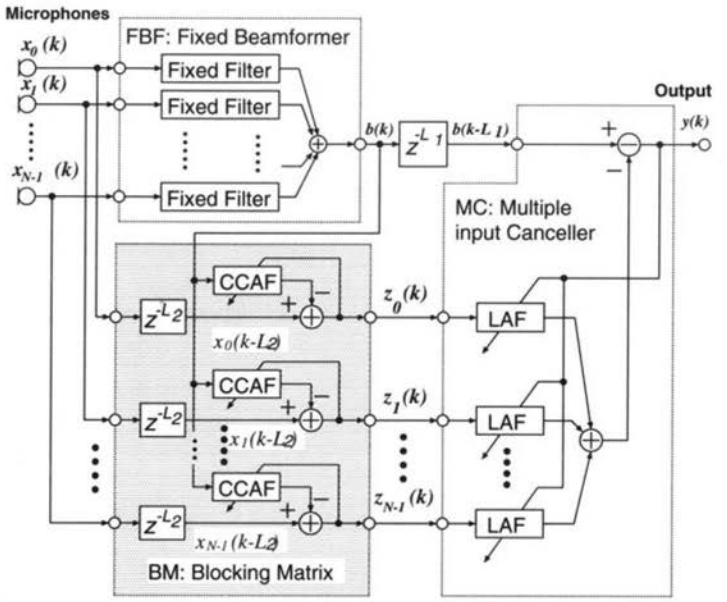

Fig. 5.5. GJBF with a CCAF-LAF Structure.

5.4.2 CCAF-LAF Structure

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

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

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

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

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

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

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

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

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

5.4.3 CCAF-NCAF Structure

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

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

where

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

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

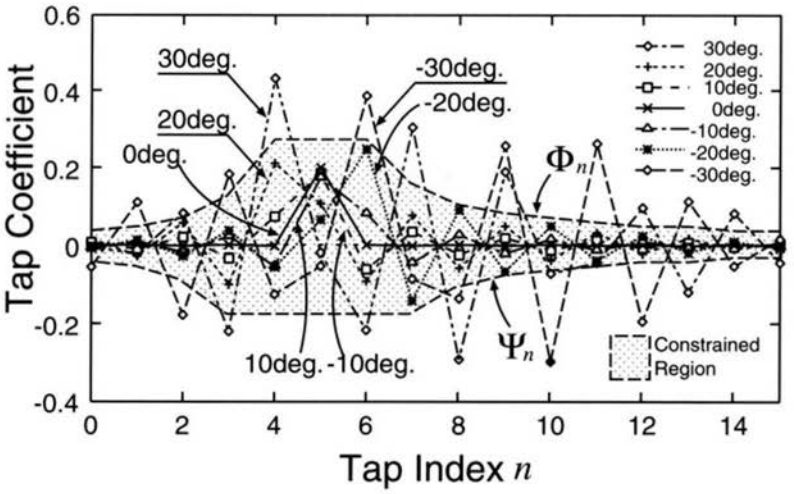

Fig. 5.6. An example of CCAF coefficients to minimize signals from different DOAs and their constraints. When the CCAF coefficients are constrained in the hatched region, up to 20° error in look direction could be allowed.

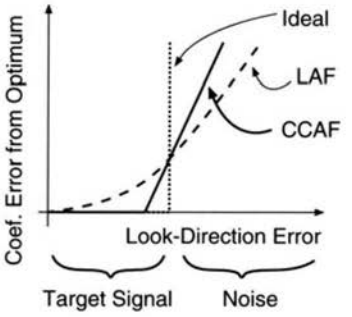

Fig. 5.7. Comparison of selectivity in LAF and CCAF.

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

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

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

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

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

5.4.4 CCAF-NCAF Structure with an AMC

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

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

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

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

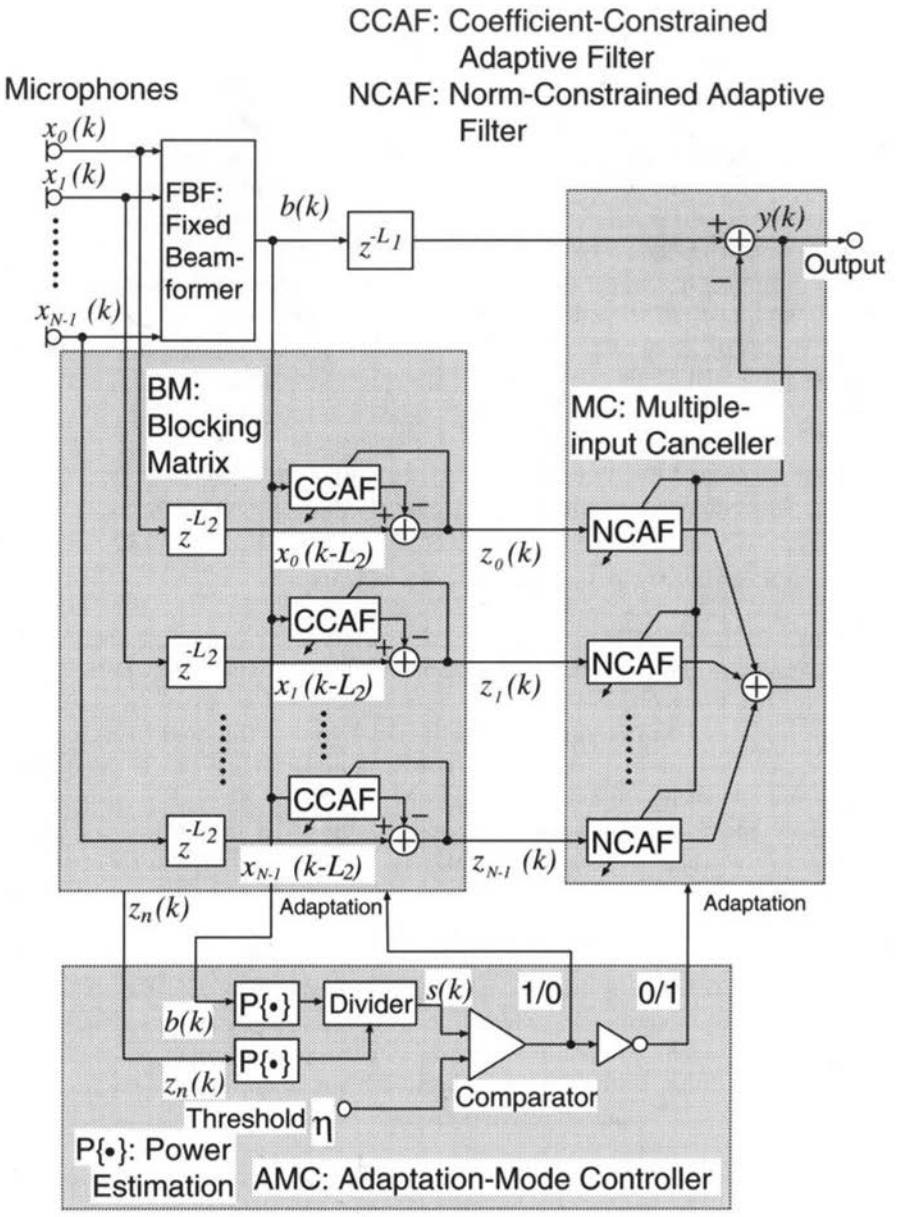

Fig. 5.8. CCAF-NCAF structure with an AMC.

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

5.5 Software Evaluation of a Robust Adaptive Microphone Array

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

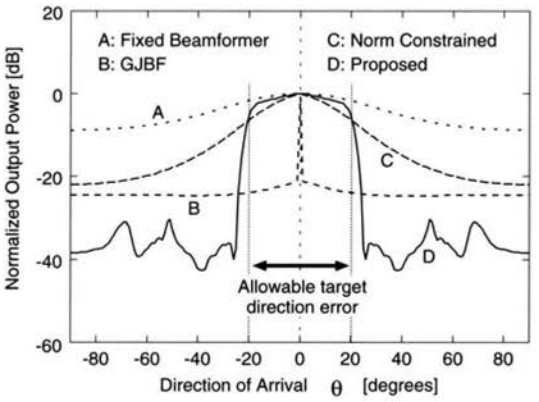

Fig. 5.9. Normalized output power after convergence as a function of DOA.

5.5.1 Simulated Anechoic Environment

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

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

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

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

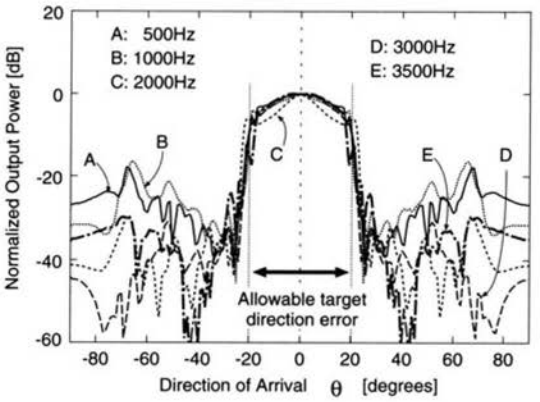

Fig. 5.10. Sensitivities after convergence as a function of DOA at different frequencies.

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

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

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

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

5.5.2 Reverberant Environment

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

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

Objective Evaluation

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

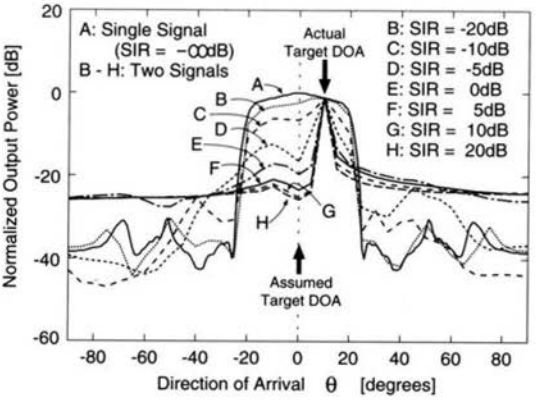

Fig. 5.11. Normalized output power after convergence as a function of DOA with different SIRs.

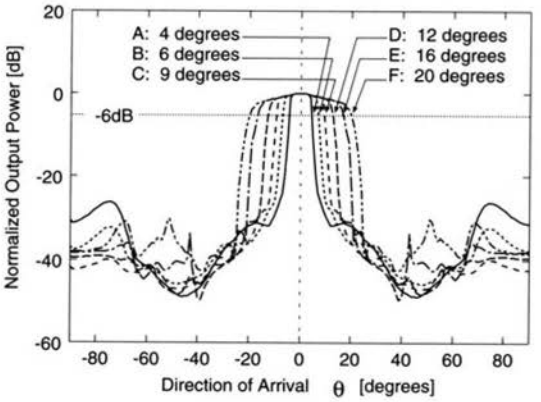

Fig. 5.12. Normalized output power after convergence for different allowable target directions.

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

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

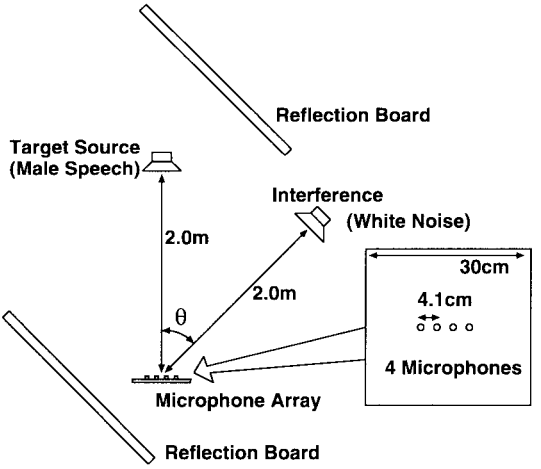

Fig. 5.13. Experimental set-up.

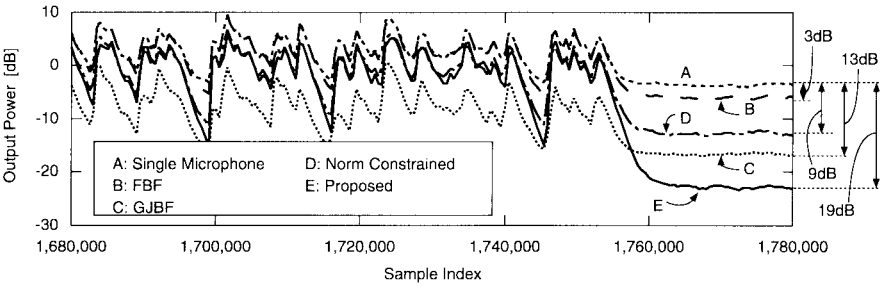

Fig. 5.14. Output Powers for a male speech signal and white noise.

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

Subjective Evaluation

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

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

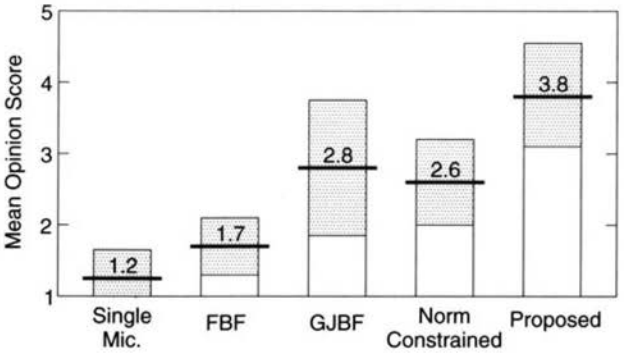

Fig. 5.15. Mean opinion score results.

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

5.6 Hardware Evaluation of a Robust Adaptive Microphone Array

5.6.1 Implementation

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

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

5.6.2 Evaluation in a Real Environment

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

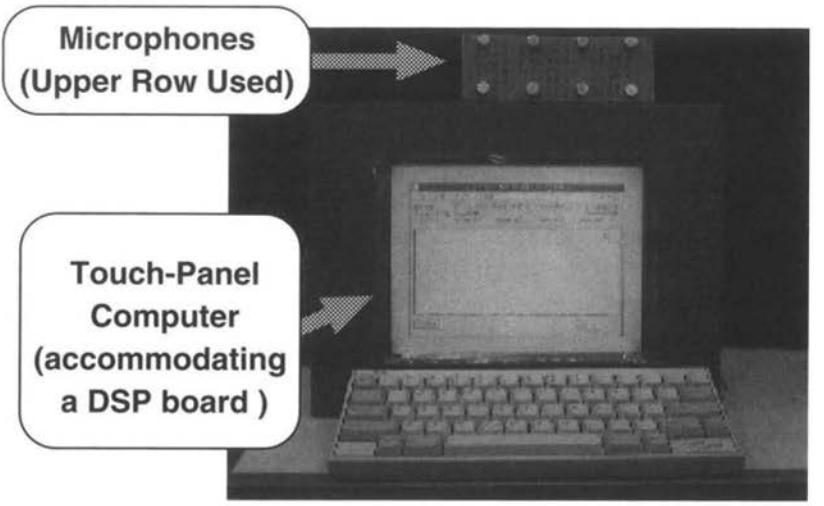

Fig. 5.16. Real-time DSP system.

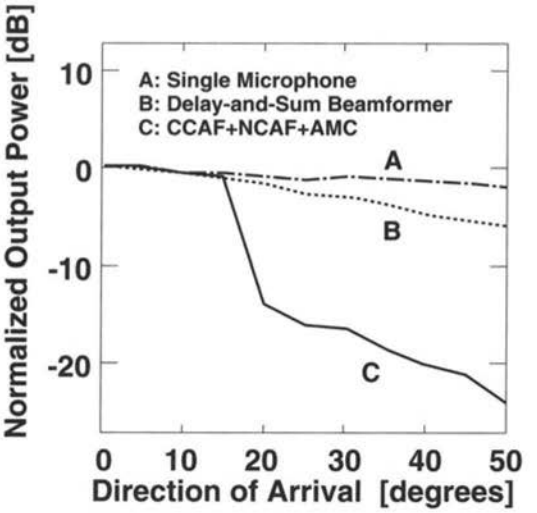

Fig. 5.17. Directivity patterns (i.e., the output powers normalized by the power at the center) measured in 5-degree intervals.

Directivity

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

Noise Reduction

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

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

5.7 Conclusion

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

References

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

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

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

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

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

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

The overall impulse response, $h_n(\boldsymbol{q}_s, t)$, is the result of cascading two filters: the room impulse response and the microphone channel response. The former characterizes all acoustic paths between the source and microphone locations, including the direct path. It is a function of $\boldsymbol{p}_n$ as well as the source location, $\boldsymbol{q}_s$, and is highly dependent on these parameters. In general, the room impulse response is affected by environmental conditions, such as temperature and humidity. It also varies with the movement of furniture and individuals inside the room. While such variations are significant, it is reasonable to assume that these factors remain constant over short periods. Hence, a room impulse response may be considered time-invariant for short periods when the source and microphone are spatially fixed. The microphone channel response accounts for the electrical, mechanical and acoustical properties of the microphone system. In general, the microphone's directivity pattern makes its response a function of its orientation as well as its spatial placement relative to the source. The additive term, $v_n(t)$, is the result of channel noise in the microphone system and any propagating ambient noise such as that due to fans or other mechanical equipment. The propagating noise is usually more significant than the channel noise and tends to dominate this term. Generally, $v_n(t)$ is assumed to be uncorrelated with $s(t)$.

Figure 8.1 illustrates a close-up view of the response that was measured in a typical conference room. The direct-path component and some of the strong reflected components are highlighted in this plot. The peaks corresponding to the reflected sound waves are comparable in size to the direct-path peak. These peaks, which occur within 20 ms of the direct-path, are responsible for many of the erroneous results produced by short-time TDE's, which operate on data blocks as small as 25 ms. The large secondary peaks in the room response are highly correlated with the false peaks in the GCC function [5].

The purpose of TDE is to evaluate the temporal disparity between the direct-path components in the two received microphone signals. To this end, it will be useful to rewrite the impulse response specifically in terms of its direct-path component. Equation 8.1 is modified to:

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

where r_n is the source-microphone separation distance, τ_n is the direct path time delay, and $g_n(\boldsymbol{q}_s, t)$ is the modified impulse response which encompasses the original response minus the direct path component. The microphone signal model is now expressed explicitly in terms of the parameter of interest, namely the time delay, τ_n.

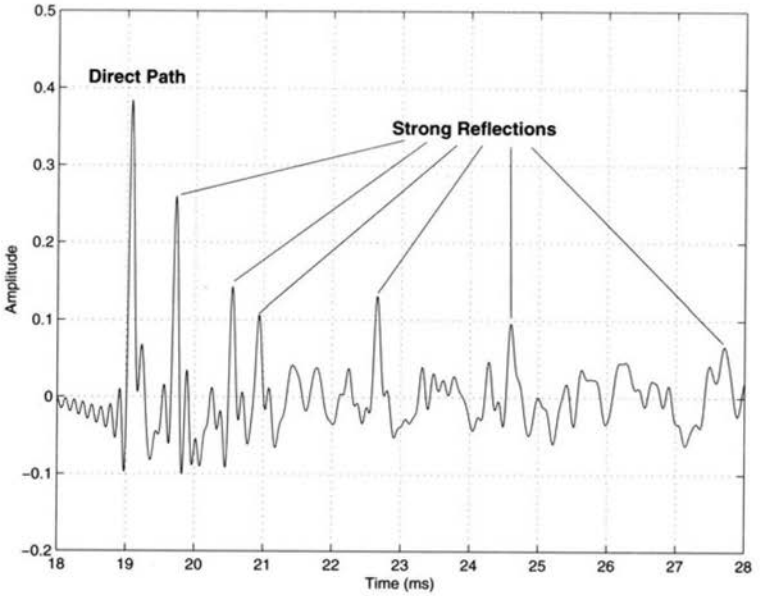

Fig. 8.1. A close-up of a 10-millisecond segment of a room impulse response measured in a typical conference room. The direct-path component and some strong reflected components are highlighted.

8.3.2 The GCC and PHAT Weighting Function

For a pair of microphones, $n = 1, 2$, their associated TDOA, τ_{12}, is defined as

$$\tau_{12} \equiv \tau_2 - \tau_1. \tag{8.3}$$

Applying this definition to their associated received microphone signal models yields

$$\begin{aligned} x_1(t) &= \frac{1}{r_1} s(t - \tau_1) \star g_1(\boldsymbol{q}_s, t) + v_1(t) \\ x_2(t) &= \frac{1}{r_2} s(t - \tau_1 - \tau_{12}) \star g_2(\boldsymbol{q}_s, t) + v_2(t). \end{aligned} \tag{8.4}$$

If the modified impulse responses for the microphone pair are similar, then (8.4) shows that a scaled version of $s(t - \tau_1)$ is present in the signal from microphone 1 and a time-shifted (and scaled) version of $s(t - \tau_1)$ is present in the signal from microphone 2. The cross-correlation of the two signals should show a peak at the time lag where the shifted versions of $s(t)$ align, corresponding to the TDOA, τ_{12}. The cross correlation of signals and is defined as:

$$c_{12}(\tau) = \int_{-\infty}^{+\infty} x_1(t) x_2(t + \tau) dt \tag{8.5}$$

The GCC function, $R_{12}(\tau)$, is defined as the cross correlation of two filtered versions of $x_1(t)$ and $x_2(t)$ [29]. With the Fourier transforms of these filters denoted by $G_1(\omega)$ and $G_2(\omega)$, respectively, the GCC function can be expressed in terms of the Fourier transforms of the microphone signals

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \big(G_1(\omega)X_1(\omega)\big)\big(G_2(\omega)X_2(\omega)\big)^* e^{j\omega\tau} d\omega \tag{8.6}$$

Rearranging the order of the signals and filters and defining the frequency dependent weighting function, $\Psi_{12} \equiv G_1(\omega)G_2(\omega)^*$, the GCC function can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{12}(\omega)X_1(\omega)X_2(\omega)^* e^{j\omega\tau} d\omega \tag{8.7}$$

Ideally, $R_{12}(\tau)$ will exhibit an explicit global maximum at the lag value which corresponds to the relative delay. The TDOA estimate is calculated from

$$\hat{\tau}_{12} = \underset{\tau \in D}{\operatorname{argmax}}\, R_{12}(\tau). \tag{8.8}$$

The range of potential TDOA values is restricted to a finite interval, D, which is determined by the physical separation between the microphones. In general, $R_{12}(\tau)$ will have multiple local maxima which may obscure the true TDOA peak and subsequently, produce an incorrect estimate. The amplitudes and corresponding time lags of these erroneous maxima depend on a number of factors, typically ambient noise levels and reverberation conditions.

The goal of the weighting function, Ψ_{12}, is to emphasize the GCC value at the true TDOA value over the undesired local extrema. A number of such functions have been investigated. As previously stated, for realsitic acoustical conditions the PHAT weighting [29] defined by

$$\Psi_{12}(\omega) \equiv \frac{1}{|X_1(\omega)X_2^*(\omega)|} \tag{8.9}$$

has been found to perform considerably better than its counterparts designed to be statistically optimal under specific non-reverberant, noise conditions. The PHAT weighting whitens the microphone signals to equally emphasize all frequencies. The utility of this strategy and its extension to steered-beamforming form the basis of the SRP-PHAT algorithm that follows.

8.3.3 ML TDOA-Based Source Localization

Consider the 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, $\boldsymbol{q}_s$, the true TDOA relative to the i^{th} sensor pair will be denoted by $T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s)$, and is calculated from the expression

$$T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s) = \frac{|\boldsymbol{q}_s - \boldsymbol{p}_{i2}| - |\boldsymbol{q}_s - \boldsymbol{p}_{i1}|}{c} \tag{8.10}$$

where c is the speed of sound in air. The estimate of this true TDOA, the result of a TDE procedure involving the signals received at the two microphones, will be given by $\hat{\tau}_i$. In practice, the TDOA estimate is a corrupted version of the true TDOA and in general, $\hat{\tau}_i \neq T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s)$.

For a single microphone pair and its TDOA estimate, the locus of potential source locations in 3-space which satisfy (8.10) corresponds to one-half of a hyperboloid of two sheets. This hyperboloid is centered about the midpoint of the microphones and has $\boldsymbol{p}_{i2} - \boldsymbol{p}_{i1}$ as its axis of symmetry.

For sources with a large source-range to microphone-separation ratio, the hyperboloid may be well-approximated by a cone with a constant direction angle relative to the axis of symmetry. The corresponding estimated direction angle, $\hat{\theta}_i$, for the microphone pair is given by:

$$\hat{\theta}_i = \cos^{-1}\left(\frac{c \cdot \hat{\tau}_i}{|\mathbf{m}_{i1} - \mathbf{m}_{i2}|}\right) \tag{8.11}$$

In this manner each microphone pair and TDOA estimate combination may be associated with a single parameter which specifies the angle of the cone relative to the sensor pair axis. For a given source and TDOA estimate, $\hat{\theta}_i$ is referred to as the DOA relative to the 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, $\boldsymbol{q}_s$. Ideally, the estimate will be an element of the intersection of all the potential source loci. In practice, however, for more than two pairs of sensors this intersection is, in general, the empty set. This disparity is due in part to imprecision in the knowledge of system parameters (TDOA estimate and sensor location measurement errors) and in part to unrealistic modeling assumptions (point source radiator, ideal medium, ideal sensor characteristics, etc.). With no ideal solution available, the source location must be estimated as the point in space which best *fits* the sensor-TDOA data or more specifically, minimizes an error criterion that is a function of the given data and a hypothesized source location. If the time-delay estimates at each microphone pair are assumed to be independently corrupted by zero-mean additive white Gaussian noise of equal variance then the ML location estimate can be shown to be the position which minimizes the least squares error criterion

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

The location estimate is then found from

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

The criterion in (8.12) will be referred to as the LS-TDOA error. As stated earlier, the evaluation of $\hat{\boldsymbol{q}}_s$ in this manner involves the optimization of a non-linear function and necessitates the use of search methods. Closed-form approximations to this method were given earlier.

8.3.4 SRP-Based Source Localization

The microphone signal model in (8.2) shows that for an array of N microphones in the reception region of a source, a delayed, filtered, and noise corrupted version of the source signal, $s(t)$, is present in each of the received microphone signals. The delay-and-sum beamformer time aligns and sums together the $x_n(t)$, in an effort to preserve unmodified the signal from a given spatial location while attenuating to some degree the noise and convolutional components. It is defined as simply as

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

where Δ_n are the *steering delays* appropriate for focusing the array to the source spatial location, $\boldsymbol{q}_s$, and compensating for the direct path propagation delay associated with the desired signal at each microphone. In practice, the delays relative to a reference microphone are used instead of the absolute delays. This makes all shifting operations causal, which is a requirement of any practical system, and implies that $y(t, \boldsymbol{q}_s)$ will contain an overall delayed version of the desired signal which in practice is not detrimental. The use of a single reference microphone means that the steering delays may be determined directly from the TDOA's (estimated or theoretical) between each microphone and the reference. This implies that knowledge of the TDOA's alone is sufficient for steering the beamformer without an explicit source location.

In the most ideal case with no additive noise and channel effects, the output of the deal-and-sum beamformer represents a scaled and potentially delayed version of the desired signal. For the limited case of additive, uncorrelated, and uniform variance noise and equal source-microphone distances this simple beamformer is optimal. These are certainly very restrictive conditions. In practice, convolutional channel effects are nontrivial and the additive noise is more complicated. The degree to which these noise and reverberation components of the microphone signals are suppressed by the delay-and-sum beamformer is frequently minimal and difficult to analyze. Other methods have been developed to extend the delay-and-sum concept to the more general filter-and-sum approach, which applies adaptive filtering to the microphone

signals before they are time-aligned and summed. Again, these methods tend to not be robust to non-theoretical conditions, particularly with regard to the channel effects.

The output of an N-element, filter-and-sum beamformer can be defined in the frequency domain as

$$Y(\omega, \boldsymbol{q}) = \sum_{n=1}^{N} G_n(\omega) X_n(\omega) e^{j\omega \Delta_n} \tag{8.15}$$

where $X_n(\omega)$ and $G_n(\omega)$ are the Fourier Transforms of the 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,

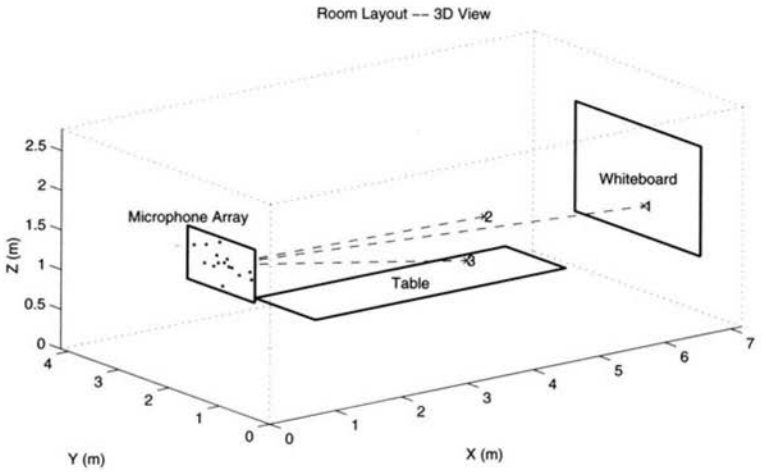

Fig. 8.2. Conference room layout.

$P(\boldsymbol{q})$ is maximized over a region of potential source locations. As will be shown in the next section, relative to the search space indicative of the standard SRP approach, the SRP-PHAT functional significantly deemphasizes extraneous peaks and dramatically sharpens the resolution of the true peak. These desirable features result in a decreased sensitivity to noise and reverberations and more precise location estimates than the existing localization methods offer. Additionally, this is achieved using a very short analysis interval.

8.4 Experimental Comparison

While more extensive results are available in [5], an experiment is offered here to evaluate and compare the relative characteristics and performance of three different source locators: SRP, SRP-PHAT and ML-TDOA. Five second recordings were made for three source locations in a 7 by 4 by 3 m conference room at Brown University using a 15-element microphone array. Figure 8.2 illustrates the room layout. Pre-recorded speech, which was acquired using a close-talking microphone, was played through a loudspeaker while simultaneously recording the signals from the array. The use of the loudspeaker was preferable to an actual talker since the loudspeaker could be precisely located and would be fixed over the duration of the recordings. The talkers were males uttering a unique string of alpha-digits. Source 1 was most distant from the array and was positioned at standing height in front of a white-board. The other two sources were positioned at a seated level

around a conference table, which was located approximately in the center of the room.

The microphone array was composed of eight omni-directional electret condenser microphones, which were randomly distributed on a plane within a .33 by 0.36 m rectangle. The microphones were attached to a rectangular sheet of acoustic foam, which was supported by an aluminum frame. This frame was mounted on a tripod that was placed parallel to the back wall at a distance of 0.9 m. The acoustic foam damps some of the multi-path reflections from this wall and isolates the microphones from vibrations traveling along the mountings.

The loudspeaker faced the array and the volume level was adjusted at each location to maximize SNR conditions. SNR levels at each microphone averaged about 25 dB for the three source locations. Source 3, with its location the closest to the microphone array, had SNRs as high as 36 dB. With such high SNRs, all microphones signals in the conference room dataset have minimal contributions from the background noise, which was primarily produced by the fans inside the computer equipment.

The measured reverberation time of the room was determined to be 200 ms. This qualifies as a mildly reverberant room. However, the near-end peaks in the impulse responses (as in Figure 8.1) combined with a 200 ms reverberation time do, in fact, have a significant impact on localization. This will be demonstrated by the following performance comparisons.

Given the size of the array aperture relative to the source ranges, all three talkers can be considered to lie in the far field of array. Under such conditions, range estimates are ambiguous, and only the azimuth and elevation angles can be estimated reliably. Accordingly, this experiment will focus on DOA measures as opposed to 3-D Cartesian coordinates. Results obtained with more extensive arrays and near-field sources are available in [5].

The recorded data was segmented into 25 ms frames using a half-overlapping Hanning window. SNR-based speech detection was performed for each frame. All frames where any of the eight microphone channels had SNR within 12dB of the background noise were eliminated. Out of the 399 frames per recording, 313, 340, and 297 were retained for sources 1,2, and 3, respectively. The DOA's of the sources were estimated by minimization of the LS-TDOA error and maximization of SRP and SRP-PHAT evaluated over azimuth and elevation relative to the array's origin. The frequency range used to compute both the steered responses and the GCC's was 300 Hz to 8 kHz. These functions were computed over a range of $-60°$ to $+60°$ for both azimuth and elevation with a $0.1°$ resolution.

By taking all possible combinations, 28 microphone pairs were formed using the 8-element array. Hence, for each data frame, 28 TDOA estimates were made for each of the three speech recordings using GCC-PHAT. Figure 8.3 illustrates the LS-TDOA error as a function of azimuth and elevation for a segment of nine successive frames recorded for source 1. The white point in

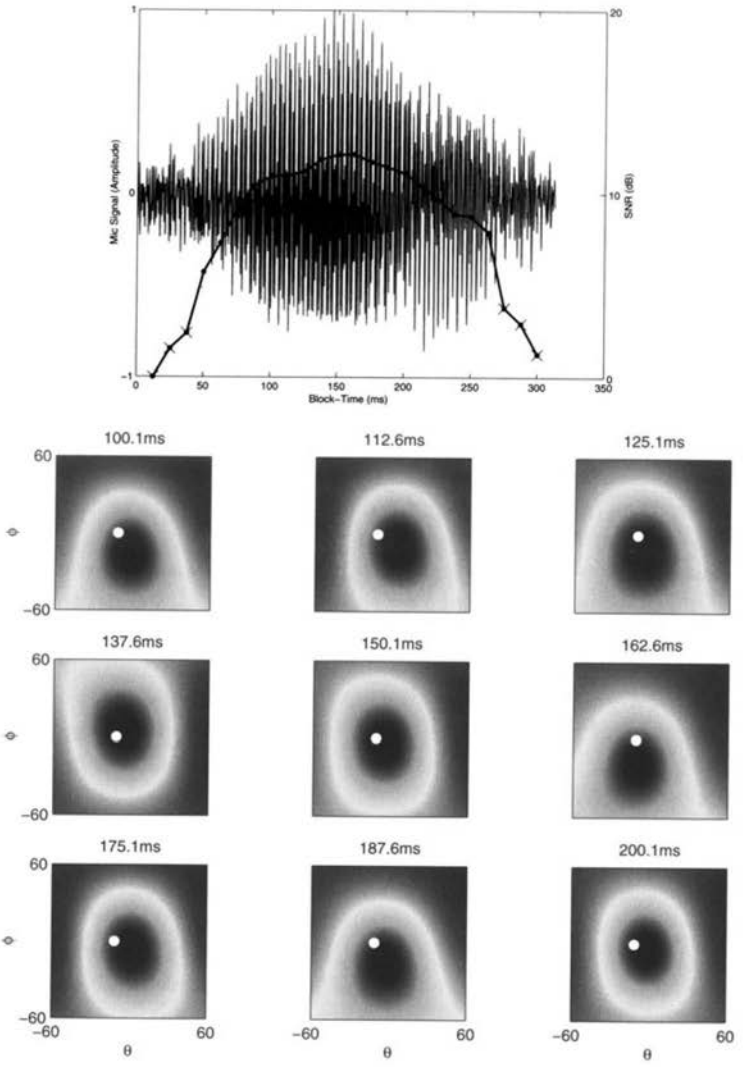

Fig. 8.3. Speech segment (top) with nine frames of the LS-TDOA error surfaces.

each contour plot marks the true DOA. The dark area in the center of the images represents the minima of the LS-TDOA error. At the top of this figure is a plot of the amplitude of the corresponding speech segment, which is the letter "R", spoken as in "Are we there yet?" Superimposed on this speech signal is a curve representing the average power of the signals from the array, with the scale of its vertical axis labeled on the right side of the graph. Each point along this power curve corresponds to the average frame SNR. The three frames at the beginning and end of this speech segment

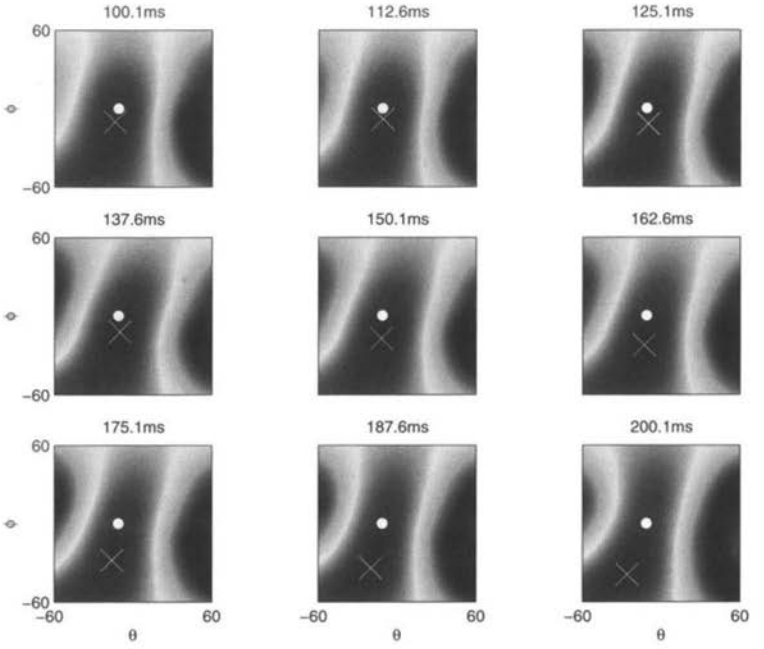

Fig. 8.4. Delay-and-sum beamformer SRP over nine, 25 ms frames.

lacked sufficient SNR to included in the analysis. These plots show that the LS-TDOA error is generally a smooth surface with a global minimum over the angular range of $\pm 60°$. However, from frame to frame the minima vary from the true source location. This inaccuracy is caused by erroneous TDOA estimates. Note also that because of the smooth nature of the error space, the resolution of the DOA estimates is considerably limited.

Figures 8.4 and 8.5 illustrate the error spaces of the SRP and SRP-PHAT as evaluated for the same nine 25 ms frames of speech. Relative to the prior figure the contour images are now inverted in darkness to emphasize the maxima. The plots of the delay-and-sum beamformer SRP in Figure 8.4 bear a noticeable similarity in general shape to their LS-TDOA counterparts. The maximum value in each SRP image, marked by an X, occurs at points distant from the actual DOA, indicated by a white dot. The main beam of the delay-and-sum beamformer is broad and fluctuates considerably over the duration of the speech segment. As a result, many inaccurate location estimates are produced by this method. In contrast to the LS-TDOA and SRP cases, the peaks of SRP-PHAT plots in Figure 8.5 match the actual DOA almost exactly. The main beam of the PHAT beamformer is sharp and consistent over each frame. This produces contour images which appear quite different from the LS-TDOA and SRP versions. The PHAT filters, when applied to the filter-and-sum beamformer, yield an error space that is superior to that of

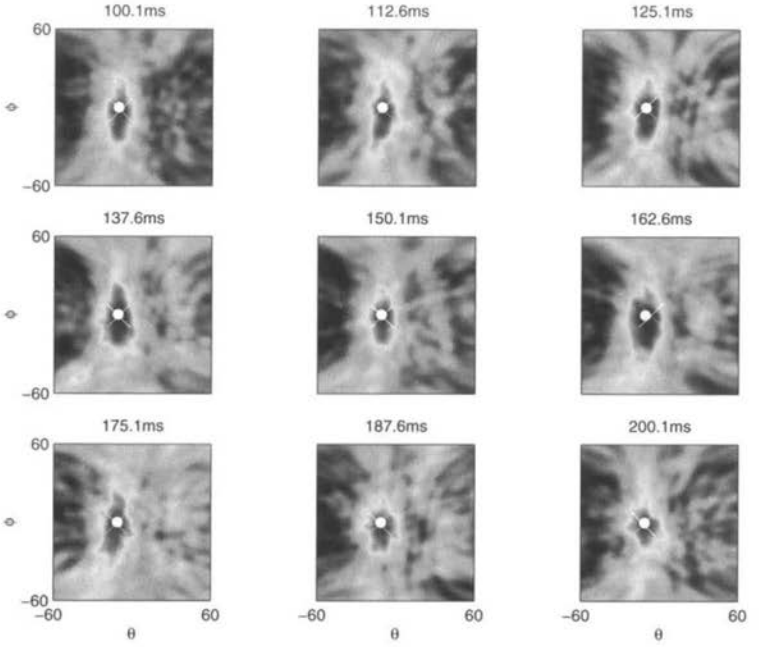

Fig. 8.5. SRP-PHAT response over nine, 25 ms frames.

the delay-and-sum beamformer or the TDOA-based criterion. This qualitative observation will now be corroborated through a numerical performance comparison.

For the DOA estimates produced for each of the three source locations, an RMS DOA error was computed from

$$E_{\text{RMS}}(\hat{\theta}, \hat{\phi}) = \sqrt{(\hat{\theta} - \theta)^2 + (\hat{\phi} - \phi)^2} \tag{8.22}$$

where ϕ and θ are the true azimuth and elevation angles and $\hat{\phi}$ and $\hat{\theta}$ are their estimated counterparts. Figure 8.6 illustrates the results. These plots show the fraction of DOA estimates in each case which exceed a given RMS error threshold. Using this metric, the SRP-PHAT consistently outperforms the other two methods for each of the source locations. The ML-TDOA exhibits definite advantages over the SRP. While the SRP-PHAT's results are nearly identical for all the source locations, including the most distant source 1, the ML-TDOA locator is highly dependent on source location. For example, 60% percent of the estimates from source 1 had error greater than 10° while 50% percent from source 2 and 15% percent from source 3 had error greater 10°. In contrast, nearly all the estimates produced by SRP-PHAT had error less than 10°. About 90% of the estimates from sources 2 and 3, and 80% from source 1 had errors less than 4°.

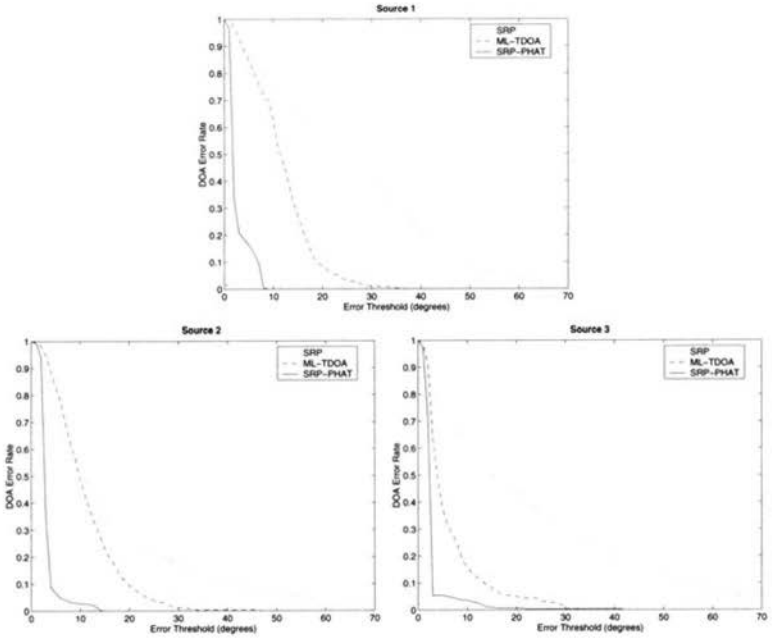

Fig. 8.6. Localizer DOA error rates for three different sources.

The results of this limited experiment illustrate the performance advantages of the SRP-PHAT localizer relative to more traditional approaches for talker localzation with microphone arrays. Other experiments conducted under more general and adverse conditions are consistent with the results here and serve to confirm the utility of combining steered-beamforming and a uniform-magnitude spectral weighting for this purpose.

While the TDOA-based localization method performed satisfactorily for a talker relatively close to the array, it was severely impacted by even the mild reverberation levels encountered when the source was more distant. This result is due to the fact that signal-to-reverberation ratios decrease with increasing source-to-microphone distance. As the reverberation component of the received signal increases relative to the direct path component, the validity of the single-source model inherent in the TDE development is no longer valid. As a result TDOA-based schemes rapidly exhibit poor performance as the talker moves away from the microphones. The SRP-PHAT algorithm is relatively insensitive to this effect. As the results here suggest the proposed algorithm exhibits no marked performance degradation from the near to distant source conditions tested.

The SRP-PHAT algorithm is computationally more demanding than the TDOA-based localization methods. However, its significantly superior performance may easily warrant the additional processing expense. Additionally,

while not discussed here, it is possible to alter the algorithm to dramatically reduce its computational load while maintaining much of its benefit.

References

1. H. Silverman, W. Patterson, J. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 251–254, April 1997.
2. M. Brandstein, J. Adcock, and H. Silverman, "Microphone array localization error estimation with application to sensor placement," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3807–3816, 1996.
3. J. Flanagan and H. Silverman, eds., *International Workshop on Microphone-Array Systems: Theory and Practice*, Brown University, Providence RI, USA, October 1992.
4. B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 881–884, March 1999.
5. J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*, PhD thesis, Brown University, Providence RI, USA, May 2000.
6. W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," in *Signal Processing* (J. Griffiths, P. Stocklin, and C. V. Schooneveld, eds.), pp. 577–590, Academic Press, 1973.
7. G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Am.*, vol. .62, pp. 922–926, October 1977.
8. W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform Theory*, vol. IT-19, pp. 608–614, September 1973.
9. W. Hahn, "Optimum signal processing for passive sonar range and bearing estimation," *J. Acoust. Soc. Am.*, vol. 58, pp. 201–207, July 1975.
10. M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 1210–1217, October 1983.
11. V. M. Alvarado, *Talker Localization and Optimal Placement of Microphones for a Linear MIcrophone Array using Stochastic Region Contraction.* PhD thesis, Brown University, Providence RI, USA, May 1990.
12. H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone-array data," *Computer, Speech, and Language*, vol. 6, pp. 129–152, April 1992.
13. D. Johnson and D. Dudgeon, *Array Signal Processing- Concepts and Techniques*, Prentice Hall, 1993.
14. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, second ed., 1991.
15. R. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*, PhD thesis, Stanford University, Stanford CA, USA, 1981.
16. J. Krolik, "Focussed wide-band array processing for spatial spectral estimation," in *Advances in Spectrum Analysis and Array Processing* (S. Haykin, ed.), vol. 2, pp. 221–261, Prentice Hall, 1991.

17. H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 823–831, August 1985.
18. K. Buckley and L. Griffiths, "Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. .36, pp. 953–964, July 1988.
19. A. Vural, "Effects of pertubations on the performance of optimum/adaptive arrays," *IEEE Trans. Aerosp. Electron.*, vol. AES-15, pp. 76–87, January 1979.
20. R. Compton Jr., *Adaptive Antennas*, Prentice Hall, 1988.
21. T. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation in coherent signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 806–811, August 1985.
22. M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 45–50, January 1997.
23. P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 231–234, April 1997.
24. Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 937–940, March 1999.
25. R. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. Aerosp. Electron.*, vol. AES-8, pp. 821–835, November 1972.
26. J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1661–1669, December 1987.
27. H. Lee, "A novel procedure for accessing the accuracy of hyperbolic multilateration systems," *IEEE Trans. Aerosp. Electron.*, vol. AES-11, pp. 2–15, January 1975.
28. N. Marchand, "Error distributions of best estimate of position from multiple time difference hyperbolic networks," *IEEE Trans. Aerosp. Navigat. Electron.*, vol. .11, pp. 96–100, June 1964.
29. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-24, pp. 320–327, August 1976.
30. M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer, Speech, and Language*, vol. 9, pp. 153–169, April 1995.
31. M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, April 1997.
32. S. Bédard, B. Champagne, and A. Stéphenne, "Effects of room reverberation on time-delay estimation performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-94)*, Adelaide, Australia, pp. II:261–264, April 1994.
33. M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 375–378, April 1997.

34. H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 187–190, April 1997.
35. M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event localization," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 288–292, May 1997.
36. P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 231–234, April 1997.
37. M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.
38. A. Stéphenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 3055–3058, May 1995.
39. N. Strobel and R. Rabenstein, "Classification of time delay estimates in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 3081–3084, March 1999.
40. B. Friedlander, "A passive localization algorithm and its accuracy analysis," *IEEE Jour. Oceanic Engineering*, vol. OE-12, pp. 234–245, January 1987.
41. Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Processing*, vol. 42, pp. 1905–1915, August 1994.
42. D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 371–374, April 1997.
43. L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *Fundamentals of Acoustics*, John Wiley & Sons, third ed., 1982.
44. L. Ziomek, *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*, CRC Press, 1995.

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.

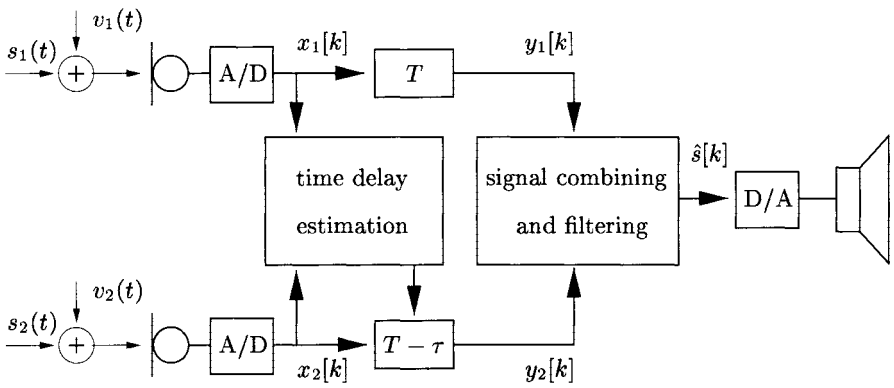

Fig. 12.1. Block diagram of the symmetric two-microphone noise reduction system.

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.

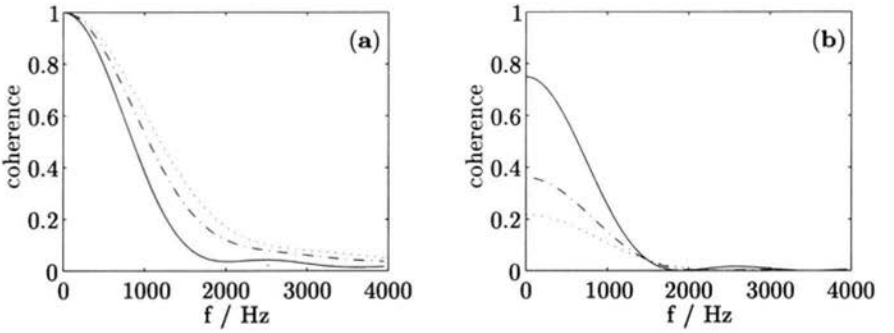

Fig. 12.2. Magnitude squared coherence of two microphone signals in an ideally diffuse noise field [17]. The microphones have a cardioid directivity pattern and different look directions. The microphone distance is 0.1 m. 0° equals broadside orientation.
(**a**): both microphones are turned clockwise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

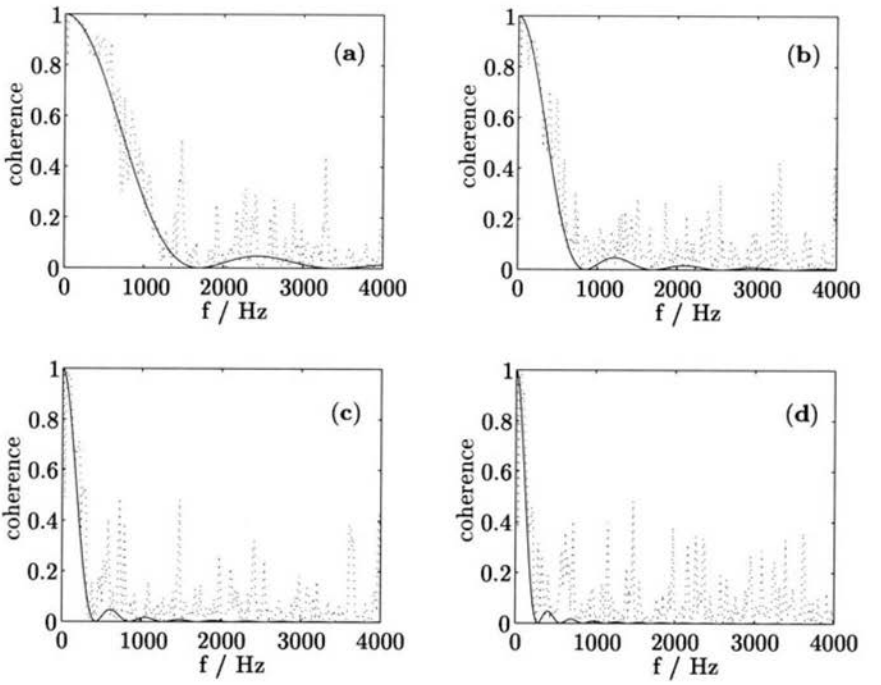

Fig. 12.3. Coherence of the ideally diffuse sound field (solid) and measured coherence (dotted) of office noise for omnidirectional microphones and microphone distances $d_{\text{mic}} = 0.1\,\text{m}$ (**a**), $d_{\text{mic}} = 0.2\,\text{m}$ (**b**), $d_{\text{mic}} = 0.4\,\text{m}$ (**c**), and $d_{\text{mic}}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

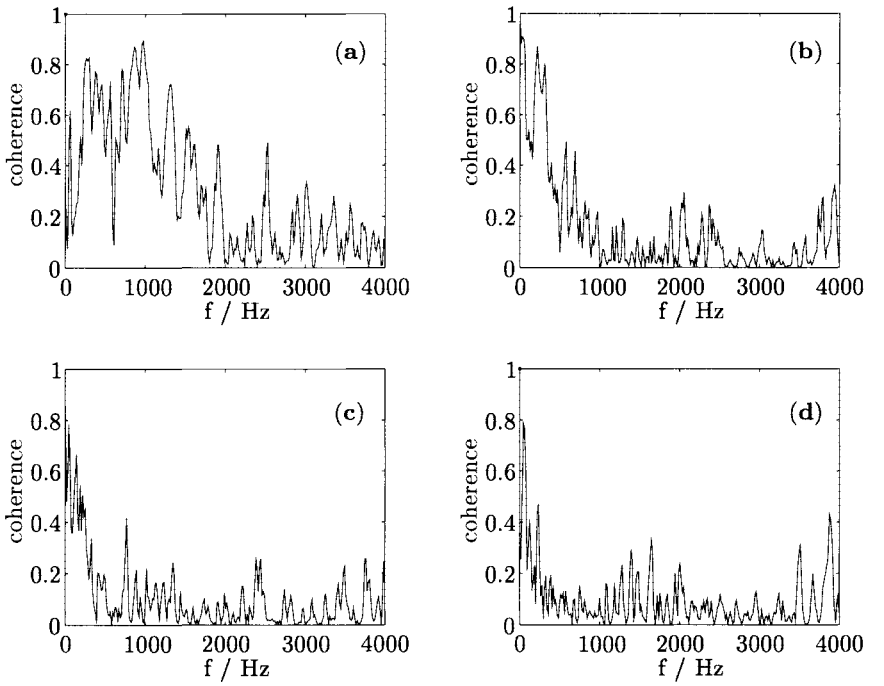

Fig. 12.4. Measured coherence of noise in a car for hypercardioid microphones and microphone distances $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$ (**a**), $d_{\mathrm{mic}} = 0.2\,\mathrm{m}$ (**b**), $d_{\mathrm{mic}} = 0.4\,\mathrm{m}$ (**c**), and $d_{\mathrm{mic}} = 0.6\,\mathrm{m}$ (**d**).

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,

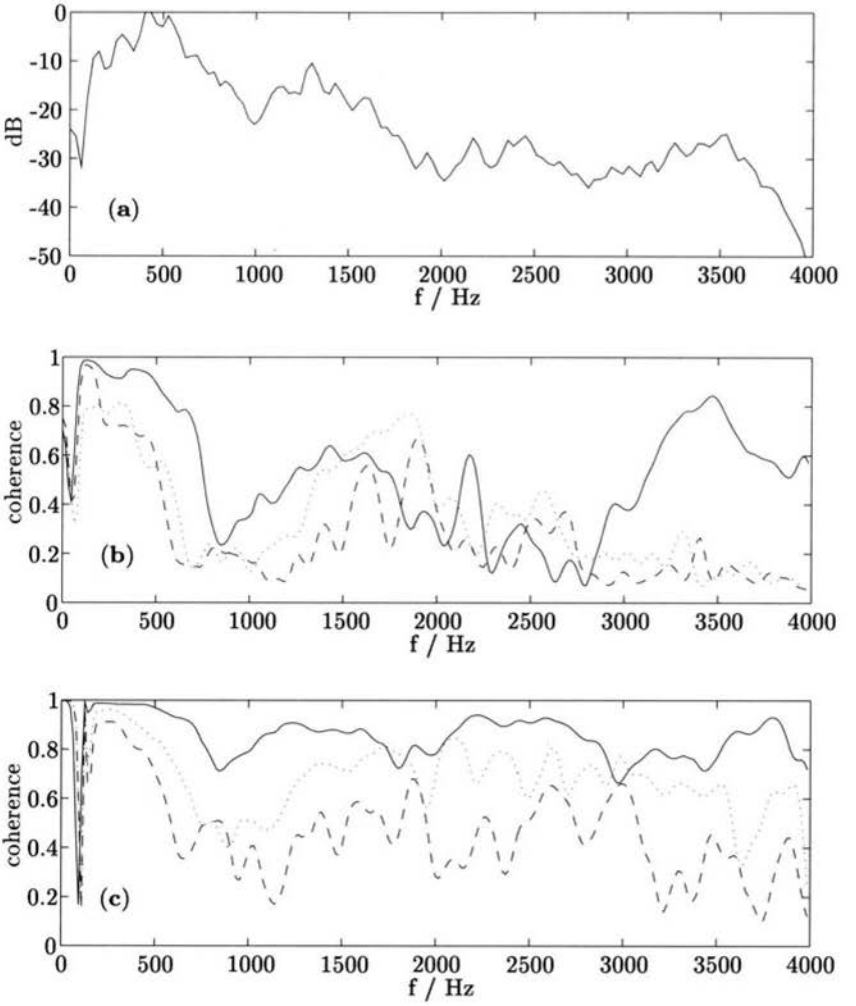

Fig. 12.5. Long term speech spectrum (**a**) and long term average of the MSC of a speech signal in an office room for omnidirectional (**b**) and for hypercardioid (**c**) microphones and various distances ((——): 0.5 m; (·····): 1 m; (– – –): 2 m) from speaker to microphones. The distance between microphones 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. $H_{\text{sopt}}(\Omega) = P_{s1s2}(\Omega)/P_{x2x2}(\Omega)$ and $H_{\text{vopt}}(\Omega) = P_{v1v2}(\Omega)/P_{x2x2}(\Omega)$ are linear MMSE estimators for the speech component $s_1[k]$ and the (undesired) noise component $v_1[k]$ of the reference signal $x_1[k]$, respectively. The estimation error between the desired speech signal $s_1[k]$ and the output of the Wiener filter $\widehat{s}[k]$ can therefore be written in terms of the estimation errors of the subfilters

$$e_r[k] = s_1[k] - \widehat{s}[k] = \left(s_1[k] - \widehat{\widehat{s}}_1[k]\right) - \widehat{\widehat{v}}_1[k]\,. \tag{12.6}$$

The overall minimum mean square error can be then decomposed into

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

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

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

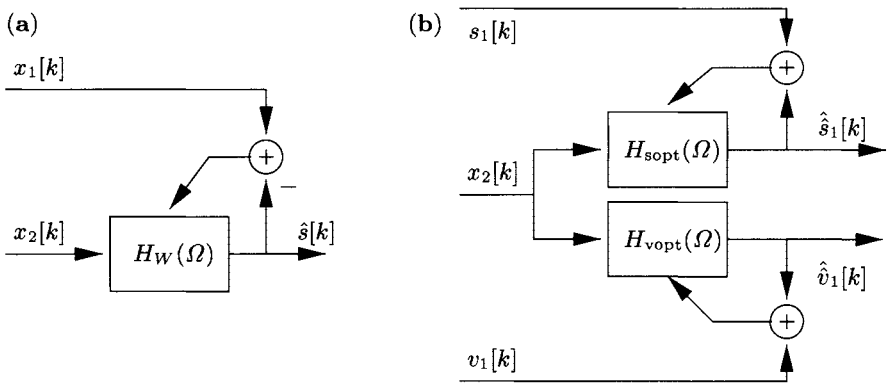

Fig. 12.6. The Wiener filter (**a**) and its subfilter decomposition (**b**).

where we used $E\left\{\widehat{\widehat{v}}_1^2[k]\right\} = E\left\{v_1[k]\widehat{\widehat{v}}_1[k]\right\}$ which holds for the MMSE filter.

Using Parceval's relation, the minimum mean square errors of the unconstrained subfilters $H_{\text{sopt}}(\Omega)$ and $H_{\text{vopt}}(\Omega)$ can be rewritten in the frequency domain as [18]

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

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1v_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{\text{vopt}}(\Omega)\, P^*_{v_1v_2}(\Omega)\, d\Omega \quad (12.10)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1v_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1v_2}(\Omega)|^2}{P_{s_2s_2}(\Omega) + P_{v_2v_2}(\Omega)} d\Omega \quad (12.11)$$

and

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

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1s_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{\text{sopt}}(\Omega)\, P^*_{s_1s_2}(\Omega)\, d\Omega \quad (12.12)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1s_1}(\Omega)\, d\Omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{s_1s_2}(\Omega)|^2}{P_{s_2s_2}(\Omega) + P_{v_2v_2}(\Omega)} d\Omega \quad (12.13)$$

where $P_{xy}(\Omega)$ denotes the (cross) power spectral density of the signals in the subscript.

Applying these results to (12.9), we obtain for the overall MMSE

$$E\left\{e_r^2[k]\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1 s_1}(\Omega)\, d\Omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1 v_2}(\Omega)|^2 - |P_{s_1 s_2}(\Omega)|^2}{P_{s_2 s_2}(\Omega) + P_{v_2 v_2}(\Omega)} d\Omega\,. \tag{12.14}$$

The representation of the MMSE in the frequency domain shows that a successful application (i.e. a small MMSE) of the two-microphone Wiener filter to the speech enhancement task requires a high correlation of the speech components $s_1[k]$ and $s_2[k]$ and a low correlation of the noise components $v_1[k]$ and $v_2[k]$. In the following sections we consider two special cases.

12.3.1 No Near End Speech

During speech pause, $s_1[k] \equiv s_2[k] \equiv 0$, the MMSE $E\left\{e_r^2[k]\right\}$ is given by

$$E\left\{e_r^2[k]\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|P_{v_1 v_2}(\Omega)|^2}{P_{v_2 v_2}(\Omega)} d\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{v_1 v_1}(\Omega)\, C_{v_1 v_2}(\Omega)\, d\Omega\,. \tag{12.15}$$

The residual noise at the output of the optimal filter depends on the power spectral density $P_{v_1 v_1}(\Omega)$ of the noise and the coherence $C_{v_1 v_2}(\Omega)$ of the noise components.

12.3.2 High Signal to Noise Ratio

If the microphone signals have a high SNR the MMSE can be written as a function of the power spectral density of the speech signal and the coherence $C_{s_1 s_2}(\Omega)$ of the speech components. The approximations $P_{s_2 s_2}(\Omega) \gg P_{v_2 v_2}(\Omega)$ and $P_{s_2 s_2}(\Omega) \gg |P_{v_1 v_2}(\Omega)|^2$ lead to

$$E\left\{e_r^2[k]\right\} \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{s_1 s_1}(\Omega)\,(1 - C_{s_1 s_2}(\Omega))\, d\Omega\,. \tag{12.16}$$

A prerequisite for high speech quality is thus a coherence of the speech components which is close to one. Incoherent speech components generated, for instance, by reverberation will be attenuated. Whether this attenuation constitutes an improvement or a reduction of the perceived speech quality depends on the ratio of coherent and incoherent speech sounds and the noise level. If the speech components are less coherent than the ambient noise the SNR will not be improved. A sufficient amount of coherent speech is therefore of paramount importance for a good performance of the coherence based two-microphone speech enhancement system. The coherence of speech signals can be improved by using directional microphones.

12.4 A Noise Reduction Application

In this section we describe a noise reduction algorithm which is based on the Wiener filter as discussed in Section 12.3. The algorithm uses a time domain implementation of the Wiener filter and was developed for the voice intercom of a computed tomography scanner.

The voice communication between a patient in a computed tomography (CT) scanner and the operator at the control desk is disturbed by acoustic noise which originates from the CT scanner. The acoustic noise in the gantry tunnel is due to numerous cooling fans and to the rotating x-ray imaging system. To reduce the fatigue of the operator, it is very desirable to reduce the level of noise transmitted from the gantry of the scanner to the control desk. Since the noise is highly non-stationary, single microphone speech enhancement methods do not perform well in this environment.

Figure 12.7 illustrates the application of the two-microphone noise reduction algorithm in the computed tomography scanner. The microphones which pick up the patient's speech are mounted inside the gantry tunnel at a distance of 0.4 m. The microphone signals are sampled and processed on a DSP. The enhanced signal $\hat{s}[k]$ is then played back on a loudspeaker at the control desk.

12.4.1 An Implementation Based on the NLMS Algorithm

Figure 12.8 shows a block diagram of the time domain implementation. The microphone signals are bandlimited to 3600 Hz and sampled with $f_s = 8000$ Hz. Preemphasis filters whiten the input signals and thus improve the convergence of the adaptive filter. They also help to improve the

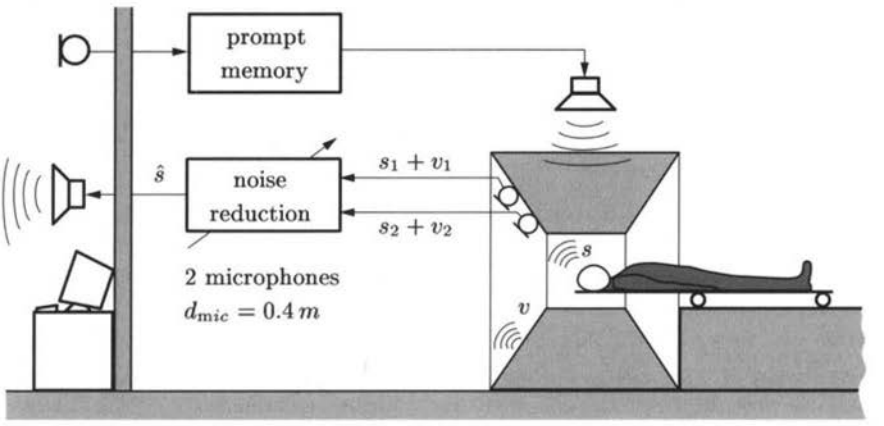

Fig. 12.7. Application of the two-microphone noise reduction system for voice communication in a computed tomography scanner.

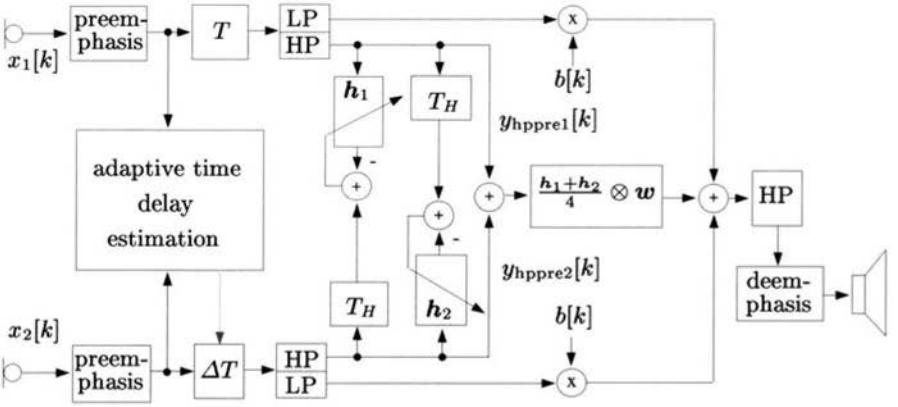

Fig. 12.8. Block diagram of the two-microphone noise reduction system. LP: lowpass filter, HP: highpass filter, $\Delta T = T - \tau$.

reproduction of high frequency speech components in a fixed-point implementation.

An adaptive time delay estimation algorithm compensates time delays between the two input signals. The time delay compensation is based on the correlation of the two microphone signals and on the SNR. When a high SNR is detected, a recursively smoothed correlation function is searched for maxima and the time delay is determined [17]. To avoid noticeable lowpass or comb-filtering effects, the magnitude of the delay error after delay compensation should be smaller than $1/(4f_s)$.

The spectral density and the coherence properties of the microphone signals suggest processing the speech signal in two frequency bands. Above 800 Hz, a linear phase adaptive Wiener filter is used which suppresses incoherent signal components (noise and reverberated speech) and passes highly coherent speech signals. Processing the frequency band below 800 Hz with this adaptive filter would result in noticeable fluctuations of the residual noise. These fluctuations are caused by the correlation of the noise signals at low frequencies. The noise in the band between 240 and 800 Hz is therefore suppressed by an adaptive scalar factor $b[k]$. This factor is controlled by the speech activity of the person inside the gantry tunnel. The speech activity is determined by the SNR estimator which is also used to increase the robustness of the time delay estimation algorithm [19]. The frequency band below 240 Hz is attenuated by 20 dB by means of a second order recursive highpass filter. The deemphasis filter at the output of the speech enhancement system restores the spectral characteristics of the speech signal. This noise reduction system is currently used in the Siemens SOMATOM PLUS 4 CT scanner[1].

[1] Siemens and SOMATOM PLUS 4 are registered trademarks of Siemens AG, Germany.

The computational complexity of the algorithm is about 10 MIPS on a 24 bit fixed-point DSP. In what follows we briefly describe the algorithm.

12.4.2 Processing in the 800 – 3600 Hz Band

The Wiener filter is approximated by two antiparallel linear phase NLMS adapted FIR filters (see Figure 12.8). The adaptive filters with the coefficient vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ of order $L_H = 64$ are updated using a linear phase version [20] of the NLMS algorithm ($T_H = L_H/2$),

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

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

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

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

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

and

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

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

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 $\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\text{hppre1}}[k]$ and $\left(\boldsymbol{I} + \boldsymbol{I}^R\right) \boldsymbol{y}_{\text{hppre2}}[k]$ 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, $(y_{\text{hppre1}}[k] + y_{\text{hppre2}}[k])/2$, we use the mean of the two adaptive coefficient vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ and an additional smoothing window $\boldsymbol{w} = (w_0, w_1, ..., w_{L_H})^T$

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

The symbol $\otimes$ denotes the pointwise multiplication of two vectors. The window function is used to smooth the frequency response of the adaptive filter. A Kaiser window [21] with a shape parameter β_{Kaiser} in the range of $3 \leq \beta_{Kaiser} \leq 5$ results in good speech quality and increased noise reduction. Since the coefficient vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ represent linear phase filters, the averaging of the vectors $\boldsymbol{h}_1[k]$ and $\boldsymbol{h}_2[k]$ in (12.24) yields an average of the amplitude spectra of these filters without errors due to mismatching phase spectra.

12.4.3 Processing in the 240 – 800 Hz Band

The attenuation factor $b[k]$ ($b_{min} \leq b[k] \leq b_{max}$) is controlled by a speech activity detector and adjusted according to the estimated SNR. Whenever the estimated SNR is below a preselected threshold the attenuation is slowly and successively increased until a maximum attenuation of 40 dB (corresponding to $b_{min} = 0.01$) is reached. Whenever the estimated SNR is above the threshold the attenuation is rapidly decreased to a minimum value of 3 dB ($b_{max} = 0.5$). The SNR threshold is set to 3 dB. Thus, the attenuation factor $b[k]$ is computed using the recursive system

$$b[k+1] = b[k]\beta_1 + b_{max}\,(1-\beta_1)\,,\ SNR > \text{threshold} \tag{12.25}$$

$$b[k+1] = b[k]\beta_2 + b_{min}\,(1-\beta_2)\,,\ SNR \leq \text{threshold}\,. \tag{12.26}$$

The smoothing constants β_1 and β_2 are set to $\beta_1 = 0.9996$ and $\beta_2 = 0.99999$.

12.4.4 Evaluation

We assess the performance of the noise reduction algorithm in terms of

- distortion of the speech signal;
- noise reduction during speech activity;
- noise reduction during speech pause.

These criteria can be measured during simulation of the speech enhancement system. For the purpose of measuring the above properties, the adaptive filter is duplicated such that the undisturbed speech signal and the noise signal can be processed independently [22]. Figure 12.9a outlines this approach. It requires separate recordings of the noise and the speech signals.

The speech signal distortion can then be measured as the segmental SNR of the filtered signal $\widetilde{s}[k]$ with respect to the unprocessed delayed speech signal $s[k - T_H]$

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

with

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

M denotes the segment length and K the number of segments. To reduce the influence of speech pauses, we average the SNR of only those speech signal frames which exhibit an SNR larger than 0 dB. Since we use a linear phase filter, the segmental SNR measures the amplitude distortion of the speech signal and is therefore well correlated with perceived distortions.

The attenuation of the noise signals during speech activity NR_{active} and during speech pause NR_{pause} is measured as the power ratio of the noise signals before and after the adaptive filter

$$NR_{active} = 10 \log_{10} \left(\overline{P_v[k - T_H]} / \overline{P_{\tilde{v}}[k]} \right) , \ \overline{P_s}[k] \neq 0 \tag{12.29}$$

$$NR_{pause} = 10 \log_{10} \left(\overline{P_v[k - T_H]} / \overline{P_{\tilde{v}}[k]} \right) , \ \overline{P_s}[k] = 0 , \tag{12.30}$$

where $\overline{P_v[k]}$ and $\overline{P_{\tilde{v}}[k]}$ denote the average power of the unprocessed and the processed noise signal, respectively. $\overline{P_s}[k]$ denotes the short term power of the speech signal.

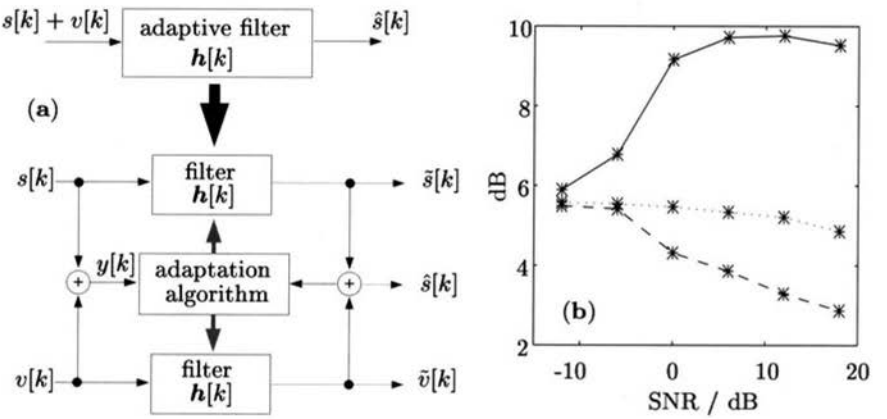

Fig. 12.9. Method for computing objective measures (**a**). Distortion of speech signal and noise reduction during speech activity and during speech pause vs. input SNR of the adaptive filter in the upper band (**b**). Step size: $\alpha = 0.1$, filter order $L_H = 64$.
(——): Segmental SNR of speech;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

Figure 12.9b plots the objective measures as a function of the input SNR. For this experiments, adaptive filters of order $L_H = 64$ and a rectangular window $\boldsymbol{w}$ were used (equivalent to $\beta_{Kaiser} = 0$). It can be seen that the speech signal distortion increases as the input SNR decreases. When the incoherent noise becomes dominant at about 0 dB, the speech SNR significantly degrades. The noise reduction during speech pause is about 5 – 6 dB. It can be improved by 2 – 3 dB if a tapered smoothing window is used. The overall noise reduction during speech pause including the adaptive scalar weighting in the lower band and the highpass filter is about 14 dB during speech pause.

12.4.5 Alternative Implementations of the Coherence Based Postfilter

The speech enhancement system as outlined above exploits the coherence properties of the microphone signals. Coherence based noise reduction systems can be also implemented in the frequency domain [3,9,5,6]. Since the coherence based approach does not rely on an explicit noise power spectral density estimate, the performance of these systems is limited by the coherence of the noise and the speech signals. To improve the noise reduction especially for low frequencies, the combination with spectral weighting techniques has been proposed. In [5] the coherence function is also employed to detect speech pauses and to enable noise power spectral estimation during speech pauses. In [23] the cross power spectral density of the microphone signals is used to derive an explicit noise power spectral density estimate, which is then used in a two-channel spectral subtraction. Combined with small superdirective endfire arrays, this system led to significant improvements of speech intelligibility in conjunction with cochlear implants [24].

12.5 Combined Noise and Acoustic Echo Reduction

A hands-free conferencing system has to cope not only with ambient noise but also with acoustic echoes. The feedback of the far-end speech signal via the loudspeaker, the room, and the microphone (the "LRM system") back to the far-end side necessitates an echo suppression device to guarantee the stability of the electro-acoustic loop and to supply sufficient echo reduction. While the stability of the electro-acoustic loop can be treated as a control problem, the echo reduction aims at making the echo imperceptible. The echo reduction problem is therefore closely linked to psychoacoustics, especially to masking effects in the human auditory system.

The noise and the echo reduction problems were addressed independently for many years (see e.g. [25–28] and [29,30] for reviews of these methods). To achieve optimal performance it has been recognized, however, that the echo control and noise reduction problem should be tackled in a combined approach [31,8,32–34]. The combined treatment yields algorithms which deliver better performance at less computational costs than systems based on

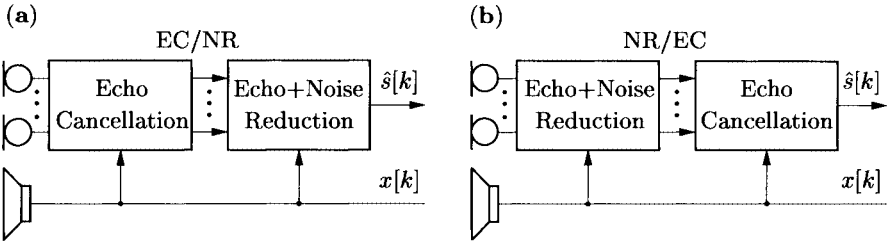

Fig. 12.10. Processing options for combined systems: EC precedes ENR (**a**), ENR precedes EC (**b**).

separate algorithms [8,35]. In this section we will outline one solution to the combined echo and noise reduction problem and show how the coherence based noise reduction postfilter can be extended to achieve a high level of echo reduction. This requires the integration of echo cancellers into the two-microphone speech enhancement system.

When acoustic echo cancellation is combined with an echo and noise reduction postfilter it must be asked in which order these two processing operations should be performed [36]. Figures 12.10a and 12.10b depict two principal cases: the configuration EC/NR where the acoustic echo cancellation (EC) precedes a speech enhancement postfilter (NR), and vice versa, the configuration NR/EC.

Although the echo canceller can benefit from the noise reduction in the NR/EC configuration, there are good reasons why the configuration of Figure 12.10a, where the echo cancellation precedes the noise reduction, is preferable. The main advantage of the EC/NR configuration is that the noise reduction postfilter is not presented with the disturbing and possibly highly coherent echo that is found in the microphone signals and that there is no time varying noise reduction filter in the echo path. Besides that, if the echo canceller does not deliver sufficient echo attenuation, the residual echo can be treated similar to the background noise signal and can be further attenuated by the postfilter. This idea is successfully exploited in a frequency selective echo reduction technique, called *echo shaping* [37], which does not require complete cancellation of the echo by the echo canceller. Instead, the total echo attenuation is split between the echo cancellers and the postfilter. A disadvantage of the EC/NR approach is, however, that it requires cancellation of the microphone signals of the array by individual cancellers.

The combination of acoustic echo cancellation with an adaptive microphone array is a challenging task by itself [38–40]. If we provide a canceller for each microphone channel, the echo cancellers converge as well as in the single microphone case. However, the computational load may be too large. If a single echo canceller is placed after the summation point of the array, the adaptation of the echo canceller might be severely disturbed when the look direction of the array is adapted to the speaker position.

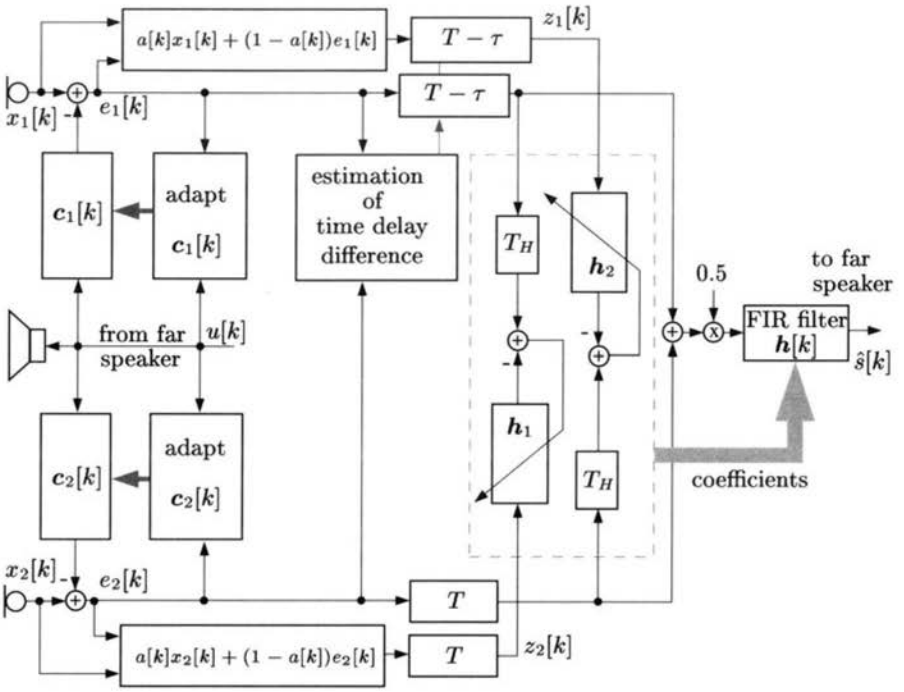

Fig. 12.11. Block diagram of the two microphone combined echo and noise reduction algorithm.

To avoid the adaptation and the complexity problems, we adopt a strategy as follows: The echo cancellers are placed at the microphone inputs but they are equipped with a reduced number of filter taps. The echo attenuation of the cancellers will be reduced but since they have fewer filter taps their speed of convergence will be improved. The reduced echo attenuation of the echo cancellers is compensated for by the array gain [41] and the noise reduction postfilter which will also reduce acoustic echoes, especially the incoherent late reverberation portion. To improve the echo suppression, we apply additional echo attenuation in the postfilter by using the *echo shaping* technique for the postfilter adaptation. The *echo shaping* technique attenuates only those frequencies for which the echo dominates the near-end signal. It leads to significantly increased echo attenuation, to acceptable near-end signal distortions during double talk, and to modest computational demands as compared to single-microphone speech enhancement systems.

Figure 12.11 depicts a block diagram of the combined echo and noise reduction algorithm. The cancellers at the microphone inputs are NLMS adapted FIR filters with adaptive stepsize control [42]. The echo cancelled signals are time delay compensated. Because some of the echo is incoherent, the noise reduction postfilter will also reduce acoustic echoes. To amplify this

effect and to reduce coherent echoes, the amount of echo in the inputs of the adaptive filters is deliberately increased by using a linear combination of the microphone signal $x_n[k]$ and the echo compensated signal $e_n[k]$ as an input to the adaptive filters. The reference signal of the adaptive filters is the echo compensated signal of the other microphone channel. Neglecting the time delay compensation (or letting $\tau = T$), the input to the adaptive filter $\boldsymbol{h}_2$ is given by

$$z_1[k] = a[k]x_1[k] + (1 - a[k])e_1[k] \tag{12.31}$$

$$= s_1[k] + v_1[k] + d_1[k] - (1 - a[k])\hat{d}_1[k] \,, \tag{12.32}$$

where $d_1[k]$ denotes the echo signal and $\hat{d}_1[k]$ the echo estimate of the echo canceller. The linear combination is controlled by the time varying factor $a[k]$. For $a[k] > 0$ there will be more echo in the input signal of the adaptive filter than in the reference signal since the reference signal is the echo compensated signal. To match the echo level of the reference signal, the adaptive filters will attenuate the echo if it dominates the near-end signals. Therefore, this mechanism provides for additional echo attenuation whenever the echo is disturbing. An algorithm for the adaptation of the "mixing factor" $a[k]$ is outlined in [17].

12.5.1 Experimental Results

The two-microphone algorithm as explained above was evaluated in a desktop conferencing experiment. Since we aim for low coherence of the acoustic echoes and high inherent robustness, the acoustic interface was designed such that there is little coupling between the loudspeaker and the microphones. Also, to increase the performance of the system, directional microphones or small superdirective endfire arrays [43] should be used. Figure 12.12 explains the experimental setup. The near-end speaker and the microphones are placed at a table in accordance with ITU-T recommendation P.34 [44]. The microphones have a hypercardioid directivity pattern and the loudspeaker is placed in a direction of low sensitivity of the microphones. The specific arrangement of the microphones and the loudspeaker combined with the gains of the loudspeaker and the microphone amplifiers resulted in an echo return loss of 9 dB. A similar setup (with similar results) was also used in a car. In this case the two microphones were mounted at the sun visor and the loudspeaker was attached to the dashboard.

Figure 12.13a plots the echo return loss enhancement (ERLE) for *single talk* and a stationary LRM system as a function of the number of compensator taps. For $a[k] \equiv 0$, the postfilter reduces noise and incoherent echo components only. The additional echo reduction delivered by the postfilter is then about 6 dB, independent of the compensator order. A significant increase of the echo attenuation is achieved when the *echo shaping* algorithm is turned on ($a[k]$ adaptive). It can be shown with the noise-free case that

for $a[k] \equiv 1$ the postfilter delivers the same amount of echo suppression as the echo canceller alone [17]. Hence, the slope of the ERLE vs. compensator order plot is about twice as steep as the slope of the plot for the echo canceller only. Figure 12.13b plots the ERLE as a function of the SNR. Again we find a significant increase of the echo attenuation for the *echo shaping* technique. For *double talk* most of the echo reduction is delivered by the echo canceller since the coherent near-end speech is passed by the filter. Only for frequencies where the residual echo dominates the near-end signal is significant echo reduction applied. If the far-end speaker is not active (no echo), the combined system behaves exactly like the two-microphone noise reduction systems of Section 12.4.1. For the office situation the average noise reduction is about 10 dB.

12.6 Conclusions

This chapter has presented a two-microphone postfilter approach to noise reduction and to combined echo and noise reduction. It was shown that the performance of these systems is closely linked to the spatial coherence of the speech, the noise, and the echo signals. The coherence based processing is useful only above a cutoff frequency which depends on the microphone distance. However, as a high coherence of the speech signal is also of great importance, the microphone distance cannot be made arbitrarily large. Best results are therefore obtained when microphones or small (endfire) arrays with a high directivity are used in conjunction with the proposed postfilter.

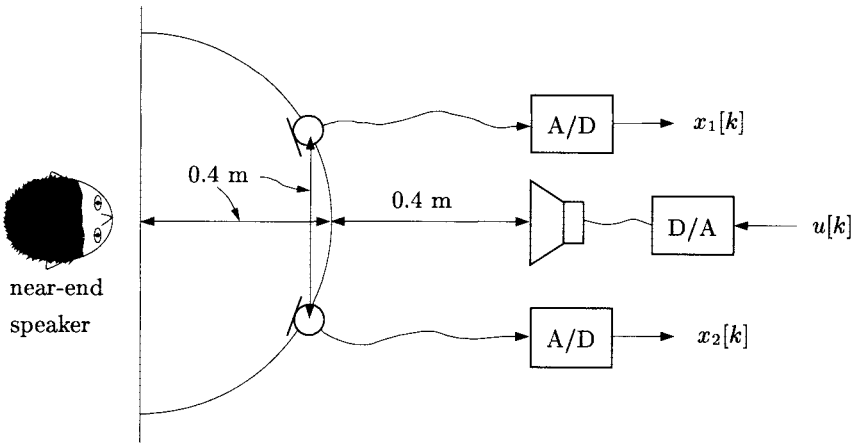

Fig. 12.12. Setup of the near-end speaker, the microphones, and the loudspeaker for conferencing experiments. The microphones have a hypercardioid directivity pattern.

The array and postfilter approach has been successfully deployed in the voice intercom of a computed tomography scanner.

Acknowledgements

The author is grateful to M. Dörbecker (Ericsson Eurolab Nuremberg), H. Hagena and T. Lotter (Aachen University of Technology) for reviewing the manuscript.

References

1. B. Widrow, J.R. Glover, J.M. McCool, et al., "Adaptive Noise Cancelling: Principles and Applications," in *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.
2. W. Armbrüster, R. Czarnach, and P. Vary, "Adaptive Noise Cancellation with Reference Input - Possible Applications and Theoretical Limits," in *Signal Processing III, Theories and Applications*, pp. 391–394, Elsevier, 1986.
3. J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.
4. E.R. Ferrara and B. Widrow, "Multichannel Adaptive Filtering for Signal Enhancement," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-29, no. 3, pp. 766–770, 1981.
5. R. Le Bouquin and G. Faucon, "On Using the Coherence Function for Noise Reduction," in *Signal Processing V: Theories and Applications*, pp. 1103–1106, Elsevier, 1990.

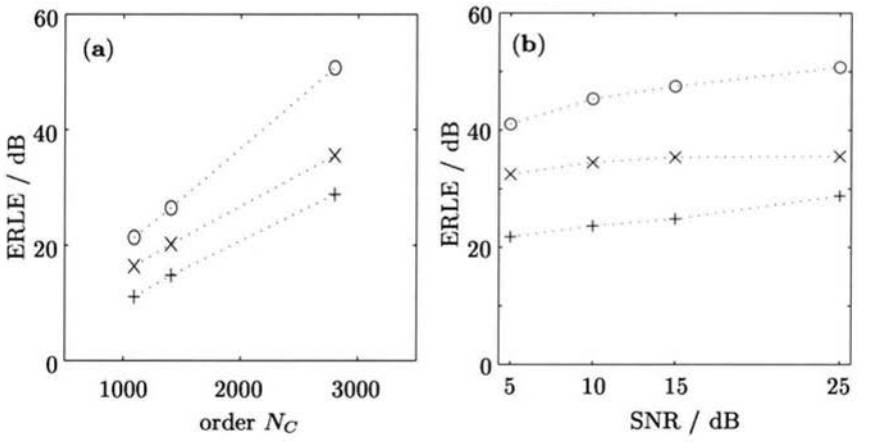

Fig. 12.13. Mean ERLE for single talk vs. the compensator order N_C for SNR $\approx$ 30 dB (**a**) and vs. the SNR for $N_C = 2800$ (**b**).
+: mean ERLE of echo canceller only;
x: mean ERLE of combined system with $a[k] \equiv 0$;
o: mean ERLE of combined system with $a[k]$ adaptive.

6. R. Le Bouquin and G. Faucon, "Study of a noise cancellation system based on the coherence function," in *Signal Processing VI: Theories and Applications*, pp. 1633–1636, Elsevier, 1992.
7. R. Martin and P. Vary, "A Symmetric Two Microphone Speech Enhancement System – Theoretical Limits and Application in a Car Environment," in *Proc. Fifth IEEE Signal Processing Workshop*, pp. 4.5.1–4.5.2, 1992.
8. R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation, and Noise Reduction: A Two Microphone Approach," in *Proc. Third Int. Workshop on Acoustic Echo Control*, Lannion, France, pp. 125–132, 1993.
9. R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-88)*, New York NY, USA, pp. 2578–2581, Apr. 1988.
10. R. Zelinski, "Noise Reduction Based on Microphone Array with LMS Adaptive Post-Filtering," *Elect. Lett.*, vol. 26, no. 24, pp. 2036–2037, 1990.
11. C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. SAP*, vol. 6, no. 3, pp. 240–259, 1998.
12. J.S. Bendat and A.G. Piersol, *Measurement and Analysis of Random Data*, Wiley, 1966.
13. G.C. Carter, "Coherence and Time Delay Estimation," in *Proc. IEEE*, vol. 75, no. 2, pp. 236–255, 1987.
14. H. Kuttruff, *Room Acoustics*, Elsevier Science, 3rd edn., 1990.
15. B.F. Cron and C.H. Sherman, "Spatial-Correlation Functions for Various Noise Models," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1732–1736, 1962.
16. P. Dämmig, "Zur Messung der Diffusität von Schallfeldern durch Korrelation," *Acustica*, vol. 7, pp. 387, 1957.
17. R. Martin, "Hands-free Telephones based on Multi-Microphone Echo Cancellation and Noise Reduction," PhD Thesis (in German), Institute of Communication Systems and Data Processing, Aachen University of Technology, Aachen, Germany, 1995.
18. T. Kailath, *Lectures on Wiener and Kalman Filtering*, CISM Courses and Lectures No. 140, Springer Verlag, 1981.
19. R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals," in *Proc. EUROSPEECH*, Berlin, Germany, pp. 1093–1096, 1993.
20. C.F.N. Cowan and P.M. Grant, *Adaptive Filters*, Prentice Hall, 1985.
21. J.F. Kaiser, "Nonrecursive Digital Filter Design Using the I_o-sinh Window Function," in *Proc. IEEE Int. Symp. on Circuits and Syst.*, pp. 20–23, 1974.
22. R. Le Bouquin, G. Faucon, and A. Akbari Azirani, "Proposal of a Composite Measure for the Evaluation of Noise Cancelling Methods in Speech Processing," in *Proc. EUROSPEECH*, Berlin, Germany, pp. 227–230, 1993.
23. M. Dörbecker and S. Ernst, "Combination of Two-Channel Spectral Subtraction and Adaptive Wiener Post-Filtering for Noise Reduction and Dereverberation," in *Proc. EUSIPCO*, Trieste, Italy, pp. 995–998, 1996.
24. M. Dörbecker, *Multi-channel Algorithms for the Enhancement of Noisy Speech for Hearing Aids*, PhD Thesis (in German), Institute of Communication Systems and Data Processing, Aachen University of Technology, Aachen, Germany, 1998.
25. C. Breining, P. Dreiseitel, E. Hänsler, et al., "Acoustic Echo Control. An application of very high order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, 1999.

26. S.L. Gay, J. Benesty (eds.),*Acoustic Signal Processing for Telecommunication*, Kluwer, 2000.
27. E. Hänsler, "The Hands-Free Telephone Problem - An Annotated Bibliography," *Signal Processing*, vol. 27, pp. 259–271, 1992.
28. E. Hänsler, "The Hands-Free Telephone Problem - An Annotate Bibliography Update," *Annales des Télécommunication*, vol. 49, no. 7-8, pp. 360–367, 1994.
29. J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993.
30. Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.
31. Y. Grenier, M. Xu, J. Prado, and D. Liebenguth, "Real-Time Implementation of an Acoustic Antenna for Audioconferencing," in *1st Intl. Workshop on Acoustic Echo Control*, Berlin, Germany, 1989.
32. R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation, and Noise Reduction: A Two Microphone Approach," *Annales des Télécommunications*, vol. 49, no. 7–8, pp. 429–438, 1994.
33. B. Ayad and G. Faucon, "Acoustic Echo and Noise Cancelling for Hands-Free Communication Systems," in *Proc. Fourth Int. Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 91–94, 1995.
34. G. Faucon and R. Le Bouquin Jeannes, "Joint System for Acoustic Echo Cancellation and Noise Reduction," in *Proc. EUROSPEECH*, Madrid, Spain, pp. 1525–1528, 1995.
35. R. Martin and J. Altenhöner, "Coupled Adaptive Filters for Acoustic Echo Control and Noise Reduction," in *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 3043–3046, May 1995.
36. R. Martin and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands-Free Telephony - State of the Art and Perspectives," in *Proc. EUSIPCO*, Trieste, Italy, pp. 1107–1110, 1996.
37. R. Martin and S. Gustafsson, "The Echo Shaping Approach to Acoustic Echo Control," *Speech Communication*, vol. 20, pp. 181–190, 1996.
38. W. Herbordt and W. Kellermann, "GSAEC - Acoustic Echo Cancellation Embedded into the Generalized Sidelobe Canceller," in *Proc. EUSIPCO*, Tampere, Finland, 2000.
39. W. Kellermann, "Strategies for Combining Acoustic Echo Cancellation and Adaptive Beamforming Microphone Arrays," in *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 219–222, 1997.
40. R. Martin, S. Gustafsson, and M. Moser, "Acoustic Echo Cancellation for Microphone Arrays Using Switched Coefficient Vectors," in *Proc. 5th Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, London, England, pp. 85–88, 1997.
41. W. Kellermann, "Some Properties of Echo Path Impulse Responses of Microphone Arrays and Consequences for Acoustic Echo Cancellation," in *Proc. Fourth Int. Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 39–43, 1995.
42. C. Antweiler, *Orthogonalizing Algorithms for Digital Compensation of Acoustic Echoes*. PhD Thesis (in German), Institute of Communication Systems and Data Processing, Aachen University of Technology, Aachen, Germany, 1995.
43. M. Dörbecker, "Small Microphone Arrays with Optimized Directivity for Speech Enhancement," in *Proc. EUROSPEECH*, Rhodes, Greece, pp. 327–330, 1997.

44. ITU-T Recommendation P.34, *Transmission Characteristics of Hands-Free Telephones*, Melbourne, Australia, 1988.

13 Acoustic Echo Cancellation for Beamforming Microphone Arrays

Walter L. Kellermann

University Erlangen – Nürnberg, Germany

Abstract. Acoustic feedback from loudspeakers to microphones constitutes a major challenge for digital signal processing in interfaces for natural, full-duplex human–machine speech interaction. Two techniques, each one successful on its own, are combined here to jointly achieve maximum echo cancellation in real environments: For one, acoustic echo cancellation (AEC), which has matured for single-microphone signal acquisition, and, secondly, beamforming microphone arrays, which aim at dereverberation of desired local signals and suppression of local interferers, including acoustic echoes. Structural analysis shows that straightforward combinations of the two techniques either multiply the considerable computational cost of AEC by the number of array microphones or sacrifice algorithmic performance if the beamforming is time-varying. Striving for increased computational efficiency without performance loss, the integration of AEC into time-varying beamforming is examined for two broad classes of beamforming structures. Finally, the combination of AEC and beamforming is discussed for multi-channel recording and multi-channel reproduction schemes.

13.1 Introduction

For natural human–machine interaction, acoustic interfaces are desirable that support seamless full-duplex communication without requiring the user to wear or hold special devices. For that, the general scenario of Figure 13.1 foresees several loudspeakers for multi-channel sound reproduction and a microphone array for acquisition of desired signals in the local acoustic environment. Acoustic signal processing is employed to support services such as speech transmission, speech recognition, or sound field synthesis offered by communication networks or autonomous interactive systems. Such hands-free acoustic interfaces may be tailored for incorporation into a wide variety of communication terminals, including teleconferencing equipment, mobile phones and computers, car information systems, and home entertainment equipment.

For signal acquisition, microphone arrays allow spatial filtering of arriving signals and, thus, desired signals can be enhanced and interferers can be suppressed. With full-duplex communication, echoes of the loudspeaker signals will join local interferers to corrupt the desired source signals. Beamforming, however, does not exploit the available loudspeaker signals as reference information for suppressing the acoustic echoes. This is accomplished by acoustic

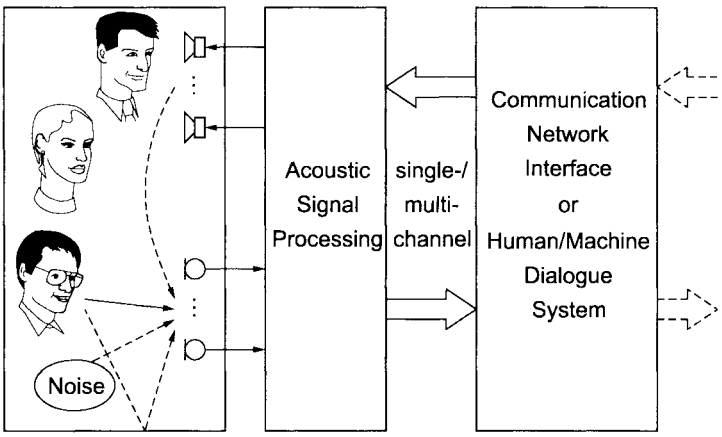

Fig. 13.1. Acoustic interface for natural human–machine communication.

echo cancellation (AEC) algorithms [1–3]. For discussing the combination of AEC with microphone arrays, the concept of AEC is first reviewed in Section 13.2 and beamforming methods are categorized in Section 13.3 with respect to the properties determining the interaction with AEC. Then, generic concepts for the combination of AEC and beamforming are discussed in Section 13.4. Structures for integrating AEC into beamforming are investigated in Section 13.5. Finally, the extension from single-channel reproduction to the case of multiple reproduction channels is outlined.

13.2 Acoustic Echo Cancellation

The concept of AEC is first considered for the case of a single loudspeaker and a single microphone according to Figure 13.2. To remove the echo from the microphone signal $x(n)$ (with n denoting discrete time), AEC aims at generating a replica $\widehat{v}(n)$ for the signal $v(n)$, which is an echoed version of the loudspeaker signal $u(n)$. Aside from the echo $v(n)$, $x(n)$ contains components originating from local desired sources and local interferers, $s(n)$ and $r(n)$, respectively. Introducing the residual echo

$$e(n) = v(n) - \widehat{v}(n), \tag{13.1}$$

the estimate for the desired signal $\widehat{s}(n)$ can be written as:

$$\widehat{s}(n) = x(n) - \widehat{v}(n) = s(n) + e(n) + r(n). \tag{13.2}$$

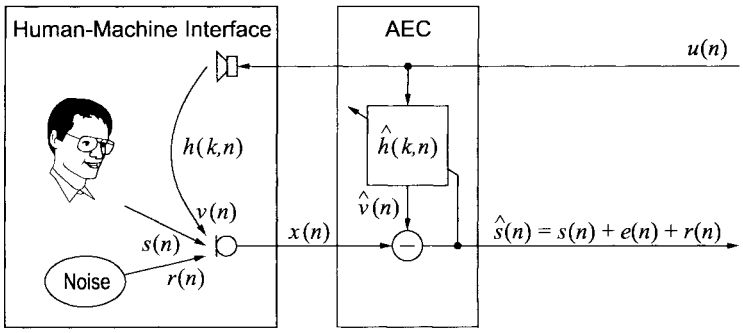

Fig. 13.2. Basic structure for single-channel AEC.

The amount of echo attenuation achieved by AEC is expressed by the *echo return loss enhancement* ($ERLE$)[1]:

$$ERLE_{log}(n) = 10 \cdot \log \frac{\mathcal{E}\{v^2(n)\}}{\mathcal{E}\{e^2(n)\}} \quad [\text{dB}], \tag{13.3}$$

with $\mathcal{E}\{\cdot\}$ denoting the expectation operator. As long as potential nonlinearities of the loudspeaker system can be neglected [4], the loudspeaker-enclosure-microphone(LEM) system is completely characterized by its generally time-varying impulse response $h(k,n)$. Indeed, the impulse response may vary drastically and unpredictably over time, as a slight change in position of any object can alter many coefficients significantly [2]. The number of impulse response samples that must be modeled for an $ERLE_{log}$ value of x dB is estimated by [2,5]

$$L_{AEC} \approx \frac{x}{60} \cdot f_s \cdot T_{60}, \tag{13.4}$$

where f_s denotes the sampling frequency, and T_{60} is the reverberation time[2]. Based on this estimate, more than $L_{AEC} = 1000$ impulse response coefficients must be perfectly matched to assure 20 dB of $ERLE_{log}$ for a typical office with $T_{60} = 400$ ms and an echo canceller operating at $f_s = 8$ kHz.

As a model for the LEM system, a digital FIR filter structure with a time-varying impulse response $\hat{h}(k,n)$ of length L_{AEC} is employed, so that the estimated echo $\hat{v}(n)$ is given by

$$\hat{v}(n) = \hat{\mathbf{h}}^T(n) \cdot \mathbf{u}(n) \tag{13.5}$$

[1] As $v(n)$ and $e(n)$ are not accessible in practical situations, $ERLE$ must be estimated from $\hat{s}(n)$ and $x(n)$ [2].

[2] As characteristic parameter of an enclosure, the reverberation time T_{60} is the time until the sound energy decays by 60dB after switching off the source.

where ^{T} denotes transposition and

$$\widehat{\mathbf{h}}(n) = \left[\widehat{h}(0,n), \widehat{h}(1,n), \ldots, \widehat{h}(L_{AEC}-1,n)\right]^T, \tag{13.6}$$

$$\mathbf{u}(n) = \left[u(n), u(n-1), \ldots, u(n-L_{AEC}+1)\right]^T. \tag{13.7}$$

The misalignment between the FIR model $\widehat{\mathbf{h}}(n)$ and the LEM system $\mathbf{h}(n)$ is described by the logarithmic system error norm $D_{log}(n)$:

$$D_{log}(n) = 10 \cdot \log \frac{||\mathbf{h}(n) - \widehat{\mathbf{h}}(n)||_2^2}{||\mathbf{h}(n)||_2^2}, \tag{13.8}$$

with $||\cdot||_2$ denoting the l_2 norm[3].

13.2.1 Adaptation algorithms

For identifying the time-varying impulse response $h(k,n)$, adaptive filtering algorithms derive an optimum vector $\widehat{\mathbf{h}}_{opt}(n)$ by minimizing a mean square error criterion based on the input $u(n)$ and the estimation error $e(n)$ (assuming here, for simplicity, $s(n) = r(n) = 0$). Three fundamental algorithms are introduced below for the general case of complex signals (for a comprehensive treatment of adaptive FIR filtering see, e.g., [6,7]). Adaptation control in the context of AEC is addressed and frequency domain implementations are outlined briefly.

Fundamental algorithms. Minimizing the mean squared error $E\left\{|e(n)|^2\right\}$ for (at least) wide-sense stationary signals and a time-invariant echo path $h(k,n) = h(k)$ leads to the Wiener-Hopf equation for the optimum echo canceller $\widehat{\mathbf{h}}_{opt}$ [7]

$$\widehat{\mathbf{h}}_{opt} = \mathbf{R}_{\mathbf{uu}}^{-1} \cdot \mathbf{r}_{\mathbf{uv}} \tag{13.9}$$

with the time-invariant correlation matrix $\mathbf{R}_{\mathbf{uu}}$ and the crosscorrelation vector $\mathbf{r}_{\mathbf{uv}}$ given by

$$\mathbf{R}_{\mathbf{uu}} = E\left\{\mathbf{u}(n)\mathbf{u}^H(n)\right\}, \tag{13.10}$$

$$\mathbf{r}_{\mathbf{uv}} = E\left\{\mathbf{u}(n)v^*(n)\right\}, \tag{13.11}$$

respectively. (* denotes complex conjugation and ^{H} conjugate complex transposition.) For nonstationary environments, iterative or recursive algorithms are required to approach the Wiener solution in (13.9). As the most popular adaptation algorithm, the **NLMS**(*Normalized Least Mean Square*) algorithm [6,7] updates the filter according to

$$\widehat{\mathbf{h}}(n+1) = \widehat{\mathbf{h}}(n) + \alpha \frac{\mathbf{u}(n)}{\mathbf{u}^H(n)\mathbf{u}(n)} e^*(n) \tag{13.12}$$

[3] If the length of $\mathbf{h}(n)$ is greater than L_{AEC}, then $\widehat{\mathbf{h}}(n)$ must be complemented with zeros accordingly.

and may be understood as a stochastic approximation of the steepest descent algorithm, with $\mathbf{u}(n)$ approximating the negative gradient vector, and a step-size parameter α, $0 < \alpha < 2$. Obviously, for correlated signals such as speech, $\mathbf{u}(n)$ will not cover uniformly the L_{AEC}-dimensional vector space, which implies that the convergence to minimum system error $D_{log}(n)$ in (13.8) is slow [7]. The popularity of the NLMS is based on its robust convergence behavior [2] and its low computational complexity (about $2L_{AEC}$ multiplications per sampling interval T (MUL's per T) are needed for implementing (13.1), (13.5), and (13.12)).

To improve the convergence for speech signals, the *Affine Projection Algorithm* (**APA**) uses P previous input vectors

$$\mathbf{U}(n) = [\mathbf{u}(n), \mathbf{u}(n-1), \ldots, \mathbf{u}(n-P+1)] \tag{13.13}$$

to compute an error vector

$$\mathbf{e}(n) = \mathbf{v}(n) - \mathbf{U}^T(n) \cdot \widehat{\mathbf{h}}^*(n), \tag{13.14}$$

where

$$\mathbf{e}(n) = [e(n), e(n-1), \ldots, e(n-P+1)], \tag{13.15}$$

$$\mathbf{v}(n) = [v(n), v(n-1), \ldots, v(n-P+1)]. \tag{13.16}$$

The filter coefficients are then updated according to

$$\widehat{\mathbf{h}}(n+1) = \widehat{\mathbf{h}}(n) + \alpha \mathbf{U}(n) \left[\mathbf{U}^H(n)\mathbf{U}(n) - \delta \mathbf{I}\right]^{-1} \mathbf{e}^*(n), \tag{13.17}$$

with the regularization parameter δ ($\delta \geq 0$) and $\mathbf{I}$ denoting the identity matrix. Thus, the APA can be interpreted as a generalization of the NLMS algorithm, which in turn corresponds to an APA with $P = 1, \delta = 0$. The gradient estimate for the APA is equal to the projection of the system misalignment vector $\mathbf{h}(n) - \widehat{\mathbf{h}}(n)$ onto the P-dimensional subspace spanned by $\mathbf{U}(n)$. Thus, the complementary orthogonal component of the misalignment vector becomes smaller with increasing P. The computational complexity of the APA amounts to approximately $(P+1) \cdot L_{AEC} + O(P^3)$ MUL's per T, where, typically, $P = 2, \ldots, 32$, and L_{AEC} is given by (13.4). Fast versions of the APA reduce the computational load to $2L_{AEC} + 20P$, but require additional measures to assure numerical stability [2,6].

As the most powerful and computationally demanding adaptation method, the **RLS**(*Recursive Least Squares*) algorithm directly minimizes a weighted sum of previous error samples

$$J(\widehat{\mathbf{h}}, n) = \sum_{k=1}^{n} \beta(k) |e(k)|^2, \quad \text{with} \quad 0 < \beta \leq 1. \tag{13.18}$$

The solution has the form of (13.9), however with time-dependent estimates for $\mathbf{R}_{\mathbf{uu}}(n)$, $\mathbf{r}_{\mathbf{u}v}(n)$ given by

$$\widehat{\mathbf{R}}_{\mathbf{uu}}(n) = \sum_{k=1}^{n} \beta(k)\mathbf{u}(k)\mathbf{u}^H(k), \tag{13.19}$$

$$\widehat{\mathbf{r}}_{\mathbf{u}v}(n) = \sum_{k=1}^{n} \beta(k)\mathbf{u}(k)v^*(k). \tag{13.20}$$

The update equation reads here

$$\widehat{\mathbf{h}}(n+1) = \widehat{\mathbf{h}}(n) + \widehat{\mathbf{R}}_{\mathbf{uu}}^{-1}(n)\mathbf{u}(n)e^*(n). \tag{13.21}$$

If an exponential window $\beta(k) = \lambda^{n-k}$ with the forgetting factor $0 < \lambda < 1$ is used, the inversion of $\widehat{\mathbf{R}}_{\mathbf{uu}}(n)$ is avoided by exploiting the matrix inversion lemma that allows recursive update of the inverse [7]. Then, the complexity of the RLS algorithm is on the order of L_{AEC}^2 MUL's per T [6]. Similarly to the APA, fast versions for the RLS algorithm have been proposed which reduce computational complexity to $7L_{AEC}$ MUL's per T. However, the large filter order L_{AEC} and the nonpersistent excitation $u(n)$ require extra efforts to assure stable convergence [6]. A simplified version of fast RLS algorithms is the *Fast Newton Algorithm* [6], which reduces the complexity to $L_{AEC} \cdot P$ MUL's per T, with P being a predictor order that should be matched to the correlation properties of the input $u(n)$. (For speech signals, $P \approx 10$ is a typical value at $f_s = 8$kHz.)

Adaptation control. Adaptation control has to satisfy two contradicting requirements. On one hand, changes in the echo path $h(k,n)$ should be tracked as fast as possible. This requires a large stepsize, α, for the NLMS and APA algorithms in (13.12) and (13.17)), and a rapidly decaying β for the RLS algorithm in (13.21), respectively. On the other hand, the adaptation must be robust to interfering local sources $s(n)$ and noise $r(n)$, which requires a small stepsize, α, and a slowly decaying β, respectively [2,7]. When a local talker is active, adaptation should be stalled immediately to avoid divergence of $\widehat{\mathbf{h}}(n)$. Therefore, a fast and reliable detection of local source activity and estimation of background noise levels is decisive for efficient AEC operation. Correspondingly, a significant amount of computational complexity is invested in monitoring parameters and signals which support adaptation control [2]. With properly tuned adaptation control, acoustic echoes are attenuated by, typically, about 25 dB of $ERLE_{log}$ during steady state using the above adaptation algorithms.

Frequency subband and transform domain structures. To reduce computational load and to speed up convergence of adaptation algorithms

which do not inherently decorrelate $u(n)$ (e.g., the NLMS algorithm), frequency subband and transform domain structures have been developed [1,8]. Subband structures decompose the fullband signals $u(n)$ and $x(n)$ into M subbands which are usually downsampled by $R < M$ [3,9]. The adaptive subband filters operate at a reduced sampling rate and require fewer coefficients which leads to overall computational savings by a factor of close to R^2/M compared to fullband adaptive filtering. After subtraction, the subband signals are synthesized to yield again a fullband signal $\widehat{s}(n)$. While the additional complexity for the analysis/synthesis filter banks is relatively small for large L_{AEC}, the introduced signal delay for $\widehat{s}(n)$ is objectionable in some applications [2,31].

Transform-domain structures draw their computational advantage over direct time-domain implementations from the fast Fourier transform (FFT) and its use for fast convolution [1,6,8]. Block-exact adaptation algorithms, which behave exactly like their time-domain counterparts, have been proposed for all the fundamental algorithms above. For the long impulse responses at issue, the system model $\widehat{h}(k,n)$ is often partitioned into shorter subsystems to reduce the signal delay [2].

13.2.2 AEC for multi-channel sound reproduction

Considering a multi-channel reproduction unit (see Figure 13.1) broadcasting K different sound channels $\mathbf{u}_\kappa(n)$ $(\kappa = 0, \ldots, K-1)$ with usually time-varying mutual correlation, any microphone records the sum of K echo signals produced by different echo paths $h_\kappa(k,n)$,

$$v(n) = \sum_{\kappa=0}^{K-1} \mathbf{h}_\kappa(n)^T \cdot \mathbf{u}_\kappa(n), \tag{13.22}$$

with $\mathbf{h}_\kappa(n), \mathbf{u}_\kappa(n)$ being defined according to (13.6) and (13.7). Correspondingly, K echo cancellers, $\widehat{\mathbf{h}}_\kappa(n)$, are needed to model the respective echo paths. As only one error signal, $e(n)$, is available, the K inputs, $u_\kappa(n)$, must be mutually decorrelated without perceptible distortion to allow identification of the individual $\widehat{\mathbf{h}}_\kappa(n)$. This difference to single-channel AEC defines an even more challenging system identification problem, which has been considered only for the stereo case ($K = 2$) so far [1,10–12]. Current adaptation schemes still exhibit slower convergence and multiply computational load by more than K compared to their single-channel AEC counterparts.

13.2.3 AEC for multi-channel acquisition

A straightforward extension of the single-loudspeaker/ single-microphone scenario to an N-microphone acquisition system essentially multiplies the number of adaptive filters by N. The N-channel echo cancellation is captured by

extending the signals in (13.2) to N-dimensional column vectors,

$$\widehat{\mathbf{s}}(n) = \mathbf{x}(n) - \widehat{\mathbf{v}}(n) = \mathbf{s}(n) + \mathbf{r}(n) + \mathbf{e}(n) \tag{13.23}$$

$$= \mathbf{s}(n) + \mathbf{r}(n) + \mathbf{v}(n) - \widehat{\mathbf{H}}^T(n)\mathbf{u}(n) \tag{13.24}$$

with $\mathbf{u}(n)$ according to (13.7), with $\mathbf{e}(n), \mathbf{r}(n), \widehat{\mathbf{s}}(n), \mathbf{s}(n), \mathbf{v}(n), \widehat{\mathbf{v}}(n), \mathbf{x}(n)$ as column vectors of the form

$$\mathbf{x}(n) = [x_0(n), \ldots, x_{N-1}(n)]^T, \tag{13.25}$$

and with $\widehat{\mathbf{H}}(n)$ as a matrix containing the impulse responses $\widehat{\mathbf{h}}_\nu(n)$ as columns according to

$$\widehat{\mathbf{H}}(n) = \left[\widehat{\mathbf{h}}_0(n), \ldots, \widehat{\mathbf{h}}_\nu(n), \ldots, \widehat{\mathbf{h}}_{N-1}(n)\right]. \tag{13.26}$$

While this implies a corresponding multiplication of the computational cost for filtering, the cost for adaptation and its control is not necessarily multiplied by N. All operations depending only on the input data, $u(n)$, have to be carried out only once for all N channels, which would include the matrix inversion in the APA or RLS algorithms, (13.17) and (13.21), respectively. However, some fast versions draw their efficiency from interweaving matrix inversion and update equations [6] and, therefore, do not completely support this separation. Frequency subband and transform domain algorithms [1,6,8,9] support this separation at least by requiring the analysis transform of $u(n)$ only once for all channels.

13.3 Beamforming

This section only aims at categorizing beamforming algorithms with respect to their interaction with AEC. For a comprehensive treatment of fundamental techniques see, e.g., [13,14], while the current state of beamforming technology with microphone arrays is covered in several other chapters of this book.

13.3.1 General structure

Consider a microphone array capturing N real-valued sensor signals, $x_\nu(n)$, which are filtered by linear time-varying systems with impulse responses $g_\nu(k,n)$ and then summed up (Figure 13.3). The resulting beamformer output, $y(n)$, can be written as

$$y(n) = \mathbf{G}^T(n) \cdot \mathbf{X}(n) = \mathbf{G}^T(n) \cdot [\mathbf{S}(n) + \mathbf{R}(n) + \mathbf{V}(n)], \tag{13.27}$$

with the column vector $\mathbf{G}(n)$ representing the concatenated impulse response vectors $\mathbf{g}_\nu(n)$

$$\mathbf{G}(n) = \left[\mathbf{g}_0^T(n), \ldots, \mathbf{g}_{N-1}^T(n)\right]^T, \tag{13.28}$$

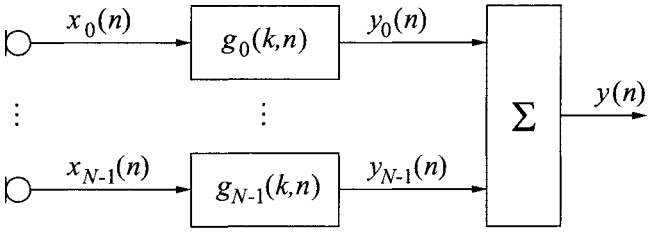

Fig. 13.3. General structure for a beamforming microphone array

where all $\mathbf{g}_\nu(n)$ are of length L_{BF}:

$$\mathbf{g}_\nu(n) = [g_\nu(0,n), \ldots, g_\nu(L_{BF}-1,n)]^T . \tag{13.29}$$

The column vector $\mathbf{X}(n)$ (and, equally, $\mathbf{R}(n), \mathbf{S}(n), \mathbf{V}(n)$) contains the latest L_{BF} signal samples of each microphone signal

$$\mathbf{X}(n) = \left[\mathbf{x}_0^T(n), \ldots, \mathbf{x}_{N-1}^T(n)\right]^T \tag{13.30}$$

with

$$\mathbf{x}_\nu(n) = [x_\nu(n), \ldots, x_\nu(n-L_{BF}+1)]^T . \tag{13.31}$$

In the scenario of Figure 13.1, beamforming aims at spatial filtering to dereverberate the components $\mathbf{s}(n)$ originating from the desired source(s) and to suppress interfering signals $\mathbf{r}(n)$ and echoes $\mathbf{v}(n)$.

For ideal dereverberation of a single source, the desired signal as it is emitted by the source, $s^{(0)}(n)$, should be retrieved except for some delay $n_0 > 0$:

$$\mathbf{G}^T(n) \cdot \mathbf{S}(n) = s^{(0)}(n-n_0). \tag{13.32}$$

Assuming that delayed versions of $s^{(0)}(n)$ are contained in $\mathbf{s}_\nu(n)$ defined by (13.31), the filters $g_\nu(k,n)$ have to equalize the corresponding delays and the sum of the filters has to provide a flat frequency response for all signals arriving from the source direction. Obviously, delay equalization requires knowledge about the location of the desired source(s). For the following, it is assumed that the source location is given by *a priori* knowledge or separately determined by some source localization algorithm (see, e.g., Chapters 8-10). For an anechoic environment and with the desired signal components being delay-equalized by the array geometry, the total impulse response, $g(k,n)$, of the beamformer to the desired source $s^{(0)}(n)$ should ideally fulfill

$$g(k,n) = \sum_{\nu=1}^{N} g_\nu(k,n) \stackrel{!}{=} \delta(k-k_0) \tag{13.33}$$

to assure a constant frequency response with unity gain and constant group delay k_0.

For interference suppression, the beamformer should minimize its response to all undesired signal components, which include here local interferers and loudspeaker echoes. Using, the mean squared error (MSE) as optimization criterion, this reads:

$$\mathcal{E}\left\{\left(\mathbf{G}^T(n)\cdot[\mathbf{R}(n)+\mathbf{V}(n)]\right)^2\right\}\stackrel{!}{=}\min. \tag{13.34}$$

Based on this general concept and with AEC in mind, basic methods for time-invariant or time-varying beamforming are outlined below.

13.3.2 Time-invariant beamforming

Time-invariant beamforming, i.e., $\mathbf{G}(n)=\mathbf{G}$, $\mathbf{g}_\nu(n)=\mathbf{g}_\nu$, is used for applications where the beamformer does not have to change the 'look direction' and where the potential nonstationarity of the involved signals, $\mathbf{s}(n),\mathbf{r}(n),\mathbf{v}(n)$, is not accounted for.

As the most basic beamforming method, the delay-and-sum beamformer (DSB) realizes in its simplest form a tapped delay line with a single non-zero coefficient for each filter $\mathbf{g}_\nu(n)$ [13,14]. If the required delays for the desired 'look direction' do not coincide with integer multiples of the sampling period, interpolation filters are required for realizing fractional delays [15–17]. Accounting for the wideband nature of speech and audio signals, nested arrays are often employed using different sets of sensors for different frequency bands to approximate a constant ratio between aperture width and signal wavelength [17–19]. As a generalization of DSB, filter-and-sum beamforming (FSB) aims for a frequency-independent spatial selectivity within each frequency band as detailed in Chapter 1 and [20]. Both beamforming concepts, DSB and FSB, were first developed on the basis of the far-field assumption [18], but may also be extended to near-field beamforming as described in Chapter 1. Time-invariant DSB and FSB are mostly signal-independent, i.e., no attention is paid to the power spectral densities of the signals $\mathbf{s}(n)$, $\mathbf{r}(n)$, $\mathbf{v}(n)$ and the direction of arrival (DOA) of interferers.

Such 'beamsteering' techniques are obviously appropriate for human–machine interfaces in reverberant environments with a restricted range of movement for a single desired source and where, due to reverberation, unwanted signal components of comparable level must be expected from all directions.

Nevertheless, time-invariant beamforming can incorporate additional spatial information to suppress dominant interferers [21,22]. Moreover, knowledge about long-term statistics of the noise field can be accounted for [23] and may lead to statistically optimum beamformers with superdirective behaviour for low frequencies as described in Chapter 2 and [24].

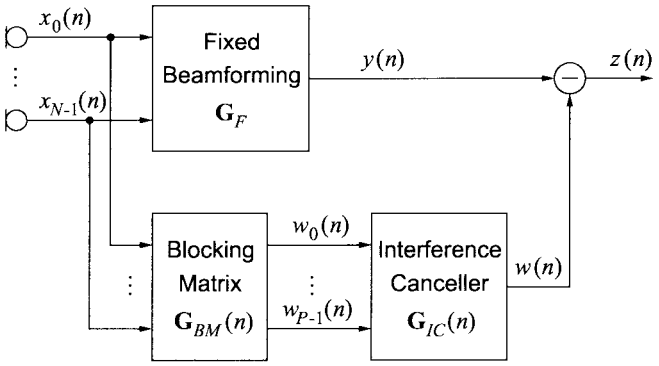

Fig. 13.4. Generalized sidelobe canceller structure for adaptive beamforming.

13.3.3 Time-varying beamforming

For nonstationary environments with both nonstationary signal characteristics and potentially moving sources, the beamformer should be able to track the time-variance of the signal characteristics and the spatial arrangement of the interfering sources. For that purpose, adaptive beamforming methods design filters $g_\nu(k,n)$ which minimize a statistical error criterion based on the array output, $y(n)$, with constraints for the DOA of a desired source (or 'target') such as formulated in (13.33) and (13.34) [13,14,25–27]. See also Chapter 5.

Generalized Sidelobe Canceller (GSC). As an example for an efficient implementation of adaptive beamformers that minimize a mean square error (MSE) criterion subject to a linear constraint, the generalized sidelobe canceller structure [13,25] is considered (Figure 13.4). Here, the adaptive beamforming is separated into two parallel paths: The upper path is a time-invariant, signal-independent beamformer, $\mathbf{G}_F$, steered toward the desired source. In the lower path, the first stage implements a blocking-matrix, $\mathbf{G}_{BM}(n)$, which, ideally, completely suppresses the components of the desired source, $\mathbf{s}(n)$, by a linear combination of the microphone channels [13] or filtering [28]. This topic is also detailed in Chapter 5. The $P \leq N$ outputs, $w_i(n), i = 0, \ldots, P-1$, are then used by the adaptive interference canceller, $\mathbf{G}_{IC}(n)$, to form an estimate for the interference component in $y(n)$. Optimization of $\mathbf{G}_{IC}(n)$ becomes an unconstrained Wiener filtering problem when the MSE criterion of (13.9) is used, and ideally leads to removal of all components in $y(n)$ which are correlated to $w_i(n)$. For identifying the optimum $\mathbf{G}_{IC}(n)$, the same adaptation algorithms as for echo cancellation can be used, i.e., (13.12),(13.17),(13.21), with gradient-type algorithms like the NLMS algorithm being most common.

13.3.4 Computational complexity

For both time-invariant and time-varying beamforming, the computational load is essentially proportional to the number of sensors N. The FIR filter lengths typically do not exceed $L_{BF} = 128$ [17,20,29,30]. With increasing filter length, computational savings are obtained by frequency-domain implementations of the filtering [20,29]. As with AEC, for adaptive beamforming implementations a significant share of computational complexity is dedicated to fast and reliable source activity detection which forms the basis of adaptation control.

13.4 Generic structures for combining AEC with beamforming

First, the combination of AEC with beamforming is motivated by comparing practical requirements with typical performance of AEC and beamforming. Then, the main properties of two generic options for a combination are discussed in some detail.

13.4.1 Motivation

Although AEC and beamforming are two distinct signal processing concepts, their goals meet with regard to acoustic echoes. While AEC subtracts from $\mathbf{x}(n)$ an echo estimate, $\widehat{\mathbf{v}}(n)$, based on $u(n)$ as reference information, beamforming suppresses echoes within $\mathbf{x}(n)$ as undesired interference by its spatial filtering capability. With beamforming being undisputed for its effectiveness in suppressing local noise and reverberance of local desired sources, the need for a complementary AEC unit for acoustic echo suppression is discussed in the following.

As a guideline for desired echo suppression for telecommunication, [31] requires $ERLE_{log} \geq 45$ dB during single-talk and at least 30 dB during double-talk, assuming a 'natural' echo attenuation of up to 6 dB between the loudspeaker signal, $u(n)$, and the microphone signal, $x(n)$. Echo suppression methods other than AEC, e.g., noise reduction, loss insertion, or nonlinear devices, impair full-duplex communication and, thus, are only acceptable as supplementary measures [2]. For full-duplex speech dialogue systems employing automatic speech recognition, the echo attenuation requirements are not as well-defined and will depend on the desired recognition rate as well as on the robustness of the speech recognizer with respect to speech-like interference. In view of these requirements, the echo attenuation provided by microphone arrays and the echo path gain for a microphone array are examined below.

Array gain. The echo attenuation provided by a microphone array is usually identified with the array gain for the desired sources relative to echoes as interference. For signal-independent time-invariant beamforming, the directivity index quantifying the array gain of the desired direction over the average of all other directions [26] does typically not exceed 20 dB over a wide frequency range, and is much smaller at low frequencies (< 500 Hz) due to usual geometrical aperture constraints [19,26]. This contrasts with the fact that acoustic echoes usually exhibit their maximum energy at low frequencies [2]. As a remedy, differential beamforming realizes superdirective array gains at low frequencies and allows for a directivity index of up to 12 dB in practical implementations [1,27]. On the other hand, for adaptive beamforming, interference suppression is usually also limited to about 20 dB for reverberant environments if distortion of the desired source signal $s^{(0)}(n)$ should be precluded. See Chapters 2 and 5 as well as [19,32].

Echo path gain. For microphone array applications, the echo path gain between $u(n)$ and the beamformer output, $y(n)$, will often be higher than for single-microphone systems (-6 dB), because the sum of the distances from the loudspeaker to the listener, and from the desired source to the microphone array, will usually be greater (e.g. in teleconferencing). The user will typically increase the gains for the loudspeaker signal and the microphone array correspondingly to compensate for the decay of the sound level ($\approx$ 6 dB per doubling of distance in the far-field). If the microphone array and loudspeaker are relatively close, then the required echo attenuation will be increased accordingly.

13.4.2 Basic options

Restricting the scenario to a single reproduction channel, $u(n)$, and a single acquisition channel, $\widehat{s}(n)$, a combination of AEC and beamforming is obviously conceivable in two fundamentally different ways as shown in Figure 13.5. Here, 'AEC first' realizes one adaptive filter for each microphone in $\widehat{\mathbf{H}}^{(I)}(n)$ of (13.26), whereas 'Beamforming first' uses a single-channel AEC, $\widehat{\mathbf{h}}^{(II)}(n)$, which obviously has to include the beamformer, $\mathbf{G}(n)$, into its echo path model.

13.4.3 'AEC first'

This structure suggests that $\widehat{\mathbf{H}}^{(I)}(n)$ may operate without any repercussions from the beamforming so that the AEC problem corresponds to that described by (13.23). On the other hand, with perfect echo cancellation, the beamforming will be undisturbed by acoustic echoes and will concentrate on suppressing local interferers and reverberation.

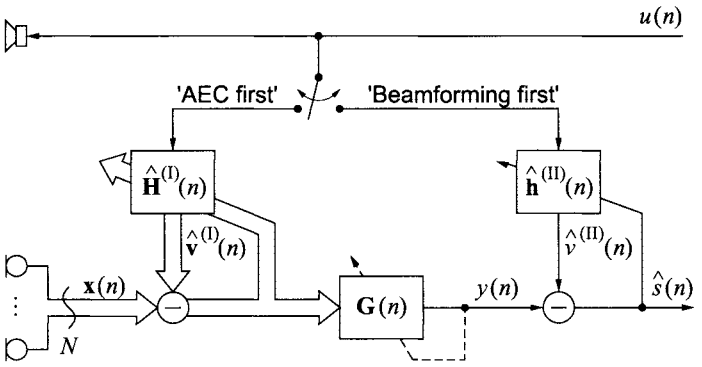

Fig. 13.5. Generic structures for combining AEC with beamforming.

AEC properties. Although AEC could operate independently from the beamforming, synergies with beamforming should be exploited with regard to detection of local source activity and computational complexity.

Local source activity detection. As noted above, the adaptation of $\widehat{\mathbf{H}}^{(I)}(n)$, requires a fast and reliable detection of local source activity to avoid divergence. With single-channel AEC, the detection is based on comparing estimates for

$$Q_\nu(n) = \frac{E\left\{v_\nu^2(n)\right\}}{E\left\{(r_\nu(n) + s_\nu(n))^2\right\}} \tag{13.35}$$

to a given threshold. With subsequent beamforming, this decision can be derived from estimates of

$$Q(n) = \frac{E\left\{\left(\mathbf{G}^T(n)\mathbf{V}(n)\right)^2\right\}}{E\left\{\left(\mathbf{G}^T(n)\left[\mathbf{R}(n) + \mathbf{S}(n)\right]\right)^2\right\}} \tag{13.36}$$

which reflect local source activity much clearer than $Q_\nu(n)$ as $\mathbf{r}_\nu(n)$, $\mathbf{v}_\nu(n)$ are suppressed relative to $\mathbf{s}(n)$ by beamforming. Thus, $Q(n)$ reduces uncertainty in local source activity detection and allows adaptation during time intervals where adaptation might have been stalled if its control was based on $Q_\nu(n)$.

Computational complexity. At least the filtering and the filter coefficient update of the AEC adaptation will require N-fold computational cost compared to a single-channel AEC. Even with continuing growth of the performance-cost ratio of signal processing hardware, this computational load will remain prohibitive in the near future for many cost-sensitive or very large systems employing $N = 5, \ldots, 512$ sensors [17,19,26,30,33,34]. One option to alleviate the computational burden is to reduce the length L_{AEC} in (13.4) of the FIR filter models, $\widehat{\mathbf{h}}_\nu$, and to rely on the beamformer for suppressing the residual

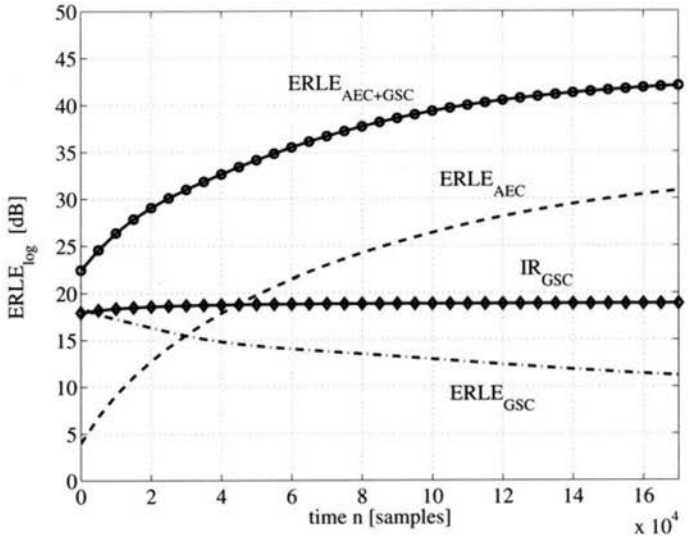

Fig. 13.6. Example for convergence of $ERLE_{log}$ components and local interference suppression(IR) for 'AEC first' structure ($N = 8$, $T_{60} \approx 300$ ms, $f_s = 12$ kHz, $L_{AEC} = 2500$, $L_{BM} = 16$, $L_{IC} = 50$).

echoes, $\mathbf{e}(n)$. Shortening $\widehat{\mathbf{h}}_\nu$ implies however, that the adaptation of the AEC is disturbed by an increased noise component, which is due to the unmodeled tail of the true echo path impulse response, $\mathbf{h}_\nu(n)$ [2].

Beamforming performance. For a signal-independent beamformer, the presence and performance of the AEC has no impact on the beamforming. The signal-independent spatial filtering will increase echo suppression according to its directivity while suppression of local interferers remains unaffected.

Signal-dependent beamformers use $\mathbf{w}(n) = \mathbf{x}(n) - \widehat{\mathbf{v}}^{(I)}(n)$ for optimizing the beamforming filters $\mathbf{G}(n)$. Thereby, at the cost of local interference suppression, the beamformer will concentrate on suppressing echo components, $\mathbf{e}(n)$, if their levels exceed that of local interferers, $\mathbf{r}(n)$, and it will further suppress residual echoes as long as they are not negligibly small compared to the local interferers. For illustration, the typical convergence behaviour for 'AEC first' using a GSC beamformer is shown in Figure 13.6 for $\mathbf{r}(n)$, $\mathbf{s}(n)$, $u(n)$ being white noise signals, and for alternating adaptation of $\mathbf{G}_{IC}(n)$, and $\widehat{\mathbf{H}}^{(I)}(n)$ (see also [32]). Due to its short filters, the beamformer converges almost instantaneously to about $ERLE_{GSC} = 18$ dB, and thereby provides a significant amount of $ERLE_{log}$ long before $\widehat{\mathbf{H}}^{(I)}(n)$ has converged. At the same time, suppression of local interference, IR_{GSC}, remains essentially con-

stant over time, as it converges very rapidly to almost 20 dB and is not allowed to converge much further to preclude distortion of the desired signal.

13.4.4 'Beamforming first'

In this structure, the beamformer is essentially independent from the AEC so that the beamforming performance agrees with Section 13.3 for acoustic echoes being perceived as another source of interference. AEC is realized by a single adaptive filter $\widehat{\mathbf{h}}^{(II)}(n)$ as in Figure 13.5 which is attractive with regard to computational complexity. However, the system identification problem faced by $\widehat{\mathbf{h}}^{(II)}(n)$ requires closer examination.

Echo path for AEC. Incorporating the beamformer, $\mathbf{G}(n)$, into the echo path model means that, ideally, the adaptive filter, $\widehat{\mathbf{h}}^{(II)}(n)$, models the sum of N echo paths from the loudspeaker input, $u(n)$, to the beamformer output, $y(n)$, (see Figure 13.3)

$$\widehat{\mathbf{h}}_{opt}^{(II)}(n) = \mathbf{f}(n) = \sum_{\nu=1}^{N} \mathbf{f}_\nu(n), \tag{13.37}$$

with the impulse responses, $\mathbf{f}(n)$, given by ($\star$ denotes linear convolution):

$$\mathbf{f}_\nu(n) = [f_\nu(0,n), \ldots, f_\nu(L_{AEC+BF}-1, n)]^T, \tag{13.38}$$

$$f_\nu(k,n) = h_\nu(k,n) \star g_\nu(k,n). \tag{13.39}$$

Thus, the impulse response length of $\widehat{\mathbf{h}}^{(II)}(n)$ depends on the beamforming, and, if any $g_\nu(k,n)$ is time-varying, $\widehat{\mathbf{h}}^{(II)}(n)$ has to track this time-variance as well[4]. The required length, L_{AEC+BF}, for $\widehat{\mathbf{h}}^{(II)}(n)$ is essentially the sum of the length L_{BF} and the necessary length for the acoustic path (including loudspeaker and microphone), L_{AEC}:

$$L_{AEC+BF} = L_{AEC} + L_{BF} - 1. \tag{13.40}$$

Note that for a given desired $ERLE_{log}$, L_{AEC} can be chosen smaller than given by (13.4) depending on the expected contribution of beamforming to $ERLE_{log}$ (see also [35]).

Signal-independent, time-invariant beamformers. Due to the time-invariance of $g_\nu(k,n)$, the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ only has to track the time-variance of $\mathbf{h}_\nu(n)$ and, thus, the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ is identical to the adaptation of one of the N filters $\widehat{\mathbf{h}}_\nu^{(I)}(n)$ in the 'AEC first' structure except for the different filter length L_{AEC+BF}.

[4] Note that the time-varying components $h_\nu(k,n)$ cannot be identified separately, although $g_\nu(k,n)$ is known ('knapsack problem').

Signal-dependent, time-varying beamformers. Here, the main problem is that the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ has to track the time-variance of $\mathbf{G}(n)$. As for the adaptation algorithms discussed in Section 13.2.1 an increasing filter order involves a reduced tracking capability [7], the high-order filter, $\widehat{\mathbf{h}}^{(II)}(n)$, cannot follow the time-variance of the low-order filters of $\mathbf{G}(n)$ ($L_{AEC+BF} \gg L_{BF}$). Therefore, $\widehat{\mathbf{h}}^{(II)}(n)$ can find a useful echo path model only when $\mathbf{G}(n)$ remains time-invariant for a sufficiently long time. In Figure 13.7, the adaptation behaviour of the 'beamforming first' structure is analyzed for a speech conversation with a GSC as adaptive beamformer [28,32]. Inspecting the time domain signals $u(n)$ and $s(n)$ in Figures 13.7a and 13.7b shows that a 'double-talk' period occurs for time $n = 3.5 \ldots 4.0 \cdot 10^5$. Figure 13.7c illustrates which component is adapted at a given time. To track slight movements of the desired local source, the blocking matrix, $\mathbf{G}_{BM}(n)$, is adapted if only the local source and noise are present [28,32]. The system error of (13.8) depicted in Figure 13.7d converges monotonically when $\widehat{\mathbf{h}}^{(II)}(n)$ is adapted. When the interference canceller, $\mathbf{G}_{IC}(n)$, or the blocking matrix, $\mathbf{G}_{BM}(n)$, are adapted the system error rises instantaneously ($n = 2 \ldots 3.5 \cdot 10^5$). This is not critical as long as $u(n) = 0$, however, during double-talk ($n = 3.5 \ldots 4.0 \cdot 10^5$), a large residual error, $e(n)$, arises (Figure 13.7e,f) as $\widehat{\mathbf{h}}^{(II)}(n)$ cannot reconverge. Consequently, with the 'beamforming first' structure, the benefits of AEC are missing when they are desired most, i.e., during double-talk and during transitions from far-end activity to local activity and vice-versa (at other times primitive echo suppression methods, such as loss insertion [2], are less objectionable).

13.5 Integration of AEC into time-varying beamforming

As time-varying beamforming, $\mathbf{G}(n)$, cannot be tracked satisfactorily by the adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$, a compromise is desirable for AEC which avoids the computational burden of $\widehat{\mathbf{H}}^{(I)}(n)$ for large N and provides improved echo cancellation compared to $\widehat{\mathbf{h}}^{(II)}(n)$. For this, the beamformer is decomposed into a time-invariant and a time-varying part in the sequel, with AEC acting only on the output of the time-invariant part. Two options for arranging the time-invariant and the time-varying stage are examined: First, a cascade with the time-invariant stage followed by the time-varying stage, and second, a parallel arrangement of the two stages.

13.5.1 Cascading time-invariant and time-varying beamforming

As illustrated in Figure 13.8, instead of a single beamformer output, $y(n)$, (see Figure 13.3), $M < \ldots \ll N$ beamformer output signals $\mathbf{y}(n) =$

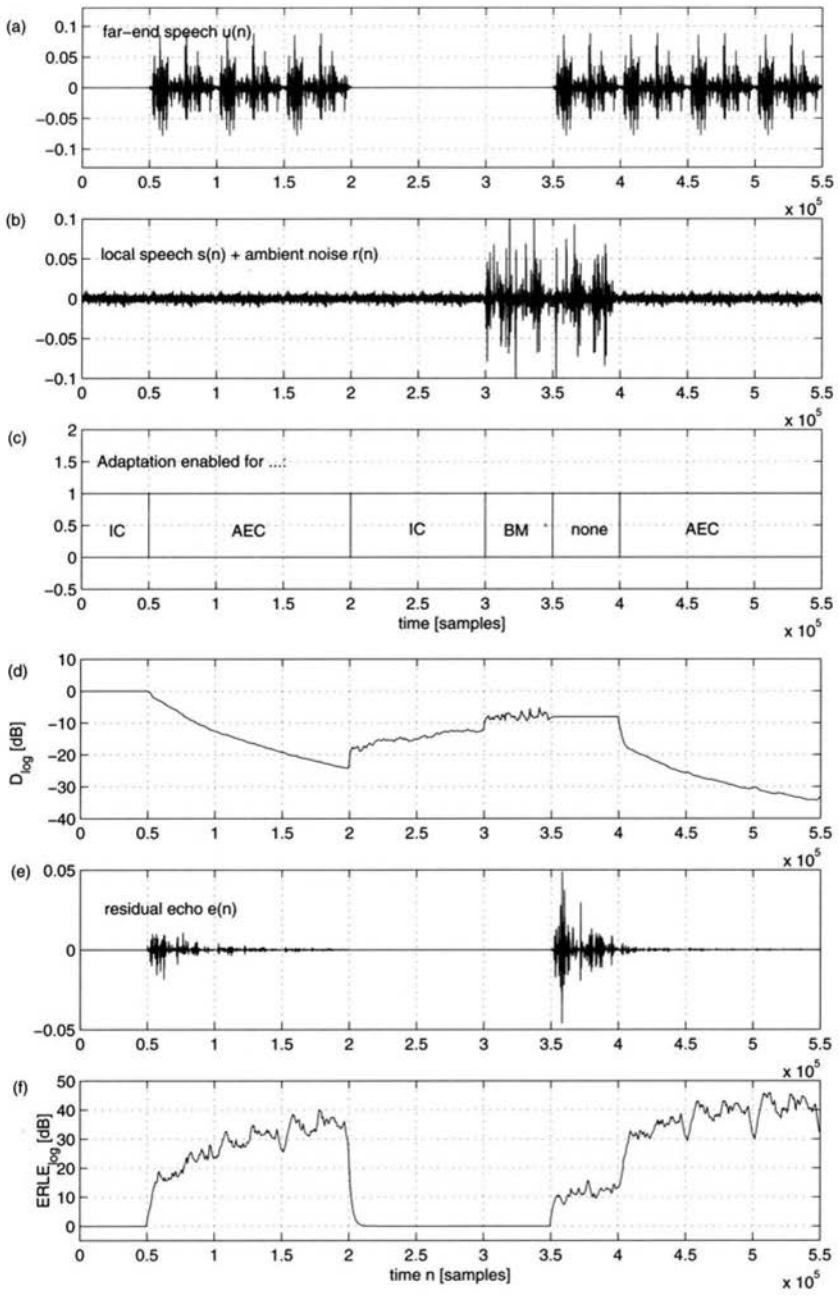

Fig. 13.7. Adaptation of $\widehat{\mathbf{h}}^{(II)}(n)$ in 'beamforming first' structure ($N = 8$, $T_{60} \approx 50$ ms, $f_s = 12$ kHz, $L_{AEC+BF} = 300$, $L_{BM} = 16$, $L_{IC} = 50$, adaptation by NLMS algorithm)

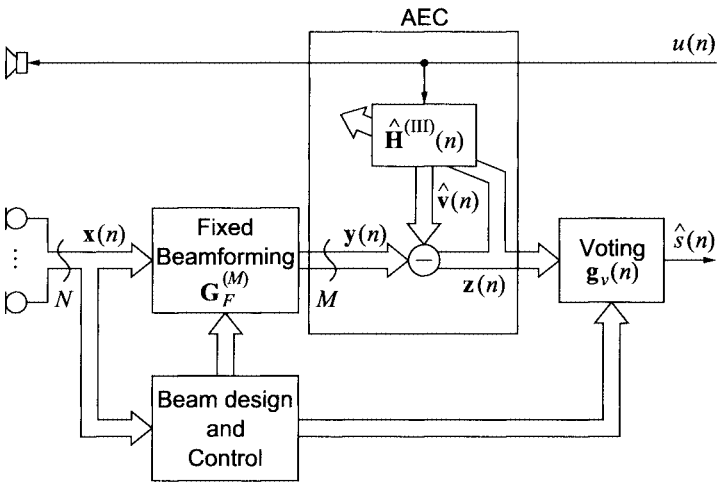

Fig. 13.8. AEC integrated into cascaded beamforming.

$[y_0(n), \ldots y_{M-1}(n)]^T$ are produced by M sets of fixed beamforming filters $\mathbf{G}_F^{(M)}$ according to

$$\mathbf{y}(n) = \mathbf{G}_F^{(M)^T} \cdot \mathbf{X}(n), \tag{13.41}$$

where $\mathbf{X}(n)$ is given by (13.30) and

$$\mathbf{G}_F^{(M)} = [\mathbf{G}_{F,0}^T, \ldots, \mathbf{G}_{F,\mu}^T, \ldots, \mathbf{G}_{F,M-1}^T] \tag{13.42}$$

with $\mathbf{G}_{F,\mu}$ according to (13.28). For AEC, $\widehat{\mathbf{H}}^{(III)}(n)$ realizes M adaptive echo cancellers $\widehat{\mathbf{h}}_\mu(n), \mu = 0, \ldots, M-1$, which exhibit the same performance as $\widehat{\mathbf{h}}^{(II)}(n)$ with time-invariant $\mathbf{G}(n)$ (see Section 13.4.4). Thus, if $M < N$ and $L_{AEC+BF} \approx L_{AEC}$, AEC operates at a reduced computational cost compared to $\widehat{\mathbf{H}}^{(I)}(n)$ (see Section 13.4.3). The time-varying part of the beamforming implements a weighted sum ('voting') using time-varying weights, $g_{v,\mu}(n)$:

$$\widehat{s}(n) = \mathbf{g}_v^T(n) \cdot \mathbf{z}(n) \tag{13.43}$$

with

$$\mathbf{g}_v(n) = [g_{v,0}(n), \ldots, g_{v,\mu}(n), \ldots, g_{v,M-1}(n)]^T, \tag{13.44}$$

$$\mathbf{z}(n) = [z_0(n), \ldots, z_\mu(n), \ldots, z_{M-1}(n)]^T. \tag{13.45}$$

Fixed beamformer design. The fixed beamformers of $\mathbf{G}_F^{(M)}$ may be designed to account for various situations, for instance, different beamformers could be employed for the presence or absence of echo, $\mathbf{v}(n)$, and of certain

local interferers, $\mathbf{r}(n)$. This concept is easily extended to cover several desired sources or moving desired sources, which is especially attractive for teleconferencing [5,17,18,22,26]. For the actual design of $\mathbf{G}_{F,\mu}$, techniques based on both time-invariant or time-varying beamforming can be applied. Updating may be attractive to allow for long-term flexibility.

$\mathbf{G}_F^{(M)}$ *based on time-invariant beamforming.* As a straightforward approach, $M_0 > M$ signal-independent fixed beams may be formed to cover several possible interference scenarios and/or all possible desired source positions. The output of these M_0 beamformers is monitored and a subset of M beamformers is used for $\mathbf{G}_F^{(M)}(n)$ to produce potentially desired signals $\mathbf{y}(n)$. As an example, in a teleconferencing studio with $M_0 = 7$ seats and three local participants being present, only $M = 3$ beams should produce desired signals (for examples see [17,18,22,26]).

$\mathbf{G}_F^{(M)}$ *based on adaptive beamforming.* Signal-dependent adaptive beamforming can be used to identify fixed beamformers for typical interference scenarios. To this end, an adaptive beamformer operates at a normal adaptation rate with its filter coefficients acting as a training sequence for finding M representative fixed beamformers. *A priori* knowledge of the desired source location(s) for incorporating constraints is necessary as well as initial training [5].

Initializing and updating $\mathbf{G}_F^{(M)}$. The monitoring of M_0 fixed beams, or the learning of optimum beamformers for deciding upon $\mathbf{G}_F^{(M)}$ can be carried out during an initial training phase only, or continuously. Continuous monitoring is recommended when changes are expected that demand the updating of $\mathbf{G}_F^{(M)}$. Monitoring of M_0 beams helps also to establish reliable estimates for background noise levels and supports detection of local talker activity so that convergence speed and robustness of AEC adaptation can be improved. Generally, as long as updating of $\mathbf{G}_F^{(M)}$ occurs less frequently than significant changes in the acoustic path, the model of time-invariant beamforming is justified with respect to AEC behavior. Aiming at minimum computational complexity for AEC, more frequent updates of $\mathbf{G}_F^{(M)}$ may be accepted for reduced M. The update should preferably occur at the beginning of 'far-end speech only' periods, as then, the AEC $\hat{\mathbf{H}}^{(III)}(n)$ can immediately adapt to the new echo path.

Voting. The time-varying weights, $g_{v,\mu}(n)$, in (13.44) must be chosen to allow for a fast reaction ($\leq$ 20 ms) to newly active local sources or changing interference scenarios, while at the same time avoiding the perception of switching, e.g., by applying a sigmoïd-like gain increase over time. For maximum spatial selectivity, only one beam signal should have a nonzero weight,

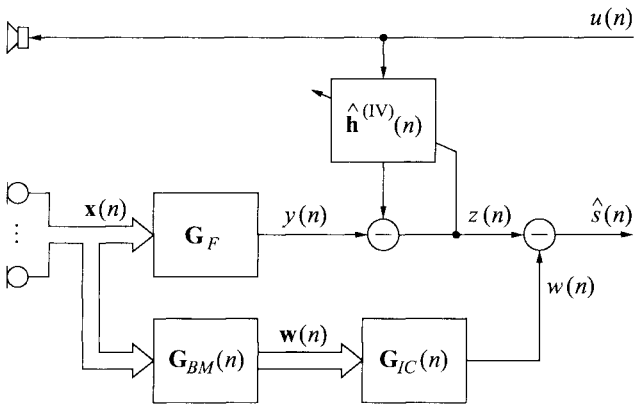

Fig. 13.9. GSC with embedded AEC.

$g_{v,\mu}(n)$, in the stationary case. Frequent toggling between beams is subjectively objectionable and should be prevented by hysteresis mechanisms (see also [17,26]).

13.5.2 AEC with GSC-type beamforming structures

As a popular representative of adaptive beamformers, the GSC (see Section 13.3.3) is also an example for a parallel arrangement of time-varying and time-invariant beamforming. If an integrated AEC should only see time-invariant beamforming in the echo path, it has to act on the output of the fixed beamformer, $y(n)$, as depicted in Figure 13.9 [32]. Obviously, only a single adaptive filter, $\widehat{\mathbf{h}}^{(IV)}(n)$, is necessary which faces the same system identification task as $\widehat{\mathbf{h}}^{(II)}(n)$ for time-invariant beamforming (see Section 13.4.4), which in turn is essentially identical to the plain single-microphone AEC problem. However, residuals of acoustic echoes, $\mathbf{v}(n)$, will also be contained in $w(n)$ unless they are eliminated by $\mathbf{G}_{BM}(n)$ or $\mathbf{G}_{IC}(n)$. Here, leaving echo suppression to the interference canceller, $\mathbf{G}_{IC}(n)$, seems to be the obvious solution. Recall that $\mathbf{G}_{IC}(n)$ minimizes the quadratic norm of $\widehat{s}(n)$ to remove all components from $z(n)$ that are correlated with $\mathbf{w}(n)$. If $\widehat{\mathbf{h}}^{(IV)}(n)$ is perfectly adjusted, no echo components remain in $z(n)$ and the echo estimate within $w(n)$ should be zero. On the other hand, local interference components in $\mathbf{w}(n)$ should be linearly combined using nonzero filter coefficients, so that $w(n)$ can remove interference residuals from $z(n)$. Clearly, a conflict in the design of $\mathbf{G}_{IC}(n)$ arises [32].

For illustration, consider a stationary situation for a given frequency, ω_0, in a 2-D plane containing a linear beamforming array with time-invariant $\mathbf{G}_F$, $\mathbf{G}_{BM}$, and $\mathbf{G}_{IC}$. A local interferer, $\mathbf{r}(n)$, arrives as a planar wave from θ_0 and passes the blocking matrix which is transparent for $\mathbf{r}(n)$ ($\mathbf{G}_{BM}^T \cdot \mathbf{r}(n) = \mathbf{r}(n)$).

Then, for perfect interference cancellation, $\mathbf{G}_{IC}(n)$ has to model the response of the fixed beamformer, $\mathcal{F}\{\mathbf{G}_{IC}\}(\theta_0,\omega_0) = \mathcal{F}\{\mathbf{G}_F\}(\theta_0,\omega_0)$, with $\mathcal{F}\{\mathbf{G}_{(\cdot)}\}(\theta,\omega)$ denoting the frequency response for a plane wave of frequency ω arriving from θ. If, on the other hand, an acoustic echo arrives from the same direction, θ_0, with nonzero spectral support at ω_0, this should be perfectly suppressed if $z(n)$ contains no echo, which means $\mathcal{F}\{\mathbf{G}_{IC}\}(\theta_0,\omega_0) = 0$. Obviously, this conflict requires a compromise at the cost of either local interference suppression or echo attenuation. Here, adaptation algorithms will automatically favor the dominant signal component in $\mathbf{w}(n)$. Even if echo and local interference do not arrive from the same direction, the finite number of degrees of freedom limits the ability of $\mathbf{G}_{IC}$ to suppress echo and local interference simultaneously. This is especially true for reverberant environments which possess a very large (if not infinite) number of DOAs for both echoes and local interference.

To avoid the conflict of interests within $\mathbf{G}_{IC}$, a suppression of the acoustic echoes, $\mathbf{v}(n)$, using $\mathbf{G}_{BM}(n)$ seems attractive. Considering the options, it is obvious that the output, $\mathbf{w}(n)$, should be freed from $\mathbf{v}(n)$ without suppressing $\mathbf{r}(n)$ or impairing the suppression of $\mathbf{s}(n)$. This means that no additional filtering on $\mathbf{x}(n)$ is allowed. As an alternative, estimates for the echoes, $\mathbf{v}(n)$, could be subtracted from $\mathbf{w}(n)$, which requires one adaptive filter for each of the $P \leq N$ channels and is similar to the generic concept of Section 13.4.3.

13.6 Combined AEC and beamforming for multi-channel recording and multi-channel reproduction

Multi-channel recording means that the output of the acquisition part of the acoustic interface in Figure 13.1 consists of several ($L > 1$) channels which, e.g., are necessary to convey spatial information for remote multi-channel sound reproduction, but may also support local signal processing. In Figures 13.5, 13.6, and 13.7 this translates to an L-dimensional output vector $\widehat{\mathbf{s}}(n)$. With respect to the beamforming, it means a duplication of the filtering and adaptation for each channel using the techniques outlined in Section 13.3. Both, time-invariant and adaptive beamforming will typically use L different 'look directions.' Regarding the generic methods to combine AEC with beamforming (Section 13.4), this means that for the 'AEC first' structure, the AEC part, $\widehat{\mathbf{H}}^{(I)}(n)$, remains unchanged while only the beamforming has to be duplicated. For the 'beamforming first' structure, the AEC realized by $\widehat{\mathbf{h}}^{(II)}(n)$ has to be duplicated as well.

When AEC is integrated into cascaded beamforming (see Section 13.5.1) the extension to the multi-channel recording case is included in the concept. The number of parallel fixed beams simply must equate or exceed the number of recorded channels, $M \geq L$, and the voting must be chosen accordingly. The AEC structure, $\widehat{\mathbf{H}}^{(III)}(n)$, remains unchanged. If the AEC is embedded into

a GSC-like structure, both the beamforming, $\mathbf{G}(n)$, and the AEC structure, $\widehat{\mathbf{h}}^{(IV)}(n)$, have to be implemented L times. However, removal of the acoustic echoes in the blocking matrix is necessary only once if performed directly on the microphone signals, $\mathbf{x}(n)$.

Multi-channel reproduction introduces a K-channel AEC problem (as described in Section 13.2.2), wherever a single echo cancellation filter is employed for single-channel reproduction, regardless of whether echo is to be removed from a microphone output or from a beamformer output. Essentially, this deteriorates convergence behavior and increases computational complexity for all structures discussed in Sections 13.4 and 13.5, accordingly.

Finally, for a system with both multi-channel reproduction and multi-channel recording as suggested in Figure 13.1, the complexity for combined AEC and beamforming obeys the superposition principle with respect to filtering and filter adaptation. Synergies are obtained by the common use of control information for several channels. The nature of the problems, however, does not change compared to the basic scenarios studied in Sections 13.2.2, 13.4, and 13.5 so that the corresponding results remain meaningful.

13.7 Conclusions

Beamforming and acoustic echo cancellation have been shown to jointly contribute to the suppression of acoustic feedback occurring in hands-free acoustic man-machine interfaces. While for time-invariant beamforming a single adaptive AEC filter suffices in the case of single-channel reproduction and single-channel recording, time-varying beamformers demand multiple adaptive filters if echo cancellation performance is not to degrade severely. However, realizing a time-varying beamformer as a cascade of time-invariant beamforming and time-varying voting requires only a few adaptive echo cancellers even for microphone arrays with many sensors. Implementing a combination of AEC and beamforming for a multi-channel reproduction and multi-channel recording system involves a corresponding increase in computational complexity. Signal processing performance, however, is still determined by the solutions for the elementary problems.

Acknowledgement

The author wishes to thank Wolfgang Herbordt for providing the simulation results and Susanne Koschny for preparing the illustrations.

References

1. S.L. Gay and J. Benesty, eds., *Acoustic Signal Processing for Telecommunication*, Kluwer, 2000.

2. C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, July 1999.
3. M.M. Sondhi and W. Kellermann, "Echo cancellation for speech signals," in *Advances in Speech Signal Processing*, (S. Furui and M.M. Sondhi, eds.), Marcel Dekker, 1991.
4. A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, pp. 1747–1760, 2000.
5. W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.219–222, Apr. 1997.
6. G.-O. Glentis, K. Berberidis, and S. Theodoridis. "Efficient least squares adaptive algorithms for FIR transversal filtering," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 13–41, July 1999.
7. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.
8. J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
9. W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-88)*, New York NY, USA, pp.2570–2573, Apr. 1988.
10. M.M. Sondhi, D.R. Morgan, and J.L. Hall, "Stereophonic echo cancellation: An overview of the fundamental problem," *IEEE Signal Processing Letters*,vol. 2, no. 8, pp. 148–151, Aug. 1995.
11. S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp.3059–3062, May 1995.
12. J.Benesty, D.R. Morgan, and M.M. Sondhi, "A hybrid mono/stereo acoustic echo canceler," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 468–475, Sept. 1998.
13. B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
14. D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1993.
15. R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice Hall, 1983.
16. R.G. Pridham and R.A. Mucci, "Digital interpolation beamforming for low-pass and bandpass signals," *Proceedings of the IEEE*, vol. 67, no. 6, pp. 904–919, June 1979.
17. W. Kellermann, "A self-steering digital microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-91)*, Toronto, Canada, pp.3581–3584, May 1991.
18. J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985.
19. C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

20. T. Chou, "Frequency-independent beamformer with low response error," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp.2995–2998, May 1995.
21. Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, Dec. 1986.
22. P. Chu, "Desktop mic array for teleconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp.2999–3002, May 1995.
23. M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.239–242, Apr. 1997.
24. P. Chu, "Superdirective microphone array for a set-top videoconferencing system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.235–238, Apr. 1997.
25. L.J. Griffiths and C.W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
26. J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, 1991.
27. G. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, pp. 229–240, 1996.
28. O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix unsing constrained adaptive filters, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-96)*, Atlanta GA, USA, pp.925–928, May 1996.
29. I. Claesson, S.E. Nordholm, B.A. Bengtsson, and P. Eriksson, "A multi-DSP implementation of a broad-band adptive beamformer for use in a hands-free mobile radio telephone," *IEEE Trans. on Vehicular Technology*, vol. 40, no. 1, pp. 194–202, Feb. 1991.
30. S. Oh, V. Viswanathan, and P. Papamichalis, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-92)*, San Francisco CA, USA, pp. I:281–I:284, Mar. 1992.
31. ITU-T recommendation G.167, *Acoustic echo controllers*, Mar. 1993.
32. W. Herbordt and W. Kellermann, "GSAEC - Acoustic echo cancellation embedded into the generalized sidelobe canceller," in *Signal Processing X: Theories and Applications (Proceedings of EUSIPCO-2000)*, Tampere, Finland, vol.3, pp.1843–1846, Tampere, Finland, Sept. 2000.
33. S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson, "Noise reduction using an adaptive microphone array in a car – a speech recognition evaluation," in *Conference Recordings of the ASSP Workshop on Application of Digital Signal Processing to Audio and Acoustics*, New Paltz NY, USA, Oct. 1993.
34. H. Silverman, W. R. Patterson, J.L. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp.251–254, Apr. 1997.

35. W. Kellermann, "Some properties of echo path impulse responses of microphone arrays and consequences for acoustic echo cancellation," in *Conf. Rec. of the Fourth International Workshop on Acoustic Echo Control*, Røros, Norway, June 1995.

14 Optimal and Adaptive Microphone Arrays for Speech Input in Automobiles

Sven Nordholm[1], Ingvar Claesson[2], and Nedelko Grbić[2]

[1] Curtin University of Technology, Perth, Australia
[2] Blekinge Institute of Technology, Ronneby, Sweden

Abstract. In this chapter, speech enhancement and echo cancellation for hands-free mobile telephony are discussed. A number of microphone array methods have been tested and results from car measurements are given. Traditional methods such as linearly constrained beamforming and generalized sidelobe cancelers are discussed as well as array gain optimization methods. An *in situ* calibrated method which gives an overall improved performance is also presented. Algorithms such as Least-Squares (LS) and Normalized-Least-Mean-Squares (NLMS) are used to find optimal weights. Improved performance using an LS-method is shown, but at the cost of increased numerical complexity limiting its implementation in real-time applications. By introducing subband processing this issue can be avoided. The results show a noise suppression of approximately 18 dB and hands-free loudspeaker suppression of the same order.

14.1 Introduction: Hands-Free Telephony in Cars

The increased use of mobile telephones in cars has created a greater demand for hands-free, in-car installations. The advantages of hands-free telephones are safety and convenience. In many countries and regions hand-held telephony in cars is prohibited by legislation. The car manufacturers also prohibit such use since it will interact with other electronic devices in the car such as air bags, navigation equipment, etc. This means that a mobile telephone should be properly installed and an external antenna should be used. However, by installing the microphone far away from the user, a number of disadvantages, such as poor sound quality and acoustic feedback from the far-end side, are introduced. This means that some form of filtering is required in order to obtain sound quality comparable to that of hand-held telephony. This filtering operation must suppress the loudspeaker, as well as background noise and room reverberation, without causing severe speech distortion. A number of potential methods will be presented to address this problem.

For automobile applications, there has also been the desire to replace many hand-controlled functions with voice controls. The signal degradations in this context have many similarities to those encountered in distant-talker speech recognition applications. A study of recognition in car environments was presented in [1,2]. However this topic is beyond the primary goal here and is the specific subject of Chapter 15.

Hands-free car installations result in noisy near-end speech as well as an acoustic feedback of the far-end speech. The near-end disturbances, resulting in substantial speech distortion, are mainly room reverberation and car noise. Furthermore, acoustic feedback, generated at the near-end side, is a problem for the far-end side talker, who will hear his voice echoed with 100-200 ms delay, making speech conversation substantially more difficult. Three major tasks must be addressed in order to improve the quality of hands-free mobile telephones: noise suppression, room reverberation suppression, and acoustic feedback suppression of the hands-free loudspeaker. Because of the cabin conditions, room reverberation suppression is not a critical issue in most standard automobiles. In trucks, buses, people movers and 4-WD with their larger interiors, it may need to be considered. The measurements presented here are from a normal-sized station wagon. The acoustic channel is non-minimum phase and thus quite hard to deconvolve [3]. Matched filtering approaches which do not require explicit channel deconvolution [3,4] and several other methods detailed earlier in this text are available for reverberation suppression under more adverse conditions.

Speech enhancement in hands-free mobile telephony can be performed using spectral subtraction [5–8] or temporal filtering such as Wiener filtering, noise cancellation and multi-microphone methods using a variety of different array techniques [9–11]. Room reverberation in this context is most effectively handled using array techniques or by proper microphone design and placement. Acoustic feedback for hands-free mobile telephony is usually addressed by conventional echo cancellation techniques [12–15] although subband echo cancellation has been popular lately, see for instance [15,16].

A broad-band microphone array is able to perform all the given tasks, i.e. speech enhancement, echo cancellation and reverberation suppression, in a concise and effective manner. This is due to the fact that the spatial domain is utilized as well as the time domain. An effective combination of spatial and temporal processing will lead to a very efficient solution. Hence, improved speech enhancement performance is achieved compared to single microphone solutions. The microphone array technique also handles the acoustic feedback in an efficient way. The hands-free loudspeaker represents a single source despite having been filtered by the channel associated with the car's interior. Similarly, the main talker (driver) represents an additional single source within the cabin. These two sources will have different locations. The echo-suppression level and speech distortion will depend on how "well apart" these two sources are placed [17].

The outline of this chapter is as follows:

- Section 2, Optimum and Adaptive Beamforming topics are reviewed from a hands-free mobile telephone perspective, specifically:
  1. Signal Model
  2. Constrained Minimum Variance Beamforming and Generalized Sidelobe Canceler (GSC)

3. *In situ* Calibrated Microphone Array
4. Time-Domain Minimum-Mean-Square-Error Beamformer
5. Frequency-Domain Minimum-Mean-Square-Error Beamformer
6. Optimal Near-field Signal-to-Noise plus Interference Beamformer (SNIB)

- Section 3, Subband Implementation of the Microphone Array
  1. Description of LS-Subband Beamforming
- Section 4, Multi-resolution Time-Frequency Adaptive Beamforming
  1. Complexity Comparisons
- Section 5, Evaluation and examples
- Section 6, Summary and conclusions

14.2 Optimum and Adaptive Beamforming

The hands-free mobile telephony problem in an automobile is well suited for optimum or adaptive beamforming. The user is in a fixed and known location relative to the array and enclosure. The geometrical array configuration is known. It is further possible to place the loudspeaker in a position that is advantageous from a beamforming perspective. Early approaches to this task involved the direct adoption of adaptive antenna array methods in use since the 1960's [18]. However, this proved not to be a straightforward task and required the development of approaches specific to the application environment, e.g. [11,19,10,20].

The most common means of applying adaptive beamforming concepts is to treat the problem as one of constrained optimization. These methods rely on geometrical constraints, where the location of the source is known either perfectly or with some accuracy. They require sensor calibration and stable hardware (e.g. to avoid low temperature drift). The algorithms may be extended using different robustness constraints [21–23] and noise sub-space constraints [3]. The problem may also be viewed as several multi-dimensional filtering problems. These filters are combined with an interference cancelling structure [10,24,25].

14.2.1 Common Signal Modeling

In order to provide a consistent description it will be useful to introduce a simple signal model which is general in the sense that microphone elements and sources with any spectral content can be placed arbitrarily. The D different point signal sources $s_d(t)$, $d = 1, 2, \ldots, D$, with spectral densities $R_{s_d s_d}(\omega)$ are assumed to be mutually uncorrelated, i.e. the cross power spectral density $R_{s_d s_e}(\omega)$ is zero if $d \neq e$. All sources impinge on an array of N microphone elements, corrupted with non-directional independent diffuse additive noise $v(t)$. The transfer function between source d and array element n is denoted $G_{d,n}(\omega)$ and is either measured or modeled. In this model, a spherical source

in a free-field and homogeneous medium is assumed. In a real world situation with measured data such an assumption may be suspect. The signal received at the n^{th} microphone element, $x_n(t)$, is

$$x_n(t) = \sum_{d=1}^{D} s_d(t) \star g_{d,n}(t) + v(t) \quad n = 1, 2, \ldots, N, \tag{14.1}$$

where $\star$ denotes convolution. Each source signal is treated as a point source filtered by an LTI system. An implication of these assumptions is that variations in the acoustic channel are slow relative to the filter update rate. In the sequel all signals are assumed to be bandlimited and sampled with a discrete-time index k.

14.2.2 Constrained Minimum Variance Beamforming and the Generalized Sidelobe Canceler

In minimum variance beamforming, the objective is to minimize the output of a (broadband) array while maintaining a constant gain constraint towards the desired source, in this case the talker of interest.

The output of the beamformer is given as

$$y[k] = \sum_{n=1}^{N} \mathbf{w}_n^H \mathbf{x}_n[k] \tag{14.2}$$

where the weight vector and input data vector both are of length L.

The expression to be minimized is the "variance" of the assumed zero-mean output, $E(|y[k]|^2)$, with respect to the filter weights given by

$$r_{yy}[0] = E(y[k]y^H[k]) = \textstyle\sum_{n=1}^{N} \mathbf{w}_n^H E(\mathbf{x}_n[k]\mathbf{x}_n^H[k])\mathbf{w}_n. \tag{14.3}$$

If it is assumed, without loss of generality, that point source one is the talker of interest, then the major task is to find the constraint on the weight vector such that $y[k] = s_1(t)|_{t=kT}$, i.e. the output is distortion free. A natural way to do this is to express the minimization in the frequency domain and include matched filtering [3]. For this process, there is a strict requirement of accurate signal modeling or robust constraints, otherwise super-resolution will cancel the source [23,10].

The Generalized Sidelobe Canceler can be viewed as a constrained beamformer which has been converted to a non-constrained beamformer by means of a blocking matrix. Thus, the problem is separated into two tasks: the design of a fixed beamformer to determine the response for the desired source and a matrix filter that blocks the desired source from entering. In the simplest case of a free-field, far-field source and a perfectly calibrated array this blocking matrix will amount to a point constraint [26,27]. For the near field situation and a reverberant enclosure, special measures must be taken. The

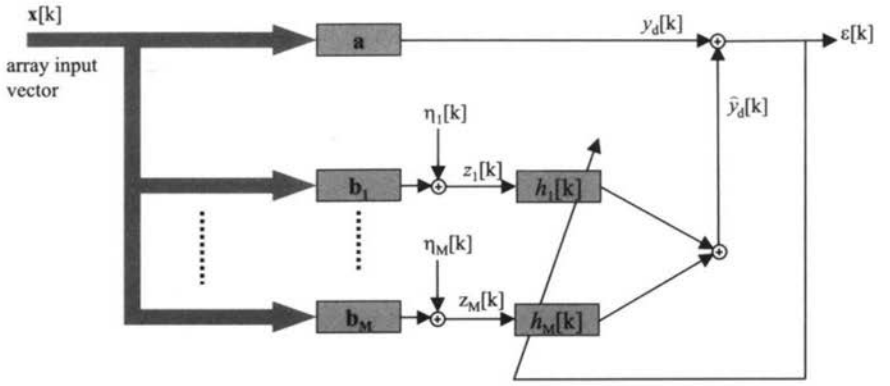

Fig. 14.1. Structure of the Generalized Sidelobe Canceler.

original form of the GSC only provides for a point constraint and is excessively sensitive to calibration and direction errors [23,22]. A number of methods have been proposed which are more suitable for microphone array applications [27,28,21,10,29]. Details of this appropriate GSC implementation will now be presented.

The GSC structure shown in Figure 14.1 consists of two main parts: an upper fixed beamformer and a blocking matrix with subsequent interference cancelers. In order to avoid attenuation of the desired signal it is critical that the input to these interference cancelers contain only the undesired signals.

The input signal vector, $\mathbf{x}[k]$, is filtered by the upper beamformer, $\mathbf{a}$, steering towards the talker of interest and creating an output $y_d[k]$,

$$y_d[k] = \mathbf{a}^H \mathbf{x}[k]. \tag{14.4}$$

This beamforming filter in its simplest form consists of a vector of ones. More generally, it consists of FIR filters forming a multi-dimensional filter. The blocking matrix should form signals that are orthogonal to the desired signal. Thus, the input to the interference cancelers should contain only undesired signals, (and some injected noise)

$$z_m[k] = \mathbf{b}_m^T \mathbf{x}[k] \quad (+\eta_m[k]). \tag{14.5}$$

When designing the lower beamformers, $\mathbf{b}_m$, which implement the signal blocking matrix, the requirement is that the desired signal should be blocked totally. This is not practically feasible. To do so would require knowledge of the transfer function from the desired source to the input of the lower beamformers with extremely high precision. By choosing to relax this requirement and viewing the problem from a filter design perspective where the desired signal should be suppressed below a certain level determined by an artificial injected noise level, $\eta_m[k]$, one may overcome these limitations [10]. This injected noise is not actually present, it is only included in the filter weight

updating algorithm used in the adaptive implementation of the interference canceler. The desired signal will not be picked up and attenuated by the interference canceler as long as the injected noise dominates over the desired signal. This approach is also valid for the background noise free case [25].

Another approach to this constrained optimization problem is the use of subspace techniques such as that suggested in [3]. This method requires several adaptive steps and also a Voice Activity Detector (VAD). The speech distortion is related to how well the transfer functions from the desired source to each microphone element, $G_{1i}(\omega)$, are identified. The upper beamformer is then created as a matched spatial temporal filter and the blocking matrix is created as a projection matrix that is orthogonal to the transfer function vector $(G_{11}(\omega), G_{12}(\omega), \ldots, G_{1N}(\omega))$. This implies that, as long as this orthogonality constraint is valid, no target signal will leak through. All of this assumes that a good estimate of the transfer function vector is used, the talker can be represented by a point source, and the conditions are time invariant.

Experience using the GSC has shown that it provides a very good suppression of background noise, but that control of the signal distortion and calibrating for a combined array are problematic [10]. Another observation reached from implementation experience is the importance of using a precise VAD. The interference canceler is very effective at exploiting correlations with the target and adjusting its weights to suppress or heavily distort the desired signal. A combination of VAD and leaky LMS was used in the implementation [10] to give a reasonable result. Still it was difficult to obtain satisfactory results with long term tests in a car, i.e. over a few days of measurements using an initial calibration. This suggests the need to have a means for very simple *in situ* calibration.

14.2.3 *In Situ* Calibrated Microphone Array (ICMA)

The basic idea when developing this scheme was to find a robust yet effective strategy to an undistorted version of the desired signal with significant suppression of background noise and unwanted sources. A primary goal was to overcome the environmental sensitivity inherent in the constrained optimization strategies outlined above. A way to achieve this is to record calibration sequences through the actual system in a real situation with all of its imperfections. These calibration sequences contain information regarding the statistical properties of the speaker, from both a spatial and temporal point of view. All calibration signals are gathered from the correct position and with the actual hardware. The adaptive system, as such, is then designed not to suppress signals close to the calibration point, i.e it should have low sensitivity to perturbation errors and avoid super-resolution. This can be achieved by moving the source (spatial dithering) around the nominal point during calibration or exploiting temporal dithering in the A/D converters. Calibration sequences are recorded from both the jammer position(s) and the target position when *no car noise is present.* These signals are stored in memory for

later use as training signals in an adaptive phase. As will be shown it is only necessary to save the second order statistics of the calibration signals in the implementation phase. This approach gives an inherent calibration where it is possible to average and weigh interesting frequency bands, microphones, and spatial points. The methodology does not rely on any geometric *a priori* information or array element similarities with accurate positioning. The result is a system that is tailored for the actual situation. The system has been studied from a theoretical [17] and implementation perspectives [30–32]. The ICMA uses a Minimum-Mean-Square-Error (MMSE) optimization that is approximated by either an NLMS implementation [30,31] or an LS solution [33]. An LS or Recursive-Least-Squares (RLS) solution becomes practical when using a subband implementation. The ICMA design can be viewed as an MMSE beamformer where there is separate access to the undesired noise and desired speech signal.

14.2.4 Time-Domain Minimum-Mean-Square-Error Solution

Assume that the input to the beamformer consists of a sum of known calibration sequence observations, $s_n[k]$, $n = 1 \ldots N$, sent out from the position of interest, and noise-plus-interference signals, $x_n[k]$, $n = 1 \ldots N$, consisting of the actual environment signal observations. The time-domain objective can be formulated as

$$\mathbf{w}_{opt} = \arg\min_{\mathbf{w}} \quad E\left[(y[k] - s_r[k])^2\right] \tag{14.6}$$

where the output, $y[k]$, from the beamformer is given by

$$y[k] = \sum_{n=1}^{N} \mathbf{w}_n^H \left(\mathbf{x}_n[k] + \mathbf{s}_n[k]\right). \tag{14.7}$$

The desired signal, $s_r[k]$, is chosen from a single calibration array sensor observation, $s_n[k]$, or a separate reference microphone signal chosen as the reference sensor. In theory the true source signal would be desirable to use instead of a sensor observation, but the true source signal is simply not accessible in a noise-filled car. The optimal weights which minimizes the mean square error between the output and the reference signal are found by [34]

$$\mathbf{w}_{opt} = \left[\mathbf{R_{ss}} + \mathbf{R_{vv}}\right]^{-1} \mathbf{r}_s. \tag{14.8}$$

where $\mathbf{R_{ss}}$ is defined as

$$\mathbf{R_{ss}} = \begin{pmatrix} \mathbf{R}_{s_1 s_1} & \mathbf{R}_{s_1 s_2} & \cdots & \mathbf{R}_{s_1 s_N} \\ \mathbf{R}_{s_2 s_1} & \mathbf{R}_{s_2 s_2} & \cdots & \mathbf{R}_{s_2 s_N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{s_N s_1} & \mathbf{R}_{s_N s_2} & \cdots & \mathbf{R}_{s_N s_N} \end{pmatrix} \tag{14.9}$$

and

$$\mathbf{R}_{s_m s_n} = \begin{pmatrix} r_{s_m s_n}[0] & r_{s_m s_n}[1] & \dots & r_{s_m s_n}[L-1] \\ r^*_{s_m s_n}[1] & r_{s_m s_n}[0] & \dots & r_{s_m s_n}[L-2] \\ \vdots & \vdots & \ddots & \vdots \\ r^*_{s_m s_n}[L-1] & r^*_{s_m s_n}[L-2] & \dots & r_{s_m s_n}[0] \end{pmatrix} \tag{14.10}$$

with

$$r_{s_m s_n}[l] = E\{s_m[k]s_n[k+l]\} \qquad l = 0, 1, \cdots, L-1 \tag{14.11}$$

The noise correlation matrix, $\mathbf{R_{vv}}$, is defined similarly and consists of the correlation estimates of all noise plus interference signals. The filter weights, $\mathbf{w}$, are arranged according to

$$\mathbf{w}^T = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ \dots \ \mathbf{w}_N^T] \tag{14.12}$$

where

$$\mathbf{w}_n^T = [w_n[0] \ w_n[1] \ \dots \ w_n[L-1]] \qquad n = 1, 2, \cdots, N. \tag{14.13}$$

The cross correlation vector, $\mathbf{r_s}$, is defined as

$$\mathbf{r_s} = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \dots \quad \mathbf{r}_N] \tag{14.14}$$

with

$$\mathbf{r}_n = [r_n[0] \quad r_n[1] \quad \dots \quad r_n[L-1]] \qquad n = 1, 2, \cdots, N \tag{14.15}$$

and each element as

$$r_n[l] = E[s_n[k]s_r[k+l]] \tag{14.16}$$

$$n = 1, 2, \cdots, N, \quad r \in [1, 2, \cdots, N], \quad l = 0, 1, \cdots, L-1.$$

14.2.5 Frequency-Domain Minimum-Mean-Square-Error Solution

The formulation of the MMSE beamformer can be expressed in terms of individual frequency bands. The optimal beamformer consists of the frequency-dependent weights that minimize the mean square error across the individual frequency bands. This is provided that the different frequency bands are essentially independent and that the fullband signal can be represented accurately via this subband decomposition. The frequency domain design criterion

can be formulated in a fashion similar to that of the time domain. For each subband with center frequency, f, the criterion will be

$$\mathbf{w}_{opt}^{(f)} = \arg\min_{\mathbf{w}^{(f)}} \quad E\left[|y^{(f)}[k] - s_r^{(f)}[k]|^2\right] \tag{14.17}$$

where the output, $y^{(f)}[k]$, from the beamformer is given by

$$y^{(f)}[k] = \sum_{n=1}^{N} w_n^{(f)} \left[x_n^{(f)}[k] + s_n^{(f)}[k]\right] \tag{14.18}$$

where $x_n^{(f)}[k]$, $s_n^{(f)}[k]$ and, $y^{(f)}[k]$ are narrow-band signals containing essentially only components of frequency f. The single sensor observation, $s_r[k]^{(f)}$, is again one of the microphone observations chosen as the reference sensor. The optimal weights, which minimize the mean square error between the output and the reference signal for each frequency band, are found by

$$\mathbf{w}_{opt}^{(f)} = \left[\mathbf{R}_{ss}^{(f)} + \mathbf{R}_{vv}^{(f)}\right]^{-1} \mathbf{r}_s^{(f)}. \tag{14.19}$$

where

$$\mathbf{R}_{ss}^{(f)} = E\{\mathbf{s}^{(f)}[k]\mathbf{s}^{(f)}[k]^H\} \tag{14.20}$$

where

$$\mathbf{s}^{(f)}[k] = \left[s_1^{(f)}[k] \quad s_2^{(f)}[k] \quad \cdots \quad s_N^{(f)}[k]\right]^T. \tag{14.21}$$

Here each signal, $s_n^{(f)}[k]$, $\quad n = 1, 2, \cdots, N$, is the narrow-band observation when only the source signal of interest is active. The correlation matrix $\mathbf{R}_{vv}^{(f)}$ is defined similarly where each microphone observation consists of only the noise and interference signals. The cross correlation vector, $\mathbf{r}_s^{(f)}$, is defined as

$$\mathbf{r}_s^{(f)} = [r_1^{(f)} \quad r_2^{(f)} \quad \ldots \quad r_N^{(f)}] \tag{14.22}$$

with each element as

$$r_n^{(f)} = E[s_n^{(f)}[k]s_r^{(f)}[k]] \tag{14.23}$$

$$n = 1, 2, \cdots, N, \quad r \in [1, 2, \cdots, N].$$

The frequency dependent weights, $\mathbf{w}^{(f)}$, are defined as complex valued vectors of dimension N.

14.2.6 Optimal Near-Field Signal-to-Noise plus Interference Beamformer

The output signal-to-noise plus interference power ratio (SNIR) is defined as

$$Q = \frac{\text{average signal output power}}{\text{average noise-plus-interference output power}} \tag{14.24}$$

and the beamformer which maximizes the ratio, Q, is the optimal SNIB. Expressing the mean signal output power as a function of the filter weights in the beamformer and finding the optimal weights which maximize Q is done below.

Time-Domain Formulation The beamformer output power when only the signal of interest, $s[k]$, is active, is found from the zero lag of the autocorrelation function, $r_{y_s y_s}[0]$, as

$$r_{y_s y_s}[0] = \mathbf{w}^H \mathbf{R}_{\mathbf{ss}} \mathbf{w} \tag{14.25}$$

The matrix, $\mathbf{R}_{\mathbf{ss}}$, is defined in (14.9). The weights, $\mathbf{w}$, are arranged as in (14.12) and (14.13).

An expression for the noise-plus-interference power, $r_{y_v y_v}[0]$, is found from

$$r_{y_v y_v}[0] = \mathbf{w}^H \mathbf{R}_{\mathbf{vv}} \mathbf{w} \tag{14.26}$$

when all the surrounding noise sources are active and the source signal of interest is inactive.

Now, the optimal weights are found by maximizing the ratio of two quadratic forms, according to

$$\mathbf{w}_{opt} = \arg\max_{\mathbf{w}} \left\{ \frac{\mathbf{w}^H \mathbf{R}_{\mathbf{ss}} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{\mathbf{vv}} \mathbf{w}} \right\}. \tag{14.27}$$

Frequency-Domain Formulation The formulation of the optimal signal-to-noise plus interference beamformer may be derived for individual frequency subbands. The weights that maximize the quadratic ratios at individual frequencies constitute the optimal beamformer that maximizes the total output power ratio, provided the subband signals are independent.

For frequency, f, the quadratic ratio between the output signal power and the output noise-plus-interference power is

$$\mathbf{w}_{opt}^{(f)} = \arg\max_{\mathbf{w}^{(f)}} \left\{ \frac{\mathbf{w}^{(f)H} \mathbf{R}_{ss}^{(f)} \mathbf{w}^{(f)}}{\mathbf{w}^{(f)H} \mathbf{R}_{vv}^{(f)} \mathbf{w}^{(f)}} \right\} \tag{14.28}$$

where the matrices, $\mathbf{R}_{ss}^{(f)}$ and $\mathbf{R}_{vv}^{(f)}$, are defined as in (14.20). The weights, $\mathbf{w}^{(f)}$, are defined as the complex valued vectors of dimension N.

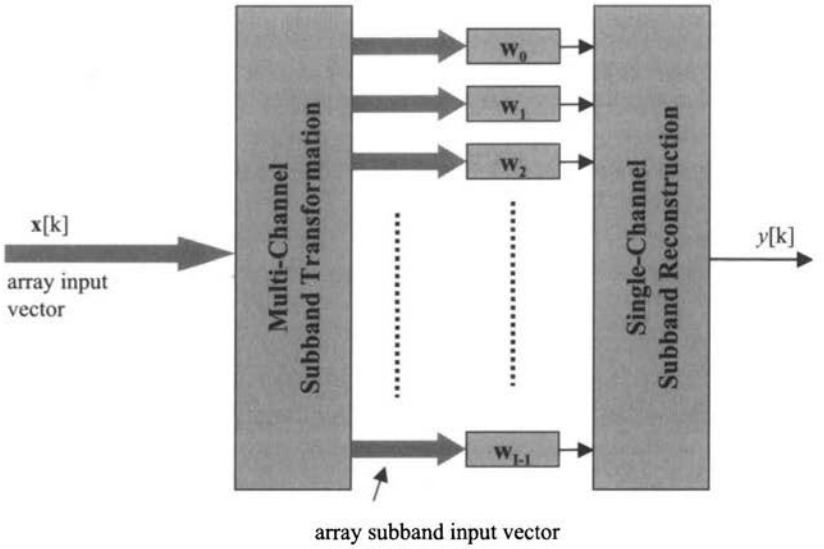

Fig. 14.2. Subband Beamforming Structure.

14.3 Subband Implementation of the Microphone Array

Noise and echo suppression exhibit significant gains when an LS solution is used in place of the NLMS algorithm [33]. However, computational considerations make use of the LS criterion impractical for the wide-band problem. Subband frequency transformations as shown in Figure 14.2 provide efficient means of allowing for the use of second order methods (such as RLS), while keeping computational complexity low. The frequency-domain algorithms have a least-squares objective function, as described in (14.29).

An uniform DFT analysis-synthesis filterbank [35] will be employed here. The filterbank is used to decompose the full-rate sampled signals, $x_n[k]$, into I subband signals [36]. The subband signals are essentially generated from a common bandpass filter with varying center frequency, $\frac{2\pi i}{I}$, and cover the entire frequency range. As a special case, when the number of subbands equals the length of the prototype filter, the subband decomposition will equal the overlapped Short-Time Fourier Transform (STFT) and the prototype filter is chosen as a simple, uniform moving average. The subband signals are decimated to a lower sampling rate allowing for a polyphase implementation. This provides an analysis-synthesis structure with approximately the same computational complexity as the STFT [35].

14.3.1 Description of LS-Subband Beamforming

The MMSE beamforming scheme formulated in (14.17) may be reexpressed in the time domain using an LS formulation as subband number

$$\mathbf{w}_{ls,opt}^{(i)}(N) = \arg\min_{\mathbf{w}^{(i)}} \left\{ \sum_{k=0}^{K-1} \left[|y^{(i)}[k] - s_r^{(i)}[k]|^2 \right] \right\} \tag{14.29}$$

where i indicates the subband index, K is the number of data points considered, and where $y^{(i)}[k]$ is given by (14.18) with $f = 2\pi i/I$. The reference source signal, $s_r^{(i)}[k]$, is not directly available, but a calibration sequence gathered in a quiet environment can be used in its place. This calibration signal contains the source's temporal and spatial information. Since $s_r^{(i)}[k]$ is independent of the actual data $x_n^{(i)}[k]$, at least for large K, the LS problem can be expressed as the sum of two components by

$$\mathbf{w}_{ls,opt}^{(i)}(K) = \arg\min_{\mathbf{w}^{(i)}} \left\{ \sum_{k=0}^{K-1} \left[|\mathbf{w}^{(i)H}\mathbf{s}^{(i)}[k] - s_r^{(i)}[k]|^2 + |\mathbf{w}^{(i)H}\mathbf{x}^{(i)}[k]|^2 \right] \right\}. \tag{14.30}$$

The equation may be rewritten as

$$\mathbf{w}_{ls,opt}^{(i)}(K) = \arg\min_{\mathbf{w}^{(i)}} \left\{ \mathbf{w}^{(i)H} \left[\hat{\mathbf{R}}_{ss}^{(i)}(K) + \hat{\mathbf{R}}_{xx}^{(i)}(K) \right] \mathbf{w}^{(i)} - \mathbf{w}^{(i)H}\hat{\mathbf{r}}_s^{(i)}(K) - \hat{\mathbf{r}}_s^{(i)H}(K)\mathbf{w}^{(i)} + \hat{r}_{s_r}^{(i)} \right\} \tag{14.31}$$

where the source correlation estimates can be precalculated in the calibration phase from

$$\begin{aligned} \hat{\mathbf{R}}_{ss}^{(i)}(K) &= \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{s}^{(i)}[k]\mathbf{s}^{(i)H}[k] \\ \hat{\mathbf{r}}_s^{(i)}(K) &= \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{s}^{(i)}[k]s_r^{(i)H}[k] \end{aligned} \tag{14.32}$$

where

$$\mathbf{s}^{(i)}(K) = [s_1^{(i)}[k], \quad s_2^{(i)}[k], \quad \cdots \quad s_N^{(i)}[k]]^T$$

is the microphone data when the source signal alone is active. The least-squares minimization of (14.31) is found by

$$\mathbf{w}_{ls,opt}^{(i)}(K) = \left[\hat{\mathbf{R}}_{ss}^{(i)}(K) + \hat{\mathbf{R}}_{xx}^{(i)}(K) \right]^{-1} \hat{\mathbf{r}}_s^{(i)}(K) \tag{14.33}$$

where

$$\hat{\mathbf{R}}_{xx}^{(i)}(K) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{x}^{(i)}[k]\mathbf{x}^{(i)H}[k] \tag{14.34}$$

is the observed data correlation matrix estimate. This implies that an estimate of the calibration data may be used as part of the solution.

14.4 Multi-Resolution Time-Frequency Adaptive Beamforming

The performance of the algorithm stated in previous section requires that the number of subbands is large enough for the frequency-domain representation to be accurate. The number of subbands is proportional to the length of the equivalent time-domain filters, and the more subbands chosen, the more degrees of freedom inherent in the beamformer. The number of subbands is also related to the delay caused by the frequency transformations since a large number of subbands necessitates a longer prototype filter, which in its turn will cause an increased delay.

The algorithm is easily extended to a combination of time- and frequency-domain representations. Each subband signal can be seen as a time-domain signal sampled at a reduced sampling rate and containing only frequencies in that particular subband. By applying a time-domain algorithm in each subband, the degrees of freedom for the filters are increased while the number of subbands can be held constant. The lengths of the corresponding filters may differ across the subbands to produce a multi-resolution framework.

14.4.1 Memory Saving and Improvements

The proposed beamformer consists of a source signal information gathering phase followed by the operation phase. Information about the source signal is represented through the frequency-dependent, source-only correlation matrix estimates, $\hat{\mathbf{R}}_{ss}^{(i)}$. These estimates are calculated and stored for each of the I subbands. When there are known unwanted sources, such as hands-free loudspeakers, which have a fixed location in relation to the microphones and the enclosure, correlation estimates from these signals are also estimated and saved. Estimates of the frequency-dependent cross-correlation vectors, $\hat{\mathbf{r}}_s^{(i)}$, are also maintained. The number of elements, P, required in memory to store the fullband time-domain solution is:

$$P^{time} = [NL(NL+2)]^2$$

where N is the number of microphone channels and L is the fullband FIR filter length. For the frequency-domain representation the number of storage elements needed is

$$P^{freq} = I[N\frac{L}{I}(N\frac{L}{I}+2)]^2$$

where index I is the number of subbands. As an example, Figure (14.3), shows the ratio of the number of storage elements required for the time- and the frequency-domain implementations as a function of the fullband time-domain filter length and subbands values of $I = 16, 32, 64, 128, 256, 512$. The number of channels is $N = 6$. Even for moderate filter lengths, the size of the mem-

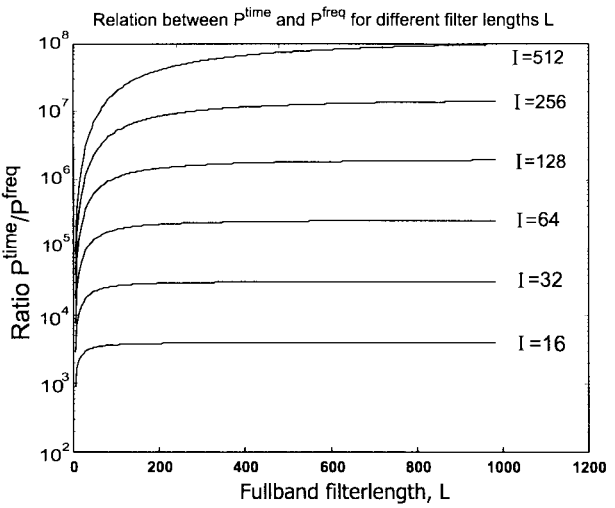

Fig. 14.3. The ratio P^{time}/P^{freq} for filter lengths L varying from 4 to 1024, and the number of subbands, I, is varying from 16 to 512. The number of channels is $N = 6$.

ory is reduced substantially with the frequency-domain implementation. The number of multiplications is proportional to the number of stored elements, and the relationship between computational costs for the time-domain and frequency-domain implementations is the same as in Figure 14.3.

14.5 Evaluation and Examples

14.5.1 Car Environment

A performance evaluation of the beamformer was made in a hands-free situation with a six-element microphone array mounted on the passenger-side visor of a Volvo station wagon. Data was gathered on a multi-channel DAT-recorder with a 12 kHz sampling rate and a 300-3400 Hz bandwidth. In order to facilitate simultaneous driving and recording, an artificial talker was mounted in the passenger seat to simulate a real person engaging in a conversation. Initially, a white noise sequence within the bandwidth was emitted from the artificial talker, in a non-moving car with the engine turned off. This sequence served as the desired sound source calibration signal in all of the following simulations. Interference signals were recorded by emitting an independent sequence of bandlimited, white noise from the hands-free loudspeaker. This recording functioned as the point-source interference calibration signal and was referred to as the echo signal. In order to gather background noise estimates, the car was driven at a speed of 110 km/h over a paved road. The car cabin noise environment consisted of a number of unwanted sound sources,

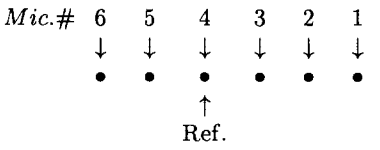

Fig. 14.4. Geometry of the six-element, linear array with an adjacent microphone spacing of 5 cm.

mostly with broad spectral content, e.g. wind and tire noise. Recordings with real speech signals, from both the artificial talker and the hands-free loudspeaker, were recorded both individually and while driving. These recordings served as the beamformer evaluation signals. All of the performance measures presented in Section 14.5.5 were based on these real speech recordings.

14.5.2 Microphone Configurations

The sensors used in this evaluation were high-quality Sennheiser microphones mounted flat on the visor. The speaker was positioned 35 cm from the center of the array and oriented perpendicular to the its axis. The mounting of the six-element linear array is given in Figure 14.4. The spacing between adjacent elements in the array was 5 cm.

14.5.3 Performance Measures

There are two objectives for the beamformer: minimize the distortion caused by the beamforming filters (measured by the deviation between the beamformer output and the source signal) and maximize noise and interference suppression. In order to measure the performance the normalized distortion quantity, D, is introduced as

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} |C_d \hat{P}_{y_s}(w) - \hat{P}_{x_s}(w)| dw \tag{14.35}$$

where $w = 2\pi f$ and f is the normalized frequency. The constant, C_d, is defined as

$$C_d = \frac{\int_{-\pi}^{\pi} \hat{P}_{x_s}(w) dw}{\int_{-\pi}^{\pi} \hat{P}_{y_s}(w) dw} \tag{14.36}$$

where $\hat{P}_{x_s}(w)$ is a spectral power estimate of a single sensor observation and $\hat{P}_{y_s}(w)$ is the spectral power estimate of the beamformer output when the source signal alone is active. The constant C_d normalizes the mean output spectral power to that of the single sensor spectral power. The single sensor

observation is chosen as the reference microphone observation, i.e. microphone 4 in the array. This distortion measure is essentially an estimate of the mean output spectral power deviation from the observed single sensor spectral power, and under ideal circumstances will be zero.

To measure normalized noise suppression, the quantity, S_N, is computed from

$$S_N = C_s \frac{\int_{-\pi}^{\pi} \hat{P}_{y_N}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{x_N}(w)dw} \tag{14.37}$$

where

$$C_s = \frac{1}{C_d} \tag{14.38}$$

and $\hat{P}_{y_N}(w)$ and $\hat{P}_{x_N}(w)$ are the spectral power estimates of the beamformer output and the reference sensor observation, respectively, when the surrounding noise alone is active.

Similarly, the normalized interference suppression quantity, S_I, is given by

$$S_I = C_s \frac{\int_{-\pi}^{\pi} \hat{P}_{y_I}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{x_I}(w)dw} \tag{14.39}$$

with $\hat{P}_{y_I}(w)$ and $\hat{P}_{x_I}(w)$ being the spectral power estimates when the interference and desired signals, respectively, are active alone. Both of these suppression measures are normalized to the amplification (or attenuation) caused by the beamformer relative to the reference sensor observation when the source signal is active alone. Accordingly, when the beamformer scales the source signal by a specific amount, the noise and interference suppression quantities are adjusted appropriately.

14.5.4 Spectral Performance Measures

In order to evaluate the performance within individual subbands, the above definitions may be made frequency-dependent by omitting the integration operations, i.e.

$$S_N(w) = \frac{C_s \hat{P}_{y_N}(w)}{\hat{P}_{x_N}(w)} \tag{14.40}$$

and

$$S_I(w) = \frac{C_s \hat{P}_{y_I}(w)}{\hat{P}_{x_I}(w)} \tag{14.41}$$

where the definition of C_s and the power spectral estimates are the same as above.

In practice, the above measures were calculated using Welch's averaged periodogram spectral estimation method with non-overlapping Hanning windows of length 256. The integrals were approximated by discrete summation over the periodograms. All measures were calculated from the time-domain signals, which implies that any distortions created by the frequency transformations were also taken into account.

14.5.5 Evaluation on car data

In this evaluation, 8 s white noise calibration signals were used. These were emitted individually from the artificial talker and the hands-free loudspeaker, as the source and interference calibration signals, respectively. The calibration input sequence used to generate all the optimal beamformer weights consisted of these signals along with the car cabin noise signals, gathered at a specific time instant, t.

In order to evaluate the optimal beamformers, input signals were created by emitting independent speech signals from the artificial talker and the hands-free loudspeaker and recording the microphone observations with car cabin noise taken at time instant $t + 8$ s. The beamformer output was generated by filtering the inputs with the fixed filter weights found from the calibration sequences.

In the time-domain implementations, the FIR filter length was chosen as $L = 256$. For the frequency-domain implementations, the total number of subbands was set to $I = 64$. By setting the prototype filter length in the filterbank to 256, the same filter order as for the corresponding time-domain filters was obtained. This comes from the fact that the number of time-domain lags used in the frequency transformation equals the prototype filter length.

14.5.6 Evaluation Results

Performance measures in dB of the distortion, noise, and interference quantities, as described in section 14.5.3, are presented in Table 14.1. In general, the optimal SNIB beamformers have better suppression levels for both noise and interference when compared to the LS beamformers. However, the LS beamformers have much lower distortions values. Additionally, the subband-LS beamformer has performance comparable to the fullband-LS solution as the number of subband weights is increased.

Evaluation plots are now presented for the least-squares beamformer. Figure 14.5 illustrates the short-time (20 ms) power estimates in dB derived from an 8 s sequence of the single-reference microphone observation without any processing, followed by 8 s of the beamformer output signal acquired using the time-domain least-squares beamformer. Source speech, hands-free interference and car cabin noise are all active simultaneously. The near-end

Table 14.1. Distortion, noise, and interference performance measures of the beamformer output.

Performance [dB]	D	S_N	S_I
Time domain			
SNIB	-19.4	18.1	30.7
NLMS	-24.9	4.04	3.78
LS	-30.6	15.2	17.2
Frequency domain			
SNIB	-19.8	18.0	23.7
NLMS 1-tap	-21.1	8.68	5.00
NLMS 2-tap	-20.9	7.95	5.55
NLMS 3-tap	-20.8	7.45	4.96
NLMS 4-tap	-20.7	7.19	4.68
NLMS 5-tap	-20.7	7.11	4.54
NLMS 6-tap	-24.8	7.05	4.45
LS 1-tap	-28.6	12.9	13.6
LS 2-tap	-28.8	13.4	14.4
LS 3-tap	-30.0	13.8	15.2
LS 4-tap	-30.4	14.2	15.4
LS 5-tap	-30.5	14.3	15.7
LS 6-tap	-30.7	14.3	15.8

speech, coming from the location of interest is denoted in the plot as "Speech Male/Female" while the far-end speech echo, i.e. the interfering hands-free loudspeaker, is denoted by "Echo Male/Female".

Figures 14.6 and 14.7 show the spectral power estimates in dB of the reference microphone observation and the normalized spectral estimate of the least-squares beamformer outputs when the noise and the interference signals are active individually. These plots correspond to the numerator and the denominator of (14.40) and (14.41), respectively.

14.6 Summary and Conclusions

A number of optimal, time- and frequency-domain beamformers based on different error criteria were presented. The beamformers were evaluated in a real environment, a car hands-free telephony situation. Simulations with real speech signals acquired by a linear microphone array show that noise reduction of 18 dB and echo suppression of 30 dB can be achieved, simultaneously. This was accomplished by the time-domain version of the optimal signal-to-noise plus interference beamformer. With the time-domain least-squares implementation, noise suppression of 15 dB and hands-free suppression of 17 dB were found. The least-squares implementation yields ten times less distortion, as compared to the optimal signal-to-noise plus interference beamformer.

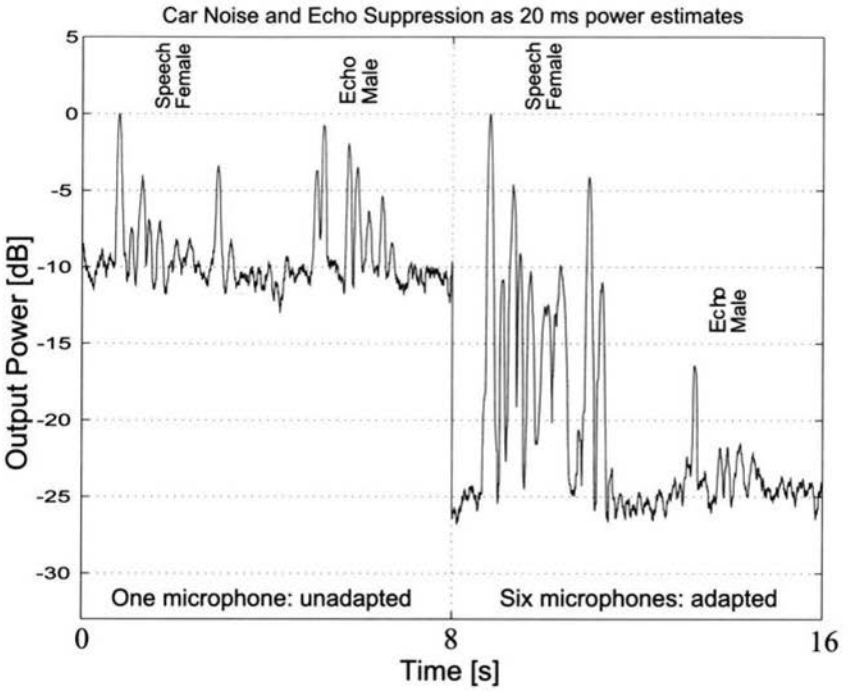

Fig. 14.5. Short-time (20 ms) power estimates of an unprocessed single microphone observation followed by the time-domain least-squares beamformer output signal.

The frequency-domain implementations show a similar relation between the optimal beamfomers. Better suppression is achieved with the optimal signal-to-noise plus interference beamformer, but the distortion is much higher than that for the least-squares implementations.

The subband least-squares beamformer evaluation showed that the performance on the real speech recordings is very close to that of the optimal time-domain least-squares beamformer. The noise and echo suppression were 14 dB and 16 dB, respectively, while the computational complexity was substantially reduced, thereby making it amenable to real-time processors. The distortion caused by the proposed method is the same as with the optimal time-domain least-squares beamformer.

Further research includes blind speech source extraction where the desired cross-correlation vector may be interchanged with a nonlinear function of the averaged beamformer output, for each frequency. The performance relies on the difference between the probability density functions of the source speech and the background noise. Implementations at an early stage show encouraging results. Source tracking is implicitly possible since a calibration sequence is unnecessary, and the objective function is made invariant to source movements.

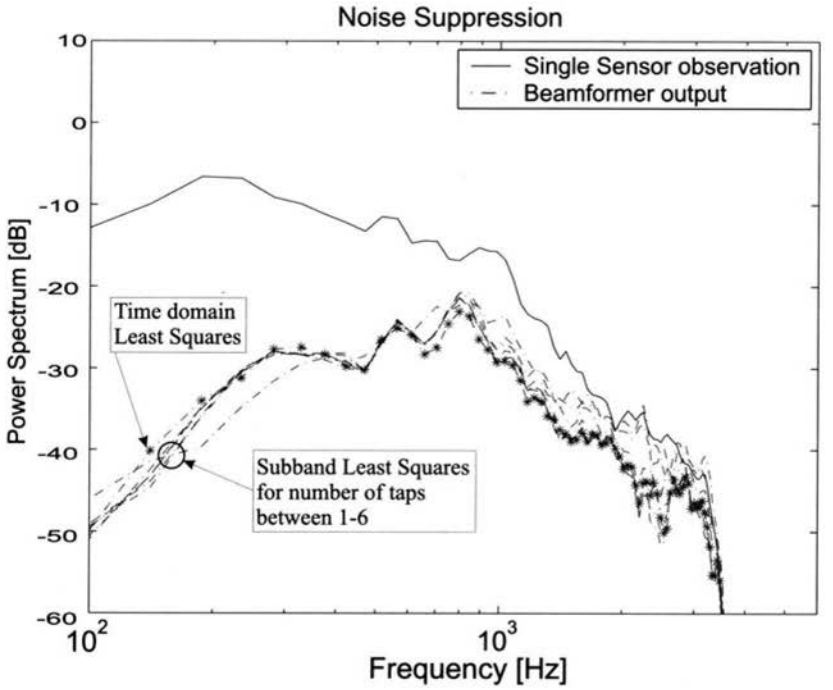

Fig. 14.6. Power spectrum of unprocessed single microphone observation (solid line) and power spectrum of the least-squares beamformer output signals (dashed-dotted lines) when only car cabin noise is present. The time-domain least-squares beamformer is marked by dashed-dots with stars.

References

1. S. Nordebo, B. Bengtsson, I. Claesson, S. Nordholm, A Roxström, M. Blomberg, and K. Elenius, "Noise reduction using an adaptive microphone array for speech recognition in a car," in *Proc. RVK93, Radio Vetenskaplig Konferens*, Lund, Sweden, Apr. 1993.
2. S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson, "Noise reduction using an adaptive microphone array in a car– a speech recognition evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz NY, USA Oct. 1993.
3. S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 425–437, Sept. 1997.
4. J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, Oct. 1993.
5. J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Macmillan, 1993.

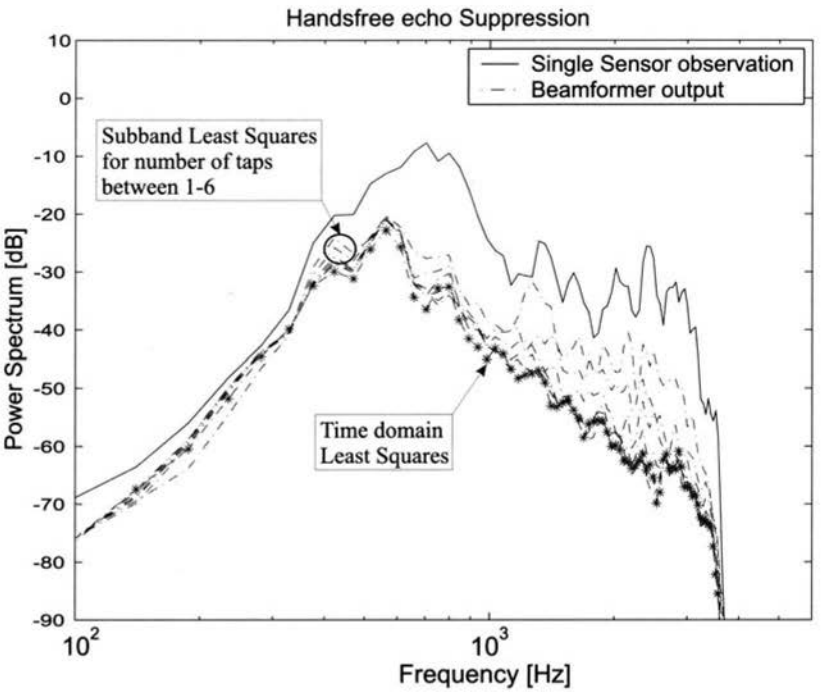

Fig. 14.7. Power spectrum of unprocessed single microphone observation (solid line) and power spectrum of the least-squares beamformer output signals (dashed-dotted lines) when only hands-free speech interference is present. The time-domain least-squares beamformer is marked by dashed-dots with stars.

6. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP–27, no. 2, pp. 113–120, April 1979.
7. J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, Minneapolis MN, USA, vol. II, pp. 363–366, April 1993.
8. H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction, truly linear convolution and a spectrum dependant adaptive averaging method," *Submitted for publication in IEEE Trans. Speech Audio Processing*, June 1999.
9. Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-34, no. 6, pp. 1391–1400, Dec. 1986.
10. S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Vehicular Tech.*, vol. 42, no. 4, pp. 514–518, Nov. 1993.
11. Y. Grenier and M. Xu, "An adaptive array for speech input in cars," in *Proc. Int. Symp. Automotive Technology and Automation (ISATA)*, 1990.
12. M. Sondhi and D.A. Berkley, "Silencing echoes in the telephone network," *Proc. IEEE*, vol. 68, pp. 948–963.

13. D.G. Messerschmidt, "Echo cancellation in speech and data Transmission," *IEEE J. Sel. Areas Commun.*, vol. SAC-2, pp. 283–297, Mar. 1982.
14. M. Sondhi and W. Kellermann, "Adaptive echo cancellation for speech signals," in *Advances in speech signal processing*, (S. Furui and M. Sondhi, eds.), ch. 11, Dekker, 1992.
15. C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt and J. Tilp, "Acoustic echo control, an application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, pp. 42–69, July 1999.
16. S. L. Gay and J. Benesty, eds., *Acoustic signal processing for telecommunication*, Kluwer, 2000.
17. S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals. an analytical evaluation," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 241–252, May 1999.
18. B. Widrow and S. D. Stearns, *Adaptive signal processing*, Prentice Hall, 1985.
19. I. E. Claesson, S. E. Nordholm, B. A. Bengtsson, and P. F. Eriksson, "A multi-DSP implementation of a broad-band adaptive beamformer for use in a hands-free mobile radio telephone," *IEEE Trans. Vehicular Tech.*, vol. 40, no. 1, pt. 2, pp. 194–202, Feb. 1991.
20. S. Nordebo, I. Claesson, and S. Nordholm, "An adaptive microphone array Employing calibration signals recorded on-site" in *Proc. ICSPAT94*, Dallas TX, USA, Oct. 1994.
21. M. H. Er, "A robust formulation for an optimum beamformer subject to amplitude and phase perturbations," *Signal Processing*, vol. 19, no. 1, pp. 17–26, Jan. 1990.
22. H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-35, pp. 1365–1376, Oct. 1987.
23. E. Walach, "On superresolution effects in maximum likelihood adaptive antenna arrays," *IEEE Trans. Antennas Propagat.*, vol. 32, pp. 259–263, March 1984.
24. I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. 40, no. 9, pp. 1093–1096, Sept. 1992.
25. S. Nordholm, I. Claesson, and S. Nordebo, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE J. Oceanic Eng.*, vol. 19, no. 4, pp. 583–590, Oct. 1994.
26. L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. 30, pp. 27–34, Jan. 1982.
27. M. H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 31, no. 6, pp. 1378–1393, Dec. 1983.
28. M. H. Er and A. Cantoni, "A new set of linear constraints for broad-band time domain element space Processors," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 2, pp. 320–329, Mar. 1986.
29. O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

30. M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with a microphone array," *IEEE Trans. Vehicular Tech.*, vol. 48, no. 5, pp. 1518–1526, Sep. 1999.
31. M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, vol. 1, pp. 239–242, 1997.
32. N. Grbić, M. Dahl, and I. Claesson, "Neural network based adaptive microphone array system for speech enhancement," *IEEE World Congress on Computational Intelligence*, Anchorage AK, USA, vol. 3, pp. 2180–2183, May 1998.
33. N. Grbić, *Speech signal extraction - a multichannel approach*, University of Karlskrona/Ronneby, Sweden, Nov. 1999.
34. S. Haykin, *Adaptive filter theory*, Prentice Hall, 3rd edition, 1996.
35. P.P. Vaidyanathan, *Multirate systems and filter-banks*, Prentice Hall, 1993.
36. S.K. Mitra, *Digital signal processing*, McGraw-Hill, 1998.

15 Speech Recognition with Microphone Arrays

Maurizio Omologo, Marco Matassoni, and Piergiorgio Svaizer

ITC-IRST (Istituto per la Ricerca Scientifica e Tecnologica), Povo (Trento), Italy

Abstract. Microphone arrays can be advantageously employed in Automatic Speech Recognition (ASR) systems to allow distant-talking interaction. Their beamforming capabilities are used to enhance the speech message, while attenuating the undesired contribution of environmental noise and reverberation. In the first part of this chapter the state of the art of ASR systems is briefly reviewed, with a particular concern about robustness in distant-talker applications. The objective is the reduction of the mismatch between the real noisy data and the acoustic models used by the recognizer. Beamforming, speech enhancement, feature compensation, and model adaptation are the techniques adopted to this end. The second part of the chapter is dedicated to the description of a microphone-array based speech recognition system developed at ITC-IRST. It includes a linear array beamformer, an acoustic front-end for speech activity detection and feature extraction, a recognition engine based on Hidden Markov Models and the modules for training and adaptation of the acoustic models. Finally the performance of this system on a typical recognition task is reported.

15.1 Introduction

During the last decade research on ASR technology has made significant advances. As a result, high performance systems are now available for situations where there is a good match between testing and training conditions [1,2]. However, these same systems tend to suffer from a limited robustness to variability in their operating environment [3–8].

One of the most attractive potential features of ASR technology is the flexibility afforded through hands-free interaction. Not being encumbered by a hand-held or head-mounted microphone may be of considerable utility to the user. Of particular concern is the distant-talker case where the user is beyond the normal acquisition range of the system microphone[1] (e.g. at more than one meter in the case of an omnidirectional microphone). For ASR applications of moderate/high language complexity this represents a very ambitious task.

The development of distant-talker ASR will allow for the expansion of voice activated technology into a number of areas where it has until now

[1] The distinctions among different types of microphones will not be treated here. However, the choices available and their relative characteristics should be specifically addressed in the design of the application.

been ineffective (e.g. noisy offices or factory floors) and will improve its functionality for applications where it has already seen some use (e.g. computer dictation or the home). In the first example, voice messages of a varying nature would have to be recognized as isolated word commands immersed in a background of multi-talker speech and noise. The second example involves a rather clean environment, but a large vocabulary of words to identify.

In the real-world applications it would also be necessary to account for various factors related to the means of interaction. The talker's position may be unknown and time-varying in an unpredictable fashion. Because of sound attenuation and talker radiation effects, the quality of the input signal may be influenced by even subtle head movements. Moreover, environmental noise and room acoustics play an important role, especially in the case of highly reverberant conditions and unstationary noise sources. In the most adverse noisy conditions, a talker will tend to speak more loudly than usual and thereby modify the underlying characteristics of the speech signal produced relative to normal speaking conditions. This is known as the Lombard effect [9]. Additionally, when the language/dialogue model becomes more complex, the variability in talking style may increase and one can expect that the talker will often speak in spontaneous mode.

For these reasons and others, there are many challenging and as yet unsolved problems in this field. In the last few years, some work has been devoted to the application of multi-microphone based processing for distant-talker speech recognition. Compared to the number of labs working on improving the robustness of single-channel ASR systems, this effort is relatively small. This fact may be due to the incipient nature of microphone array technology and the increase in hardware complexity that is required for a multi-channel front-end. However, judging by the advancements in ASR performance that may be attributed to improvements in input signal quality brought about by microphone array processing, this work is well justified.

The remainder of this chapter is organized into two sections. The following section summarizes the current state of research activity in the field of ASR, particularly with regard to the distant-talker situation. The final section details a specific microphone-array based recognition system, namely an Italian language recognizer developed at ITC-IRST.

15.2 State of the Art

15.2.1 Automatic Speech Recognition

Automatic speech recognition can be viewed as a problem of conversion from speech into text by a decoding process that involves several processing stages. The characteristics and the difficulty of an ASR application differ substantially based upon various features. These include vocabulary size and confusability, speaker independence, language complexity, and input speech quality.

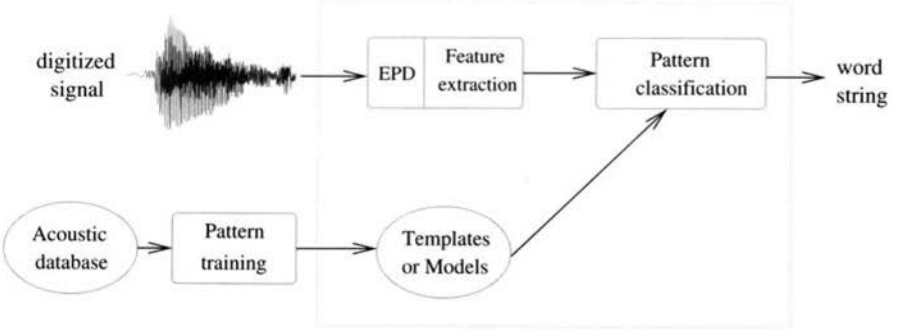

Fig. 15.1. General block diagram of a pattern recognition based ASR system.

Research on ASR has been conducted for more than four decades. The related literature is very large; good overviews of the most significant achievements can be found in [1,2]. Many different techniques for ASR have been investigated. Currently, the most widely used approaches are based on some form of statistical pattern recognition. Thanks to these modern methods, the growth in hardware capability, and the availability of very large speech corpora for training, the last decade has witnessed high level performance achieved on recognition tasks of progressively increasing complexity.

The purpose of this section is to give a very brief introduction to the basic problem and to the most common solutions which have been adopted, with specific reference to the distant talker problem.

Pattern recognition based ASR A general block diagram of an ASR system based on the pattern recognition approach is shown in Figure 15.1. It is assumed that the speech message has been transduced by a microphone into an electrical signal and then converted into an equivalent digital representation with an adequate sampling rate and quantization level (e.g. 16 kHz and 16 bit, respectively). In general, at a preprocessing level ASR systems include a pre-emphasis step in the form of a single tap high-pass filter. The goal of which is to emphasize high-frequency formants which typically have a reduced magnitude due to a negative spectral tilt in the speech signal, particularly voiced sounds.

A speech activity detection process, also called End-Point Detection (EPD), is employed to isolate speech events from other segments and background noise. Several techniques are available for EPD, e.g. [10–14]. These are usually based on criteria such as short-term energy and zero-crossing rate. However, they may also rely on the same acoustic features used during the recognition process.

The objective of Feature Extraction (FE) is to convert the input signal into some form of compressed parametric representation. The most common examples of FE are based on short-time spectral analysis. Speech can be

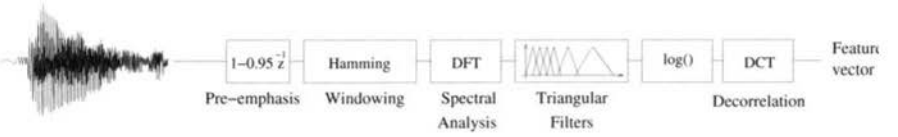

Fig. 15.2. Block diagram of the computation of Mel-based cepstral coefficients.

considered statistically stationary over short periods of time (a few tens of milliseconds). As a consequence, the analysis frame size employed for FE is generally over 15 ms with a step size in the range of 5-30 ms. The FE process produces a sequence of vectors of dimensionality normally ranging between 8 and 12. These vectors are frequently augmented by including data to characterize the first and second order time derivatives of the given features.

A number feature sets have been investigated for ASR. Popular ones include Mel-scaled Cepstral Coefficients (MCC's) (see Figure 15.2), LPC coefficients, PLP coefficients [5,15]. Currently, MCC's are the most widely used acoustic features for ASR. Figure 15.2 outlines their method of production. Basically, a triangular filter-bank is applied to the output of a short-term spectral analysis. The logarithmic-like Mel scale models the frequency resolution of the human ear and for this reason is preferred to a linear scale in the filter-bank. A Discrete Cosine Transform (DCT) is then applied to decorrelate the log-filter-bank output. The resulting MCC's may then be statistically modeled through Gaussians with diagonal covariance matrices. This property is useful in the case of HMM-based recognition discussed below.

Note that non-linearities and approximations are included in the processing that derives the acoustic features from the signal or its power spectrum representation. For instance, the output of the filter-bank used for MCC computation provides a rough approximation to the FFT-based spectral analysis method. As a consequence, when addressing the impact of microphone arrays to distant-talker ASR, it is possible that improvements in signal quality produced by the array processing may be rendered ineffective during successive stages of the recognition chain because of these approximations.

In the pattern recognition approach to ASR, the acoustic feature vector sequence derived from the unknown speech is compared to the feature sequence of reference speech. Among the various ways to perform this comparison, three methods have been primarily utilized: Hidden Markov Model (HMM), Dynamic Time Warping (DTW), and Artificial Neural Networks (ANN). A detailed discussion of these techniques, the functional relationships between them, and the hybrid solutions which have been studied (e.g. ANN/HMM), goes beyond the scope of this chapter and can be found in [1,2].

The literature reports that for simple tasks (e.g. connected digit recognition) in controlled and matched environments (i.e. the user interacts with a close-talking microphone and the system has been trained on clean speech), satisfactory recognition performance can be obtained using any of the above

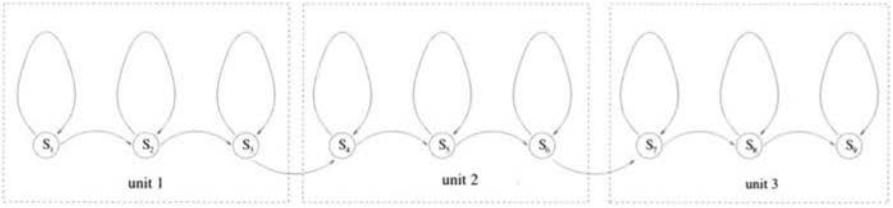

Fig. 15.3. Concatenation of three HMM-based phone models, characterized by a left-to-right three-state topology.

methods. However, HMM's are currently considered to be the most effective and stable framework for speech modeling in a general context. This is particularly the case for large vocabulary tasks and when statistical language modeling as well as integration of the recognizer with a dialogue manager are required by the application. As a consequence, the remainder of this chapter will assume the use of HMM's as the means for pattern classification. Accordingly, it will be necessary to present a modicum of detail regarding the procedure.

HMM framework In the statistical ASR paradigm, a generic utterance consists of a sequence of unknown words and the recognizer finds the most probable word string, given the observed feature vector sequence provided by the front-end processing. HMM's serve as the statistical model used to classify the utterances and quantify their observation probabilities.

Bayes' rule is used to decompose the required probability into two components: the *a priori* probability of observing the sequence of words (the "language model") and the probability of observing the feature vector sequence given that word string (the "acoustic model").

Each word is represented by a chain of basic sounds called phones. An HMM is adopted for each phone. In practice, the model consists of a number of states with the sequencing through them determined by a set of transition probabilities. Each state produces observations which are characterized by a set of state-dependent observation statistics. These are frequently modeled as mixtures of Gaussian densities. Figure 15.3 shows an example of a three-state, Markov model. In this case the non-zero transition probabilities are constrained to produce a left-to-right topology which is very common in ASR applications.

The *Training Problem* involves learning the appropriate HMM parameters given a reference ensemble of feature sequences associated with the desired word. An efficient procedure known as the Baum-Welch algorithm is available for this purpose. For what concerns the *Scoring Problem* during recognition, once the most likely state sequence is selected, the related recognized text is provided. The Viterbi algorithm is commonly used to efficiently evaluate word string likelihoods.

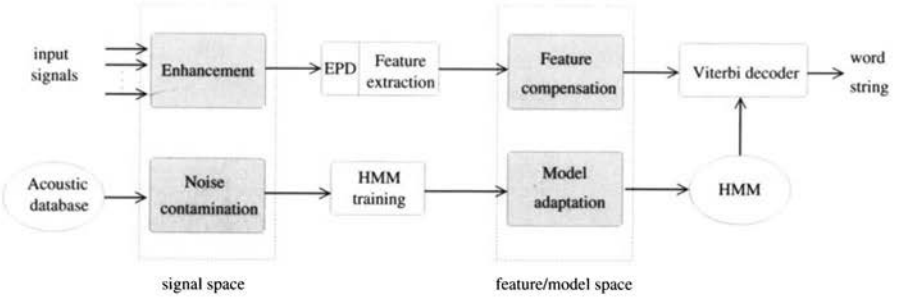

Fig. 15.4. Block diagram of a robust ASR system operating in an adverse environment. Highlighted blocks, operating either in the signal space or in the feature/model space, contribute to reduce the mismatch between noisy data and the acoustic models used by the recognizer.

For a more thorough development of HMM theory and practice, see [16]. As to the main concerns of this chapter, note that the theory of HMM and its application to ASR are based on a number of signal assumptions, among which is that of statistical independence of the observations over time. This is not satisfied in many cases, particularly for speech acquired in the presence of noise and reverberations.

15.2.2 Robustness in ASR

The crucial aspect for most ASR techniques is robustness. In practice, performance very often degrades rapidly when these systems are used with speech input taken from a noisy environment or, in general, with speech input having characteristics which differ from those observed during the training phase. Mismatch between training and test conditions can be caused by many factors. Some examples are background noise, transducers, channel noise, inter-speaker variability, and spontaneous-speech phenomena. Hence, flexibility and robustness with respect to these sources of variability is one of the main objectives of current ASR research [3–8]. Generally, the training of speech recognizers is accomplished by using large speech corpora. In principle, for each noisy environmental condition and, in this case, for each microphone and talker position, a specific corpus should be used. This solution being impractical, a fundamental task for researchers and technology developers becomes that of exploiting as much as possible from existing corpora, tools, techniques, and *a priori* knowledge, in order to build robust recognizers.

Current approaches to improving robustness of noisy speech recognizers can be classified into a number of categories as reported in [4,7,17]. Four of these general approaches (signal enhancement, feature compensation, model adaptation, and noise contamination) are summarized below. Figure 15.4 shows how these methods may be introduced as modules into a generic HMM-based architecture to improve system robustness.

Enhancement techniques are used to increase the quality of the signal provided to the recognizer [18–21]. Their impact on ASR is not obvious. There is no direct relationship between the SNR (or perceptual quality) of the resulting signal and speech recognition performance, even when the recognizer is trained using the same preprocessing. In this regard, a further critical issue is that of end-point detection. Several algorithms (generally based on energy-thresholding techniques) have been proposed, which can be applied successfully with a SNR as low as 10 dB. However, most of these algorithms generally become unusable with lower SNR conditions [5].

Many techniques have been proposed which address the parametric representation of the signal. These aim at constructing a compact and robust feature set and processing it to compensate for the mismatch between acoustic spaces of the clean and noisy speech. As an example, a very simple feature normalization technique, often combined with MCC's, is Cepstral Mean Subtraction (CMS). It consists of removing from each cepstral coefficient sequence the mean evaluated across the whole utterance (or over an extended interval). CMS aids in reducing the influence of slow variations in the acoustic feature vectors, like those related to convolutional channel effects (e.g. change of microphone) and to speaker-dependent biases. Other relevant compensation techniques, operating either in the feature space or in the model space, are signal bias removal [22,23], stochastic matching [24,17], noise modeling and masking [25–27], Parallel Model Combination [28,29].

In the recent years increasing attention has been devoted to acoustic model adaptation. When the application requirements allow, these techniques attempt to adapt system parameters to the speaker and environment by exploiting data samples representative of the actual acoustic conditions. For systems based on continuous-density HMM's, most popular adaptation schemes rely on maximum *a posteriori* (MAP) estimation [30] or maximum likelihood linear regression (MLLR) [31–33] of the model parameters.

Finally, it is possible to adopt an approach known in the literature as *training data contamination* [4,34,35], which provides a way of training acoustic models which are more robust and representative of the given real noisy environment than those derived through training on the corresponding clean speech. In practice, training data are produced by injecting real or artificial noise into the clean speech material. Clearly, this approach is time-consuming and may become impractical when the size of training set grows large. However, it does offer some advantage. For instance, it is free from the negative spectrum problem typical of noise subtraction schemes [36].

15.2.3 Microphone Arrays and Related Processing for ASR

The utility of a microphone array as input to a speech recognition system lies in its ability to acquire a higher quality signal than that provided by a single far-field microphone. The signal enhancement is obtained by emphasizing the talker's speech as well as by reducing noise and reverberation components.

These methods are the specific topics of other chapters in this work. The relevant approaches will only be summarized very briefly here.

In order to reduce mismatch effects in the recognizer, a first requirement is that of having uniform improvement levels across the complete speech spectrum. Additionally, it would be desirable to have these spectral enhancements be independent of the talker's position. Unfortunately, microphone array frequency responses are characterized by significant variations across source angles and distances. Consequently, it is necessary to reduce as much as possible this variability, which may introduce significant discrepancies with respect to the training conditions.

In practice, an enhanced output can be derived from a microphone array by the application of beamforming techniques. The simplest and most commonly used approach is the delay-and-sum beamformer, which reduces the output power for directions other than that of the steering location by means of destructive interference. Figure 15.5 illustrates a typical result from this procedure. The delay-and-sum beamformer may be used passively to realign the signals given a set of delay estimates, or actively by aiming the array towards a specific direction. In the former case, delay estimates are derived from Time Delay Estimation (TDE) techniques [37], as shown in the case of talker location in [38], where a Cross-Power Spectrum Phase analysis was adopted.

Applying Time Delay Compensation (TDC) processing represents a good solution for the case of an isotropic (diffuse) noise field, as no spatial coherence is exploitable to suppress undesired components. Very much dependent on the number of array elements and their geometry relative to the source position, delay-and-sum beamforming provides only moderate directivity gains. Additional drawbacks [39–43,8], are grating lobes in the directivity pattern and a low-pass effect due to both the beam narrowness at high frequencies and to imperfect steering caused by imprecise inter-channel delay estimates.

In the case of coherent noise sources linearly constrained adaptive beamformers, such as those proposed by Frost [44] or the Generalized Sidelobe Canceler [45], have the specific objective of eliminating noise contributions in directions outside the directivity lobe. The main limitation of these schemes in a reverberant environment is the issue of signal cancellation. Since the degradations are correlated with the desired signal, the suppression process introduces distortions to the desired signal. Superdirective beamformers have also been proposed [46] to suppress interfering signals effectively.

A technique specifically developed to address the reverberation phenomenon of enclosures is the Matched Filter Array . This technique utilizes the acoustic impulse responses of the environment to create constructive interference between direct and reflected components of the speech signal [47,48].

Beamforming techniques may also be combined with adaptive postfilters [46] (e.g. based on Wiener theory) for further noise reduction. However,

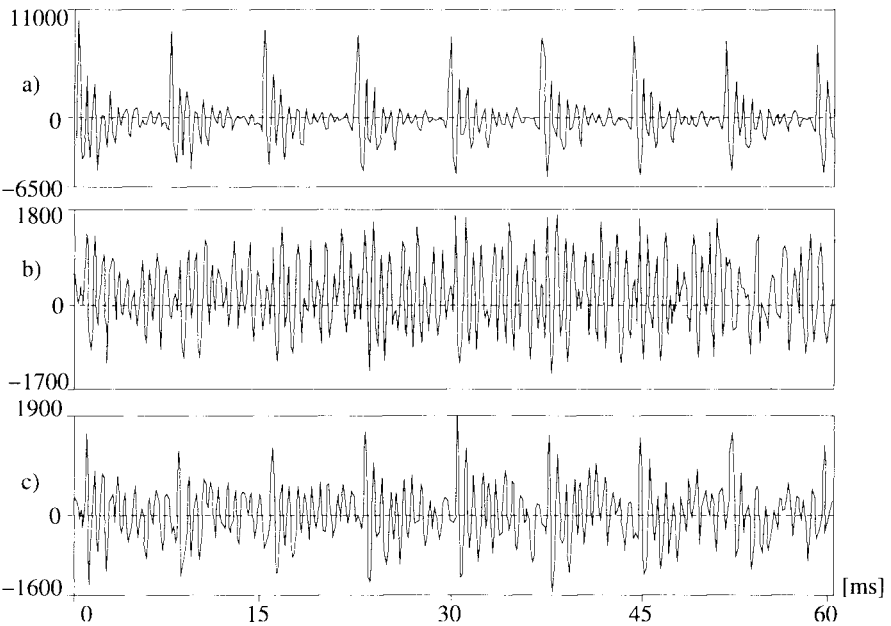

Fig. 15.5. Portion of a vowel /a/ uttered at 4 m distance from a linear array of six microphones. The upper waveform is the close-talker signal, the middle plot represents the reverberated signal acquired by one of the microphones, and the lower plot shows the corresponding delay-and-sum beamformed signal.

the use of post-filtering, like its above counterparts, can introduce artifacts into the reconstructed signal, particularly for the case of reverberant environments [49], and may consequently limit recognition improvements.

15.2.4 Distant-Talker Speech Recognition

A sizeable body of work on distant-talker ASR have been produced in recent years [50–56]. The use of either single microphones or multi-microphone systems has focused primarily on experimental contexts (e.g. car environment) for tasks generally characterized by a small-size vocabulary and by a low complexity language. Simple multi-microphone products are available commercially and have replaced traditional input devices in some ASR applications. However, these devices are of limited practical utility and are typically only effective for talkers no more than one meter away from the array, at a fixed and a quite narrow range of angles, and in a rather quiet environment.

This section provides a brief overview of the literature relating to various research topics related to distant-talker ASR. The intention is not to give a thorough description of all the techniques which have been investigated, but rather to indicate the general issues and approaches. Section 15.3 will explore in detail the system developed at IRST labs.

Array Geometry This first topic concerns the array characteristics as well as their influence on recognition performance. Clearly, the number of microphones represents an important factor. In an effort to limit hardware complexity, most investigators utilize 16 or fewer microphones. Optimal array design techniques have been addressed by other chapters of this book and, with specific reference to ASR, in [57]. For practical ASR systems, the bulk of array geometries investigated have either been of the linear equi-spaced or harmonic nesting varieties. The latter is the most common in practice, despite the fact that a real advantage in application to hands-free ASR is not evident [58,59]. In general, demonstrating a potential advantage inherent in a specific geometry is difficult. The speech quality improvement due to the array configuration is counterbalanced by several approximations (e.g. in the generation of artificial signals, when simulation is used, and in the acoustic feature extraction), and by effects related to the application of compensation/adaptation techniques applied to these features or to the acoustic modeling.

Beamforming The literature reports on the use of various beamforming techniques for distant-talker ASR. The delay-and-sum beamformer is the most commonly used, despite its limited speech enhancement capabilities. The joint use of delay-and-sum beamforming and a talker localization module is investigated in [60]. The use of adaptive beamforming techniques (e.g. GSC), generally under the assumptions of fixed talker and noise source positions, is also common. Some examples are shown in [61,51,50,62]. As expected, in the presence of coherent noise sources, adaptive beamformers yield more robust recognition performance than delay-and-sum beamformers. However, many authors report that the improvements are lower than what would be expected on the basis of the SNR or of the reconstructed signal's perceived quality. This observation seems to be more common for data acquired in real environments.

End-Point Detection This topic is very critical, even in moderately noisy and reverberant environments. Most of the experiments reported in the literature are based on the use of a "push-to-talk" speech acquisition method. In the past, a few works addressed the impact of EPD on distant-talker recognition performance. In [50] adaptive energy thresholds were applied to the output of the delay-and-sum beamformer to identify speech boundaries. In [38], a CSP-based coherence measure between two input channels was used to detect a generic acoustic event. Its effectiveness in speech activity detection will be confirmed in the next section. The application of a similar coherence measure to speech/noise classification is also documented in [63].

Acoustic Features In ASR research and applications, there is the tendency today to adopt a standard acoustic feature set (e.g. Mel or LPC cepstral co-

efficients). This is also the case for distant-talker ASR research. Most of the systems described in the literature are based on the use of MCC's or LPC cepstral coefficients together with their first/second order derivatives and cepstral mean subtraction. However, some other acoustic features more robust to noise, such as PLP or short-time modified coherence (SMC) [64,46], have been investigated. In the interest of acoustic feature robustness to reverberation effects, possible future approaches may be inspired by techniques which selectively process the linear prediction residual [65] or incorporate speech production modeling into the given multi-channel framework [66,67].

Recognition Engine The majority of recognizers investigated in the literature are based on the use of traditional HMM-based solutions. Training HMM's on artificially contaminated speech may lead to robust solutions when a large noisy database is not available, as shown in [68] and in [69]. In some cases, MAP, MLLR or ANN-based adaptation techniques have been adopted [70,71,54,53] to further reduce the mismatch between training and test conditions which remains after the microphone array based processing. In [56], MLLR is also compared favorably to Parallel Model Combination. Broadly speaking, these approaches allow the system to learn more about the speaker characteristics, the environmental noise, and the "artifacts"(e.g. low-pass effects typical of delay-and-sum beamforming) introduced by the multi-channel processing. Overall, the use of adaptation techniques has had a significant positive impact on hands-free ASR system performance.

An alternative approach that deserves to be mentioned is proposed in [61]. Here a new dimension is added to the search space used by the Viterbi algorithm to account for the different directions from which the talker may be speaking. In this way, changes in source position should be implicitly detected on the basis of a maximum likelihood criteria. The resulting system is more flexible, but possesses a considerable complexity increase and requires a consistent HMM training to be performed initially.

Another alternative approach is reported in [72], where an ANN is used to perform a transformation/normalization of the acoustic features extracted from the delay-and-sum beamformed signal. During training, the ANN learns information related to the talker position. The influence of the talker location is addressed in [71], where the effectiveness of a location independent ANN is demonstrated.

Speech Corpora and Tasks A variety of recognition tasks have been investigated in the literature. The most common is connected-digits. However, the choice of the experimental task is probably not as relevant as the way this task is created. In order to derive speech material for training and especially for test experiments, three main approaches have been adopted:

- Speech data is collected from a sampling of *real* talkers using multi-channel recording hardware. This method requires much more effort than

its alternatives, but it represents the most reliable means for investigation in this field and for comparing results to theory.

- Speech data (e.g. extracted from a given clean database) is played back through a loudspeaker. Again, multi-channel hardware is used to record the signals. With this method there is the advantage of repeatability of the same utterances under different conditions, array geometries, etc. However, the experiments may be influenced by artifacts inherent in the recording process, such as the dependency on loudspeaker response, the different radiation modeling, and other variabilities in the environmental conditions. It is worth mentioning the work done in Bell Laboratory's VarEchoic Chamber [73], by which any reverberant condition can be investigated without the risk of changing other environmental characteristics across recording sessions.
- Speech data is reproduced by simulation, typically based on a simplified additive/convolutive channel modeling. In this case, the reverberation effects on the various input channels are generally recreated by convolving the close-talker signal with real impulse responses measured using a time-stretched pulse [74,75] or with artificial impulse responses derived by applying the Image Method [76]. A simulation-based experiment has a clear limitation due to the fact that many phenomena occurring in a real environment may be neglected. Moreover, the use of simulation both in training and in test may provide misleading results due to biases in the artificial data generation.

15.3 A Microphone Array-Based ASR System

This section describes the distant-talking recognition system being developed and experimented with at IRST labs [70,60,59,77,68,78,79]. Figure 15.6 shows a block diagram of the system, consisting of: a microphone array and the related TDC processing, an acoustic front-end for speech activity detection and acoustic feature extraction, a recognition engine module (Viterbi decoding) and related modules for HMM adaptation. Each module of the system as well as the experimental framework will be described below together with the most relevant recognition results so far obtained.

15.3.1 System Description

Speech Acquisition Distant-talker speech signals were acquired by a linear array of six equi-spaced (at 15 cm) omnidirectional microphones. Each channel was synchronously sampled at a 16 kHz rate with 16-bit accuracy. Delays estimated between the channels (through CSP-based time delay estimation) were used to align the signals in a delay and sum beamformer.

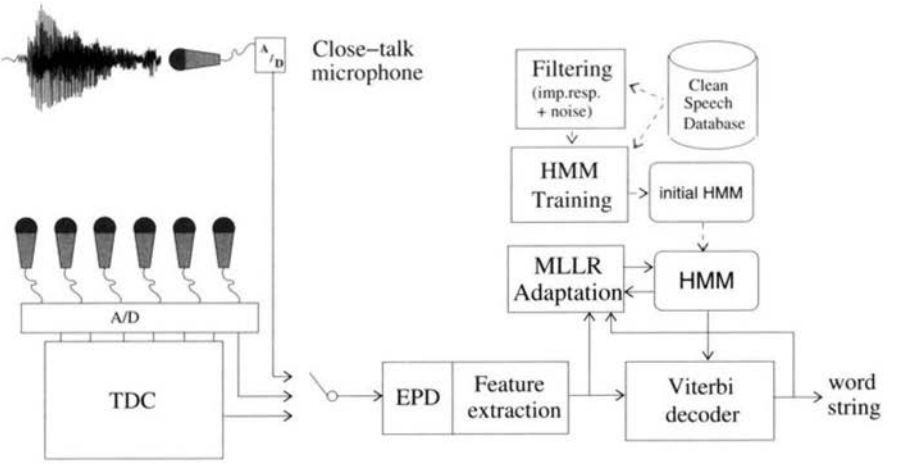

Fig. 15.6. Block diagram of the distant-talking ASR system developed at IRST. It includes signal acquisition, Time Delay Compensation processing, the acoustic front-end, a recognition engine, and the modules for training and adaptation of the acoustic models. Close-talker and single far-field microphones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

Speech-Activity Detection The use of a microphone array adds a spatial dimension to the domain of the time/frequency analysis of conventional single input systems. Besides source localization and selective acquisition by beamforming, an additional benefit of multi-microphone systems is the capability of discerning between coherent directive sources (e.g. a talker facing the microphones) and spatially diffuse, low coherence disturbances. The discriminating feature is a coherence measure between the signals of different microphones, such as the phase correlation [37,38]. Coherence measure computation is here extended to several microphone pairs to provide a more robust speech activity function. This function is effective for low SNR and reverberant signals, where an energy-based approach would not be.

Figure 15.7 illustrates an example of this procedure. The upper plot depicts the noisy speech signal acquired by a single microphone in the array. The middle plot represents the corresponding phase correlation between two channels of the array as a function of time (horizontal axis) and mutual delay in samples between the channels (vertical axis). A darker gray level denotes higher coherence. The lower plot shows the coherence measure at the true delay versus time.

In practice, the EPD technique proposed here is based on a preliminary selection of an inter-channel delay for each microphone pair. Given the inter-channel delays associated with the various microphone pairs, the appropriate coherence functions are summed to derive a speech activity function. Adaptive thresholds are then applied in order to determine speech boundaries.

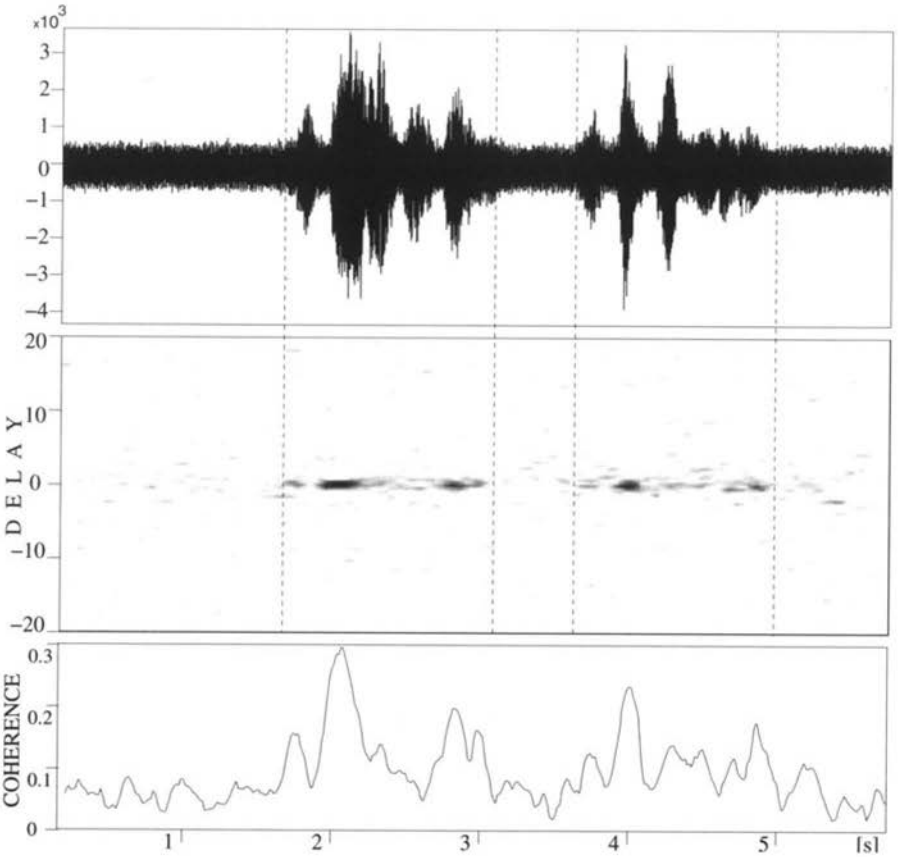

Fig. 15.7. Example of coherence computation for the signals of a microphone pair. It includes one of the noisy speech waveforms acquired by the microphones, a grey level coherence measure representation at various inter-channel delays, and the coherence levels at the correct delay (0 samples).

Acoustic Feature Extraction In the experimental set-up described here, the input to the feature extractor (see Figure 15.6) is either the output of the TDC processing derived from the microphone array data, the signal acquired by a single microphone within the array, or that acquired by the close-talker microphone.

In any case, the input signal is pre-emphasized and blocked into 20 ms, half-overlapping frames. For each frame, 8 MCC's and the log-energy are extracted. CMS is then applied to each feature sequence. The resulting normalized MCC's and log-energy, together with their first and second order time derivatives, are arranged into a single observation vector of 27 components.

HMM Recognizer Acoustic modeling is based on a set of 34 phone units. Each unit is modeled with a three state left-to-right continuous-density HMM, with output probability distributions represented by the means of mixtures having 16 Gaussian components with diagonal covariance matrices. Phone HMM's are trained either using a clean speech database or a noisy version, obtained by simulation as described below. Once the HMM's have been trained, the resulting models are adapted to the real environment by applying a MLLR adaptation technique [31–33].

15.3.2 Speech Corpora and Task

Various speech corpora have either been collected or produced in order to perform the experiments to be discussed.

Clean Speech Corpus The initial step of HMM training is accomplished through the standard Baum-Welch procedure. For this purpose, phonetically rich sentences representing a portion of APASCI [80] were used. This corpus was acquired in a quiet room (SNR $\geq$ 40 dB) using a high quality close-talker microphone. The training set consisted of 2100 utterances collected from a total of 100 speakers (50 males and 50 females).

Multi-Channel Real Noisy Corpus The multi-channel noisy corpus consists of speech material collected in an office of size (5.5 m $\times$ 3.6 m $\times$ 3.5 m) characterized by a moderate amount of reverberation ($T_{60} \simeq 0.3$ s) as well as by the presence of coherent noise due to secondary sources (e.g. computers, air conditioning, etc). Multi-channel recording of each utterance was accomplished by using both a close-talker (CT) directional microphone and the linear array described above.

Speech material was collected from 8 speakers (4 males and 4 females) during a series of recording sessions with variable environmental noise conditions. Each speaker uttered 50 connected digit strings (400 digit occurrences), both at frontal position F150 (1.5 m distance from the array) and at lateral position L250 (2.5 m distance, left of the array). Four of the individuals also uttered the same string set at position L150 (1.5 m distance, 60° right of the array).

Utterances were recorded with background noise segments of varying length at the beginning and end of each digit sequence. SNR, expressed as the ratio of the average speech to noise energy measured at the array microphones, was 12.6 dB mean with 3 dB standard deviation for the frontal recordings, and 10.7 dB mean and 2.8 dB standard deviation for the lateral recordings. As reference, SNR estimated on the CT microphone signals possessed a 28 dB mean and 3.8 dB standard deviation.

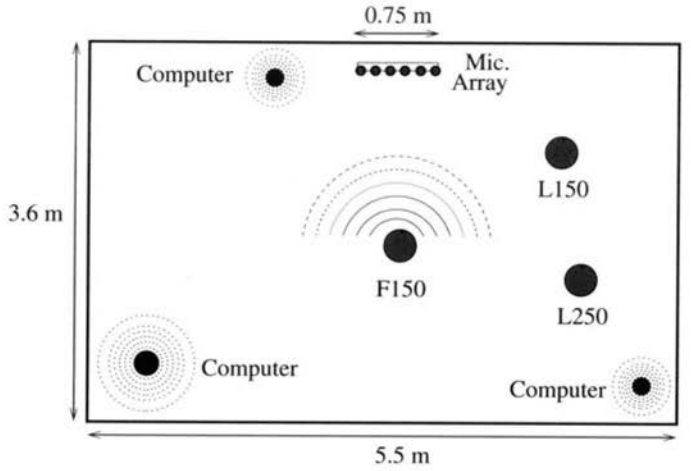

Fig. 15.8. Map of the experimental room showing the positions of the talker, the microphone array, and the computers (noise sources). The label at each position indicates, in a compact form, the orientation (F for frontal, L for lateral) and the distance in cm from the array.

Contaminated Speech Corpora A set of training databases consisting of acoustically realistic signals was artificially recreated using the APASCI clean corpus along with knowledge (e.g. room impulse responses and background noise signals) of the real operating environment. For this purpose, a simplified additive/convolutive model was adopted as follows:

$$s_{co}(t) = h_r(t) \star s_{cl}(t) + k \cdot n(t) \tag{15.1}$$

where $h_r(t)$ is an impulse response of the room, k is a scaling factor, $n(t)$ is background noise acquired in the room, s_{cl} is the clean speech, s_{co} is the contaminated speech, and $\star$ denotes convolution. The effect of background noise is accounted for by scaling the noise recorded inside the room using an appropriate amplitude to reproduce different SNR's (ranging from 0 to 21 dB) and then adding the result to reverberant speech. The reverberation effects of a room can be simulated in several ways. In this case, it was achieved by convolving the close-talker signal with impulse responses measured using a time-stretched pulse.

15.3.3 Experiments and Results

Experimental results involving connected-digit recognition are reported below. These are expressed in terms of Word Recognition Rate (WRR), computed as the average performance obtained by testing on material obtained from all the speakers and positions. As a result, each test experiment consists of the recognition of 8000 digits.

Table 15.1. Word recognition rates obtained on a connected-digit recognition task using different phone models, input devices, and end-point detection methods.

Models | Input	FarMic	Arr_{CSP_EPD}	Arr_{ID_EPD}
Clean	33.7	57.1	61.2
Rob	91.5	95.2	96.3
AdaRob	95.6	98.3	98.6

Experiments were conducted using either the microphone array (Arr) or a single microphone of the array (FarMic). For comparison purposes, results obtained testing on material acquired with the close-talker microphone (and using clean models) was approximately 99%. This reference result represents a sort of upper bound of any experiment conducted.

As shown in Table 15.1, training with filtered clean speech (Rob) improves recognition performance tangibly, even in the case of a single far microphone input. This result is consistent with other work [68,78,79]. The results confirm that the use of the microphone array, in combination with the TDC module ensures superior recognition performance relative to a single microphone. However, the advantage of using the array is more relevant in the case of robust models. In this case, more than 40% relative improvement was obtained (from 91.5% to 95.2%).

A second issue investigated was the impact of the speech activity detection method on the recognition performance. In addition to difficulties due to the distance between the talker and microphones, the system is prone to insertions in this experimental framework. This is due to the adoption of a digit-loop grammar with no information about string length. As a consequence, pauses inside a digit sequence and long noise segments, preceding and following the speech utterance, can cause many errors because of mismatched acoustic modeling. The right hand column in Table 15.1 (Arr_{ID_EPD}) shows the results obtained using an "ideal" end point detector. These were acquired using utterance boundaries identified manually and leads to a relative performance improvement of about 20% compared to results obtained using the coherence-based EPD method described earlier (Arr_{CSP_EPD}). Resolving this performance disparity is a goal for the future development of a more accurate EPD algorithm.

The results show the further improvement provided by adopting on-line incremental HMM adaptation (AdaRob). On-line adaptation is more suitable for real-time applications where environmental conditions, talker position, etc. may vary substantially with time. The Table shows that in the best case, that is starting from robust models and exploiting manually segmented speech boundaries, 98.6% WRR was obtained, not far from the close-talker reference performance. In previous work [68,79] it was shown that, when starting from

clean models, both batch and on-line adaptation techniques do not achieve this performance level. Finally, note that the adaptation produces a score of 95.6% WRR in the case of a single far-microphone.

15.4 Discussion and Future Trends

Hands-free interaction represents the most natural form of human communication. Research on hands-free speech recognition is drawing scientists together to form an important discipline with numerous potential applications. In particular, various multi-modal/multi-media interaction scenarios have be conceived thanks to the enhanced functionality added to traditional ASR systems.

Because of growing research and prototype development, the field of distant-talker speech recognition using microphone arrays has developed dramatically. As seen in this chapter, the introduction of a microphone array into an ASR system has the potential to improve performance significantly. However, this is at the cost of hardware and software complexity. Additional improvements are possible through the use of adaptation/compensation techniques and specific methods for acoustic model training. Through these approaches, performance increases can be achieved even using a single microphone. Hence, it seems reasonable that future research will focus on the use of arrays consisting of few microphones and the joint application of effective techniques for an on-line reduction of the mismatch between the operating conditions and those under which the system was trained.

Given the current state of the art, future research is needed in all the directions highlighted in the previous sections, from microphone array processing, to speech activity detection, to robust acoustic features, to adaptation of the recognizer to the real environmental conditions. Furthermore, new approaches will have to be envisaged to deal with the various environmental uncertainties which characterize distant-talker speech recognition applications.

Distributed multi-microphone systems [69], with instantaneous selection of the most reliable microphone input, may represent a promising approach. Along these lines, a specific sub-band recognizer or full-band recognizer may be associated with each of the given microphones. This would be with the purpose of realizing a competitive parallel recognition framework, where the recognized word string is selected among different hypotheses.

Another approach that deserves future study is that of incorporating speech production modeling into the multi-channel system and applying nonlinear analysis techniques as those proposed in [66,67] and detailed in Chapter 7. In this way, the system may be made less sensitive to the influence of variabilities related to reverberation effects, imperfect talker location, or a talker's head movements and may better focus attention on the speech propagating in the environment.

Finally, experimental tasks and activities are important aspects to highlight. Because of relevant differences in the experimental frameworks and the type of speech material (and languages) which are adopted, results obtained by the various research teams are often not comparable to one another. Moreover, results are often provided only on the basis of simulation experiments, while real world experiments are always needed to confirm a given theory. In the past, the most relevant and widely known activities for the development of basic speech recognition technology were carried out under the ARPA-CSR program. This produced the development of common speech material and standard evaluation criteria. Hence, the creation of a common framework for all the research centers operating in this field, may allow for significant advances in this exciting discipline.

References

1. L.R. Rabiner, B.H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.
2. R. De Mori, *Spoken dialogues with computers*, Academic Press, 1998.
3. A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, Kluwer, 1992.
4. Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.
5. J.C. Junqua and J.P. Haton, *Robustness in automatic speech recognition.* Kluwer, 1996.
6. C.H. Lee, F.K. Soong, and K.K. Paliwal, *Automatic speech and speaker recognition.* Kluwer, 1996.
7. S. Furui, "Recent advances in robust speech recognition," in *Proc. of ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 11–20, 1997.
8. M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, pp. 75–95, 1998.
9. J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.
10. L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals* Prentice Hall, 1978.
11. L.R. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Sys. Tech. Journal*, vol. 54, no. 2, pp. 297–315, 1975.
12. L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, pp. 777–785, 1981.
13. H. Ney, "An optimization algorithm for determining the endpoints of isolated utterances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-81)*, Atlanta GA, USA, pp. 720–723, 1981.
14. J.C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, pp. 406–412, 1994.

15. D. O'Shaughnessy, *Speech Communications*, IEEE Press, 2000.
16. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
17. C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, vol. 25, pp. 29–47, 1998.
18. J.S. Lim, *Speech Enhancement*, Prentice Hall, 1983.
19. Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. on Signal Processing*, vol. 40, pp. 1303–1316, 1992.
20. S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction* Wiley and Teubner, 1996.
21. S. Boll, "Speech enhancement in the 1980s, Noise suppression with pattern matching," in *Advances in Speech Signal Processing*, (S. Furui and M.M. Sondhi, eds.), pp.309–325, Marcel Dakker, 1992.
22. M. Rahim and B.H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 19–30, 1996.
23. C. Lawrence and M. Rahim, "Integrated bias removal techniques for robust speech recognition," *Computer Speech and Language*, vol. 13, pp. 283–298, 1999.
24. A. Sankar and C.H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 190–202, 1996.
25. A. Nadas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.
26. I. Sanches, "Noise-compensated hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5 , pp. 533–540, 2000.
27. Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 255–266, 2000.
28. M.J.F. Gales, *Model-based techniques for noise robust speech recognition*, PhD thesis, Cambridge University, Cambridge, England, 1995.
29. M. J. F. Gales and S. J. Young, "Robust speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 352-359, 1996.
30. J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
31. C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.
32. M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.
33. M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
34. S. Das, R. Bakis, A. Nadas, D. Nahamoo, and M. Picheny, "Influence of background noise and microphone on the performance of the ibm tangora speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, Minneapolis MN, USA, pp. 95–98, Apr. 1993.

35. B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 31, no. 4, pp. 793–897, 1983.
36. S. Furui, "Robust speech recognition under adverse conditions," in *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 31–42, 1992.
37. C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
38. M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.
39. Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, 1986.
40. R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-88)*, New York NY, USA, pp. 2578–2581, Apr. 1988.
41. S. Haykin, ed., *Advances in spectrum analysis and array processing.* Prentice Hall, 1995.
42. M. W. Hoffman and K. M. Buckley, "Robust time-domain processing of broadband microphone array data," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 3, pp. 193–203, 1995.
43. S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, no. 3-4, pp. 215-27, 1996.
44. O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. of IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
45. L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1 , pp. 27–34, 1982.
46. J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Multi-microphone noise reduction techniques for hands-free speech recognition - a comparative study," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, pp. 171–174, 1999.
47. J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, 1993.
48. E.E. Jan, P. Svaizer, and J.L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. of IEEE ISCAS*, pp. 1460–1463, 1995.
49. C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech and Audio Proc.*, vol. 6, no. 3, pp. 240–259, 1998.
50. D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-90)*, Albuquerque NM, USA, pp. 833–836, Apr. 1990.
51. Y. Grenier, "A microphone array for car environments," *Speech Communication*, vol. 12, pp. 25–39, 1993.

52. T.M. Sullivan and R.M. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-93)*, Minneapolis MN, USA, pp. 91–94, Apr. 1993.
53. P. Raghavan, R.J. Renomeron, C. Che, D.S. Yuk, and J.L. Flanagan, "Speech recognition in a reverberant environment using matched filter array (MFA) processing and linguistic-tree maximum likelihood linear regression (LT-MLLR) adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 777–780, Mar. 1999.
54. T.B. Hughes, H.S. Kim, J.H. DiBiase, and H.F. Silverman, "Performance of an HMM Speech Recognizer using a real-time tracking microphone array as input," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 3 , pp. 346–349, 1999.
55. T. Takiguchi, S. Nakamura, and K. Shikano, "Speech recognition for a distant moving speaker based on HMM composition and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 1403–1406, June 2000.
56. J. Kleban and Y. Gong, "HMM adaptation and microphone array processing for distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 1411–1414, June 2000.
57. D. V. Rabinkin, R. J.Renomeron, J. C. French, and J. L. Flanagan, "Optimum microphone placement for array sound capture," *Proc. of the SPIE*, vol. 3162, pp. 227-39, 1997.
58. M. Inoue, S. Nakamura, T. Yamada, and K. Shikano, "Microphone array design measures for hands-free speech recognition," in *Proc. of EUROSPEECH*, pp. 331–334, 1997.
59. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of different microphone array configurations for hands-free speech recognition in noisy and reverberant environments," in *Proc. of EUROSPEECH*, pp. 347–350, 1997.
60. M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 227–230, Apr. 1997.
61. S. Nakamura, T. Yamada, P. Heracleous, and K. Shikano, "Recognition of distant-talking speech based on 3-D trellis search using a microphone array and adaptive beamforming," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, pp. 219–222, 1999.
62. S. Oh, and V. Viswanathan, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-92)*, San Francisco CA, USA, pp. 281–284, Mar. 1992.
63. R. Le Bouquin, "Enhancement of noisy speech signals, application to mobile radio communications," *Speech Communication*, vol. 18, pp. 3–19, 1996.
64. D. Mansour and B.H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 6, pp. 795–804, 1989.
65. B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendano, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 405–408, May 1998.
66. M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA pp. 3613–3616, May 1998.

67. M. Brandstein, "An event-based method for microphone array speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 953–956, Mar. 1999.
68. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 449–452, Mar. 1999.
69. Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 197–200, June 2000.
70. D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation," in *Proc. of ICSLP*, pp. 1329–1332, 1996.
71. Q. Lin, C.W. Che, D.S. Yuk, L. Jin, B. de Vries, J. Pearson, and J.L. Flanagan, "Robust distant-talking speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-96)*, Atlanta GA, USA, pp. 21-24, May 1996.
72. C. Che, Q. Lin, J. Pearson, B. de Vries, and J.L. Flanagan, "Microphone arrays and neural networks for robust speech recognition," in *Proc. ARPA Human Language Technology (HLT)*, pp. 342–348, 1994.
73. W. Ward, G. Elko, R. Kubli, and W. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. of Wallance Clement Sabine Centennial Symposium*, pp. 343–346, 1994.
74. N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1484–1488, 1981.
75. Y. Suzuki, F. Asano, H. Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1995.
76. J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
77. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Experiments of HMM adaptation for hands-free connected digit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle WA, USA, pp. 473–476, May 1998.
78. D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of filtered clean speech for robust HMM training," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp. 99–102, 1999.
79. M. Matassoni, M. Omologo, and D. Giuliani, "Hands-free speech recognition using a filtered clean corpus and incremental HMM adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-00)*, Istanbul, Turkey, pp. 1407–1410, June 2000.
80. B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," in *Proc. of ICSLP*, pp. 1391–1394, 1994.

18 Future Directions in Microphone Array Processing

Dirk Van Compernolle

Katholieke Universiteit Leuven, Leuven, Belgium

18.1 Lessons From the Past

Antenna array processing has had long-standing impact on phased array radars, sonars and radio astronomy for several decades. The gigantic antenna arrays that were constructed for deep space observation must stand out as some of the most impressive engineering achievements of any discipline. Success in these related fields of signal processing have without any doubt stimulated interest in microphone array processing. And these successes did not only generate interest, they did much more—they created high expectations. Another interest generating stimulus came from a very different field, i.e. the one of anatomy and physiology. Nature has endowed virtually all species with two ears. One good reason, of course, is that there is always a second as backup when one of the two fails. But at the same time we all know that our sense for orientation is helped considerably by the use of two ears instead of one and that it helps us understand each other in the midst of a noisy crowd.

After 20 years of active research, however, we cannot claim that microphone array processing has had the success many of us hoped for, and many will wonder when the great breakthrough in microphone array processing will finally come, if ever. Nevertheless, progress in computer technology has helped us in a big way. In the early days only analog schemes of limited signal processing complexity were possible. This was followed by early years of high cost DSP computations, where computational cost seemed to impede widespread use of the technology. Today we have affordable DSPs that allow us to implement all but the most complex schemes cheaply in digital signal processing technology in real-time. But this in itself was not enough. Apart from breaking through the computing bottleneck, our understanding of the problems at hand has significantly progressed, as witnessed in this book. Most of the results presented are from recent years and give new insight into both the potential and the limitations of microphone array processing. However, too often the same problems that were considered too hard ten or twenty years ago are still set apart for 'future research'. Admitted weaknesses to proposed solutions are similar to the ones that we have been struggling with for a long time. Generally speaking we may say that many proposed solutions add to our understanding but lack robustness in order to make a bright future for themselves.

So should we not ask ourselves if there is a fundamental issue with microphone array processing? And my answer is 'nothing is fundamentally wrong'. Microphone array processing has only proven to be quite a bit harder than other previously successful array processing applications. We have known the problems from the beginning, but have underestimated the impact of some of them in real-life situations.

The basic problems fall into a small number of categories: (i) the speech signal is broadband; (ii) in many practical situations the desired source is in a reverberant space in the near or mid field, is moving, and cannot be assumed to be a point source; and (iii) the speech signal changes rapidly, it is intermittent, and shares many characteristics with the competing and interfering signals.

It is very difficult to tackle all these issues at once. It is especially difficult to come up with tractable mathematical models for this complex environment. The result of this complex situation is that a lot of research effort has gone into, and continues to go into the search for optimal beamforming strategies that rely on extra assumptions and constraints. Sadly enough, this all too often leads to solutions that lack robustness when evaluated in a variety of real-life situations. It may be that a far field assumption is required, it may be that less reverberation would be sufficient, or it may be that a perfect predictive speech detector will bring the breakthrough. Surely these mathematical developments are relevant and give us a better understanding of broadband beamforming in general. Simultaneously we should admit to ourselves that robustness has been, and still is today, one of the main issues.

The drive to achieve (mathematically) optimal solutions is a natural underpinning of our science and engineering nature. But is microphone array processing not too complex to be solved with optimal approaches? Should we not expect real breakthroughs to come from so-called robust solutions that are clearly sub-optimal for any given circumstance, but applicable in a relatively wide range of situations? Also is it not obvious that there will not be a single solution, but that we need quite different solutions depending on the target application(s)? These observations go hand in hand with one of the major problems that has faced microphone arrays since their debut: size and cost. A large size always seemed a must from the requirement of uniform broadband beamforming. Some of the first microphone arrays, especially the one constructed in the auditorium at Bell Labs, were magically impressive by their shear size and number of microphones. They were great fun as a research project. Also they resulted in functional solutions. At the same time the price of such systems seems exorbitant. Later on we saw many arrays on the order of, say, 1 meter. Any such design is still only applicable to a very limited number of applications such as conference rooms. In the majority of potential applications such a bulky design has no place. No industry has screamed more for tiny and low cost solutions than the hearing aid industry. Here spacing of a few cm are the maximum and processing power is an or-

der of magnitude less than in desktop applications. In all these situations we should not be surprised about the small size and limited number of sensors (two) in human hearing. It is far from optimal, but it works.

18.2 A Future Focused on Applications

If we ask ourselves what will the future bring for microphone array processing, we must envisage a range of widely differing solutions of different sizes and costs.

In the sequel I analyze the potential of the most important market segments and go looking for killer applications. If commercialization has not yet started, the question is of course what hampered commercial introduction and when, if ever, will we see usage of microphone arrays in each of these application fields.

18.2.1 Automotive

If any 'killer application' exists for microphone arrays, then it should be voice input in the car. It has all the right ingredients. Mobile telephony and speech recognition scream for hands-free voice input in a noisy environment. Signal-to-noise ratios obtained by single microphones are just not sufficient. Thus microphone arrays seem the logical solution. There are extra features that should help. The speakers inside a car are not mobile and their position is reasonably constant from session to session. Ultimately, a market potential of tens of millions of units per year should be commercially convincing. All of this should be sufficient for successful uptake of microphone array technology, but is it?

Today, penetration of microphone arrays in cars is minimal, except for a few top brands that are not all that noisy by themselves, and therefore have the least need for it. The major concern of car equipment manufacturers and car manufacturers alike is cost: multiple microphones, multiple wires, extra DSP power required, etc. Every cent in every component counts when putting a car together and microphone arrays have been judged as too expensive. Also, at least for the foreseeable future one should envisage that most microphones are mounted into existing cars, further complicating the story for arrays.

Given this large cost concern I do not believe that large arrays spanning the entire car will ever be viable. On the contrary, the car is an ideal environment for a microphone array that behaves like a traditional directional microphone but with a slightly steerable beam. Such an embedded array can be mounted as any regular microphone by not so specialized technicians. I do believe that the development of better microphones for usage inside the car will be a point of focus for microphone developers in the coming years.

18.2.2 Desktop

Cost has been the stumbling block as well for desktop microphone arrays in conjunction with speech recognition in the PC environment. People just do not like it if you tell them that the accuracy of a $50 speech recognition software package will drastically improve if they buy a $150 microphone array to go with it. It is still unclear if we will ever overcome the cost hurdle in this case. It may just be a question of whether large enough volumes will ever be reached such that current prices can be lowered drastically.

Microphone arrays for the desktop have just started to appear on the market. The reviews so far are ambiguous. In *quiet* environments they work as well as any headset worn. So if you do not want to be physically hooked up to your computer, this is the way to go the reviews say. At the same time the reviews will warn you that the existing commercial array microphones do not work well in considerable noise, and that one should not move around. Current reviews unanimously advise a wireless headset if one needs to move around a lot.

It seems therefore that current commercial implementations only solve a small part of the problem. All of the designs rely primarily on fixed beamforming, most often with limited directionality adjustment. On top of this, some additional noise suppression may be used. The 'speech seeking' part seems to be insufficient in all of the produced arrays. Also the quality and speed of tracking is substandard. It just shows how great the robustness issue really is when bringing microphone array technology to consumer products.

All in all there is reason for optimism, however. Desktop arrays are very pragmatic in their designs. These microphones are built for applications that use a PC screen or monitor, and they sit perfectly well on top of a monitor or attach to the front of it. Overall size is limited to about 20 cm, all computing is done inside the array, and the array connects to other equipment just as any other microphone would do. We have come a very long way to bring prices down enough such that a single enclosure with multiple elements, A/D converters and a DSP can be made at prices competing with traditional high end microphones. And let us not forget that these are first generation devices and that volumes are still very small.

Given some more time, I believe that there is hope that microphone arrays will capture a part of this market. Who knows, 5 years from now microphone arrays may be standard equipment on laptop and desktop computers. There is also a chicken and egg situation here. A wider usage of speech recognition would put more pressure on hardware manufactures to include higher end microphones, including arrays. On the other hand, one of the main hurdles in improving performance and subsequent acceptance of speech recognition is the low quality audio input on most systems today.

18.2.3 Hearing Aids

Hearing aids form a market by themselves. Restrictions on size and computational power are an order of magnitude more stringent than in other areas, leading to substantially different designs. Array sizes of 5 to 20 cm have been used in experiments with hearing aids, but have overall been met with disapproval. Nevertheless, here we have also seen the introduction of a range of new multi-microphone based products in the last couple of years. Many of these products do not use classical arrays, but a combination of microphones with different characteristics, used as inputs to a noise suppression stage. Perhaps even more obvious than in the automotive or desktop case, the evolution is towards an adaptive speech seeking and noise suppressing microphone. The distinction here between microphone technology and array technology is not entirely clear (but that does not really matter).

18.2.4 Teleconferencing

Teleconferencing was for some time seem as one of the potential killer applications. But I think that this is no longer true. On the one hand, the expansion of the teleconferencing market seems to have come down to slow growth and we see nothing of the explosion that some had hoped for. Therefore, the hope for a massive market does not seem justified. Acoustic echo cancellation is the crucial issue and it can not be solved by array processing. When using arrays, as with any multi-microphone input, the problem becomes significantly worse. Special microphone designs, including radial arrays, have been constructed and will continue to play a role in this market. Large wall mounted microphone arrays, however, are unlikely to find their way into teleconferencing rooms in any big way.

18.2.5 Very Large Arrays

Teleconferencing was one of the potential markets for large arrays. Another one is the virtual microphone in large auditoria. However, this can not be considered a booming market either. Design and manufacturing of these arrays is costly and a large degree of optimization may be required from site to site, making the picture even worse. Hence large microphone arrays are doomed to remain a niche market. They will certainly survive in high profile demonstration projects, and as a research topic they will carry on for many years to come. Another (quite niche) market for very large arrays exists in the acoustic monitoring industry.

18.2.6 The Signal Subspace Approach - An Alternative to Spatial Filtering ?

Finally, we should ask ourselves the question if we should not look for alternative solutions to plain spatial filtering. We may think in two directions:

blind signal separation and signal subspace approaches. These techniques do not require sensitive geometric information about the array layout but work with any configuration.

These techniques should result in higher configuration robustness. But at the same time they are computationally very demanding and, while making fewer assumptions about the layout, they make in general more assumptions about the signals. Practical implementations have not appeared so far, but demonstration results are often impressive. So we should keep an eye open for these techniques. It is unlikely we will find them in products in the coming years, but in later generation array processing techniques, they may become the standard way to go.

18.3 Final Remarks

The near-term trend is in one direction: small arrays with few microphones and a high degree of robustness that behave as speech seeking, directional, and noise canceling microphones. Depending on the target application designs may vary from less than a 1 cm in diameter for the hearing aid market, over 5 cm for the car, to a maximum of 20 cm for desktop. After all, human hearing does very well with two ears spaced about 20 cm apart. These designs will not reach maximal noise suppression in any theoretical sense. Their goal is clear: a few dB gain in signal-to-noise ratio across the board at a cost which is only marginally above that of other microphones. A market of several million units for such medium cost devices is realistic and therefore economically viable. Economic potential for large arrays is much more limited and will therefore remain a niche market.