Signals and Communication Technology

Patrick A. Naylor
Nikolay D. Gaubitch
(Eds.)

Speech Dereverberation

Springer

Signals and Communication Technology

Patrick A. Naylor · Nikolay D. Gaubitch
Editors

Speech Dereverberation

Springer

Patrick A. Naylor, PhD
Nikolay D. Gaubitch, PhD

Department of Electrical and Electronic Engineering
Imperial College London
Exhibition Road
London SW7 2AZ
United Kingdom

p.naylor@imperial.ac.uk
ndg@imperial.ac.uk

ISSN 1860-4862
ISBN 978-1-84996-055-7 e-ISBN 978-1-84996-056-4
DOI 10.1007/978-1-84996-056-4
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2010930018

Cover design: WMXDesign, Heidelberg, Germany

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book owes its existence to the numerous students and co-workers who, over the years, have worked with me and inspired me. I gratefully acknowledge their contributions.

I dedicate this book to my wife Catharine who has taught me so much.

Patrick A. Naylor

To my mother.

Nikolay D. Gaubitch

Preface

Speech dereverberation has been on the agenda of the signal processing community for several years. It is only in the last decade, however, that the topic has really taken off, as seen from the growing number of publications appearing in the journals and at conferences. One of the reasons that the topic has become more popular is the rapidly growing availability in the marketplace of computationally capable mobile devices, such as phones, PDAs and laptop computers, for which hands-free (distant talking) operation is desirable. This is all the more significant when seen in the context of the confluence of computing and communication terminals exploiting low-cost VoIP-enabled telephony applications. Additionally, it is also true to say that user expectations of computing and communication devices is a strongly increasing function with time, perhaps only moderated by considerations of value versus cost – people are more forgiving of technology limitations if they are not paying (much) for the service they are employing. Factors such as these have combined to motivate the signal processing community to provide robust solutions for speech enhancement in general and to work on in particular, what for many is a new task, dereverberation.

Since we began our research in this field, we have been receiving inquiries from curious researchers seeking a digestible review on the state-of-the-art in the field of speech dereverberation. Until now, the answer has always been that, although there have been several books that treat the subject of speech processing, microphone array processing, and audio processing, which have included chapters on speech dereverberation, there has not been a publication that gives a comprehensive overview of the topic. We believe that the field has now reached a maturity that allows the compilation of such a book, solely dedicated to the topic of speech dereverberation. It was this belief and the context of the situation that motivated our initiative in this writing project.

Before you decide to skip the rest of this Preface on the grounds that its authors have lost their grip on reality, let us momentarily clarify the level of maturity to which we are referring. The three main axes of speech enhancement were highlighted by Walter Kellermann at the 1999 International Workshop on Acoustic Echo and Noise Control to be echo cancellation, noise reduction and dereverberation. Of these three it is a likely consensus that dereverberation is the more difficult

task. Modelling of room acoustics is more complicated than either the modelling of speech production or of noise generation processes and their additive combination with speech signals, at least in the manner in which such models are currently applied in DSP algorithm development. Dereverberation is also normally formulated as a blind (or unsupervised) problem, somewhat related to, but nevertheless distinct from, blind source separation. Computational limitations both in power and precision also present real challenges in this field. So, it is inevitable that, given the difficulty of the problem and the fact that attention on this problem has not been strongly focused for as long as it has on either echo cancellation or noise reduction, the level of maturity in the understanding of the dereverberation problem and its solutions is far below that of the other related problems. At this stage of dereverberation technology, we could argue that there are more open questions than solutions; those solutions that are available strive towards, but do not always achieve, the levels of robustness found in many of the more mature technologies.

This book, therefore, by no means offers any ultimate solutions to the speech dereverberation problem. Nonetheless, it aims to provide an in-depth overview of the state-of-the-art in speech dereverberation methods with contributing chapters from some of the key researchers in the field. It also gives what we believe to be a valuable introduction to some topics relevant to dereverberation, such as room acoustics and psychoacoustics, though we have not aimed to cover these subjects in detail.

The book is aimed at researchers and graduate students who would like to pursue research in this field, giving an accessible introduction to the topic including numerous references to other publications. It could make an excellent complementary text for postgraduate courses in speech processing. However, it is not exclusively limited to this group of readers. The attempt to solve the very difficult problem of speech dereverberation has involved a large variety of signal processing tools. Such tools include multirate signal processing, adaptive filtering, Bayesian inference, and linear prediction, to mention but a few. The applications of these techniques can be found useful in other fields of engineering.

We would like to thank all the contributing authors for their excellent chapters. We would also like to express our special thanks to Dr. Emanuël Habets for his carefully reading of our drafts and for his helpful discussions and contributions throughout this project.

London, February 2009

Patrick A. Naylor
Nikolay D. Gaubitch

Contents

List of Contributors

Herbert Buchner
Deutsche Telekom Laboratories, Berlin University of Technology, Germany
e-mail: hb@buchner-net.com

Marc Delcroix
NTT Communication Science Laboratories, Japan
e-mail: marc.delcroix@gmail.com

Christine Evers
University of Edinburgh, UK
e-mail: c.evers@ed.ac.uk

Steven Fortune
University of Edinburgh, UK
e-mail: steven.fortune@ed.ac.uk

Sharon Gannot
Bar-Ilan University, Israel
e-mail: gannot@macs.biu.ac.il

Nikolay D. Gaubitch
Imperial College London, UK
e-mail: ndg@imperial.ac.uk

Emanuël A.P. Habets
Imperial College London, UK
e-mail: e.habets@imperial.ac.uk

Takafumi Hikichi
NTT Communication Science Laboratories, Japan
e-mail: hikichi@cslab.kecl.ntt.co.jp

James R. Hopgood
University of Edinburgh, UK
e-mail: james.hopgood@ed.ac.uk

Walter Kellermann
University of Erlangen-Nuremberg, Germany
e-mail: wk@lnt.de

Andy Khong
Nanyang Technological University, Singapore
e-mail: AndyKhong@ntu.edu.sg

Keisuke Kinoshita
NTT Communication Science Laboratories, Japan
e-mail: kinoshita@cslab.kecl.ntt.co.jp

Masato Miyoshi
NTT Communication Science Laboratories, Japan
e-mail: miyo@cslab.kecl.ntt.co.jp

Tomohiro Nakatani
NTT Communication Science Laboratories, Japan
e-mail: nak@cslab.kecl.ntt.co.jp

Patrick A. Naylor
Imperial College London, UK
e-mail: p.naylor@imperial.ac.uk

Mark R.P. Thomas
Imperial College London, UK
e-mail: mark.r.thomas@imperial.ac.uk

Jimi Y.-C. Wen
Imperial College London, UK
e-mail: jimi.wen@imperial.ac.uk

Takuya Yoshioka
NTT Communication Science Laboratories, Japan
e-mail: takuya@cslab.kecl.ntt.co.jp

Chapter 1
Introduction

Patrick A. Naylor and Nikolay D. Gaubitch

Abstract Acoustic reverberation will be introduced in this chapter in the context of telecommunication. The adverse effects on speech caused by reverberation are problematic, in particular, in hands-free terminals operating typically at arms-length from the talker's lips. This introductory chapter will provide a system description of room reverberation and will formulate mathematically the dereverberation problem in its most direct form so as to introduce and underpin the more detailed presentation in subsequent chapters. Elements of room acoustics will also be introduced where needed, though detailed study of acoustics is not the aim of this text.

At the time of writing this, dereverberation is a topic of study with many important research questions remaining as yet unanswered. Whilst reviewing the relevant literature later in this chapter, it is intended both to describe the state-of-the-art and to highlight some of the significant open issues. Whereas the former aims to consolidate, perhaps for the first time, the known achievements to date of the research community, the latter aims to highlight potential avenues of future research.

1.1 Background

One can confidently deduce that the phenomenon of reverberation has been known to mankind since the time of prehistoric cave dwellers. Sound reflection effects are believed to have influenced prehistoric cave art [90, 91]. Reverberation is also used in several well known cases by other species, such as bats for navigation during flight.

There is evidence of comprehension of the notion of reflected speech occurring in Plato's *Republic* [76]: *"And what if sound echoed off the prison wall opposite them? When any of the passers-by spoke, don't you think they'd be bound to assume that the sound came from a passing shadow?"*. Pioneering scientific work on sound and acoustics in the 19th century was undertaken by, for example, Rayleigh [80] and

Imperial College London, UK

Sabine [81]. In the 20th century initial efforts in the understanding of reverberation of speech were provided by Bolt [8] and the effects of single echoes by Haas [37].

The level to which humans employ reverberation during everyday life is unclear. There is some evidence to suggest that, through the use of two ears, spatial processing is used to enhance speech intelligibility and enables a useful degree of source separation to be achieved in human speech perception [12].

In music audio processing, the sense of 'space' that can be created by stereo or surround sound reproduction adds greatly to realism and often makes recorded music more attractive and enjoyable. We might then ask ourselves the question: *since reverberation is present in everyday life experience of sound and in some important cases is effective in aiding speech communication, why should we be interested in removing reverberation from speech using dereverberation processing?* The answer to this question is dependent on the application context.

There is a continuously growing demand for high quality hands-free speech input for various telecommunication applications [71, 82]. One driving force behind this development is the rapidly increasing use of portable devices such as mobile telephones, Personal Digital Assistant (PDA) devices and laptop computers equipped for Voice Over Internet Protocol (VoIP) [63]. Furthermore, there is a continuous worldwide expansion of broadband internet access [5]. These factors have paved the way for several advanced speech applications such as wideband teleconferencing with automatic camera steering, automatic speech-to-text conversion, speaker identification, voice-controlled device operation and car interior communication systems [82]. Another important application where speech obtained from a distant talker is of interest is that of hearing aids [82].

1.2 Effects of Reverberation

When speech signals are obtained in an enclosed space by one or more microphones positioned at a distance from the talker, the observed signal consists of a superposition of many delayed and attenuated copies of the speech signal due to multiple reflections from the surrounding walls and other objects, as illustrated in Fig. 1.1. We here define the direct-path as the acoustic propagation path from the talker to the microphone without reflections. We also note that a delay of the superimposed copies arises because all other propagation paths are longer than the direct-path and that additional attenuation occurs at each reflection due to frequency dependent absorption. The perceptual effects of reverberation can be summarized as:

1. *The box effect* – the reverberated speech signal can be viewed as the same source signal coming from several different sources positioned at different locations in the room and thus arriving at different times and with different intensities [3]. This adds spaciousness to the sound [56] and makes the talker sound as if positioned "inside a box".
2. *The distant talker effect* – the perceived spaciousness explained in the previous point makes the talker sound far away from the microphone.

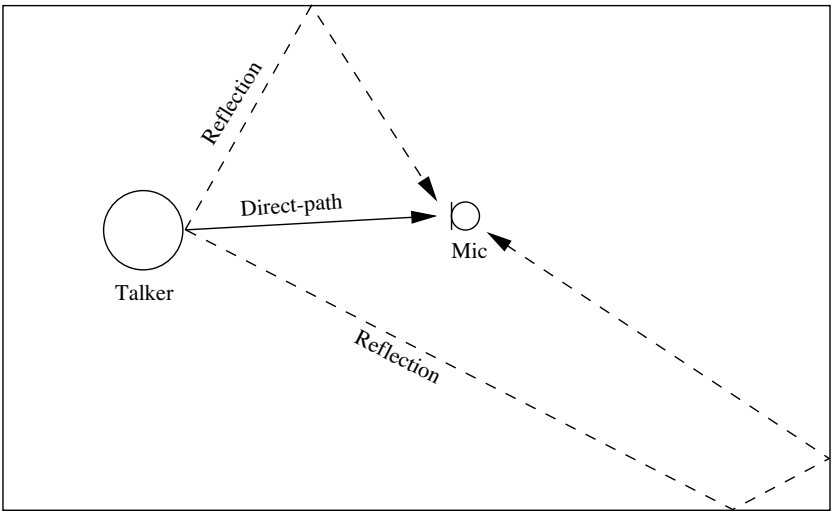


Fig. 1.1 Schematic illustration of room reverberation

When these effects are carefully controlled and moderately applied, the reverberation can add a pleasant sense of the acoustic space in which the sound resides. This is valuable and important in audio rendering but almost always unhelpful in voice communication. When the reverberation effects are severe, intelligibility of speech is degraded. Reverberation alters the characteristics of the speech signal, which is problematic for signal processing applications including speech recognition, source localization and speaker verification, and significantly reduces the performance of algorithms developed without taking room effects into consideration. The deleterious effects are magnified as the distance between the talker and the microphones is increased.

1.3 Speech Acquisition

The problems associated with reverberation can sometimes be overcome in practice by utilizing a headset by means of which the microphone is held close to the mouth. Alternatively, a microphone with a fixed directional sensitivity characteristic positioned in front of the talker can be used. The advent of bluetooth technology has made high quality, low cost wireless headsets feasible. Nevertheless, these solutions impose restrictions on the flexibility and comfort of the talker, which are the main desired features in the use of hands-free equipment [71]. In some applications, such as teleconferencing with multiple talkers on one end, these headset based solutions may not be practical. Therefore, a signal processing approach independent of the relative talker-microphone configuration is certainly preferable.

In hands-free speech acquisition, the talker's lips are typically located at a distance of 0.3-3 m from the microphone. In such a scenario, the speech signal is affected by the user's surrounding environment, which results in the following three distinct effects [71, 82]:

(i) *Additive measurement noise* due to, for example, other audible talkers or passing traffic. When the noise level is comparable to or greater than the speech level, it is difficult for a listener to distinguish the desired speech signal from the noise, and thus intelligibility and listening comfort are reduced.
(ii) *Acoustic echoes* due to speech from a far-end talker which is picked up by the near-end microphones and retransmitted back to the far-end talker with delay. This results in the talker hearing an echo of their own voice, which greatly disturbs the communication.
(iii) *Reverberation* that arises whenever sound is produced in enclosed spaces, such as offices and other rooms, due to reflections from walls and surrounding objects.

These components jointly contribute to an overall degradation in the quality of the observed speech signals, which significantly reduce the perceived speech quality for the listener and the performance of applications such as speech recognizers [71]. Speech enhancement and acoustic echo cancellation are two widely researched fields that address problems (i) and (ii) respectively. Several significant contributions have been made in these areas [6, 7, 11, 13, 30, 41, 60] and many algorithms have been implemented and are in use in commercial applications [77]. The problem of reverberation on the other hand, received much less attention in the literature until recently. Nevertheless, finding solutions to this problem is essential for the future development of applications with hands-free speech acquisition. This indeed motivates the focus of the forthcoming chapters of this book.

1.4 System Description

A generic system diagram for multichannel dereverberation is shown in Fig. 1.2. The speech signal, $s(n)$, from the talker propagates through acoustic channels, $H_m(z)$ for $m = 1$ to M. The output of each channel is observed using M microphones to give signals $x_m(n)$. All noise in the system is assumed additive and is represented by $v_m(n)$.

The observed signal, $x_m(n)$, at microphone m can be described as the superposition of (i) the direct-path signal, which propagates by line-of-sight from the talker to the microphone with corresponding attenuation and propagation delay and (ii) a theoretically infinite set of reflections of the talker signal arriving at the microphone at later time instances [56] with attenuation dependent on the properties of the reflecting surfaces. This can be expressed as

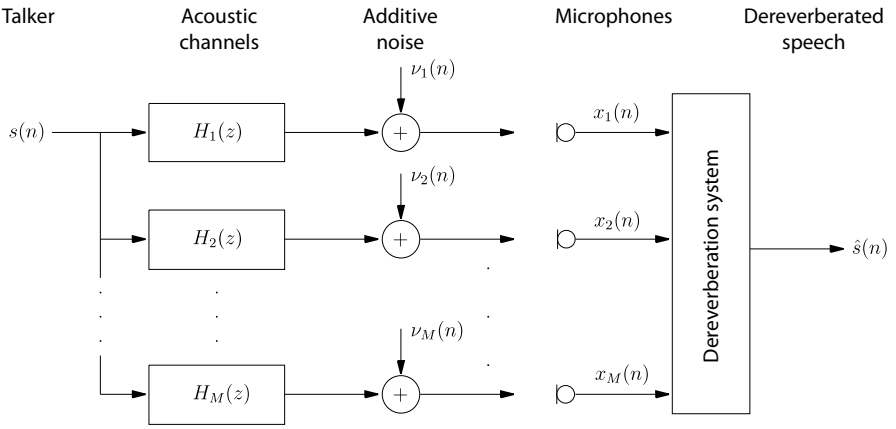


Fig. 1.2 Generic multichannel reverberation-dereverberation system model

$$x_m(n) = \sum_{i=0}^{\infty} h_{m,i}(n)s(n-i), \tag{1.1}$$

where the acoustic channel impulse responses $h_{m,i}(n)$ represent the attenuation and the propagation delay corresponding to the direct signal and all the reflected components.

The aim of speech dereverberation is to find a system with input $x_m(n),\ m = 1,\ldots,M$ and output $\hat{s}(n)$, which is a 'good' estimate of $s(n)$. The definition of 'good' is this context is application dependent. It may, for example, be desired to estimate $s(n)$ with minimum Mean Square Error (MSE). Alternatively, other criteria may be relevant, such as those related to perceptual quality. This is a blind problem since the acoustic channels $H_m(z)$ are unknown.

Recent efforts in acoustic signal processing have produced several algorithms for speech dereverberation and reverberant speech enhancement. These methods can be divided broadly into three main categories:

1. *Beamforming* – the signals received at the different microphones are filtered and weighted so as to form a beam of enhanced sensitivity in the direction of the desired source and to attenuate sounds from other directions. Beamforming is dependent on the availability of multi-microphone inputs. Beamforming is a multiple input single output process.
2. *Speech enhancement* – the speech signals are modified so as to represent better some features of the clean speech signal according to an *a priori* defined model of the speech waveform or spectrum. Speech enhancement is often a single input single output process, though many speech enhancement techniques benefit from the use of multiple inputs.
3. *Blind deconvolution* – the acoustic impulse responses are identified blindly, using only the observed microphone signals, and then used to design an inverse filter that compensates for the effect of the acoustic channels.

1.5 Acoustic Impulse Responses

The Acoustic Impulse Response (AIR) characterizes the acoustics of a given enclosure and therefore study of the AIR is a natural approach to dereverberation. This section will introduce some of the characteristics of AIRs. The focus is on the AIRs of rooms where reverberation has a significant effect on telecommunication applications. Further relevant details of room acoustics are given in Chap. 2. Whereas AIR is used to refer to acoustic impulse responses in general, there are some cases where it is more appropriate to limit the acoustic context to be within a room, in which case, the impulse response is referred to as a Room Impulse Response (RIR). In this book we will use AIR and also RIR, depending on the acoustic scenario being considered.

Several models of room impulse responses have been considered in the literature, including both Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) structures [40, 47, 48, 65, 66, 74]. The choice of AIR model will generally influence the algorithmic development.

An often-used quantification of the impulse response of a room is the reverberation time, originally introduced by Sabine [56]. The reverberation time, T_{60}, is defined as the time taken for the reverberant energy to decay by 60 dB once the sound source has been abruptly shut off. The reverberation time for a room is governed by the room geometry and the reflectivity of the reflecting surfaces.

The reverberation time is approximately constant when measured at any location in a given room. However, the impulse response is spatially variant and will vary as the talker, the microphones or other objects in the room change location [56]. A particular characteristic that varies with the talker-microphone separation is the relation between the energy of the direct-path component and the energy of the reflected components of the AIR. The critical distance is the distance such that these two energies are equal.

Figure 1.3 shows an example room impulse response. Direct-path propagation from the sound source to the microphone gives rise to an initial short period of near-zero amplitude, sometimes referred to as the direct-path propagation delay, followed by a peak. The amplitude of this peak due to direct-path propagation may be greater or less than the amplitude of the later reflections depending on the source-microphone distance and the reflectivity of the surfaces in the room. The example of Fig. 1.3 shows a relatively strong direct-path component, indicating that the source-microphone distance is relatively short.

The early and the late reflections are indicated in the figure as two distinct regions of the AIR. The early reflections are often taken as the first 50 ms of the impulse response [56], and constitute well defined impulses of large magnitude relative to the smaller magnitude and diffuse nature of the late reflections. The propagation from the talker's lips to the microphone is represented by the convolution of the speech signal with the AIR. The AIR early reflections cause spectral changes and lead to a perceptual effect referred to as coloration [56]. In general, closely spaced echoes are not distinguished by human hearing due to masking properties of the ear, and it has been shown that early reflections can have a positive impact on the intelligibility

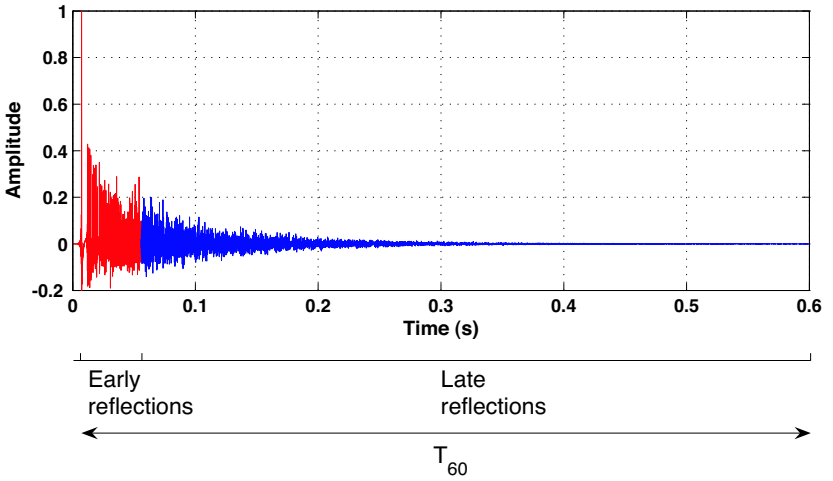


Fig. 1.3 An example room impulse response

of speech with an effect similar to increasing the strength of the direct-path sound [9, 56, 71]. However, coloration can degrade the quality of recorded speech [56]. The late reflections are referred to as the tail of the impulse response and constitute closely spaced, decaying impulses, which are seemingly randomly distributed. The late reflections cause a 'distant' and 'echo-ey' sound quality, we refer to as the reverberation tail and provides the major contribution to what is generally perceived of as reverberation in everyday experience.

In terms of spectral characteristics, the room transfer function is proportional to the sound pressure [56, 92] and has been studied extensively in the room acoustics literature, where many properties have been established [56]. One property of interest in the context of dereverberation is the average magnitude difference between minimum and maximum spectral points, which has been shown to extend beyond 10 dB [56, 83]. Since the room transfer function changes depending on the location of the source and the microphone, it can been described as a random process [56, 79, 92]. Neely and Allen [68] concluded that the AIRs in most real rooms posses non-minimum phase characteristics.

Rooms are generally stable systems with the coefficients $h_{m,i}(n)$ tending to zero with increasing index i and therefore, it is sufficient to consider only the first L_h coefficients in (1.1). The choice of L_h is often linked to the reverberation time of the room. Taking into account any additive noise sources, the observed signal at the m^{th} microphone can be written in a vector form

$$x_m(n) = \mathbf{h}_m^T(n)\mathbf{s}(n) + \nu_m(n), \tag{1.2}$$

where $\mathbf{h}_m(n) = [h_{m,0}(n)\ h_{m,1}(n)\ \ldots\ h_{m,L_h-1}(n)]^T$ is the L_h-tap impulse response of the acoustic channel from the source to microphone m, $\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L_h+1)]^T$ is the speech signal vector and $\nu(n)$ is observation noise.

In the frequency domain this can be expressed, equivalently, as

$$X_m(e^{j\omega}) = H_m(e^{j\omega})S(e^{j\omega}) + \mathcal{N}_m(e^{j\omega}), \tag{1.3}$$

where $X_m(e^{j\omega})$, $H_m(e^{j\omega})$, $S(e^{j\omega})$ and $\mathcal{N}_m(e^{j\omega})$ are the Fourier transforms of $x_m(n)$, $\mathbf{h}_m(n)$, $s(n)$ and $\nu_m(n)$, respectively.

Having introduced these properties of the room impulse response, the following can be deduced regarding the processing of reverberant speech:

1. Hand-free telephony users can be expected to move around their acoustic environment and so the AIRs will vary with time.
2. The use of measured impulse responses is not feasible for dereverberation due to the dependence on talker-microphone position and on the room geometry.
3. If the talker-microphone separation is much smaller than the critical distance, the effects of reverberation are likely to be negligible. Thus, dereverberation is of greatest importance when the source-microphone distance (D) is larger than the critical distance (D_c), $D \geq D_c$.
4. The reverberation time in typical office-sized rooms can be expected to vary in the range 0.1-1 s. Consequently, this involves FIR filters of several thousand taps for typical sampling frequencies.
5. Although undermodelling of the channel is possible, late reflections are important in dereverberation, in particular in the case when $D \geq D_c$.
6. The non-minimum phase property and the large spectral dynamic range of room transfer functions will raise challenges in designing AIR equalization filters.

1.6 Literature Overview

This section presents an overview of the existing literature, which deals explicitly with enhancement of reverberant speech with the aim to serve as an introduction to the topic and to provide an annotated bibliography. A more thorough treatment with additional bibliographic records of several methods mentioned here is provided in the relevant chapters.

1.6.1 Beamforming Using Microphone Arrays

Beamforming techniques are fundamentally important and among the first multi-channel processing approaches for enhancement of speech acquisition in noisy and reverberant environments [11]. The most direct and straightforward technique is the Delay-and-sum Beamformer (DSB) in which the microphone signals are delayed, to compensate for different times of arrival, and then weighted and summed [15, 89] as a convex combination. The output of the DSB can be written as

$$\bar{x}(n) = \sum_{m=1}^{M} w_m x_m(n - \tau_m), \tag{1.4}$$

where τ_m is the propagation delay in samples from the source to the m^{th} sensor and w_m is the weighting applied to the m^{th} sensor. In this way, the coherent components across channels, due to the direct-paths, are added constructively, while incoherent components, due to reverberation or noise, are attenuated [15].

The DSB can also be interpreted as forming a beam of sensitivity in the chosen direction. From this spatial filtering interpretation it can be seen intuitively that the beamformer approach works best for strongly localized sources and is less effective when the sound field is diffuse. The design of the weights is the spatial equivalent to the design of temporal FIR filters; the number of microphones is analogous to the number of taps and the spacing between sensors is analogous to the sampling frequency [15, 89]. Consequently, there is a spatial sampling criterion analogous to the time domain Nyquist sampling criterion, which relates the distance between microphones to the frequency components in the signal such that spatial aliasing can be avoided. This is defined as [15]

$$\|\mathbf{q}_{\text{mic},m} - \mathbf{q}_{\text{mic},m+1}\|_2 < \frac{c}{2f}, \tag{1.5}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\mathbf{q}_{\text{mic},m}$ is the three-dimensional position vector of the m^{th} microphone and c the speed of sound. Talantzis and Ward studied an alternative design of optimal weights in [85]. It can be seen from the expression in (1.5) that, for broadband signals such as speech, a linear array may not be the optimal solution. Consequently, several designs have been proposed with three or four subarrays and with different microphone spacing such that each of these subarrays covers a different bandwidth [11, 15].

Several variants of the DSB exist. For example, the DSB can be extended into the filter-and-sum beamformer in which the scalar weights are replaced each by an FIR filter [15]. Alternatively, in an approach employing frequency subbands by Allen *et al.* [4], the signals are co-phased in each frequency band and the gain is adjusted based on the cross-correlation between the channels to remove incoherent components before the summation. A two-dimensional microphone array was proposed by Flanagan *et al.* [19], which uses a DSB with a 'track-while-scan' approach where the area under consideration is quantized into overlapping regions that are scanned sequentially and speech characteristics are incorporated to distinguish a speech source from noise. The extension to three-dimensional arrays has also been considered [20] and also the use of spherical microphone arrays [58, 61]. Adaptive beamforming approaches have been studied, which automatically adjust the weights of the beamformer [15, 43] and which may also include constraints in the adaptation rule [22]. Generally, beamformers have been found to be efficient in applications to suppress localized additive noise sources [11]. Reverberation can be partially reduced, as will be shown in Chap. 2. However, since diffuse reverberant sound comes from all possible directions in a room [56], it will always enter the look-direction of the beam and hence will be only partially suppressed.

Improvements to beamforming applied in reverberant environments can be achieved using multiple beamformers where, instead of only forming a single beam in the direction of the desired source, a three-dimensional array can be used to form additional beams that are steered in the direction of the strong initial reflections [20, 70]. The additional reflections are treated as virtual sources in a similar way to the source-image method for simulation of room acoustics [3] described in Chap. 2. Another approach is the matched filter beamformer where the microphone signals are convolved with the time-reversed room impulse responses [1, 20, 32, 53, 54]. However, both these methods require at least partial knowledge of the room impulse response and can rather be treated as an alternative to inverse filtering.

1.6.2 Speech Enhancement Approaches to Dereverberation

An early technique in the class of speech enhancement dereverberation was proposed by Oppenheim and Schafer [72, 73]. The authors first introduce the observation that simple echoes are observed as distinct peaks in the cepstrum of the speech signal. Consequently, they use a peak picking algorithm to identify these peaks and attenuate them with, for example, a comb filter. An alternative to this was also considered, where a lowpass weighting function was applied to the cepstrum assuming that most of the energy of speech is in the lower quefrencies. However, this approach was not found suitable for more complex reverberation models [73].

A class of techniques emerged from the observation that the linear prediction residual signal contains the effects of reverberation, comprising peaks corresponding to excitation events in voiced speech together with additional peaks due to the reverberant channel [10, 98]. These techniques aim to suppress the effects of reverberation without degrading the original characteristics of the residual such that dereverberated speech can be synthesized using the processed residual and the all-pole filter resulting from prediction analysis of the reverberant speech. It is assumed in these methods that the effect of reverberation on the Autoregressive (AR) coefficients is insignificant [10]. It was shown in [29] that the AR coefficients of the clean speech can be estimated accurately from multichannel observations.

An early idea based on linear prediction processing was proposed in a patent by Allen [2] where the author suggested that synthetic clean speech could be generated from reverberant speech by identifying the Linear Predictive Coding (LPC) parameters from one or more reverberant observations. Griebel and Brandstein *et al.* [33, 34] used wavelet extrema clustering to reconstruct an enhanced prediction residual. In [35] the authors employ coarse room impulse response estimates and apply a matched filter type operation to obtain weighting functions for the reverberant residuals. Yegnanarayana *et al.* [97] used multichannel time-aligned Hilbert envelopes to represent the strength of the peaks in the prediction residuals. The Hilbert envelopes are then summed and the result used as a weight vector, which is applied to the prediction residual of one of the microphones. In [98] the authors derive a

weighting function based on the signal-to-reverberant ratio in different regions of the prediction residual. Gillespie *et al.* [31] demonstrate the kurtosis of the residual to be a useful reverberation metric, which they then maximize using an adaptive filter. This method was extended by Wu and DeLiang [94], who added a spectral subtraction stage to further suppress the remaining reverberation. Although these methods do attenuate the impulses due to reverberation in the prediction residual, they also significantly reduce naturalness in the dereverberated speech. This problem was ameliorated using a spatiotemporal averaging approach, where the speech signals are first spatially averaged and the prediction residual is further enhanced using temporal averaging of neighbouring larynx cycles [24, 27, 28, 86]. A further discussion on the processing of the linear prediction residual and the spatiotemporal averaging method will be given in Chap. 4.

A related method was proposed by Nakatani *et al.* [67]. This assumes a sinusoidal speech model. First the fundamental frequency of the speech signal is identified from the reverberant observations, then the remaining sinusoidal components are identified. Using the identified magnitude and phases of these sinusoids, an enhanced speech signal is synthesized. Subsequently, the reverberant and the dereverberated speech signals are used to derive an equivalent equalization filter. The processing is performed in short frames and the inverse filter is updated in each frame. It is shown that this inverse filter tends to the AIR equalization filter. However, this method may be computationally demanding [67].

Spectral subtraction has been widely applied, with some success, in noise reduction [6, 13]. Spectral subtraction was applied to dereverberation by Lebart *et al.* [57] and extended to the multichannel case by Habets [38, 39]. The authors assume a statistical model of the room impulse response comprising Gaussian noise modulated by a decaying exponential function. The decay rate of this exponential function is governed by the reverberation time. It is then shown that, if the reverberation time can be blindly estimated and in combination with multichannel spatial averaging, the power spectral density of the impulse response can be identified and subsequently removed by spectral subtraction. This method has shown promising results [38, 93], provided that the assumed unknowns are available, and will be elaborated in Chap. 3.

In summary, several speech enhancement approaches to dereverberation have appeared in the literature. These do not assume explicit knowledge of the room impulse response. However, blind identification of other features is often required. Nevertheless, many of these methods are computationally efficient and suitable for real-time implementation.

1.6.3 Blind System Identification and Inversion

The effects of reverberation can be removed if the AIR from the talker to at least one microphone can be identified and inverted so as to give a perfect equalizer for the acoustic channel. This approach presents several technical challenges that are

the subject of much current research. Significant progress has been made towards addressing these difficulties though, at the time of writing this, many issues related to algorithm design and implementation remain open.

1.6.3.1 Blind System Identification

Blind multichannel system identification using second order statistics is usually based on the cross-relation between two observations x_1 and x_2 and the corresponding two AIRs h_1 and h_2, where the time index is temporarily omitted for brevity. The cross-relation is given by [95]: $x_1 * h_2 = (s * h_1) * h_2 = x_2 * h_1$, which leads to the system of equations $\mathbf{Rh} = \mathbf{0}$, where in general for M channels $\mathbf{R}$ is a correlation-like matrix [50] and $\mathbf{h} = [\mathbf{h}_1^T \ \mathbf{h}_2^T \ \ldots \ \mathbf{h}_M^T]^T$ is a vector of the concatenated AIRs. It can be seen from this system of equations that the desired solution is the eigenvector corresponding to the zeroth eigenvalue in $\mathbf{R}$ or, in the presence of noise, the smallest eigenvalue. Several alternative solutions have been proposed. A Least Squares (LS) approach for solving this problem is given in [95]. An eigendecomposition method was proposed by Gürelli and Nikias [36]. Gannot and Moonen [23] use eigendecomposition methods for blind system identification both in the full-band and in frequency subbands. Huang and Benesty proposed the use of adaptive filters and derived multichannel LMS and Newton adaptive filters both in the time domain [49, 51, 52] and in the frequency domain [50].

This type of blind system identification requires that the following identifiability conditions are satisfied [95]:

1. The unknown channels must not include common zeros.
2. The correlation matrix of the source signal must be full rank.

Blind acoustic system identification algorithms additionally have to overcome the following challenges:

1. Acoustic channels are normally time-varying and therefore system identification must be performed adaptively.
2. AIRs have a duration typically corresponding to thousands of coefficients, and estimation of systems with such high order requires robust algorithms with high numerical precision and that typically present high computational requirements.
3. Noise in the observations can cause the adaptive algorithms to misconverge. Some approaches have been developed to improve robustness [25, 26, 42, 52];
4. Many approaches assume knowledge of the order of the unknown system. This issue has been addressed, for example, in [23] and [21];
5. Solutions for $\mathbf{h}$ are normally found only to within a scale factor [23, 52, 95].

Other approaches include Subramaniam's [84] proposed use of the cepstrum for blind system identification between two channels. It is shown that the channels can be reconstructed from their phases using an iterative approach, where the phases are identified from the cepstra of the observed data [75, 84] but that the method is

sensitive to zeros close to the unit circle – a situation which often arises in acoustic systems as was shown in [59]. A method introduced by Triki and Slock [88] comprises multichannel Linear Prediction (LP) to whiten the input signal and subsequent multichannel linear prediction which is used to identify the channels. A different approach to multichannel LP for dereverberation was taken in [14]. Recent developments of this class of methods will be discussed in more detail in Chap. 9. Finally, in [48] it is proposed to use an autoregressive model of channel impulse response, which is assumed to be stationary, in contrast to the FIR model employed in all the above methods. Furthermore, it is assumed that the source signal is a locally stationary AR process but that it is globally nonstationary. In this way, the parameters of the all-pole channel filter can be identified by observing several frames of the input signal and collecting information regarding the poles either by using a histogram approach or a more robust Bayesian probabilistic framework. Over several frames, the poles due to the stationary channel become apparent and the channel can thus be identified. One major advantage of this method is that, by using an AR model of the channel, the order of the channel is reduced compared to the FIR channel models. Further extensions based on this idea have been developed in [16–18]. This approach will be discussed in more depth in Chap. 8. Nevertheless, problems of sensitivity to noise and channel order estimation are common to all approaches and the subject of much current research, which will be discussed in Chaps. 5, 6 and 8.

1.6.3.2 Inverse Filtering

If the acoustic impulse responses from the talker to the microphones, $\mathbf{h}_m(n)$, are available, for example, from a blind system identification algorithm, dereverberation can be achieved in principle by an inverse system, $\mathbf{g}_m$, satisfying $\mathbf{h}_m^T(n)\mathbf{g}_m = \kappa\delta(n-\tau)$, where κ and τ are, respectively, arbitrary scale and delay factors. However, direct inversion of an acoustic channel presents several significant technical challenges.

1. AIRs have duration typically corresponding to thousands of coefficients and inversion of systems with such high order requires robust algorithms with high numerical precision and that typically present high computational requirements.
2. Acoustic channels typically exhibit non-minimum phase characteristics [68].
3. Acoustic channels may contain spectral nulls, which after inversion give strong peaks in the spectrum causing narrow band noise amplification.

Several alternative approaches have been studied for single channel inversion. For example, single channel LS inverse filters can be designed by minimizing the error $\hat{\mathbf{g}}_m = \min_{\mathbf{g}_m} \|\mathbf{h}_m^T(n)\mathbf{g}_m - \delta(n-\tau)\|_2^2$ [64, 66]. Homomorphic inverse filtering has also been investigated [3, 64, 78, 87], where the impulse response is decomposed into a minimum phase component, $\mathbf{h}_{\mathrm{mp},m}(n)$ and an all-pass component, $\mathbf{h}_{\mathrm{ap},m}(n)$, such that $\mathbf{h}_m(n) = \mathbf{h}_{\mathrm{ap},m}^T(n)\mathbf{h}_{\mathrm{mp},m}(n)$. Consequently, magnitude and phase are equalized separately, where an exact inverse can be found for the magnitude, while the

phase can be equalized, e.g., using matched filtering [55, 78]. An important result is that equalization of only the magnitude results in audible distortion [68, 78].

In the multichannel case, an exact inverse can be found by application of multichannel least squares design [51, 62]. The Multiple-input/output INverse Theorem (MINT) approach was the first such multichannel inversion method proposed by Miyoshi and Kaneda [62], which was also implemented in a subband version [96]. Adaptive versions have also been considered in [69]. If there are no common zeros between the two channel transfer functions, a pair of inverse filters, $\mathbf{g}_1$ and $\mathbf{g}_2$ can be found such that: $\mathbf{h}_1^T(n)\mathbf{g}_1 + \mathbf{h}_2^T(n)\mathbf{g}_2 = \delta(n)$. Thus, exact inverse filtering can be performed, with inverse filters of length similar to the channel length [51, 62]. Undermodelled estimates of $\mathbf{h}_m(n)$ are problematic for this type of inversion, and it has been observed that true channel inverses are of limited value for practical dereverberation when the channel estimate contains even moderate estimation errors. Regularized multichannel equalization was shown to increase the equalization robustness to noise and estimation errors [44–46, 99]. Acoustic channel equalization will be discussed in Chaps. 7 and 9.

1.7 Outline of the Book

The remainder of this book is organized as follows:

Chapter 2 reviews the acoustic characteristics of typical rooms and discusses measurement and simulation of acoustic impulse responses. Furthermore, subjective and objective measures of reverberation in speech are discussed.

Chapter 3 introduces a statistical model of the room impulse response and uses that to develop a multichannel spectral subtraction based algorithm for speech dereverberation.

Chapter 4 reviews the use of processing of the linear prediction residual for dereverberation of speech. A spatiotemporal averaging method for linear prediction residual processing is introduced and its application to speech dereverberation is demonstrated.

Chapter 5 develops a multichannel eigendecomposition method for blind identification of room impulse responses in the presence of coloured noise. The identified impulse responses are then used to design equalization filters for speech dereverberation.

Chapter 6 introduces a class of adaptive blind system identification methods with implementations both in the time and frequency domain. The adverse effects of noise on these algorithms are explored and several approaches to added noise robustness are presented and discussed.

Chapter 7 presents a multichannel Acoustic Transfer Function (ATF) equalizer design framework using oversampled and decimated subbands. The method is shown to allow for approximate equalization of long, non-minimum phase ATFs at low computational cost.

Chapter 8 considers blind dereverberation in time-varying acoustic environments. The source and the AIRs are presented by parametric models that are employed in combination with Bayesian inference to estimate the room acoustic parameters.

Chapter 9 uses multichannel linear prediction to derive an equalization filter without necessarily estimating the acoustic impulse responses first. This equalization filter results in excessive whitening of the speech and, consequently, four different methods to overcome this problem are presented.

Chapter 10 presents TRINICON – a generic framework for Multi-Input Multi-Output (MIMO) signal processing. It is applied here to derive two dereverberation algorithms: one where the AIRs are first identified blindly and used to design equalization filters, and the second where the equalization filters are identified directly from the reverberant observations.

References

1. Affes, S., Grenier, Y.: A signal subspace tracking algorithm for microphone array processing of speech. IEEE Trans. Speech Audio Process. **5**(5), 425–437 (1997)
2. Allen, J.B.: Synthesis of pure speech from a reverberant signal. U.S. Patent No. 3786188 (1974)
3. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
4. Allen, J.B., Berkley, D.A., Blauert, J.: Multimicrophone signal-processing technique to remove room reverberation from speech signals. J. Acoust. Soc. Am. **62**(4), 912–915 (1977)
5. BBC, U.: Iceland comes first in broadband. [Online] (2006). URL `http://news.bbc.co.uk/1/hi/technology/4903776.stm`
6. Benesty, J., Makino, S., Chen, J. (eds.): Speech enhancement. Springer (2005)
7. Benesty, J., Sondhi, M.M., Huang, Y. (eds.): Springer handbook of speech processing. Springer (2007)
8. Bolt, R.H., MacDonald, A.D.: Theory of speech masking by reverberation. J. Acoust. Soc. Am. **21**(6), 577–580 (1949)
9. Bradley, J.S., Sato, H., Picard, M.: On the importance of early reflections for speech in rooms. J. Acoust. Soc. Am. **113**(6), 3233–3244 (2003)
10. Brandstein, M.S., Griebel, S.M.: Nonlinear, model-based microphone array speech enhancement. In: S.L. Gay, J. Benesty (eds.) Acoustic Signal Processing For Telecommunication, pp. 261–279. Kluwer Academic Publishers (2000)
11. Brandstein, M.S., Ward, D.B. (eds.): Microphone arrays: Signal processing techniques and applications, 1 edn. Springer (2001)
12. Cherry, C.: On human communications, third edn. MIT Press (1980)
13. Davis, G.M. (ed.): Noise reduction in speech applications. CRC Press (2002)
14. Delcroix, M., Hikichi, T., Miyoshi, M.: Precise dereverberation using multichannel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15**(2), 430–440 (2007)
15. Elko, G.W.: Microphone array systems for hands-free telecommunication. Speech Communication **20**(3-4), 229–240 (1996)
16. Evers, C., Hopgood, J.R.: Parametric modelling for single-channel blind dereverberation of speech from a moving speaker. IET Communications **2**, 59–74 (2008)
17. Evers, C., Hopgood, J.R., Bell, J.: Acoustic models for blind source dereverberation using sequential Monte Carlo methods. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2008)

18. Evers, C., Hopgood, J.R., Bell, J.: Blind speech dereverberation using batch and sequential Monte Carlo methods. In: Proc. Int. Symp. on Circuits and Systems (2008)
19. Flanagan, J.L., Johnston, J.D., Zahn, R., Elko, G.W.: Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. **78**(5), 1508–1518 (1985)
20. Flanagan, J.L., Surendran, A.C., Jan, E.E.: Spatially selective sound capture for speech and audio processing. Speech Communication **13**(1-2), 207–222 (1993)
21. Furuya, K., Kaneda, Y.: Two-channel blind deconvolution for non-minimum phase impulse responses. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1315–1318 (1997)
22. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
23. Gannot, S., Moonen, M.: Subspace methods for multi-microphone speech dereverberation. EURASIP J. on App. Signal Process. **2003**(11), 1074–1090 (2003)
24. Gaubitch, N.D.: Blind identification of acoustic systems and enhancement of reverberant speech. Ph.D. thesis, Imperial College London (2007)
25. Gaubitch, N.D., Hasan, M.K., Naylor, P.A.: Generalized optimal step-size for blind multichannel LMS system identification. IEEE Signal Process. Lett. **13**(10), 624–627 (2006)
26. Gaubitch, N.D., Hasan, M.K., Naylor, P.A.: Noise robust adaptive blind identification using spectral constraints. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. V–93–V–96. Toulouse, France (2006)
27. Gaubitch, N.D., Naylor, P.A.: Spatiotemporal averaging method for enhancement of reverberant speech. In: Proc. IEEE Int. Conf. Digital Signal Processing (DSP). Cardiff, UK (2007). DOI 10.1109/ICDSP.2007.4288655
28. Gaubitch, N.D., Naylor, P.A., Ward, D.B.: Multi-microphone speech dereverberation using spatio-temporal averaging. In: Proc. European Signal Processing Conf. (EUSIPCO), pp. 809–812. Vienna, Austria (2004)
29. Gaubitch, N.D., Ward, D.B., Naylor, P.A.: Statistical analysis of the autoregressive modeling of reverberant speech. J. Acoust. Soc. Am. **120**(6), 4031–4039 (2006)
30. Gay, S.L., Benesty, J. (eds.): Acoustic signal processing for telecommunication. Kluwer Academic Publishers (2000)
31. Gillespie, B.W., Malvar, H.S., Florêncio, D.A.F.: Speech dereverberation via maximum-kurtosis subband adaptive filtering. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 6, pp. 3701–3704 (2001)
32. Grenier, Y., Affes, S.: Microphone array response to speaker movements. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 247–250 (1997)
33. Griebel, S.M.: A microphone array system for speech source localization, denoising and dereverberation. Ph.D. thesis, Harvard University, Cambridge, Massachusetts (2002)
34. Griebel, S.M., Brandstein, M.S.: Wavelet transform extrema clustering for multi-channel speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Pocono Manor, Pennsylvania (1999)
35. Griebel, S.M., Brandstein, M.S.: Microphone array speech dereverberation using coarse channel estimation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 201–204 (2001)
36. Gürelli, L., Nikias, C.L.: EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. IEEE Trans. Signal Process. **43**(1), 143–149 (1995)
37. Haas, H.: The influence of a single echo on the audibility of speech. J. Audio Eng. Soc. **20**, 145–159 (1972)
38. Habets, E.A.P.: Multi-channel speech dereverberation based on a statistical model of late reverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. iv/173–iv/176. Philadelphia (2005)
39. Habets, E.A.P.: Single- and multi-microphone speech dereverberation using spectral enhancement. Ph.D. thesis, Technische Universiteit Eindhoven (2007). URL `http://alexandria.tue.nl/extra2/200710970.pdf`
40. Haneda, Y., Makino, S., Kaneda, Y.: Common acoustical pole and zero modeling of room transfer functions. IEEE Trans. Speech Audio Process. **2**(2), 320–328 (1994)

41. Hänsler, E., Schmidt, G. (eds.): Topics in acoustic echo and noise control. Springer (2006)
42. Hasan, M.K., Benesty, J., Naylor, P.A., Ward, D.B.: Improving robustness of blind adaptive multichannel identification algorithms using constraints. In: Proc. European Signal Processing Conf. (EUSIPCO). Antalya, Turkey (2005)
43. Haykin, S.: Adaptive filter theory, 4 edn. Prentice Hall, Upper Saddle River, N.J. (2001)
44. Hikichi, T., Delcroix, M., Miyoshi, M.: Inverse filtering for speech dereverberation less sensitive to noise. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4 (2006)
45. Hikichi, T., Delcroix, M., Miyoshi, M.: On robust inverse filter design for room transfer function fluctuations. In: Proc. European Signal Processing Conf. (EUSIPCO) (2006)
46. Hikichi, T., Delcroix, M., Miyoshi, M.: Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. EURASIP J. Advances in Signal Processing **2007**, 1–12 (2007)
47. Hopgood, J.R., Rayner, P.J.W.: A probabilistic framework for subband autoregressive models applied to room acoustics. In: Proc. IEEE Workshop Statistical Signal Processing, pp. 492–495 (2001)
48. Hopgood, J.R., Rayner, P.J.W.: Blind single channel deconvolution using nonstationary signal processing. IEEE Trans. Speech Audio Process. **11**(5), 476–488 (2003)
49. Huang, Y., Benesty, J.: Adaptive multi-channel least mean square and Newton algorithms for blind channel identification. Signal Processing **82**(8), 1127–1138 (2002)
50. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. **51**(1), 11–24 (2003)
51. Huang, Y., Benesty, J., Chen, J.: A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. IEEE Trans. Speech Audio Process. **13**(5), 882–895 (2005)
52. Huang, Y., Benesty, J., Chen, J.: Optimal step size of the adaptive multichannel LMS algorithm for blind SIMO identification. IEEE Signal Process. Lett. **12**(3), 173–176 (2005)
53. Jan, E., Flanagan, J.L.: Microphone arrays for speech processing. In: Int. Symposium on Signals, Systems, and Electronics, pp. 373–376 (1995)
54. Jan, E., Svazier, P., Flanagan, J.L.: Matched-filter processing of microphone array for spatial volume selectivity. In: Proc. Int. Symp. on Circuits and Systems, vol. 2, pp. 1460–1463 (1995)
55. Kennedy, R.A., Radlović, B.D.: Iterative cepstrum-based approach for speech dereverberation. In: Proc. Int. Symposium on Signal Processing and Its Applications (ISSPA), vol. 1, pp. 55–58 (1999)
56. Kuttruff, H.: Room acoustics, 4 edn. Taylor & Francis (2000)
57. Lebart, K., Boucher, J.M., Denbigh, P.N.: A new method based on spectral subtraction for speech dereverberation. Acta Acoustica **87**, 359–366 (2001)
58. Li, Z., Duraiswami, R.: Flexible and optimal design of spherical microphone arrays for beamforming. IEEE Trans. Audio, Speech, Lang. Process. **15**(2), 702–714 (2007)
59. Lin, X., Gaubitch, N.D., Naylor, P.A.: Two-stage blind identification of SIMO systems with common zeros. In: Proc. European Signal Processing Conf. (EUSIPCO). Florence, Italy (2006)
60. Loizou, P.C.: Speech enhancement theory and practice. Taylor & Francis (2007)
61. Meyer, J., Agnello, T.: Spherical microphone array for spatial sound recording. In: Audio Engineering Society, 115th Convention, preprint 5975. New York (2003)
62. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
63. Mobile Operators Association: History of cellular mobile communications. [Online] (2005). URL `http://www.mobilemastinfo.com/information/history.htm`
64. Mourjopoulos, J., Clarkson, P., Hammond, J.: A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 7, pp. 1858–1861 (1982)
65. Mourjopoulos, J., Paraskevas, M.A.: Pole and zero modeling of room transfer functions. J. Sound Vib. **146**(2), 281–302 (1991)
66. Mourjopoulos, J.N.: Digital equalization of room acoustics. J. Audio Eng. Soc. **42**(11), 884–900 (1994)

67. Nakatani, T., Miyoshi, M., Kinoshita, K.: Single-microphone blind dereverberation. In: J. Benesty, S. Makino, J. Chen (eds.) Speech Enhancement, 1 edn. Springer Verlag (2005)
68. Neely, S.T., Allen, J.B.: Invertibility of a room impulse response. J. Acoust. Soc. Am. **66**(1), 165–169 (1979)
69. Nelson, P.A., Orduña-Brustamante, F., Hamada, H.: Inverse filter design and equalization zones in multichannel sound reproduction. IEEE Trans. Speech Audio Process. **3**(3), 185–192 (1995)
70. Nishiura, T., Nakanura, S., Shikano, K.: Speech enhancement by multiple beamforming with reflection signal equalization. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 189–192 (2001)
71. Omologo, M., Svazier, P., Matassoni, M.: Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Communication **25**(1), 75–95 (1998)
72. Oppenheim, A.V., Schafer, R.W.: Digital signal processing, 1 edn. Prentice Hall (1975)
73. Oppenheim, A.V., Schafer, R.W., Stockham, T.G.: Nonlinear filtering of multiplied and convolved signals. IEEE Trans. Audio Electroacoust. **AU-16**(3), 437–466 (1968)
74. Paatero, T.: Modeling of long and complex responses using Kautz filters and time-domain partitions. In: Proc. European Signal Processing Conf. (EUSIPCO), pp. 313–316. Vienna, Austria (2004)
75. Petropulu, A.P., Nikias, C.L.: Blind deconvolution using signal reconstruction from partial higher order cepstral information. IEEE Trans. Signal Process. **41**(6), 2088–2095 (1993)
76. Plato: The republic. Penguin Books Ltd (2003)
77. Polycom: Polycom communicator. [Online] (2006). URL `http://www.polycom.com/`
78. Radlović, B.D., Kennedy, R.A.: Nonminimum-phase equalization and its subjective importance in room acoustics. IEEE Trans. Speech Audio Process. **8**(6), 728–737 (2000)
79. Radlović, B.D., Williamson, R.C., Kennedy, R.A.: Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Acoust., Speech, Signal Process. **8**(3), 311–319 (2000)
80. Rayleigh, J.W.S.: The theory of sound. Dover Publications (1976)
81. Sabine, W.C.: Collected papers on acoustics. Dover Publications (1964)
82. Schmidt, G.: Applications of acoustic echo control – an overview. In: Proc. European Signal Processing Conf. (EUSIPCO), pp. 9–16. Vienna, Austria (2004)
83. Schroeder, M.R.: Statistical parameters of the frequency response curves of large rooms. J. Audio Eng. Soc. **35**(5), 299–305 (1987)
84. Subramaniam, S., Petropulu, A.P., Wendt, C.: Cepstrum-based deconvolution for speech dereverberation. IEEE Trans. Acoust., Speech, Signal Process. **4**(5), 392–396 (1996)
85. Talantzis, F., Ward, D.B.: Robustness of multi-channel equalization in an acoustic reverberant environment. J. Acoust. Soc. Am. **114**(2), 833–841 (2003)
86. Thomas, M.R.P., Gaubitch, N.D., Gudnason, J., Naylor, P.A.: A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY (2007)
87. Tohyama, M., Lyon, R.H., Koike, T.: Source waveform recovery in a reverberant space by cepstrum dereverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 157–160 (1993)
88. Triki, M., Slock, D.T.M.: Delay-and-predict equalization for blind speech dereverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toulouse, France (2006)
89. VanVeen, B.D., Buckley, K.M.: Beamforming: a versatile approach to spatial filtering. IEEE Signal Process. Mag. **5**(2), 4–24 (1988)
90. Waller, S.J.: Sound and rock art. Nature **363** (1993)
91. Waller, S.J.: Psychoacoustic influences of the echoing environments of prehistoric art. J. Acoust. Soc. Am. **112** (2002)
92. Ward, D.B.: On the performance of acoustic crosstalk cancellation in a reverberant environment. J. Acoust. Soc. Am. **110**(2), 1195–1198 (2001)

93. Wen, J.Y.C., Gaubitch, N.D., Habets, E.A.P., Myatt, T., Naylor, P.A.: Evaluation of speech dereverberation algorithms using the MARDY database. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Paris, France (2006)
94. Wu, M., Wang, D.: A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. **14**(3), 774–784 (2006)
95. Xu, G., Liu, H., Tong, L., Kailath, T.: A least-squares approach to blind channel identification. IEEE Trans. Signal Process. **43**(12), 2982–2993 (1995)
96. Yamada, K., Wang, J., Itakura, F.: Recovering of broad band reverberant speech signal by subband MINT method. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 969–972 (1991)
97. Yegnanarayana, B., Prasanna, S.R.M., Rao, K.S.: Speech enhancement using excitation source information. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 541–544 (2002)
98. Yegnanarayana, B., Satyanarayana, P.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Acoust., Speech, Signal Process. **8**(3), 267–281 (2000)
99. Zhang, W., Gaubitch, N.D., Naylor, P.A.: Computationally efficient equalization of room impulse responses robust to system estimation errors. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2008)

Chapter 2
Models, Measurement and Evaluation

Patrick A. Naylor, Emanuël A.P. Habets, Jimi Y.-C. Wen, and Nikolay D. Gaubitch

Abstract It is the science of room acoustics that offers an understanding of the physical processes by which sound waves propagate in enclosed spaces and the manner in which acoustic reflections combine to give the effect that we refer to as reverberation. This chapter aims to summarize some of the main concepts of room acoustics that are relevant to the subsequent material. In particular, this discussion will focus on models and simulation techniques for room acoustics that aid the description of reverberation and the development of dereverberation algorithms. Examples will be given involving both simulated room impulse responses and measured responses of real rooms. The issue of evaluation of dereverberation processing will then be addressed. Measures that aim to characterize the quantity and perceived effect of reverberation in a speech signal will be described and discussed. The chapter ends by considering the well known delay-and-sum beamformer, which is often considered to be a baseline spatial filtering approach, and presents an analysis of the dereverberation performance levels that can be expected at such a baseline.

2.1 An Overview of Room Acoustics

Consider a single omnidirectional source of sound located within an enclosed space such as an office or living room with walls and other surfaces that reflect sound to some extent. Let us assume that the source starts to emit at some instant in time $t = t_0$ and that the room was silent for $t < t_0$. The sound emanating from the source will be reflected multiple times in a manner that depends on the geometry of the source and the room as well as the nature of the reflective surfaces. This process produces a sound energy distribution that becomes increasingly uniform with time $t > t_0$ across a wide range of frequencies of interest.

Imperial College London, UK

2.1.1 The Wave Equation

Let us begin by considering a sound field as a superposition of plane waves. The propagation of such waves within a room can be considered to be a linear process after applying several simplifications including the assumptions that the medium in which the waves travel is homogeneous, at rest, and that its characteristics are independent of the wave amplitude. Then the propagation of acoustic waves through a material can be described by the second order partial differential wave equation. The wave equation describes the evolution of sound pressure $p(\mathbf{q},t)$, without any driving source, as a function of position $\mathbf{q} = (q_x, q_y, q_z)$ and time t and is given by

$$\nabla^2 p(\mathbf{q},t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{q},t)}{\partial t^2} = 0, \tag{2.1}$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.2}$$

The wave equation accurately describes the pressure in a realistic sound field provided that the wave amplitude is small such that $|p(\mathbf{q},t)| \ll \rho_0 c^2$ where ρ_0 is the density of the propagation medium at equilibrium and c is the propagation speed. In practice, two types of inhomogeneities occur [33]. The medium may exhibit scalar inhomogeneities giving rise to a spatial distribution of sound speed and density, for example, due to temperature variations in the medium. The medium may also exhibit vector inhomogeneities giving rise to a spatial distribution of medium velocity, for example, due to the presence of fans or air conditioning. However, the effects of these inhomogeneities are usually sufficiently small so that they can be ignored in many practical situations.

The Fourier transform of sound pressure, $p(\mathbf{q},t)$, is given by

$$P(\mathbf{q},\omega) = \int_{-\infty}^{\infty} p(\mathbf{q},t) e^{-j\omega t} \mathrm{dt}. \tag{2.3}$$

Therefore, the wave equation can be expressed in the frequency domain by taking the Fourier transform of (2.1) to give the Helmholtz equation

$$\nabla^2 P(\mathbf{q},\omega) + k^2 P(\mathbf{q},\omega) = 0, \tag{2.4}$$

where

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda} \tag{2.5}$$

is the wavenumber, ω is the angular frequency and λ is the wavelength.

2.1.2 Sound Field in a Reverberant Room

When sound is produced in a room or other reverberant environment, a listener will hear a mixture of direct sound and reverberant sound. The direct-path component is the sound that travels from the source to the listener without reflection whereas the reverberant component is the sound that travels from the source to the listener via one or more reflections. The effect of increasing the distance between the sound source and the listening location is to reduce the energy of the direct-path component. The energy of the reverberant sound is not in general affected by the source-listener distance but instead is dependent on the acoustic properties of the room.

For a single sound source in a room, the resulting sound pressure at a point $\mathbf{q} = (q_x, q_y, q_z)$ and frequency ω can be written as the sum of two components [25, 40]

$$P(\mathbf{q}, \omega) = P_{\mathrm{d}}(\mathbf{q}, \omega) + P_{\mathrm{r}}(\mathbf{q}, \omega), \tag{2.6}$$

where subscripts d and r indicate direct and reverberant components respectively.

The sound energy density, defined as the sound energy per unit volume, due to the direct-path component is then given by

$$E_{\mathrm{d}} = \frac{\mathscr{E}\{P_{\mathrm{d}}(\mathbf{q}, \omega) P_{\mathrm{d}}^*(\mathbf{q}, \omega)\}}{\rho_0 c^2} = \frac{Q W_s}{4\pi c D^2}, \tag{2.7}$$

where W_s is the power output from the sound source in watts, D is the distance from the source and Q describes the directivity of the source such that $Q = 1$ for an omnidirectional source. The spatial expectation operator, $\mathscr{E}\{\cdot\}$, is discussed in more detail in Sect. 2.2.6 and here indicates the expected value over spatial locations spanned by $\mathbf{q}$.

Similarly, the sound energy density due to the reverberant component is given by

$$E_{\mathrm{r}} = \frac{\mathscr{E}\{P_{\mathrm{r}}(\mathbf{q}, \omega) P_{\mathrm{r}}^*(\mathbf{q}, \omega)\}}{\rho_0 c^2} = \frac{4 W_s}{c R}, \tag{2.8}$$

with the room constant, R, given by

$$R = \frac{\bar{\alpha} A}{1 - \bar{\alpha}}, \tag{2.9}$$

where $\bar{\alpha}$ and A denote the average absorption coefficient of the surfaces in the room and the total absorption surface area, respectively.

It can be seen, therefore, that the energy density of the reverberant sound is independent of the distance D, whilst the direct sound energy density is related to D by an inverse square law.

2.1.3 Reverberation Time

A widely used and important characteristic of an acoustic space is the reverberation time. The reverberation time can be measured by exciting a room with a broadband signal until a steady state uniform sound energy distribution is obtained, then switching off the sound source and recording the resulting decay of the squared sound pressure against time. This is known as the Energy Decay Curve (EDC). The reverberation time, T_{60}, is defined for a diffuse sound field as the time in seconds required for the EDC to decay by 60 dB.

This concept originates from the early work of Sabine [25] who determined that the reverberation time was proportional to the volume of the room, V, and inversely proportional to the amount of absorption in the room. Sabine's method [42] estimates the reverberation time, neglecting the effect of attenuation due to propagation through the air, as

$$T_{60} = \frac{24\ln(10)}{c}\frac{V}{\alpha_{\text{Sabine}}A} \quad \text{s.} \tag{2.10}$$

In this expression, $\alpha_{\text{Sabine}}A$ represents the total absorption and is, in the field of architectural acoustics, formed from the sum of products of Sabine's sound absorption coefficients and their corresponding areas. For example, in a concert hall, different absorption coefficients are used for regions of the hall such as audience seating, balconies or other sound reflecting surfaces. Alternatively, the absorption may be calculated from an average absorption coefficient $\bar{\alpha}$ with the total corresponding reflecting surface area.

The reverberation time is alternatively given by Eyring's reverberation formula [25] as

$$T_{60} = -\frac{24\ln(10)}{c}\frac{V}{\ln(1-\alpha_{\text{Eyring}})A} \quad \text{s,} \tag{2.11}$$

where α_{Eyring} is the Eyring sound absorption coefficient. As in the Sabine case, the denominator has to take into account the various region of the hall by applying appropriate absorption coefficients over the corresponding surface areas for each of the regions, and combine them taking into account the natural logarithm function. The Eyring reverberation time may also be calculated from an average absorption coefficient $\bar{\alpha}$ and a total corresponding reflecting surface area. The Eyring absorption coefficients can be derived from the Sabine coefficients as given in [3]. This same article also gives some fascinating historical insights into the original of Eyring's formula.

When $\bar{\alpha}$ is small, the expansion

$$-\ln(1-\bar{\alpha}) = \bar{\alpha} + \frac{\bar{\alpha}^2}{2} + \frac{\bar{\alpha}^3}{3} + \ldots \tag{2.12}$$

shows that Eyring's and Sabine's reverberation times become approximately equal. Furthermore, the reverberation time for a given room is seen from these expres-

sions to be independent of the position within the room of the sound source and the measurement location.

If the Acoustic Impulse Response (AIR) of the room, $h(t)$, is known, the EDC can be obtained from the Schroeder integral [25]

$$\mathrm{EDC}(t) = \int_t^\infty h^2(\tau)\mathrm{d}\tau. \tag{2.13}$$

An example is given in Fig. 2.1, which shows the EDC for a measured impulse response in the MARDY database [47]. The dB scale is referred to $\mathrm{EDC}(0)$. Four regions of this plot can now be identified.

1. Close to the time origin, the plot is approximately constant and close in value to the maximum – the reference level of 0 dB. This occurs because the measured response includes the direct-path propagation delay that is manifested in the leading samples of the response containing only noise.
2. Shortly after the time origin, the plot shows a sharp drop in energy, which corresponds to the transition from the region of the direct-path component and early reflections to the region of free decay for which the sound energy in the room is diffuse.
3. Between about 50 and 300 ms the plot shows a slope with near-constant negative gradient. This region corresponds to the free decay.
4. After 350 ms, the plot can be seen to begin to flatten. This occurs when the EDC has decayed sufficiently so that its energy approaches the energy of measurement noise, which, in the case of this measurement, is approximately -48 dB.

In order to determine the T_{60} from an EDC plot, the impulse response should be measured at a distance greater than the critical distance (see Sect. 2.1.4). This is so that any effects due to the direct-path component are ignored since these are dependent on the geometry of the source and microphone in the room – factors from which T_{60} is independent. Estimation of T_{60} should also be made from measurements at levels greater than the measurement noise floor in order to avoid the effects of such noise. Accordingly, taking these factors into account, useful estimates of T_{60} can be obtained from EDC plots such as Fig. 2.1 by measuring the slope of only the free decay section, being the part that has near constant gradient. This can be found to be -89.4 dB per second in this case, which, for a decay of 60 dB, corresponds to $T_{60} = 0.67$ s.

To give additional insight, the impulse response can be split into frequency subbands and the EDC computed in each subband to give the Energy Decay Relief (EDR) as a function of both time and frequency, $\mathrm{EDR}(t, f)$ [21, 25]. This is typically presented as a 2-D surface plot and enables the frequency dependence of reverberation time to be studied.

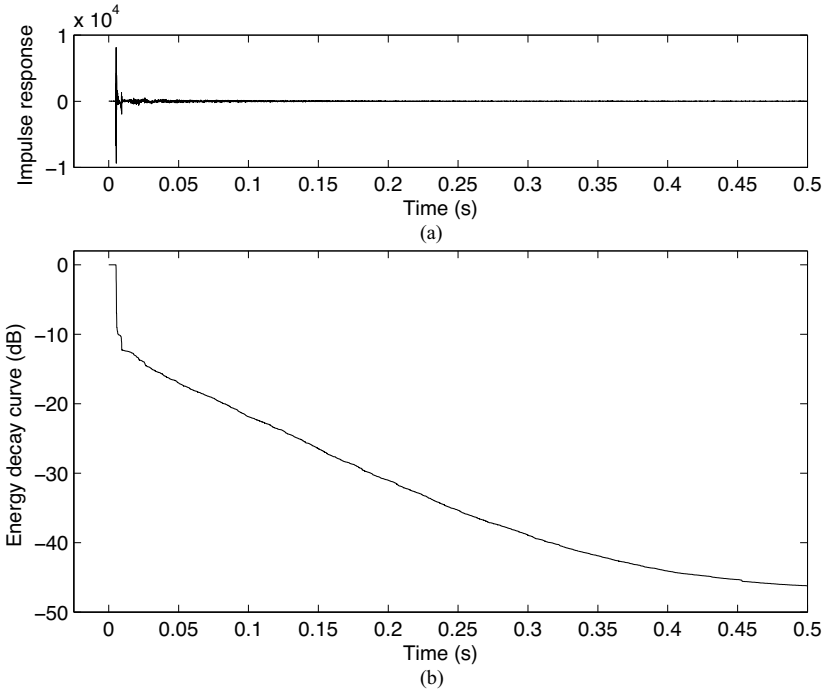


Fig. 2.1 Example (a) room impulse response and (b) its corresponding energy decay curve

2.1.4 The Critical Distance

The critical distance is defined as the distance D_c from the source at which the sound energy density due to the direct-path component, E_d, and the sound energy density due to the reverberant component, E_r, are equal. It is evaluated by equating (2.7) and (2.8) to give

$$\frac{Q}{4\pi D_c^2} = \frac{4}{R}, \tag{2.14}$$

so that

$$D_c = \sqrt{\frac{QR}{16\pi}} \text{ m}. \tag{2.15}$$

As shown in [25], the critical distance can also be expressed in terms of Q, V and the reverberation time T_{60} as

$$D_c \approx 0.1\sqrt{\frac{QV}{\pi T_{60}}} \text{ m}. \tag{2.16}$$

An example of sound energy density in a room as a function of the distance from the source is shown in Fig. 2.2 for a 1 watt omnidirectional source in a room of

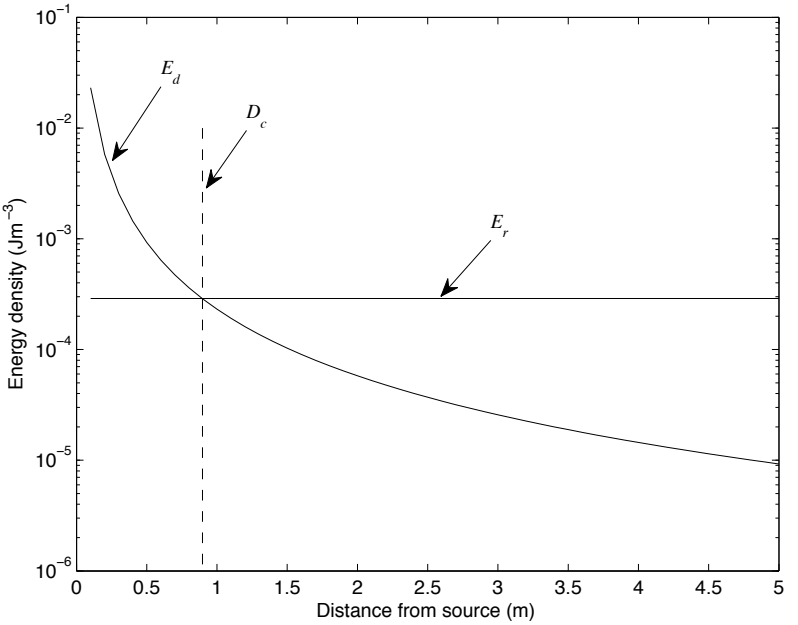


Fig. 2.2 Direct energy density, E_d and reverberant energy density, E_r against distance from a 1 watt source in a room of dimensions $3 \times 4 \times 5$ m with $\bar{\alpha} = 0.3$ ($T_{60} \approx 0.29$ from (2.11) with $\alpha_{\mathrm{Eyring}} = \bar{\alpha}$) and $c = 344$ m/s. The *vertical dashed line* indicates the critical distance, $D_\mathrm{c} \approx 0.9$ m, computed using the approximate formula in (2.16)

dimensions $3 \times 4 \times 5$ m with $\bar{\alpha} = 0.3$ (giving $T_{60} \approx 0.29$ from (2.11) with $\alpha_{\mathrm{Eyring}} = \bar{\alpha}$) and $c = 344$ m/s. The critical distance corresponds to the intersection of E_d and E_r. The vertical dashed line marks the critical distance, where in this case $D_\mathrm{c} \approx 0.9$ m, computed using the approximate formula in (2.16).

2.1.5 Analysis of Room Acoustics Dependent on Frequency Range

There are several different techniques for studying room acoustics [25] and, in general, each technique is only applicable to a limited range of the audible spectrum; no single analytical or numerical tool can currently model the entire audible spectrum from 20 Hz to close to 20 kHz. The audible spectrum can be conveniently divided into the following four regions (i-iv), defined in terms of the speed of sound, c, and the largest dimension of the room, $L_{\max}$.

(i) At very low frequencies for which $f < c/2L_{\max}$, there is no resonant support for the sound in the room. This frequency range can be analyzed using non-harmonic solutions to the wave equation. As an example, a room with dimensions $3 \times 5 \times 7$ m, and a sound velocity of 344 m/s has no resonant frequencies below 25 Hz.

(ii) The next frequency range, $f \sim c/L_{\max}$, is that for which the wavelength of the sound source is comparable to the dimensions of the room and spans from the lowest resonant mode to the Schroeder frequency. The Schroeder frequency plays an important role in room acoustics [26]. It conceptually separates the range of frequencies for which distinct resonances can be observed from the range for which the resonances are too close in frequency to be distinct since, at any observation frequency in this range, the effects of several neighbouring resonances are superimposed. More specifically, consider a room of volume V with resonances having an average 3 dB bandwidth of $\langle \Delta f \rangle$. The Schroeder frequency is the value of f for which the resonant frequencies of the room are separated such that the three resonant frequencies lie with one resonant bandwidth and is given [25] by solving for f in

$$\langle \Delta f \rangle = 3\frac{c^3}{4\pi V f^2}. \tag{2.17}$$

The resulting Schroeder frequency can be written as

$$f_{\mathrm{g}} \approx \frac{5500}{\sqrt{V\bar{\zeta}}}\ \mathrm{Hz}, \tag{2.18}$$

where $\bar{\zeta}$ is the average value of the damping constants associated with each resonant frequency of the room. A further approximate expression for the Schroeder frequency can be written as

$$f_{\mathrm{g}} \approx 2000\sqrt{\frac{T_{60}}{V}}\ \mathrm{Hz}, \tag{2.19}$$

using the commonly quoted approximation relating the average damping constant to the reverberation time

$$T_{60} = \frac{3\ln(10)}{\bar{\zeta}}. \tag{2.20}$$

In this frequency range, wave acoustics are applicable for describing the acoustical properties of a room. Wave acoustics assume a harmonic sound source and are based on solutions of the wave equation. For the same example room having a reverberation time of 0.5 s, this frequency range spans from 25 to 138 Hz.

(iii) In the range from the Schroeder frequency, f_{g} to approximately $4f_{\mathrm{g}}$, the wavelengths are often too short for accurate modelling using wave acoustics and too long for geometric acoustics and so a statistical treatment is usually employed. This range is from 138 to 552 Hz for a room of $3 \times 5 \times 7$ m and $T_{60} = 0.5$ s. For a car passenger compartment of volume $V = 2.5\ \mathrm{m}^3$ and $T_{60} = 0.05$ s, this range is from 282 to 1131 Hz.

(iv) At high frequencies $f \gg c/L_{\max}$, for which the dimensions of the room are large compared with the wavelength of the sound, geometrical or ray room acoustics apply. This covers a wide range of audio frequencies in standard rooms. Hence, in this frequency range, specular reflections and the sound ray approach to acous-

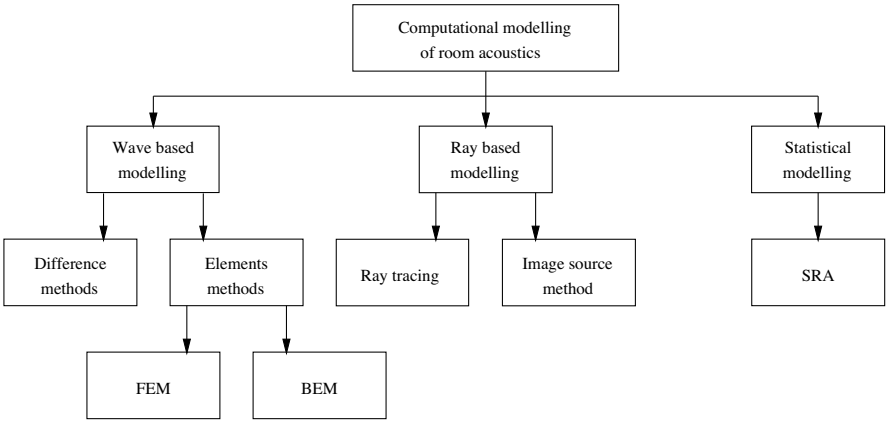


Fig. 2.3 Methods for modelling and simulating room acoustics

tics prevail. Because the sound is represented by energy waves rather than complex pressure waves, geometrical acoustics often neglect wave related effects such as diffraction and interference.

2.2 Models of Room Reverberation

A consequence of the complex nature of room reverberation is the desire for simple and accurate models of the reverberation process that can be used in, for example, the analysis or synthesis of room acoustics. Such models can be considered in three classes: wave-based models, ray-based models and statistical models [43]. It has already been shown that different analysis techniques are appropriate for different ranges of frequency of sound in rooms. A combination of modelling techniques is therefore necessary to achieve accuracy over the full audio spectrum. For speech signals, however, simpler modelling is usually performed because of the limited bandwidth of the signal, particularly when considering telecommunications applications.

In the following, the main models of room reverberation will be discussed in the context of Fig. 2.3, which shows a hierarchical overview. The subsequent discussion will begin with a short intuitive outline and will then consider specific modelling approaches. We will first consider wave based modelling which is based on the wave equation and is a fundamental approach. We will then discuss the concept of ray based modelling which leads to ray tracing methods and, subsequently, to the image method for modelling reverberation using virtual sources. Lastly, we will describe statistical models of reverberation that lead to the framework of Statistical Room Acoustics (SRA).

2.2.1 Intuitive Model

Intuitive consideration of the impulse response of a room leads to the deduction that the response must decay with time and will contain little or no deterministic structure after a sufficiently large number of reflections. Moorer [29] noted the resemblance between a concert hall impulse response and a synthetic response formed from a white noise signal multiplied by an exponentially decaying envelope. It was reported that the result of convolving anechoic signals with such a synthetic response gave a natural-sounding reverberation effect. This observation leads to a more detailed discussion of Polack's model in Sect. 2.2.6.

2.2.2 Finite Element Models

Analytical solutions for the wave equation can normally only be found in simple cases, such as for rectangular rooms with rigid walls. In other situations, it is therefore attractive to consider numerical wave-based methods. Finite Element Method (FEM) and the related Boundary Element Method (BEM) [22, 36] can be used for modelling and simulation of room acoustics. In both these numerical approaches, the elements interact with each other to represent wave propagation. The size of the elements has to be chosen to be much smaller than the size of the wavelength for all frequencies of interest and, therefore, at high frequencies the required number of elements becomes very large, resulting in a high computational complexity. These methods are therefore more suitable for low frequencies and small enclosures.

Another method for room acoustics simulation is provided by the Finite-Difference Time-Domain (FDTD) method [4, 44]. The main principle of this approach is that derivatives in the wave equation are replaced by corresponding finite differences. The FDTD method produces impulse responses that are better suited for auralization than FEM and BEM. On the other hand, the main benefit of the element methods over FDTD methods is that one can create a denser mesh structure where required, such as locations near corners or other acoustically complex regions. In all wave-based methods, it is usually highly challenging to incorporate appropriate boundary conditions and geometrical description of the objects within the acoustic environment. Hence, application of these approaches in the literature has been limited.

2.2.3 Digital Waveguide Mesh

The flexibility and suitability of Digital Waveguide Mesh (DWM) techniques for simulating room acoustics have been demonstrated in recent years [2, 31, 45]. Modelling of wave propagation effects, such as diffusion and scattering, for example, is intrinsic to the approach, and results can be obtained that are both accurate and

computationally tractable. Recently, integrated design software has been developed including Roomweaver [2], which makes DWM room simulation more intuitive and straightforward. Several open issues remain the subject of research including accurate modelling of dispersion and the dependence on frequency of reflector properties.

2.2.4 Ray-tracing

Acoustic propagation in geometrical room acoustics can be simplified into a form in which sound waves are represented by rays and reflections are specular. Such a simplification is valid when the diffraction and interference effects found in wave propagation are insignificant, such as when the wavelength of the sound is small compared to the dimensions of the reflecting surfaces in the room and large compared to any structural details or surface texture.

Ray-tracing techniques have been proposed [23] in which rays of sound are emitted from the source and arrive at the point of measurement after zero or more specular reflections. The measurement accumulates the rays to build up the acoustic response from source to measurement position. It is advantageous to control the number of rays used in a simulation, which can be of the order 10^5 or more, by limiting the model to include only first and second order reflections in order to maintain the computational requirement of the simulation at a modest level. Ray-tracing methods and their accuracy has been further considered in, for example, [12, 24].

2.2.5 Source-image Model

A further example of ray-based modelling of room reverberation is the source-image method, originally proposed by Allen and Berkley [1]. This is one of the most commonly used techniques for simulating room acoustics in the context of speech reverberation. Given a single omnidirectional source in a reverberant room, the effect of reverberation is represented using a set of source images. The location of the image sources is determined by the dimensions of the room, which is assumed to be rectilinear. All image sources simultaneously emit the same signal as the true source. The signals emitted by the image sources arrive at the measurement location at times and with intensities that depend on the distances between the image sources and the measurement location. A reflection coefficient, ϕ, is applied to account for the sound reflected by the surfaces and is related to the absorption coefficient.

An illustration of the source-image method for a two-dimensional case is depicted in Fig. 2.4, where the room is indicated with a bold rectangle. In practice, the images extend over a three-dimensional lattice. Due to different distances from each image to the measurement location, the signals arrive at the microphone at different times and with different intensities. Moreover, finite reflection coefficients

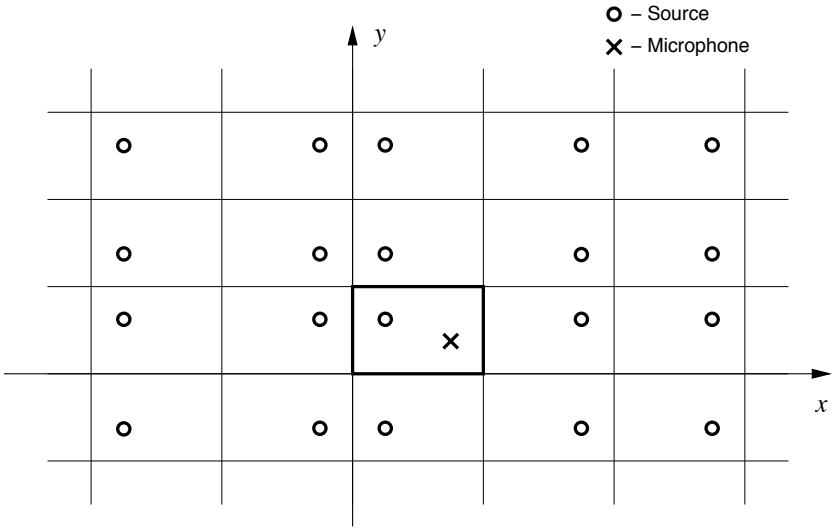


Fig. 2.4 Example image sound sources for a rectangular room

$\phi_{x1}, \phi_{x2}, \phi_{y1}, \phi_{y2}, \phi_{z1}$ and ϕ_{z2}, which are assumed to be independent, are applied to account for sound reflected by each the six walls and are related to the average wall absorption coefficient $\bar{\alpha}$ in (2.11) according to $\phi = \frac{1}{6}(\phi_{x1} + \phi_{x2} + \phi_{y1} + \phi_{y2} + \phi_{z1} + \phi_{z2}) = \sqrt{1-\bar{\alpha}}$. Thus, for a rectangular room with dimensions (L_x, L_y, L_z), a receiver positioned at (x, y, z), and a source positioned at $(\tilde{x}, \tilde{y}, \tilde{z})$, the i^{th} tap of the AIR can be computed as [1]

$$h_i = \sum_{\boldsymbol{\epsilon}=0}^{1} \sum_{\boldsymbol{\rho}=-\infty}^{\infty} \phi_{x1}^{|q-u|} \phi_{x2}^{|q|} \phi_{y1}^{|r-v|} \phi_{y2}^{|r|} \phi_{z1}^{|s-w|} \phi_{z2}^{|s|} \times \frac{\delta(n - (|\boldsymbol{D}_\varepsilon + \boldsymbol{D}_\rho|/c))}{4\pi|\boldsymbol{D}_\varepsilon + \boldsymbol{D}_\rho|}, \tag{2.21}$$

where $\boldsymbol{D}_\varepsilon = (x - \tilde{x} + 2u\tilde{x}, y - \tilde{y} + 2v\tilde{y}, z - \tilde{z} + 2w\tilde{z})$ and $\boldsymbol{D}_\rho = (qL_x, rL_y, sL_z)$ such that $|\boldsymbol{D}_\varepsilon + \boldsymbol{D}_\rho|$ defines the Euclidean distance between the receiver and each source image. The triplets $\boldsymbol{\epsilon} = (u, v, w)$ and $\boldsymbol{\rho} = (q, r, s)$ indicate that each summation in fact consists of three different summations. Although, $\boldsymbol{\rho} = (q, r, s)$ is given in the interval $[-\infty, \infty]$, in practice the limits are finite and are governed by the chosen order of source images. The expression in (2.21) follows from the solution of the wave equation for a rectangular enclosure [1].

In the original implementation of (2.21), impulses calculated at fractional sampling delays are rounded to the nearest sample, which at lower sampling frequencies can introduce significant errors. Peterson [34] proposed to apply a low-pass filter to each impulse obtained in (2.21), which better satisfies the sampling theorem and provides a more accurate representation of the simulated impulse responses, in particular for multi-microphone scenarios. This is equivalent to a fractional delay [27]

implementation of each impulse. Consequently, a Hanning-windowed ideal low-pass filter is applied to each impulse resulting from a simulated image according to [34]

$$w_{\mathrm{L},n} = \begin{cases} 0.5\,(1+\cos(2\pi n/L_{\mathrm{w}}))\,\mathrm{sinc}(2\pi f_{\mathrm{co}} n), & -L_{\mathrm{w}}/2 < n < L_{\mathrm{w}}/2 \\ 0, & \text{otherwise,} \end{cases} \tag{2.22}$$

where f_{co} is the filter cut-off frequency, often taken as $f_{\mathrm{co}} = f_s/2$ and L_{w} is the window length, which in [34] is taken as 4 ms, so that $L_{\mathrm{w}} = \frac{4 f_s}{1000}$.

2.2.6 Statistical Room Acoustics

Within the framework of SRA, the amplitudes and phases of all reflected acoustic plane waves in a room are considered randomly distributed such that they form a uniform, diffuse sound field at any point in the room. Subsequently, it is assumed that the Acoustic Transfer Function (ATF) from the source to the m^{th} microphone can be expressed as the sum of a direct-path component comprising only the direct-path propagation, $H_{\mathrm{d},m}(e^{j\omega})$, and a reverberant component comprising all reflections, $H_{\mathrm{r},m}(e^{j\omega})$, such that

$$H_m(e^{j\omega}) = H_{\mathrm{d},m}(e^{j\omega}) + H_{\mathrm{r},m}(e^{j\omega}), \quad m = 1,2,\ldots,M. \tag{2.23}$$

The following assumptions are used in SRA and are valid for many practical situations over the bandwidth important for speech communication.

1. The dimensions of the room are large relative to the wavelength at all frequencies of interest.
2. The average spacing between the resonant frequencies of the room is smaller than one third of their bandwidth. This can be satisfied at all frequencies above the Schroeder frequency.
3. Sound sources and microphones are situated in the room interior at least a half wavelength from the surrounding walls.

Under these conditions and due to the different propagation directions and the random relation of the phases of the direct-path component and all the reflected waves, it can be assumed that the direct and the reverberant components are uncorrelated [25, 32]. We next employ the spatial expectation, $\mathscr{E}\{\cdot\}$, which was first introduced in Sect. 2.1.2, defined in this context as the expectation over all allowed source and microphone positions in a room. The spatial expectation of the energy density spectrum of the ATF can now be written as

$$\mathscr{E}\{|H_m(e^{j\omega})|^2\} = \mathscr{E}\{|H_{\mathrm{d},m}(e^{j\omega})|^2\} + \mathscr{E}\{|H_{\mathrm{r}}(e^{j\omega})|^2\}, \tag{2.24}$$

since the spatial expectation of the cross-terms vanish [15, 32]. The spatial expectation, $\mathscr{E}\{\cdot\}$, gives a result that is, in general, independent of the source and micro-

phone positions and can be found using the methods of Radlović *et al.* [40] and Gustafsson *et al.* [15]. An initial geometry is first defined in terms of the source position, $\mathbf{q}_{\text{src}}(0)$, and an initial position for each microphone, $\mathbf{q}_{\text{mic},m}(0)$. Random translation vectors, $\boldsymbol{\theta}(i)$, and rotation vectors, $\boldsymbol{\Theta}(i)$, are used to generate the i^{th} realization of the geometry

$$\mathbf{q}_{\text{src}}(i) = \boldsymbol{\Theta}(i)\mathbf{q}_{\text{src}}(0) + \boldsymbol{\theta}(i), \tag{2.25}$$

$$\mathbf{q}_{\text{mic},m}(i) = \boldsymbol{\Theta}(i)\mathbf{q}_{\text{mic},m}(0) + \boldsymbol{\theta}(i), \tag{2.26}$$

$$i = 1,\ 2,\ \ldots,\ \mathcal{N}, \tag{2.27}$$

such that the spatial relationships between microphones and between the source and each microphone remain constant while the absolute position and orientation in the room of the source-microphone configuration are varied randomly. The spatial expectation, $\mathscr{E}\{\cdot\}$, is then estimated from an average over all $\mathcal{N}$ realizations.

Polack [37] developed a time-domain model that describes the AIR as one realization of a non-stationary stochastic process

$$h(t) = b(t)\exp^{-\bar{\zeta}t}, \quad \text{for } t \geq 0, \tag{2.28}$$

where $b(t)$ is a zero-mean stationary Gaussian random process that is characterized by its power spectral density $B(f)$, and $\bar{\zeta}$ is the average damping constant that is related to the reverberation time T_{60} by

$$\bar{\zeta} = \frac{3\ln(10)}{T_{60}}, \tag{2.29}$$

as also indicated by (2.20).

The time-domain response can only be Gaussian if a sufficient number of reflections overlap at any time along the response. Therefore, Polack's model becomes valid only after a certain amount of time that is called the mixing time. After the mixing time the peaks in the AIR no longer correspond to the arrivals of individual reflections. Since the reflection density increases with time the situation is similar to that found in the frequency domain, except that the spreading of a reflection in the time domain cannot be defined solely with respect to the intrinsic properties of the room (unlike the bandwidth of a mode). The spreading of a reflection in the time domain can only be expressed with reference to the bandwidth of the source excitation or microphone. If the criterion is that at least 10 reflections overlap within a characteristic time interval (taken equal to 24 ms in [37]) the mixing time is given by

$$t_{\text{mix}} = 1000\sqrt{V}\ \text{s}. \tag{2.30}$$

This expression was also proposed in [41] as a reasonable approximation for the transition time between early reflections and late reverberation. Polack also showed that the exponentially decaying stochastic model can be established within the framework of geometrical acoustics and billiard theory [37, 38]. In this context the mixing time is defined as the time it takes for a set of initially adjacent sound rays to

spread uniformly across the room. By that time (if the origin is taken as the time of emission of a sound pulse by the source), the reverberation process has become diffuse, i.e., the sound energy density and the direction of the intensity vector are uniformly distributed across the room. The mixing character of a room depends on its geometry and the diffusing properties of the boundaries. When mixing is achieved, the echo density increases exponentially with time, rather than being proportional to t^2 [38]. Consequently, the value $1000\sqrt{V}$ can be considered as an upper limit for the mixing time in typical 'mixing' rooms.

2.3 Subjective Evaluation

Subjective speech quality measures can be obtained using subjective listening tests in which human participants rate the performance of a system or quality of a signal in accordance with an opinion scale [18]. The International Telecommunications Union (ITU-T) has standardized the most commonly used methods for measuring the subjective quality of speech transmission over voice communication systems. For both listening-only and conversational tests the ITU-T recommends the use of a speech quality rating on a five-point category scale, which is commonly known as the listening-quality scale [18]. An alternative speech quality scale that is used in listening-only tests is the listing-effort scale. In conversational tests a binary conversation difficulty scale is usually employed. These scales are listed in Table 2.1.

Table 2.1 ITU-T recommended speech quality measurement scales [18]

Listening-quality scale:

Quality of the speech/connection	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Listening-effort scale:

Effort required to understand the meaning of sentences	Score
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

Conversation difficulty scale:

Did you and your partner have any difficulty in hearing over the connection?	Yes 1 / No 0

A listening test is performed by a number of subjects that listen to recordings that are degraded by an acoustic channel, and enhanced by the algorithm under

test. The subjects provide their opinion on the quality of each signal, or the effort required to understand it, using the listening-quality scale or listening-effort scale, respectively. In conversational tests, subjects use a voice communication system before providing their opinion on its quality. The Mean Opinion Score (MOS) is the averaged opinion score across subjects and indicates the subjective quality of the system or algorithm under test. To obtain a realistic variability in the opinion scores, a large numbers of subjects is required. Therefore, the main drawback of subjective testing is cost [35]. Even with a large number of subjects, the variance of MOS can still be high. Furthermore, the quality that is expected by a customer will be different depending on whether the device is an expensive conference system or a cheap mobile telephone. The constraints imposed by the need to limit the cost and the amount of subjects also limit the ability to test the system or algorithm under different environmental conditions. Hence, it is highly desirable to find automatic assessment systems based on objective measures by which an indication of quality can be obtained.

2.4 Channel-based Objective Measures

As will be seen later in this book, dereverberation algorithms can be considered in two classes: (i) algorithms that affect the reverberant signal in a manner that can be represented by a linear filter whose response is either known or can be deduced, and (ii) algorithms that affect the reverberant signal in a manner for which a linear transfer function does not exist or cannot be deduced. In the former, the level of reverberation can be found using channel-based measures discussed in this section. In the latter, the channel impulse response is not available, so the level of reverberation must be found using only signal-based measures which will be discussed later in Sect. 2.5.

There are a number of objective measures of speech quality including, for example, the Perceptual Evaluation of Speech Quality (PESQ) [19] and PEMO-Q [17], which give a general indication of the expected perceived quality of speech. Many such measures were originally introduced in the context of speech coding and are intrusive in that they operate by comparing an observed signal (usually the output of some processing, coding or transmission system) with a reference signal (usually an original clean speech signal). Nonintrusive measures, in contrast, operate from the observed signal only, without the need for a reference signal.

When the nature of the processing through which the observed signal has passed is known, it is possible to develop targeted objective measures that aim to characterize that specific processing. In the case of reverberation, the nature of the 'processing' is known to be a linear convolution with the AIR. This leads to measurements based on the ratio between direct and reverberant components. When dereverberation processing is subsequently employed, it is possible to consider the total impulse response that describes the combination of the AIR and the effect of the dereverberation algorithm in cascade. In the case of an ideal dereverberation algorithm, this

total response tends to a sinc function with zero crossings separated by the sampling period and with system dependent scaling and delay. The performance of non-ideal dereverberation algorithms can be assessed by measuring how closely to the ideal case they perform. In general, objective measures are often used to assess the improvement brought about through processing by comparing the measures evaluated before and after the processing.

2.4.1 Normalized Projection Misalignment

An important measure in the context of system identification is Normalized Projection Misalignment (NPM), championed by Morgan *et al.* [30]. This is a method for quantifying the accuracy of an estimated impulse response in a manner that is independent of any multiplicative scaling of the estimate. There are several applications, of which dereverberation is one, in which it is desirable to be able to estimate an unknown system but for which the scaling of the estimate is irrelevant; it can be ignored, deduced or assumed. In the case of blind system identification, it is common to employ algorithms that aim to identify the unknown system only to within a scale factor, hence it is necessary to ignore the scaling of the estimate when evaluating its accuracy. In the specific case of dereverberation, an equalizer could potentially be designed, based on such an estimate of the acoustic system, and applied to the reverberant speech in order to invert the effect of the room reverberation. The effect of scaling in the estimate can be straightforwardly compensated by a gain factor in the equalizer in order to bring the output signal to a convenient level.

NPM is usually quoted in decibels and is defined as

$$\text{NPM} = 10\log_{10}\left(\frac{||\mathbf{h}-\beta\hat{\mathbf{h}}||_2^2}{||\mathbf{h}||_2^2}\right)\ \text{dB}, \tag{2.31}$$

where $\mathbf{h}$ and $\hat{\mathbf{h}}$ represent the true and estimated impulse response vectors, respectively. The scalar β is a gain factor dependent on $\mathbf{h}$ and $\hat{\mathbf{h}}$ and is chosen such that the NPM is minimized. It is computed as

$$\beta = \frac{\mathbf{h}^T\hat{\mathbf{h}}}{\hat{\mathbf{h}}^T\hat{\mathbf{h}}}. \tag{2.32}$$

A geometric interpretation of NPM is to consider it not as a measure of the squared distance from $\mathbf{h}$ to $\hat{\mathbf{h}}$ but as a measure of the squared distance from $\mathbf{h}$ to the projection of $\mathbf{h}$ onto $\hat{\mathbf{h}}$.

2.4.2 Direct-to-reverberant Ratio

The most direct objective measure is the Direct to Reverberant Ratio (DRR) and is defined as

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_\mathrm{d}} h^2(n)}{\sum_{n=n_\mathrm{d}+1}^{\infty} h^2(n)}\right)\ \mathrm{dB}, \tag{2.33}$$

in which samples of the channel impulse response, $h(n)$, indexed from zero up to n_d are assumed to represent only the direct-path propagation, while samples of the channel impulse response with indices greater than n_d represent only the reverberation due to reflected paths.

It is intuitively helpful to visualize the direct-path propagation as being represented by the largest magnitude tap in the early part of the channel impulse response. However, this intuitive scenario is only correct when the propagation time from source to microphone is an integer number of sample periods. In general, finite rate sampling of the AIR results in the direct-path propagation being represented by samples of a sinc function corresponding to the sampling kernel and centred according to the direct-path propagation delay.

When synthetic AIRs are used, the direct-path can be computed separately. However, when dealing with measured impulse responses the direct-path component, and therefore the related energy, cannot be determined precisely. Therefore, $n_\mathrm{d}/f_\mathrm{s}$ is often taken 8 to 16 ms larger than the approximate arrival time of the direct sound.

It should be noted that the DRR depends on the distance between the source and the microphone, and on the reverberation time of the room. We can express the DRR using (2.7) and (2.8) as

$$\mathrm{DRR} = 10\log_{10}\left(\frac{QR}{16\pi D^2}\right), \tag{2.34}$$

where Q is the directivity factor, R is the room constant given by (2.9), and D is the source-microphone distance. Note that the room constant is inversely proportional to the reverberation time.

The DRR is also related to the spectral deviation, which is a measure of the deviation of the spectrum from white and is straightforwardly defined as the standard deviation of the energy spectrum of the AIR in dB [20].

2.4.3 Early-to-total Sound Energy Ratio

The earliest attempt to define an objective criterion of what may be described as the distinctness of sound is called *definition* (originally *Deutlichkeit*) or early-to-

total sound energy ratio. The range of time within the impulse response taken to correspond to early reflections is typically the first 50 to 80 ms. This time, in milliseconds, is often used as a subscript such that, in the case of $n_e/f_s = 50$ ms, the definition can be written as

$$D_{50} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_e} h^2(n)}{\sum_{n=0}^{\infty} h^2(n)}\right) \text{ dB}. \tag{2.35}$$

2.4.4 Early-to-late Reverberation Ratio

Another objective criterion is known as the Early to Late reverberation Ratio (ELR) or *clarity index* (originally *Klarheitsmaß*) and it is defined as

$$C = 10\log_{10}\left(\frac{\sum_{n=0}^{n_e} h^2(n)}{\sum_{n=n_e+1}^{\infty} h^2(n)}\right) \text{ dB}, \tag{2.36}$$

where n_e/f_s is also usually chosen to be in the range of 50 to 80 ms. Similarly to *definition*, the time (in milliseconds) is often used as a subscript, i.e., in the case $n_e/f_s = 50$ ms the ELR is denoted by C_{50}. The division of the impulse response into an early and a late portion is motivated by way in which the human auditory system interprets multipath signal components as a single signal if the arrival times of components differ by less than around 50 ms. Therefore, the relative strength of the early reflections compared to the late reflections gives a measure of how much of the nondirect-path energy will be perceived of as coloration of the direct-path component, compared to echoey reverberation.

2.5 Signal-based Objective Measures

In cases for which the reverberating system's impulse response is available, it is natural to compute performance measures that employ this information. There are, however, some cases for which the effect of a dereverberation algorithm cannot be characterized in terms of an impulse response, such as [9, 14, 50]. In these cases, the processing is not Linear Time-Invariant (LTI) and hence cannot be described in the normal manner using an impulse response. Accordingly, performance measures must then be computed from the signals alone, without reference to the reverberating system.

2.5.1 Log Spectral Distortion

The Log Spectral Distortion (LSD) is one of the most straightforward and long-standing speech distortion measures and it has been shown in [16] to be moderately well suited to the assessment of dereverberation algorithms. It is computed as the RMS value of the difference of the log spectra of the original clean speech signal $s(n)$ and the signal under test $x(n)$ which is normally the output of a processing algorithm. It is common to use the FFT-based short-time spectra $X(l,k)$ and $S(l,k)$ of $x(n)$ and $s(n)$ respectively, where l denotes the time frame and k the frequency bin. Frames of duration between 32 and 64 ms are typically employed with overlapping between adjacent frames of between 50 and 75%. The RMS value of the difference between $S(l,k)$ and $X(l,k)$, in the l^{th} frame, is defined as

$$\text{LSD}(l) = \left(\frac{2}{K} \sum_{k=0}^{\frac{K}{2}-1} |\mathcal{L}\{X(l,k)\} - \mathcal{L}\{S(l,k)\}|^2 \right)^{\frac{1}{2}} \text{dB}, \tag{2.37}$$

where $\mathcal{L}\{X(l,k)\} = \max\{20\log_{10}(|X(l,k)|), \delta\}$ is the log spectrum confined to about 50 dB dynamic range ($\delta = \max_{l,k}\{20\log_{10}(|X(l,k)|)\} - 50$), and likewise $\mathcal{L}\{S(l,k)\}$. The mean LSD is obtained by averaging (2.37) over all frames containing speech.

2.5.2 Bark Spectral Distortion

Bark Spectral Distortion (BSD) was presented in [46] as an intrusive measure for predicting the subjective quality of speech coders. It has also been employed subsequently as an objective evaluation measure in other areas of speech signal processing, most notably speech enhancement, for which the intention is usually to benefit from the perceptual significance of the Bark scale to try to improve the measurement of speech quality as it would be perceived by a human listener. Studies in [7, 46] have shown strong correlation between the BSD measure and MOS scores. In the broad area of speech quality assessment, and in particular in the evaluation of speech coders, more recent developments such as PESQ [19] and subsequent developments are now usually preferred. Nevertheless, BSD is a natural counterpart to LSD, introducing the influence of human perception on speech quality assessment methods.

BSD operates in the perceptually motivated Bark spectral domain that incorporates human auditory models [46]. It is therefore able to operate in cases where time-domain waveform preservation is not expected, such as in many speech coders and also in the case of reverberation.

To describe the use of BSD, consider a known speech signal, $s(n)$, and an observation of the speech signal, $x(n)$, in a reverberant environment. The BSD metric makes use of $\mathcal{B}_s$ and $\mathcal{B}_x$, the Bark spectra of $s(n)$ and $x(n)$, respectively. In order

to measure the effectiveness of a speech dereverberation algorithm, BSD can be computed and compared before and after dereverberation processing.

The three steps involved in computing the Bark spectrum are: (i) critical band filtering, (ii) equal loudness pre-emphasis and (iii) phon to sone conversion. For each frame of the signals s and x, these three steps are computed starting from the magnitude squared spectrum. The resulting Bark spectra are denoted $\mathcal{B}_s(l,k)$ and $\mathcal{B}_x(l,k)$, respectively, where l is the frame index and k is the Bark frequency bin.

Having computed the Bark spectra, the BSD score can be obtained using

$$\text{BSD} = \frac{1}{N_{\text{frm}}} \sum_{l=0}^{N_{\text{frm}}-1} \frac{\sum_{k=1}^{K} \left(\mathcal{B}_s(l,k) - \mathcal{B}_x(l,k)\right)^2}{\sum_{k=1}^{K} \left(\mathcal{B}_s(l,k)\right)^2}, \tag{2.38}$$

where N_{frm} denotes the number of analysis frames. The resulting BSD score for a speech signal is the average of the BSD scores for all of the analysis frames.

Modified Bark Spectral Distortion (MBSD) incorporates a noise-masking threshold into the BSD [49] with the aim that the MBSD measure will be able to differentiate between audible and inaudible distortions. The MBSD measure assumes that loudness differences below the noise masking threshold are not audible and are therefore excluded from the calculation of the perceptual distortion. MBSD uses a simple cognition model to calculate the distortion value [49].

Several approximations of the theoretical formulation of BSD are often necessary for practical implementations of the measure. Most current implementations make approximations, the validity of which is restricted to narrowband speech, in particular in order to simplify the modelling of the equal loudness curves. With wideband speech becoming more common through, for example, Voice Over Internet Protocol (VoIP) applications, there is a need for validated wide-band quality measures and a wide-band BSD measure would be a valuable development.

2.5.3 Reverberation Decay Tail

Two effects due to reverberation can be observed in speech. As has been previously discussed, the human auditory system interprets multipath signal components as a single signal if the components' times of arrival differ by less than around 50 ms. As a consequence, the early reflections in an AIR give rise to the first effect – the perception of colouration of the speech. The nature of the colouration corresponds to filtering with filter coefficients from approximately the first 50 ms of the AIR. The second effect is the commonly understood symptom of reverberation in which the sound 'rings on' for a short time before decaying, resulting in smearing such that sounds are spread over a longer time giving rise to an impression of space and distance. This second effect is referred to here as the reverberation tail effect. Since there are two effects, each with a different perceptual impact, it is logical to consider evaluating the effects separately.

R_{DT} is an objective measure of the tail effect proposed in [48], which jointly characterizes the relative energy in the tail of the AIR and the rate of its decay. It is intended to be substantially independent of colouration and does not require estimation of the AIR. The analysis of the test signal and the reference signal are performed in the Bark spectral domain.

The measure first requires the definition of an *end-point* as an instant of time at which the speech energy falls abruptly and a *flat-region* as a period of time immediately following and end-point during which there is no significant increase in speech energy due to speech onset. The function

$$\Delta\mathcal{B}_x(l,p,k) = \mathcal{B}_x(l,k) - \mathcal{B}_x(l+p,k) \tag{2.39}$$

describes the difference between the sone values at time frame l and $l+p$ in the k^{th} Bark bin. An end-point is defined in bin k at frame l_{ep} when a sufficiently large negative gradient is found, satisfying

$$\Delta\mathcal{B}_x(l_{\mathrm{ep}},-1,k) > -\delta_1. \tag{2.40}$$

For each end-point, the number of frames I is counted for which

$$\Delta\mathcal{B}_x(l_{\mathrm{ep}}+i,1,k) > \delta_2 \quad i = 1,2,\ldots,I. \tag{2.41}$$

A flat-region may only follow an end-point. A flat-region is defined in any Bark bin as a range of time frames within which the sone loudness in that bin varies less than δ_{min} and is lower in amplitude than a given floor δ_{floor} for a duration of at least J frames, satisfying

$$\Delta\mathcal{B}_x(l_{\mathrm{ep}}+I,j,k) < \delta_{\mathrm{min}} \quad \text{and} \quad \mathcal{B}_x(l_{\mathrm{ep}}+I+j,k) > \delta_{\mathrm{floor}}, \tag{2.42}$$

with $j = 1,2,\ldots,J$. Typical values of I, J, δ_1, δ_2, δ_{min} and δ_{floor} are 2, 5, 0.2, 0.1, 0.1 and 0.2 for 32 ms frames. The measure uses the reference speech signal to determine end-points and flat-regions.

Next, the decay curve

$$d(l,k) = A_k e^{\lambda_k n}, \quad n = 1,2,\ldots,J, \tag{2.43}$$

of length J is fitted to the reverberant speech over all flat-regions and in all Bark bins $k = 1,\ 2,\ \ldots,\ K$. The value of $l = 0$ is specifically omitted from the fitting procedure.

For a particular end-point, the R_{DT} measure employs the following three terms: $A_{\mathrm{avg}} = \frac{1}{K}\sum_{k=1}^{K} A_k$ represents the average absolute decay tail energy; $\lambda_{\mathrm{avg}} = \frac{1}{KA_{\mathrm{avg}}}\sum_{k=1}^{K} A_k\lambda_k$ represents the average decay rate, $\mathcal{D}_{\mathrm{avg}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{D}_k$ represents the average direct-path energy estimated from the reference signal, where $\mathcal{D}_k$ is the direct-path energy estimate in the k^{th} Bark bin.

The R_{DT} measure is then computed as

$$R_{\mathrm{DT}} = \frac{A_{\mathrm{avg}}}{\lambda_{\mathrm{avg}} \mathcal{D}_{\mathrm{avg}}}. \tag{2.44}$$

Note that a high R_{DT} value corresponds to either a high relative energy in the tail or a slower decay rate.

In [47], the R_{DT} measure was tested using three dereverberation methods. The results were compared to subjective tests involving 26 subjects. The results showed a high correlation between the R_{DT} values and the amount of reverberation perceived by the subjects.

2.5.4 Signal-to-reverberant Ratio

The Signal to Reverberation Ratio (SRR) is a signal-based measure of reverberation that can be computed even when the effect of a dereverberation algorithm cannot be represented in the impulse response of an LTI system. It requires knowledge of the original speech after propagation through the direct-path, s_{d}, which is usually difficult and often impossible to obtain when dealing with measured signals but easily available in an intrusive situation when the original signal is known. Typically, the SRR is computed using the signals before and after processing, and an improvement in SRR due to the processing can then be determined. The SRR requires the direct-path signal component, $s_{\mathrm{d}}(n)$, and is therefore an intrusive measure. It and can be written as

$$\mathrm{SRR} = 10 \log_{10} \left(\frac{\|\mathbf{s}_{\mathrm{d}}\|_2^2}{\|\hat{\mathbf{s}} - \mathbf{s}_{\mathrm{d}}\|_2^2} \right) \mathrm{dB}, \tag{2.45}$$

where $\mathbf{s}_{\mathrm{d}} = [s_{\mathrm{d}}(0)\ s_{\mathrm{d}}(1)\ \ldots\ s_{\mathrm{d}}(L_s - 1)]^T$, $s_{\mathrm{d}}(n) = \mathbf{h}_{\mathrm{d}}^T \mathbf{s}(n)$ and $\hat{\mathbf{s}} = (\hat{\mathbf{s}}_{\mathrm{d}} + \hat{\mathbf{x}}_{\mathrm{r}})$ is the L_s-sample signal to be evaluated. For example, the SRR may be measured first using $\hat{\mathbf{s}} = \mathbf{x}$ at the input of a dereverberation algorithm and then with $\hat{\mathbf{s}}$ at the output of the algorithm in order to quantify the effect of the algorithm. It is sometimes convenient to use the segmental SRR, $\mathrm{SRR}_{\mathrm{seg}}$. This is found by computing $\mathrm{SRR}(l)$ as the SRR of short, possibly overlapping, signal segments each of length L_s typically corresponding to a duration of 32 ms. An average of such SRR values in dB is then taken over all segments to give

$$\mathrm{SRR}_{\mathrm{seg}} = \frac{1}{N_{\mathrm{seg}}} \sum_{l=0}^{N_{\mathrm{seg}}-1} \mathrm{SRR}(l), \tag{2.46}$$

where N_{seg} is the total number of speech segments.

2.5.4.1 Relationship Between DRR and SRR

Subject to correct level normalization as will be discussed below, the SRR is equivalent to the DRR when the source $s(n)$ is spectrally white. In the case when $\hat{\mathbf{s}}_{\mathrm{d}} = \mathbf{s}_{\mathrm{d}}$

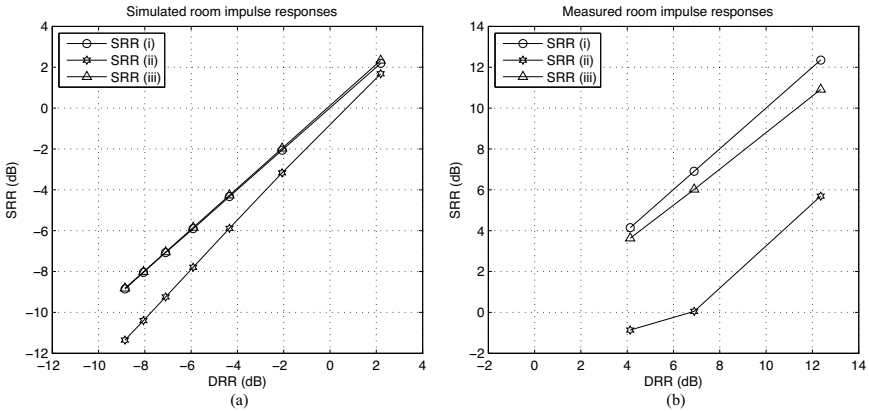


Fig. 2.5 Comparison of DRR and SRR for (i) white Gaussian noise input, (ii) speech input and (iii) prewhitened speech input, with (a) simulated impulse responses and (b) measured impulse responses

and evoking Parseval's theorem, in the frequency domain we have

$$\sum_k |S(k)|^2 |H_\mathrm{d}(k)|^2 / \sum_k |S(k)|^2 |H_\mathrm{r}(k)|^2. \tag{2.47}$$

When $S(k) = S$, independent of k, $|S|^2$ can be taken outside the summation in both numerator and denominator and cancelled. An illustrative example is when $s(n) = \delta(n)$, so that $S(k) = 1\,\forall k$, in which case (2.45) reduces directly to the formulation of the DRR in (2.33). In practice, when speech signals are considered, a prewhitening filter can be employed [39], as will be shown below.

These effects are illustrated in Fig. 2.5, which shows a comparison of DRR and SRR for (a) a room of dimensions $6 \times 5 \times 4$ m simulated using the source-image method [1, 34] and (b) for real measured room impulse responses from MARDY [47]. The SRR calculated for a white noise input is shown in curve (i) and is seen to correspond almost exactly to DRR. Curve (ii) shows SRR calculated for five sentences of male speech, sampled at 20 kHz from the APLAWD database [28]. Lastly the results with prewhitened speech are shown in curve (iii). The prewhitening filters were computed over all five sentences using a 10^{th} order linear predictor; separate filters were obtained for $\mathbf{s}_\mathrm{d}$ and $\hat{\mathbf{s}}$ and were applied to each of the signals respectively. It is clear that whitening the speech signal is significantly effective.

2.5.4.2 Level Normalization in SRR

A dereverberation algorithm aims to attenuate the level of reverberation and may affect either or both of the direct-path signal $s_\mathrm{d}(n)$ or the reverberant component $x_\mathrm{r}(n)$ in order to improve the SRR. Therefore we can write that

$$\hat{s}(n) = \gamma s_{\mathrm{d}}(n) + \bar{x}_{\mathrm{r}}(n), \tag{2.48}$$

where $\bar{x}_{\mathrm{r}}(n)$ is the attenuated reverberant component and γ is a scalar assumed stationary over the duration of the measurement. We also assume that any processing delay has been appropriately compensated as is generally assumed in other measurements such as the SNR.

The measurement of the reverberant component's energy and the assessment of its impact on the speech signal must be done relative to the energy of the direct-path component. This can be conveniently accomplished by normalization in order to match the level of the direct-path component before and after processing. The aim of this normalization is to adjust the magnitude of $\hat{\mathbf{s}}$ such that the direct-path signal energy is unchanged by the dereverberation algorithm. This can be achieved by determining γ. The motivation for this comes from the observation that signal-based measures are not, in general, scale independent, and therefore, unless the scaling is correctly normalized, misleading results can be obtained.

This problem can be formulated as a search for a scalar $\hat{\gamma}$ such that the Normalized Signal-to-Reverberant Ratio (NSRR)

$$\mathrm{NSRR} = 10\log_{10}\left(\frac{\|\mathbf{s}_{\mathrm{d}}\|_2^2}{\|(1/\hat{\gamma})\hat{\mathbf{s}} - \mathbf{s}_{\mathrm{d}}\|_2^2}\right)\ \mathrm{dB} \tag{2.49}$$

is a good estimate of DRR.

It is necessary to estimate γ from the available signals. For baseline comparison purposes, straightforward approaches can be employed to determine γ using

$$\gamma_{norm} = \frac{\|W\{\hat{\mathbf{s}}\}\|_{norm}}{\|W\{\mathbf{s}_{\mathrm{d}}\}\|_{norm}} \tag{2.50}$$

corresponding to RMS and peak matching for $norm = 2$ and $norm = \infty$ respectively, and employing uniform and A-weighting [25] for $W\{.\}$ representing a corresponding weighting filter. These approaches lead to incorrect calculation of SRR as will be shown below.

A good solution to the normalization problem can be obtained using γ_{ls} from the least squares minimization

$$\gamma_{\mathrm{ls}} = \arg\min_{\hat{\gamma}} \|\hat{\mathbf{s}} - \hat{\gamma}\mathbf{s}_{\mathrm{d}}\|_2^2. \tag{2.51}$$

The solution for γ_{ls} is found by minimizing $J = E\{\|\hat{\mathbf{s}} - \hat{\gamma}\mathbf{s}_{\mathrm{d}}\|_2^2\}$ arising from (2.51), where $E\{\cdot\}$ denotes mathematical expectation. To minimize J, we differentiate it with respect to $\hat{\gamma}$ and set the result to zero, which gives

$$\frac{\partial J}{\partial \hat{\gamma}} = -2E\{\mathbf{s}_{\mathrm{d}}^T[\hat{\mathbf{s}} - \hat{\gamma}\mathbf{s}_{\mathrm{d}}]\} = 0. \tag{2.52}$$

The final step is to approximate expectations with sample averages giving γ_{ls} to be the value of $\hat{\gamma}$ satisfying (2.52) as

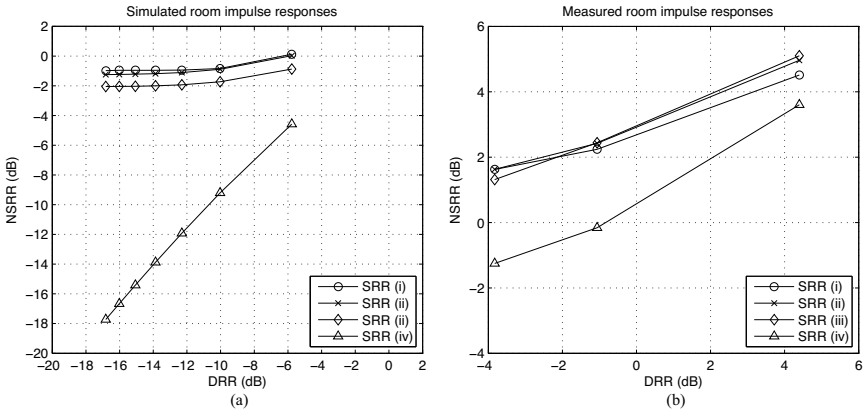


Fig. 2.6 Comparison with DRR and NSRR calculated using (i) peak normalization, (ii) RMS normalization, (iii) A-weighted RMS normalization, (iv) least squares optimal normalization, with (a) simulated impulse responses and (b) measured impulse responses

$$\gamma_{\text{ls}} = \frac{\mathbf{s}_{\text{d}}^T \hat{\mathbf{s}}}{\mathbf{s}_{\text{d}}^T \mathbf{s}_{\text{d}}}, \tag{2.53}$$

which is a projection of $\hat{\mathbf{s}}$ onto the direct component $\mathbf{s}_{\text{d}}$.

The effect of $\hat{\gamma}$ is seen by substituting (2.48) into J to obtain

$$\begin{aligned} J &= E\{\|\gamma \mathbf{s}_{\text{d}} + \bar{\mathbf{x}}_{\text{r}} - \hat{\gamma}\mathbf{s}_{\text{d}}\|_2^2\} \\ &= E\{(\gamma - \hat{\gamma})^2 \|\mathbf{s}_{\text{d}}\|_2^2\} + E\{2(\gamma - \hat{\gamma})\mathbf{s}_{\text{d}}^T \bar{\mathbf{x}}_{\text{r}}\} + E\{\|\bar{\mathbf{x}}_{\text{r}}\|_2^2\}. \end{aligned} \tag{2.54}$$

Clearly, J is minimized when $\gamma = \hat{\gamma}$.

2.5.4.3 SRR Computation Example

Figure 2.6 shows a comparison of DRR with NSRR computed from (2.49) with $\hat{\gamma}$ obtained using four different level normalization schemes. These results were obtained for the same experimental setup as described earlier in Sect. 2.5.4.1. The test signal $\hat{s}$ was generated as in (2.48) with γ chosen arbitrarily and $\bar{x}_{\text{r}}(n) = x_{\text{r}}(n)$. The speech signals were prewhitened with prewhitening filters computed from $\mathbf{s}_{\text{d}}$ and $(1/\hat{\gamma})\hat{\mathbf{s}}$ and applied, after the level normalization, to each of the signals, respectively. Curves (i), (ii) and (iii) show SRR with the normalization factor γ from (2.50) with peak normalization, RMS normalization and A-weighted RMS normalization, respectively. Curve (iv) shows SRR with least squares optimal normalization. It can be seen that the match between DRR and least squares optimal normalized SRR is much smaller over a wide range of DRRs, whereas other normalization schemes substantially that overestimate offer little discrimination between different values of DRR. These discrepancies are more severe at lower DRR values.

2.5.4.4 SRR Summary

Two effects require consideration when employing SRR. First, the signal characteristics affect the SRR calculation such that good estimates of DRR are obtained when the signal is white. Prewhitening of speech with a 10^{th}-order predictor has been seen to be sufficient for the cases studied here. Second, the level of the signals must be correctly normalized. It has been shown that level normalization using RMS, A-weighted RMS and peak matching are not appropriate. A least squares optimal normalization scheme has been formulated and it has been shown that this can be expressed as a projection of the signal onto the direct-path component. The above simulation results confirm that the least squares optimal level normalization and prewhitening enable DRR to be estimated without the requirement for impulse response measurements.

2.5.5 Experimental Comparisons

Signal-based objective measures have been experimentally compared to fundamental measurements associated with reverberation in order to gain an insight into the effectiveness of the objective measures. We here consider a comparison of the objective measures with the fundamental measures of reverberation time T_{60} and spectral deviation, which is defined as the standard deviation of the energy spectrum in dB of the AIR [20]. A speech signal of 40 s duration containing utterances of both male and female speech from the TIMIT database [8] has been employed. Reverberation is applied by convolving the original speech signal with a AIR generated using the image method and the resulting comparisons are shown graphically for several of the more significant objective measures below.

The relationship of segmental SRR to reverberation time is shown in Fig. 2.7(a) for source-microphone distances of 0.5 and 2.0 m, and the relationship of Segmental SRR to spectral deviation is shown in Fig. 2.7(b). It can be seen that SRR varies monotonically with reverberation time and shows a clear dependent relationship with spectral deviation in the range of Segmental SRR between around -10 and 10 dB.

Graphs in Figs. 2.8(a) and 2.8(b) show the relation between the BSD and reverbereration time and between the BSD and the spectral distortion, respectively. The relation between the BSD and the reverberation time depends on the distance D between the source and the microphone, as shown in Fig. 2.8(a). For $D = 0.5$ m there is seen to be an almost linear relation; for $D = 2$ m the relation is non-linear. For Fig. 2.8(b), the BSD values were calculated using different reverberation times and source-microphone distances. The results demonstrate that only very low BSD values correspond to a decrease in spectral deviation; BSD values less than 0.2 for a significant reduction. Hence, a decrease in BSD does not necessarily correspond to a significant reduction in spectral deviation.

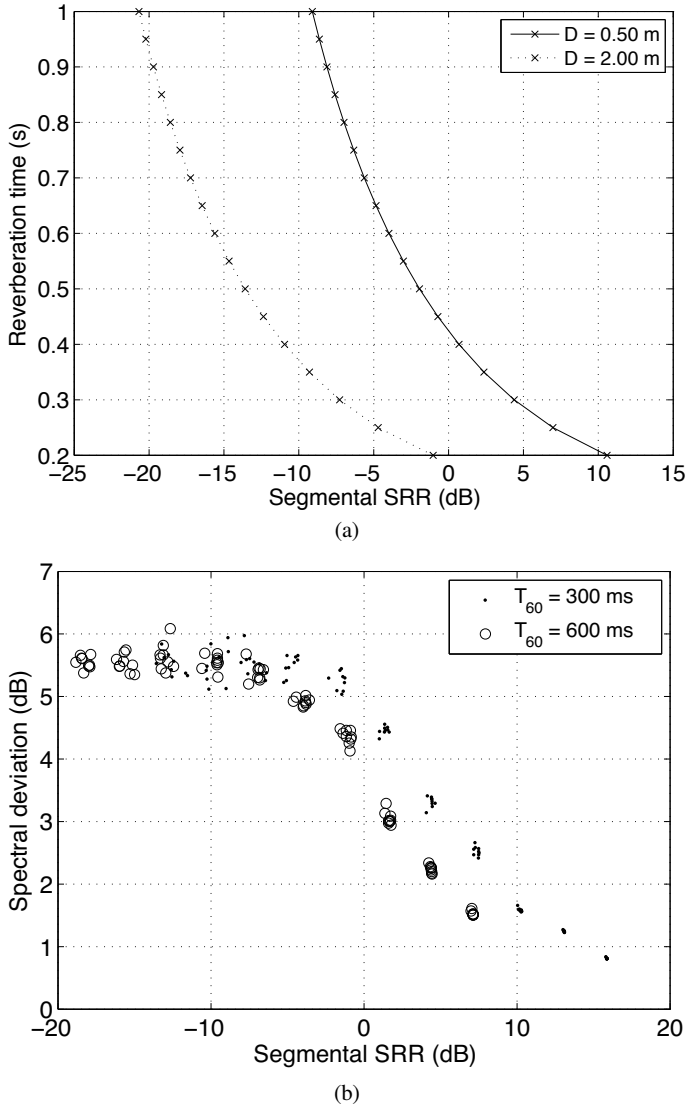


Fig. 2.7 Segmental SRR *vs.* (a) reverberation time and (b) spectral deviation

The relations between the reverberation decay tail measure, R_{DT}, and the reverberation time is shown in Fig. 2.9(a) and can be seen to be close to linear. Figure 2.9(b) shows the relation between the R_{DT} measure and the spectral deviation for $T_{60} = 300$ ms and $T_{60} = 600$ ms, from which it can be seen that the R_{DT} measure depends on the amount of colouration, which is indicated by the spectral deviation. Although the R_{DT} measure was intended to be independent of coloration, as

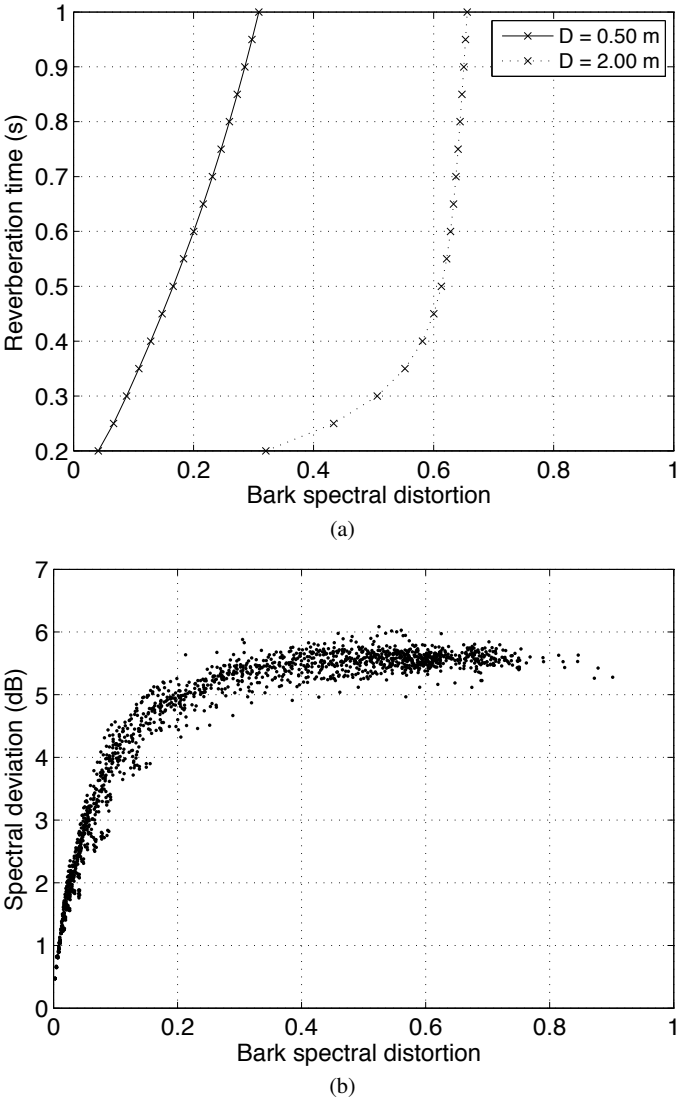


Fig. 2.8 Bark spectral distortion *vs.* (a) reverberation time and (b) spectral deviation

discussed in [48], the colouration process considered there was limited to that of strong early reflections, which causes a strong modulation in the power spectrum of the AIR. The R_{DT} measure is inversely proportional to the average direct-path energy $\mathcal{D}_{\mathrm{avg}}$ in (2.44). Since the average direct-path energy is inversely proportional to D^2, as is the direct energy, the R_{DT} measure can be expected to be proportional to D^2.

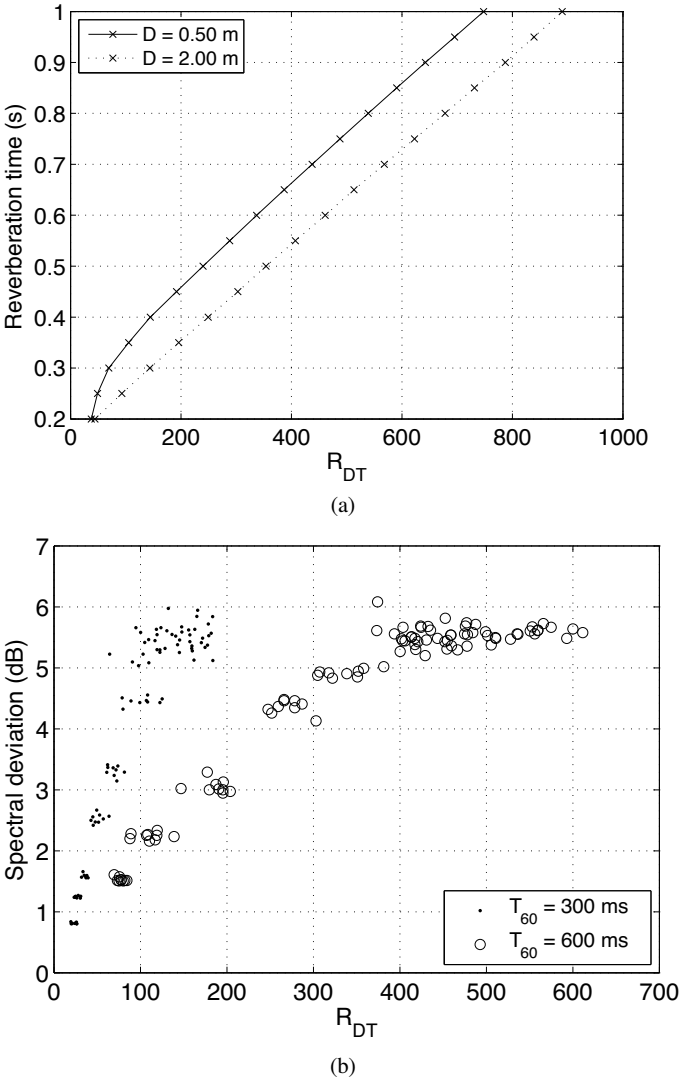


Fig. 2.9 R_{DT} *vs.* (a) reverberation time and (b) spectral deviation

2.6 Dereverberation Performance of the Delay-and-sum Beamformer

The Delay-and-sum Beamformer (DSB) is a fundamentally important approach to dereverberation and provides a helpful reference against which baseline other methods can be compared. The performance of the DSB is now analyzed.

The DSB applies spatial filtering to capture the direct-path signal from the direction of the source whilst suppressing reverberant sound from other directions. The source direction is assumed known or could be adaptively estimated, although the accuracy of direction-of-arrival estimation is often degraded in strong reverberation [5]. A closed form expression is now presented for the expected improvement in DRR that can be achieved with a DSB compared to a single microphone. The following expression is evaluated

$$\mathscr{E}\{\overline{\mathrm{DRR}}\} = 10\log_{10}\left(\frac{\mathscr{E}\{\mathrm{DRR}_{\mathrm{DSB}}\}}{\mathscr{E}\{\mathrm{DRR}'\}}\right)\ \mathrm{dB}, \tag{2.55}$$

where $\mathscr{E}\{\cdot\}$ is the spatial expectation operator, DRR' is the DRR of the best microphone which is defined as the microphone closest to the source and $\mathrm{DRR}_{\mathrm{DSB}}$ is the DRR at the output of the DSB. Using tools from SRA it can be shown [10] that the expected improvement in DRR that can be achieved with a DSB is

$$\mathscr{E}\{\overline{\mathrm{DRR}}\} = 10\log_{10}\left(\frac{D_{\min}^2\sum_{m=1}^{M}\sum_{l=1}^{M}\frac{1}{D_m D_l}}{\sum_{m=1}^{M}\sum_{l=1}^{M}\frac{\sin k\|\mathbf{q}_{\mathrm{mic},m}-\mathbf{q}_{\mathrm{mic},l}\|_2^2}{k\|\mathbf{q}_{\mathrm{mic},m}-\mathbf{q}_{\mathrm{mic},l}\|_2^2}\cos(k[D_m-D_l])}\right)\ \mathrm{dB}, \tag{2.56}$$

where D_m is the distance between the source and the m^{th} microphone, $D_{\min} = \min_m(D_m)$ is the distance from the source to the closest microphone and $\mathbf{q}_{\mathrm{mic},m}$ is the m^{th} microphone coordinate vector in three dimensions. The wave number is $k = 2\pi f/c$ with f denoting the frequency and c being the speed of sound in air, which here is taken as $c = 344$ m/s. The validity of the result in (2.56) depends on the common SRA assumptions for diffuse sound fields [25, 40] as described in Sect. 2.2.6.

The following observations can be made from the expression in (2.56): (i) the expected improvement that can be achieved with the DSB depends only on the distance between the source and the array and the separation of the microphones, (ii) consequently, the improvement is independent of the reverberation time and (iii) in the special case when the microphones are separated by exactly a half wavelength at each frequency and the distance between the source and the microphones is large, the denominator tends to zero and perfect dereverberation is achieved.

2.6.1 Simulation Results: DSB Performance

Two simulation results are presented to validate the theoretical expression in (2.56) and to gain some insight in the expected performance of the DSB for dereverberation. For these simulations, the source-image method [1] and the modification proposed in [34] were used to generate finite room impulse responses, $\mathbf{h}_m$. The room transfer function, $H_m(e^{j\omega})$, was then found by taking the Fourier transform of $\mathbf{h}_m$. A room with the dimensions $6.4 \times 5 \times 4$ m was modelled and a

linear array of M microphones with the spacing between adjacent microphones $\|\mathbf{q}_{\text{mic},m} - \mathbf{q}_{\text{mic},m+1}\| = 0.2$ m. The reverberation time was set to $T_{60} = 0.5$ s. Frequencies between 300–3400 Hz were considered, and sources and microphones were kept at least a half wavelength away from the walls to satisfy the conditions set for the statistical room model [10, 40].

Experiment 1: Effect of Source-microphone Distance

In the first experiment, an array with $M = 5$ microphones was employed. The distance between the array and the source was gradually increased from 0.5 to 3 m in steps of 0.5 m. The distance from the source to the array is defined here as the distance to the closest microphone. The result is shown in Fig. 2.10, where the improvement in DRR, calculated with the expression in (2.56), is plotted with a dashed line and the experimental result is shown with a solid line. It can be observed here that the improvement increases with the distance when the source is close to the array but then flattens out with increased distance. This can be related to the theoretical expression by noting that the improvement is mainly governed by the microphone separation when the distance to the array is large.

Experiment 2: Effect of Number of Microphones

For the second experiment, the distance between the source and the array was kept fixed at 2 m while the number of microphones was increased. The result of this is shown in Fig. 2.11, where the improvement in DRR, calculated with the expression in (2.56), is plotted with a dashed line and the experimental result is shown with a solid line. This result demonstrates that the improvement in DRR is approximately linearly proportional to the number of microphones.

In summary, the DSB is a straightforward approach that can provide moderate reduction in reverberation. Consequently, beamformers are often used as preprocessing or post-processing multichannel techniques and as benchmark methods for newly developed algorithms [6, 11, 13].

2.7 Summary and Discussion

Although it is indeed the science of room acoustics that offers an understanding of the physical acoustic processes that we refer to as reverberation, choosing appropriate models and simulation techniques from the relatively large number of those available could be said to be something of an art. This chapter has aimed to highlight the main classes of models and simulation techniques for room acoustics and has identified the circumstances under which each might be appropriately employed. We have not aimed to present all the techniques in detail, for obvious reasons, but

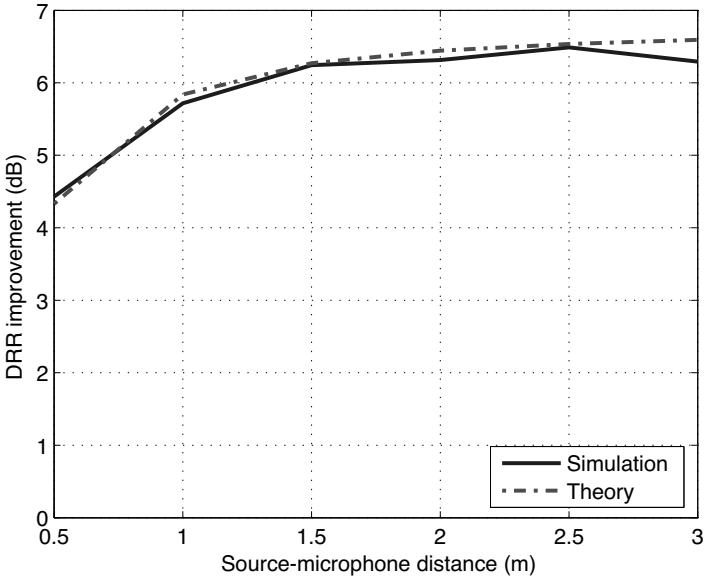


Fig. 2.10 DRR improvement *vs.* source-microphone distance for an array of $M = 5$ microphones

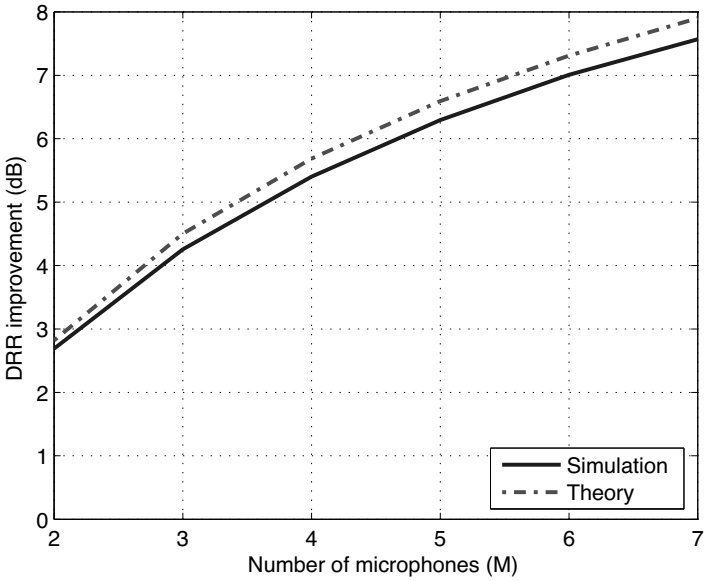


Fig. 2.11 DRR improvement *vs.* number of microphones at a fixed spacing of 2 m

hope that the references within the text will provide sufficient information to satisfy those wishing to work with the techniques or to satisfy their interest through further reading.

Quantitative characterization of reverberation is not a straightforward task, nor is evaluation of the effect of signal processing algorithms on the level of reverberation contained in a speech signal easy. We have considered the task of such evaluation in two sets of circumstances; in one case the AIR is available and in the other it is not. For many researchers, it will be more direct and straightforward to use techniques based on the AIR but, for the cases when the AIR or its estimate are not available, the use of the signal-based measures provides a reasonable strategy for objective measurement. We have also aimed to give a few pointers to methods of subjective testing.

Many speech researchers encounter the subject of speech dereverberation because of their need to improve Automatic Speech Recognition (ASR) in reverberant, and possible also noisy, environments. In this context, evaluation of the level of reverberation, and the ability of a signal processing algorithm to enhance the signal, is normally very effectively achieved by measurement of the error rate obtained in the ASR. Two points of discussion arise here. First, it would be surprising if a dereverberation algorithm that improved ASR performance always made the reverberant speech sound better. Second, evaluating reverberation, and several other degradation types, is more straightforward in signals intended for machines than in the context of human perception.

To end the beginning of this book, we have analysed the important tool that is the delay-and-sum beamformer from the perspective of its dereverberation performance. The DSB is often considered a baseline method and provides us with a firm foundation from which to explore alternative approaches and techniques in the following chapters.

References

1. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
2. Beeson, M.J., Murphy, D.T.: Roomweaver: A digital waveguide mesh based room acoustics research tool. In: Proc. Int. Conf. on Digital Audio Effects, pp. 268–273. Naples, Italy (2004)
3. Beranek, L.L.: Analysis of Sabine and Eyring equations and their application to concert hall audience and chair absorption. J. Acoust. Soc. Am. **120**(3), 1399–1410 (2006)
4. Botteldoore, D.: Finite-difference time-domain simulation of low-frequency room acoustic problems. J. Acoust. Soc. Am. **98**(6), 3302–3308 (1995)
5. Brandstein, M.S., Griebel, S.M.: Nonlinear, model-based microphone array speech enhancement. In: S.L. Gay, J. Benesty (eds.) Acoustic Signal Processing For Telecommunication, pp. 261–279. Kluwer Academic Publishers (2000)
6. Brandstein, M.S., Ward, D.B. (eds.): Microphone arrays: Signal processing techniques and applications, 1 edn. Springer (2001)
7. Deller, J., Proakis, J., Hansen, J.: Discrete-time processing of speech signals. New York: MacMillan (1993)

8. Garofolo, J.: Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database. Tech. rep., National Institute of Standards and Technology (NIST), Gaithersburg,Maryland (1988)
9. Gaubitch, N., Naylor, P.A., Ward, D.B.: On the use of linear prediction for dereverberation of speech. In: Proc. Int. Workshop Acoust. Echo Noise Control, pp. 99–102 (2003)
10. Gaubitch, N.D., Naylor, P.A.: Analysis of the dereverberation performance of microphone arrays. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Eindhoven, The Netherlands (2005)
11. Gillespie, B.W., Malvar, H.S., Florêncio, D.A.F.: Speech dereverberation via maximum-kurtosis subband adaptive filtering. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 6, pp. 3701–3704 (2001)
12. Giner, J., Militello, C., Garcia, A.: Ascertaining confidence within the ray-tracing method. J. Acoust. Soc. Am. **106**(2), 816–822 (1999)
13. Griebel, S.M.: A microphone array system for speech source localization, denoising and dereverberation. Ph.D. thesis, Harvard University, Cambridge, Massachusetts (2002)
14. Griebel, S.M., Brandstein, M.S.: Wavelet transform extrema clustering for multi-channel speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control. Pocono Manor, Pennsylvania (1999)
15. Gustafsson, T., Rao, B.D., Trivedi, M.: Source localization in reverberant environments: modeling and statistical analysis. IEEE Trans. Speech Audio Process. **11**(6), 791803 (2003)
16. Habets, E.A.P.: Single- and multi-microphone speech dereverberation using spectral enhancement. Ph.D. thesis, Technische Universiteit Eindhoven (2007)
17. Huber, R., Kollmeier, B.: PEMO-Q – a new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio, Speech, Lang. Process. **14**(6), 1902–1911 (2006). DOI 10.1109/TASL.2006.883259
18. ITU-T: Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunications Union (ITU-T) (1996)
19. ITU-T: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Recommendation P.862, International Telecommunications Union (ITU-T) (2001)
20. Jetzt, J.J.: Critical distance measurement of rooms from the sound energy spectral response. J. Acoust. Soc. Am. **65**(5), 1204–1211 (1979)
21. Jot, J.M., Cerveau, L., Warusfel, O.: Analysis and synthesis of room reverberation based on a statistical time-frequency model. In: Proc. Audio Eng. Soc. Convention (1997)
22. Kleiner, M., Dalenbäck, B., Svensson, P.: Auralization – an overview. J. Acoust. Soc. Am. **41**(11), 861–875 (1993)
23. Krokstad, A., Strom, S., Sørsdal, S.: Calculating the acoustical room response by the use of a ray tracing technique. J. Sound Vib. **8**, 118–125 (1968)
24. Kulowski, A.: Error investigation for the ray tracing technique. Appl. Acoust. **15**(4), 263–274 (1982)
25. Kuttruff, H.: room acoustics, 4 edn. Taylor & Francis (2000)
26. Kuttruff, K.H., Schroeder, M.R.: On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima. J. Acoust. Soc. Am. **34**(1), 76 – 80 (1962)
27. Laakso, T.I., Valimaki, V., Karjalainen, M., Laine, U.K.: Splitting the unit delay. IEEE Signal Process. Mag. **13**(1), 30–60 (1996)
28. Lindsey, G., Breen, A., Nevard, S.: SPARs archivable actual word databases. Tech. rep., University College London (1987)
29. Moorer, J.: About this reverberation business [computer music]. Computer Music Journal **3**(2), 13–28 (1979)
30. Morgan, D., Benesty, J., Sondhi, M.: On the evaluation of estimated impulse responses. IEEE Signal Processing Lett. **5**(7), 174–176 (1998). DOI 10.1109/97.700920
31. Murphy, D.T., Howard, D.M.: 2-D digital waveguide mesh topologies in room acoustics modelling. In: Proc. Int. Conf. on Digital Audio Effects, pp. 211–216. Verona, Italy (2000)

32. Nelson, P.A., Elliott, S.J.: Active control of sound. Academic, London (1993)
33. Ostashev, V.E. (ed.): Acoustics in moving inhomogeneous media. E and FN Spon (1997)
34. Peterson, P.M.: Simulating the response of multiple microphones to a single acoustic source in a reverberant room. J. Acoust. Soc. Am. **80**(5), 1527–1529 (1986)
35. Picovici, D., Mahdi, A.: Towards non-intrusive speech quality assessment for modern telecommunications. In: First Joint IEI/IEE Symposium of Telecom Systems Research (2001)
36. Pietrzyk, A.: Computer modeling of the sound field in small rooms. In: Proc. of the 15th AES Int. Conf. on Audio, Acoustics and Small Spaces, vol. 2, pp. 24–31. Copenhagen, Denmark (1998)
37. Polack, J.D.: La transmission de l'énergie sonore dans les salles. Thèse de doctorat d'etat, Université du Maine, Le Mans (1988)
38. Polack, J.D.: Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics. Appl. Acoust. **38**(2), 235–244 (1993)
39. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals. Prentice-Hall, Englewood Cliffs, NJ (1978)
40. Radlović, B.D., Williamson, R.C., Kennedy, R.A.: Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Acoust., Speech, Signal Process. **8**(3), 311–319 (2000)
41. Reichardt, W., Lehmann, U.: Raumeindruck als oberbegriff von rumlichkeit und halligkeit, erluterungen des raumeindrucksmasses. Acustica **40**, 174–183 (1978)
42. Sabine, W.C.: Collected papers on acoustics. Peninsula Publishing (1993 (Originally 1921))
43. Savioja, L.: Modeling techniques for virtual acoustics. Doctoral dissertation, Helsinki University of Technology, Espoo, Finland (1999)
44. Savioja, L., Backman, J., Järvinen, A., Takala, T.: Waveguide mesh method for low-frequency simulation of room acoustics. In: Proc. of the 15th Int. Congr. Acoust. (ICA'95), vol. 2, pp. 1–4. Trondheim, Norway (1995)
45. Savioja, L., Rinne, T.J., Takala, T.: Simulation of room acoustics with a 3-D finite difference mesh. In: Proc. Int. Computer Music Conf., pp. 463–466. Denmark (1994)
46. Wang, S., Sekey, A., Gersho, A.: An objective measure for predicting subjective quality of speech coders. IEEE J. Sel. Areas Commun. **10**(5), 819–829 (1992). DOI 10.1109/49.138987
47. Wen, J.Y.C., Gaubitch, N.D., Habets, E.A.P., Myatt, T., Naylor, P.A.: Evaluation of speech dereverberation algorithms using the MARDY database. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Paris, France (2006)
48. Wen, J.Y.C., Naylor, P.A.: An evaluation measure for reverberant speech using tail decay modelling. In: Proc. European Signal Processing Conf. (EUSIPCO). Florence, Italy (2006)
49. Yang, W., Dixon, M., Yantorno, R.: A modified bark spectral distortion measure which uses noise masking threshold. In: IEEE Speech Coding Workshop, pp. 55–56. Pocono Manor (1997)
50. Yegnanarayana, B., Satyanarayana, P.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Speech Audio Process. **8**(3), 267–281 (2000)

Chapter 3
Speech Dereverberation Using Statistical Reverberation Models*

Emanuël A.P. Habets

Abstract In speech communication systems, such as voice-controlled systems, hands-free mobile telephones and hearing aids, the received microphone signals are degraded by room reverberation, ambient noise and other interferences. This signal degradation can decrease the fidelity and intelligibility of speech and the word recognition rate of automatic speech recognition systems.

The reverberation process is often described using deterministic models that depend on a large number of unknown parameters. These parameters are often difficult to estimate blindly and are dependent on the exact spatial position of the source and receiver. In recently emerged speech dereverberation methods, which are feasible in practice, the reverberation process is described using a statistical model. This model depends on smaller number of parameters such as the reverberation time of the enclosure, which can be assumed to be independent of the spatial location of the source and receiver. This model can be utilized to estimate the spectral variance of part of the reverberant signal component. Together with an estimate of the spectral variance of the ambient noise, this estimate can then be used to enhance the observed noisy and reverberant speech.

In this chapter we provide a brief overview of dereverberation methods. We then describe single and multiple microphone algorithms that are able to jointly suppress reverberation and ambient noise. Finally, experimental results demonstrate the beneficial use of the algorithms developed.

Imperial College London, UK

* This research was supported by the Israel Science Foundation (grant no. 1085/05) and by the Technology Foundation STW, Applied Science Division of NWO and the Technology Programme of the Dutch Ministry of Economic Affairs.

3.1 Introduction

Speech signals that are received by a microphone at a distance from the speech source usually contain reverberation, ambient noise and other interferences. Reverberation is the process of multi-path propagation of an acoustic sound from its source to a microphone. The received microphone signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*) and reflections that arrive after the early reverberation (commonly called *late reverberation*). The combination of the direct sound and early reverberation is sometimes referred to as the *early speech component*. Early reverberation mainly contributes to spectral colouration, while late reverberation changes the waveform's temporal envelope as exponentially decaying tails are added at sound offsets. The colouration can be characterized by the spectral deviation σ, which is defined as the standard deviation of the log-amplitude frequency response of the Acoustic Impulse Response (AIR) [46].

For the development of dereverberation algorithms it is of great importance to have a good understanding of the effects of reverberation on speech perception. The reduction in speech intelligibly caused by late reverberation is especially noticeable for non-native listeners [72] and for listeners with hearing impairments [58]. The detrimental effects of reverberation on speech intelligibility have been attributed to two types of masking. Bolt and MacDonald [10] and Nábělek *et al.* [57] found evidence of *overlap-masking*, whereby late reverberation of a preceding phoneme masks a subsequent phoneme, and of *self-masking*, which refers to the time and frequency alterations of an individual phoneme.

In a reverberant room, speech intelligibility initially decreases with increasing source-microphone distance, but beyond the so-called critical distance speech intelligibility is approximately constant. The critical distance is the distance at which the direct-path energy is equal to the energy of all reflections. For an omnidirectional microphone the critical distance D_c is approximately given by [69]

$$D_{\mathrm{c}} = \sqrt{\frac{\ln(10^6)\,V}{4\pi\, c\, T_{60}}}, \tag{3.1}$$

where c is the sound velocity in ms^{-1}, V is the volume of the room in m^3 and T_{60} is the reverberation time in seconds. To obtain sufficiently intelligible speech it is typically recommended that the source-microphone distance is smaller than 0.3 times the critical distance. In a living room with dimensions 7 m $\times$ 5 m $\times$ 3 m and $T_{60} = 0.5$ s, the critical distance $D_c \approx 0.82$ m. Hence, the speech intelligibility would be affected even when the source-microphone distance is larger than 0.25 m.

Consonants play a more significant role in speech intelligibility than vowels. If the consonants are heard clearly, the speech can be understood more easily. In 1971 Peutz [60] proposed a measure called the articulation loss of consonants ($\mathrm{Al}_{\mathrm{cons}}$) that quantifies the reduction in perception of consonants due to reverberation. For distances smaller than the critical distance the measure depends on the source-

microphone distance, the reverberation time, and the volume of the room. Beyond the critical distance the measure depends only on the reverberation time. The speech intelligibility can be increased by decreasing the articulation loss, which can be achieved by decreasing the source-microphone distance or the reverberation time, or by increasing the room volume.

In 1982 Allen [4] reported a formula to predict the *subjective preference* of reverberant speech. The main result is given by the equation

$$P = P_{\max} - \sigma T_{60}, \tag{3.2}$$

where P is the subjective preference in some arbitrary units, $P_{\max}$ is the maximum possible preference, and σ is the spectral deviation in decibels (dB). According to this formula, decreasing either the spectral deviation σ or the reverberation time T_{60} results in an increased subjective preference of reverberant speech.

It would be convenient to assume that reverberation solely reduces intelligibility, but this assumption is incorrect [71]. Strong reflections that arrive shortly after the direct sound actually reinforce the direct sound and are therefore considered useful with regard to speech intelligibility. This reinforcement, which is often referred to as the *precedence effect*, is what makes it easier to hold conversations in closed rooms rather than outdoors.

While investigating the detrimental effects of reverberation on speech, it has become clear that the speech fidelity and intelligibility are mostly degraded by late reverberation. In addition, speech intelligibility is degraded by ambient noise. Therefore, we define the effective noise as the sum of the late reverberant component and the ambient noise component. In this chapter we describe a spectral enhancement method to suppress late reverberation and ambient noise, i.e., to estimate the early speech component. Due to the joint suppression of late reverberation and ambient noise, the effective noise is reduced and the fidelity and intelligibility of speech can be improved.

This chapter is organized as follows. In Sect. 3.2 a short review of dereverberation methods is provided. In Sect. 3.3 two statistical reverberation models are discussed. In Sect. 3.4 we derive a spectral estimator which can be used to jointly suppress late reverberation and ambient noise. In Sect. 3.5 we investigate the possibility of using multiple microphones in conjunction with spectral enhancement techniques for dereverberation. The spectral estimator derived in Sect. 3.4 requires an estimate of the spectral variance of the late reverberant signal component. In Sect. 3.6 such an estimator is derived using a statistical reverberation model. Estimation of the model parameters is discussed in Sect. 3.7. Experimental results that demonstrate the beneficial use of the described dereverberation methods are presented in Sect. 3.8. Finally, a summary and directions for further research are provided in Sect. 3.9.

3.2 Review of Dereverberation Methods

Reverberation reduction processes can be divided into many categories. They may, for example, be divided into single or multi-microphone techniques and into those primarily affecting colouration or those affecting late reverberation. We categorized the reverberation reduction processes depending on whether or not the AIR needs to be estimated. We then obtain two main categories, *viz. reverberation cancellation* and *reverberation suppression*.

3.2.1 Reverberation Cancellation

The first category, i.e., reverberation cancellation, consists of methods known as blind deconvolution. Much research has been undertaken on the topic of blind deconvolution; see [43] and the references therein. Multichannel methods appear particularly interesting because theoretically exact inverse-filtering can be achieved if the AIRs can be estimated and they do not have any common-zeros in the z-plane [56]. To achieve dereverberation without *a priori* knowledge of the room acoustics, many traditional methods assume that the source signal is independent and identically-distributed (i.i.d.). However, the i.i.d. assumption does not hold for speech-like signals. When applying such traditional deconvolution methods to speech, the speech generating process is deconvolved and the resulting speech signal is excessively whitened. Delcroix *et al.* proposed a method that consists of a multichannel equalizer and a compensation filter that reconstructs the colouration of the speech signal that is whitened by the equalizer [21]. Although perfect dereverberation is possible in theory, the method is sensitive to estimation errors of the covariance matrix that is required to compute the equalizer and the compensation filter. Another interesting method was developed by Gürelli and Nikias [33] and explores the null-space of the spatial correlation matrix, calculated from the received signals. It was shown that the null-space of the correlation matrix contains information on the acoustic transfer functions. This method has also potential in the speech processing framework and was extended by Gannot and Moonen [28]. In [44] the speech signal is modelled using a block stationary auto-regressive process while the room acoustics are modelled using an all-pole model. Bayesian parameter estimation techniques were then used to estimate the unknown parameters.

While good results can be achieved the methods in this category suffer from several limitations: (1) they have been shown to be insufficiently robust to small changes in the AIR [63, 73], (2) channels cannot be identified uniquely when they contain common zeros, (3) observation noise causes severe problems, and (4) some methods require knowledge of the order of the unknown system [45]. Detailed treatments on the problems involved are presented in Chaps. 5–7 and 9.

3.2.2 Reverberation Suppression

Methods in the second category, i.e., reverberation suppression, do not require an estimate of the AIR and explicitly exploit the characteristics of speech, the effect of reverberation on speech, or the characteristics of the AIR. Methods based on processing of the Linear Prediction (LP) residual signal belong to this category [30, 32, 78]. The peaks in the LP residual signal correspond to excitation events in voiced speech together with additional random peaks due to reverberation. These random peaks can be suppressed by, for example, averaging adjacent larynx-cycles, as proposed in [30].

Other, so-called, spatial processing methods use multiple microphones placed at different locations. They often use a limited amount of *a priori* knowledge of the AIR such as, for example, the direction of arrival of the desired source. The microphone signals can be processed to enhance or attenuate signals emanating from particular directions. The well-known delay and sum beamformer is a good example of such a method and belongs to the reverberation suppression category.

Recently, spectral enhancement methods have been used for speech dereverberation [37, 39, 41, 42, 49, 74]. Spectral enhancement of noisy speech has been a challenging problem for many researchers for over 30 years and is still an active research area, see, for example, [6, 17, 23, 24] and references therein. Spectral enhancement of noisy speech is often formulated as estimation of speech spectral components from a speech signal degraded by statistically independent additive noise. One of the earlier methods, and perhaps the most well-known method, is spectral subtraction [9, 50]. This method generally results in random narrow-band fluctuations in the residual noise, also known as musical tones, which are annoying and disturbing to the perception of the enhanced signal. Many variations have been developed to cope with musical tones [8, 9, 31, 36, 70]. Spectral subtraction makes minimal assumptions about the signal and noise, and when carefully implemented, it produces enhanced signals that may be acceptable for certain applications. Lebart *et al.* proposed a single-channel speech dereverberation method based on spectral subtraction to reduce the effect of overlap-masking [49]. The method estimates the short-term Power Spectral Density (PSD) of late reverberation based on a statistical reverberation model. This model exploits the fact that the envelope of the AIR decays exponentially and depends on a single parameter that is related to the reverberation time of the room. In [38] the authors showed that the estimated short-term PSD of late reverberation can be improved using multiple microphones. Additionally, the fine-structure of the speech signal is partly restored due to spatial averaging of the received power spectra.

A more advanced spectral enhancement method is the so-called statistical method, which is often designed to minimize the expected value of some distortion measure between the clean and estimated signals [11, 17, 25, 55]. This method requires reliable statistical models for the speech and noise signals, a perceptually meaningful distortion measure and an efficient signal estimator. A statistical speech model and perceptually meaningful distortion measure that are fully appropriate for spectral enhancement have not yet been determined. Hence, the variety of statistical

methods for spectral enhancement differ mainly in the statistical model [15, 25, 55], distortion measure [26, 52, 77] and the particular implementation of the spectral enhancement algorithm [23]. In this chapter we describe a statistical method for the enhancement of noisy and reverberant speech based on a Gaussian model for the speech and interferences and a squared error distortion measure.

3.3 Statistical Reverberation Models

Since the acoustic behaviour in real rooms is too complex to model explicitly, we make use of Statistical Room Acoustics (SRA). SRA provides a statistical description of the transfer function of the system between the source and the microphone in terms of a few key quantities, e.g., source-microphone distance, room volume and reverberation time. The crucial assumption of SRA is that the distribution of amplitudes and phases of individual plane waves, which sum up to produce sound pressure at some point in a room, is so close to random that the sound field is fairly uniformly distributed throughout the room volume. The validity of this description is subjected to a set of conditions that must be satisfied to ensure the accuracy of calculations. Our analysis therefore implicitly assumes that the following conditions hold [48, 63, 73]:

1. The dimensions of the room are relatively large compared to the longest wavelength of the sound of interest.
2. The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V this condition is fulfilled for frequencies that exceed the Schroeder frequency: $f_{\mathrm{g}} = 2000\sqrt{T_{60}/V}$.
3. The source and the microphone are located in the interior of the room, at least a half-wavelength away from the walls.

3.3.1 Polack's Statistical Model

Sabine's [65] major contribution was the introduction of statistical methods to calculate the reverberation time of an enclosed space without considering the details of the space geometry. Schroeder extended Sabine's fundamental work [66, 67] and derived a frequency domain model and a set of statistical properties about the frequency response of the random impulse response.

Polack [61] developed a time-domain model complementing Schroeder's frequency domain model. In this model, an AIR is described as a realization of a non-stationary stochastic process. This model is defined as

$$h(n) = \begin{cases} b(n)\mathrm{e}^{-\bar{\zeta}n}, & \text{for } n \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.3}$$

where n denotes the discrete time index, $b(n)$ is a zero-mean stationary Gaussian noise sequence and $\bar{\zeta}$ is linked to the reverberation time T_{60} through

$$\bar{\zeta} \triangleq \frac{3\ln(10)}{T_{60} f_{\mathrm{s}}}, \tag{3.4}$$

where f_{s} denotes the sampling frequency in Hz. In contrast to the model in (3.3), the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air [48]. This dependency can be taken into account by using a different model for each frequency band. In addition, it should be noted that Polack's statistical reverberation model is only valid in cases for which the distance between the source and the measurement point is greater than the critical distance D_c.

In the early 90s, Polack [62] proved that the most interesting properties of room acoustics are statistical when the number of 'simultaneously' arriving reflections exceeds a limit of about 10. In this case, the echo density is high enough such that the space can be considered to be in a fully diffused or mixed state. The essential requirement is ergodicity, which requires that any given reflection trajectory in the space will eventually reach all points. The ergodicity assumption is determined by the shape of the enclosure and the surface reflection properties. It should be noted that non-ergodic shapes will exhibit much longer mixing times and may not even have an exponential decay. Nevertheless, while it may not be true that all acoustic environments can be modelled using this statistical model, it is sufficiently accurate for most spaces.

The energy envelope of the AIR can be expressed as

$$\mathscr{E}\{h^2(n)\} = \sigma^2 \mathrm{e}^{-2\bar{\zeta}n}, \tag{3.5}$$

where σ^2 denotes the variance of $b(n)$, and $\mathscr{E}\{\cdot\}$ denotes spatial expectation. Here the spatial expectation is defined as the ensemble average over different realizations of the stochastic process in (3.3). Under the assumption that the space is ergodic, we may evaluate the ensemble average in (3.5) by spatial averaging so that different realizations of this stochastic process are obtained by varying either the position of the receiver or the source [47]. Note that the same stochastic process will be observed for all allowable positions (in terms of the third SRA condition) provided that the time origin is defined with respect to the signal emitted by the source and not with respect to the arrival time of the direct sound at the receiver.

3.3.2 Generalized Statistical Model

In many cases the source-microphone distance is smaller than the critical distance D_c, i.e., the Direct to Reverberant Ratio (DRR) is larger than 0 dB. In these cases Polack's statistical model, although useful when the source-microphone distance is larger than the critical distance, is not an accurate model of the AIR. In [39], a

generalized statistical model was proposed, which can be used when the source-microphone distance is smaller than the critical distance. To model the contribution of the direct-path, the AIR $h(n)$ is divided into two segments, *viz.* $h_{\mathrm{d}}(n)$ and $h_{\mathrm{r}}(n)$:

$$h(n) = \begin{cases} h_{\mathrm{d}}(n), & \text{for } 0 \leq n < n_{\mathrm{d}}, \\ h_{\mathrm{r}}(n), & \text{for } n \geq n_{\mathrm{d}}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

The value n_{d} is chosen such that $h_{\mathrm{d}}(n)$ contains the direct-path and $h_{\mathrm{r}}(n)$ contains all reflections. Later we define the parameter n_{d} according to the frame rate of the time-frequency transformation. In practice, the direct-path is deterministic and could be modelled using a Dirac pulse. Unfortunately this would preclude us from creating a statistical model. To be able to model the energy related to the direct-path the following model is proposed:

$$h_{\mathrm{d}}(n) = b_{\mathrm{d}}(n)\mathrm{e}^{-\bar{\zeta}n}, \tag{3.7}$$

where $b_{\mathrm{d}}(n)$ is a white zero-mean Gaussian stationary noise sequence and $\bar{\zeta}$ is linked to the reverberation time T_{60} through (3.4). The reverberant component $h_{\mathrm{r}}(n)$ is described using the following model:

$$h_{\mathrm{r}}(n) = b_{\mathrm{r}}(n)\mathrm{e}^{-\bar{\zeta}n}, \tag{3.8}$$

where $b_{\mathrm{r}}(n)$ is a white zero-mean Gaussian stationary noise sequence. Under the SRA conditions the direct and reverberant component of the AIR are uncorrelated [63]. Therefore, it is reasonable to assume that $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$ are uncorrelated, i.e., $\mathscr{E}\{b_{\mathrm{d}}(n)b_{\mathrm{r}}(n+\tau)\} = 0$ for $\tau \in \mathbb{Z}$.

The energy envelope of $h(n)$ can be expressed as

$$\mathscr{E}\{h^2(n)\} = \begin{cases} \sigma_{\mathrm{d}}^2\mathrm{e}^{-2\bar{\zeta}n}, & \text{for } 0 \leq n < n_{\mathrm{d}} \\ \sigma_{\mathrm{r}}^2\mathrm{e}^{-2\bar{\zeta}n}, & \text{for } n \geq n_{\mathrm{d}} \\ 0, & \text{otherwise,} \end{cases} \tag{3.9}$$

where σ_{d}^2 and σ_{r}^2 denote the variances of $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$, respectively. When $\sigma_{\mathrm{d}}^2 < \sigma_{\mathrm{r}}^2$, the contribution of the direct-path can be neglected. Therefore, it is assumed that $\sigma_{\mathrm{d}}^2 \geq \sigma_{\mathrm{r}}^2$. Note that the generalized statistical model is equivalent to Polack's statistical model in the case $\sigma_{\mathrm{d}}^2 = \sigma_{\mathrm{r}}^2$.

3.4 Single-microphone Spectral Enhancement

In this section the spectral enhancement of a noisy and reverberant microphone signal is discussed. We start by formulating the spectral enhancement problem in Sect. 3.4.1. In Sect. 3.4.2 we show how the spectrum of the early speech component

can be estimated using the Minimum Mean Square Error (MMSE) Log Spectral Amplitude (LSA) estimator proposed by Cohen in [13]. This estimator depends on the so-called *a priori* Signal to Interference Ratio (SIR) that needs to be estimated in practice. In Sect. 3.4.3 we describe how the *a priori* SIR can be estimated.

3.4.1 Problem Formulation

The reverberant signal results from the convolution of the anechoic speech signal and a causal AIR. In this section we assume that the AIR is time-invariant and that its length is infinite. The reverberant speech signal at discrete-time n can be written as

$$z(n) = \sum_{l=-\infty}^{n} s(l)h(n-l). \tag{3.10}$$

To simplify the following discussion it is assumed that the direct sound arrives at time instance n, i.e., the direct-path is modelled by $h(0)$. It should be noted that this assumption can be made without loss of generality. Since our main goal is to suppress late reverberation we split the AIR into two components (see Fig. 3.1) such that

$$h(n) = \begin{cases} 0, & n < 0, \\ h_{\mathrm{e}}(n), & 0 \leq n < n_{\mathrm{e}}, \\ h_{\ell}(n), & n \geq n_{\mathrm{e}}, \end{cases} \tag{3.11}$$

where n_{e} is chosen such that $h_{\mathrm{e}}(n)$ consists of the direct-path and a few early reflections and $h_{\ell}(n)$ consists of all later reflections. The fraction $n_{\mathrm{e}}/f_{\mathrm{s}}$ can be used to define the time instance (relative to the time of arrival of the direct sound) from where the late reverberation is suppressed. Its value can be determined by the listener depending on his or her subjective preference but should be larger than the mixing time of the room, which is defined as the time it takes for initially adjacent sound rays to spread uniformly across the room [61]. In practice, $n_{\mathrm{e}}/f_{\mathrm{s}}$ usually ranges from 30 to 60 ms.

Using (3.11) we can write the microphone signal $x(n)$ as

$$x(n) = \underbrace{\sum_{l=n-n_{\mathrm{e}}+1}^{n} s(l)h_{\mathrm{e}}(n-l)}_{z_{\mathrm{e}}(n)} + \underbrace{\sum_{l=-\infty}^{n-n_{\mathrm{e}}} s(l)h_{\ell}(n-l)}_{z_{\ell}(n)} + v(n), \tag{3.12}$$

where $z_{\mathrm{e}}(n)$ is the early speech component, $z_{\ell}(n)$ denotes the late reverberant speech component, and $v(n)$ denotes the additive ambient noise component. The joint suppression of $z_{\ell}(n)$ and $v(n)$ decreases the effective noise level, and can increase the speech fidelity and intelligibility. Since the response of the first part of the AIR, i.e., $z_{\mathrm{e}}(n)$, remains unaltered we do not reduce the colourations caused by the early reflections.

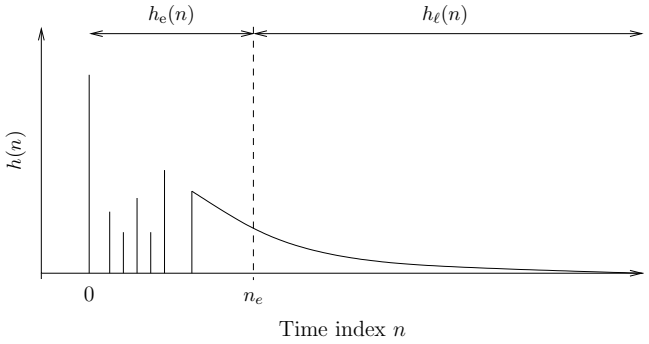


Fig. 3.1 Schematic representation of the acoustic impulse response

Estimating $z_e(n)$ is a challenging problem because both $s(n)$ and $h(n)$ are unknown. Here we formulate the problem of estimating $z_e(n)$, or in other words suppressing $z_\ell(n)$, using spectral enhancement. A block diagram of the spectral enhancement system is depicted in Fig. 3.2. The noisy and reverberant speech signal is denoted by $x(n)$, and is first transformed to the time-frequency domain by applying the short-time Fourier transform (STFT). Specifically,

$$X(\ell,k) = \sum_{n=0}^{K-1} x(n+\ell R)w(n)\mathrm{e}^{-j\frac{2\pi}{K}nk}, \tag{3.13}$$

where $j=\sqrt{-1}$, $w(n)$ is the analysis window of size K, and R is the number of samples separating two successive frames. The spectral component $X(\ell,k)$ can be used to estimate the spectral variance $\lambda_v(\ell,k) = E\{|V(\ell,k)|^2\}$ of the ambient noise and to estimate the spectral variance $\lambda_{z_\ell}(\ell,k) = E\{|Z_\ell(\ell,k)|^2\}$ of the late reverberant signal component $z_\ell(n)$. In the following, we assume that the spectral variance of the ambient noise is slowly time varying. Therefore, the spectral variance $\lambda_v(\ell,k)$ of the ambient noise can be estimated using the algorithm proposed by Martin in [54] or using the Improved Minima Controlled Recursive Averaging (IMCRA) algorithm proposed by Cohen in [14]. In contrast to $\lambda_v(\ell,k)$, the spectral variance $\lambda_{z_\ell}(\ell,k)$ of late reverberant signal component is highly time-varying due to the non-stationarity of the anechoic speech signal. In the application that is considered in this chapter, it is possible to estimate $\lambda_{z_\ell}(\ell,k)$ from the microphone signal. An estimator for $\lambda_{z_\ell}(\ell,k)$ is derived in Sect 3.6. For now, we assume that an estimate of the late reverberant spectral variance is available.

Using statistical signal processing, the spectral enhancement problem can be formulated as deriving an estimator $\hat{Z}_e(\ell,k)$ for the speech spectral coefficients such that the expected value of a certain distortion measure is minimized [17]. Let $\mathcal{H}_1(\ell,k)$ and $\mathcal{H}_0(\ell,k)$ denote the hypotheses for speech presence and absence in the spectral coefficient $Z_e(\ell,k)$, respectively. Such that

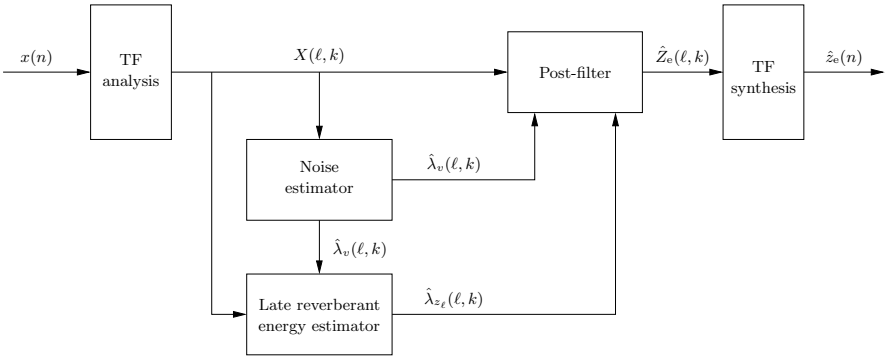


Fig. 3.2 Block diagram of the single-microphone spectral enhancement system for late reverberation and noise suppression

$$\mathcal{H}_1(\ell,k):\ X(\ell,k) = Z_{\mathrm{e}}(\ell,k) + Z_\ell(\ell,k) + V(\ell,k), \tag{3.14}$$

$$\mathcal{H}_0(\ell,k):\ X(\ell,k) = Z_\ell(\ell,k) + V(\ell,k). \tag{3.15}$$

Let $\hat{p}(\ell,k) = P(\mathcal{H}_1(\ell,k))$ denote an estimate for the probability that the desired speech component is present and let $\hat{\lambda}_{z_{\mathrm{e}}}(\ell,k)$ denote an estimate of the variance of the early speech spectral coefficient $Z_{\mathrm{e}}(\ell,k)$ under $\mathcal{H}_1(\ell,k)$. We can now calculate an estimator for $Z_{\mathrm{e}}(\ell,k)$ that minimizes the expected value of the distortion measure given $\hat{p}(\ell,k)$, $\hat{\lambda}_{z_{\mathrm{e}}}(\ell,k)$, the estimated late reverberant spectral variance $\hat{\lambda}_{z_\ell}(\ell,k) = E\{|\hat{Z}_\ell(\ell,k)|^2\}$, the estimated ambient noise spectral variance $\hat{\lambda}_v(\ell,k) = E\{|\hat{V}(\ell,k)|^2\}$ and the spectral coefficient $X(\ell,k)$:

$$\hat{Z}_{\mathrm{e}}(\ell,k) = \underset{\hat{Z}_{\mathrm{e}}(\ell,k)}{\operatorname{argmin}}\, E\left\{d\left(Z_{\mathrm{e}}(\ell,k), \hat{Z}_{\mathrm{e}}(\ell,k)\right)\right\}. \tag{3.16}$$

In the sequel we restrict ourselves to the squared error distortion measure, i.e.,

$$d\left(Z_{\mathrm{e}}(\ell,k), \hat{Z}_{\mathrm{e}}(\ell,k)\right) = \left|g(\hat{Z}_{\mathrm{e}}(\ell,k)) - \tilde{g}\left(Z_{\mathrm{e}}(\ell,k)\right)\right|^2, \tag{3.17}$$

where $g(Z_{\mathrm{e}})$ and $\tilde{g}(Z_{\mathrm{e}})$ are specific functions of Z_{e} that determine the fidelity criterion of the estimator. For the squared error distortion measure, the estimator $\hat{Z}_{\mathrm{e}}(\ell,k)$ is calculated from

$$\begin{aligned} g(\hat{Z}_{\mathrm{e}}(\ell,k)) &= E\left\{\tilde{g}(Z_{\mathrm{e}}(\ell,k)) \,\middle|\, X(\ell,k), \hat{p}(\ell,k)\right\} \\ &= \hat{p}(\ell,k)\, E\left\{\tilde{g}(Z_{\mathrm{e}}(\ell,k)) \,\middle|\, X(\ell,k), \mathcal{H}_1(\ell,k)\right\} \\ &\quad + (1-\hat{p}(\ell,k))\, E\left\{\tilde{g}(Z_{\mathrm{e}}(\ell,k)) \,\middle|\, X(\ell,k), \mathcal{H}_0(\ell,k)\right\}. \end{aligned} \tag{3.18}$$

Finally, given the estimated spectral component $\hat{Z}_{\mathrm{e}}(\ell,k)$ the early speech component $\hat{z}_{\mathrm{e}}(n)$ can be obtained using the inverse STFT,

$$\hat{z}_{\mathrm{e}}(n) = \sum_{\ell} \sum_{k=0}^{K-1} \hat{Z}_{\mathrm{e}}(\ell,k)\tilde{w}(n-\ell R)\mathrm{e}^{j\frac{2\pi}{K}k(n-\ell R)}, \tag{3.19}$$

where $\tilde{w}(n)$ is a synthesis window that satisfies the so-called completeness condition:

$$\sum_{\ell} w(n-\ell R)\tilde{w}(n-\ell R) = \frac{1}{K} \quad \text{for all } n. \tag{3.20}$$

Given analysis and synthesis windows that satisfy (3.20) we can reconstruct $\hat{z}(n)$ from its STFT coefficients $\hat{Z}(\ell,k)$. In practice, a Hamming window is often used for the synthesis window. A reasonable choice for the analysis window is the one with minimum energy [76], given by

$$w(n) = \frac{\tilde{w}(n)}{K\sum_{\ell}\tilde{w}^2(n-\ell R)}. \tag{3.21}$$

The inverse STFT is efficiently implemented using the weighted overlap-add method [20].

3.4.2 MMSE Log-spectral Amplitude Estimator

In the previous Section it was shown that the received microphone signal is degraded by late reverberation and ambient noise. In this section, a spectral amplitude estimator is developed that can be used to estimate the early spectral speech component $Z_{\mathrm{e}}(\ell,k)$ in the presence of late reverberation and ambient noise.

While there are many fidelity criteria that are of interest for speech enhancement it has been found that the MMSE of the log-spectral amplitude is advantageous to other MMSE estimators in the case of noise suppression [17]. The MMSE-LSA estimator is found by using the following functions:

$$g(\hat{Z}_{\mathrm{e}}(\ell,k)) = \log_{\mathrm{e}}(|\hat{Z}_{\mathrm{e}}(\ell,k)|), \tag{3.22}$$

$$\tilde{g}(Z_{\mathrm{e}}(\ell,k)) = \begin{cases} \log_{\mathrm{e}}(|Z_{\mathrm{e}}(\ell,k)|) & \text{under } \mathcal{H}_1(\ell,k) \\ \log_{\mathrm{e}}(G_{\min}(\ell,k)\,|X(\ell,k)|) & \text{under } \mathcal{H}_0(\ell,k). \end{cases} \tag{3.23}$$

The MMSE-LSA estimator is obtained by substituting (3.22) and (3.23) into (3.18). Using a Gaussian model for the spectral coefficients, the MMSE-LSA gain function yields [13]

$$G_{\text{MMSE-LSA}}(\ell,k) = \{G_{\text{LSA}}(\ell,k)\}^{p(\ell,k)}\,\{G_{\min}(\ell,k)\}^{1-p(\ell,k)}, \tag{3.24}$$

where $G_{\text{LSA}}(\ell,k)$ is the LSA gain function derived by Ephraim and Malah [26] and $G_{\min}(\ell,k)$ is the lower bound for the gain when the signal is absent and specifies the maximum amount of suppression in those frames. An efficient estimator for the speech presence probability $\hat{p}(\ell,k)$ was developed in [13]. Let $\xi(\ell,k)$ denote the *a*

priori SIR,

$$\xi(\ell,k) = \frac{\lambda_{z_\mathrm{e}}(\ell,k)}{\lambda_{z_\ell}(\ell,k) + \lambda_v(\ell,k)}, \tag{3.25}$$

and $\gamma(\ell,k)$ denote the *a posteriori* SIR,

$$\gamma(\ell,k) = \frac{|X(\ell,k)|^2}{\lambda_{z_\ell}(\ell,k) + \lambda_v(\ell,k)}. \tag{3.26}$$

Here $X(\ell,k)$ denotes the spectral coefficient of the microphone signal and $\lambda_{z_\mathrm{e}}(\ell,k)$, $\lambda_{z_\ell}(\ell,k)$, and $\lambda_v(\ell,k)$ denote the spectral variances of the early speech component, late reverberation, and ambient noise, respectively. While the *a posteriori* SIR can be calculated directly, the *a priori* SIR cannot because the spectral variance $\lambda_{z_\mathrm{e}}(\ell,k)$ of the early speech component in (3.25) is unobservable. The estimation of the *a priori* SIR is treated in Section 3.4.3.

The LSA gain function depends on the *a posteriori* and *a priori* SIR and is given by [26]

$$G_{\mathrm{LSA}}(\ell,k) = \frac{\xi(\ell,k)}{1+\xi(\ell,k)} \exp\left(\frac{1}{2}\int_{\zeta(\ell,k)}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right), \tag{3.27}$$

where

$$\zeta(\ell,k) = \frac{\xi(\ell,k)}{1+\xi(\ell,k)}\,\gamma(\ell,k). \tag{3.28}$$

To avoid speech distortions $G_{\min}$ is usually set between -12 and -18 dB. However, in practice the late reverberation plus ambient noise needs to be reduced more than 12–18 dB. Therefore, we like to control the maximum suppression of the late reverberant speech component and ambient noise separately. Due to the time-varying nature of the interferences the lower-bound becomes time and frequency dependent. Under the assumption that the interferences are uncorrelated a modified lower-bound is given by

$$G_{\min}(\ell,k) = \frac{G_{\min,z_\ell}\,\hat{\lambda}_{z_\ell}(\ell,k) + G_{\min,v}\,\hat{\lambda}_v(\ell,k)}{\hat{\lambda}_{z_\ell}(\ell,k) + \hat{\lambda}_v(\ell,k)}, \tag{3.29}$$

where $G_{\min,z_\ell}$ and $G_{\min,v}$ are used to control the maximum suppression of late reverberation and ambient noise, respectively. When $G_{\min,z_\ell} = 0$ the late reverberation is suppressed down to the residual level of the ambient noise, as shown in [40]. The results of an informal listening test using stationary ambient noise confirmed that the sound level of the residual interference was stationary in case the modified lower-bound $G_{\min}(\ell,k)$ was used, while the sound level of the residual interference fluctuated when the constant lower bound $G_{\min}$ was used.

An estimate of the early spectral speech component $Z_\mathrm{e}(\ell,k)$ can now be obtained using the amplitude estimate and the phase of the noisy and reverberant spectral coefficient $X(\ell,k)$, i.e.,

$$\hat{Z}_\mathrm{e}(\ell,k) = G_{\text{MMSE-LSA}}(\ell,k)\,X(\ell,k). \tag{3.30}$$

3.4.3 *a priori* SIR Estimator

In this section we focus on the *a priori* SIR estimation. The *a priori* SIR in (3.25) can be written as

$$\frac{1}{\xi(\ell,k)} = \frac{1}{\xi_{z_\ell}(\ell,k)} + \frac{1}{\xi_v(\ell,k)}, \tag{3.31}$$

with

$$\xi_\vartheta(\ell,k) = \frac{\lambda_{z_e}(\ell,k)}{\lambda_\vartheta(\ell,k)}, \tag{3.32}$$

where $\vartheta \in \{z_\ell, v\}$. Hence, the total *a priori* SIR can be calculated using the *a priori* SIRs of each interference separately [34, 35, 40]. By doing this, one gains control over (1) the trade-off between the interference reduction and the distortion of the desired signal, and (2) the *a priori* SIR estimation approach for each interference. In some cases, it might be desirable to reduce one of the two interferences at the cost of larger speech distortion, while reducing the other interference less to avoid distortion. In this Section it is shown how the decision-directed approach, proposed by Ephraim and Malah in [25], can be used to estimate the individual *a priori* SIRs.

In the case when the early speech component and the late reverberant signal are very small, the *a priori* SIRs $\xi_{z_\ell}(\ell,k)$ may be unreliable since $\lambda_{z_e}(\ell,k)$ and $\lambda_{z_\ell}(\ell,k)$ are close to zero. In the following, we assume that there is always a certain amount of ambient noise, i.e., $\lambda_v(\ell,k) > 0$. We propose to calculate $\xi(\ell,k)$ using only the most important and reliable *a priori* SIRs as follows:

$$\xi(\ell,k) = \begin{cases} \xi_v(\ell,k), & 10\log_{10}\left(\frac{\lambda_v(\ell,k)}{\lambda_{z_\ell}(\ell,k)}\right) > \beta^{\mathrm{dB}}, \\ \dfrac{\xi_{z_\ell}(\ell,k)\,\xi_v(\ell,k)}{\xi_{z_\ell}(\ell,k)+\xi_v(\ell,k)}, & \text{otherwise}, \end{cases} \tag{3.33}$$

where the threshold β^{dB} specifies the level difference between $\lambda_v(\ell,k)$ and $\lambda_{z_\ell}(\ell,k)$ in dB. When the noise power level is β^{dB} higher than the late reverberant power level, the total *a priori* SIR, $\xi(\ell,k)$, will be equal to $\xi_v(\ell,k)$. Otherwise $\xi(\ell,k)$ will depend on both $\xi_v(\ell,k)$ and $\xi_{z_\ell}(\ell,k)$.

The decision-directed based estimator [12, 25] is given by

$$\hat{\xi}(\ell,k) = \max\left\{\eta\,\frac{G^2_{\mathrm{LSA}}(\ell-1,k)\,|X(\ell-1,k)|^2}{\lambda_{z_\ell}(\ell,k)+\lambda_v(\ell,k)} + (1-\eta)\,\psi(\ell,k), \xi_{\min}\right\}, \tag{3.34}$$

where $\psi(\ell,k) = \gamma(\ell,k) - 1$ is the *instantaneous* SIR, $\gamma(\ell,k)$ is the *a posteriori* SIR as defined in (3.26), and $\xi_{\min}$ is a lower-bound on the *a priori* SIR that controls the residual interference level when hypothesis $\mathcal{H}_1$ is assumed to be true (i.e., when the desired speech is assumed to be active). The weighting factor η $(0 \le \eta \le 1)$ controls the tradeoff between the amount of noise reduction and distortion [12, 25]. To estimate $\xi_\vartheta(\ell,k)$ we use the following expression:

$$\hat{\xi}_{\vartheta}(\ell,k) = \max\left\{\eta_{\vartheta}\frac{G^2_{\mathrm{LSA}}(\ell-1,k)\,|X(\ell-1,k)|^2}{\lambda_{\vartheta}(\ell-1,k)} + (1-\eta_{\vartheta})\;\psi_{\vartheta}(\ell,k), \xi_{\min,\vartheta}\right\}, \tag{3.35}$$

where

$$\begin{aligned}\psi_{\vartheta}(\ell,k) &= \frac{\lambda_{z_\ell}(\ell,k)+\lambda_v(\ell,k)}{\lambda_{\vartheta}(\ell,k)}\psi(\ell,k)\\ &= \frac{|Y(\ell,k)|^2 - \left[\lambda_{z_\ell}(\ell,k)+\lambda_v(\ell,k)\right]}{\lambda_{\vartheta}(\ell,k)},\end{aligned} \tag{3.36}$$

and $\xi_{\min,\vartheta}$ is the lower bound on the *a priori* SIR $\xi_{\vartheta}(\ell,k)$.

3.5 Multi-microphone Spectral Enhancement

Single-microphone systems only exploit the spectral and temporal diversity of the received signal. Reverberation and most ambient noise sources, of course, also induce spatial diversity. To be able to additionally exploit this diversity, multiple microphones must be used and their outputs must be combined by a suitable spatial processor, e.g., a delay-and-sum beamformer, a filter-and-sum beamformer or an adaptive beamformer. Although spatial processors yield a significant improvement of the speech quality, the reverberation suppression is limited and the noise suppression is insufficient when the noise field is non-coherent or diffuse. In addition to the spatial processor a single-channel post-filter should be used to achieve satisfactory results.

In this section we will elaborate on the use of multiple microphone signals for speech dereverberation. In Sect. 3.5.1 we formulate the multi-microphone speech dereverberation problem. In Sect. 3.5.2 we describe two multi-microphone speech enhancement systems for ambient noise and reverberation suppression. In Sect. 3.5.3 we propose a method to enhance the speech presence probability estimation when multiple microphone signals are available.

3.5.1 Problem Formulation

In Sect. 3.4 we exploited the spectral and temporal diversity of the received signal to estimate the early speech component using a single microphone signal. When the signals of multiple microphones are combined using a suitable spatial processor it is possible to 'focus' on the desired source. The effect of early and late reflections can be suppressed to a degree depending on the spatial processor employed.

The reverberant signal at the m^{th} microphone results from the convolution of the anechoic speech signal $s(n)$ and a causal AIR $h_m(n)$. Here we assume that the AIR is time-invariant and that its length is infinite. The reverberant speech signal at

discrete-time n can be written as

$$z_m(n) = \sum_{l=0}^{\infty} h_m(l)\, s(n-l). \tag{3.37}$$

The m^{th} microphone signal is given by

$$x_m(n) = z_m(n) + v_m(n), \tag{3.38}$$

where $v_m(n)$ denote the additive ambient noise received by the m^{th} microphone.

In the STFT domain we can write (3.38) as

$$X_m(\ell,k) = Z_{\text{e},m}(\ell,k) + Z_{\ell,m}(\ell,k) + V_m(\ell,k), \tag{3.39}$$

where $Z_{\text{e},m}(\ell,k)$, $Z_{\ell,m}(\ell,k)$, and $V_m(\ell,k)$ denote the early and late spectral speech components and the ambient noise at the m^{th} microphone, respectively.

Our objective is to obtain an estimate of the early speech component without using detailed knowledge of the AIRs. Instead of estimating $Z_{\text{e},m}(\ell,k)$ with $m \in \{1,\ldots,M\}$, we propose to estimate a spatially filtered version of all early speech components.

3.5.2 Two Multi-microphone Systems

In this section we describe two multi-microphone systems that can be used to suppress ambient noise and reverberation. The first system consists of a Minimum Variance Distortionless Response (MVDR) beamformer followed by a single-channel post-filter. The second system consists of a non-linear spatial processor followed by a single-channel post-filter that was especially designed for speech dereverberation in [39].

3.5.2.1 MVDR Beamformer and Single-channel MMSE Estimator

This multi-microphone system consists of two stages. First, an MVDR beamformer is applied to the microphone signals. Second, a single-channel MMSE estimator is applied to the output of the MVDR beamformer.

Let us define $\mathbf{X}(\ell,k) = [X_1(\ell,k), X_2(\ell,k), \ldots, X_M(\ell,k)]^T$ and $\mathbf{V}(\ell,k) = [V_1(\ell,k), V_2(\ell,k), \ldots, V_M(\ell,k)]^T$. The MVDR filter, denoted by $\mathbf{W}(\ell,k) = [W_1(\ell,k), W_2(\ell,k), \ldots, W_M(\ell,k)]^T$, is found by solving the following minimization problem:

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \underset{\mathbf{W}(k)}{\operatorname{argmin}} \left\{ (\mathbf{W}(k))^H \mathbf{\Lambda}_{\mathbf{VV}}(\ell,k)\mathbf{W}(k) \right\}$$
$$\text{subject to } (\mathbf{W}(k))^H \mathbf{C}(k) = 1, \quad (3.40)$$

where $(\cdot)^H$ denotes the Hermitian transpose, $\mathbf{\Lambda}_{\mathbf{VV}}(\ell,k) = E\{\mathbf{V}(\ell,k)\mathbf{V}^H(\ell,k)\}$ denotes the spatial PSD matrix of the noise, and $\mathbf{C}(k)$ denotes a pre-defined constraint column vector of length M.

A major question remains how to define the constraint $\mathbf{C}(k)$ and thereby the signal which is undistorted by the MVDR beamformer. One solution would be to estimate the reverberant speech component $Z_m(\ell,k)$ for $m \in \{1,\ldots,M\}$ (see, for example, [27]). In this case, the beamformer only reduces noise (and therefore no reverberation). Here we chose to align the direct sound signals of the desired source at the output of the MVDR beamformer. Due to the spatial directivity of the beamformer the spectral coloration induced by early reflections is slightly reduced.

Let us assume that the desired source is located in the far-field, such that the propagation of the direct sound can be modelled by $\mathbf{H}_{\mathrm{d}}(k) = \mathrm{e}^{-j\omega_k\tau_1}\tilde{\mathbf{H}}_{\mathrm{d}}(k)$, where $\tilde{\mathbf{H}}_{\mathrm{d}}(k) = [1, \mathrm{e}^{-j\omega_k\tau_{12}}, \ldots, \mathrm{e}^{-j\omega_k\tau_{1M}}]^T$, $\omega_k = 2\pi f_{\mathrm{s}}k/K$, τ_1 denotes the propagation time of the desired source signal to the first microphone and τ_{1m} $(2 \leq m \leq M)$ denotes the relative delay [also known as time difference of arrival (TDOA)] of the desired source signal between the m^{th} and the first microphone. The aim of the constraint of the MVDR beamformer is to align the direct-paths of the desired source at the output of the MVDR beamformer. Therefore, the constraint vector $\mathbf{C}(k)$ can be defined as

$$\mathbf{C}(k) = \tilde{\mathbf{H}}_{\mathrm{d}}(k). \quad (3.41)$$

Estimation of the TDOAs is beyond the scope of this chapter in which we assume that the TDOAs are known.

The solution of the minimization problem in (3.40) is given by

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \frac{\mathbf{\Lambda}_{\mathbf{VV}}^{-1}(\ell,k)\mathbf{C}(k)}{\mathbf{C}^H(k)\mathbf{\Lambda}_{\mathbf{VV}}^{-1}(\ell,k)\mathbf{C}(k)}. \quad (3.42)$$

The output of the MVDR beamformer is given by

$$\begin{aligned} Q(\ell,k) &= (\mathbf{W}_{\mathrm{MVDR}}(\ell,k))^H \mathbf{X}(\ell,k) \\ &= Q_z(\ell,k) + Q_v(\ell,k), \end{aligned} \quad (3.43)$$

where $Q_z(\ell,k)$ and $Q_v(\ell,k)$ denote the residual reverberant and noise component at the beamformer's output. The spectral variance of $Q(\ell,k)$ is given by

$$\lambda_q(\ell,k) = E\{Q(\ell,k)(Q(\ell,k))^*\} \quad (3.44)$$
$$= \lambda_{q_z}(\ell,k) + \lambda_{q_v}(\ell,k), \quad (3.45)$$

where $(\cdot)^*$ denotes the complex conjugate, $\lambda_{q_z}(\ell,k)$ and $\lambda_{q_v}(\ell,k)$ denote the spectral variances of the residual reverberant and noise component at the beamformer's output. In addition, we can express $\lambda_{q_z}(\ell,k)$ as

$$\begin{aligned}\lambda_{q_z}(\ell,k) &= E\{Q_z(\ell,k)(Q_z(\ell,k))^*\} \\ &= \lambda_{q_e}(\ell,k) + \lambda_{q_\ell}(\ell,k),\end{aligned} \tag{3.46}$$

where $\lambda_{q_e}(\ell,k)$ and $\lambda_{q_\ell}(\ell,k)$ denote the residual early and late reverberation at the output of the beamformer. The spectral variance of the noise at the output of the MVDR beamformer is given by

$$\lambda_{q_v}(\ell,k) = \frac{1}{\mathbf{C}^H(k)\mathbf{\Lambda}_{\mathbf{VV}}^{-1}(\ell,k)\mathbf{C}(k)}. \tag{3.47}$$

Assuming that the residual early and late reverberant signal components are mutually uncorrelated we can reduce the residual late reverberation at the output of the MVDR beamformer using a spectral enhancement technique.

Let us now consider the case in which the ambient noise field is spatially white, i.e., $\mathbf{\Lambda}_{\mathbf{VV}}(\ell,k) = \sigma_v^2\,\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. In this case the MVDR beamformer reduces to the well-known delay and sum beamformer, i.e.,

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \frac{1}{M}\tilde{\mathbf{H}}_d(\ell,k). \tag{3.48}$$

Although the output of the beamformer is not completely dereverberated the signal will contain less reverberation than any one of the observed microphone signals. Using statistical room acoustics, Gaubitch and Naylor derived an analytic expression to calculate the DRR improvement of the delay and sum beamformer compared to the best microphone [29]. Their result demonstrates that the reverberation reduction of the delay and sum beamformer is limited, especially when the source-microphone distance is larger than the critical distance.

Here we employ a single-channel MMSE log spectral amplitude estimator as described in Sect. 3.4 to estimate the residual early speech component at the beamfomer's output. In order to compute the LSA gain function (3.27) we redefine the *a priori* and *a posteriori* SIR as

$$\xi(\ell,k) = \frac{\lambda_{q_e}(\ell,k)}{\lambda_{q_\ell}(\ell,k) + \lambda_{q_v}(\ell,k)} \tag{3.49}$$

and

$$\gamma(\ell,k) = \frac{|Q(\ell,k)|^2}{\lambda_{q_\ell}(\ell,k) + \lambda_{q_v}(\ell,k)}, \tag{3.50}$$

respectively. The spectral variance $\lambda_{q_v}(\ell,k)$ of the residual noise can be estimated either by estimating $\mathbf{\Lambda}_{\mathbf{VV}}(\ell,k)$ during noise only periods and using (3.47) or by using a minimum statistics approach [14, 54]. The late reverberant spectral variance $\lambda_{q_\ell}(\ell,k)$ can be obtained from $Q(\ell,k)$ in a similar way to how $\lambda_{z_\ell}(\ell,k)$ can be obtained from $Z(\ell,k)$, as described in Sect. 3.6.

3.5.2.2 Non-linear Spatial Processor

In [39] it was shown that the output signal of the delay and sum beamformer may contain undesired signal components that result from the spatial correlation between the acoustic channels. The spatial correlation mainly causes problems at low frequencies and becomes more severe when the inter-microphone distance is small. To avoid the creation of these undesired components, a non-linear spatial processor was proposed that can be used when the noise field is spatially white. The spatial processor computes the amplitude and phase spectrum independently. Firstly, the observed spectra are delayed according to the DOA of the desired source. Secondly, the amplitude spectrum is computed from the squared value of the average PSDs:

$$Q(\ell,k) = \left(\frac{1}{M} \sum_{m=1}^{M} \left| X_m(\ell,k) \mathrm{e}^{j\omega_k \tau_{1m}} \right|^2 \right)^{\frac{1}{2}}, \tag{3.51}$$

where τ_{1m} denotes the TDOA of the desired source signal between the m^{th} and the first microphone (by definition $\tau_{11} = 0$). Finally, the phase spectrum is computed by averaging the phase spectra of the properly delayed signals:

$$\varphi(\ell,k) = \arg \left\{ \frac{1}{M} \sum_{m=1}^{M} X_m(\ell,k) \mathrm{e}^{j\omega_k \tau_{1m}} \right\}. \tag{3.52}$$

It should be noted that the phase spectrum is equal to the phase spectrum of the delay and sum beamformer. The output of the non-linear spatial processor is given by

$$Y_{\text{NL}}(\ell,k) = Q(\ell,k) \mathrm{e}^{j\varphi(\ell,k)}. \tag{3.53}$$

Due to the averaging of the PSDs the proposed spatial processor is unable to reduce any noise. The PSD of the noise in $Y_{\text{NL}}(\ell,k)$ is given by $\frac{1}{M} \sum_{m=1}^{M} |V_m(\ell,k)|^2$.

We can now apply the single-microphone spectral enhancement algorithm that was described in Sect. 3.4 to $Y_{\text{NL}}(\ell,k)$. The spectral variance $\lambda_{z_\ell}(\ell,k)$ of the late reverberant speech component can be estimated using $Y_{\text{NL}}(\ell,k)$ in a way similar to how $\lambda_{z_\ell}(\ell,k)$ can be estimated from $X(\ell,k)$. Using statistical room acoustics it can be shown that the expected value of the spatially averaged acoustic transfer functions is white. Since the statistical reverberation models in Sect. 3.3 are based on this assumption, the result obtained sounds better than the single-microphone spectral enhancement. Furthermore, due to the spatial averaging, the spectral colouration that is caused by the early reflections is slightly reduced.

3.5.3 Speech Presence Probability Estimator

In order to compute the MMSE-LSA gain function (3.24) we require an estimate of the *a posteriori* speech presence probability $p(\ell,k)$. The *a posteriori* speech pres-

ence probability $p(\ell,k)$ can be obtained from Bayes' rule, which, under a Gaussian model for the spectral coefficients, reduces to [13]

$$p(\ell,k) = \left\{1 + \frac{1-p(\ell,k|\ell-1)}{p(\ell,k|\ell-1)}\,(1+\xi(\ell,k))\exp\left(-\zeta(\ell,k)\right)\right\}^{-1}, \tag{3.54}$$

where $p(\ell,k|\ell-1)$ denotes the *a priori* speech presence probability, $\xi(\ell,k)$ is the *a priori* SIR and $\zeta(\ell,k)$ is defined in (3.28). In this section we develop an efficient estimator for the *a priori* speech presence probability $p(\ell,k|\ell-1)$, which exploits the strong correlation of speech presence in neighbouring frequency bins of consecutive frames and the strong spatial coherence of the desired signal.

We propose to estimate the *a posteriori* speech presence probability using four probabilities that are obtained using a soft-decision approach. Three probabilities, i.e., $P_{\text{local}}(\ell,k)$, $P_{\text{global}}(\ell,k)$, and $P_{\text{frame}}(\ell)$, are proposed by Cohen in [13], and are based on the time-frequency distribution of the estimated *a priori* SIR, $\xi(\ell,k)$. These probabilities reflect the strong correlation of speech presence in neighbouring frequency bins of consecutive frames. Since the spatial coherence of the desired direct sound is much larger than the spatial coherence of the reverberant sound, we propose to relate the fourth probability, denoted by $P_{\text{spatial}}(\ell,k)$, to the spatial coherence of the received signals. In [42] we proposed to determine $P_{\text{spatial}}(\ell,k)$ using Mean Square Coherence (MSC). Firstly, we smooth the MSC estimate in time and frequency to reduce its variance. Secondly, we map the MSC value to the probability $P_{\text{spatial}}(\ell,k)$. The latter can easily be achieved since the MSC value lies between zero and one.

The MSC is defined as

$$\Phi_{\text{MSC}}(\ell,k) \triangleq \frac{|\lambda_{x_{21}}(\ell,k)|^2}{\lambda_{x_1}(\ell,k)\lambda_{x_2}(\ell,k)}, \tag{3.55}$$

where $\lambda_{x_{21}}(\ell,k) = E\{X_2(\ell,k)(X_1(\ell,k))^*\}$ denotes the cross spectral density, and $\lambda_{x_1}(\ell,k)$ and $\lambda_{x_2}(\ell,k)$ are the power spectral densities. In addition, we know that $0 \leq \Phi_{\text{MSC}}(\ell,k) \leq 1$.

Let η $(0 \leq \eta_{\text{s}} \leq 1)$ denote a smoothing parameter. Then, the power and cross spectral density are estimated using

$$\hat{\lambda}_{x_i}(\ell,k) = \eta_{\text{s}}\hat{\lambda}_{x_i}(\ell-1,k) + (1-\eta_{\text{s}})|X_i(\ell,k)|^2, \quad i \in \{1,2\} \tag{3.56}$$

and

$$\hat{\lambda}_{x_{21}}(\ell,k) = \eta_{\text{s}}\hat{\lambda}_{x_{21}}(\ell-1,k) + (1-\eta_{\text{s}})X_2(\ell,k)(X_1(\ell,k))^*, \tag{3.57}$$

respectively. The MSC is further smoothed over different frequencies using

$$\tilde{\Phi}_{\text{MSC}}(\ell,k) = \sum_{i=-w_{\text{MSC}}}^{w_{\text{MSC}}} b(i)\Phi_{\text{MSC}}(\ell,k+i), \tag{3.58}$$

where $b(i)$ are the coefficients of a normalized window ($\sum_{i=-w_{\mathrm{MSC}}}^{w_{\mathrm{MSC}}} b(i) = 1$) of size $2w_{\mathrm{MSC}}+1$ that determine the frequency smoothing. In the case when more than two microphone signals are available one could average the MSC over different microphone pairs (with equal inter-microphone distance) to improve the estimation procedure even further.

The spatial speech presence probability $\hat{P}_{\mathrm{spatial}}(\ell,k)$ is related to (3.58) by

$$\hat{P}_{\mathrm{spatial}}(\ell,k) = \begin{cases} 0, & \text{for } \tilde{\Phi}_{\mathrm{MSC}}(\ell,k) \leq \Phi_{\min}, \\ 1, & \text{for } \tilde{\Phi}_{\mathrm{MSC}}(\ell,k) \geq \Phi_{\max}, \\ \frac{\tilde{\Phi}_{\mathrm{MSC}}(\ell,k)-\Phi_{\min}}{\Phi_{\max}-\Phi_{\min}}, & \text{otherwise}, \end{cases} \tag{3.59}$$

where $\Phi_{\min}$ and $\Phi_{\max}$ are the minimum and maximum threshold values for $\tilde{\Phi}_{\mathrm{MSC}}(\ell,k)$, respectively.

Finally, an estimate of the *a priori* speech presence probability is obtained by

$$\hat{p}(\ell,k|\ell-1) = \hat{P}_{\mathrm{local}}(\ell,k)\hat{P}_{\mathrm{global}}(\ell,k)\hat{P}_{\mathrm{frame}}(l)\hat{P}_{\mathrm{spatial}}(\ell,k). \tag{3.60}$$

3.6 Late Reverberant Spectral Variance Estimator

In this section we derive a spectral variance estimator for the late reverberant spectral component, $Z_\ell(\ell,k)$, using the generalized statistical reverberation model described in Sect. 3.3.

Before the spectral variance $\lambda_{z_\ell}(\ell,k) = E\{|Z_\ell(\ell,k)|^2\}$ can be estimated, we need to obtain an estimate of the spectral variance of the reverberant spectral component $Z(\ell,k)$ denoted by $\lambda_z(\ell,k)$. Assuming that the spectral coefficients of the reverberant signal and the noise are mutually independent Gaussian random variables, an estimate of the spectral variance $\lambda_z(\ell,k)$ can be obtained by calculating the following conditional expectation:

$$\begin{aligned} \hat{\lambda}_z(\ell,k) &= E\{|Z(\ell,k)|^2 \,|\, X(\ell,k)\} \\ &= |G_{\mathrm{SP}}(\ell,k)X(\ell,k)|^2, \end{aligned} \tag{3.61}$$

where $G_{\mathrm{SP}}(\ell,k)$ denotes the MMSE spectral power gain function. This gain function is given by [3]

$$G_{\mathrm{SP}}(\ell,k) = \frac{\xi_{\mathrm{SP}}(\ell,k)}{1+\xi_{\mathrm{SP}}(\ell,k)} \left(\frac{1}{\gamma_{\mathrm{SP}}(\ell,k)} + \frac{\xi_{\mathrm{SP}}(\ell,k)}{1+\xi_{\mathrm{SP}}(\ell,k)} \right), \tag{3.62}$$

where

$$\xi_{\mathrm{SP}}(\ell,k) = \frac{\lambda_z(\ell,k)}{\lambda_v(\ell,k)} \tag{3.63}$$

and

$$\gamma_{\mathrm{SP}}(\ell,k) = \frac{|X(\ell,k)|^2}{\lambda_v(\ell,k)} \tag{3.64}$$

denote the *a priori* and *a posteriori* SIRs, respectively. The *a priori* SIR is estimated using the decision-directed approach proposed by Ephraim and Malah [25]. Estimates of the spectral variance, $\lambda_v(\ell,k)$, of the noise in the received signal $x(n)$ can be estimated using so-called minimum statistics approaches [14, 54].

In order to derive an estimator for the spectral variance of the late reverberant signal component $z_\ell(n)$ we start by analyzing the autocorrelation of the reverberant signal $z(n)$. The autocorrelation of the reverberant signal $z(n)$ at discrete time n and lag τ for a fixed source-microphone configuration is defined as

$$r_{zz}(n,n+\tau;h) = E\{z(n)z(n+\tau)\}, \tag{3.65}$$

where $E\{\cdot\}$ denotes ensemble averaging. Using (3.37), we have for one realization of h,

$$\begin{aligned} r_{zz}(n,n+\tau;h) = &\sum_{l=n-n_{\mathrm{d}}+1}^{n} \sum_{l'=n-n_{\mathrm{d}}+1+\tau}^{n+\tau} E\{s(l)s(l')\}h_{\mathrm{d}}(n-l)h_{\mathrm{d}}(n+\tau-l') \\ &+ \sum_{l=-\infty}^{n-n_{\mathrm{d}}} \sum_{l'=-\infty}^{n-n_{\mathrm{d}}+\tau} E\{s(l)s(l')\}h_{\mathrm{r}}(n-l)h_{\mathrm{r}}(n+\tau-l'). \end{aligned} \tag{3.66}$$

Using (3.6)–(3.8) and the fact that $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$ consist of a zero-mean white Gaussian noise sequence, it follows that

$$\mathscr{E}\{h_{\mathrm{d}}(n-l)h_{\mathrm{d}}(n+\tau-l')\} = \sigma_{\mathrm{d}}^2 \mathrm{e}^{-2\bar{\zeta}n} \mathrm{e}^{\bar{\zeta}(l+l'-\tau)} \delta(l-l'+\tau), \tag{3.67}$$

and

$$\mathscr{E}\{h_{\mathrm{r}}(n-l)h_{\mathrm{r}}(n+\tau-l')\} = \sigma_{\mathrm{r}}^2 \mathrm{e}^{-2\bar{\zeta}n} \mathrm{e}^{\bar{\zeta}(l+l'-\tau)} \delta(l-l'+\tau), \tag{3.68}$$

where $\delta(\cdot)$ denotes the Kronecker delta function. It should be noted that $\mathscr{E}\{b_{\mathrm{d}}(n)b_{\mathrm{r}}(n+\tau)\} = 0$ implies that $\mathscr{E}\{h_{\mathrm{d}}(n)h_{\mathrm{r}}(n+\tau)\} = 0$. Under the assumption that the stochastic processes h and s are mutually independent the spatially averaged autocorrelation results in

$$\begin{aligned} r_{zz}(n,n+\tau) &= \mathscr{E}\{r_{zz}(n,n+\tau;h)\} \\ &= r_{z_{\mathrm{d}}z_{\mathrm{d}}}(n,n+\tau) + r_{z_{\mathrm{r}}z_{\mathrm{r}}}(n,n+\tau), \end{aligned} \tag{3.69}$$

with

$$r_{z_{\mathrm{d}}z_{\mathrm{d}}}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=n-n_{\mathrm{d}}+1}^{n} E\{s(l)s(l+\tau)\}\, \sigma_{\mathrm{d}}^2 \mathrm{e}^{2\bar{\zeta}l}, \tag{3.70}$$

and

$$r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=-\infty}^{n-n_\mathrm{d}} E\{s(l)s(l+\tau)\}\, \sigma_\mathrm{r}^2 \mathrm{e}^{2\bar{\zeta}l} \tag{3.71}$$

$$= \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=n-2n_\mathrm{d}+1}^{n-n_\mathrm{d}} E\{s(l)s(l+\tau)\}\, \sigma_\mathrm{r}^2 \mathrm{e}^{2\bar{\zeta}l} + \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=-\infty}^{n-2n_\mathrm{d}} E\{s(l)s(l+\tau)\}\, \sigma_\mathrm{r}^2 \mathrm{e}^{2\bar{\zeta}l}. \tag{3.72}$$

The first term in (3.69) depends on the direct signal between time $n-n_\mathrm{d}+1$ and n, and the second depends on the reverberant signal.

Let us consider the spatially averaged autocorrelation at time $n-n_\mathrm{d}$:

$$r_{zz}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) = r_{z_\mathrm{d}z_\mathrm{d}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) + r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau), \tag{3.73}$$

with

$$r_{z_\mathrm{d}z_\mathrm{d}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) = \sigma_\mathrm{d}^2\, \mathrm{e}^{-2\bar{\zeta}(n-n_\mathrm{d})} \sum_{l=n-2n_\mathrm{d}+1}^{n-n_\mathrm{d}} E\{s(l)s(l+\tau)\}\mathrm{e}^{2\bar{\zeta}l}, \tag{3.74}$$

and

$$r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) = \sigma_\mathrm{r}^2\, \mathrm{e}^{-2\bar{\zeta}(n-n_\mathrm{d})} \sum_{l=-\infty}^{n-2n_\mathrm{d}} E\{s(l)s(l+\tau)\}\mathrm{e}^{2\bar{\zeta}l}. \tag{3.75}$$

Using (3.74) and (3.75) the term $r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau)$ can be expressed as

$$r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau) = \kappa\, \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}} r_{z_\mathrm{d}z_\mathrm{d}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) + \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}}\, r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau), \tag{3.76}$$

with $\kappa = \sigma_\mathrm{r}^2/\sigma_\mathrm{d}^2$. Here $\kappa \leq 1$, since it is assumed that $\sigma_\mathrm{d}^2 \geq \sigma_\mathrm{r}^2$. Using (3.73) we can rewrite (3.76) as

$$r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}}\,(1-\kappa)\, r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) + \kappa\, \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}}\, r_{zz}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau). \tag{3.77}$$

The late reverberant component can now be obtained using

$$r_{z_\ell z_\ell}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}(n_\mathrm{e}-n_\mathrm{d})} r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{e}+n_\mathrm{d},n-n_\mathrm{e}+n_\mathrm{d}+\tau). \tag{3.78}$$

Note that for $\kappa = 1$, i.e., $\sigma_\mathrm{d}^2 = \sigma_\mathrm{r}^2$, (3.77) and (3.78) result in

$$r_{z_\ell z_\ell}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n_\mathrm{e}} r_{zz}(n-n_\mathrm{e},n-n_\mathrm{e}+\tau). \tag{3.79}$$

Given an estimate of the reverberation time T_{60}, the parameter $\bar{\zeta}$ can be calculated using (3.4). The parameter $\kappa = \sigma_{\mathrm{r}}^2/\sigma_{\mathrm{d}}^2$ is related to the DRR, which is defined as

$$\frac{E_{\mathrm{d}}}{E_{\mathrm{r}}} = \frac{\sum_{l=0}^{n_{\mathrm{d}}} h^2(l)}{\sum_{l=n_{\mathrm{d}}+1}^{\infty} h^2(l)}. \tag{3.80}$$

It should be noted that the DRR can be estimated directly from the AIR using (3.80). However, in many practical situations the AIR is not known in advance. Therefore, we will discuss the blind estimation of the reverberation time T_{60} and κ in Section 3.7. Using the model in (3.6) the direct and reverberant energy can be expressed as

$$E_{\mathrm{d}} = \sum_{l=0}^{n_{\mathrm{d}}} \sigma_{\mathrm{d}}^2 \mathrm{e}^{-2\bar{\zeta}l} = \frac{\sigma_{\mathrm{d}}^2}{2\bar{\zeta}}\left(1 - \mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}\right) \tag{3.81}$$

and

$$E_{\mathrm{r}} = \sum_{l=n_{\mathrm{d}}+1}^{\infty} \sigma_{\mathrm{r}}^2 \mathrm{e}^{-2\bar{\zeta}l} = \frac{\sigma_{\mathrm{r}}^2}{2\bar{\zeta}} \mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}, \tag{3.82}$$

respectively, where σ_{d}^2 and σ_{r}^2 denote the variances of $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$, respectively. Now the parameter κ can be expressed in terms of E_{d} and E_{r}:

$$\kappa = \frac{\sigma_{\mathrm{r}}^2}{\sigma_{\mathrm{d}}^2} = \frac{1 - \mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}}{\mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}} \frac{E_{\mathrm{r}}}{E_{\mathrm{d}}}. \tag{3.83}$$

In general the DRR is frequency dependent, as shown in [48]. Hence, to improve the accuracy of the model we propose to make κ frequency dependent. Furthermore, we should keep in mind that the DRR, and thus κ, depends on the distance between the source and microphone. Therefore, spatial averaging can only be performed over those microphone signals that have the same source-microphone distance.

In practice the signals can be considered as stationary over periods of time that are short compared to the reverberation time T_{60}. This is justified by the fact that the exponential decay is very slow and that speech is quasi-stationary. We consider that $n_{\mathrm{e}} \ll T_{60} f_{\mathrm{s}}$ and that $n_{\mathrm{e}}/f_{\mathrm{s}}$ is larger than the time span over which the speech signal can be considered stationary, which is usually around 20–40 ms [22]. In the following we assume that n_{d} is equal to the number of samples separating two successive STFT frames, denoted by R. Under these assumptions and by taking the frequency dependency of κ and $\bar{\zeta}$ into account, the counterparts of (3.77) and (3.78) in terms of the spectral variances are:

$$\lambda_{z_{\mathrm{r}}}(\ell,k) = \mathrm{e}^{-2\bar{\zeta}(k)R}\left(1 - \kappa(k)\right)\lambda_{z_{\mathrm{r}}}(\ell-1,k) + \kappa(k)\,\mathrm{e}^{-2\bar{\zeta}(k)R}\lambda_z(\ell-1,k), \tag{3.84}$$

and

$$\lambda_{z_\ell}(\ell,k) = \mathrm{e}^{-2\bar{\zeta}(k)(n_{\mathrm{e}}-R)}\lambda_{z_{\mathrm{r}}}(\ell - \frac{n_{\mathrm{e}}}{R} + 1, k). \tag{3.85}$$

Note that the value n_e should be chosen such that n_e/R is an integer value.

By substituting $\lambda_z(\ell,k) = \lambda_{z_d}(\ell,k) + \lambda_{z_r}(\ell,k)$ in (3.84) and rearranging the terms we obtain

$$\lambda_{z_r}(\ell,k) = \mathrm{e}^{-2\bar{\xi}(k)R}\lambda_{z_r}(\ell-1,k) + \kappa(k)\mathrm{e}^{-2\bar{\xi}(k)R}\lambda_{z_d}(\ell-1,k). \tag{3.86}$$

Using (3.83) we obtain

$$\lambda_{z_r}(\ell,k) = \mathrm{e}^{-2\bar{\xi}(k)R}\lambda_{z_r}(\ell-1,k) + \frac{E_r}{E_d}\left(1-\mathrm{e}^{-2\bar{\xi}(k)R}\right)\lambda_{z_d}(\ell-1,k). \tag{3.87}$$

This equation shows that the spectral variance of the reverberant signal component at time frame ℓ consists of $\mathrm{e}^{-2\bar{\xi}(k)R}$ times the spectral variance of the reverberant signal component at time frame $\ell-1$ and $\frac{E_r}{E_d}\left(1-\mathrm{e}^{-2\bar{\xi}(k)R}\right)$ times the spectral variance of the direct speech component at time frame $\ell-1$. While the first term models the energy decay in the room, the second term models the energy growth due to the power of the source ($\lambda_{z_d}(\ell,k)/E_d$). As expected, only the source can increase the reverberant energy in the room and the absorption of the energy is completely determined by the reverberation time of the room.

3.7 Estimating Model Parameters

In order to estimate the late reverberant spectral variance an estimate of the reverberation time T_{60} of the room and the direct to reverberation ratio is required.

3.7.1 Reverberation Time

Partially blind methods to estimate the reverberation time have been developed in which the characteristics of the room are learnt using neural network approaches [19]. Another method uses a segmentation procedure for detecting gaps in the signals and then tracks the sound decay curve [49, 74]. A blind method has been proposed by Ratnam *et al.* based on a maximum-likelihood estimation procedure [64]. In [53] Löllmann and Vary proposed a maximum-likelihood estimator which takes additive noise into account. Most of these methods can also be applied to band-pass filtered versions of the original signal in order to estimate the reverberation time in different frequency bands.

In general, it is reasonable to assume that the reverberation time is approximately constant in the room. Therefore, in communication systems that involve echo cancellation, the reverberation time can be estimated using the estimated echo path [41]. For some applications such as audio or video-conferencing where a fixed setup is used, the reverberation time can be estimated using a calibration process.

3.7.2 Direct-to-reverberant Ratio

In many practical situations the distance between the source and the microphone will vary. Since the DRR depends on the distance between the source and the microphone, it is important that the parameter κ can be estimated online.

The parameter κ was introduced to prevent over-estimation of the reverberant spectral variance $\lambda_{z_\mathrm{r}}(\ell,k)$ when the source-microphone distance is smaller than the critical distance. In the case when κ is too large, the spectral variance $\hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ could become larger than $|Z(\ell,k)|^2$, which indicates that over-estimation has occurred. In this case, the value of κ should be lowered. In addition we know that during the free decay, which occurs after an offset of the source signal, $\hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ should be approximately equal to $|Z(\ell,k)|^2$. Estimation of κ could therefore be performed after a speech offset. Unfortunately, the detection of speech offsets is rather difficult. However, from the above discussion is has become clear that κ should at least fulfill the following condition: $|Z(\ell,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(\ell,k) \geq 0$.

The parameter κ can be estimated adaptively using the following strategy: (1) when speech is detected and $|Z(\ell,k)|^2 < \hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ the value of κ is lowered, (2) when $|Z(\ell,k)|^2 > \hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ the value of κ is raised slowly and (3) when $|Z(\ell,k)|^2 = \hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ the value of κ is assumed to be correct. This strategy can be implemented as follows:

$$\hat{\kappa}(\ell) = \hat{\kappa}(\ell-1) + \frac{\mu_\kappa}{P_z(\ell-1)} \sum_{k=0}^{\frac{K}{2}-1} \left(|Z(\ell-1,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(\ell-1,k) \right) \tag{3.88}$$

where $P_z(\ell-1) = \sum_{k=0}^{\frac{K}{2}-1} |Z(\ell-1,k)|^2$, and μ_κ ($0 < \mu_\kappa < 1$) denotes the step-size. After each update step, $\hat{\kappa}(\ell)$ is constrained, such that $0 < \hat{\kappa}(\ell) \leq 1$. Experimental results that demonstrate the feasibility of this estimator can be found in Sect. 3.8.

3.8 Experimental Results

In this section we present and discuss the experimental results that were obtained using single and multiple microphones. A uniformly linear microphone array was used with inter-microphone spacing $D_\mathrm{i} = 5$ cm. The source-array distance D is defined as the distance between the source and the center of the array, and ranges from 0.25 to 3 m. The dimensions of the room are 5 m $\times$ 6 m $\times$ 4 m (length $\times$ width $\times$ height). The experimental setup is depicted in Fig. 3.3. The APLAWD database [51] was used for evaluation with the sampling frequency set to $f_\mathrm{s} = 8$ kHz; it contains anechoic recordings comprising ten repetitions of five sentences uttered by five male and five female talkers. The reverberant microphone signals were obtained by convolving the anechoic recordings with different AIRs. The AIRs are generated using the image method for modelling small room acoustics [5], modified to accommodate fractional sample delays according to [59], with reverberation times from 250 to 1000 ms. The additive noise $v(n)$ was speech-like noise, taken

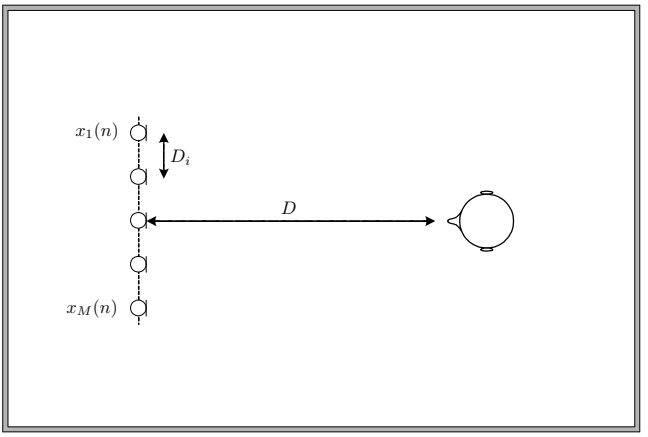


Fig. 3.3 Experimental setup with a uniform linear microphone array

from the NOISEX-92 database [75]. The spectral variance of the noise was estimated from the noisy microphone signal $x(n)$ using the IMCRA approach [14]. All *a priori* SIRs were estimated using the decision-directed approach. In all experiments we assumed that the reverberation time T_{60} of the room is known. Its value was determined using the Schroeder method, described in [68]. The parameter κ was estimated adaptively using the method described in Sect. 3.7.2. The parameters that were used for these experiments are shown in Table 3.1.

The segmental SIR and Bark Spectral Distortion (BSD), as defined in Chap. 2, are used for the evaluation.

Table 3.1 Parameters used in experiments

$f_s = 8000$ Hz	$n_e = 40$ ms	$G_{\min}^{\mathrm{dB}} = 18$ dB	$\beta^{\mathrm{dB}} = 3$ dB
$\eta = 0.95$	b = Hanning window	$w_{\mathrm{MSC}} = 9$	$\Phi_{\min} = 0.2$
$\Phi_{\max} = 0.65$	$\eta_s = 0.35$		

3.8.1 Using One Microphone

In this section we evaluate the performance of the single-microphone dereverberation method in the presence of noise using two objective measures. A summary of the complete single-microphone spectral enhancement algorithm that suppresses late reverberation and ambient noise is summarized in Algorithm 3.1.

We first evaluate the objective measures when $T_{60} = 0.5$ s and $D = 1$ m. The Signal to Noise Ratio (SNR) of the microphone signal ranges from 10 to 30 dB. In

Algorithm 3.1 Summary of the single-microphone spectral enhancement algorithm that suppresses late reverberation and ambient noise

1. **STFT**: Calculate the STFT of the noisy and reverberant signal $x(n)$.
2. **Estimate model parameters**: Firstly, decay-rate $\bar{\zeta}(k)$ is calculated using (3.4). Secondly, the parameter κ is estimated using (3.88).
3. **Estimate ambient noise**: Estimate $\lambda_v(\ell,k)$ using the method described in [18].
4. **Estimate late reverberant energy**: Calculate $G_{\mathrm{SP}}(\ell,k)$ using (3.62)–(3.64). Estimate $\lambda_z(\ell,k)$ using (3.61), and calculate $\hat{\lambda}_{z_\ell}(\ell,k)$ using (3.85).
5. **Post-filter**:
   (a) Calculate the *a posteriori* SIR using (3.26) and the individual *a priori* SIRs using (3.35)–(3.36) with $\vartheta \in \{z_\ell, v\}$, the total *a priori* SIR can then be calculated using (3.33).
   (b) Estimate the *a priori* speech presence probability $p(\ell,k|\ell-1)$ using the method described in [15] and calculate $\hat{p}(\ell,k)$ using (3.54).
   (c) Calculate the gain function $G_{\text{MMSE-LSA}}(\ell,k)$ using (3.27), (3.29), and (3.24).
   (d) Calculate $\hat{Z}_{\mathrm{e}}(\ell,k)$ using (3.30).
6. **Inverse STFT**: Calculate the output $\hat{z}_{\mathrm{e}}(n)$ by applying the inverse STFT to $\hat{Z}_{\mathrm{e}}(\ell,k)$.

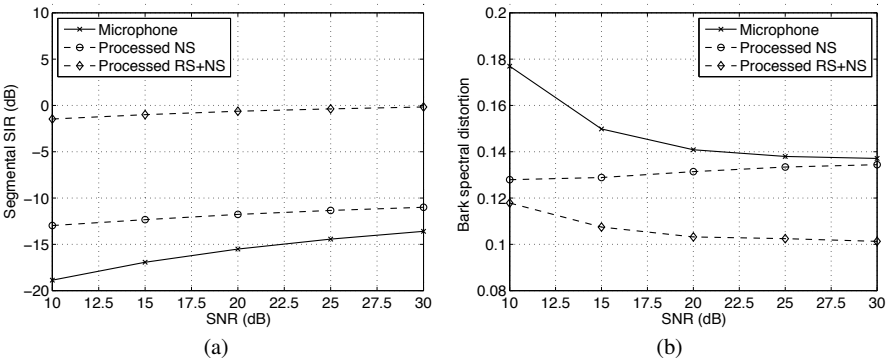


Fig. 3.4 (a) Segmental SIRs and (b) BSDs of the unprocessed microphone signal, the processed signal after noise suppression (NS), and the processed signal after joint reverberation and noise suppression (RS+NS). The SNR of the received signal varies between 10 and 30 dB ($D = 1$ m,$T_{60} =$ 500 ms, and $n_{\mathrm{e}}/f_{\mathrm{s}} = 40$ ms)

Fig. 3.4 the segmental SIR and BSD are depicted for the (unprocessed) reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). Joint reverberant and noise suppression significantly improves the segmental SIR (approximately 10 dB) and the BSD (approximately 0.04–0.06) compared to noise suppression only. After the noise suppression is applied, the reverberation becomes more pronounced. When, in addition to the noise, the late reverberation is suppressed, the subjective sound quality is significantly improved and the residual ambient noise sounds stationary. When listening to the processed signal, minor artifacts were audible when the SNR was larger than 15 dB. In Fig. 3.5,

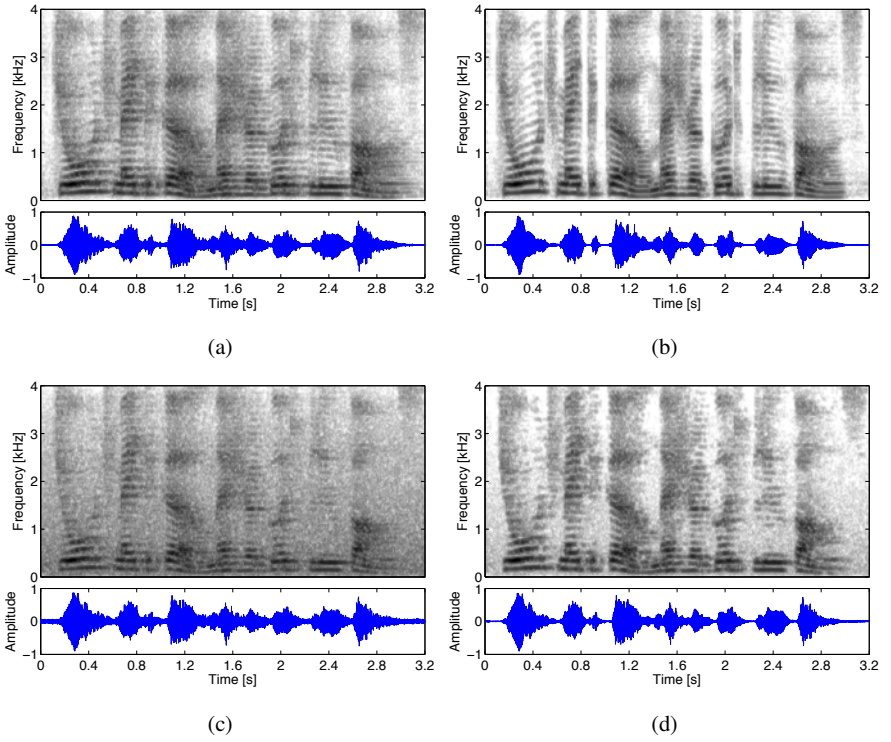
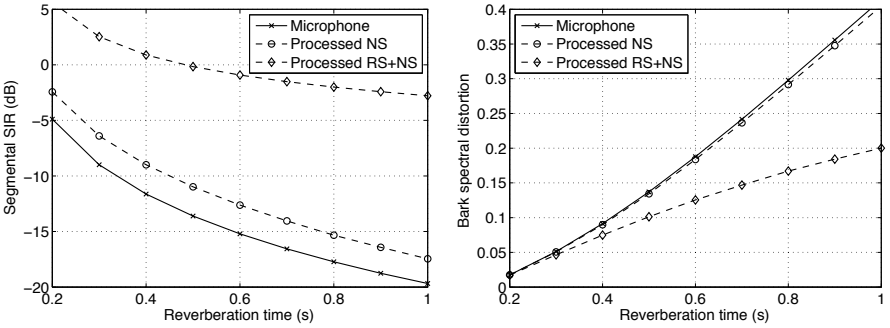


Fig. 3.5 Spectrograms and time-domain waveforms of (a) reverberant signal $z(n)$, (b) early speech signal $z_{\mathrm{e}}(n)$, (c) microphone signal (SNR $= 15$ dB, $T_{60} = 0.5$ s, $D = 1$ m), and (d) estimated early speech signal $\hat{z}_{\mathrm{e}}(n)$

spectrograms and time-domain waveforms are presented for one speech fragment. In both the spectrogram and time-domain waveform of the reverberant signal smearing of the speech, caused by the late reflections can be observed. In the enhanced speech signal, the smearing is significantly reduced as a result of the suppression of late reverberation. In addition, it can be seen that the noise is suppressed.

In the second experiment we evaluate the algorithms for SNR $= 30$ dB and $D =$ 1 m. The reverberation time T_{60} ranges from 0.2 to 1 s. In Fig. 3.6 the segmental SIR and BSD are depicted for the reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). Since the SNR is relatively high, the segmental SIR mainly depends on the reverberation suppression. The results of this experiment demonstrate that the algorithm is able to suppress a significant amount of late reverberation for short and long reverberation times. The results of an informal listening test indicated that for long reverberation times ($T_{60} > 0.5$ s), a larger value of n_{e} is preferred to maintain a natural sounding speech signal.



(a) (b)

Fig. 3.6 (a) Segmental SIRs and (b) BSDs of the unprocessed microphone signal, the processed signal after noise suppression (NS), and the processed signal after joint reverberation and noise suppression (RS+NS). The reverberation time varies between 0.2 and 1 s (SNR = 30 dB, $D = 1$ m, and $n_e/f_s = 40$ ms)

In the third experiment we evaluate the algorithms for SNR = 30 dB and $T_{60} = 0.5$ s. The source-microphone distance D ranges from 0.25 to 4 m. In the current setup the critical distance D_c equals 0.9 m. In Fig. 3.7 the segmental SIR and BSD are depicted for the reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). Since the SNR is relatively high, the segmental SIR mainly depends on the reverberation suppression. The results shown here demonstrate that the algorithm is able to suppress a significant amount of late reverberation over a wide range of source-microphone distances that are smaller and larger than the critical distance. While the BSD measures mainly show an improvement when the source-microphone distances are large, the segmental SIR improvement is almost constant. It should be noted that, for a source-microphone distance smaller than the critical distance, the value of n_e/f_s can be decreased without affecting the amount of speech distortion significantly.

3.8.2 Using Multiple Microphones

In this section we evaluate the performance of three multi-microphone dereverberation methods in the presence of spatially white noise (SNR = 30 dB) using two objective measures. Since the SNR is relatively high, the segmental SIR mainly depends on the reverberation suppression. The first multi-microphone method is the Delay-and-sum Beamformer (DSB). The second method is the delay and sum beamformer in conjunction with the single-channel post-filter described in Algorithm. 3.1 and is denoted by (DSB-PF). The third method is based on the non-linear spatial processor in conjunction with the same single-channel post-filter and is denoted by

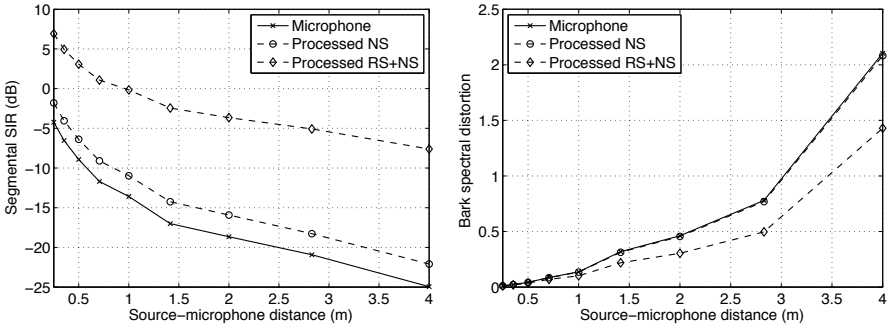
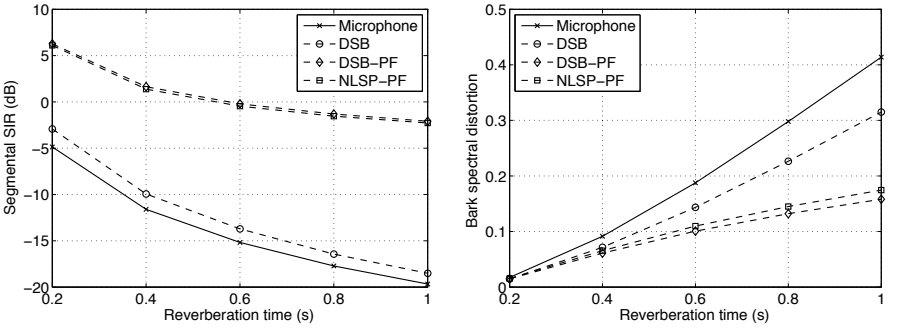


Fig. 3.7 (a) Segmental SIRs and (b) BSDs of the unprocessed microphone signal, the processed signal after noise suppression (NS), and the processed signal after joint reverberation and noise suppression (RS+NS). The source-microphone varies between 0.25 and 4 m (SNR = 30 dB, T_{60} = 500 ms, and $n_e/f_s = 40$ ms)

(NLSP-PF). As a reference the signal of the microphone that is closest to the desired source was evaluated.

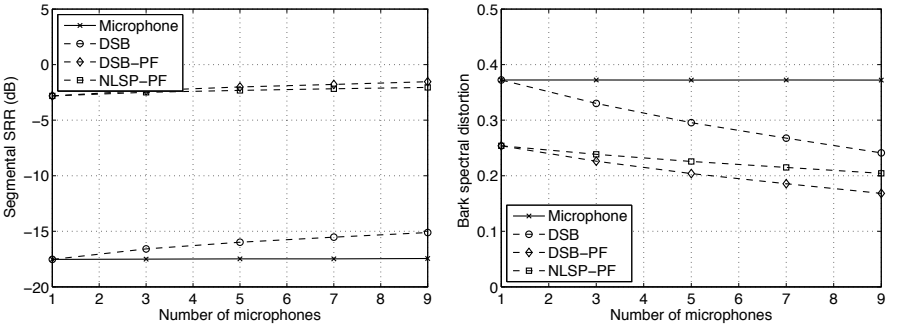
In the first experiment the number of microphones used was $M = 5$ and the source-microphone distance was set to $D = 1.5$ m. The reverberation time T_{60} ranged from 0.2 to 1 s. In Fig. 3.8 the segmental SIR and BSD are depicted for the reference microphone signal, the output of the DSB, the result of the DSB-PF method, and the result of the NLSP-PF method. These results show the limited performance of the DSB. A significant improvement is achieved by applying the single-channel post-filter to the output of the delay and sum beamformer. According to the objective measures employed the NLSP-PF method performs slightly worse compared to the DSB-PF method. However, the results of an informal listing test indicated that the output of the NLSP-PF method sounds more natural and contains less audible distortions than the output of the DSB-PF method. This could be explained by the fact that the objective measures used in this work are unable to reflect certain perceptual characteristics of the evaluated signals that are important in the context of speech dereverberation.

In the second experiment the reverberation time $T_{60} = 0.5$ s was used, and the source-microphone distance was set to $D = 1.5$ m. The number of microphones M ranged from 1 to 9. The segmental SIR and BSD values obtained are shown in Fig. 3.9. As in the previous experiment we can see that the single-channel post-filter significantly increases the dereverberation performance. The segmental SIR was increased by more than 14.5 dB compared to the reference microphone. It is noted that the segmental SIR increases slightly when more than one microphone is used. However, the BSD is significantly reduced by using multi-microphone signals. In terms of the segmental SIR and BSD the best result is obtained by the DSB-PF system. Judging from these results one might argue that the DSB-PF method performs better than the NLSP-PF method. However, as before the results from an



(a) (b)

Fig. 3.8 (a) Segmental SIRs and (b) BSDs of the reference microphone signal, the DSB signal, the DSB-PF signal, and the NLSP-PF signal. The reverberation time varies between 0.2 and 1 s ($D = 1.5$ m, SNR $= 30$ dB, and $n_e/f_s = 40$ ms)



(a) (b)

Fig. 3.9 (a) Segmental SIRs and (b) BSDs of the reference microphone signal, the DSB signal, the DSB-PF signal, and the NLSP-PF signal. The number of microphones ranges from 1 to 9 ($D = 1.5$ m, $T_{60} = 0.5$ s, SNR $= 30$ dB, and $n_e/f_s = 40$ ms)

informal listening test indicated that the results obtained by the NLSP-PF method sound more natural and contain less artifacts than the results obtained by the DSB-PF method.

3.9 Summary and Outlook

In this chapter single and multi-microphone speech dereverberation methods that are entirely or partly based on spectral enhancement were described. The quality of the received speech signal can be improved by reducing the effective noise that consists

of late reverberation and ambient noise. It was shown that quantifiable properties of the AIR, such as the reverberation time and DRR, can be used to dereverberate the received speech signal partly. In order to use spectral enhancement methods for speech dereverberation, an estimate of the late reverberant spectral variance is required. In Sect. 3.6 such an estimator was derived using a generalized statistical reverberation model. When the source-receiver distance is smaller than the critical distance the proposed estimator that is based on the generalized statistical model is advantageous over the estimator that is based on Polack's statistical model [39].

In the development of the speech enhancement method we assumed that the spectral coefficients of the speech and noise are Gaussian. Furthermore, we used the minimum mean squared error distortion measure and the log-amplitude fidelity criterion that was successfully used for noise suppression. However, it has yet to be determined if the MSE distortion measure and log-amplitude fidelity criterion provide the best results in the case of reverberation and noise suppression. Recently, the generalized autoregressive conditional heteroscedasticity (GARCH) model was shown to be useful for statistically modelling speech signals in the STFT domain [16]. A Markov-switching time-frequency GARCH model was proposed in [1, 2] for modelling non-stationary signals in the time-frequency domain. The model takes into account the strong correlation of successive spectral magnitudes and is more appropriate than the decision-directed approach for speech spectral variance estimation in noisy environments. Should this or other statistical speech models be used in the development of novel spectral speech dereverberation algorithms, they might further increase the suppression of late reverberation and noise and decrease the amount of speech distortion. In the course of this chapter, two modifications of the standard MMSE-LSA estimator were discussed. The first modification concerns the spectral gain function and allows a larger suppression of late reverberation when the early speech component is inactive and results in a constant residual ambient noise level. The second modification concerns the speech presence probability estimator, which is improved by analyzing the magnitude squared coherence of the observed sound field.

We also investigated the use of multiple microphones for speech dereverberation and described two multi-microphone systems. The first system consists of an MVDR beamformer followed by a single-channel post-filter. Although this system can be useful in the presence of coherent noise sources, we could not directly exploit the spatial diversity of the reverberant signal to estimate the late reverberant spectral variance. In a spatially white noise field, the MVDR beamformer reduces to the well-known delay and sum beamformer. It has been shown in [39] that due to the spatial correlation between the AIRs, the residual reverberation at the output of the beamformer might contain undesired signal components. These components are especially pronounced at low frequencies and become larger when the inter-microphone distances are small. A second multi-microphone system that does not suffer from the spatial correlation between the AIRs was described. The latter consists of a non-linear spatial processor followed by a single-channel post-filter. The non-linear spatial processor can only be employed when the noise field is spatially white. Although practically feasible multi-microphone solutions have been found,

further research is required to investigate the tradeoff between noise suppression and reverberation suppression.

Finally, experimental results demonstrated the beneficial use of the single-microphone spectral dereverberation method described and showed that a large amount of reverberation and noise can be reduced with little speech distortion.

Acknowledgment

The author thanks Dr. Sharon Gannot and Dr. Israel Cohen for the valuable discussions and helpful suggestions.

References

1. Abramson, A., Cohen, I.: Markov-switching GARCH model and application to speech enhancement in subbands. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4. Paris, France (2006)
2. Abramson, A., Cohen, I.: Recursive supervised estimation of a Markov-switching GARCH process in the short-time Fourier transform domain. IEEE Trans. Signal Process. **55**(7), 3227–3238 (2007)
3. Accardi, A.J., Cox, R.V.: A modular approach to speech enhancement with an application to speech coding. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 201–204 (1999)
4. Allen, J.B.: Effects of small room reverberation on subjective preference. J. Acoust. Soc. Am. **71**(S1), S5 (1982)
5. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
6. Benesty, J., Makino, S., Chen, J. (eds.): Speech Enhancement. Springer (2005)
7. Benesty, J., Sondhi, M.M., Huang, Y. (eds.): Springer Handbook of Speech Processing. Springer (2007)
8. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 208–211 (1979)
9. Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech, Signal Process. **ASSP-27**(2), 113–120 (1979)
10. Bolt, R.H., MacDonald, A.D.: Theory of speech masking by reverberation. J. Acoust. Soc. Am. **21**, 577–580 (1949)
11. Burshtein, D., Gannot, S.: Speech enhancement using a mixture-maximum model. IEEE Trans. Speech Audio Process. **10**(6), 341351 (2002)
12. Cappe, O.: Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. **2**(2), 345–349 (1994). DOI 10.1109/89.279283
13. Cohen, I.: Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. **9**(4), 113–116 (2002)
14. Cohen, I.: Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003). DOI 10.1109/TSA.2003.811544

15. Cohen, I.: From volatility modeling of financial time-series to stochastic modeling and enhancement of speech signals. In: J. Benesty, S. Makino, J. Chen (eds.) Speech Enhancement, chap. 5, pp. 97–114. Springer (2005)
16. Cohen, I.: Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. Signal Processing **86**(4), 698–709 (2006)
17. Cohen, I., Gannot, S.: Spectral enhancement methods. In: Benesty et al. [7], chap. 45. Part H
18. Cohen, I., Gannot, S., Berdugo, B.: An integrated real-time beamforming and post filtering system for nonstationary noise environments. EURASIP J. on App. Signal Process. **11**, 1064–1073 (2003)
19. Cox, T.J., Li, F., Darlington, P.: Extracting room reverberation time from speech using artificial neural networks. J. Audio Eng. Soc. **49**(4), 219–230 (2001)
20. Crochiere, R.E., Rabiner, L.R.: Multirate Digital Signal Processing. Prentice-Hall (1983)
21. Delcroix, M., Hikichi, T., Miyoshi, M.: Precise dereverberation using multichannel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15**(2), 430–440 (2007)
22. Deller, J.R., Proakis, J.G., Hansen, J.H.L.: Discrete-Time Processing of Speech Signals. New York: MacMillan (1993)
23. Ephraim, Y., Cohen, I.: Recent advancements in speech enhancement. In: R.C. Dorf (ed.) The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing, third edn. CRC Press (2006)
24. Ephraim, Y., Lev-Ari, H., Roberts, W.J.J.: A brief survey of speech enhancement. In: The Electronic Handbook, second edn. CRC Press (2005)
25. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. **32**(6), 1109–1121 (1984)
26. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. **33**(2), 443–445 (1985)
27. Gannot, S., Cohen, I.: Adaptive beamforming and postfiltering. In: Benesty et al. [7], chap. 48
28. Gannot, S., Moonen, M.: Subspace methods for multimicrophone speech dereverberation. EURASIP J. on App. Signal Process. **2003**(11), 1074–1090 (2003)
29. Gaubitch, N.D., Naylor, P.A.: Analysis of the dereverberation performance of microphone arrays. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC) (2005)
30. Gaubitch, N.D., Naylor, P.A., Ward, D.B.: On the use of linear prediction for dereverberation of speech. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 99–102 (2003)
31. Goh, Z., Tan, K.C., Tan, T.G.: Postprocessing method for suppressing musical noise generated by spectral subtraction. IEEE Trans. Speech Audio Process. **6**(3), 287–292 (1998). DOI 10.1109/89.668822
32. Griebel, S.M., Brandstein, M.S.: Wavelet transform extrema clustering for multi-channel speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 52–55. Pocono Manor, Pennsylvania (1999)
33. Gürelli, M.I., Nikias, C.L.: EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. IEEE Trans. Signal Process. **43**(1), 134–149 (1995)
34. Gustafsson, S., Martin, R., Jax, P., Vary, P.: A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. IEEE Trans. Speech Audio Process. **10**(5), 245–256 (2002)
35. Gustafsson, S., Martin, R., Vary, P.: Combined acoustic echo control and noise reduction for hands-free telephony. Signal Processing **64**(1), 21–32 (1998)
36. Gustafsson, S., Nordholm, S., Claesson, I.: Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech Audio Process. **9**(8), 799–807 (2001)
37. Habets, E.A.P.: Multi-channel speech dereverberation based on a statistical model of late reverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 173–176. Philadelphia, USA (2005)
38. Habets, E.A.P.: Speech dereverberation based on a statistical model of late reverberation using a linear microphone array. In: Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), pp. d7–d8. Piscataway, NJ, USA (2005)

39. Habets, E.A.P.: Single- and multi-microphone speech dereverberation using spectral enhancement. Ph.D. thesis, Technische Universiteit Eindhoven (2007)
40. Habets, E.A.P., Cohen, I., Gannot, S.: MMSE log spectral amplitude estimator for multiple interferences. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4. Paris, France (2006)
41. Habets, E.A.P., Cohen, I., Gannot, S., Sommen, P.C.W.: Joint dereverberation and residual echo suppression of speech signals in noisy environments. IEEE Trans. Audio, Speech, Lang. Process. **16**(8), 1433–1451 (2008)
42. Habets, E.A.P., Gannot, S., Cohen, I.: Dual-microphone speech dereverberation in a noisy environment. In: Proc. IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT), pp. 651–655. Vancouver, Canada (2006)
43. Haykin, S.: Blind Deconvolution, fourth edn. Prentice-Hall Information and System Sciences. Prentice-Hall (1994)
44. Hopgood, J.: Nonstationary signal processing with application to reverberation cancellation in acoustic environments. Ph.D. thesis, Cambridge University (2001)
45. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. **51**(1), 11–24 (2003)
46. Jetzt, J.J.: Critical distance measurement of rooms from the sound energy spectral response. J. Acoust. Soc. Am. **65**(5), 1204–1211 (1979)
47. Jot, J.M., Cerveau, L., Warusfel, O.: Analysis and synthesis of room reverberation based on a statistical time-frequency model. In: Proc. Audio Eng. Soc. Convention (1997)
48. Kuttruff, H.: Room Acoustics, 4th edn. Taylor & Frances (2000)
49. Lebart, K., Boucher, J.M., Denbigh, P.N.: A new method based on spectral subtraction for speech dereverberation. Acta Acoustica **87**, 359–366 (2001)
50. Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**(12), 1586–1604 (1979)
51. Lindsey, G., Breen, A., Nevard, S.: SPAR's archivable actual-word databases. Technical report, University College London (1987)
52. Loizou, P.C.: Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Trans. Speech Audio Process. **13**(5), 857–869 (2005). DOI 10.1109/TSA.2005.851929
53. Löllmann, H.W., Vary, P.: Estimation of the reverberation time in noisy environments. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4 (2008)
54. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001). DOI 10.1109/89.928915
55. Martin, R.: Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Process. **13**(5), 845–856 (2005). DOI 10.1109/TSA.2005.851927
56. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
57. Nábělek, A.K., Letowski, T.R., Tucker, F.M.: Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Am. **86**(4), 1259–1265 (1989)
58. Nábělek, A.K., Mason, D.: Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. J. Speech Hear. Res. **24**, 375–383 (1981)
59. Peterson, P.M.: Simulating the response of multiple microphones to a single acoustic source in a reverberant room. J. Acoust. Soc. Am. **80**(5), 1527–1529 (1986)
60. Peutz, V.M.A.: Articulation loss of consonants as a criterion for speech transmission in a room. J. Audio Eng. Soc. **19**(11), 915–919 (1971)
61. Polack, J.D.: La transmission de l'énergie sonore dans les salles. Ph.D. thesis, Université du Maine, La Mans, France (1988)
62. Polack, J.D.: Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics. Appl. Acoust. **38**(2), 235–244 (1993)
63. Radlović, B.D., Kennedy, R.A.: Nonminimum-phase equalization and its subjective importance in room acoustics. IEEE Trans. Speech Audio Process. **8**(6), 728–737 (2000)

64. Ratnam, R., Jones, D.L., Wheeler, B.C., O'Brien, Jr., W.D., Lansing, C.R., Feng, A.S.: Blind estimation of reverberation time. J. Acoust. Soc. Am. **114**(5), 2877–2892 (2003)
65. Sabine, W.C.: Collected Papers on acoustics (Originally 1921). Peninsula Publishing (1993)
66. Schroeder, M.R.: Statistical parameters of the frequency response curves of large rooms. J. Audio Eng. Soc. **35**, 299–306 (1954)
67. Schroeder, M.R.: Frequency correlation functions of frequency responses in rooms. J. Acoust. Soc. Am. **34**(12), 1819–1823 (1962)
68. Schroeder, M.R.: Integrated-impulse method measuring sound decay without using impulses. J. Acoust. Soc. Am. **66**(2), 497–500 (1979)
69. Schroeder, M.R.: The "schroeder frequency" revisited. J. Acoust. Soc. Am. **99**(5), 3240–3241 (1996). DOI 10.1121/1.414868
70. Sim, B.L., Tong, Y.C., Chang, J.S., Tan, C.T.: A parametric formulation of the generalized spectral subtraction method. IEEE Trans. Speech Audio Process. **6**(4), 328–337 (1998)
71. Steinberg, J.C.: Effects of distortion upon the recognition of speech sounds. J. Acoust. Soc. Am. **1**, 35–35 (1929)
72. Takata, Y., Nábělek, A.K.: English consonant recognition in noise and in reverberation by Japanese and American listeners. J. Acoust. Soc. Am. **88**, 663–666 (1990)
73. Talantzis, F., Ward, D.B.: Robustness of multichannel equalization in an acoustic reverberant environment. J. Acoust. Soc. Am. **114**(2), 833–841 (2003)
74. Tashev, I., Malvar, H.S.: A new beamformer design algorithm for microphone arrays. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. iii/101–iii/104 (2005)
75. Varga, A., Steeneken, H.J.M.: Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication **3**(3), 247–251 (1993). DOI 10.1016/0167--6393(93)90095--3
76. Wexler, J., Raz, S.: Discrete Gabor expansions. Signal Processing **21**(3), 207–220 (1990)
77. Wolfe, P.J., Godsill, S.J.: Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. on App. Signal Process. **2003**(10), 1043–1051 (2003)
78. Yegnanarayana, B., Satyanarayana, P.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Speech Audio Process. **8**(3), 267–281 (2000)

Chapter 4
Dereverberation Using LPC-based Approaches

Nikolay D. Gaubitch, Mark R.P. Thomas, and Patrick A. Naylor

Abstract A class of reverberant speech enhancement techniques involve processing of the linear prediction residual signal following Linear Predictive Coding (LPC). These approaches are based on the assumption that reverberation is mainly confined to the prediction residual and affects the LPC coefficients to a lesser extent. This chapter begins with a study on the effects of reverberation on the LPC parameters where mathematical tools from statistical room acoustics are used in the analysis. Consequently, a general framework for dereverberation using LPC is formulated and several existing methods utilizing this approach are reviewed. Finally, a specific method for processing a reverberant prediction residual is presented in detail. This method uses a combination of spatial averaging and larynx cycle-based temporal averaging. Experiments with a microphone array in a small office demonstrate the dereverberation and noise suppression of the spatiotemporal averaging method, showing up to a 5 dB improvement in segmental SRR and 0.33 in the normalized Bark spectral distortion score.

4.1 Introduction

Speech dereverberation algorithms can be classified into one of the three main categories:

(i) *Beamforming* – the observed signals received at the different microphones are delayed, weighted and summed, so as to form a beam in the direction of the desired source and to attenuate sounds form other directions.
(ii) *Speech enhancement* – the observed speech signals are modified so as to better represent some features of the clean speech signal according to *a priori* models of the speech waveform or spectrum.

Imperial College London, UK

(iii) *Blind system identification and equalization* – the acoustic impulse responses are identified blindly (using only the observed signals) and then used to design an equalization filter that compensates for the effect of the acoustic impulse responses.

As discussed in Chap. 1, blind system identification and equalization methods can, in theory, perform exact dereverberation. However, these methods are difficult to apply in practice due to several factors, including high computational complexity and sensitivity to noise. Speech enhancement methods and beamforming, on the other hand, are often seen as more directly practical techniques, at least at the current point in the development of these technologies. Although they provide incomplete dereverberation, they allow practical online implementations. One important class of speech enhancement algorithms for reverberation reduction is based on linear predictive coding of speech and is the focus of this chapter.

Linear Predictive Coding (LPC) is an established and powerful analysis tool for speech and audio signals [2, 9, 24, 28], which represents speech as an excitation signal that excites an all-pole filter that can be represented compactly. Therefore, LPC is employed in several speech processing applications such as speech recognition, although not normally directly, speech coding and in pitch modification [9].

It has been observed that when LPC analysis is applied to reverberant speech, the effects of reverberation mainly reside in the prediction residual [6, 14, 40]. This is particularly true in the case of multi-microphone systems where accuracy of the estimation of the clean speech LPC coefficients from reverberant observations is improved [13]. Dereverberation is achieved by processing the LPC residual signal and then synthesizing a speech signal with reduced reverberation from the output of a filter employing the LPC coefficients obtained from the reverberant speech whose input is an enhanced version of the prediction residual signal. LPC based methods have been found to provide only moderate reduction in dereverberation but possess several additional beneficial features. Firstly, the 'blindness' of the dereverberation problem is reduced to some extent because the general structure of the LPC residual is known in the form of models that have become established over many years of research. Secondly, such methods are less sensitive to processing errors since these are effectively smoothed at the synthesis stage. Third, they do not require knowledge of the room impulse responses, which are difficult and computationally expensive to estimate. Therefore, the LPC-based methods offer practical algorithms with the possibility of online implementation [33].

This chapter is organized as follows: Sect. 4.2 reviews the fundamentals of LPC. In Sect. 4.3, the effects of reverberation on the LPC coefficients and on the prediction residual are discussed. The use of LPC for reverberant speech enhancement is elaborated in Sect. 4.4. One current algorithm based on prediction residual enhancement is developed in detail in Sect. 4.5 and its performance is demonstrated with illustrative example simulations. The concepts presented in this chapter are concluded in Sect. 4.6.

4.2 Linear Predictive Coding of Speech

Linear predictive coding of speech is the key to the dereverberation methods presented in this chapter. The fundamental properties of LPC analysis of speech signals are now reviewed. The interested reader is referred to one of the many excellent texts on speech signal processing (e.g. [9, 28]) for more details on this topic.

A speech signal $s(n)$ can be expressed in terms of a p^{th} order linear predictor according to [28]

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n), \tag{4.1}$$

where a_i are the predictor coefficients and $e(n)$ is the prediction error. The LPC coefficients can be used to form the prediction error filter

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^i \tag{4.2}$$

and the corresponding all-pole filter

$$\begin{aligned} V(z) &= \frac{1}{1+\sum_{i=1}^{p} a_i z^i} \\ &= \frac{1}{A(z)}. \end{aligned} \tag{4.3}$$

Thus, the problem of LPC is to determine the predictor coefficients a_i given the signal $s(n)$.

One commonly used approach is to find the coefficients that minimize the mean squared prediction error. A cost function is formed from (4.1) as

$$\begin{aligned} J &= E\left\{e^2(n)\right\} \\ &= E\left\{\left(s(n) - \sum_{i=1}^{p} a_i s(n-i)\right)^2\right\}, \end{aligned} \tag{4.4}$$

where $E\{\cdot\}$ is the expectation operator.

The error is minimized in each analysis frame, defined over some range of n, in a least squares sense by setting the derivative of J to zero with respect to each of the LPC coefficients

$$\frac{\partial J}{\partial a_i} = 0, \tag{4.5}$$

which results in

$$\sum_{u=1}^{p} a_u E\left\{s(n-i)s(n-u)\right\} = E\{s(n)s(n-u)\}, \quad 1 \le u \le p. \tag{4.6}$$

The set of the p linear equations in (4.6) are often referred to as the normal equations and can be written in matrix form as

$$\begin{bmatrix} r_{ss,0} & r_{ss,1} & \cdots & r_{ss,p-1} \\ r_{ss,1} & r_{ss,0} & \cdots & r_{ss,p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{ss,p-1} & r_{ss,p-2} & \cdots & r_{ss,0} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_{ss,1} \\ r_{ss,2} \\ \vdots \\ r_{ss,p} \end{bmatrix}, \tag{4.7}$$

where

$$r_{ss,i} = E\{s(n)s(n-i)\} \tag{4.8}$$

is the autocorrelation of $s(n)$ at time lag i. The autocorrelation function can be expressed equivalently in terms of the signal spectrum as

$$r_{ss,i} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(e^{j\omega})|^2 e^{j\omega i} \, \mathrm{d}\omega, \quad i = 1,2,\ldots,p, \tag{4.9}$$

where $S(e^{j\omega})$ is the Fourier transform of $s(n)$. Equation (4.7) can be written more compactly as

$$\mathbf{R}_{ss}\mathbf{a} = \mathbf{r}_{ss}, \tag{4.10}$$

where

$$\mathbf{a} = [a_1 \; a_2 \; \ldots \; a_p]^T$$

are the prediction coefficients.

Consequently, the least squares optimal estimate of the LPC coefficients can be found by

$$\hat{\mathbf{a}} = \hat{\mathbf{R}}_{ss}^{-1}\hat{\mathbf{r}}_{ss}, \tag{4.11}$$

with $\hat{\mathbf{R}}_{ss}$ and $\hat{\mathbf{r}}_{ss}$ being estimates of $\mathbf{R}_{ss}$ and $\mathbf{r}_{ss}$, respectively, where the expectations are calculated with sample averages [24].

In the so-called autocorrelation method of LPC, an appropriate definition of the frame (in terms of the choice of the range of n) over which J is formulated results in $\mathbf{R}_{ss}$ exhibiting Toeplitz symmetry and, as such, (4.11) can be solved efficiently with the Levinson–Durbin algorithm [9, 24]. In the derivation of the calculation of the optimal LPC coefficients it is assumed that $s(n)$ is stationary over the frame of analysis. However, speech signals are intrinsically time-varying but can be considered stationary for a duration of 10–30 ms. Hence, the LPC coefficients are usually calculated over windowed frames of 10 to 30 ms duration, possibly overlapping, resulting in a time-varying filter $V(z)$ [9].

The parameters obtained from the LPC can be linked to a model of the speech production system given in Fig. 4.1. In this simplified model, the all-pole filter $V(z)$ represents the vocal tract, and the prediction residual, $e(n)$, approximately represents the vocal tract excitation sequence, comprising a quasi-periodic pulse train for voiced speech and random noise for unvoiced speech. Further investigations into this model enable specific inclusion of other effects such as modelling of lip radiation and sophisticated models of glottal excitation.

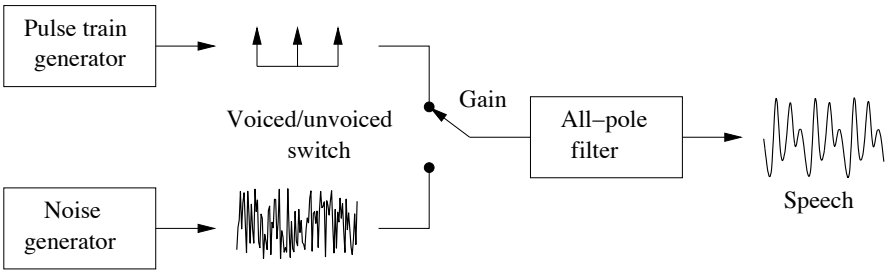


Fig. 4.1 Speech production model

4.3 LPC on Reverberant Speech

In this section, we present a study on the relationship between the LPC parameters obtained from anechoic speech and those obtained from reverberant speech.

The speech signal observed at the m^{th} sensor in an array of M microphones can be expressed as

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n) + \nu_m(n), \quad m = 1,2,\ldots,M, \tag{4.12}$$

where

$$\mathbf{h} = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T$$

is the L-tap room impulse response from the source to the m^{th} sensor,

$$\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L+1)]^T$$

is the input vector of clean speech samples and $\nu_m(n)$ denotes additive measurement noise. For the study of the effects of reverberation on LPC, we assume a noise-free environment, $\nu_m(n) = 0,\ \forall m$. For additional studies on the effects of noise on the LPC see [22, 30]. In linear prediction terms, the observation of the reverberant speech at the m^{th} sensor from (4.12) can be written

$$x_m(n) = \sum_{i=1}^{p} b_{m,i} x_m(n-i) + e_m(n), \quad m = 1,2,\ldots,M, \tag{4.13}$$

and the LPC coefficients are found as in (4.11)

$$\hat{\mathbf{b}}_m = \hat{\mathbf{R}}_{xx,m}^{-1} \hat{\mathbf{r}}_{xx,m}, \tag{4.14}$$

where

$$\hat{\mathbf{b}}_m = [\hat{b}_{m,1}\ \hat{b}_{m,2}\ \ldots\ \hat{b}_{m,p}]^T$$

is a vector with the prediction coefficients and $e_m(n)$ is the LPC residual obtained from the m^{th} sensor signal. We are interested in the relationship between the prediction residuals, $e(n)$, and LPC coefficients obtained from clean speech, $\hat{\mathbf{a}}$, and the

corresponding prediction residuals, $e_m(n)$, and LPC coefficients, $\hat{\mathbf{b}}_m$, found from reverberant observations. This will be discussed next in the context of both a single microphone and multiple microphones.

4.3.1 Effects of Reverberation on the LPC Coefficients

We utilize tools from Statistical Room Acoustics (SRA) theory [21, 27, 29] for the analysis of the LPC coefficients in reverberant speech. SRA provides a means for describing the sound field in a room that is more mathematically tractable compared to, for example, wave theory [29]. SRA has been employed by several researchers for the analysis of signal processing algorithms in reverberant environments including acoustic channel equalization [4, 29, 31], blind source separation [32], sound source localization [18] and acoustic crosstalk cancellation [36].

There are several ways to obtain the LPC coefficients from reverberant speech when a multiple microphone observation is available. The following three cases are considered in our study:

(i) LPC coefficients calculated from a single microphone observation
(ii) LPC coefficients calculated from an M-channel observation ($M > 1$)
(iii) LPC coefficients obtained at the output of a Delay-and-sum Beamformer (DSB)

It will be shown, in terms of spatial expectation, that the LPC coefficients obtained from reverberant speech are approximately equal to those from clean speech for cases (i) and (ii), while the LPC coefficients obtained from the output of the DSB can differ due to spatial correlation between the microphones. Furthermore, it will be demonstrated that the M-channel LPC coefficients from (ii) provide the best estimate of the clean speech coefficients compared to the other two cases.

4.3.1.1 Single Microphone

Consider the single microphone case where $M = 1$. Applying the LPC analysis from Sect. 4.2 on the reverberant speech signal, $x(n)$ (we drop the subscript m for the single channel case to improve clarity of presentation), the LPC coefficients are obtained with

$$\hat{\mathbf{b}} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xx}. \tag{4.15}$$

The i^{th} autocorrelation coefficient of $\mathbf{R}_{xx}$ is given by

$$r_{xx,i} = E\{x(n)x(n-i)\}, \tag{4.16}$$

which can be written equivalently in the frequency domain as

$$\begin{aligned} r_{xx,i} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{j\omega})|^2 e^{j\omega i}\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 |S(e^{j\omega})|^2 e^{j\omega i}\,\mathrm{d}\omega, \quad i = 1,2,\ldots,p \end{aligned} \tag{4.17}$$

where $H(e^{j\omega})$ and $S(e^{j\omega})$ are the Fourier transforms of $\mathbf{h}$ and $s(n)$, respectively.

In order to study the LPC coefficients of reverberant speech, the spatial expectation [27] is taken on both sides of (4.15) giving

$$\mathscr{E}\{\hat{\mathbf{b}}\} = \mathscr{E}\{\mathbf{R}_{xx}^{-1}\mathbf{r}_{xx}\}, \tag{4.18}$$

where $\mathscr{E}\{\cdot\}$ denotes spatial expectation, which is explained in detail in Chap. 2. Furthermore, the expectation of each term of (4.18) has to be considered separately. We use the same approach as that used in [29], where the zeroth order Taylor series expansion is employed in the approximation

$$\mathscr{E}\{g(x)\} \cong g(\mathscr{E}\{x\}),$$

so (4.18) can be written as

$$\mathscr{E}\{\hat{\mathbf{b}}\} \cong \mathscr{E}\{\mathbf{R}_{xx}\}^{-1}\mathscr{E}\{\mathbf{r}_{xx}\}. \tag{4.19}$$

This reduces the problem to studying the properties of the LPC coefficients in terms of the autocorrelation function. Consequently, we consider the spatial expectation of $r_{xx,i}$ in (4.17)

$$\mathscr{E}\{r_{xx,i}\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathscr{E}\{|H(e^{j\omega})|^2\}|S(e^{j\omega})|^2 e^{j\omega i}\,\mathrm{d}\omega, \tag{4.20}$$

where the term $S(e^{j\omega})$ is taken outside the spatial expectation since it is independent of the source-microphone position.

According to SRA (as discussed in Chap. 2) the spatial expectation of the energy density spectrum of the Acoustic Transfer Function (ATF) can be written in terms of a direct component and a reverberant component

$$\mathscr{E}\{|H_m(e^{j\omega})|^2\} = |H_{d,m}(e^{j\omega})|^2 + \mathscr{E}\{|H_r(e^{j\omega})|^2\}. \tag{4.21}$$

Furthermore, the direct and the reverberant components can be expressed as, respectively, [29]

$$|H_{d,m}(e^{j\omega})|^2 = \frac{1}{(4\pi D)^2} \tag{4.22}$$

and

$$\mathscr{E}\{|H_r(e^{j\omega})|^2\} = \left(\frac{1-\bar{\alpha}}{\pi A\bar{\alpha}}\right), \tag{4.23}$$

where D is the distance between the source and microphone, A is the total surface area of the room and $\bar{\alpha}$ is the average wall absorption coefficient. The SRA expression for the expected energy density spectrum of the ATF is then

$$\begin{aligned}\mathscr{E}\{|H(e^{j\omega})|^2\} &= \frac{1}{(4\pi D)^2} + \left(\frac{1-\bar{\alpha}}{\pi A \bar{\alpha}}\right) \\ &= \kappa. \end{aligned} \tag{4.24}$$

Since κ is independent of frequency, by substitution of (4.24) into (4.20) the autocorrelation coefficient in (4.20) becomes

$$\begin{aligned}\mathscr{E}\{r_{xx,i}\} &= \frac{\kappa}{2\pi}\int_{-\pi}^{\pi} |S(e^{j\omega})|^2 e^{j\omega i}\,\mathrm{d}\omega \\ &= \kappa r_{xx,i}, \end{aligned} \tag{4.25}$$

for $i = 1, 2, \ldots, p$. Finally, substituting the result from (4.25) into (4.19) gives

$$\mathscr{E}\{\hat{\mathbf{b}}\} \cong \hat{\mathbf{a}}. \tag{4.26}$$

This result states that if LPC analysis is applied to reverberant speech, the coefficients $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are not necessarily equal at a single observation point in space. However, in terms of spatial expectation, the LPC coefficients from reverberant speech are approximately equal to those from clean speech. The accuracy of the approximation depends on the accuracy of the estimation of the spatial expectation of the autocorrelation function. Intuitively, the result in (4.26) suggests that using a microphone array in a manner so as to approximate the taking of the spatial expectation will give a more accurate estimation of the LPC coefficients than the use of a single observation alone.

4.3.1.2 Joint Multichannel Optimization

From (4.13), a joint M-channel cost function can be formulated as [6, 13]

$$J_M = \frac{1}{M}\sum_{m=1}^{M} E\{e_m^2(n)\} \tag{4.27}$$

$$= \frac{1}{M}\sum_{m=1}^{M} E\left\{\left(x_m(n) - \sum_{i=1}^{p} b_{m,i} x_m(n-i)\right)^2\right\}. \tag{4.28}$$

The optimum set of coefficients that minimize this error, similarly to (4.11), is given by

$$\hat{\mathbf{b}}_M = \bar{\mathbf{R}}_{xx}^{-1}\bar{\mathbf{r}}_{xx}, \tag{4.29}$$

with

$$\bar{\mathbf{R}}_{xx} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{R}_{xx,m} \tag{4.30}$$

and

$$\bar{\mathbf{r}}_{xx} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{r}_{xx,m}, \tag{4.31}$$

where $\bar{\mathbf{R}}_{xx}$ and $\bar{\mathbf{r}}_{xx}$ are, respectively, the $p \times p$ mean autocorrelation matrix and the $p \times 1$ mean autocorrelation vector across the M microphones.

Replacing the autocorrelation matrix and the autocorrelation vector in (4.18) with the M-microphone averages in (4.30) and (4.31), respectively, and then following the steps of the derivation of (4.26), it can be seen that the spatial expectation of the LPC coefficients obtained from minimization of (4.27) is approximately equal to those from clean speech,

$$\mathscr{E}\{\hat{\mathbf{b}}_M\} \cong \hat{\mathbf{a}}. \tag{4.32}$$

This result implies that the optimal LPC coefficients obtained using a spatial expectation over M channels are equivalent to the spatial expectation of the LPC coefficients in the single microphone case in (4.26). However, at each individual position the M-channel case provides a more accurate estimation of the clean speech LPC coefficients than that obtained with a single reverberant channel, as will be shown by simulations in Sect. 4.3.3. This is because the averaging of the autocorrelation functions in (4.30) and (4.31) is equivalent in effect to the calculation of the spatial expectation operation in the single channel case (4.19).

4.3.1.3 LPC at the Output of a Delay-and-sum Beamformer

A different approach to the multichannel LPC technique described above is to perform spatial averaging on the speech signals using, for example, a delay-and-sum beamformer. The output of a delay-and-sum beamformer can be written as [35]

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n-\tau_m), \tag{4.33}$$

where τ_m is the propagation delay in samples from the source to the m^{th} microphone. Assuming that the time-delays of arrival are known for all microphones, linear prediction can be performed on the beamformer output, $\bar{x}(n)$, as for the single channel chase

$$\hat{\mathbf{b}}_{\text{DSB}} = \mathbf{R}_{\bar{x}\bar{x}}^{-1}\mathbf{r}_{\bar{x}\bar{x}}. \tag{4.34}$$

The spatial expectation of the LPC coefficients calculated by linear prediction from the output of the DSB is

$$\mathscr{E}\{\hat{\mathbf{b}}_{\text{DSB}}\} \cong \mathbf{T}\hat{\mathbf{a}} - \mathbf{t}, \tag{4.35}$$

with

$$\mathbf{T} = \mathbf{I} - \frac{1}{\bar{\kappa}}\mathbf{R}_{ss}^{-1}\boldsymbol{\Gamma}\left(\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Gamma}^H\frac{1}{\bar{\kappa}}\mathbf{R}_{ss}^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^H$$

and

$$\mathbf{t} = (\bar{\kappa}\mathbf{R}_{ss} + \boldsymbol{\Xi})^{-1}\boldsymbol{\xi},$$

where these terms are defined in the derivation given in Appendix A.

The result in (4.35) states that in terms of spatial expectation, the LPC coefficients obtained by LPC analysis of the DSB output, $\bar{x}(n)$, differ from those obtained from clean speech. This difference depends on the spatial cross-correlation between the acoustic channels. It can be seen from (4.57) that the inter-channel correlation and its significance are governed by the reverberation time, the distance between adjacent microphones, the source-microphone separation and is dependent on the array size if the speaker is in the near-field of the microphone array. Of particular interest is the separation of adjacent microphones in the array. From (4.57) it is evident that the term $\psi(\omega)$ and, consequently, the matrix $\boldsymbol{\Xi}$ and the vector $\boldsymbol{\xi}$ will tend to zero as the source-microphone separation is increased. Therefore, for large inter-microphone separation the matrix $\mathbf{T}$ tends to the identity matrix $\mathbf{I}$ and the vector $\mathbf{t}$ tends to zero so that the result in (4.35) tends to the result in (4.26). Furthermore, if estimates of $\mathbf{T}$ and $\mathbf{t}$ were available and since $\mathbf{T}$ is a square matrix, the effects of the spatial cross-correlation could be compensated as $\hat{\mathbf{a}} \cong \mathbf{T}^{-1}(\mathscr{E}\{\hat{\mathbf{b}}_{\mathrm{DSB}}\} + \mathbf{t})$. However, estimating these parameters is difficult in practice. Finally, for the special case where the distance between the microphones is exactly a multiple of a half wavelength at each frequency and the speaker is far from the microphones, then $\psi(\omega) = 0, \forall\omega$ and thus $\boldsymbol{\Xi}$ and $\boldsymbol{\xi}$ from (4.58) and (4.59) are equal to zero. Therefore, the matrix $\mathbf{T}$ becomes exactly the identity matrix $\mathbf{I}$ and the vector $\mathbf{t}$ is exactly zero. This then results in the expression in (4.35) becoming equivalent to that in (4.26).

4.3.2 Effects of Reverberation on the Prediction Residual

Consider a frequency domain formulation of the source-filter speech production model discussed in Sect. 4.2. The speech signal is written as

$$S(e^{j\omega}) = E(e^{j\omega})V(e^{j\omega}), \tag{4.36}$$

where $E(e^{j\omega})$ is the Fourier transform of the prediction residual and $V(e^{j\omega})$ is the transfer function of the all-pole filter from (4.3) evaluated for $z = e^{j\omega}$.

Now consider the speech signal produced in a reverberant room as defined in (4.12) which, in the frequency domain, leads to

$$\begin{aligned} X(e^{j\omega}) &= S(e^{j\omega})H(e^{j\omega}) \\ &= E(e^{j\omega})V(e^{j\omega})H_m(e^{j\omega}). \end{aligned} \tag{4.37}$$

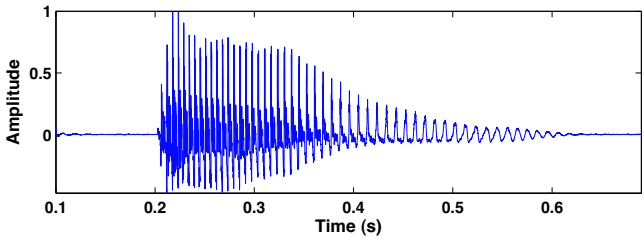


Fig. 4.2 Speech sample used in the experiments comprising the time-domain waveform of the diphthong /eI/ as in the alphabet letter 'a' uttered by a male talker

Referring to (4.26), an inverse filter, $B(e^{j\omega}) = 1 + \sum_{k=1}^{p} b_k e^{j\omega k}$, can be obtained such that $\mathscr{E}\{B(e^{j\omega})\} \cong A(e^{j\omega})$, where $A(e^{j\omega})$ is given by (4.2) for $z = e^{j\omega}$. Filtering the reverberant speech signal with this inverse filter, the coefficients of which are obtained from the reverberant speech signal, results in

$$E_m(e^{j\omega}) \cong E(e^{j\omega}) H_m(e^{j\omega}), \tag{4.38}$$

where $E_m(e^{j\omega})$ is the Fourier transform of the prediction residual, $e_m(n)$, obtained from the reverberant speech observation at the m^{th} microphone. Thus, in the time domain, the prediction residual obtained from reverberant speech is approximately equal to the clean speech residual convolved with the room impulse response. The approximation in (4.38) arises from the LPC. Therefore, if the LPC coefficients used were identical to those from clean speech, the approximation would be an equivalence.

4.3.3 Simulation Examples for LPC on Reverberant Speech

Having established the theoretical relationship between the LPC coefficients obtained from clean speech and those obtained from reverberant speech observations, simulation results are now presented to demonstrate and to validate the theoretical analysis. In summary, two specific points will be demonstrated:

1. On average, over all positions in the room, the LPC coefficients obtained from a single microphone as in (4.15) and those calculated from M-microphones as in (4.29) are not affected by reverberation, while the LPC coefficients from the DSB become more dissimilar from the clean speech coefficients with increased reverberation time.
2. The M-channel LPC coefficients from (4.29) are the most accurate estimates of the clean speech LPC coefficients from the three cases studied.

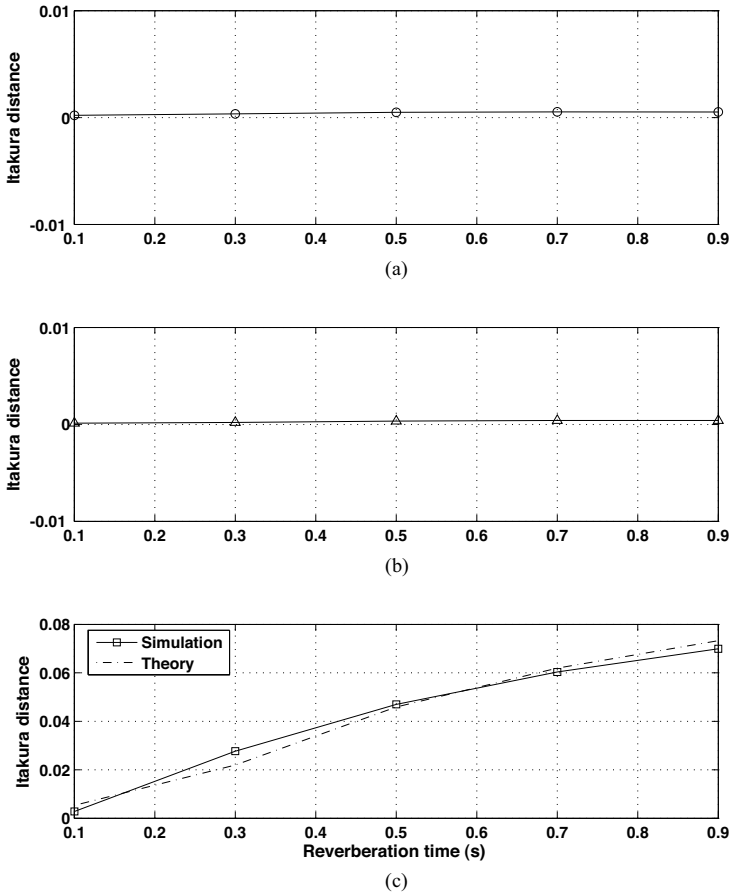


Fig. 4.3 Itakura distance *vs.* reverberation time for the spatially expected LPC coefficients of (a) a single channel, (b) $M = 7$ channels and (c) DSB output simulation (*squares*) and the theoretical expression for the DSB output (4.35) (*dashed line*)

The Itakura distance is used as a similarity measure between two sets of LPC coefficients, defined as [9]

$$d_I = \log\left(\frac{\mathbf{b}^T\mathbf{R}_{ss}\mathbf{b}}{\mathbf{a}^T\mathbf{R}_{ss}\mathbf{a}}\right), \tag{4.39}$$

where $\mathbf{R}_{ss}$ is the autocorrelation matrix of the clean speech signal defined in (4.7), $\mathbf{a}$ is the set of clean speech LPC coefficients and $\mathbf{b}$ are the LPC coefficients under test. The Itakura distance can be interpreted as the log ratio of the minimum mean

squared errors obtained with the true and the estimated coefficients. The denominator represents the optimal solution for clean speech and thus $d_I \geqslant 0$.

The speech sample depicted in Fig. 4.2, comprising the diphthong /eI/ as in the alphabet letter 'a' uttered by a male talker, was used as an example. The LPC analysis was performed using selective linear prediction [25] with a frame length equal to the length of the vowel and a prediction order $p = 21$ with sampling frequency $f_s = 16$ kHz. The prediction order was chosen using the relation $p = \frac{f_s}{1000} + 5$ as recommended in [9]. This gives a pole pair per kHz of Nyquist sampling frequency and some additional poles to model the glottal pulse. Selective linear prediction was employed in the frequency range 0.3–7 kHz, in order to avoid errors due to bandlimiting filters.

The spatial expectation was calculated from $\mathcal{N} = 200$ realizations of the source-array positions within a non-changing acoustic environment and an average autocorrelation function was calculated for each of the cases under consideration. This was repeated, in each new case varying the reverberation time, T_{60}, from 0.1 to 0.9 s. For each case the Itakura distance was calculated for the spatial expectation of the coefficients. Figure 4.3 shows the Itakura distance of the spatially expected LPC coefficients versus reverberation time for (a) a single channel, (b) $M = 7$ channels and (c) the DSB output simulation (solid line) and the theoretical expression for the DSB output in (4.35) (dashed line). It can be seen that the experimental outcome closely corresponds to the theoretical results where the coefficients from the M-channel case and from a single channel are close to the clean speech coefficients. In contrast, the difference between the results from the DSB output and the clean speech increases in a manner proportional to the reverberation time.

The next experiment illustrates the individual outcomes for the three cases at the different locations. The LPC coefficients were computed at each individual source-array position using (4.15), (4.29) and (4.34) and the Itakura distance was then calculated. Figure 4.4 shows the resulting plot in terms of the mean Itakura distance versus increasing reverberation time for (a) a single channel, (b) $M = 7$ channels and (c) the DSB output. The error bars indicate the range between the maximum and the minimum errors, while the crosses indicate the mean value for all $\mathcal{N}$ locations. It can be seen that the M-channel LPC provides the best approximation of the clean speech LPC coefficients. It can also be seen that the estimation error for the LPC coefficients obtained from the DSB output becomes greater with increasing reverberation time. Although this result may appear counterintuitive, it conforms with the theoretical expression in (4.35) and will be clarified further in the following experiment. Figure 4.5 shows examples of the spectral envelopes from the LPC coefficients obtained from reverberant observations using LPC for (a) a single channel, (b) $M = 7$ channels and (c) the DSB output. Each case is compared to the resulting spectral envelope from clean speech.

As discussed in Sect. 4.3.1, the discrepancy in the estimated LPC coefficients at the output of the DSB from those obtained with clean speech is governed mainly by the separation of the microphones. This final experiment demonstrates the effect of the separation between adjacent microphones on the expected LPC coefficients obtained at the output of a DSB. All parameters of the room, the source and the micro-

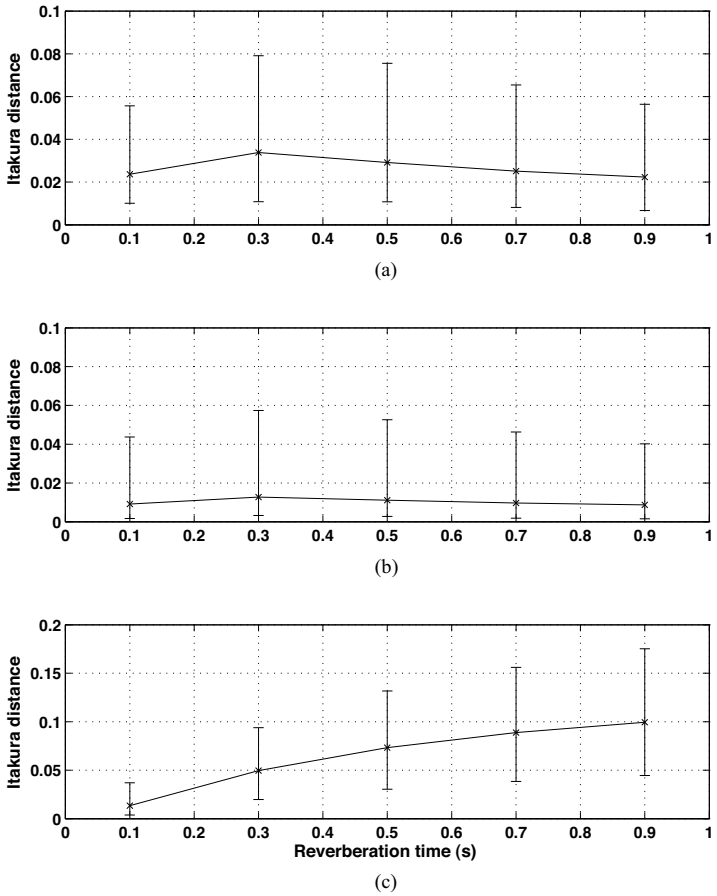


Fig. 4.4 Itakura distance *vs.* reverberation time in terms of the LPC coefficients for each individual outcome for (a) a single channel, (b) M-channels and (c) the DSB output. The *error bars* indicate the maximum and minimum error while *crosses* show the mean value

phone array were kept fixed, while the separation, $\|\mathbf{q}_{\text{mic},m} - \mathbf{q}_{\text{mic},m+1}\|_2$, $1 \leq m < M$, between adjacent microphones in the linear array was increased from 0.05 to 0.3 m in steps of 0.05 m. The results are shown in Fig. 4.6 where the Itakura distance is plotted against microphone separation for (a) the theoretical results calculated with (4.35) (dashed line) and the simulated results (crosses) for the spatially expected LPC coefficients at the output of the DSB and (b) the LPC coefficients for each individual outcome. Error bars indicate the maximum and the minimum errors, while crosses indicate the mean value. It is seen from these results that the estimates at the output of the DSB become more accurate as the distance between the microphones is increased. At a microphone separation of $\|\mathbf{q}_{\text{mic},m} - \mathbf{q}_{\text{mic},m+1}\|_2 = 0.3$ m the results

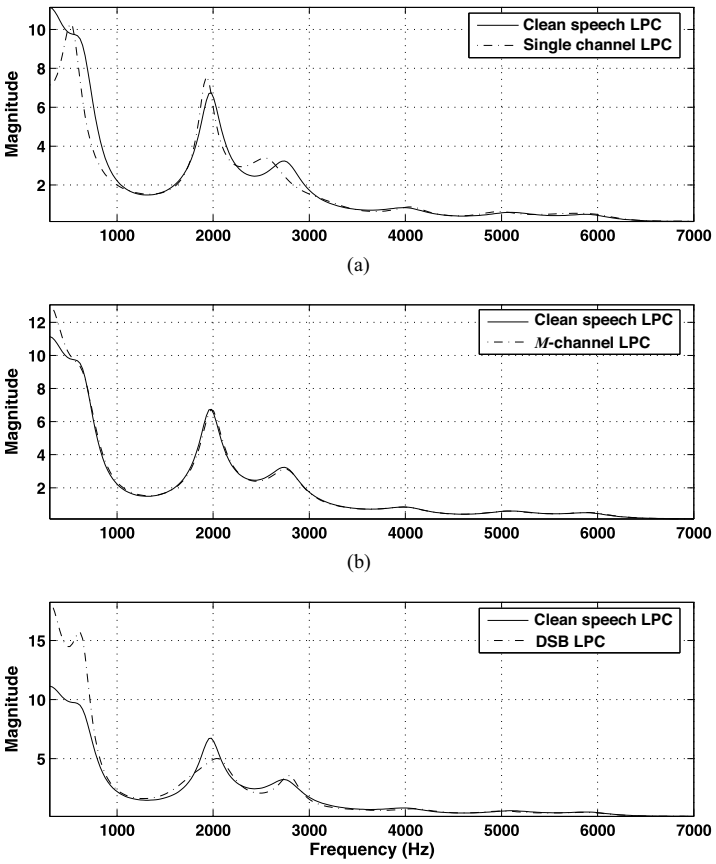


Fig. 4.5 Spectral envelopes calculated from the LPC coefficients of clean speech compared with spectral envelopes obtained from the LPC coefficients of (a) a single channel, (b) $M = 7$ channels and (c) the DSB output

are comparable to the M-channel case both in terms of spatial expectation and of the individual outcomes. This is because the spatial correlation between microphones becomes negligible.

Finally, Fig. 4.7 shows a single channel example of portions of clean voiced speech, reverberant voiced speech and the corresponding prediction residuals. This example is from a simulated rectangular room with dimensions $6.4 \times 5 \times 4$ m, a source positioned at 1.5 m from the microphone, and reverberation time set to $T_{60} = 0.5$ s. The Acoustic Impulse Response (AIR) is shown in Fig. 4.8. It can be seen in Fig. 4.7 (d) that the reverberant residual contains several peaks of similar strength as the true excitation peaks. This leads to a definition adopted in the remainder of this

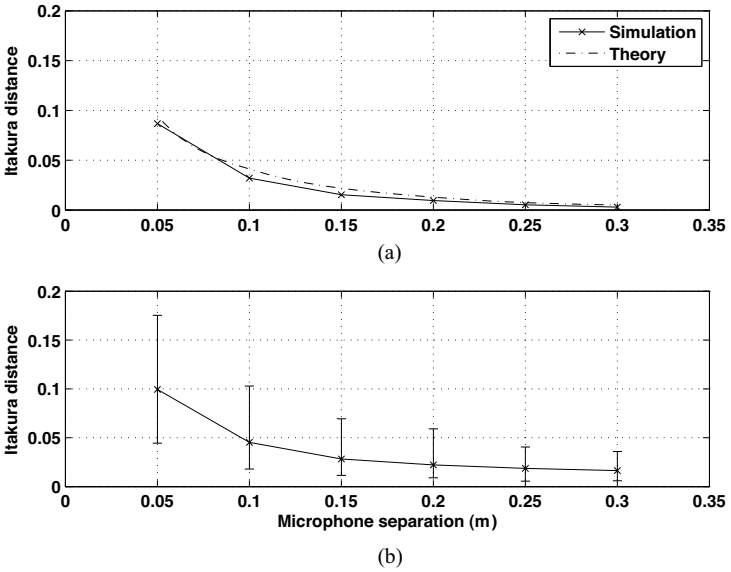


Fig. 4.6 Itakura distance *vs.* microphone separation for (a) the theoretical results calculated with (4.35) (*dashed line*) and the simulated results (*crosses*) for the spatially expected LPC coefficients at the output of the DSB and (b) the LPC coefficients for each individual outcome; *error bars* indicate the maximum and the minimum errors and *crosses* show the mean value

chapter: *an erroneous peak* is a pulse in the prediction residual of reverberant speech that is of comparable strength to the true excitation peak. As source-microphone separation and reverberation time increase, the contribution of such erroneous peaks becomes more significant, resulting in large distortion in the prediction residual.

In summary, statistical room acoustics theory has been used for the analysis of the LPC modelling of reverberant speech. Investigating three scenarios, it has been shown that, in terms of spatial expectation, the LPC coefficients calculated from reverberant speech are approximately equivalent to those from clean speech both in the single channel case and in the case when the coefficients are calculated jointly from an M-channel observation. Furthermore, it was shown that the LPC coefficients calculated at the output of a DSB differ from the clean speech coefficients due to spatial correlation, which is governed by the room characteristics and the microphone array geometry. It was also demonstrated that LPC coefficients calculated jointly in the M-channel observation provide the best approximation of the clean speech coefficients at individual source-microphone positions. Thus, the M-channel joint calculation of the LPC coefficients is preferred where such an equivalence is important. Finally, the findings in this chapter are of particular interest in speech dereverberation methods using prediction residual processing, where the main and crucial assumption is that reverberation mostly affects the prediction residual. Since

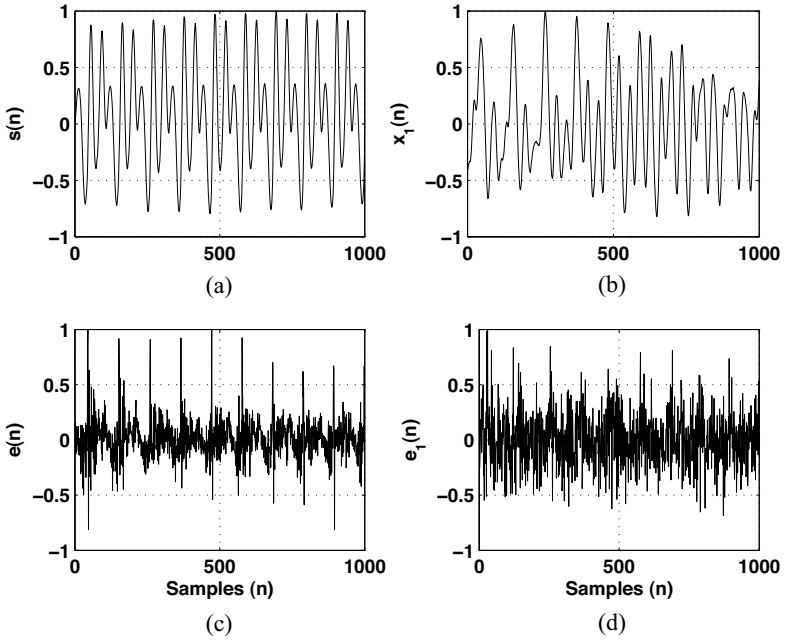


Fig. 4.7 (a) Clean speech, (b) reverberant speech, (c) clean speech prediction residual and (d) reverberant speech prediction residual

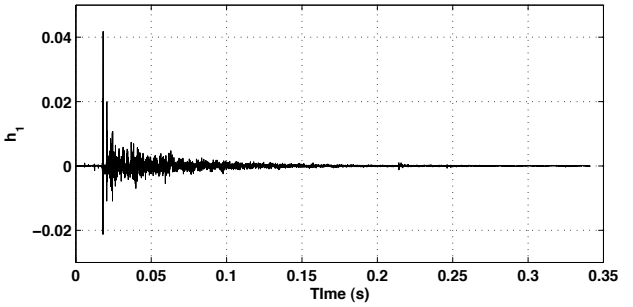


Fig. 4.8 Measured AIR used in the example of Figs. 4.7(b) and (d) for source-microphone separation $D = 1.5$ m

most of these methods utilize microphone arrays for the residual processing, M-channel joint calculation of the LPC coefficients should normally be deployed to ensure the validity of this assumption.

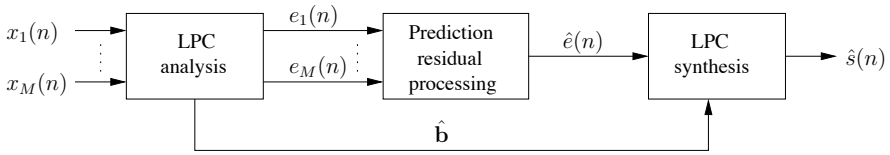


Fig. 4.9 Reverberant speech enhancement using LPC

4.4 Dereverberation Employing LPC

We have so far established that LPC coefficients can be calculated from reverberant speech signals and that the major effect of reverberation resides in the prediction residual. Consequently, signal processing to reduce reverberation can be applied to the prediction residual directly, and the *a priori* information about the structure of the prediction residual for voiced speech, obtained from the source-filter model [28], can be utilized to enhance the reverberant speech. An approach to reducing reverberation can therefore be achieved by attenuating the erroneous peaks in the prediction residuals obtained from the reverberant observations and then synthesizing the speech signal using the processed prediction residual with the all-pole filter calculated from the reverberant speech. The general procedure of the use of linear prediction for reverberant speech enhancement is shown in Fig. 4.9. LPC analysis is performed on the reverberant observations, $x_m(n)$, $m = 1,2,\ldots,M$ in order to obtain the prediction residuals $e_m(n)$ and a set of LPC coefficients, $\hat{\mathbf{b}}$, which are an estimate of the clean speech coefficients, $\mathbf{a}$. The prediction residuals from the M microphone speech signals are then processed to find an estimate of the clean speech residual, $\hat{e}(n) \approx e(n)$. Finally, a clean speech estimate, $\hat{s}(n)$, is found by synthesis using $\hat{e}(n)$ and $\hat{b}_i$ such that

$$\hat{s}(n) = \sum_{i=1}^{p} \hat{b}_i \hat{s}(n-i) + \hat{e}(n). \tag{4.40}$$

An attractive feature of the prediction residual processing methods is that they can reduce the effects of reverberation without specific knowledge of the acoustic transfer function, which is generally not available and both difficult and computationally expensive to estimate. Therefore, it makes these algorithms practical and suitable for online implementation. Another advantage of manipulating the prediction residual instead of the speech signal is that any estimation errors are smoothed by the characteristics of the all-pole synthesis filter [40].

Various methods for processing the prediction residual resulting from the LPC analysis of reverberant speech have been proposed in the literature and they will be summarized and discussed in this section. Most of these methods make use of multi-microphone systems, which is beneficial since in the case of time-aligned signals, peaks due to the original excitation are correlated across the channels, while those due to the acoustic impulse response are not.

4.4.1 Regional Weighting Function

Yegnanarayana *et al.* [40] provided a comprehensive study of the reverberant speech prediction residual. They have demonstrated that reverberation affects the prediction residual differently in different speech segments, depending on the energy in the signal and whether a segment is voiced or unvoiced. Motivated by these observations, the authors propose to use *a regional weighting function* based on the signal-to-reverberant ratio (SRR) in each region and also a *global weighting function* derived from the short term signal energy. For the derivation of the SRR based weightings, the entropy function and the normalized error are used. This algorithm is only appropriate in the case of small amounts of reverberation; however, it is able to operate on a single channel.

4.4.2 Weighting Function Based on Hilbert Envelopes

Yegnanarayana and Satyanarayana [39] propose *a weighting function based on Hilbert envelopes*. The authors use the Hilbert envelopes of the prediction residuals of multiple channels to represent the strength of the peaks in the residuals. These Hilbert envelopes are then time-aligned and summed, resulting in a signal that emphasizes the positions of the true excitation peaks. This weighting function is applied to the prediction residual of one of the channels.

4.4.3 Wavelet Extrema Clustering

An approach proposed by Brandstein and Griebel [5, 16], based on an idea inherited from speech de-noising, is *wavelet extrema clustering*. This method is based on the assumption that the prediction residual peaks due to the true excitation sequence are correlated among the channels, while the remaining impulses from the multipath are not. The prediction residuals are transformed into the wavelet domain and extrema clusters among the channels at each scale are identified and used to reconstruct an estimate of the clean residual signal.

4.4.4 Weight Function from Coarse Channel Estimates

An alternative approach by Griebel and Brandstein [17] uses a *weight function based on coarse estimates of the room impulse responses*. The authors show that coarse estimates of the acoustic impulse responses can be obtained by averaging the phase transform version of the generalized cross-correlation [20] from the multiple microphones. These estimates are then applied in a matched filtering operation to obtain a

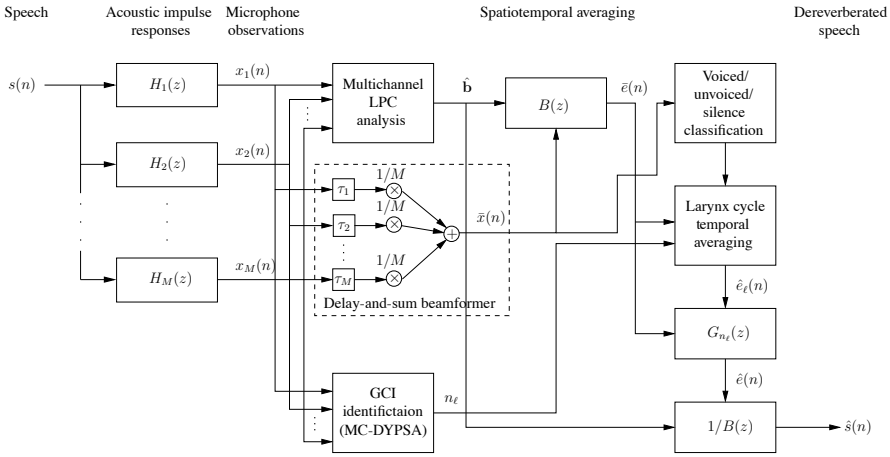


Fig. 4.10 System diagram of the spatiotemporal averaging method for enhancement of reverberant speech

weighting function for the prediction residuals of the M microphones. The enhanced speech signals at the output of each microphone are finally used in a beamformer to produce the dereverberated speech signal. This method requires a large number of microphones for the coarse channel estimates. The results in [17], for example, were generated using $M = 15$ microphones.

4.4.5 Kurtosis Maximizing Adaptive Filter

An adaptive algorithm was proposed by Gillespie *et al.* [14] using a *kurtosis maximizing subband adaptive filter*. The authors demonstrate that the kurtosis of the prediction residual decreases as a function of increased reverberation, which was also suggested in [40]. They use this observation to derive an adaptive filter that maximizes the kurtosis of the prediction residual. The filter is applied directly to the observed signal rather than to the prediction residual and so avoids the LPC synthesis stage. This also lessens the dependence on the LPC coefficients to some extent. The adaptive filter is implemented in a multichannel subband framework for increased efficiency. This is an example of a hybrid approach drawing on elements of both speech enhancement and blind system identification and inversion. An extension to this method was presented in [37], where it was combined with spectral subtraction to remove residual reverberation due to late reflections.

4.5 Spatiotemporal Averaging Method for Enhancement of Reverberant Speech

Many methods for LPC based enhancement tend to ignore the specific characteristics of the prediction residual both for the peaks due to the original excitation and the information contained between the Glottal Closure Instants (GCIs). Modifying the excitation peaks or excessively flattening the waveform between such peaks will result in distortions in the reconstructed speech signal and render it less natural [38]. Furthermore, most methods do not consider the unvoiced/silent speech segments in the dereverberation process. In the following, a method is described that addresses these issues; we refer to this method as SMERSH – Spatiotemporal averaging Method for Enhancement of Reverberant SpeecH [11, 12, 33].

The observed speech signals are first spatially averaged using the DSB defined in (4.33). Consider Fig. 4.11, which shows a portion of the prediction residual obtained from (a) clean speech, (b) reverberant speech, and (c) speech at the output of a DSB. The effect of reverberation on the prediction residual can be clearly seen in the form of many random peaks of similar strength to the periodic peaks occurring at the GCIs in clean speech. The following specific observations can be made from the prediction residuals in this and in other examples:

(i) The prediction residual obtained by performing LPC on the DSB output differs from that obtained from the clean speech by seemingly random peaks that are left unattenuated after the spatial averaging; these appear uncorrelated among consecutive larynx cycles.
(ii) The main features in consecutive larynx cycles of the clean speech prediction residual change slowly and show high inter-cycle correlation.
(iii) Strong periodic peaks in the prediction residual from the DSB output appear to represent the GCIs seen in the clean speech.

Property (i) arises from the quasi-periodic nature of voiced excitation. Property (ii) is well-known in speech processing and has been applied in, for example, the Time-Domain Pitch Synchronous Overlap Add Method (TD-PSOLA) for pitch modification [9]. Motivated by these observations, it is suggested that applying a moving average operation on neighbouring larynx cycles in voiced speech will suppress the uncorrelated features and, hence, enhance the prediction residual. There are two key issues to consider. First, it is necessary to identify correctly the peaks that belong to the original excitation so as to segment the larynx cycles. Secondly, peaks attributed to GCIs are important to speech quality [38] and should remain unchanged; therefore they should be excluded from the averaging process.

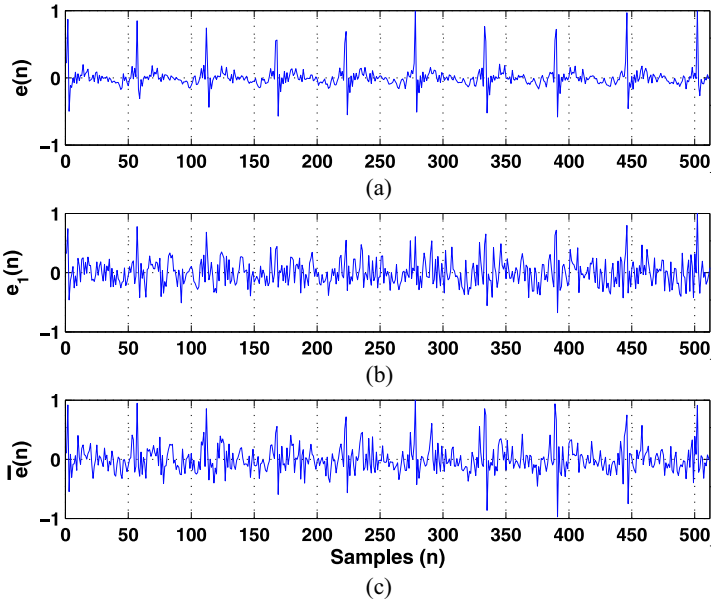


Fig. 4.11 Prediction residuals obtained from (a) clean speech, (b) reverberant speech and (c) spatially averaged speech

The algorithm comprises four major components:

- Time-delay-of-arrival estimation, which is used to time align the speech signals, such that their direct-path components of each channel coincide, prior to spatial averaging.
- GCI detection so that the prediction residual can be segmented into individual larynx cycles, where a larynx cycle is taken between two consecutive GCIs.
- Temporal averaging of two or more neighboring larynx cycles to obtain an enhanced larynx-cycle.
- Voiced/univoiced/silcence detection.

Each of these components will now be discussed in detail.

4.5.1 Larynx Cycle Segmentation with Multichannel DYPSA

One of the key components in SMERSH is the accurate larynx cycle segmentation of the prediction residual. This requires a GCI detection algorithm that is robust to noise and reverberation. One such algorithm is the DYnamic Programming Phase-Slope Algorithm (DYPSA) [26], which comprises three main parts:

1. *Group delay function* – this is defined as the average slope of the unwrapped phase spectrum of the short time Fourier transform of the prediction residual. GCI candidates are selected based on the negative-going zero crossings of the group delay function.
2. *Phase-slope projection* – this is introduced to generate GCI candidates when a local maximum is followed by a local minimum without the group delay function crossing zero. The midpoint between such turning points is identified and projected onto the time axis with unit slope. In this way, GCIs whose negative-going slope does not cross the zero point (i.e., those GCIs missed by the group delay function) are correctly identified.
3. *Dynamic programming* – this uses known characteristics of voiced speech (such as pitch consistency and waveform similarity across larynx cycles) and forms a cost function that is minimized in order to select a subset of the GCI candidates that are most likely to correspond to the true GCIs. The subset of candidates is selected according to the minimization problem defined as

$$\min_{\mathbf{\Omega}} \sum_{r=1}^{|\mathbf{\Omega}|} \boldsymbol{\lambda}^T \mathbf{c}_{\Omega}(r), \tag{4.41}$$

where $\mathbf{\Omega}$ is a subset of GCIs of size $|\mathbf{\Omega}|$, $\boldsymbol{\lambda}$ is a vector of experimentally determined weighting factors and $\mathbf{c}_{\Omega}(r)$ is a vector of cost elements evaluated at the r^{th} GCI of the subset $\mathbf{\Omega}$.

Multichannel DYPSA (MC-DYPSA) was proposed in [34] to exploit the spatial diversity of acoustic transfer functions [21]. When the channels are time-aligned, the direct-path signal is common to all channels but reverberation components are less likely to show correlation. MC-DYPSA applies parts (i) and (ii) above to each channel independently and creates an additional cost element based upon the inter-channel correlation, penalizing those which occur in a small number of channels and encouraging those in close temporal proximity across channels. This is passed to the dynamic programming stage and the most likely GCIs, at sample instants n_ℓ, are identified. Experiments in [34] have shown that GCI estimation from a reverberant speech signal for $T_{60} = 500$ ms is on average 16% more accurate with MC-DYPSA than single-channel DYPSA applied to the output of an 8-channel DSB and 29% more accurate than DYPSA on a single channel, providing 83% accuracy [34].

4.5.2 Time Delay of Arrival Estimation for Spatial Averaging

Both MC-DYPSA and spatial averaging rely on the correct inter-channel time alignment so as to maximize the correlation of the direct-path signal across channels. The Generalized Cross-Correlation Phase Transform (GCC-PHAT) [20] is a simple and sufficiently accurate method for the estimation of delay between two channels from moderately reverberant speech signals [7, 8].

Let the reference channel be $x_{ref}(n)$ and the measurement channel $x_m(n)$. The delay estimate at the m^{th} channel, $\hat{\tau}_m$, is determined by maximizing the cross-correlation between channels

$$\hat{\tau}_m = \arg\max_{\tau} r_{x_{ref}x_m}(\tau), \tag{4.42}$$

with

$$r_{x_{ref}x_m}(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{X_{ref}(e^{j\omega})X_m^*(e^{j\omega})}{|X_{ref}(e^{j\omega})||X_m^*(e^{j\omega})|} e^{j\omega\tau} \mathrm{d}\omega, \tag{4.43}$$

where X^* denotes the complex conjugate of X and $r_{x_{ref}x_m}(\tau)$ is a weighted inverse Fourier transform of the signal cross-spectra for time lag τ.

The GCC-PHAT method has been shown to be accurate enough for moderate reverberation although it is suboptimal under ideal conditions as it places equal weighting on each frequency [3, 20]. The process is repeated for $M-1$ pairs of microphones to determine the inter-channel delay between microphone $m=1$ and microphone $m=2,3,\ldots,M$.

The spatial averaging of the speech signals is performed with the DSB defined in (4.33) and the estimated delays from (4.42) according to:

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n-\hat{\tau}_m). \tag{4.44}$$

4.5.3 Voiced/Unvoiced/Silence Detection

Voiced/unvoiced/silence detection is performed on a speech signal that has been processed with the DSB in (4.44). Voiced segments are determined using a voiced-unvoiced-silence detector based on five measurements [1]: (1) zero crossing rate, (2) energy, (3) autocorrelation coefficient, (4) the first LPC coefficient and (5) normalized prediction error (in dB). Each measure is computed over 32 ms frames with 60% overlap, forming a sequence of feature vectors. These vectors are then clustered using an unsupervised Expectation Maximization (EM) algorithm [10]. The three clusters are labelled as silence, unvoiced and voiced according to their mean vectors and variances. The unvoiced cluster is chosen to be the one with an autocorrelation coefficient closest to zero mean and 0.5 variance. Of the remaining two clusters, the one corresponding to the high speech energy is chosen to be voiced. Every vector in the sequence is then evaluated under each of the three Gaussians and classified according to which cluster produces the highest likelihood.

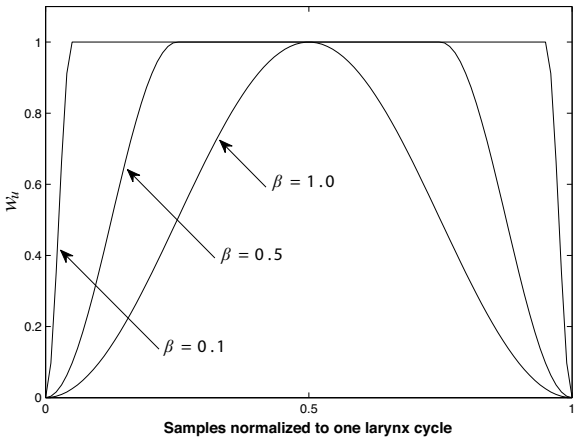


Fig. 4.12 Larynx weighting function defined in (4.45) with different values for β

4.5.4 Weighted Inter-cycle Averaging

In order to leave the glottal pulse undisturbed at GCIs, a weighting function is applied on each larynx cycle prior to the averaging. The weighting function should, ideally, exclude only the true glottal pulse. However, in practice, GCIs are identified to an uncertainty in the order of 1 ms [26] and the glottal pulse is not a true impulse but is somewhat spread in time [9]. A weighting function that has been found suitable, with a reasonable trade-off between the issues described above, is the time-domain Tukey window defined as [19]

$$w_u = \begin{cases} 0.5 + 0.5\cos\left(\frac{2\pi u}{\beta(\mathcal{L}-1)} - \pi\right), & u < \frac{\beta\mathcal{L}}{2}, \\ 0.5 + 0.5\cos\left(\frac{2\pi}{\beta} - \frac{2\pi u}{\beta(\mathcal{L}-1)} - \pi\right), & u > \mathcal{L} - \frac{\beta\mathcal{L}}{2} - 1, \\ 1.0, & \text{otherwise}, \end{cases} \tag{4.45}$$

where $\mathcal{L}$ is the length of one larynx cycle (in samples) and $0 \leq \beta \leq 1$ is the taper ratio of the window. An example of the weighting function with three different values for β is shown in Fig. 4.12. The taper ratio offers a tunable parameter with the beneficial ability to control the amount of the larynx cycle to be included in the averaging process and can be adjusted, for example, in some proportion to the estimation error variance of the GCI identification algorithm. Following the averaging procedure, the inverse weight function with weights , $1 - w_u$, is applied to the larynx cycle under consideration to restore the original glottal pulse shape.

Thus, each enhanced larynx cycle in a voiced speech segment is obtained by averaging the current weighted larynx cycle under consideration with $\mathcal{I}$ of its neigh-

bouring weighted larynx cycles. The result is then added to the original larynx cycle weighted with the inverse weight function. The final expression for the ℓ^{th} enhanced larynx cycle becomes

$$\hat{\mathbf{e}}(n_\ell) = (\mathbf{I} - \mathbf{W})\bar{\mathbf{e}}(n_\ell) + \frac{1}{2\mathcal{I}+1} \sum_{i=-\mathcal{I}}^{\mathcal{I}} \mathbf{W}\bar{\mathbf{e}}(n_{\ell+i}), \tag{4.46}$$

where

$$\bar{\mathbf{e}}(n_\ell) = [\bar{e}(n_\ell)\ \bar{e}(n_\ell+1)\ \ldots\ \bar{e}(n_\ell+\mathcal{L}-1)]^T$$

is the ℓ^{th} larynx cycle at the output of the DSB with its GCIs at time n_ℓ,

$$\hat{\mathbf{e}}(n_\ell) = [\hat{e}(n_\ell)\ \hat{e}(n_\ell+1)\ \ldots\ \hat{e}(n_\ell+\mathcal{L}-1)]^T$$

is the ℓ^{th} larynx cycle of the enhanced residual, $\mathbf{I}$ is the identity matrix and

$$\mathbf{W} = \text{diag}\{w_0\ w_1\ \ldots\ w_{\mathcal{L}-1}\}$$

is a diagonal weighting matrix with the weights calculated with (4.45). Larynx cycles are not strictly periodic but may be assumed to vary by a few samples over a neighborhood defined by $\mathcal{L}$. Therefore, $\mathcal{L}$ is set to equal the length of the larynx cycle being processed; other larynx cycles used in the averaging that have less than $\mathcal{L}$ samples are padded with zeros, while those with more than $\mathcal{L}$ samples are truncated.

The choice of $\mathcal{I}$ is important. If too many cycles are included, the averaging will remove uncorrelated portions from the original excitation, whereas if too few cycles are considered, erroneous peaks due to reverberation will remain. For the results presented here, the number of cycles for averaging was set to $\mathcal{I} = 4$. It was found through several experiments that this is a good choice in general and that $\mathcal{I} > 4$ provides less accurate results.

This averaging process can only be applied in the form as so far described to segments of voiced speech, leaving reverberation components of unvoiced speech and silence unaffected. Furthermore, in the case of an erroneous GCI, the algorithm will produce incorrect results. To improve robustness, a dereverberating L_g-tap equalization filter with taps

$$\mathbf{g} = [g_0\ g_1\ \cdots\ g_{L_g-1}]^T$$

for the ℓ^{th} larynx cycle is defined, which performs the equivalent operation of temporal averaging. An estimate of $\mathbf{g}$ is found by solving the following optimization problem

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}} \|\mathbf{g}^T\bar{\mathbf{e}}(n_\ell) - \hat{e}(n_\ell)\|_2^2, \tag{4.47}$$

whose least squares solution can be found to be

$$\hat{\mathbf{g}} = \mathbf{R}_{\bar{e}\bar{e}}^{-1}\mathbf{r}_{\bar{e}\hat{e}}, \tag{4.48}$$

where $\mathbf{R}_{\bar{e}\bar{e}}$ is an autocorrelation matrix formed from $\bar{\mathbf{e}}(n_\ell)$ and $\mathbf{r}_{\bar{e}\hat{e}}$ is a cross-correlation vector formed from $\bar{\mathbf{e}}(n_\ell)$ and $\hat{\mathbf{e}}(n_\ell)$.

The filter in (4.48) is used to update a time-varying filter

$$\hat{\mathbf{g}}(n_\ell) = \gamma\hat{\mathbf{g}}(n_{\ell-1}) + (1-\gamma)\hat{\mathbf{g}}, \tag{4.49}$$

where $0 \leq \gamma \leq 1$ is a forgetting factor with typical values in the range $\{0.1 - 0.3\}$, initialized to $\hat{\mathbf{g}}(0) = [1\ 0\ ...\ 0]^T$. It is updated only during voiced speech, with the latest iteration used for periods of unvoiced speech or silence. The complete SMERSH is summarized in Algorithm 4.1.

4.5.5 Dereverberation Results

We now present results to demonstrate the performance of the spatiotemporal averaging method for enhancement of reverberant speech. The performance is compared with the delay-and-sum beamformer using data captured in an office room. A microphone array consisting of eight AKG C417 microphones spaced linearly at 0.05 m intervals, was placed in a $3.3 \times 2.9 \times 2.9$ m room with reverberation time (T_{60}) of 0.3 s. Utterances of the sentence "George made the girl measure a good blue vase" by five male and five female talkers were taken from the APLAWD database [23] and played through a GENELEC 8030 loudspeaker at distances 0.5 to 2 m from the centre of the microphone array. The AIRs between the loudspeaker and each microphone was estimated using the Maximum Length Sequence (MLS) method [21].

Recording and channel alignment were made at a sampling frequency of $f_s =$ 48 kHz. The remainder of the processing was performed at $f_s = 16$ kHz and with the samples high pass filtered at 100 Hz. The recorded speech, the speech at the output of the DSB and the speech processed with SMERSH were evaluated against the clean speech samples using the segmental Signal to Reverberation Ratio (SRR) defined in (2.45) and (2.46) and Bark Spectral Distortion (BSD) defined in (2.38) using 30 ms frames with 50% overlap. The MLS-derived channel estimates were truncated to determine a direct-path impulse response, $h_{\mathrm{d}}(n)$, which was convolved with the clean speech signal to align the unprocessed and processed signals, denoted $s_{\mathrm{d}}(n) = h_{\mathrm{d}}(n) * s(n)$ which were used in the evaluation procedure as discussed in Chap. 2.

The results in terms of segmental SRR, averaged over all ten talkers in APLAWD, are shown in Fig. 4.13 for (a) reverberant speech at the microphone closest to the talker, (b) speech at the output of the DSB and (c) speech processed with SMERSH. Corresponding BSD results are shown in Fig. 4.14. Reverberation and noise reduction of up to 5.0 dB and 0.33 in BSD score are observed at a distance of 2 m, corresponding to 2.7 dB and 0.07 over the DSB. Informal listening tests carried out indicate that, perceptually, reverberation effects are reduced and the talker appears to be closer to the microphone. The results show a strong correlation with the simulations in [12].

Algorithm 4.1 SMERSH: Spatiotemporal Averaging Method for Enhancement of Reverberant Speech

1. Calculate the LPC coefficients using M-channel LPC as in (4.29):

$$\hat{\mathbf{b}}_M = \bar{\mathbf{R}}^{-1}\bar{\mathbf{r}}.$$

2. Find time delays of arrival using GCC-PHAT in (4.42):

$$\hat{\tau}_m = \arg\max_{\tau} r_{x_{ref}x_m}(\tau).$$

3. Apply the DSB to the M microphone signals to obtain $\bar{x}(n)$ as in (4.33):

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n-\hat{\tau}_m).$$

4. Apply the filter from (4.3) with coefficients $\hat{\mathbf{b}}_M$ to the DSB output to obtain the prediction residual:

$$\bar{e}(n) = \hat{\mathbf{b}}_M^T\bar{\mathbf{x}}(n).$$

5. Use MC-DYPSA to identify the GCIs from the excitation peaks, n_ℓ in the prediction residual $\bar{e}(n)$, and segment into larynx cycles $\bar{\mathbf{e}}(n_\ell)$.
6. For each larynx cycle, $\ell = 1, 2, \ldots$:

   6.1 Calculate the temporally averaged larynx cycle according to (4.46):

$$\hat{\mathbf{e}}(n_\ell) = (\mathbf{I}-\mathbf{W})\bar{\mathbf{e}}(n_\ell) + \frac{1}{2\mathcal{I}+1}\sum_{i=-\mathcal{I}}^{\mathcal{I}}\mathbf{W}\bar{\mathbf{e}}(n_{\ell+i}).$$

   6.2 Update the time-varying filter $\hat{\mathbf{g}}(n_\ell)$ according to (4.47):

$$\hat{\mathbf{g}}(n_\ell) = \gamma\hat{\mathbf{g}}(n_{\ell-1}) + (1-\gamma)\hat{\mathbf{g}},$$

   with the ℓ^{th} filter calculated as in (4.48)

$$\hat{\mathbf{g}} = \mathbf{R}_{\bar{e}\bar{e}}^{-1}\mathbf{r}_{\bar{e}\hat{e}}.$$

   6.3 Apply filter to the prediction residual from the DSB output until the beginning of the next larynx cycle, $\ell+1$:

$$\hat{e}(n) = \hat{\mathbf{g}}^T(n_\ell)\bar{\mathbf{e}}(n_\ell).$$

   6.4 Obtain an estimate of the clean speech signal, $\hat{s}(n)$, by synthesis using the enhanced residual $\hat{e}(n)$ and the filter from (4.3) with LPC coefficients, $\hat{\mathbf{b}}_M$:

$$\hat{s}(n) = [\hat{\mathbf{b}}_M^{-1}]^T\hat{\mathbf{e}}(n),$$

   where $\hat{\mathbf{b}}_M^{-1}$ represents the all-pole filter coefficients corresponding to $\hat{\mathbf{b}}_M$.

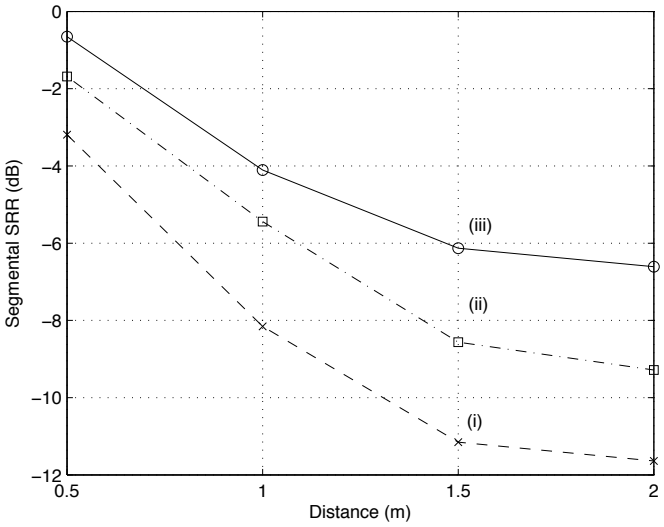


Fig. 4.13 Segmental SRR *vs.* distance for (i) reverberant speech, (ii) DSB processed speech and (iii) speech processed with SMERSH

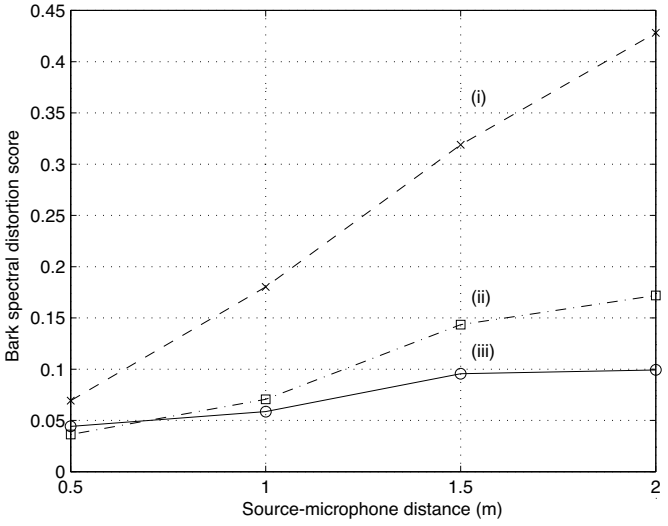


Fig. 4.14 BSD *vs.* distance for (i) reverberant speech, (ii) DSB processed speech and (iii) speech processed with SMERSH

4.6 Summary

We have discussed reverberant speech enhancement with LPC-based approaches. Linear prediction coding of speech was first reviewed and statistical room acoustic theory was used for the analysis of the LPC applied to reverberant speech, showing that reverberation primarily affects the prediction residual and to lesser extent the LPC coefficients. By investigating three scenarios it has been shown that, in terms of spatial expectation, the LPC coefficients calculated from reverberant speech are approximately equivalent to those from clean speech both in the single channel case and in the case when the coefficients are calculated jointly from a multichannel observation. Furthermore, it was shown that the LPC coefficients calculated at the output of a delay-and- sum beamformer differ from the clean speech coefficients due to spatial correlation, which is governed by the room characteristics and the microphone array geometry. It was also demonstrated that LPC coefficients calculated jointly on the multichannel observation provide the best approximation of the clean speech coefficients at individual source-microphone positions. Therefore, multichannel joint calculation of the LPC coefficients is the preferred option where such an equivalence is important.

The general concept of linear prediction in dereverberation was described and existing approaches for prediction residual enhancement were reviewed. A multi-microphone method to prediction residual enhancement based on spatial averaging of the observed signals and temporal averaging of neighbouring larynx cycles was described in detail. The performance of the algorithm was illustrated through experiments in a real office with a T_{60} of 0.3 s. The experiments demonstrated the dereverberation and noise suppression of the spatiotemporal averaging method, showing up to a 5 dB improvement in segmental SRR and 0.33 in normalized Bark spectral distortion score.

Appendix A

Consider a speech signal source, $s(n)$, observed with M microphones and combined using a DSB to give a signal $\bar{x}(n)$ according to (4.33). In the frequency domain this can be expressed as

$$\begin{aligned}\bar{X}(e^{j\omega}) &= \left(\frac{1}{M}\sum_{m=1}^{M} H_m(e^{j\omega})e^{-j2\pi f\tau_m}\right) S(e^{j\omega}) \\ &= \bar{H}(e^{j\omega})S(e^{j\omega}),\end{aligned} \tag{4.50}$$

where $\bar{X}(e^{j\omega})$, $S(e^{j\omega})$ are the Fourier transforms of $\bar{x}(n)$ and $s(n)$, respectively, $H_m(e^{j\omega})$ is the ATF with respect to the m^{th} microphone and $\bar{H}(e^{j\omega})$ is the averaged ATF at the output of the DSB. The optimum estimate of the LPC coefficients, $\mathbf{b}_{\text{DSB}}$, at the beamformer output are calculated as in Sect. 4.3.1

$$\hat{\mathbf{b}}_{\mathrm{DSB}} = \mathbf{R}_{\bar{x}\bar{x}}^{-1}\mathbf{r}_{\bar{x}\bar{x}}, \tag{4.51}$$

where $\mathbf{R}_{\bar{x}\bar{x}}$ and $\mathbf{r}_{\bar{x}\bar{x}}$ are a $p \times p$ autocorrelation matrix and a $p \times 1$ autocorrelation vector respectively with the i^{th} correlation coefficient being

$$\begin{aligned} r_{\bar{x}\bar{x},i} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{X}(e^{j\omega})|^2 e^{j\omega i}\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{H}(e^{j\omega})|^2 |S(e^{j\omega})|^2 e^{j\omega i}\,\mathrm{d}\omega,\ i = 1,2,\ldots,p \end{aligned} \tag{4.52}$$

From this point on, we omit the frequency index for reasons of clarity. The expected energy density spectrum of the averaged ATFs can be written

$$\mathscr{E}\{|\bar{H}|^2\} = \frac{1}{M^2}\left[\sum_{m=1}^{M}\mathscr{E}\{|H_m|^2\} + \sum_{m=1}^{M}\sum_{\substack{l=1\\ l\neq m}}^{M}\mathscr{E}\{H_m H_l^*\}e^{-j2\pi f(\tau_m-\tau_l)}\right]. \tag{4.53}$$

From (4.24), the expected energy density for the m^{th} channel is

$$\mathscr{E}\{|H_m|^2\} = \frac{1}{(4\pi D_m)^2} + \left(\frac{1-\bar{\alpha}}{\pi A\bar{\alpha}}\right), \tag{4.54}$$

where D_m is the Euclidean distance between the source and the m^{th} microphone, $\bar{\alpha}$ is the average room absorption coefficient and A is the total room surface area.

The expected cross-correlation between the m^{th} and the l^{th} microphones can be shown to be [13, 36]

$$\mathscr{E}\{H_m H_l^*\} = \frac{e^{jk(D_m-D_l)}}{16\pi^2 D_m D_l} + \left(\frac{1-\bar{\alpha}}{\pi A\bar{\alpha}}\right)\frac{\sin k\Delta_{lm}}{k\Delta_{lm}}, \tag{4.55}$$

where

$$\Delta_{lm} = \|\mathbf{q}_{\mathrm{mic},m} - \mathbf{q}_{\mathrm{mic},l}\|_2 \tag{4.56}$$

is the distance between the m^{th} and the l^{th} microphones.

By substituting (4.54) and (4.55) into (4.53) and with $\tau_m = D_m/c$, we obtain the following expression for the mean energy density at the DSB output:

$$\mathscr{E}\{|\bar{H}|^2\} = \bar{\kappa} + \psi(\omega), \tag{4.57}$$

with

$$\bar{\kappa} = \frac{1}{(4\pi M)^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\frac{1}{D_m D_l} + \left(\frac{1-\bar{\alpha}}{M\pi A\bar{\alpha}}\right)$$

and

$$\psi(\omega) = \left(\frac{1-\bar{\alpha}}{M^2\pi A\bar{\alpha}}\right)\sum_{m=1}^{M}\sum_{\substack{l=1\\ l\neq m}}^{M}\frac{\sin k\Delta_{lm}}{k\Delta_{lm}}\cos(k[D_m - D_l]),$$

where $\bar{\kappa}$ is a frequency independent component and $\psi(\omega)$ is a component due to spatial correlation.

Now, let

$$\xi_u = \frac{1}{2\pi}\int_{-\pi}^{\pi} \psi(\omega)|S(e^{j\omega})|^2 e^{j\omega u}\,\mathrm{d}\omega \tag{4.58}$$

and

$$\Xi_{u,v} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \psi(\omega)|S(e^{j\omega})|^2 e^{j\omega(u-v)}\,\mathrm{d}\omega \tag{4.59}$$

be the u^{th} element of a vector $\boldsymbol{\xi}$ and the $(u,v)^{\text{th}}$ element of a matrix $\boldsymbol{\Xi}$ respectively. The expected value of the u^{th} element of $\mathbf{r}_{\bar{x}\bar{x}}$ from (4.52) then becomes

$$\mathscr{E}\{r_{\bar{x}\bar{x},u}\} = \bar{\kappa} r_u + \xi_u, \quad u = 1,2,\ldots,p, \tag{4.60}$$

where $r_{ss,u}$ is the u^{th} element of the vector $\mathbf{r}_{ss}$ in (4.11). Similarly, the expected value of the $(u,v)^{\text{th}}$ element of $\mathbf{R}_{ss}$ is

$$\mathscr{E}\{r_{\bar{x}\bar{x},uv}\} = \bar{\kappa} r_{ss,uv} + \Xi_{uv}, \quad u,v = 1,2,\ldots,p, \tag{4.61}$$

where $r_{ss,uv}$ is the $(u,v)^{\text{th}}$ element of the matrix $\mathbf{R}_{ss}$ defined in (4.11). The expected set of coefficients for the DSB output is therefore

$$\mathscr{E}\{\hat{\mathbf{b}}_{\text{DSB}}\} \cong (\bar{\kappa}\mathbf{R}_{ss} + \boldsymbol{\Xi})^{-1}(\bar{\kappa}\mathbf{r}_{ss} + \boldsymbol{\xi}). \tag{4.62}$$

Since $\boldsymbol{\Xi}$ is a Hermitian symmetric matrix, it can be factored as

$$\boldsymbol{\Xi} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^H, \tag{4.63}$$

where $\boldsymbol{\Gamma}$ is a matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. Using the matrix inversion lemma [15], we can write

$$\begin{aligned}(\bar{\kappa}\mathbf{R}_{ss} + \boldsymbol{\Xi})^{-1} = {} & \frac{1}{\bar{\kappa}}\mathbf{R}_{ss}^{-1} \\ & - \frac{1}{\bar{\kappa}^2}\mathbf{R}_{ss}^{-1}\boldsymbol{\Gamma}\left(\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Gamma}^H\frac{1}{\bar{\kappa}}\mathbf{R}_{ss}^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^H\mathbf{R}_{ss}^{-1}.\end{aligned} \tag{4.64}$$

Finally, substituting the result from (4.64) into (4.62) we obtain the result in (4.35).

References

1. Atal, B., Rabiner, L.: A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Trans. Acoust., Speech, Signal Process. **24**(3), 201–212 (1976)
2. Atal, B.S., Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. **50**(2), 637–655 (1971)

3. Benesty, J.: Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J. Acoust. Soc. Am. **107**(1), 384–391 (2000)
4. Bharitkar, S., Hilmes, P., Kyriakakis, C.: Robustness of spatial average equalization: A statistical reverberation model approach. J. Acoust. Soc. Am. **116**(6), 3491–3497 (2004)
5. Brandstein, M.S., Griebel, S.M.: Nonlinear, model-based microphone array speech enhancement. In: S.L. Gay, J. Benesty (eds.) Acoustic Signal Processing For Telecommunication, pp. 261–279. Kluwer Academic Publishers (2000)
6. Brandstein, M.S., Ward, D.B. (eds.): Microphone arrays: Signal processing techniques and applications, 1 edn. Springer (2001)
7. Chen, J., Benesty, J., Huang, Y.: Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. EURASIP J. on App. Signal Process. **2005**(1), 25–36 (2005)
8. Chen, J., Benesty, J., Huang, Y.: Time delay estimation in room acoustic environments: an overview. EURASIP J. on App. Signal Process. Special issue on advances in multimicrophone speech processing. **2006**, 1–19 (2006)
9. Deller, J.R., Hansen, J.H.L., Proakis, J.G.: Discrete-time processing of speech signals. Macmillan (1993)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., Series B **39**(1), 1–38 (1977)
11. Gaubitch, N.D.: Blind identification of acoustic systems and enhancement of reverberant speech. Ph.D. thesis, Imperial College London (2007)
12. Gaubitch, N.D., Naylor, P.A.: Spatiotemporal averaging method for enhancement of reverberant speech. In: Proc. Int. Conf. on Digital Signal Processing (DSP). Cardiff (2007)
13. Gaubitch, N.D., Ward, D.B., Naylor, P.A.: Statistical analysis of the autoregressive modeling of reverberant speech. J. Acoust. Soc. Am. **120**(6), 4031–4039 (2006)
14. Gillespie, B.W., Malvar, H.S., Florêncio, D.A.F.: Speech dereverberation via maximum-kurtosis subband adaptive filtering. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 6, pp. 3701–3704 (2001)
15. Golub, G.H., van Loan, C.F.: Matrix computations, 3 edn. John Hopkins Series in the Mathematical Sciences. John Hopkins University Press (1996)
16. Griebel, S.M.: A microphone array system for speech source localization, denoising and dereverberation. Ph.D. thesis, Harvard University, Cambridge, Massachusetts (2002)
17. Griebel, S.M., Brandstein, M.S.: Microphone array speech dereverberation using coarse channel estimation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 201–204 (2001)
18. Gustafsson, T., Rao, B.D., Trivedi, M.: Source localization in reverberant environments: Modeling and statistical analysis. IEEE Trans. Speech Audio Process. **11**(6), 791–803 (2003)
19. Harris, F.J.: On the use of windows for harmonic analysis with the discrete fourier transform. Proc. IEEE **66**(1), 51–83 (1978)
20. Knapp, G.H., Carter, G.C.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. **24**(4), 320–327 (1976)
21. Kuttruff, H.: Room acoustics, 4th edn. Taylor & Francis (2000)
22. Lim, J.S., Oppenheim, A.V.: All-pole modeling of degraded speech. IEEE Trans. Acoust., Speech, Signal Process. **26**(3), 197–210 (1978)
23. Lindsey, G., Breen, A., Nevard, S.: SPAR's archivable actual-word databases. Tech. rep., University College London (1987)
24. Makhoul, J.: Linear prediction: A tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
25. Makhoul, J.: Spectral linear prediction: Properties and applications. IEEE Trans. Acoust., Speech, Signal Process. **23**(3), 283 – 296 (1976)
26. Naylor, P.A., Kounoudes, A., Gudnason, J., Brookes, M.: Estimation of glottal closure instants in voiced speech using the dypsa algorithm. IEEE Trans. Audio, Speech, Lang. Process. **15**(1), 34–43 (2007)
27. Nelson, P.A., Elliott, S.J.: Active control of sound. Academic Press (1993)
28. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals. Prentice-Hall (1978)

29. Radlović, B.D., Williamson, R.C., Kennedy, R.A.: Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Acoust., Speech, Signal Process. **8**(3), 311–319 (2000)
30. Sambur, M.R., Jayant, N.S.: LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise. IEEE Trans. Acoust., Speech, Signal Process. **24**(6), 488–494 (1976)
31. Talantzis, F., Ward, D.B.: Robustness of multi-channel equalization in an acoustic reverberant environments:. J. Acoust. Soc. Am. **114**(2), 833–841 (2003)
32. Talantzis, F., Ward, D.B., Naylor, P.A.: Performance analysis of dynamic acoustic source separation in reverberant rooms. IEEE Trans. Audio, Speech, Lang. Process. **14**(4), 1378–1390 (2006)
33. Thomas, M.R.P., Gaubitch, N.D., Gudnason, J., Naylor, P.A.: A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2007)
34. Thomas, M.R.P., Gaubitch, N.D., Naylor, P.A.: Multichannel DYPSA for estimation of glottal closure instants in reverberant speech. In: Proc. European Signal Processing Conf. (EUSIPCO). Poznan, Poland (2007)
35. Veen, B.D.V., Buckley, K.M.: Beamforming: A versatile approach to spatial filtering. IEEE Signal Process. Mag. **5**(2), 4–24 (1988)
36. Ward, D.B.: On the performance of acoustic crosstalk cancellation in a reverberant environments:. J. Acoust. Soc. Am. **110**(2), 1195–1198 (2001)
37. Wu, M., Wang, D.: A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. **14**(3), 774–784 (2006)
38. Yegnanarayana, B., Naik, J.M., Childers, D.G.: Voice simulation: Factors affecting quality and naturalness. In: Proc. Conf. of the Association for Computational Linguists, pp. 530–533. Stanford, California, USA (1984)
39. Yegnanarayana, B., Prasanna, S.R.M., Rao, K.S.: Speech enhancement using excitation source information. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 541–544 (2002)
40. Yegnanarayana, B., Satyanarayana, P.: Enhancement of reverberant speech using LP residual signals. IEEE Trans. Acoust., Speech, Signal Process. **8**(3), 267–281 (2000)

Chapter 5
Multi-microphone Speech Dereverberation Using Eigen-decomposition

Sharon Gannot

Abstract A family of approaches for multi-microphone speech dereverberation in colored noise environments, using eigen-decomposition of the data correlation matrix, is explored in this chapter. It is shown that the Acoustic Impulse Responses (AIRs), relating the speech source and the microphones are embedded in the null subspace of the received signals. The null subspace is estimated using either the generalized singular value decomposition of the data matrix or the generalized eigenvalue decomposition of the respective correlation matrix.

In cases where the channel order is overestimated, further processing is required. A closed-form algorithm for extracting the AIR is derived. The proposed algorithm exploits the special structure of the null subspace matrix by using the total least squares criterion.

A study of the incorporation of the subspace method into a subband framework has potential to improve the performance of the proposed method, although many problems, especially the gain ambiguity problem, remain open.

The estimated AIRs can be used for dereverberation by applying conventional channel inversion methods.

An experimental study supports the potential of the proposed method, and provides insight into its limitations.

5.1 Introduction

In many speech communication applications, the recorded speech signal is affected by multipath reflections at the room walls and other objects on the propagation path from the source to the microphones. This phenomenon is usually referred to as *reverberation*. The received reverberant speech signal is often perceived by the listeners as suffering from reduced quality, and in severe cases, as unintelligible.

Bar-Ilan University, Israel

Moreover, subsequent processing of the speech signal, such as speech coding or Automatic Speech Recognition (ASR) might be rendered useless in the presence of even modest levels of reverberation.

This chapter is dedicated to multi-microphone algorithms, which are based on concepts adopted from the *blind deconvolution* (also known as *blind equalization* or *blind identification*) family of methods [19]. Blind deconvolution has been successfully applied in communication applications. It usually consists of two stages. First, the impulse response of the system, relating the source and the receivers, is blindly estimated. Then, at the subsequent inversion stage, these estimates are used for the equalization. It is assumed that perfect estimation of the impulse responses suffices for perfectly equalizing the received signal. A survey of multichannel blind identification methods is given in [40].

The core of this chapter is a discussion of methods for channel identification based on the eigen-structure of the spatiotemporal correlation matrix of the received signals.

Two of the early contributions to this field may be attributed to Moulines *et al.* [34] and to Xu *et al.* [43]. In [34] the correlation matrix of a data block is first estimated. Then, using eigenvalue decomposition, the null subspace of the correlation matrix is determined. Finally, based on the orthogonality of the channels, assumed to be Finite Impulse Response (FIR) filters, and the null subspace, the identification is performed by applying a (rather heuristic) minimization of a quadratic term, related to the null subspace. It is shown that the filters are identifiable, provided that the dimension of the correlation matrix is large enough and that the channels do not share any common zeros. The assumption that the filter order is known in advance, and the tendency of the correlation matrix towards high dimensionality, restricts the application of the proposed method to speech dereverberation problems.

Xu *et al.* [43] officially state and prove the necessary and sufficient conditions for the identifiably of the Acoustic Impulse Responses (AIRs) from the received multi-microphone data. They show that the input signal must be sufficiently 'rich' to excite all the systems' modes (i.e., the respective correlation matrix is of sufficient rank) and that the AIR polynomials must not share common zeros. They show that the AIR can be estimated by calculating the correlation matrix of the receivers' signals, provided that the length of the filters is overestimated. The correct order of the filters can be calculated by identifying the number of zero eigenvalues.

Gürelli and Nikias [15] presented an Eigenvector-Based Algorithm for Multichannel Blind Deconvolution (EVAM). They show that deconvolution may be obtained by filtering the received signals with an FIR filter. The order of these filters is assumed to be an overestimation of the actual AIRs' order. They further show that the null subspace eigenvectors of the respective correlation matrix are filtered versions of the actual AIRs. The extraneous zeros constitute filters, the order of which is equal to the amount of overestimation. These zeros are shown to be common to all null eigenvectors. This observation forms the basis of the proposed procedure for eliminating extraneous zeros. The authors propose a recursive method for successively eliminating the extraneous zeros, referred to as the fractal method. In this method, the estimated AIRs' order is gradually decreased until only one null

subspace eigenvector remains. The correct-order AIRs can then be extracted from the remaining eigenvector. The algorithm has been successfully applied to noiseless speech signals, filtered by arbitrary filters, 600 taps long.

Affes and Grenier [1] use the eigen-structure of the correlation matrix (calculated per frequency band) to show that the desired system, relating the source signal and the receiving microphones, can be extracted from the signal subspace as well. The signal subspace is estimated using the recursive tracking algorithm proposed by Yang [45]. The algorithm proposed in [1] is implemented in the frequency domain. Therefore, rather than estimating the AIR, its Fourier transform, denoted Acoustic Transfer Function (ATF), is calculated. The estimated ATFs are then embedded into a Generalized Sidelobe Canceller (GSC) beamformer, which is used for enhancing a speech signal contaminated by white noise. The use of the ATF estimator can reduce the reverberation, provided that some *a priori* information is given. In [1], it is assumed that the average norm of all ATFs can be measured in advance. It is shown that this quantity is quite robust to small speaker movements. Doclo and Moonen [5] extended the concept of signal subspace estimation to spatially non-white noise environments, using Generalized Eigenvalue Decomposition (GEVD). The resulting algorithm is able to reduce jointly the noise and mitigate the reverberation, when the average ATF norm is given. It should be emphasized, however, that when this information is not available, neither method can eliminate the reverberation entirely. Since the small movement assumption, which is an important assumption intrinsic to both methods, cannot be be guaranteed in many important applications, we will avoid using it in this chapter.

Gannot and Moonen (see [11] and its preliminary version [9]) use the received signals' correlation matrix (similar to the matrix used by Gürelli and Nikias [15]) for extracting the AIRs. Their method differs in the treatment of the overestimation of the system order. The null subspace is first obtained using the GEVD of the noisy correlation matrix and the respective noise-only correlation matrix. The special Sylvester structure of the filtering matrix is then imposed on the null subspace for deriving a Total Least Squares (TLS) estimate of the desired AIRs. Other, less robust but computationally efficient methods are derived, based on the QR decomposition of the same null subspace. The high sensitivity of the GEVD procedure to noise, especially when the involved AIRs are very long, together with the wide dynamic range of the speech signal, limit the applicability of the full-band method in realistic scenarios. Therefore, Gannot and Moonen further incorporate the TLS subspace method into a subband structure [10, 11]. The subband method proves to be efficient, although a new problem, namely the gain ambiguity problem, arises. Eneman and Moonen [7] propose a method for mitigating the gain ambiguity problem inherent to subband methods.

Several classical and more recently developed multi-microphone dereverberation algorithms are compared by Eneman and Moonen [8] in several test scenarios. The dereverberation ability of the algorithms as pre-processors for ASR systems, as well as their ability to flatten the frequency response of the ATFs, is validated. Non-encouraging results are obtained by both the full-band and subband versions of the null subspace methods, despite their higher computational load, compared to

the simpler dereverberation methods, e.g., delay-and-sum beamforming, matched filtering [26], and cepstral processing [31].

Lin *et al.* [29, 30] address the problem of the common zeros. The identifiability conditions, stated in [34, 43], are violated when the AIRs share common zeros. Hughes and Nakeghbali [25] show that polynomials with random coefficients tend to have uniformly distributed zeros close to the unit circle as the polynomial order tends to infinity. Therefore, ATFs will be more likely to have common or near common zeros for larger orders. Since, according to Polack's theory [36], AIRs can be modeled as long filters with uncorrelated taps having a decaying energy profile, it is reasonable to assume that the common zeros phenomenon is likely to occur. Note, however, that common zeros are less likely to occur when more microphone signals are used, as the zeros should be common for all channels. It is shown in [29, 30] that these common zeros have the effect of filtering the input signal. Hence, utilizing common subspace methods, only a filtered version of the input signal can be reconstructed rather than the actual input signal. However, assuming slowly time varying room responses and non-stationary input signal (recall that the input is a speech signal) the common zeros can be identified and eliminated from the entire response. Note that the overestimation of the AIRs order is manifested as common zeros in the identification procedure [11]. Hence, the AIRs' common zeros, and the common zeros resulting in from the overestimation, are indistinguishable. It is therefore assumed in [29] that the exact AIR order is known. The method for identification of common zeros has been tested using 512 taps long filters in a noiseless environment.

Javidi *et al.* [27] investigate the influence of noise on the criterion that forms the core of the subspace methods, namely that the AIRs are embedded in the null (noise) subspace eigenvectors. They show that, in presence of noise, the eigenvector producing the best channel estimate, does not necessarily correspond to the eigenvalue of the smallest power.

A class of frequency-domain adaptive approaches was proposed by Huang and Benesty [22]. The minimizer of the proposed criterion is the smallest eigenvalue of the data matrix. The minimization is carried out using Least Mean Squares (LMS)-type adaptation in the frequency-domain. The resulting AIRs are normalized to have unit norm. The robustness of the proposed Normalized Multichannel Frequency Domain Least Mean Square (NMCFLMS) algorithm is explored by Hasan *et al.* [17]. They show that by incorporating the correct delay and gain of the direct-path into the NMCFLMS algorithm, an improved convergence behavior and increased robustness of the algorithm may be obtained. Ahmad *et al.* [3] show that this performance improvement is maintained when the correct direct-path delay is replaced by its estimate obtained using the PHAT variant of the GCC algorithm [28]. The robustness of the NMCFLMS algorithm to additive noise is further explored by Hasan and Naylor [18]. They show that the misconvergence of the NMCFLMS algorithm can be attributed to the non-zero gradient of the cost function under noisy conditions. The convergence characteristic can be ameliorated by introducing frequency domain energy constraints which can counterbalance the low-pass filtering effect of the unconstrained NMCFLMS algorithm. Ahmad *et al.* [2] propose the use of alternating

sets of filters to alleviate the misconvergence phenomenon. One set is adapted using the NMCFLMS algorithm, while the second set keeps track of the best available estimate obtained just before the algorithm misconverged. A joint scheme for blind source separation and dereverberation is proposed by Huang *et al.* [23]. They show that the Multi-Input Multi-Output (MIMO) system can be recast as a set of Single-Input Multi-Output (SIMO) systems for each input signal. Assuming that from time to time each speaker occupies an interval exclusively, and that the AIRs are slowly varying in time, the AIRs can be identified by the application of the NMCFLMS algorithm.

Hikichi *et al.* [20, 21] propose another method for compensating for the excess zeros resulting in from the application of the subspace method due to channel overestimation. In their method, the AIRs are first estimated using subspace methods. Then, the exact inverse filter set is calculated using Multiple-input/output INverse Theorem (MINT) [32]. The signal processed by the inverse filter set is still reverberated due to the effect of the common polynomial. It is shown that this reverberation effect is proportional to the inverse of the minimum phase counterpart of the filter constructed by the common zeros. A method for extracting the compensating polynomial coefficients from the AIR estimates (having overestimated order) is derived, employing the multi-channel linear prediction technique.

The structure of this chapter is as follows. The general dereverberation problem is stated in Sect. 5.2. In Sect. 5.3 we review the core principle of the null subspace algorithm. This review serves as an intuitive explanation of the concepts proved in, e.g., [43]. The basic full-band algorithm, based on the Sylvester structure of the filtering matrix, is explored in Sect. 5.4. Several important extensions, namely the two-microphone noisy case, the multi-microphone case, and the case in which only part of the null eigenvectors is available, are presented in Sect. 5.5. Sect. 5.6 is dedicated to the incorporation of the null subspace algorithm into a subband structure. The experimental study, presented in Sect. 5.8 verifies the applicability of the proposed methods to the problem at hand and emphasizes their performance limitations. These limitations, related to the noise robustness, the computational burden, the common zero problem, and the gain ambiguity problem, encountered in the subband variants, are discussed in Sect. 5.9, while possible cures and new research directions are proposed.

5.2 Problem Formulation

A speech signal is received by M microphones in a noisy and reverberant environment. The received speech signal is subject to propagation through a set of AIRs before being picked up by the microphones. The M received signals are given by:

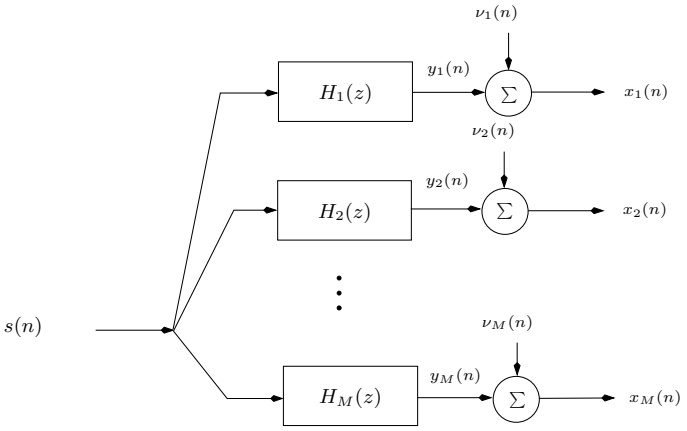


Fig. 5.1 The general reverberation model. Each of the microphone signals $x_m(n)$ is comprised of the speech signal $s(n)$ convolved with the ATFs $H_m(z)$ and an additive noise signal $\nu_m(n)$

$$\begin{aligned} x_m(n) &= y_m(n) + \nu_m(n) \\ &= h_m(n) * s(n) + \nu_m(n) \\ &= \sum_{k=0}^{L} h_m(k) s(n-k) + \nu_m(n), \end{aligned} \tag{5.1}$$

where $m = 1,\ldots,M$ is the microphone index, $n = 0,1,\ldots,N$ is the time index, the number of samples observed is $N+1$, $x_m(n)$ is the signal received at the m^{th} microphone, $y_m(n)$ is the corresponding desired signal component, $\nu_m(n)$ is the noise signal picked up by the m^{th} microphone and $s(n)$ is the desired speech signal. We further assume that the AIRs, relating the speech source and each of the M microphones, can be modeled as FIR filters of order L, time-invariant during the $N+1$ observed samples, and having the following coefficients:

$$\mathbf{h}_m^T = \left[h_m(0)\ h_m(1)\ \ldots\ h_m(L) \right].$$

Define also the z-transform of each of the M filters as:

$$H_m(z) = \sum_{k=0}^{L} h_m(k) z^{-k}, \quad m = 1,2,\ldots,M.$$

All the involved signals and AIRs are depicted in Fig. 5.1. The goal of the dereverberation task is to reconstruct the speech signal $s(n)$ from the noisy observations $x_m(n)$, $m = 1,2,\ldots,M$. In this contribution we will try to achieve this goal by first estimating the AIRs, $\mathbf{h}_m$, and then applying a signal reconstruction scheme based on these AIR estimates. Schematically, an AIR estimation block, depicted in Fig. 5.2 will be the outcome of the first stage of the dereverberation algorithm.

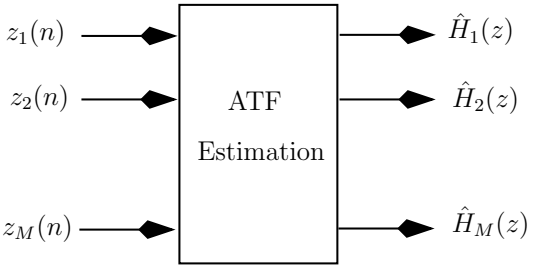


Fig. 5.2 Schematic depiction of an ATF estimation procedure. The inputs of the procedure are the microphone signals $x_m(n)$, $m = 1,2,\ldots,M$, and its outcome are the estimated ATFs $\hat{H}_m(z)$, $m = 1,2,\ldots,M$

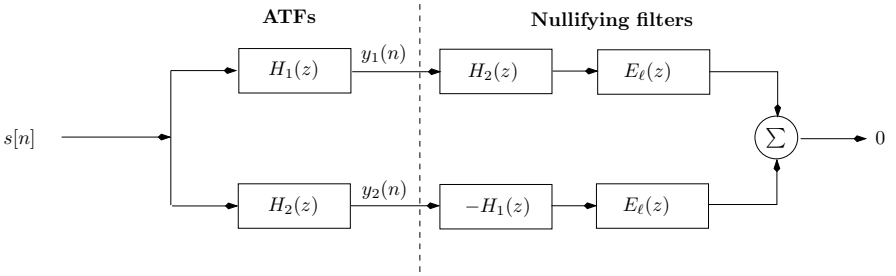


Fig. 5.3 Nullifying filters for the two-microphone noiseless case. The desired ATFs are embedded in the nullifying filters (in reverse order)

5.3 Preliminaries

In this section we lay the foundations of the algorithm by showing that the desired AIRs are embedded in the null subspace of a signal data matrix. This proof is a repetition of previously established results (e.g., [15, 34, 43]), but presented in a more intuitive way. We demonstrate the concept for the two-microphone noiseless case.

The two-microphone noiseless case is depicted in Fig. 5.3. The noiseless signals, $y_m(n)$, are taken from the left-hand side of the figure as:

$$\begin{aligned} y_1(n) &= h_1(n) * s(n), \\ y_2(n) &= h_2(n) * s(n), \end{aligned} \tag{5.2}$$

where $*$ denotes the convolution. Clearly, as depicted on the right-hand side of Fig. 5.3, the following equality holds:

$$(y_2(n) * h_1(n) - y_1(n) * h_2(n)) * e_\ell(n) = 0, \tag{5.3}$$

where $e_\ell(n), \ell = 0,1,2,\ldots$ are arbitrary and unknown filters, the number and order of which will be discussed in the sequel. It is evident that the filtered version of the desired AIRs, subject to the constraint that the arbitrary filters, $e_\ell(n)$ are common to all the microphone systems, results in a zero output.

Define the set of filtered AIRs $\tilde{h}_{m,\ell}(n) = h_m(n) * e_\ell(n),\ m = 1,2,\ldots,M$. Let $\tilde{\mathbf{h}}_{m,\ell},\ m = 1,2,\ldots,M$ be vectors comprised of the coefficients of the respective filters given by:

$$\tilde{\mathbf{h}}_{m,\ell}^T = \left[\, \tilde{h}_{m,\ell}(0)\ \tilde{h}_{m,\ell}(1)\ \ldots\ \tilde{h}_{m,\ell}(\hat{L}) \,\right]. \tag{5.4}$$

Concatenating these vectors yields:

$$\tilde{\mathbf{h}}_\ell^T = \left[\, \tilde{\mathbf{h}}_{1,\ell}^T\ \tilde{\mathbf{h}}_{2,\ell}^T\ \ldots\ \tilde{\mathbf{h}}_{M,\ell}^T \,\right], \tag{5.5}$$

where we assume at present that $M = 2$. Define also the $(\hat{L}+1)\times(N+\hat{L}+1)$ single channel data matrix

$$\mathbf{Y}_m = \begin{bmatrix} y_m(0) & \cdots & & y_m(\hat{L}-1) & y_m(\hat{L}) & \cdots & y_m(N) & 0 & \cdots & 0 \\ 0 & y_m(0) & \cdots & \vdots & \vdots & & \cdots & y_m(N) & 0 & 0 \\ \vdots & 0 & \ddots & \vdots & & & & & \ddots & \vdots \\ 0 & & & y_m(0) & & \ddots & & 0 & \ddots & \\ 0 & & \cdots & 0 & y_m(0) & \cdots & y_m(\hat{L}) & \cdots & & y_m(N) \end{bmatrix}. \tag{5.6}$$

Note that, as the correct AIR order L is unknown, an overestimated value, $\hat{L}$ is used instead, i.e., the inequality $\hat{L} \geq L$ is assumed to hold. An estimate of the correct order would be a by-product of the proposed algorithm. Define also the two-channel data matrix,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_2 \\ -\mathbf{Y}_1 \end{bmatrix}. \tag{5.7}$$

Using definitions (5.4)–(5.7) and Fig. 5.3 we have:

$$\mathbf{Y}^T \tilde{\mathbf{h}}_\ell = 0,\ \forall \ell. \tag{5.8}$$

It is therefore easily verified that

$$\tilde{\mathbf{h}}_\ell \mathbf{Y}\mathbf{Y}^T \tilde{\mathbf{h}}_\ell = 0,\ \forall \ell \tag{5.9}$$

also holds. We can now identify the $2(\hat{L}+1)\times 2(\hat{L}+1)$ sample correlation matrix of the data as $\hat{\mathbf{R}}_y = \frac{\mathbf{Y}\mathbf{Y}^T}{N+1}$. Hence, the vectors $\tilde{\mathbf{h}}_\ell$, the number of which is yet to be determined, are the eigenvectors belonging to the null subspace of $\hat{\mathbf{R}}_y$. Now, following [15], the null subspace of the correlation matrix can be calculated by virtue of the eigenvalue decomposition. Let $\lambda_\ell,\ \ell = 0,1,\ldots,2\hat{L}+1$ be the eigenvalues of the correlation matrix $\hat{\mathbf{R}}_y$. Then, by sorting them in ascending order, we have:

$$\begin{array}{l} \lambda_\ell = 0, \ell = 0,1,\ldots,\hat{L}-L \\ \lambda_\ell > 0, \text{otherwise} \end{array}. \tag{5.10}$$

Hence, as proven by Gürelli and Nikias [15], the rank of the null subspace of the correlation matrix is $\hat{L}-L+1$. This rank is useful for determining the correct AIR order, L. We note that Singular Value Decomposition (SVD) of the data matrix $\mathbf{Y}$ might be used instead of EVD of the correlation matrix $\hat{\mathbf{R}}_y$ for determining the null subspace. SVD is generally regarded as a more robust method.

Denote the null subspace vectors (eigenvectors corresponding to zero eigenvalues or singular values) by $\mathbf{v}_\ell$ for $\ell = 0,1,2,\ldots,\hat{L}-L$. Then, splitting each null subspace vector into two parts of equal length $\hat{L}+1$ we obtain

$$\mathbf{V} = \left[\mathbf{v}_0\, \mathbf{v}_1 \cdots \mathbf{v}_{\hat{L}-L}\right] = \begin{bmatrix} \tilde{\mathbf{h}}_{1,0}\, \tilde{\mathbf{h}}_{1,1} \cdots \tilde{\mathbf{h}}_{1,\hat{L}-L} \\ \tilde{\mathbf{h}}_{2,0}\, \tilde{\mathbf{h}}_{2,1} \cdots \tilde{\mathbf{h}}_{2,\hat{L}-L} \end{bmatrix}. \tag{5.11}$$

Each part of the null subspace vector can be readily identified as the filtered AIRs $\tilde{\mathbf{h}}_{m,\ell}$, given by (5.4), with the corresponding z-transform:

$$\tilde{H}_{m,\ell}(z) = \sum_{k=0}^{\hat{L}} \tilde{h}_{m,\ell}(k) z^{-k},\ l = 0,1,\ldots,\hat{L}-L,\ m = 1,2. \tag{5.12}$$

From the above discussion, these filters may be presented as the following product:

$$\tilde{H}_{m,\ell}(z) = H_m(z) E_\ell(z),\ \ell = 0,1,\ldots,\hat{L}-L,\ m = 1,2. \tag{5.13}$$

Hence, the zeros of the filters $\tilde{H}_{m,\ell}(z)$ extracted from the null subspace of the data, include the roots of the desired filters as well as some extraneous zeros. This observation was proven by Gürelli and Nikias [15] as the basis of their EVAM algorithm. The core of all eigen-decomposition based methods is formally stated in Lemma 5.1 (for the general M-channel case).

Lemma 5.1. *Let $\tilde{\mathbf{h}}_{m,\ell}$ be the partitions of the null subspace eigenvectors into M vectors of length $\hat{L}+1$, with $\tilde{H}_{m,\ell}(z)$ their equivalent filters. Then, all the filters $\tilde{H}_{m,\ell}(z)$ for $\ell = 0,\ldots,\hat{L}-L$ have L common roots, which constitute the desired ATFs $H_m(z)$, and $\hat{L}-L$ different extraneous roots, which constitute $E_\ell(z)$. These extraneous roots are common for all partitions of the same vector, i.e., $\tilde{H}_{m,\ell}(z)$ for $m = 1,\ldots,M$.* □

Under several regularity conditions (stated, for example, by Moulines *et al.* [34]), the filters $H_m(z)$ can be found. Of special interest is the observation that common roots of the filters $H_m(z)$ cannot be extracted by the method, because they might be confused with the extraneous roots that constitute $E_\ell(z)$. Although this is a drawback of the method, we will benefit from this property while constructing the subband structure in Sect. 5.6.

In matrix form, (5.13) may be written in the following manner. Define the $(\hat{L}+1)\times(\hat{L}-L+1)$ Sylvester filtering matrix (recall we assume $\hat{L} \geq L$),

$$\mathbf{H}_m = \underbrace{\begin{bmatrix} h_m(0) & 0 & 0 & \cdots & 0 \\ h_m(1) & h_m(0) & 0 & \cdots & 0 \\ \vdots & h_m(1) & \ddots & & \vdots \\ h_m(L) & \vdots & \ddots & \ddots & 0 \\ 0 & h_m(L) & & \ddots & h_m(0) \\ \vdots & 0 & & & h_m(1) \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & h_m(L) \end{bmatrix}}_{\hat{L}-L+1}. \tag{5.14}$$

Then,

$$\tilde{\mathbf{h}}_{m,\ell} = \mathbf{H}_m \mathbf{e}_\ell, \tag{5.15}$$

where, $\mathbf{e}_\ell^T = \left[e_\ell(0)\, e_\ell(1) \ldots e_\ell(\hat{L}-L) \right]$, $\ell = 0,1,\ldots,\hat{L}-L$ are vectors of the coefficients of the arbitrary unknown filters $E_\ell(z)$. Thus, the number of different filters (as shown in (5.13)) is $\hat{L}-L+1$ and their order is $\hat{L}-L$. Using Fig. 5.3 and (5.3) we obtain in matrix form:

$$\mathbf{V} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} \mathbf{E} \stackrel{\triangle}{=} \mathbf{H}\mathbf{E}. \tag{5.16}$$

$\mathbf{E}$ is an unknown $(\hat{L}-L+1) \times (\hat{L}-L+1)$ matrix, formed by concatenating the arbitrary unknown filters,

$$\mathbf{E} = \left[\mathbf{e}_0\, \mathbf{e}_1 \cdots \mathbf{e}_{\hat{L}-L} \right].$$

We note that, in the special case where the AIRs' order is known, i.e. $\hat{L} = L$, there is only one vector in the null subspace and its partitions $\tilde{\mathbf{h}}_{m,0}$, $m = 1,\ldots,M$ constitute the desired filters $\mathbf{h}_m$ up to a (common) scaling factor. In the case where $\hat{L} > L$, the actual ATFs $H_m(z)$ are embedded in $\tilde{H}_{m,\ell}(z)$, $\ell = 0,1,\ldots,\hat{L}-L$. The case $\hat{L} < L$ cannot be treated properly by the proposed method.

The special structure depicted in (5.14) and (5.16) forms the basis of our proposed algorithm.

5.4 AIR Estimation – Algorithm Derivation

In this section an AIR estimation algorithm is derived. The special structure of the null subspace, discussed in Sect. 5.3 is exploited to derive the estimation method. Initially, we concentrate on the *two-microphone noiseless* case. In Sect. 5.5 we will elaborate on several extensions of the algorithm.

Based on the special structure of (5.16) and, in particular, on the Sylvester structure of $\mathbf{H}_1$ and $\mathbf{H}_2$, found in (5.14), we now derive an algorithm for estimating the AIRs $\mathbf{h}_m$.

Note that $\mathbf{E}$ in (5.16) is an arbitrary square matrix, which implies that its inverse usually exists. Denote this inverse by $\mathbf{E}^{-1} \triangleq \text{inv}(\mathbf{E})$. Then

$$\mathbf{V}\mathbf{E}^{-1} = \mathbf{H}. \tag{5.17}$$

Denote the columns of $\mathbf{E}^{-1}$ by $\mathbf{e}_\ell^{\text{i}}$, $\ell = 0, 1, \ldots, \hat{L} - L$. Equation (5.17) can then be rewritten as,

$$\tilde{\mathbf{V}}\theta = \mathbf{0}, \tag{5.18}$$

where $\mathbf{0}$ is a vector of zeros, $\mathbf{0}^T = \begin{bmatrix} 0\,0\cdots 0 \end{bmatrix}$, $\tilde{\mathbf{V}}$ is defined as:

$$\tilde{\mathbf{V}} = \begin{bmatrix} \mathbf{V}\,\mathbf{O}\cdots\cdots\cdots\mathbf{O} & -\mathbf{S}^{(0)} \\ \mathbf{O}\,\mathbf{V}\,\mathbf{O}\cdots\cdots\mathbf{O} & -\mathbf{S}^{(1)} \\ \vdots\ \mathbf{O}\ \ddots \quad\quad \vdots & \vdots \\ \vdots\ \vdots \quad \ddots\ \ddots\ \vdots & \vdots \\ \vdots\ \vdots \quad \ddots\ \ddots\ \mathbf{O} & \vdots \\ \mathbf{O}\,\mathbf{O}\cdots\cdots\mathbf{O}\,\mathbf{V} & -\mathbf{S}^{(\hat{L}-L)} \end{bmatrix}, \tag{5.19}$$

and the vector of unknowns is defined as:

$$\theta^T = \left[(\mathbf{e}_0^{\text{i}})^T\ (\mathbf{e}_1^{\text{i}})^T \cdots (\mathbf{e}_{\hat{L}-L}^{\text{i}})^T\ \mathbf{h}_1^T\ \mathbf{h}_2^T \right]. \tag{5.20}$$

We used the following expressions: $\mathbf{O}$ is a $2(\hat{L}+1) \times (\hat{L}-L+1)$ all-zero matrix and $\mathbf{S}^{(\ell)}$, $\ell = 0, 1, \ldots, \hat{L} - L$, is a fixed shifting matrix given by:

$$\mathbf{S}^{(\ell)} = \left[\begin{array}{c|c} \begin{matrix} \mathbf{O}_{\ell\times(L+1)} \\ \mathbf{I}_{(L+1)\times(L+1)} \\ \mathbf{O}_{(\hat{L}-L-l)\times(L+1)} \end{matrix} & \mathbf{O}_{(\hat{L}+1)\times(L+1)} \\ \hline \mathbf{O}_{(\hat{L}+1)\times(L+1)} & \begin{matrix} \mathbf{O}_{\ell\times(L+1)} \\ \mathbf{I}_{(L+1)\times(L+1)} \\ \mathbf{O}_{(\hat{L}-L-l)\times(L+1)} \end{matrix} \end{array}\right].$$

$\mathbf{I}_{(L+1)\times(L+1)}$ is the $(L+1)\times(L+1)$ identity matrix and $\mathbf{O}_{k\times(L+1)}$ is a $k\times(L+1)$ all-zero matrix (k as specifically determined). A non-trivial (and exact) solution for the homogenous set of equations (5.18) may be obtained by finding the eigenvector of the matrix $\tilde{\mathbf{V}}$ corresponding to its zero eigenvalue. The AIR coefficients are given by the last $2(L+1)$ terms of this eigenvector. The first part of the eigenvector is comprised of the nuisance parameters $\mathbf{e}_\ell^{\text{i}}$, $\ell = 0, 1, \ldots, \hat{L} - L$. In the presence of noise, this somewhat non-straightforward procedure will prove to be useful.

5.5 Extensions of the Basic Algorithm

In this section we extend the algorithm proposed in Sect. 5.4 in several important directions. First, the *two-microphones contaminated by noise* case is treated. Then, the basic two-channel algorithm is extended to deal with the general *multi-microphone coloured noise* case. Finally, we treat the case when only part of the null subspace vectors can be determined. This generalization becomes relevant whenever the input signal correlation matrix becomes ill-conditioned. As the input speech signal correlation matrix tends to exhibit low level eigenvalues, the likelihood of encountering this situation is relatively high.

5.5.1 Two-microphone Noisy Case

Recall that $\mathbf{V}$ is a matrix containing the eigenvectors corresponding to zero eigenvalues of the noiseless data matrix. In the presence of additive noise, the noisy observations $x_m(n)$, given in (5.1), can be stacked into a data matrix fulfilling

$$\mathbf{X} = \mathbf{Y} + \boldsymbol{\Upsilon},$$

where $\mathbf{X}$ and $\boldsymbol{\Upsilon}$ are noisy signal and noise-only data matrices, similar to (5.7) with $y_m(n)$ replaced by $x_m(n)$ or $v_m(n)$, respectively.

Now, for long observation intervals, the following approximation holds:

$$\hat{\mathbf{R}}_x \approx \hat{\mathbf{R}}_y + \hat{\mathbf{R}}_v,$$

where $\hat{\mathbf{R}}_x = \frac{\mathbf{X}\mathbf{X}^T}{N+1}$ and $\hat{\mathbf{R}}_v = \frac{\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T}{N+1}$ are the noisy signal and noise-only signal correlation matrices, respectively. Now (5.18) will no longer be accurate. For dealing with this problem, several modifications to (5.18) are required. First, the null subspace matrix $\mathbf{V}$ should be determined in a manner slightly different from that proposed in (5.10). The white noise and colored noises cases are treated separately in the sequel. Second, the matrix $\widetilde{\mathbf{V}}$ defined in (5.19) will, in general, no longer have a zero-valued eigenvalue. A reasonable approximation for the solution, although not exact, would be to cast (5.18) into the following problem:

$$\tilde{\mathbf{V}}\theta = \varepsilon, \tag{5.21}$$

where ε is an error term, which should be minimized. To obtain this minimization, the eigenvector corresponding to the smallest eigenvalue of $\tilde{\mathbf{V}}$ is chosen, and the desired AIRs are obtained from the last part of the vector (as in the noiseless case). Note that this corresponds exactly to the TLS approach for estimating the parameter vector θ. As the matrix $\tilde{\mathbf{V}}$ is highly structured, the more efficient Structured Total Least Squares (STLS) method [24] may be used. The application of the STLS method is beyond the scope of our discussion.

5.5.1.1 White Noise Case

In the case of spatiotemporally white noise, i.e. when $\hat{\mathbf{R}}_v \approx \sigma_v^2 \mathbf{I}$ (where $\mathbf{I}$ stands for the identity matrix) the first $\hat{L} - L + 1$ eigenvalues in (5.10) will be σ_v^2 rather than 0. However, the corresponding eigenvectors remain the same, and, hence, the rest of the algorithm remains unchanged.

5.5.1.2 Colored Noise Case

The case of non-white noise was addressed in [15, 34]. In contrast to the pre-whitening of the noise correlation matrix, presented in [34], and the noise balancing method presented in [15], we treat the problem more rigorously, with the application of either the GEVD or the Generalized Singular Value Decomposition (GSVD) techniques. These alternative methods are computationally more efficient. GEVD will be applied to the noisy signal correlation matrix, $\hat{\mathbf{R}}_x$ and the noise correlation matrix $\hat{\mathbf{R}}_v$. The latter may be estimated from speech-free data segments. The null subspace matrix $\mathbf{V}$ is now constructed by choosing the generalized eigenvectors corresponding to the generalized eigenvalues of value 1. Alternatively, the GSVD of the corresponding data matrices, $\mathbf{X}$ and $\boldsymbol{\Upsilon}$, can be used. After determining the null subspace matrix, the subsequent steps of the algorithm remain unchanged.

5.5.2 Multi-microphone Case ($M > 2$)

In the multi-microphone case, a reasonable extension would be based on channel pairing (see also [15]). Each of the $\frac{M\times(M-1)}{2}$ pairs of the noiseless signals fulfills (5.22):

$$\begin{aligned} &(y_i(n) * h_j(n) - y_j(n) * h_i(n)) * e_\ell(n) = 0, \\ &i, j = 1, 2, \ldots, M,\ \ell = 0, 1, \ldots, \hat{L} - L. \end{aligned} \tag{5.22}$$

Thus, the new data matrix is constructed as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_2 & \mathbf{X}_3 & \cdots & \mathbf{X}_M & \mathbf{O} & \cdots & \mathbf{O} & \cdots & \mathbf{O} \\ -\mathbf{X}_1 & \mathbf{O} & \cdots & & \mathbf{X}_3 & \cdots & \mathbf{X}_M & & \mathbf{O} \\ \mathbf{O} & -\mathbf{X}_1 & & & -\mathbf{X}_2 & & \mathbf{O} & & \vdots \\ \vdots & \mathbf{O} & \ddots & & & & \vdots & & \mathbf{O} \\ & \vdots & \ddots & & & & \mathbf{O} & & \mathbf{X}_M \\ \mathbf{O} & \mathbf{O} & \cdots & -\mathbf{X}_1 & \cdots & & -\mathbf{X}_2 & \cdots & -\mathbf{X}_{M-1} \end{bmatrix}, \tag{5.23}$$

where $\mathbf{O}$ is an $(\hat{L}+1)\times(N+\hat{L}+1)$ all-zero matrix. This data matrix, as well as the corresponding noise matrix can be used by either the GEVD or the GSVD methods to construct the null subspace. Denoting this null subspace by $\underline{\mathbf{V}}$, we can construct a new TLS equation,

$$\underline{\widetilde{\mathbf{V}}}\,\underline{\theta} = \varepsilon,$$

where $\underline{\widetilde{\mathbf{V}}}$ is constructed following the same procedure by which $\widetilde{\mathbf{V}}$ was constructed from $\mathbf{V}$ in (5.19). The unknown parameter vector $\underline{\theta}$ is now given by:

$$\underline{\theta}^T = \left[\begin{array}{ccccccc} (\mathbf{e}_0^{\mathrm{i}})^T & (\mathbf{e}_1^{\mathrm{i}})^T & \cdots & (\mathbf{e}_{\hat{L}-L}^{\mathrm{i}})^T & \mathbf{h}_1^T & \mathbf{h}_2^T & \ldots \; \mathbf{h}_M^T \end{array} \right].$$

Note that the last $M\times(L+1)$ terms of $\underline{\theta}$, $\mathbf{h}_m$, $m=1,2,\ldots,M$, give the desired filter coefficients.

5.5.3 Partial Knowledge of the Null Subspace

In the noisy case, especially when the dynamic range of the input signal $s(n)$ is high (which is the case for speech signals), determination of the null subspace might be a troublesome task. As there are no zero eigenvalues, and as some of the eigenvalues are small due to the input signal, the borderline between the signal eigenvalues and the noise eigenvalues becomes vague. As the number of actual null subspace vectors is not known in advance, using only a subgroup of the eigenvectors, which are associated with the smallest eigenvalues, might increase the robustness of the method. Based on Lemma 5.1, it is obvious that in the noiseless case, even two null subspace vectors suffice to estimate the AIRs, merely by extracting their common zeros. Denote by $\bar{L}<\hat{L}-L$ the number of eigenvectors used. The dimensions of the matrix $\mathbf{E}$ in (5.16) becomes $(\hat{L}-L+1)\times\bar{L}$, resulting in non-invertible $\mathbf{E}$. To overcome this problem we suggest concatenating several shifted versions of (5.16):

$$\bar{\mathbf{V}} = \begin{bmatrix} \mathbf{V} & \mathbf{0} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} & & \mathbf{0} \\ \vdots & & \ddots & & \mathbf{0} \\ & & & \ddots & \vdots \\ \mathbf{0} & & & & \mathbf{V} \end{bmatrix} = \bar{\mathbf{H}} \underbrace{\begin{bmatrix} \mathbf{E} & \mathbf{0} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{E} & \mathbf{0} & & \mathbf{0} \\ \vdots & & \ddots & & \mathbf{0} \\ & & & \ddots & \vdots \\ \mathbf{0} & & & & \mathbf{E} \end{bmatrix}}_{L>\hat{L}-L+\hat{l}}$$

$$\triangleq \bar{\mathbf{H}}\bar{\mathbf{E}} \tag{5.24}$$

The dimensions of $\bar{\mathbf{E}}$ are $L\times(\hat{L}-L+\hat{l})$, where $\hat{l}$ is the number of blocks concatenated. Each block adds 1 to the row dimension and $\bar{L}$ to the columns dimension.

The matrix $\bar{\mathbf{H}}$ has a structure similar to $\mathbf{H}$ in (5.14) and (5.16) but with more columns. The resulting matrix $\bar{\mathbf{E}}$ now has more columns than rows and hence can

generally be inverted using the pseudo-inverse

$$\mathbf{E}^{+} = \bar{\mathbf{E}}^{T}(\bar{\mathbf{E}}\bar{\mathbf{E}}^{T})^{-1}, \tag{5.25}$$

resulting in

$$\bar{\mathbf{V}}\mathbf{E}^{+} = \bar{\mathbf{H}}. \tag{5.26}$$

Now the extended matrix $\bar{\mathbf{V}}$ can be used in (5.21), rather than $\mathbf{V}$, to construct a modified matrix $\widetilde{\bar{\mathbf{V}}}$, using a procedure similar to (5.19). Subsequent stages of the algorithm remain unchanged.

5.6 AIR Estimation in Subbands

The proposed method, although theoretically tractable, has several drawbacks when applied to real-life scenarios. The first problem stems from the tendency of AIRs in real room environments to be very long (2000 taps are commonly encountered even in regular offices). In such cases, the GEVD procedure is not robust enough and it is quite sensitive to small errors in the null subspace matrix [14]. Furthermore, the matrices involved become extremely large, imposing huge memory and computation requirements. Another problem arises from the wide dynamic range of the speech signal. This may result in erroneous estimates of the ATFs in the low energy bands of the input signal.

A common procedure for treating these problems is to spilt the full-band signal into several frequency subbands. We propose to incorporate the same subspace methods presented in Sects 5.4–5.5 into a subband structure. The use of subbands for splitting adaptive filters, especially in the context of echo cancellation, has gained recent interest in the literature [6, 39, 41, 42]. However, the use of subbands in subspace methods is not as common. The design of the subbands is of crucial importance. Special emphasis should be given to adjusting the subband structure to the problem at hand. In this contribution we only aim to demonstrate the ability of the method, thus a simple 6-channel subband structure, as depicted in Fig. 5.4, is used. Each of the channels is constructed by shifting a prototype low-pass FIR filter of order 150 to the appropriate position along the frequency axis. The filter was designed by applying the window method (using FDATool provided by MATLAB®). We used equi-spaced filters of equal bandwidth; however more advanced design methods can be adopted.

The M microphone signals are now filtered by the subband structure prior to the application of the subspace method. Although the resulting subband signal corresponds to a longer filter (which is the convolution of the corresponding AIR and the subband filter), the algorithm aims at the reconstruction of the actual AIR, ignoring the filter-bank roots. This can be attributed to the fact that the filter-bank roots are common to all channels. Recall that subspace methods are blind to common zeros, as discussed in Sect. 5.4. For properly exploiting the benefits of the subband

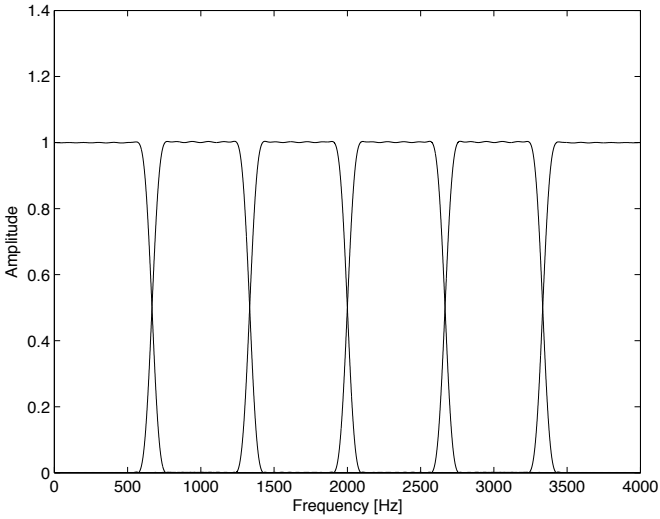


Fig. 5.4 Subband structure – six equi-spaced equi-bandwidth filters

structure, each subband signal should be decimated. We chose a critically decimated filter-bank, i.e., the decimation factor equals the number of bands. No special efforts were made to minimize the reconstruction error of the filter-bank. The decimation results in an AIR order reduction (per band) by approximately the decimation factor, relaxing the computation and memory demands of the AIR estimation task. As a direct result, the required overestimated order becomes much shorter than the corresponding order in the full-band case. Another benefit of the decimation may be a boost in performance, which is obtained because the signals processed in each subband are flatter in the frequency domain. Following the estimation of the decimated AIRs, they are combined in an appropriate synthesis system, comprised of interpolation followed by a synthesis filter-bank, which is similar to the analysis filter-bank. The overall subband system is depicted in Fig. 5.5, whereas the *AIR estimation* block was shown schematically in Fig. 5.2. The analysis and synthesis filters are denoted $P_k(z),\ k = 0,1,\ldots,K-1$ and $Q_k(z),\ k = 0,1,\ldots,K-1$, respectively.

5.7 Signal Reconstruction

The ultimate goal of the dereverberation algorithm is the reconstruction of the desired speech. In Sects. 5.4–5.6 we introduced a family of methods for estimating the AIRs. In this section we will use these estimates for equalizing the channels. Equalization schemes were extensively studied in the literature. Miyoshi and Kaneda [32] proposed a procedure, referred to as MINT, for inverting AIRs. They show that

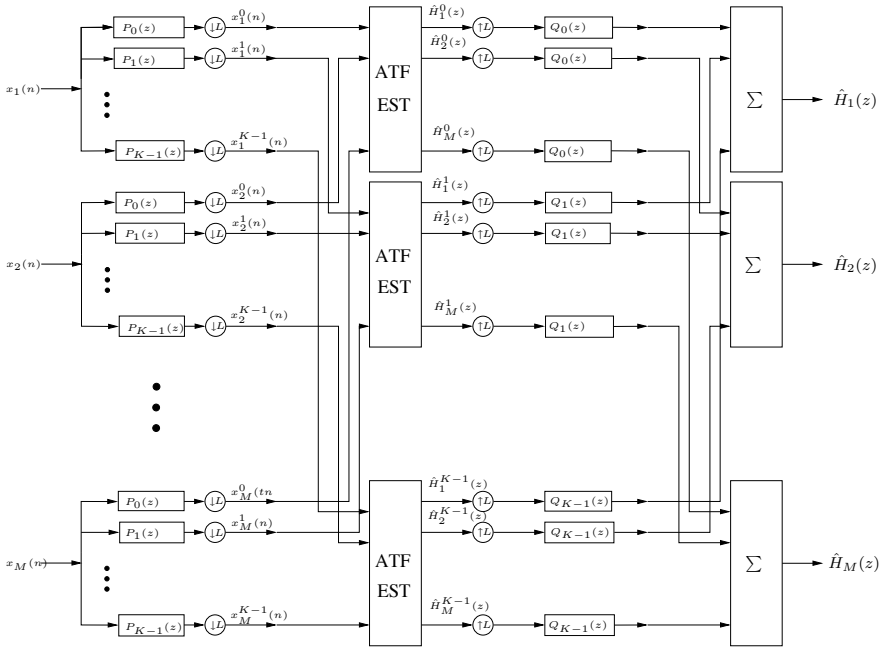


Fig. 5.5 Subband dereverberation system. $P_k(z),\ k = 0,1,\ldots,K-1$ are the analysis filters and $Q_k(z),\ k = 0,1,\ldots,K-1$ are the corresponding synthesis filters, which guarantee (almost) perfect reconstruction. The ATF estimation blocks, shown schematically in Fig. 5.2, implement the (full-band) dereverberation algorithm, independently applied in each subband

perfect reconstruction can be obtained with FIR filters when multiple observations are available, provided that the AIRs do no share any common zeros. Radlovic *et al.* [37] argued, however, that the MINT equalizer is very sensitive to uncertainties and variations in the positions of the source and microphone. The high sensitivity of the MINT procedure to inaccuracies in the AIR estimates is a direct consequence of this observation. Gaubitch *et al.* [13] extend previous work, presented by Yamada *et al.* [44], and introduce a subband version of the MINT algorithm. They show that the subband estimator is computationally more efficient and less sensitive to inaccuracies in the estimated AIRs compared to its full-band counterpart.

Let $g_m(n),\ m = 1,2,\ldots,M$ denote a set of M equalizers. Using (5.1) the estimated speech signal is given by:

$$\begin{aligned}\hat{s}(n) &= \sum_{m=1}^{M} g_m(n) * x_m(n) \\ &= \sum_{m=1}^{M} \left(g_m(n) * h_m(n) * s(n) + g_m(n) * v_m(n)\right). \end{aligned} \tag{5.27}$$

The goal of the equalization procedure is to obtain the filter set $g_m(n)$, $m = 1,2,\ldots,M$. Perfect equalization is obtained by satisfying:

$$\sum_{m=1}^{M} g_m(n) * h_m(n) = \delta(n). \tag{5.28}$$

In the z-domain this identity can be formulated as:

$$\sum_{m=1}^{M} G_m(z)H_m(z) = 1. \tag{5.29}$$

As an alternative to the MINT equalizer, a simple solution for eliminating the reverberation can be obtained by a Matched Filter Beamformer (MBF), also known as the *zero-forcing equalizer* in the communications field. A normalized version of the MBF [26] equalizer is given by

$$G_m(z) = \frac{H_m^*(1/z^*)}{\sum_{m=1}^{M} H_m(z)H_m^*(1/z^*)}, \tag{5.30}$$

where * denotes the conjugation operation. In the frequency domain the equalizer transfer function can be restated as:

$$G_m(e^{j\omega}) = \frac{H_m^*(e^{j\omega})}{\sum_{m=1}^{M} |H_m(e^{j\omega})|^2}. \tag{5.31}$$

This solution has two drawbacks. First, there is no guarantee that the noise term $\sum_{m=1}^{M} g_m(n) * v_m(n)$ is canceled out; in fact, in some cases it might even be amplified. The second drawback stems from the non-causality of the filter in (5.30). Hence, applying the filter set in (5.30) might result in a distorted reconstructed signal. The distortion is manifested as a response preceding the input signal, usually referred to as *pre-echo*. Note that applying inverse filtering, i.e., $G_m(z) = \frac{1}{H_m(z)}$ might have the same effect, since, as noted by Neely and Allen [35], typical AIRs are not minimum phase. In the sequel the performance of both MINT and MBF methods will be demonstrated and compared.

5.8 Experimental Study

The applicability of the subspace methods for mitigating the reverberation effect is verified in a series of simulations. Three Figures-Of-Merit (FOM) are used to quantify the obtained results. The first is a simple inspection of the estimated AIR and its corresponding ATF, in comparison with the actual filters. The second consists of assessment of sound spectrograms (sonograms) and time-domain waveforms comparing the input speech signal, the reverberant signal, as captured by one of the microphones, and the processed signal. This comparison tests both the system iden-

Table 5.1 NPM *vs.* SNR (both measured in dB) for white noise input. The number of microphones is $M = 2$, the AIR order is $L = 16$, and the AIR order assumed by the algorithm is $\hat{L} = 21$, i.e., the order was overestimated by 5 taps

SNR	15	20	25	30	35	40	45
NPM	−3.5	−8.6	−16.5	−28.0	−35.0	−44.0	−53

Table 5.2 NPM *vs.* SNR (both measured in dB) for speech signal input. The number of microphones is $M = 2$, the AIR order is $L = 16$, and the AIR order assumed by the algorithm is $\hat{L} = 21$, i.e., the order was overestimated by 5 taps

SNR	35	40	45	50	55	60	65
NPM	0.0	0.0	−2.0	−10.0	−11.0	−24.5	−38.0

tification procedure and the subsequent equalization procedure. The third FOM is the objective measure proposed by Morgan *et al.* [33]. This measure, denoted Normalized Projection Misalignment (NPM), is insensitive to the overall estimation scaling, which makes this FOM suitable for evaluating the estimation performance. The NPM is defined as

$$\begin{aligned}\mathrm{NPM} &= 10\log_{10}\left(\frac{1}{\|\mathbf{h}\|_2^2}\left\|\mathbf{h}-\frac{\mathbf{h}^T\hat{\mathbf{h}}}{\|\hat{\mathbf{h}}\|_2^2}\hat{\mathbf{h}}\right\|_2^2\right)\ \mathrm{dB}\\ &= 10\log_{10}\left(1-\left(\frac{\mathbf{h}^T\hat{\mathbf{h}}}{\|\mathbf{h}\|_2\|\hat{\mathbf{h}}\|_2}\right)^2\right)\ \mathrm{dB}. \end{aligned} \tag{5.32}$$

We start our experimental study with the full-band version of the subspace method, described in Sects. 5.4–5.5 and continue with the subband version described in Sect. 5.6.

5.8.1 Full-band Version – Results

In Tables 5.1–5.2 the dependency of the NPM on the SNR level is depicted. The results in Table 5.1 were obtained by using white noise input, while for the results in Table 5.2 speech input drawn from the TIMIT database [12], down-sampled to 8kHz, was used. In both cases, the length of the observation interval was 1 s, the AIRs' order was set to $L = 16$, and their coefficients were drawn from discrete uniform distribution. The number of microphones was set to $M = 2$. While applying the algorithm, the order of the filters was overestimated and set to $\hat{L} = 21$. The contaminating signal was a colored Gaussian noise. The noise power spectral density was estimated using noise-only segments, assumed to be available. For obtaining the results we used the median of 50 Monte Carlo trials. An NPM level above -10 dB can be regarded as being of unacceptable quality. It is evident by comparing the results

Table 5.3 NPM (measured in dB) *vs.* AIR order L for white noise input. The number of microphones is $M = 2$ and the SNR=50 dB

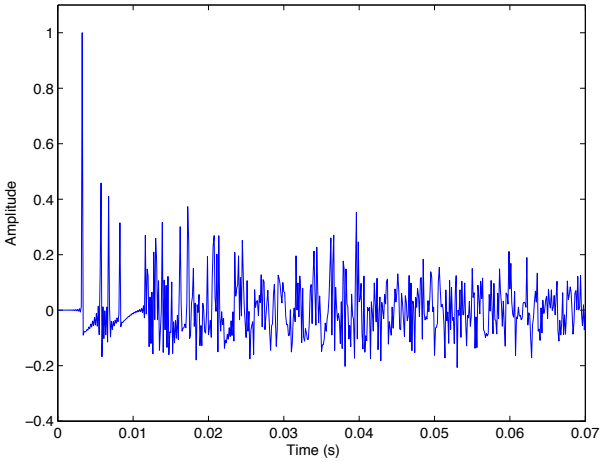
L	16	32	64	128	256
NPM	−60.0	−49.5	−33.0	−18.0	−0.5

in Tables 5.1 and 5.2 that the performance obtained by using speech signal input is significantly inferior to the performance obtained while using white noise input. The degradation might reach approximately 30 dB.

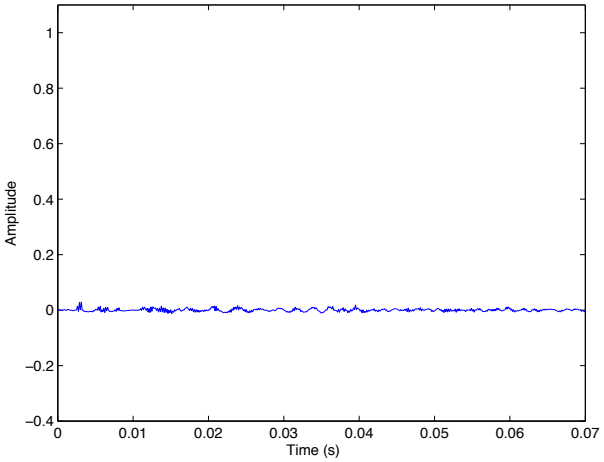
Next, we test the performance degradation as a function of the filter order. In this scenario the AIR coefficients were Gaussian distributed with a decaying envelope in accordance with Polack's time-domain model [36]. We gradually increased the filter order while maintaining SNR level of 50 dB. Again, the algorithm assumed the overestimated filter order $\hat{L} - L = 5$. The results are summarized in Table 5.3. It is shown that even for a relatively high SNR level the performance of the algorithm is subject to fast deterioration with increasing filter order. The algorithm is rendered useless for a filter order higher than $L = 128$.

Finally, we conducted several experiments using the image method [4] to simulate the AIRs. The implementation of the simulator can be found in [16]. In a noise-free scenario with white-noise input and reverberation time $T_{60} = 0.7$ s, the obtained median NPM was -73 dB for AIRs truncated to order $L = 512$, and overestimated by the algorithm to $\hat{L} = L + 5$. Due to the high computational burden, only five Monte Carlo experiments were conducted. To further demonstrate the quality of the estimation procedure in Fig. 5.6 we present the actual AIR together with its estimate for $SNR = 110$ dB and filter order $L = 600$ taps. Again, order overestimation by 5 taps was assumed. The resulting NPM for this simulation was -27 dB. Note that the decay rate of the AIR corresponds to the reverberation time $T_{60} = 0.7$ s, although the impulse response was truncated to a much shorter order of 600 taps, corresponding to a reverberation time as low as 0.07 s.

The reconstruction performance, for filter order $L = 500$, of the two equalization methods, namely MINT and MBF, is evaluated by an assessment of sound sonograms (depicted in Fig. 5.7). Although NPM is as low as -26 dB (i.e., a good match between the real filters and their estimates), it is clearly shown that the MINT equalizer fails to reconstruct the signal. This observation confirms the sensitivity analysis presented in [37]. The MBF equalizer, although superior to the MINT equalizer in this respect, suffers from several annoying artifacts. First, we observe the *pre-echo* effect, caused by the equalizer non-causality. Another artifact can be identified as *musical tones*. Although the reason for this artifact is not clear, we suspect that it can be attributed to erroneous estimation of the subspace method in sporadic frequencies.



(a)



(b)

Fig. 5.6 AIR relating the source and microphone #1 for $M = 2$ microphones, SNR=110 dB and filter order $L = 600$ taps. (a) Original AIR, (b) estimation error

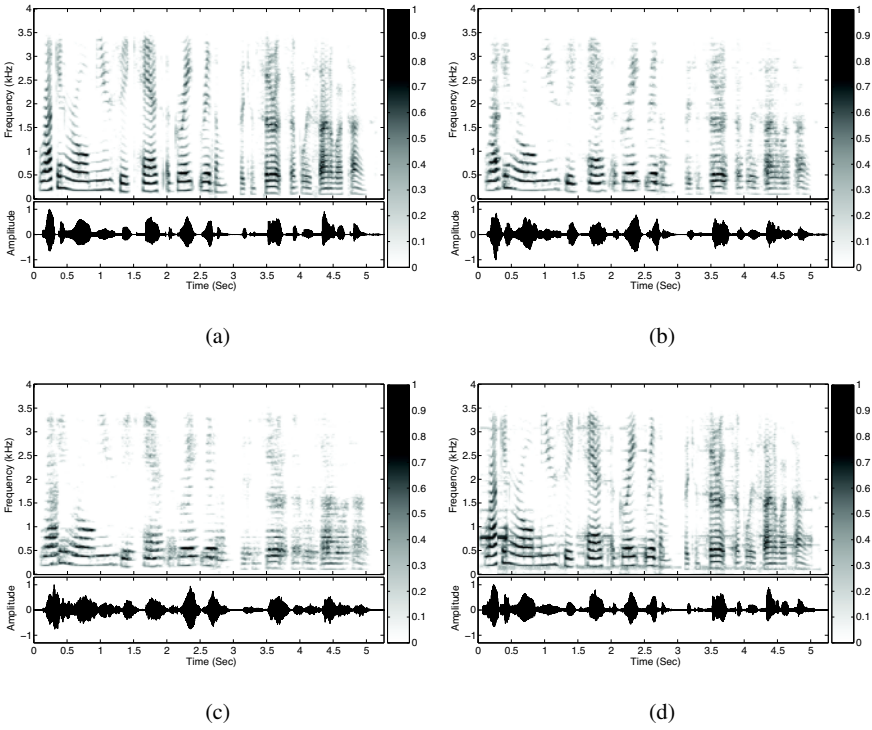


Fig. 5.7 Spectrograms and corresponding waveforms of (a) original speech signal $s(n)$, (b) reverberant speech at microphone #1 $y_1(n)$, (c) reconstructed speech signal $\hat{s}(n)$ using MINT equalization, and (d) reconstructed speech signal $\hat{s}(n)$ using MBF equalization. A two-microphone system was used in a noise-free scenario. The AIR order was set to $L = 500$ and the decay rate corresponds to $T_{60} = 0.7$ s. The full-band algorithm assumed overestimated AIR order of 5 taps

5.8.2 Subband Version – Results

For the subband estimation method 6 bands were used. A prototype filter 150 taps long was designed for keeping low overlap between bands. All bands are related to the prototype low-pass filter by a simple frequency shift. The full-band AIR order was chosen to have 24 taps, with decaying energy profile. Hence in each band we tried to estimate $24/6 = 4$ taps. In each band the overestimated filter order $\hat{L} = L + 2 = 6$ was assumed. The results for the noise-free scenario is presented in Fig. 5.8. For the reconstruction of the total frequency response we assumed the availability of the individual band gains. Hence, only the ability of the method to estimate the frequency shaping of the filters in each band is demonstrated, and the gain ambiguity phenomenon is ignored. It is shown that, although the estimation in each band is quite accurate, considerable misalignment in the reconstructed full-

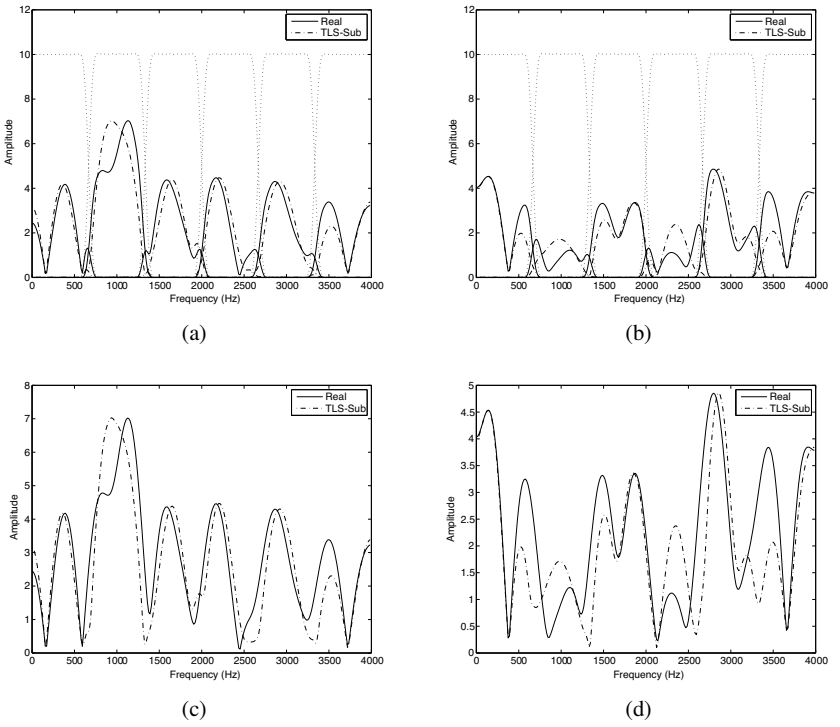


Fig. 5.8 Subband method for AIR estimation. (a) and (b) real and estimated filters per subband at microphone #1 and microphone #2, respectively. (c) and (d) the total ATFs estimated with the subband method and with gain ambiguity compensation at microphone #1 and microphone #2, respectively

band filters occurs. This phenomenon might be attributed to the small gaps between the bands.

5.9 Limitations of the Proposed Algorithms and Possible Remedies

In the previous sections we presented a family of subspace-based methods for estimating the AIRs. The method has two main variants, namely the full-band and the subband algorithms. Although mathematically tractable, the presented method has many limitations. This section is dedicated to a summary of the limitations and drawbacks presented throughout this chapter. Possible remedies to these limitations are proposed and potential research directions are discussed.

5.9.1 Noise Robustness

The main limitation of subspace methods is their lack of noise robustness. As depicted in Tables 5.1–5.2, the sensitivity of the full-band algorithm to additive noise is very high, especially when the AIR order becomes larger than a few hundred taps. This might be attributed to the sensitivity of the GEVD procedure [14]. The subband algorithm is also not very robust to noise conditions, although it is expected that, due to the shorter filter length in each subband, it might be less sensitive than the full-band method. The use of the MINT method for equalization emphasizes the robustness problem. It was observed that even with almost perfect estimation of the AIRs the MINT equalization ability might significantly differ from that of a perfect dereverberation algorithm. The subband MINT algorithm [13] is more robust to noise conditions. An interesting research direction might be a complete subband system, in which both AIR estimation and equalization are performed in subbands.

5.9.2 Computational Complexity and Memory Requirements

The proposed algorithms impose a high computational burden and memory requirements. Some of the algorithm's components, namely GEVD or GSVD, and the TLS procedure, are regarded as heavy consumers of resources. The GSVD procedure, which uses the entire data matrix, imposes severe memory requirements, which might be circumvented by calculating the GEVD of the correlation matrix instead. To achieve this, a direct calculation of the correlation matrix, which bypasses the need for the data matrix construction, is required. In any case, since AIRs tend to be very long filters (thousands of taps long), the correlation matrix dimensions become very high.

5.9.3 Common Zeros

Common zeros cannot be determined by the subspace method. The probability of encountering common zeros increases with the AIR order. However, the likelihood of this phenomenon decreases as the number of channels increases. An extension to the multi-microphone case was presented in Sect. 5.5.2. For a preliminary experimental study of the multi-microphone case, see [11].

It is shown in [29, 30] that the common zeros phenomenon can be confused with the overestimation of the AIR order. Therefore, in order to find the common zeros the exact AIR order must be known in advance.

5.9.4 The Demand for the Entire AIR Compensation

Subspace methods rely heavily on the assumption that the AIR order is overestimated. There is no subspace available when this constraint is not met. Therefore, the algorithm cannot be used for partial compensation of the reverberation effect. Such partial compensation might be very attractive for at least reducing the coloration of the signal, while giving over the dereverberation task to another algorithm. The appropriate use of an AIR model, such as Polack's model, might be helpful.

5.9.5 Filter-bank Design

The subband structure is sensitive to the subband filters' design. Two design constraints should be met. The first is the requirement for minimum overlap between channels. In critically decimated filter-banks, the overlap between bands might confuse the subspace method altogether. In this chapter we have used simple filter-bank design, whereas more advanced methods for filter-bank design that might take into account the particular constraints of the problem should be considered.

5.9.6 Gain Ambiguity

Gain ambiguity may be a major drawback of the subband algorithm. Recall that all the subspace methods' estimates are determined up to a gain factor. While in the full-band scheme the only consequence of this gain ambiguity is an overall scaling of the system output, in the subband scheme the different gain factors introduced in each subband, it is manifested as an arbitrary frequency shaping of the output signal. Hence, the reconstructed signal suffers from a distortion that might be as severe as the original reverberation. Several methods can be applied to mitigate the gain ambiguity problem. First, the original gain of the signals in each subband may be restored as an approximate gain adjustment. Another method was suggested by Rahbar *et al.* [38]. This method imposes an FIR structure on the impulse response of the dereverberation filters. The order of the these filters should be determined in advance. As this information is usually unavailable, the AIR order estimation obtained by the subspace method can be used instead. An alternative mitigation of the gain ambiguity problem was proposed in [7]. The proper use of these methods is still a topic for further research. As far as this chapter goes, the gain ambiguity problem is ignored, and the gain in each subband is assumed to be known.

5.10 Summary and Conclusions

This chapter was dedicated to multi-microphone speech dereverberation algorithms using subspace methods. The core of all the presented algorithms is the observation that the reverberating filters are embedded in the null subspace of the data received through multiple channels. The null subspace is estimated using either the generalized singular value decomposition of the data matrix or the generalized eigenvalue decomposition of the respective correlation matrix. The proposed algorithms address the problems of channel overestimation, additive colored noise, and the wide dynamic range of the speech signal. Two versions of the method, namely a full-band and subband variants, are presented. While the former is limited by the extremely high order of the AIRs, the latter suffers from the gain ambiguity inherent to subband methods.

Overall, as shown by the experimental study, both variants demonstrate a high sensitivity to the SNR level and the AIR order. At the current stage, the proposed algorithms are not capable of solving a dereverberation problem of realistic order, as only relatively short AIRs can be successfully treated.

However, several fascinating research paths that are due to be explored might, in the future, overcome all these limitations. We believe that, despite the gain ambiguity problem, subband structures might be able to bring the prospective solution for the dereverberation problem, perhaps in conjunction with other, more mature algorithms explored in other chapters of this book.

References

1. Affès, S., Grenier, Y.: A signal subspace tracking algorithm for microphone array processing of speech. IEEE Trans. Speech Audio Process. **5**(5), 425–437 (1997)
2. Ahmad, R., Gaubitch, N.D., Naylor, P.A.: A noise-robust dual filter approach to multichannel blind system identification. In: Proc. European Signal Processing Conf. (EUSIPCO). Poznan, Poland (2007)
3. Ahmad, R., Khong, A.W.H., Naylor, P.A.: A practical adaptive blind multichannel estimation algorithm with application to acoustic impulse responses. In: Proc. IEEE Int. Conf. Digital Signal Processing (DSP), pp. 31–34 (2007)
4. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
5. Doclo, S., Moonen, M.: Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 31–34. Darmstadt, Germany (2001)
6. Eneman, K., Moonen, M.: DFT modulated filter bank design for oversampled subband systems. Signal Processing **81**(9), 1947–1973 (2001)
7. Eneman, K., Moonen, M.: Ambiguity elimination in frequency-domain subspace identification. Internal Report 06-151, K. U. Leuven, Leuven, Belgium (2006)
8. Eneman, K., Moonen, M.: Multimicrophone speech dereverberation: Experimental validation. EURASIP J. Audio, Speech, Music Process. **2007**, Article ID 51831 (2007)
9. Gannot, S., Moonen, M.: Subspace methods for multi-microphone speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Darmstadt, Germany (2001)

10. Gannot, S., Moonen, M.: Speech dereverberation via subspace methods incorporating subband structure. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Kyoto, Japan (2003)
11. Gannot, S., Moonen, M.: Subspace methods for multimicrophone speech dereverberation. EURASIP J. on App. Signal Process. **2003**(1), 1074–1090 (2003). DOI http://dx.doi.org/10.1155/S1110865703305049
12. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: Acoustic-phonetic continuous speech corpus (TIMIT). CD-ROM (1991)
13. Gaubitch, N.D., Thomas, M.R.P., Naylor, P.A.: Subband method for multichannel least squares equalization of room transfer functions. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 21–24. New Paltz, NY, USA (2007)
14. Golub, G.H., van Loan, C.F.: Matrix computations, 3 edn. John Hopkins Series in the Mathematical Sciences. John Hopkins University Press (1996)
15. Gürelli, M.I., Nikias, C.L.: EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. IEEE Trans. Signal Process. **43**(1), 134–149 (1995)
16. Habets, E.A.P.: Room impulse response (RIR) generator. Online (2006). URL `http://home.tiscali.nl/ehabets/rir_generator.html`
17. Hasan, M.K., Benesty, J., Naylor, P.A., Ward, D.B.: Improving robustness of blind adaptive multichannel identification algorithms using constraints. In: Proc. European Signal Processing Conf. (EUSIPCO). Antalya, Turkey (2005)
18. Hasan, M.K., Naylor, P.A.: Analyzing effect of noise on blind adaptive multichannel identification algorithms: Robustness issue. In: Proc. European Signal Processing Conf. (EUSIPCO). Florence, Italy (2006)
19. Haykin, S. (ed.): Blind deconvolution, 4th edn. Prentice Hall (1994)
20. Hikichi, T., Delcroix, M., Miyoshi, M.: Blind dereverberation based on estimates of signal transmission channels without precise information on channel order. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 1069–1072. Philadelphia, USA (2005)
21. Hikichi, T., Delcroix, M., Miyoshi, M.: Speech dereverberation algorithm using transfer function estimates with overestimated order. Acoustical Science and Technology **27**(1), 28–35 (2006)
22. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. **51**(1), 11–24 (2003)
23. Huang, Y., Benesty, J., Chen, J.: A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. IEEE Trans. Speech Audio Process. **13**(5), 882–895 (2005)
24. Huffel, S.V., Park, H., Rosen, J.B.: Formulation and solution of structured total least norm problems for parameter estimation. IEEE Trans. Signal Process. **44**(10), 2464–2474 (1996)
25. Hughes, C.P., Nikeghbali, A.: The zeros of random polynomials cluster uniformly near the unit circle. Online (2007). URL `http://arxiv.org/abs/math.CV/0406376`. Ver. 3
26. Jan, E.E., Flanagan, J.: Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 917–920. Atlanta, Georgia, USA (1996)
27. Javidi, S., Gaubitch, N.D., Naylor, P.A.: An experimental study of the eigendecomposition methods for blind SIMO system identification in the presence of noise. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Paris, France (2006)
28. Knapp, C.H., Carter, G.C.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. **24**(4), 320–327 (1976)
29. Lin, X., Gaubitch, N.D., Naylor, P.A.: Two-stage blind identification of SIMO systems with common zeros. In: Proc. European Signal Processing Conf. (EUSIPCO). Florence, Italy (2006)
30. Lin, X., Gaubitch, N.D., Naylor, P.A.: Blind speech dereverberation in the presence of common acoustical zeros. In: Proc. European Signal Processing Conf. (EUSIPCO), pp. 389–393. Poznań, Poland (2007)

31. Liu, Q.G., Champagne, B., Kabal, P.: A microphone array processing technique for speech enhancement in a reverberant space. Speech Communication **18**(4), 317–334 (1996)
32. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
33. Morgan, D.R., Benesty, J., Sondhi, M.M.: On the evaluation of estimated impulse responses. IEEE Signal Process. Lett. **5**(7), 174–176 (1998). DOI 10.1109/97.700920
34. Moulines, E., Duhamel, P., Cardoso, J.F., Mayrargue, S.: Subspace methods for the blind identification of multichannel FIR filters. IEEE Trans. Signal Process. **43**(2), 516–525 (1995)
35. Neely, S.T., Allen, J.B.: Invertibility of a room impulse response. J. Acoust. Soc. Am. **66**(1), 165–169 (1979)
36. Polack, J.D.: La transmission de l'énergie sonore dans les salles. Thèse de doctorat d'etat, Université du Maine, La Mans (1988)
37. Radlović, B.D., Williamson, R., Kennedy, R.: Equalization in an acoustic reverberant environment: robustness results. IEEE Trans. Speech Audio Process. **8**(3), 311–319 (2000)
38. Rahbar, K., Reilly, J.P., Manton, J.H.: A frequency domain approach to blind identification of MIMO FIR systems driven by quasi-stationary signals. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1717–1720. Orlando, Florida, USA (2002)
39. Spriet, A., Moonen, M., Wouters, J.: A multichannel subband GSVD based approach for speech enhancement in hearing aids. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Darmstadt, Germany (2001)
40. Tong, L., Perreau, S.: Multichannel blind identification: From subspace to maximum likelihood methods. Proc. IEEE **86**(10), 1951–1968 (1998)
41. Weiß, S., Rice, G.W., Stewart, R.W.: Multichannel equalization in subbands. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, USA (1999)
42. Weiß, S., Stewart, R.W., Stenger, A., Rabenstein, R.: Performance limitations of subband adaptive filters. In: Proc. European Signal Processing Conf. (EUSIPCO), pp. 1245–1248. Rhodos, Greece (1998)
43. Xu, G., Liu, H., Tong, L., Kailath, T.: A least-squares approach to blind channel identification. IEEE Trans. Signal Process. **43**(12), 2982–2993 (1995)
44. Yamada, K., Wang, J., Itakura, F.: Recovering of broad band reverberant speech signal by subband MINT method. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 969–972. Toronto, Canada (1991)
45. Yang, B.: Projection approximation subspace tracking. IEEE Trans. Signal Process. **43**(1), 95–107 (1995)

Chapter 6
Adaptive Blind Multichannel System Identification

Andy W.H. Khong[1] and Patrick A. Naylor[2]

Abstract The use of adaptive algorithms for blind system identification in speech dereverberation was proposed recently. This chapter reviews adaptive multichannel system identification using minimization of the cross-relation error. One of the algorithms that adopt this approach is the Normalized Multichannel Frequency Domain Least Mean Square (NMCFLMS) algorithm. We show that, in the presence of additive noise, the coefficients of the adaptive filter employing NMCFLMS converge initially toward the true acoustic impulse responses after which they then misconverge. We provide a technique to address this misconvergence problem in NMCFLMS. This is achieved by reformulating the minimization problem into one involving a constraint. As will be shown, this constrained minimization problem requires knowledge of the direct-path components of the acoustic impulse responses and one of the main contributions of this work is to illustrate how these direct-path components can be estimated under practical conditions. We will then illustrate how these estimates can be incorporated into the proposed extended NMCFLMS (ext-NMCFLMS) algorithm so as to address the problem of misconvergence. The simulation results presented show the noise robustness of the proposed algorithm for both white Gaussian noise and speech inputs. In addition, we illustrate how errors due to the estimation of the direct-paths affect the performance of the proposed algorithm.

6.1 Introduction

Blind System Identification (BSI) techniques have been employed for applications such as communications [42], geophysical [29, 30] as well as multimedia signal processing [1]. In addition, BSI algorithms for acoustic channels have generated much interest in recent years due to the increase in quality of service for telecommunication transmission as well as innovation in consumer products including, but

[1]Nanyang Technological University, Singapore
[2]Imperial College London, UK

not limited to, tele-conferencing and video-conferencing applications. In contrast to non-blind system identification, such as for acoustic echo cancellation [6] where a known signal is used to facilitate the identification of the acoustic channel, BSI algorithms utilize signals received from the output of an unknown system for channel estimation. For the case of speech dereverberation, acoustic channels are first identified blindly using signals received from the microphones. These estimated channels are then used to design equalization filters in order to remove reverberation introduced by the acoustic channels.

Techniques for BSI can generally be classified into two main categories: Higher Order Statistics (HOS) and Second Order Statistics (SOS) methods. Comparisons between SOS and HOS methods were presented in [39]. Algorithms that are based on HOS such as those presented in [10, 24] employ fourth-order cumulants of the received microphone signals and require a large number of observation samples. In addition, these methods assume a linear time-invariant unknown system. As a result, these methods suffer from a slow rate of convergence, which in turn reduces their tracking capability. Since it is well-known that acoustic impulse responses are time-varying in nature [12, 28], HOS methods present an inherent challenge due to their inferior tracking ability [35].

To address problems associated with HOS based methods, SOS based algorithms such as presented in [34, 44] have been proposed. These algorithms utilize the Cross-Relation (CR) equality between the channels and the observed channel outputs for BSI. Subspace methods such as proposed in [33, 41] are based on the principle of orthogonality between the signal and noise subspaces. Utilizing the CR equality, these algorithms estimate the unknown system through the subspace decomposition of the received data matrix. The approach presented in [14] has been shown to be effective for the estimation of acoustic channels. Although channel decomposition can be achieved using numerically efficient algorithms such as the singular value decomposition) [15], one of the main concerns for subspace methods is that they are computationally expensive given the high order nature of acoustic impulse responses. In addition, the performance of these algorithms relies on the existence of numerically well-defined dimensions of the signal or noise subspace.

Adaptive approaches for BSI have been proposed in order to mitigate the problem of the high computational cost that exists in subspace methods as well as to address the lack of tracking capability in HOS based algorithms. One of the first adaptive BSI algorithms was proposed in [26] for communications systems utilizes the CR. This algorithm iteratively minimizes the sum square errors between the interchannel cross-correlation of the signals received from two channels using the recursive least squares (RLS) algorithm. More recently, the Normalized Multichannel Frequency Domain Least Mean Square (NMCFLMS) algorithm presented in [22] was developed and has been shown to be effective in identifying room impulse responses of length in the order of several hundred taps. This algorithm was then employed for dereverberation using a microphone array.

Although computationally appealing, one of the main problems of the NMCFLMS algorithm is that it suffers from misconvergence [18] in the presence of noise. It has been shown through simulations presented in [2, 3, 17] that the esti-

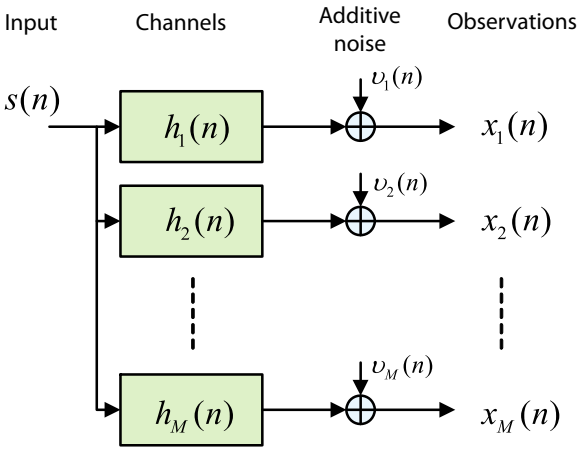


Fig. 6.1 Relationship between input and outputs of a SIMO model

mated filter coefficients converge first toward the correct impulse response of the acoustic system after which they then diverge from the true solution. Under low signal-to-noise ratio (SNR) conditions, the effect of misconvergence becomes more significant and occurs during earlier stages of adaptation.

The aim of this chapter is to review current adaptive approaches employing the CR for acoustic BSI. In addition, we study the effect of noise on the performance of one such algorithm, the NMCFLMS, and we address the misconvergence problem that exist in this algorithm. We first define in Sect. 6.2 the BSI problem and review conditions whereby the acoustic system can be identified. We next review algorithms employing the CR in Sect. 6.3. We then describe in Sect. 6.4 how additive noise affects the performance of NMCFLMS, otherwise known as the misconvergence problem. To address the problem of misconvergence, we propose in Sect. 6.5 a constraint based adaptive algorithm. The proposed extended NMCFLMS (ext-NMCFLMS) algorithm searches for the solution by introducing a constraint into the adaptation process. As will be shown in Sect. 6.5.2, this constraint requires knowledge of the direct-path components of the impulse responses. To obtain this information, the ext-NMCFLMS algorithm first extracts the delay of each direct-path component for each channel, in terms of their Time Differences Of Arrival (TDOA), using the Generalized Cross Correlation (GCC) algorithm [27]. As described in Sect. 6.5.4, in order to estimate the magnitude of the direct-path components, the ext-NMCFLMS algorithm monitors the cost function of the NMCFLMS using a novel cost function flattening point estimation (FPE) algorithm. This combined delay and magnitude information of the direct-path components is then used in the updating equation of the ext-NMCFLMS algorithm in order to address the misconvergence problem of the NMCFLMS. In Sect. 6.6 we show simulation results illustrating the noise robustness of the proposed ext-NMCFLMS using both White Gaussian Noise (WGN) and speech inputs.

6.2 Problem Formulation

We consider a Single-Input Multi-Output (SIMO) Finite Impulse Response (FIR) linear system consisting of M microphones as shown in Fig. 6.1. We define the m^{th} channel impulse response as

$$\mathbf{h}_m(n) = \big[h_{m,0}(n),\ h_{m,1}(n),\ldots,\ h_{m,L-1}(n)\big]^T, \qquad m = 1,\ldots,M, \tag{6.1}$$

and the additive noise for the corresponding channel as

$$\boldsymbol{\nu}_m(n) = \big[\nu_m(n),\ \nu_m(n-1),\ldots,\ \nu_m(n-L+1)\big]^T, \tag{6.2}$$

with $[\cdot]^T$ being the transposition operator and L being the length of the longest impulse response. The m^{th} channel output signal is then given by

$$\begin{aligned}\mathbf{x}_m(n) &= \mathbf{H}_m(n)\mathbf{s}(n) + \boldsymbol{\nu}_m(n)\\ &= \big[x_m(n),\ x_m(n-1),\ldots,\ x_m(n-L+1)\big]^T,\end{aligned} \tag{6.3}$$

where the source signal

$$\mathbf{s}(n) = \big[s(n),\ s(n-1),\ldots,\ s(n-2L+2)\big]^T \tag{6.4}$$

and the $L\times(2L-1)$ Sylvester matrix $\mathbf{H}_m(n)$ is defined as

$$\mathbf{H}_m(n) = \begin{bmatrix} h_{m,0}(n) & \ldots & h_{m,L-1}(n) & \ldots & 0\\ 0 & h_{m,0}(n) & \ldots & h_{m,L-1}(n) & 0\\ \vdots & \ddots & & \ddots & \vdots\\ 0 & \ldots & h_{m,0}(n) & \ldots & h_{m,L-1}(n)\end{bmatrix}. \tag{6.5}$$

Defining $E\{\cdot\}$ as the mathematical expectation operator, we assume that the additive noise between the M channels is uncorrelated such that

$$E\big\{\nu_m(n)\nu_l(n)\big\} = 0, \qquad \text{for } m \neq l, \tag{6.6}$$

$$E\big\{\nu_m(n)\nu_m(n-n')\big\} = 0, \qquad \text{for } n \neq n', \tag{6.7}$$

and that it is also uncorrelated with the input signal given by

$$E\big\{\nu_m(n)s(n)\big\} = 0. \tag{6.8}$$

In addition, we assume that the additive noise has a normal distribution with zero mean and variance of σ_ν^2, i.e., $\nu_m(n) \sim \mathcal{N}(0,\sigma_\nu^2)$, $\forall m$. The problem of BSI is then to estimate the acoustic impulse responses $\mathbf{h}_m(n)$ using received signals $\mathbf{x}_m(n)$ for channels $m = 1,\ldots,M$.

6.2.1 Channel Identifiability Conditions

Blind identification of a SIMO system as described above can be achieved with second order statistics of system outputs as long as the following conditions are met [44]:

1. *Channel diversity:* A multichannel model derived from a microphone array provides spatial diversity compared to a single channel model [1]. Diversity in this context refers to channels having different modes such that the acoustic channels, being modelled as finite impulse response filters, are *coprime*, i.e., they have no common zeros between their transfer functions [39]. If the acoustic channels are not coprime then one or more common factors exist across all channels. In the presence of common zeros, BSI algorithms fail to identify the channels correctly since they cannot distinguish whether these common terms are due to the input signal or the acoustic channels. The notion of near-common zeros and their effect on BSI have been studied and quantified in [25]. It has been shown that the performance of the adaptive NMCFLMS algorithm degrades with increasing number of near-common zeros for a multichannel acoustic BSI problem. In order to address the problem of near-common zeros across all acoustic channels, several microphones are employed for BSI.

2. *Full rank condition for* $\mathbf{S}(n)$*:* In order to estimate the unknown channels, the $L \times L$ Hankel matrix defined by

$$\mathbf{S}(n) = \begin{bmatrix} s(n) & s(n-1) & \dots & s(n-L+1) \\ s(n-1) & s(n-2) & \dots & s(n-L) \\ \vdots & \vdots & \dots & \vdots \\ s(n-L+1) & s(n-L) & \dots & s(n-2L+2) \end{bmatrix} \tag{6.9}$$

   for the source signal must be full-rank. This can be understood by first expressing, for the noiseless case,

$$\mathbf{S}(n)\mathbf{h}_m(n) = \mathbf{x}_m(n), \qquad \text{for } m = 1, \dots, M. \tag{6.10}$$

   Hence it can be seen that, if $\mathbf{S}(n)$ is rank deficient (6.10), will not have a unique solution even if the source signal $\mathbf{s}(n)$ is known.

We assume in the remainder of this chapter that both conditions as described above are satisfied in order to facilitate the identification of acoustic channels.

6.3 Review of Adaptive Algorithms for Acoustic BSI Employing Cross-relations

We review the formulation and development of three blind adaptive algorithms in the context of estimating acoustic channels. As will be shown, these algorithms are based on the minimization of an error formed using the cross-relation between the received microphone inputs and the estimated channels.

6.3.1 The Multichannel Least Mean Squares Algorithm

One of the first algorithms proposed for BSI in the context of acoustic channels is the Multichannel Least Mean Squares (MCLMS) algorithm [21]. We consider first a noiseless case with $\nu_m(n) = 0$, $\forall m,n$. The MCLMS algorithm utilizes the interchannel cross-relation of the received signal given by

$$\begin{aligned} x_m(n) * h_l(n) &= s(n) * h_m(n) * h_l(n) \\ &= x_l(n) * h_m(n), \end{aligned} \tag{6.11}$$

where $*$ represents linear convolution from which we obtain, in vector form,

$$\mathbf{x}_m^T(n)\mathbf{h}_l(n) = \mathbf{x}_l^T(n)\mathbf{h}_m(n), \tag{6.12}$$

for channel index $m,l = 1,\ldots,M$. Employing this cross-relation, an error between channels m and l for the adaptive filters with $m \neq l$ is given by

$$e_{ml}(n) = \mathbf{x}_m^T(n)\hat{\mathbf{h}}_l(n) - \mathbf{x}_l^T(n)\hat{\mathbf{h}}_m(n), \tag{6.13}$$

where

$$\hat{\mathbf{h}}_m(n) = \left[\hat{h}_{m,0}(n),\ldots,\hat{h}_{m,L-1}(n)\right]^T \tag{6.14}$$

is the estimated impulse response of the m^{th} channel. As a development of the well known Least Mean Squares (LMS) adaptive algorithm [20], the MCLMS algorithm then minimizes the normalized error

$$\varepsilon_{ml}(n) = \frac{e_{ml}(n)}{\left\|\hat{\mathbf{h}}(n)\right\|_2} \tag{6.15}$$

iteratively, where $\|\cdot\|_2$ is defined as the l_2-norm and

$$\hat{\mathbf{h}}(n) = \left[\hat{\mathbf{h}}_1^T(n),\ldots,\hat{\mathbf{h}}_M^T(n)\right]^T \tag{6.16}$$

is a concatenated $ML \times 1$ vector. We note that normalization of the error $e_{ml}(n)$ with $\|\hat{\mathbf{h}}(n)\|_2$ in (6.15) arises from a unit norm constraint imposed on the minimization of

$e_{ml}(n)$. This ensures that the estimated impulse responses in $\hat{\mathbf{h}}(n)$ for the MCLMS algorithm do not converge to the trivial solution of $\hat{\mathbf{h}}(n) = \mathbf{0}_{ML\times 1}$ where $\mathbf{0}_{ML\times 1}$ is the $ML \times 1$ null vector.

Using (6.15), the MCLMS algorithm is obtained by minimizing the cost function

$$\mathcal{J}(n) = \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \varepsilon_{ml}^2(n) \tag{6.17}$$

with respect to the estimated impulse responses $\hat{\mathbf{h}}_m(n)$ for channels $m = 1,\ldots,M$. The result of this minimization is the MCLMS algorithm given by [22]

$$\mathbf{R}(n) = \begin{bmatrix} \sum\limits_{m\neq 1} \mathbf{R}_{x_m x_m}(n) & -\mathbf{R}_{x_2 x_1}(n) & \ldots & -\mathbf{R}_{x_M x_1}(n) \\ -\mathbf{R}_{x_1 x_2}(n) & \sum\limits_{m\neq 2} \mathbf{R}_{x_m x_m}(n) & \ldots & -\mathbf{R}_{x_M x_2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{x_1 x_M}(n) & -\mathbf{R}_{x_1 x_M}(n) & \ldots & \sum\limits_{m\neq M} \mathbf{R}_{x_m x_m}(n) \end{bmatrix}, \tag{6.18}$$

$$\mathbf{R}_{x_m x_l}(n) = \mathbf{x}_m(n)\mathbf{x}_l^T(n), \tag{6.19}$$

$$\begin{aligned} \tilde{e}(n) &= \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} e_{ml}^2(n) \\ &= \mathcal{J}(n) \left\| \hat{\mathbf{h}}(n-1) \right\|_2^2, \end{aligned} \tag{6.20}$$

$$\hat{\mathbf{h}}(n) = \frac{\hat{\mathbf{h}}(n-1) - 2\mu \left[\mathbf{R}(n)\hat{\mathbf{h}}(n-1) - \tilde{e}(n)\hat{\mathbf{h}}(n-1) \right]}{\left\| \hat{\mathbf{h}}(n-1) - 2\mu [\mathbf{R}(n)\hat{\mathbf{h}}(n-1) - \tilde{e}(n)\hat{\mathbf{h}}(n-1)] \right\|_2}, \tag{6.21}$$

where μ in (6.21) is the step-size of the MCLMS algorithm with its optimal value being derived in [23]. We note that the normalization in (6.21) is equivalent to the normalized error in (6.15) but is imposed on every iteration of the update.

6.3.2 The Normalized Multichannel Frequency Domain LMS Algorithm

Frequency domain adaptive algorithms have been proposed to improve the computational efficiency and performance of time domain algorithms. As opposed to time domain algorithms, such as the LMS where adaptation is performed sample

by sample, frequency domain algorithms generally inherit two properties: (i) incorporation of block updating strategies and (ii) employment of the Fast Fourier Transform (FFT). A comprehensive overview of the development of frequency domain algorithms was presented in [37]. One of the first frequency domain adaptive algorithms proposed is the Fast LMS (FLMS) [13] where the overlap-save method of implementing linear convolution using FFT blocks is employed.

In a similar manner to FLMS, the MCLMS algorithm described in Sect. 6.3.1 has been extended to the frequency domain for efficient implementation. To describe the resulting NMCFLMS algorithm [22], we first define b as the frame index and $\mathbf{F}_{2L}$ as the $2L \times 2L$ Fourier matrix. We further define

$$\boldsymbol{\chi}_m(b) = \big[x_m(bL-L), x_m(bL-L+1), \ldots, x_m(bL+L-1)\big]^T \tag{6.22}$$

as the m^{th} channel time domain input frame and a $2L \times 2L$ diagonal matrix given by

$$\underline{\boldsymbol{\mathcal{D}}}_m(b) = \operatorname{diag}\big\{\mathbf{F}_{2L}\boldsymbol{\chi}_m(b)\big\}. \tag{6.23}$$

For clarity of presentation, we denote all frequency domain quantities with an underscore. Following the notation of [7], we also define $\mathbf{I}_{L\times L}$ as the $L \times L$ identity matrix and the following quantities:

$$\mathbf{W}^{01}_{L\times 2L} = \big[\mathbf{0}_{L\times L}\ \mathbf{I}_{L\times L}\big], \tag{6.24}$$

$$\mathbf{W}^{10}_{2L\times L} = \begin{bmatrix} \mathbf{I}_{L\times L} \\ \mathbf{0}_{L\times L} \end{bmatrix}, \tag{6.25}$$

$$\boldsymbol{\mathcal{W}}^{01}_{L\times 2L} = \mathbf{F}_L \mathbf{W}^{01}_{L\times 2L} \mathbf{F}^{-1}_{2L}, \tag{6.26}$$

$$\boldsymbol{\mathcal{W}}^{10}_{2L\times L} = \mathbf{F}_{2L} \mathbf{W}^{10}_{2L\times L} \mathbf{F}^{-1}_{L}, \tag{6.27}$$

$$\underline{\hat{\mathbf{h}}}_m(b) = \mathbf{F}_L \hat{\mathbf{h}}_m(b), \tag{6.28}$$

$$\underline{\hat{\mathbf{h}}}^{10}_m(b) = \mathbf{F}_{2L} \begin{bmatrix} \hat{\mathbf{h}}_m(b) \\ \mathbf{0}_{L\times 1} \end{bmatrix}. \tag{6.29}$$

The NMCFLMS update equation for BSI is then given by

$$\underline{\hat{\mathbf{h}}}^{10}_m(b) = \underline{\hat{\mathbf{h}}}^{10}_m(b-1) - \rho\big[\underline{\boldsymbol{\mathcal{P}}}_m(b) + \delta \mathbf{I}_{2L\times 2L}\big]^{-1} \times \sum_{l=1}^{M} \underline{\boldsymbol{\mathcal{D}}}^{*}_l(b)\underline{\boldsymbol{\varepsilon}}^{01}_{lm}(b), \tag{6.30}$$

where * denotes complex conjugate, $0 < \rho \leq 1$ and δ are the step-size and regularization constant, respectively, while the frequency domain cross-relation error and power spectrum are given by

$$\underline{\mathcal{E}}_{ml}^{01}(b) = \mathcal{W}_{L\times 2L}^{01}\Big[\underline{\mathcal{D}}_m(b)\mathcal{W}_{2L\times L}^{10}\hat{\underline{\mathbf{h}}}_l(b-1) - \underline{\mathcal{D}}_l(b)\mathcal{W}_{2L\times L}^{10}\hat{\underline{\mathbf{h}}}_m(b-1)\Big], \tag{6.31}$$

$$\underline{\mathcal{P}}_m(b) = \gamma\underline{\mathcal{P}}_m(b-1) + (1-\gamma)\sum_{l=1,l\neq m}^{M}\underline{\mathcal{D}}_l^*(b)\underline{\mathcal{D}}_l(b), \tag{6.32}$$

such that $\gamma = [1-1/(3L)]^L$ in (6.32) is the forgetting factor. A full derivation of the NMCFLMS algorithm is presented in [22].

Similar to the MCLMS algorithm reviewed in Sect. 6.3.1, the NMCFLMS algorithm avoids the trivial solution of $\hat{\mathbf{h}}_m(b) = \mathbf{0}_{L\times 1}$ by satisfying the unit norm constraint. It uses the initialization

$$\hat{\underline{\mathbf{h}}}_m^{10}(0) = \frac{1}{\sqrt{M}}\mathbf{1}_{2L\times 1}, \tag{6.33}$$

where $\mathbf{1}_{2L\times 1} = [1,\ldots,1]^T$ is a column vector of length $2L$. Results presented in [22] illustrate the ability of NMCFLMS to accurately identify $M = 5$ unknown room impulse responses each of length $L = 256$, which are sampled at 8 kHz. The NMCFLMS algorithm is summarized in Algorithm 6.1.

6.3.3 The Improved Proportionate NMCFLMS Algorithm

It is often observed that acoustic impulse responses contain many coefficients that have relatively small amplitude. Regions of small amplitude coefficients are attributed to the acoustic propagation delay from the source to the sensor and the late reflections from the enclosure. These features impart quasi-sparse characteristics to many acoustic impulse responses. In order to improve the convergence rate of NMCFLMS, the improved proportionate NMCFLMS (IPNMCFLMS) algorithm is proposed in [3]. This algorithm exploits the fast convergence, due to proportionality control, of the improved proportionate normalized least mean squares (IPNLMS) algorithm [8] originally proposed for sparse system identification such as used for network echo cancellation.

The IPNLMS algorithm achieves fast convergence by updating each filter coefficient with an individual step-size that is made proportional to the magnitude of the estimated impulse response. In a similar manner, the IPNMCFLMS algorithm incorporates proportionality into the NMCFLMS algorithm using a controlling factor α. To describe the IPNMCFLMS algorithm, we first define an $L\times L$ diagonal step-size control matrix $\mathbf{Q}_m(b)$ for channel index $m = 1,\ldots,M$ given by

$$\mathbf{Q}_m(b) = \mathrm{diag}\big\{q_{m,0}(b),\ldots,q_{m,L-1}(b)\big\}, \tag{6.34}$$

where elements $q_{m,p}(b)$ for elemental index $p = 0,\ldots,L-1$ are given by

Algorithm 6.1 The NMCFLMS algorithm [22]

Special matrices

$$\mathbf{W}_{2L\times L}^{10} = [\mathbf{I}_{L\times L}\ \mathbf{0}_{L\times L}]^T,$$
$$\mathbf{W}_{L\times 2L}^{01} = [\mathbf{0}_{L\times L}\ \mathbf{I}_{L\times L}],$$
$$\boldsymbol{\mathcal{W}}_{2L\times L}^{10} = \mathbf{F}_{2L}\mathbf{W}_{2L\times L}^{10}\mathbf{F}_{L}^{-1},$$
$$\boldsymbol{\mathcal{W}}_{L\times 2L}^{01} = \mathbf{F}_{L}\mathbf{W}_{L\times 2L}^{01}\mathbf{F}_{2L}^{-1}.$$

Initialization

$$0 < \rho \leq 1,$$
$$\gamma = [1-1/(3L)]^L,$$
$$\underline{\hat{\mathbf{h}}}_m^{10}(0) = \frac{1}{\sqrt{M}}\mathbf{1}_{2L\times 1}.$$

Algorithm

$$\boldsymbol{\chi}_m(b) = [x_m(bL-L),\ x_m(bL-L+1),\ldots,x_m(bL+L-1)]^T,$$
$$\underline{\boldsymbol{\mathcal{D}}}_m(b) = \mathrm{diag}\{\mathbf{F}_{2L}\boldsymbol{\chi}_m(b)\},$$
$$\underline{\boldsymbol{\varepsilon}}_{ml}^{01}(b) = ,\boldsymbol{\mathcal{W}}_{L\times 2L}^{01}\left[\underline{\boldsymbol{\mathcal{D}}}_m(b)\boldsymbol{\mathcal{W}}_{2L\times L}^{10}\underline{\hat{\mathbf{h}}}_l(b-1) - \underline{\boldsymbol{\mathcal{D}}}_l(b)\boldsymbol{\mathcal{W}}_{2L\times L}^{10}\underline{\hat{\mathbf{h}}}_m(b-1)\right],$$
$$\underline{\boldsymbol{\mathcal{P}}}_m(b) = \gamma\underline{\boldsymbol{\mathcal{P}}}_m(b-1) + (1-\gamma)\sum_{l=1,l\neq m}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\mathcal{D}}}_l(b).$$

Filter update

$$\underline{\hat{\mathbf{h}}}_m^{10}(b) = \underline{\hat{\mathbf{h}}}_m^{10}(b-1) - \rho_{\mathrm{e}}\left[\underline{\boldsymbol{\mathcal{P}}}_m(b) + \delta\mathbf{I}_{2L\times 2L}\right]^{-1}\times\sum_{l=1}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\varepsilon}}_{lm}^{01}(b).$$

$$q_{m,p}(b) = \frac{1-\alpha}{2L} + (1+\alpha)\frac{\left|\hat{h}_{m,p}(b)\right|}{2\left\|\hat{\mathbf{h}}_m(b)\right\|_1 + \phi}, \tag{6.35}$$

while ϕ is defined as the regularization parameter to avoid division by zero in (6.35). The coefficient update equation for the IPNMCFLMS algorithm is derived by first defining the matrix

$$\tilde{\mathbf{G}}_{2L\times 2L}^{10} = \mathbf{W}_{2L\times 2L}^{10}\mathbf{F}_{2L}^{-1}, \tag{6.36}$$

where

$$\mathbf{W}_{2L\times 2L}^{10} = \begin{bmatrix} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} \end{bmatrix}. \tag{6.37}$$

The IPNMCFLMS update equation can then be expressed as

$$\hat{\mathbf{h}}_m(b) = \hat{\mathbf{h}}_m(b-1) - \rho L\mathbf{Q}_m(b)\tilde{\mathbf{G}}_{2L\times 2L}^{10}\times \left[\underline{\boldsymbol{\mathcal{P}}}_m(b) + \delta\mathbf{I}_{2L\times 2L}\right]^{-1}\sum_{l=1}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\varepsilon}}_{lm}^{01}(b). \tag{6.38}$$

As can be seen from (6.35) and explained in [3], the filter update for IPNMCFLMS is performed in the time domain. More importantly, each filter coefficient is updated in a manner proportional to the estimated impulse response, which consequently gives rise to an improvement in convergence rate over NMCFLMS in many cases. The IPNMCFLMS algorithm is described by Algorithm 6.2.

Frequency domain adaptive algorithms such as NMCFLMS and IPNMCFLMS suffer from long delays when tracking time-varying acoustic impulse responses. To address this problem, the multi-delay filtering (MDF) structure [38] has been incorporated into the IPNMCFLMS algorithm where, for each microphone channel, the adaptive filter is partitioned into blocks of equal length for adaptation. As a result, the IPMCMDF algorithm presented in [3] achieves fast convergence and low delay. These beneficial properties are, respectively, due to the improved proportionate step-size control as well as the MDF structure.

6.4 Effect of Noise on the NMCFLMS Algorithm – The Misconvergence Problem

One of the main problems reported for the NMCFLMS algorithm is that it suffers from misconvergence in the presence of noise [17]. To illustrate the effect of noise on the NMCFLMS algorithm, we first define the Normalized Projection Misalignment (NPM) for the b^{th} frame given by [32]

$$\eta(b) = \frac{\left\|\mathbf{h}(b) - \alpha\hat{\mathbf{h}}(b)\right\|_2^2}{\left\|\mathbf{h}(b)\right\|_2^2}, \tag{6.39}$$

$$\alpha = \frac{\mathbf{h}^T(b)\hat{\mathbf{h}}(b)}{\hat{\mathbf{h}}^T(b)\hat{\mathbf{h}}(b)}. \tag{6.40}$$

It can be seen that the NPM measure quantifies the closeness of the estimated impulse responses $\hat{\mathbf{h}}(b)$ defined in (6.16) to the true impulse responses of the acoustic system $\mathbf{h}(b)$, where

$$\mathbf{h}(b) = \left[\mathbf{h}_1^T(b), \ldots, \mathbf{h}_M^T(b)\right]^T. \tag{6.41}$$

It is also common that BSI algorithms estimate the unknown system to within an arbitrary scaling factor [33, 40]. To account for this, the term α in (6.40) projects $\mathbf{h}(b)$ onto $\hat{\mathbf{h}}(b)$, which then allows the NPM measure to account for the unknown scaling factor.

Figure 6.2 illustrates an important result detailing the effect of noise on the performance of the NMCFLMS algorithm. In this simulation example, $M = 5$ impulse responses $\mathbf{h}_m(b)$ are generated using the method of images [4] with a sampling rate of $f_s = 8$ kHz. A reverberation time of $T_{60} = 64$ ms is used, giving rise to impulse responses of length $L = 512$. Figure 6.3 shows an illustration of the concatenated impulse responses generated, $\mathbf{h}(b)$, where we have normalized their amplitudes so

Algorithm 6.2 The IPNMCFLMS algorithm [3]

Special matrices

$$\mathbf{W}_{2L\times L}^{10} = [\mathbf{I}_{L\times L}\ \mathbf{0}_{L\times L}]^T,$$

$$\mathbf{W}_{L\times 2L}^{01} = [\mathbf{0}_{L\times L}\ \mathbf{I}_{L\times L}],$$

$$\boldsymbol{\mathcal{W}}_{2L\times L}^{10} = \mathbf{F}_{2L}\mathbf{W}_{2L\times L}^{10}\mathbf{F}_{L}^{-1},$$

$$\boldsymbol{\mathcal{W}}_{L\times 2L}^{01} = \mathbf{F}_{L}\mathbf{W}_{L\times 2L}^{01}\mathbf{F}_{2L}^{-1},$$

$$\mathbf{W}_{2L\times 2L}^{10} = \begin{bmatrix} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} \end{bmatrix},$$

$$\tilde{\mathbf{G}}_{2L\times 2L}^{10} = \mathbf{W}_{2L\times 2L}^{10}\mathbf{F}_{2L}^{-1}.$$

Initialization

$$0 < \rho \leq 1,$$

$$\gamma = [1-1/(3L)]^L,$$

$$\underline{\hat{\mathbf{h}}}_m^{10}(0) = \frac{1}{\sqrt{M}}\mathbf{1}_{2L\times 1}.$$

Step-size gain matrix

$$\mathbf{Q}_m(b) = \operatorname{diag}\{q_{m,0}(b),\ldots,q_{m,L-1}(b)\},$$

$$q_{m,p}(b) = \frac{1-\alpha}{2L} + (1+\alpha)\frac{|\hat{h}_{m,p}(b)|}{2\|\hat{\mathbf{h}}_m(b)\|_1 + \phi}.$$

Algorithm

$$\boldsymbol{\chi}_m(b) = [x_m(bL-L),\ x_m(bL-L+1),\ldots,x_m(bL+L-1)]^T,$$

$$\underline{\boldsymbol{\mathcal{D}}}_m(b) = \operatorname{diag}\{\mathbf{F}_{2L}\boldsymbol{\chi}_m(b)\},$$

$$\underline{\boldsymbol{\varepsilon}}_{ml}^{01}(b) = \boldsymbol{\mathcal{W}}_{L\times 2L}^{01}\left[\underline{\boldsymbol{\mathcal{D}}}_m(b)\boldsymbol{\mathcal{W}}_{2L\times L}^{10}\underline{\hat{\mathbf{h}}}_l(b-1) - \underline{\boldsymbol{\mathcal{D}}}_l(b)\boldsymbol{\mathcal{W}}_{2L\times L}^{10}\underline{\hat{\mathbf{h}}}_m(b-1)\right],$$

$$\underline{\boldsymbol{\mathcal{P}}}_m(b) = \gamma\underline{\boldsymbol{\mathcal{P}}}_m(b-1) + (1-\gamma)\sum_{l=1,l\neq m}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\mathcal{D}}}_l(b).$$

Filter update

$$\hat{\mathbf{h}}_m(b) = \hat{\mathbf{h}}_m(b-1) - \rho L\mathbf{Q}_m(b)\tilde{\mathbf{G}}_{2L\times 2L}^{10}\times\left[\underline{\boldsymbol{\mathcal{P}}}_m(b) + \delta\mathbf{I}_{2L\times 2L}\right]^{-1}\sum_{l=1}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\varepsilon}}_{lm}^{01}(b).$$

that the maximum of $\mathbf{h}(b)$ as defined in (6.41) equals 1 for clarity of presentation. For this illustrative example, we used a WGN input $\mathbf{s}(b)$, while an uncorrelated WGN $\boldsymbol{\nu}_m(b)$ was added to each of the received signals, as in (6.3), in order to achieve a signal-to-noise ratio (SNR) as depicted in Fig. 6.2. For each SNR, the step-size for the NMCFLMS algorithm was set to $\rho = 0.45$ and the forgetting factor of $\gamma = [1-1/(3L)]^L = 0.7165$ was used.

We note, from Fig. 6.2, that for each case of SNR, the NMCFLMS algorithm converges first towards the true solution $\mathbf{h}(b)$ after which it misconverges towards

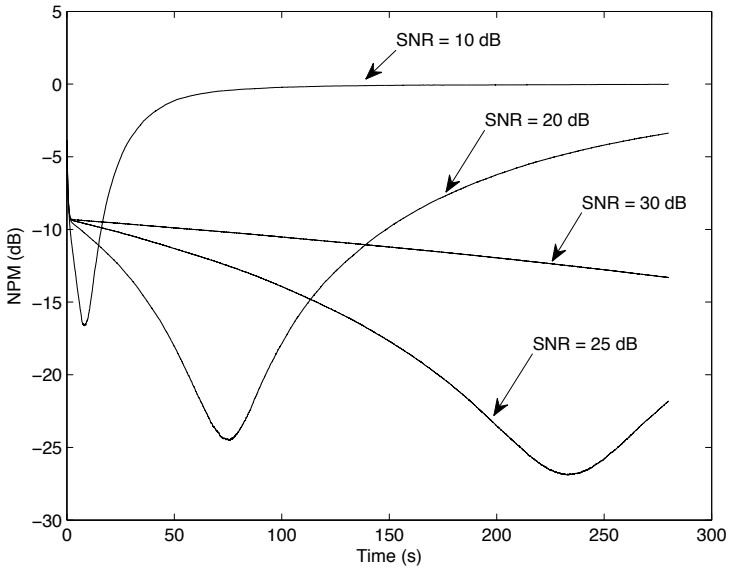


Fig. 6.2 Misconvergence of the NMCFLMS algorithm [$L = 512$, $f_s = 8$ kHz, $M = 5$, $\rho = 0.45$, $\gamma = [1 - 1/(3L)]^L = 0.7165$]

0 dB NPM. In addition, it can be seen that under low SNR conditions, the effect of misconvergence becomes more significant. Figure 6.4 illustrates, for the case of SNR $= 20$ dB, the estimated impulse responses $\hat{\mathbf{h}}(b)$ for the NMCFLMS algorithm at time $t = 256$ s after misconvergence. Comparing Figs. 6.3 and 6.4, we note that the misconverged solutions differ significantly from those of the true impulse responses, giving rise to an NPM of approximately -3.5 dB. More importantly, as can be seen from Fig. 6.4, although the condition $\left\|\hat{\mathbf{h}}(b)\right\|_2 = 1$ is satisfied, the solution $\hat{\mathbf{h}}(b) \to \mathbf{0}_{2L\times 1}$. This has the effect of NMCFLMS finding a solution $\hat{\mathbf{h}}(b)$ with the smallest energy that satisfies the unit norm constraint. This accounts for an NPM that approaches 0 dB. In the next section, we propose an algorithm robust to noise in order to address this misconvergence problem.

6.5 The Constraint Based ext-NMCFLMS Algorithm

We now propose an adaptive algorithm that addresses the problem of misconvergence for the NMCFLMS algorithm. We achieve this aim by first deriving the proposed extended NMCFLMS algorithm using first principles and introducing a penalty function into the cost function $\mathcal{J}(n)$ defined by (6.17). As will be described

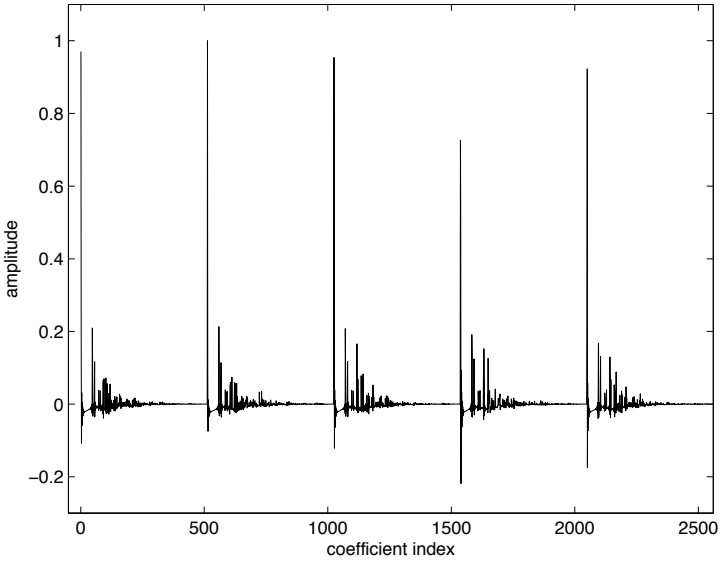


Fig. 6.3 True impulse responses to be estimated for $M = 5$ concatenated channels generated using the method of images [4]. Each impulse response is of length $L = 512$

in the following, the motivation for employing this methodology can be explained by first considering the effect of noise on the cost function $\mathcal{J}(n)$.

6.5.1 Effect of Noise on the Cost Function

We now consider the output of a multichannel system,

$$\mathbf{y}_m(n) = [y_m(n),\ y_m(n-1), \ldots, y_m(n-L+1)]^T, \tag{6.42}$$

in the presence of noise such as shown in Fig. 6.1, where

$$\mathbf{y}_m(n) = \mathbf{H}_m(n)\mathbf{s}(n), \tag{6.43}$$

$$\mathbf{x}_m(n) = \mathbf{y}_m(n) + \boldsymbol{\nu}_m(n). \tag{6.44}$$

Employing these relationships, the cross-relation error $e_{ml}(n)$ defined in (6.13) can then be expressed as [17]

$$e_{ml}(n) = e^{y}_{ml}(n) + e^{\nu}_{ml}(n), \tag{6.45}$$

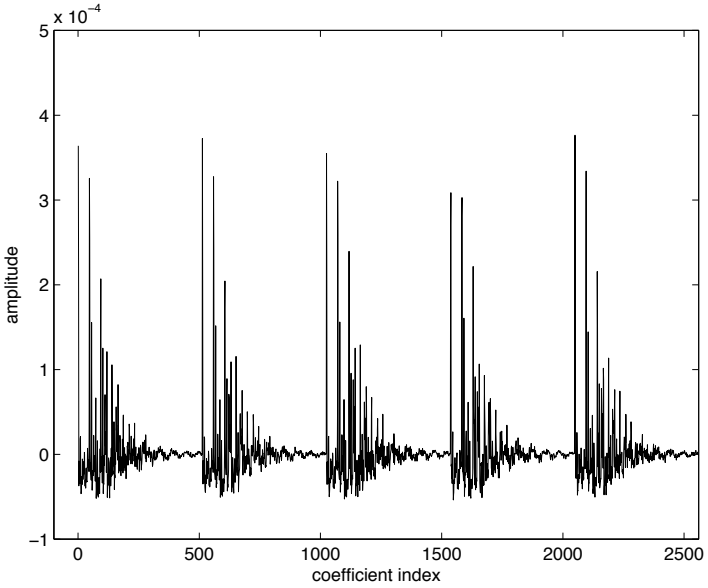


Fig. 6.4 Misconverged impulse responses $\hat{\mathbf{h}}(b)$ at time $t = 256$ s for the NMCFLMS algorithm with $M = 5$ at an SNR=20 dB. Each impulse response is of length $L = 512$

where

$$e^{y}_{ml}(n) = \mathbf{y}^{T}_{m}(n)\hat{\mathbf{h}}_{l}(n) - \mathbf{y}^{T}_{l}(n)\hat{\mathbf{h}}_{m}(n), \tag{6.46}$$

$$e^{\nu}_{ml}(n) = \boldsymbol{\nu}^{T}_{m}(n)\hat{\mathbf{h}}_{l}(n) - \boldsymbol{\nu}^{T}_{l}(n)\hat{\mathbf{h}}_{m}(n) \tag{6.47}$$

are the errors due to the received signals $\mathbf{y}_m(n)$, $\mathbf{y}_l(n)$ and additive noise $\boldsymbol{\nu}_m(n)$, $\boldsymbol{\nu}_l(n)$. We assume that the additive noise is zero mean and uncorrelated with the received signal giving $E\{y_m(n)\nu_m(n)\} = 0$. For this noisy case, the cost function $\mathcal{J}(n)$ defined in (6.17) can then be expressed as

$$\mathcal{J}(n) = \sum_{m=1}^{M-1}\sum_{l=m+1}^{M}\left\{[\varepsilon^{y}_{ml}(n)]^2 + [\varepsilon^{\nu}_{ml}(n)]^2\right\}$$

$$= \mathcal{J}_{y}(n) + \mathcal{J}_{\nu}(n), \tag{6.48}$$

where

$$\mathcal{J}_y(n) = \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \left[\varepsilon_{ml}^{y}(n)\right]^2, \tag{6.49}$$

$$\mathcal{J}_\nu(n) = \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \left[\varepsilon_{ml}^{\nu}(n)\right]^2. \tag{6.50}$$

Hence, we can see from (6.48) that the term $\mathcal{J}_\nu(n)$ can be viewed as a penalty function attached to the desired cost function $\mathcal{J}_y(n)$. The aim of the proposed algorithm is then to penalize $\mathcal{J}_y(n)$ using a penalty term when minimizing $\mathcal{J}(n)$. As evident, since $\boldsymbol{\nu}_m(n)$ is unknown in real applications, the constraint $\mathcal{J}_\nu(n)$ is unavailable and hence is not within the control of the adaptive algorithm.

6.5.2 Penalty Term Using the Direct-path Constraint

We now consider an alternative approach to introducing a penalty constraint on the cost function $\mathcal{J}(n)$ in order to address the misconvergence problem of NMCFLMS. As can be seen from Fig. 6.4, the misconverged solution $\hat{\mathbf{h}}(n)$ consist of significantly less energy compared to that of the true system as shown in Fig. 6.3. Therefore, it can be seen that minimizing $\mathcal{J}(n)$ in (6.17) results in solutions where $\mathbf{h}(n)$ misconverges to small values while satisfying $\|\hat{\mathbf{h}}(n)\|_2 = 1$. Consequently, the final misconverged solution when $n \to \infty$ occurs with elements in $\hat{\mathbf{h}}(n)$ having the smallest magnitudes whilst satisfying the unit norm constraint.

In order to prevent NMCFLMS from misconverging to small values, an additional constraint can be imposed such that the estimated direct path coefficient of each channel $\hat{h}_{\mathrm{dp},m}$ is equivalent to that of the actual direct-path coefficient $h_{\mathrm{dp},m}$ in terms of both *delay* and *magnitude* [2]. The subscript "dp" denotes the direct-path component of the respective acoustic impulse response where the direct-path is defined as the received source signal with a delay given by the source-microphone distance divided by the speed of sound. It is important to note that the direct-path of the impulse response does not necessarily correspond to the largest magnitude, since the largest magnitude might correspond to multiple reflections arriving in-phase at the microphone.

By using the above constraint-based approach, we are limiting the search space of the proposed algorithm, avoiding solutions where $\hat{\mathbf{h}}(n)$ are small. To achieve this, we start by introducing a penalty term to the cost function $\mathcal{J}(n)$ from which we then minimize

$$\mathcal{J}(n) = \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \varepsilon_{ml}^{2}(n), \tag{6.51}$$

subject to the constraint

$$\hat{h}_{\mathrm{dp},m} = h_{\mathrm{dp},m}, \tag{6.52}$$

for channel index $m = 1,\ldots,M$. Employing the method of Lagrange multipliers [31], the cost function can be reformulated as

$$\mathcal{J}_{\mathrm{dp}}(n) = \sum_{m=1}^{M-1}\sum_{l=m+1}^{M} \varepsilon_{ml}^2(n) + \beta \sum_{m=1}^{M} \big[h_{\mathrm{dp},m}(n) - \hat{h}_{\mathrm{dp},m}(n)\big]^2, \tag{6.53}$$

where the term β is the multiplier constant. Comparing (6.17) and (6.53), the penalty term $\beta\sum_{m=1}^{M}[h_{i,\mathrm{dp}}(n) - \hat{h}_{i,\mathrm{dp}}(n)]^2$ can be viewed as a correction term for the algorithm in the presence of noise. The significance of this correction term is then controlled by the multiplier term β such that under low SNR conditions, a high value of β is required to reduce the effect of noise on $\mathcal{J}(n)$.

To derive the adaptive formulation, the gradient of the penalty term

$$\mathcal{J}_p(n) = \sum_{m=1}^{M} \big[h_{\mathrm{dp},m}(n) - \hat{h}_{\mathrm{dp},m}(n)\big]^2 \tag{6.54}$$

can be obtained as

$$\begin{aligned}\nabla\mathcal{J}_p(n) &= \frac{\partial\mathcal{J}_p(n)}{\partial\hat{\mathbf{h}}_m(n)} \\ &= -2\big[h_{\mathrm{dp},m}(n) - \hat{h}_{\mathrm{dp},m}(n)\big]\mathbf{u}_m,\end{aligned} \tag{6.55}$$

with

$$\mathbf{u}_m = \big[\mathbf{0}_{1\times\tau_m-1} \;\; 1 \;\; \mathbf{0}_{1\times L-\tau_m}\big]^T, \tag{6.56}$$

while τ_m in (6.56) defines the Time Difference Of Arrival (TDOA) of the direct-path coefficient for the m^{th} channel with respect to the channel having the earliest direct-path. Employing the gradient vector given in (6.55), the update equation for the proposed ext-NMCFLMS algorithm is then given by

$$\begin{aligned}\underline{\hat{\mathbf{h}}}_m^{10}(b) = \underline{\hat{\mathbf{h}}}_m^{10}(b-1) &- \rho_{\mathrm{e}}\big[\underline{\boldsymbol{\mathcal{P}}}_m(b) + \delta\mathbf{I}_{2L\times 2L}\big]^{-1} \times \sum_{l=1}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\varepsilon}}_{lm}^{01}(b) \\ &+ 2\beta\rho_{\mathrm{e}}\mathbf{F}_{2L}\mathbf{W}_{2L\times L}^{10}\Big\{\big[h_{\mathrm{dp},m}(b) - \hat{h}_{\mathrm{dp},m}(b)\big]\mathbf{u}_m\Big\},\end{aligned} \tag{6.57}$$

where ρ_{e} is the step-size.

Comparing (6.57) with (6.30) for the NMCFLMS algorithm, we note that the additional term $2\beta\rho_{\mathrm{e}}\mathbf{F}_{2L}\mathbf{W}_{2L\times L}^{10}\Big\{\big[h_{\mathrm{dp},m}(b) - \hat{h}_{\mathrm{dp},m}(b)\big]\mathbf{u}_m\Big\}$ arises due to the penalty function introduced into $\mathcal{J}(n)$ as described by (6.53). In contrast to the algorithm proposed in [17] where the direct-path components $h_{\mathrm{dp},m}(b)$ are substituted into $\hat{h}_{\mathrm{dp},m}(b)$ at each block iteration, the proposed ext-NMCFLMS algorithm employs the penalty term in the filter update equation given by (6.57). The algorithm presented in [17] searches for the solution within the whole subspace $\mathbf{h}_m(b) \in \mathcal{R}^L$. The estimated solution is then obtained by substituting $\hat{h}_{\mathrm{dp},m}(b) = h_{\mathrm{dp},m}(b)$ at each update iteration. In contrast, the proposed ext-NMCFLMS algorithm imposes a lim-

iting constraint such that the search for solutions is constrained within the subspace containing $\hat{h}_{\mathrm{dp},m}(b) = h_{\mathrm{dp},m}(b)$, and as a consequence, the convergence rate of ext-NMCFLMS is higher compared to that of the algorithm proposed in [17].

As can be seen from (6.57), the proposed ext-NMCFLMS algorithm requires knowledge of the direct-path component $h_{\mathrm{dp},m}(b)$. This requires the estimation of the magnitude $|h_{\mathrm{dp},m}(b)|$ and the TDOA τ_m for each channel, as will be described below.

6.5.3 Delay Estimation

We now describe the estimation of τ_m for each channel. As explained, τ_m defines the TDOA of the direct-path coefficient between the m^{th} channel and the channel having the earliest direct-path, and we define $\hat{\tau}_m$ as the estimated TDOA.

One of the most popular algorithms for TDOA estimation is the GCC algorithm [27], which is realized using two pre-filters followed by a cross-correlator. The TDOA for each channel is then obtained by identifying the time-lag corresponding to the highest cross-correlation between the filtered output of the microphones $\mathbf{x}_m(b)$ for $m = 1, \ldots, M$. The estimated time delay between the m^{th} channel and that having the earliest direct-path is thus given by

$$\hat{\tau}_m = \arg\max_n \hat{\Psi}_m(n), \tag{6.58}$$

where

$$\hat{\Psi}_m(n) = \sum_{k=0}^{2L-1} \Phi_m(k)\hat{S}_{ml}(k)e^{j2\pi nk/L}, \tag{6.59}$$

$j = \sqrt{-1}$ and

$$\hat{S}_{ml}(k) = \underline{\chi}_m(k)\underline{\chi}_l^*(k) \tag{6.60}$$

is the cross-spectrum estimate between the m^{th} channel and the channel having the earliest direct-path. The variable $\underline{\chi}_m(k)$ is the k^{th} element of the $2L \times 1$ vector

$$\begin{aligned}\underline{\chi}_m(b) &= \mathbf{F}_{2L}\boldsymbol{\chi}_m(b) \\ &= \left[\underline{\chi}_m(0), \underline{\chi}_m(1), \ldots, \underline{\chi}_m(2L-1)\right]^T,\end{aligned} \tag{6.61}$$

given that $\boldsymbol{\chi}_m(b)$ is defined in (6.22) while, in a similar manner, $\underline{\chi}_l(k)$ is the k^{th} element of $\underline{\chi}_l(b)$ with $\boldsymbol{\chi}_l(b)$ being the b^{th} frame of the input corresponding to the channel having the earliest direct-path. It is important to note that, since the true impulse response is unknown, the channel having the earliest direct-path can be obtained using the estimated impulse response from the NMCFLMS algorithm.

The performance of the GCC algorithm is dependent on the choice of the pre-filters. The variable $\Phi_m(k)$ in (6.59) defines the weighting function for the pre-filter employed by the GCC algorithm. As explained in [27], the purpose of the pre-filters

is to reduce the spreading of the delta function, which in turn improves the accuracy of estimating the peak of the correlation sequence. Pre-filters that have been proposed to improve the performance of GCC include the Roth processor [36], the Smoothed COherence Transform (SCOT) [11], the Hannan–Thomson (HT) processor [16] and the Hassab-Boucher transform [19]. One of the most popular pre-filters for the GCC is the PHAse Transform (PHAT) where the variable $\Phi_m(k)$ is defined by

$$\Phi_m(k) = \frac{1}{\left|\hat{S}_{ml}(k)\right|}. \tag{6.62}$$

Substituting (6.62) into (6.59), it can be seen that the frequency components of the cross-spectrum are weighted inversely with respect to their magnitude. For reverberant speech, the GCC is performed on the Hilbert envelope of the linear prediction residual of input speech as shown in [45]. Results presented in [9, 43] showed that the PHAT processor achieves the best performance in the presence of reverberation and, as a consequence, we employ the PHAT processor for the estimation of the TDOA components τ_m in $\mathbf{u}_m$ described by (6.56).

6.5.4 Flattening Point Estimation

The proposed update equation for the ext-NMCFLMS algorithm in (6.57) requires knowledge of the true direct-path component $h_{\mathrm{dp},m}$ for channels $m = 1,\ldots,M$. In this section, we propose an online cost function flattening point estimation (FPE) algorithm for the iterative estimation of $|h_{\mathrm{dp},m}|$.

Figures 6.5 and 6.6 show, for two different SNRs of 10 and 15 dB respectively, the relationship between $\tilde{\mathcal{J}}(b)$ and NPM $\eta(b)$ of the system for the NMCFLMS algorithm where $\tilde{\mathcal{J}}(b)$ is defined as the cumulative sum of the cost function $\mathcal{J}(b)$ given by

$$\tilde{\mathcal{J}}(b) = \sum_{k=1}^{b} \left[10\log_{10}\mathcal{J}(k)\right]. \tag{6.63}$$

In these illustrative examples, $M = 5$ impulse responses each of length $L = 512$ as shown in Fig. 6.3 are generated using the method of images [4]. For each case of SNR, the step-size of the NMCFLMS algorithm is set to $\rho = 0.45$, and the forgetting factor $\gamma = [1 - 1/(3L)]^L = 0.7165$ is used. It is clear from these illustrative results that the cumulative sum of the cost function $\tilde{\mathcal{J}}(b)$ converges to a steady-state before the NMCFLMS algorithm misconverges. More importantly, it can be seen that convergence for $\tilde{\mathcal{J}}(b)$ exists even for an SNR as low as 10 dB. Hence we note that $|\hat{h}_{\mathrm{dp},m}|$ provides an estimate of the relative magnitudes of the direct-path components when $\tilde{\mathcal{J}}(b)$ converges to its steady-state.

To compute the convergence time of $\tilde{J}(b)$ at each block iteration, we evaluate the evolution of $\tilde{\mathcal{J}}(b)$ given by

$$\triangle\tilde{\mathcal{J}}(b) = \left|\tilde{\mathcal{J}}(b) - \tilde{\mathcal{J}}(b-1)\right|. \tag{6.64}$$

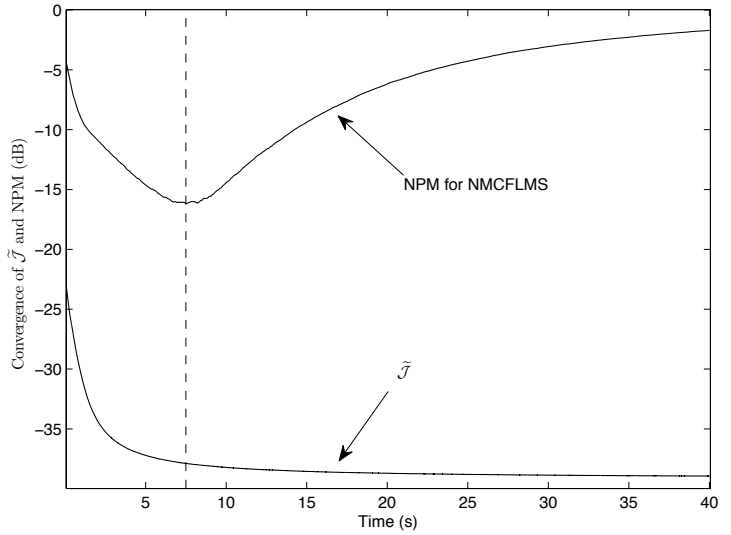


Fig. 6.5 Relationship between misconvergence for NMCFLMS and $\tilde{J}$ with SNR=10 dB

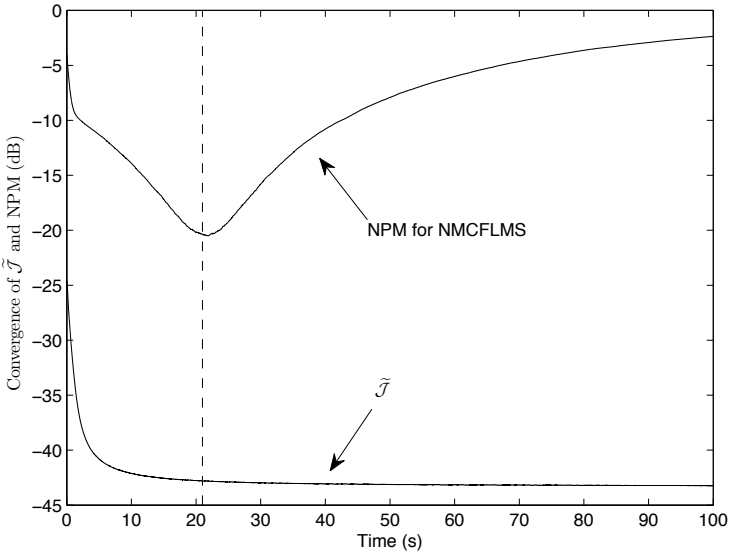


Fig. 6.6 Relationship between misconvergence for NMCFLMS and $\tilde{J}$ with SNR=15 dB

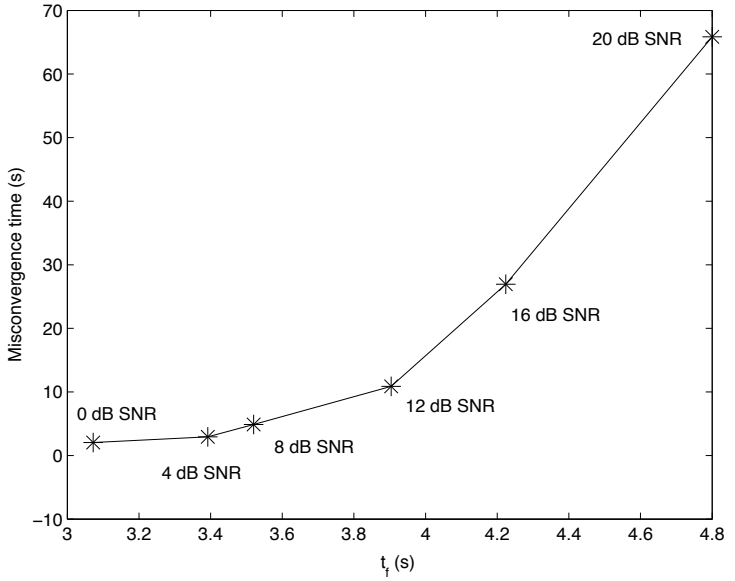


Fig. 6.7 Time to misconverge for NMCFLMS against t_f under different SNR conditions

We then define convergence for $\tilde{\mathcal{J}}(b)$ as the time taken for this change to become less than 0.05 dB as t_f given by

$$t_f \leftarrow \arg \triangle \tilde{\mathcal{J}}(b) \leq 0.05 \text{ dB}. \tag{6.65}$$

We define NPM misconvergence time as the time for NMCFLMS to reach its minimum NPM. For the NMCFLMS algorithm, Fig. 6.7 shows the variation of NPM misconvergence time with t_f under different SNR conditions. As before, for each case of SNR, the step-size for the NMCFLMS algorithm is set to $\rho = 0.45$, and the forgetting factor $\gamma = [1 - 1/(3L)]^L = 0.7165$ is used. We note from this illustrative example that t_f increases monotonically with the misconvergence time as SNR increases. As a consequence, t_f provides a good estimate of the point of misconvergence for the NMCFLMS algorithm.

Unlike the algorithm proposed in [17], where the direct path of the true impulse response is substituted, the proposed ext-NMCFLMS algorithm employs (6.64) and (6.65) and continually monitors the convergence point of $\tilde{\mathcal{J}}(b)$. Once t_f is reached, the magnitude of the estimated direct-path components are extracted and utilized according to (6.57) during the remaining adaptation of the proposed ext-NMCFLMS algorithm. By removing the often used unit-norm constraint, we find, through simulation results shown in the next section, that the knowledge of the relative magnitudes of the direct-path components is the only essential requirement to

Algorithm 6.3 The extended-NMCFLMS algorithm

Special matrices

$$\mathbf{W}_{2L\times L}^{10} = [\mathbf{I}_{L\times L}\ \mathbf{0}_{L\times L}]^T,$$
$$\mathbf{W}_{L\times 2L}^{01} = [\mathbf{0}_{L\times L}\ \mathbf{I}_{L\times L}],$$
$$\boldsymbol{\mathcal{W}}_{2L\times L}^{10} = \mathbf{F}_{2L}\mathbf{W}_{2L\times L}^{10}\mathbf{F}_L^{-1},$$
$$\boldsymbol{\mathcal{W}}_{L\times 2L}^{01} = \mathbf{F}_L\mathbf{W}_{L\times 2L}^{01}\mathbf{F}_{2L}^{-1}.$$

Initialization

$$0 < \rho \leq 1,$$
$$\gamma = [1-1/(3L)]^L,$$
$$\underline{\hat{\mathbf{h}}}_m^{10}(0) = \frac{1}{\sqrt{M}}\mathbf{1}_{2L\times 1}.$$

Algorithm

$$\boldsymbol{\chi}_m(b) = [x_m(bL-L),\ x_m(bL-L+1),\ldots,x_m(bL+L-1)]^T,$$
$$\underline{\boldsymbol{\mathcal{D}}}_m(b) = \mathrm{diag}\{\mathbf{F}_{2L}\boldsymbol{\chi}_m(b)\},$$
$$\underline{\boldsymbol{\varepsilon}}_{ml}^{01}(b) = \boldsymbol{\mathcal{W}}_{L\times 2L}^{01}\left[\underline{\boldsymbol{\mathcal{D}}}_m(b)\boldsymbol{\mathcal{W}}_{2L\times L}^{10}\underline{\hat{\mathbf{h}}}_l(b-1) - \underline{\boldsymbol{\mathcal{D}}}_l(b)\boldsymbol{\mathcal{W}}_{2L\times L}^{10}\underline{\hat{\mathbf{h}}}_m(b-1)\right],$$
$$\underline{\boldsymbol{\mathcal{P}}}_m(b) = \gamma\underline{\boldsymbol{\mathcal{P}}}_m(b-1) + (1-\gamma)\sum_{l=1,l\neq m}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\mathcal{D}}}_l(b).$$

Filter update

$$\underline{\hat{\mathbf{h}}}_m^{10}(b) = \underline{\hat{\mathbf{h}}}_m^{10}(b-1) - \rho_{\mathrm{e}}\left[\underline{\boldsymbol{\mathcal{P}}}_m(b) + \delta\mathbf{I}_{2L\times 2L}\right]^{-1}\times\sum_{l=1}^{M}\underline{\boldsymbol{\mathcal{D}}}_l^*(b)\underline{\boldsymbol{\varepsilon}}_{lm}^{01}(b)$$
$$+2\beta\rho_{\mathrm{e}}\mathbf{F}_{2L}\mathbf{W}_{2L\times L}^{10}\left\{\left[\hat{\hat{h}}_{\mathrm{dp},m}(b) - \hat{h}_{\mathrm{dp},m}(b)\right]\mathbf{u}_m\right\}.$$

avoid misconvergence in this context and their exact magnitudes are not needed. The proposed ext-NMCFLMS algorithm is given as shown in Algorithms 6.3 and 6.4.

6.6 Simulation Results

We now present simulation results to compare and evaluate the performance of the proposed ext-NMCFLMS algorithm for acoustic BSI against the NMCFLMS algorithms presented in [17, 22].

Algorithm 6.4 TDOA and Flattening point estimation for ext-NMCFLMS

TDOA estimation using GCC

$$\underline{\chi}_m(b) = \mathbf{F}_{2L}\boldsymbol{\chi}_m(b),$$

$$\hat{S}_{ml}(k) = \underline{\chi}_m(k)\underline{\chi}_l^*(k),$$

$$\Phi_m(k) = \frac{1}{|\hat{S}_{ml}(k)|},$$

$$\hat{\Psi}_m(n) = \sum_{k=0}^{2L-1} \Phi_m(k)\hat{S}_{ml}(k)e^{j2\pi nk/L},$$

$$\hat{\tau}_m = \arg\max_n \hat{\Psi}_m(n),$$

$$\mathbf{u}_m = \left[\mathbf{0}_{1\times\hat{\tau}_m-1} \quad 1 \quad \mathbf{0}_{1\times L-\hat{\tau}_m}\right]^T.$$

Flattening point estimation (FPE)

$$\tilde{\mathcal{J}}(b) = \sum_{k=1}^{b}\left[10\log_{10}\mathcal{J}(k)\right],$$

$$\triangle\tilde{\mathcal{J}}(b) = \left|\tilde{\mathcal{J}}(b)-\tilde{\mathcal{J}}(b-1)\right|,$$

$$t_f \leftarrow \arg\triangle\tilde{\mathcal{J}}(b) \leq 0.05 \text{ dB},$$

$$\hat{\hat{h}}_{\mathrm{dp},m}(b) = \hat{h}_{\mathrm{dp},m}(t_f).$$

6.6.1 Experimental Setup

All simulations are performed using impulse responses generated from the method of images [4] with a sampling rate of $f_s = 8$ kHz. The dimensions of the room are $\ell = 5 \times 4 \times 3$ m while a reverberation time of $T_{60} = 640$ ms is used, and the length of each impulse response is given by $L = 512$. A linear microphone array containing $M = 5$ microphones with uniform separation $d = 0.08$ m is used. The first microphone is placed at $\mathbf{q}_{\mathrm{mic},1} = (2.34, 2, 1.6)$ m, while the source is positioned at a range of 1 m and a bearing of 85° with respect to the centroid of the microphone array. Figure 6.8 shows the plan view of the source and microphone array placement used when generating the impulse responses. As before, the performance of the algorithms are quantified by the NPM measure defined by (6.39) and (6.40).

6.6.2 Variation of Convergence rate on β

We show the variation of the rate of convergence on the penalty gain β for the proposed ext-NMCFLMS algorithm using a WGN input signal. As shown in Fig. 6.1, additive WGN signals, $v_m(n)$ for $m = 1, \ldots, 5$ are added to the received signals such that an SNR = 15 dB is obtained for the multichannel system. The step-size of the

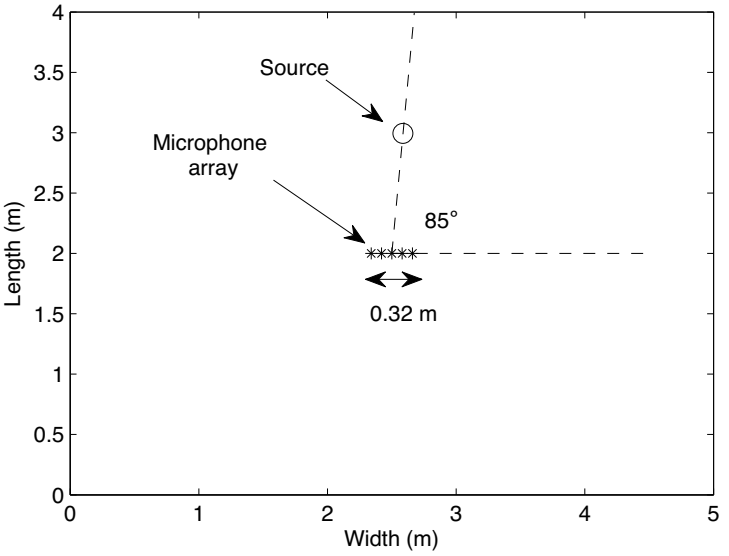


Fig. 6.8 Plan view of source and microphone array placement

ext-NMCFLMS algorithm is set to $\rho_e = 0.45$ and $\gamma = [1 - 1/(3L)]^L = 0.7165$ is used. It can be seen from Fig. 6.9 that the effect of misconvergence reduces with increasing β. In addition, it can be seen that the rate of convergence is reduced with increasing β.

Figure 6.10 shows additional results demonstrating that high β values are required at low SNRs in order to ensure the stability of the proposed ext-NMCFLMS algorithm. It can be seen from this result that for lower SNR a high value of β is required to achieve convergence. This implies that under low SNR conditions a stronger penalty term must be imposed on the cost function given by (6.53) in order to reduce the effect of additive noise in the system.

6.6.3 Degradation Due to Direct-path Estimation

We investigate the degradation in convergence performance for the ext-NMCFLMS algorithm due to direct-path estimation using the GCC and the FPE algorithms for the estimation of TDOA as well as the magnitude of the direct-path components as described in Sects. 6.5.3 and 6.5.4. Figure 6.11 compares the performance of the ext-NMCFLMS algorithm, implemented using the GCC and FPE algorithms, with that using the true direct-path components. As before, we used a WGN input signal and the step-size of the algorithm was set to $\rho_e = 0.45$ and the Lagrange multiplier

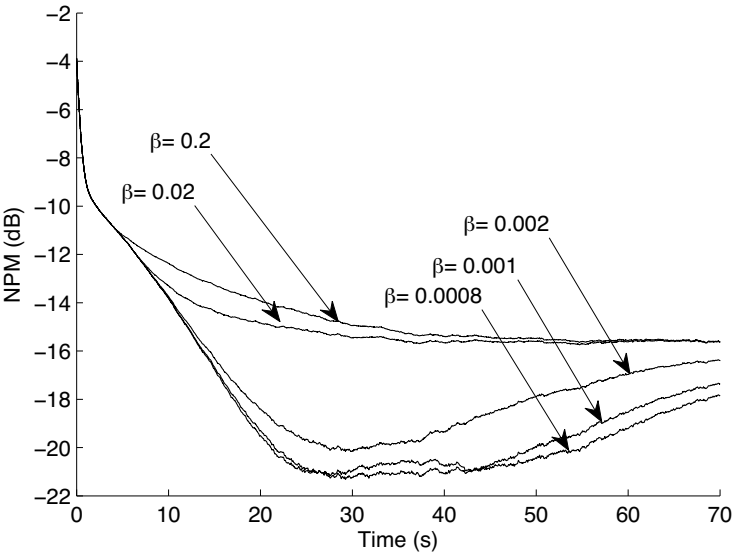


Fig. 6.9 Effect of Lagrange multiplier factor β on the convergence rate for the proposed ext-NMCFLMS algorithm

constant of $\beta = 0.02$ was used for both of these cases. The SNR for the system is 15 dB and $\gamma = [1 - 1/(3L)]^L = 0.7165$ was used for both algorithms studied.

It can be seen from this result that the ext-NMCFLMS algorithm using the true direct-path components achieves better convergence performance (both in terms of initial convergence and steady-state NPM) than that using the GCC and FPE algorithms. This degradation in convergence performance for the ext-NMCFLMS can be attributed to the inaccuracies in TDOA and magnitude estimation of the direct-path components. Inaccuracies for the TDOA estimation are due to reverberation of the room [5], while inaccuracies for the FPE algorithm arise from the estimation of the convergence point of the cumulative cost function t_f. As discussed in Sect. 6.5, it should be noted that for practical systems the direct-path components of acoustic impulse responses are unknown and, hence, some degradation in convergence performance is expected. As can be seen from Fig. 6.11, these joint inaccuracies contribute to a modest degradation in convergence performance of less than 1 dB in terms of steady-state NPM.

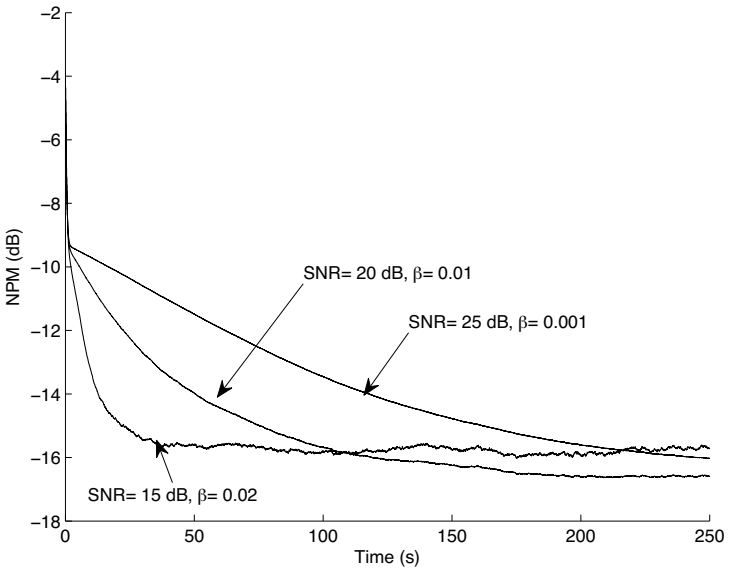


Fig. 6.10 Relationship between β and SNR of the system for the proposed ext-NMCFLMS algorithm

6.6.4 Comparison of Algorithm Performance Using a WGN Input Signal

We now compare the performance of the proposed ext-NMCFLMS algorithm with that of the NMCFLMS algorithm [22] using a WGN input signal. We have also included the performance of the algorithm proposed in [17], where the direct-path components of the impulse responses are assumed to be known *a priori* and are substituted at each time iteration into the update equation of (6.30). We denote this algorithm [17] as NMCFLMS$_{\text{dp}}$. As before, WGN is added to the received signals in order to achieve an SNR of 15 dB and we used $\beta = 0.02$ for the proposed ext-NMCFLMS algorithm. The step-sizes for all the algorithms are adjusted such that they reach the same asymptotic NPM. These correspond to $\rho = 1$, $\rho_{\text{dp}} = 0.45$ and $\rho_{\text{e}} = 0.45$ for the NMCFLMS, NMCFLMS$_{\text{dp}}$ and ext-NMCFLMS, respectively. In addition, we used $\gamma = [1 - 1/(3L)]^L = 0.7165$ for all the algorithms.

It can be seen from Fig. 6.12 that the NMCFLMS algorithm misconverges after achieving an NPM of approximately -17 dB. The proposed ext-NMCFLMS algorithm exhibits a higher rate of convergence compared to that of the NMCFLMS$_{\text{dp}}$ algorithm [17]. The higher rate of convergence for ext-NMCFLMS can be attributed to it taking the effect of additive noise into account while solving the minimization problem. This is equivalent to the ext-NMCFLMS algorithm finding a solution within the subspace determined by the constraint. During convergence, the ext-

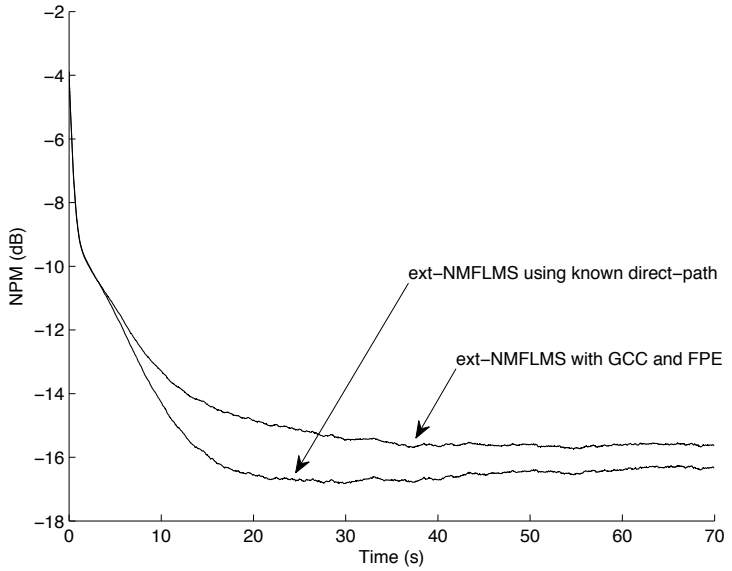


Fig. 6.11 Degradation of the proposed ext-NMCFLMS algorithm due to direct-path estimation

NMCFLMS algorithm achieves approximately 3 dB improvement in NPM over the $\text{NMCFLMS}_{\text{dp}}$ algorithm.

6.6.5 Comparison of Algorithm Performance Using Speech Input Signals

We compare the performance of the NMCFLMS, $\text{NMCFLMS}_{\text{dp}}$ and the ext-NMCFLMS algorithm as shown in Fig. 6.13 using speech input from a male talker. As before, the SNR of the multichannel acoustic system was 15 dB while the step-sizes are 0.1, 0.45 and 0.45 for the NMCFLMS, $\text{NMCFLMS}_{\text{dp}}$ and ext-NMCFLMS algorithms, respectively. These step-sizes have been adjusted such that all algorithms achieve the same asymptotic NPM. Similar to the explanation in Sect. 6.6.4, true delays and magnitudes for the direct-paths have been employed for the $\text{NMCFLMS}_{\text{dp}}$ algorithm. For the ext-NMCFLMS algorithm, we have employed the GCC with PHAT pre-filter of the Hilbert envelope of the linear prediction residual of speech to estimate the TDOA of the direct-path components [45]. The Lagrange multiplier value of $\beta = 1$ is used for the ext-NMCFLMS algorithm in this speech input example.

It can be seen from the result that after initial convergence, the NMCFLMS algorithm misconverges, while both the $\text{NMCFLMS}_{\text{dp}}$ and ext-NMCFLMS algorithms

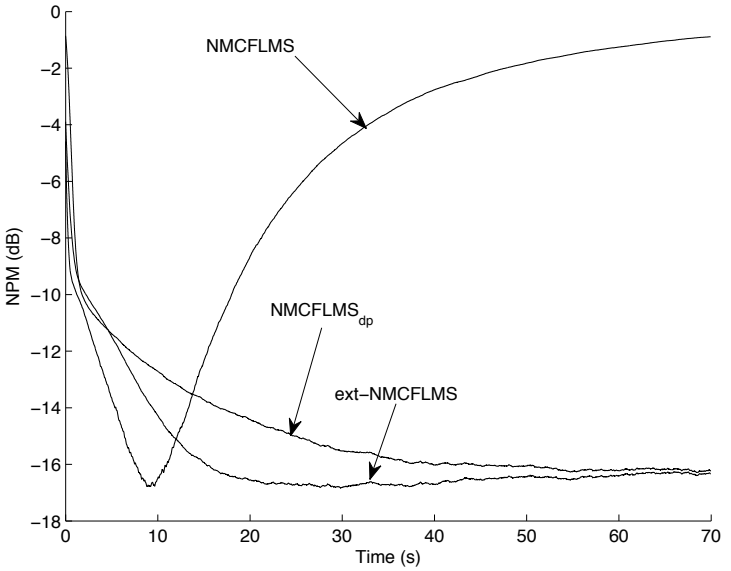


Fig. 6.12 Comparison of the convergence between the NMCFLMS, the NMCFLS$_{dp}$ and the proposed ext-NMCFLMS algorithms for a WGN input

converge to their steady-state. As before, the ext-NMCFLMS algorithm achieves a modest improvement in initial convergence over the NMCFLMS$_{dp}$ algorithm since the former takes the additive noise into account when minimizing the constraint cost function as explained in Sect. 6.5. Other tests have indicated that the ext-NMCFLMS algorithm achieves a higher rate of convergence which is robust to the presence of noise without misconvergence as compared to the NMCFLMS and NMCFLMS$_{dp}$ algorithms.

6.7 Conclusions

Estimation of acoustic impulse responses using adaptive algorithms employing the CR have been discussed and reviewed in this chapter. For dereverberation, these estimated channels can be further used to design equalization filters in order to remove reverberation introduced by the acoustic channels. We proposed the ext-NMCFLMS algorithm for the blind identification of acoustic impulse responses. The proposed algorithm achieves fast convergence compared to that proposed in [17] by imposing a limiting constraint on the cost function of the minimization problem. The result of this constrained minimization problem is the modified cost function term, which penalizes the cost due to noise. The significance of this correction term is

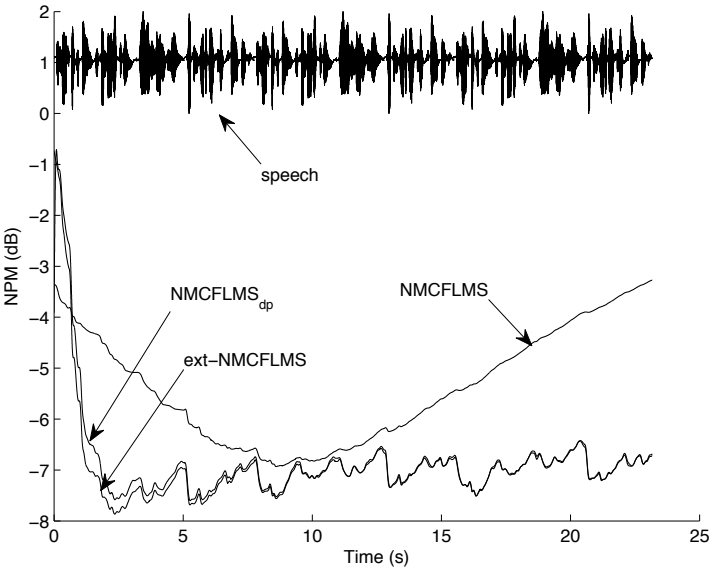


Fig. 6.13 Comparison of the convergence between the NMCFLMS, the NMCFLS$_{dp}$ and the proposed ext-NMCFLMS algorithms for speech input

then controlled by a multiplier control factor that increases with reducing SNR. In order to avoid the misconvergence problem that exists in NMCFLMS, the proposed algorithm estimates the direct-path components of the impulse responses. This is achieved using both the GCC and FPE algorithms for the delays and magnitudes, respectively. The overall contribution of employing the constraint optimization coupled with the estimation of the direct-path components enables the ext-NMCFLMS algorithm to achieve fast initial convergence and robustness to noise. Results presented using white Gaussian noise and speech signals showed the overall improvement in convergence performance for the ext-NMCFLMS over existing adaptive approaches for acoustic BSI. In addition the ext-NMCFLMS algorithm shows only a modest degradation of 1 dB NPM in convergence performance due to inaccuracies in TDOA and magnitude estimation using the GCC and FPE algorithms.

References

1. Abed-Meraim, K., Qiu, W., Hua, Y.: Blind system identification. Proc. IEEE **85**(8), 1310–1322 (1997)
2. Ahmad, R., Khong, A.W.H., Hasan, M.K., Naylor, P.A.: The extended normalized multichannel FLMS algorithm for blind channel identification. In: Proc. European Signal Processing Conf. (EUSIPCO) (2006)

3. Ahmad, R., Khong, A.W.H., Naylor, P.A.: Proportionate frequency domain adaptive algorithms for blind channel identification. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 5 (2006)
4. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
5. Benesty, J., Chen, J., Huang, Y.: Time-delay estimation *via* linear interpolation and cross correlation. IEEE Trans. Speech Audio Process. **12**(5), 509–519 (2004)
6. Benesty, J., Gänsler, T., Morgan, D.R., Sondhi, M.M., Gay, S.L.: Advances in network and acoustic echo cancellation. Springer (2001)
7. Benesty, J., Gänsler, T., Morgan, D.R., Sondhi, M.M., Gay, S.L.: General derivation of frequency-domain adaptive filtering. In: Advances in network and acoustic echo cancellation, chap. 8, pp. 157–176. Springer (2001)
8. Benesty, J., Gay, S.L.: An improved PNLMS algorithm. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 1881–1884 (2002)
9. Brandstein, M.S., Silverman, H.F.: A robust method for speech signal time-delay estimation in reverberant rooms. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 375–378 (1997)
10. Cadzow, J.A.: Blind deconvolution via cumulant extrema. IEEE Signal Process. Mag. **13**(6), 24–42 (1996)
11. Carter, G., Nuttall, A., Cable, P.: The smoothed coherence transform. Proc. IEEE **61**(10), 1497–1498 (1973)
12. Elko, G.W., Diethorn, E., Gänsler, T.: Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 67–70 (2003)
13. Ferrara, E.R.: Fast implementations of LMS adaptive filters. IEEE Trans. Acoust., Speech, Signal Process. **28**(4), 474–475 (1980)
14. Gannot, S., Moonen, M.: Subspace methods for multi-microphone speech dereverberation. EURASIP J. on App. Signal Process. **2003**(11), 1074–1090 (2003)
15. Golub, G.H., van Loan, C.F.: Matrix computations. Johns Hopkins Univ. Press (1990)
16. Hannan, E.J., Thomson, P.J.: Estimating group delay. Biometrika **60**(2), 241–253 (1973)
17. Hasan, M.K., Benesty, J., Naylor, P.A., Ward, D.B.: Improving robustness of blind adaptive multichannel identification algorithms using constraints. In: Proc. European Signal Processing Conf. (EUSIPCO) (2005)
18. Hasan, M.K., Naylor, P.A.: Analyzing effect of noise on LMS-type approaches to blind estimation of SIMO channels: robustness issue. In: Proc. European Signal Processing Conf. (EUSIPCO) (2006)
19. Hassab, J., Boucher, R.: Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal. IEEE Trans. Acoust., Speech, Signal Process. **29**(3), 549–555 (1981)
20. Haykin, S.: Adaptive filter theory, 4th edn. Information and System Science. Prentice Hall (2002)
21. Huang, Y., Benesty, J.: Adaptive multi-channel least mean square and newton algorithms for blind channel identification. Signal Processing **82**(8), 1127–1138 (2002)
22. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. **51**(1), 11–24 (2003)
23. Huang, Y., Benesty, J., Chen, J.: Optimal step size of the adaptive multichannel lms algorithm for blind SIMO identification. IEEE Signal Process. Lett. **12**(3), 173–176 (2005)
24. Inouye, Y., Hirano, K.: Cumulant-based blind identification of linear multi-input-multi-output systems driven by colored inputs. IEEE Trans. Acoust., Speech, Signal Process. **45**(6), 1543–1552 (1997)
25. Khong, A.W.H., Lin, X.S., Naylor, P.A.: Algorithms for identifying clusters of near-common zeros in multichannel blind system identification and equalization. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 389–392 (2008)

26. Kimura, T., Sasaki, H., Ochi, H.: Blind channel identification using RLS method based on second-order statistics. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 1785 – 1788 (1999)
27. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. **24**(4), 320–327 (1976)
28. Kuttruff, H.: Room acoustics, 4th edn. Taylor and Francis (2000)
29. Lee, O., Son, Y., Kim, K.: Underwater digital communication using acoustic channel estimation. In: Proc. Oceans, vol. 4, pp. 2453–2456 (2002)
30. Luo, H., Li, Y.: The application of blind channel identification techniques to prestack seismic deconvolution. Proc. IEEE **86**(10), 2082–2089 (1998)
31. Moon, T., Stirling, W.C.: Theory of constrained optimization. In: Mathematical methods and algorithms for signal processing, chap. 18, pp. 751–786. Prentice Hall (2000)
32. Morgan, D.R., Benesty, J., Sondhi, M.M.: On the evaluation of estimated impulse responses. IEEE Signal Process. Lett. **5**(7), 174–176 (1998)
33. Moulines, E., Duhamel, P., Cardoso, J.F., Mayrargue, S.: Subspace methods for the blind identification of multichannel FIR filters. IEEE Trans. Signal Process. **43**(2), 516–525 (1995)
34. Ochi, H., Oshiro, M.: Poly-phase based blind deconvolution technique using second-order statistics. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1841–1844 (1998)
35. Perez-Iglesias, H.J., Dapena, A., Castedo, L., Zarzoso, V.: Blind channel identification for Alamoutis coding systems based on eigenvector decomposition. In: Proc. European Wireless (2007)
36. Roth, P.R.: Effective measurements using digital signal analysis. IEEE Spectr. **8**, 62–70 (1971)
37. Shynk, J.J.: Frequency-domain and multirate adaptive filtering. IEEE Signal Process. Mag. **9**(1), 14–37 (1992)
38. Soo, J.S., Pang, K.K.: Multidelay block frequency domain adaptive filter. IEEE Trans. Acoust., Speech, Signal Process. **38**(2), 373–376 (1990)
39. Tong, L., Perreau, S.: Multichannel blind identification: from subspace to maximum likelihood methods. Proc. IEEE **86**(10), 1951–1968 (1998)
40. Tong, L., Xu, G., Kailath, T.: A new approach to blind identification and equalization of multipath channels. In: Proc. Asilomar Conf. on Signals, Systems and Computers, vol. 2, pp. 856–860 (1991)
41. Tsatsanis, M.K., Giannakis, G.B.: Subspace methods for blind estimation of time-varying FIR channels. IEEE Trans. Acoust., Speech, Signal Process. **45**(12), 3084–3093 (1997)
42. Tugnait, J.K.: A multidelay whitening approach to blind identification and equalization of SIMO channels. IEEE Trans. Wireless Commun. **1**(3), 456–467 (2002)
43. Villavicencio, J., Marquez, L.: Experimental comparison of correlation methods for time-delay estimation. In: Proc. IEEE Int. Conf. Electrical and Electronics Engineering, pp. 433–438 (2004)
44. Xu, G., Liu, H., Tong, L., Kailath, T.: A least-squares approach to blind channel identification. IEEE Trans. Signal Process. **43**(12), 2982–2993 (1995)
45. Yegnanarayana, B., Prasanna, S., Duraiswami, R., Zotkin, D.: Processing of reverberant speech for time-delay estimation. IEEE Trans. Speech Audio Process. **13**(6), 1110–1118 (2005)

Chapter 7
Subband Inversion of Multichannel Acoustic Systems

Nikolay D. Gaubitch and Patrick A. Naylor

Abstract Equalization of Acoustic Transfer Functions (ATFs) is an important topic with several applications in acoustic signal processing including speech dereverberation. ATFs are often modelled as finite impulse response filters with orders of thousands of taps and non-minimum phase characteristics. In practice, only approximate estimates of the actual ATFs are available due to measurement noise, limited estimation accuracy and temporal variation of the source-receiver geometry. These issues make equalization a difficult problem. In this chapter, we discuss multichannel equalization with focus on inexact ATF estimates. We present a multichannel method for the equalization filter design utilizing decimated and oversampled subbands, where the full-band acoustic impulse response is decomposed into equivalent subband filters prior to equalization. This technique is not only more computationally efficient but also more robust to impulse response inaccuracies compared with the full-band counterpart. Simulation results using simulated and measured ATFs are presented and the application of the subband method to speech dereverberation is demonstrated and evaluated.

7.1 Introduction

Equalization of Acoustic Transfer Functions (ATFs) is an important research topic with several applications in acoustic signal processing, including speech dereverberation [21] and sound reproduction [24]. Although, in theory, exact equalization is possible when multiple observations are available [18], there are many obstacles for practical application of ATF equalization algorithms. Equalization filters can be calculated either by direct estimation from the observed signals at the microphones or from measured or estimated acoustic impulse responses. The work presented here concentrates on the latter and it is assumed that an estimate of the acoustic impulse

Imperial College London, UK

response is available; obtained from, for example, blind system identification (see Chaps. 5 and 6).

Consider the L-tap acoustic impulse response of the acoustic path between a source and the m^{th} microphone in an M-element microphone array,

$$\mathbf{h}_m = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T,$$

with a z-transform $H_m(z)$ constituting the ATF. The objective of equalization is to apply an inverse system with transfer function $G_m(z)$ such that

$$H_m(z)G_m(z) = \kappa z^{-\tau}, \quad m = 1,2,\ldots,M, \tag{7.1}$$

where τ and κ are arbitrary delay and scale factors, respectively. Equivalently, considering the L_{i}-tap impulse response of $G_m(z)$,

$$\mathbf{g}_m = [g_{m,0}\ g_{m,1}\ \cdots\ g_{m,L_{\text{i}}-1}]^T,$$

(7.1) can be written in the time domain as

$$\mathbf{H}_m\mathbf{g}_m = \mathbf{d}, \tag{7.2}$$

where

$$\mathbf{H}_m = \text{diag}\{\mathbf{h}_m\ \cdots\ \mathbf{h}_m\} = \begin{bmatrix} h_{m,0} & 0 & \ldots & 0 \\ h_{m,1} & h_{m,0} & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h_{m,L-1} & \ldots & \vdots & 0 \\ 0 & h_{m,L-1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & h_{m,L-1} \end{bmatrix}$$

is a $(L+L_{\text{i}}-1)\times L_{\text{i}}$ convolution matrix, and

$$\mathbf{d} = [\underbrace{0\ \ldots\ 0}_{\tau}\ \kappa\ 0\ \ldots\ 0]^T$$

is the $(L+L_{\text{i}}-1)\times 1$ vector with the impulse response of the equalized ATF.

The problem of equalization is to find $G_m(z)$. When $H_m(z)$ is a minimum phase system, a stable inverse filter can be found by replacing the zeros of $H_m(z)$ with poles [26] such that

$$G_m(z) = \frac{1}{H_m(z)}. \tag{7.3}$$

However, ATF equalization is not that straightforward in practice because:

(i) ATFs are non-minimum phase in general [23] and so (7.3) does not give a stable causal solution for $G_m(z)$.
(ii) The average difference between maxima and minima in ATFs are in excess of 10 dB [15, 28, 30] and therefore ATFs typically contain spectral nulls that, after equalization, give strong peaks in the spectrum causing narrow band noise amplification.
(iii) Equalization filters designed from inaccurate estimates of $H_m(z)$ will cause distortion in the equalized signal [28].
(iv) The length L of $\mathbf{h}_m$ at a sampling frequency f_s is related to the reverberation time, T_{60}, in a room by $L = f_s T_{60}$ and can be several thousand taps in length [15].

Several alternative approaches, both for single and for multiple microphones, have been proposed to address these issues. There are two common methods for single channel equalization: Single Channel Least Squares (SCLS) and homomorphic equalization [19]. SCLS equalization filters are designed by minimizing an error formed from (7.2) as [19, 20]

$$\hat{\mathbf{g}}_m = \arg\min_{\mathbf{g}_m} \|\mathbf{H}_m \mathbf{g}_m - \mathbf{d}\|_2^2, \tag{7.4}$$

where $\|\cdot\|_2$ denotes Euclidean distance. The m^{th} channel least squares optimal equalization filter is then calculated as

$$\hat{\mathbf{g}}_m = \mathbf{R}_m^{-1} \mathbf{r}_m, \tag{7.5}$$

where

$$\mathbf{R}_m = \mathbf{H}_m^T \mathbf{H}_m \tag{7.6}$$

is the autocorrelation matrix of the acoustic impulse response and

$$\mathbf{r}_m = \mathbf{H}_m^T \mathbf{d} \tag{7.7}$$

is the cross-correlation between the acoustic impulse response and the desired impulse response of the equalized ATF.

In homomorphic inverse filtering [19, 23, 27, 32], the ATF is decomposed into minimum phase and all-pass components. An exact inverse can be found for the minimum phase component with (7.3), while the all-pass component can be equalized, for example, using a matched filter [27]. Equalizing only the magnitude was considered in [23, 27], but was found to result in audible residual echoes. In a comparative study between these two techniques, Mourjopoulos [19] concluded that SCLS, although sometimes less accurate than homomorphic inversion, is more efficient in practice.

Single channel methods typically result in large processing delay, which is problematic for many communications applications, extremely long and non-causal inverse filters, and provide only approximate equalization [18]. On the positive side,

single channel equalization filters are less sensitive to noise and inexact ATF estimates; this is due to the approximate nature of these filters [21]. Inherently, SCLS inverse filters only partially equalize deep spectral nulls, which can be advantageous in avoiding problems due to points (ii) and (iii) above.

In the multichannel case, the non-minimum phase problem is eliminated and exact equalization can be achieved using Bezout's theorem [14, 18]: given a set of M ATFs, $H_m(z)$, and assuming that these do not have any common zeros, a set of filters, $G_m(z)$, can be found such that [14, 18]

$$\sum_{m=1}^{M} H_m(z)G_m(z) = 1. \tag{7.8}$$

The Multiple-input/output INverse Theorem (MINT) [18] is a well-known multichannel equalization method based on (7.8). Adaptive versions have also been considered [24]. Unlike single channel equalization filters, the length of the multichannel equalization filters is of similar order to the length of the acoustic impulse responses and there is no processing delay [14, 18]. However, it has been observed that exact equalization is of limited value in practice, when the ATF estimates contain even moderate errors [21, 28].

Various alternatives have been proposed for improving robustness to ATF inaccuracies. Bharitkar *et al.* [2] use spatially averaged ATFs for the design of the equalization filter. In [12], the authors modify the desired signal, $\mathbf{d}$, in the Multichannel Equalizer (MCEQ) inverse filter design, such that the late reverberation is equalized while the early reflections are preserved. Haneda *et al.* [5, 6] form an infinite impulse response filter by decomposing the ATFs into common acoustical poles and non-common zeros. Mourjopoulos [20] uses an autoregressive (AR) model of the acoustic transfer functions rather than the all-zero model in order to reduce the filter order. The AR model of acoustic transfer functions is also exploited by Hopgood and Rayner in a single channel subband equalization approach [13]. Hikichi *et al.* [8, 9] introduce regularized multichannel equalization which adds robustness to noise and ATF fluctuations. Other recent developments in robust equalization are also discussed by Miyoshi *et al.* in Chap. 9.

In this chapter, we introduce a different method for equalization filter design based on multichannel equalization. Given a set of multichannel ATF estimates, we decompose the ATFs into their subband equivalent filters and use these to design a set of subband inverse filters. The equalization is performed in each subband before a full-band equalized signal is reconstructed. It is shown that this approach not only reduces the computational load but also reduces the sensitivity to estimation errors and the effect of measurement noise in the ATFs. An important result is that this method accommodates multichannel equalization of large order systems, taking advantage of the shorter length of multichannel equalization filters and with a low sensitivity to ATF inaccuracies similarly to single channel methods.

The remainder of the chapter is organized as follows: multichannel equalization is described in Sect. 7.2. The effects on equalization filter design for single and for multichannel scenarios using inexact ATFs are demonstrated in Sect. 7.3. The sub-

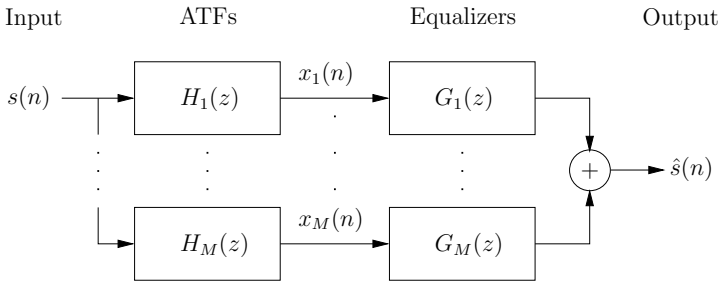


Fig. 7.1 Full-band multichannel equalization system

band equalization method is developed in Sect. 7.4. A comparative computational complexity analysis of the full-band and the subband equalizer designs is presented in Sect. 7.5. The application of the subband equalization method to speech dereverberation is discussed in Sect. 7.6 and simulation results demonstrating various aspects of the proposed algorithm are given in Sect. 7.7. Finally, conclusions are drawn in Sect. 7.8.

7.2 Multichannel Equalization

The relation in (7.8) can be written in the time domain as

$$\begin{aligned}\mathbf{d} &= \sum_{m=1}^{M} \mathbf{H}_m \mathbf{g}_m \\ &= \mathbf{H}\mathbf{g},\end{aligned} \tag{7.9}$$

where

$$\mathbf{H} = [\mathbf{H}_1 \ \mathbf{H}_2 \ \ldots \ \mathbf{H}_M]$$

and

$$\mathbf{g} = [\mathbf{g}_1^T \ \mathbf{g}_2^T \ \ldots \ \mathbf{g}_M^T]^T.$$

An optimization problem can then be formulated as

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}} \|\mathbf{H}\mathbf{g} - \mathbf{d}\|^2. \tag{7.10}$$

An estimate of the multichannel equalization filters can be calculated by solving (7.10) resulting in [14]

$$\hat{\mathbf{g}} = \mathbf{H}^{+}\mathbf{d}, \tag{7.11}$$

where $\mathbf{H}^{+}$ is the matrix pseudo-inverse [4].

The choice of equalization filter length, L_i and, consequently, the dimensions of $\mathbf{H}$, $(L+L_\mathrm{i}-1)\times ML_\mathrm{i}$, define the solution obtained with (7.11). If $L+L_\mathrm{i}-1 \leq ML_\mathrm{i}$, then

$$L_\mathrm{i} \geq \frac{L-1}{M-1}, \quad M \geq 2, \tag{7.12}$$

and the system is underdetermined such that several exact solutions exist [11]. Then, the pseudo-inverse in (7.11) is defined as

$$\mathbf{H}^{+} = \mathbf{H}^{T}(\mathbf{H}\mathbf{H}^{T})^{-1} \tag{7.13}$$

and gives the minimum norm solution to (7.10). In the special case when (7.12) results in an equivalence, the matrix $\mathbf{H}$ becomes square and the pseudo-inverse in (7.11) reduces to a standard matrix inverse. The exact solution is then unique and equivalent to that of MINT [18] . However, as pointed out in [14], it is not always possible to choose such a length for $M > 2$, since the relation in (7.12) may not give an integer result. Instead, a greater length is often chosen [7, 10, 14]. A third case arises when L_i is chosen such that $(L+L_\mathrm{i}-1) > ML_\mathrm{i}$, which results in an overdetermined system of equations and only a least squares solution can be obtained [11]. For this work, we consider the former, minimum norm exact solutions, and set the equalization filter length to

$$L_\mathrm{i} = \left\lceil \frac{L-1}{M-1} \right\rceil, \quad M \geq 2, \tag{7.14}$$

where $\lceil a \rceil$ denotes the ceiling operator giving the smallest integer greater than or equal to a. The relation between an input signal $s(n)$, ATFs $H_m(z)$, equalizers $G_m(z)$, and an output signal $\hat{s}(n)$ is depicted in Fig. 7.1 where, following the general equalization formulation in (7.1), $\hat{s}(n) = \kappa s(n-\tau)$ for ideal equalization.

7.3 Equalization with Inexact Impulse Responses

In this section the effects of equalization filter design are demonstrated for the case when using inexact $\mathbf{h}_m$, considering both single channel (approximate) equalization with SCLS and multichannel (exact) equalization with MCEQ. An inexact system impulse response,

$$\tilde{\mathbf{h}}_m = [\tilde{h}_{m,0}\ \tilde{h}_{m,1}\ \ldots\ \tilde{h}_{m,L-1}]^{T},$$

is defined here as an impulse response with system mismatch $\mathcal{M}_m > -\infty$ dB, with

$$\mathcal{M}_m = 20\log_{10}\left(\frac{\|\mathbf{h}_m - \tilde{\mathbf{h}}_m\|_2}{\|\mathbf{h}_m\|_2}\right) \text{ dB}. \tag{7.15}$$

Such inexact impulse responses occur in practical measurements and system identification due to noise, changes in the relative source-microphone configuration and

estimation error. The formulation in (7.15) does not consider scalar ambiguities that may result from some blind system identification methods since this does not affect the performance of the equalization methods described herein. In the remainder of this work system mismatch is modelled as in [3] according to

$$\tilde{\mathbf{h}}_m = (\mathbf{I} + \mathcal{E}_m)\mathbf{h}_m, \tag{7.16}$$

where

$$\boldsymbol{\mathcal{E}}_m = \mathrm{diag}\{\varepsilon_{m,0}\ \varepsilon_{m,1}\ \ldots\ \varepsilon_{m,L-1}\},$$

$\mathbf{I}$ is the identity matrix and $\varepsilon_{m,i}$ is a zero mean Gaussian distributed variable with the variance set to the desired system mismatch,

$$\mathcal{M}_m = 10\log_{10}(\mathrm{var}(\varepsilon_{m,i}))\ \mathrm{dB}.$$

The following studies the design of an equalization filter for $\mathbf{h}_m$ using $\tilde{\mathbf{h}}_m$ when $\mathcal{M}_m > -\infty$ dB. Furthermore, the equalized system is defined as

$$\hat{\mathbf{d}} = \mathbf{H}\hat{\mathbf{g}}, \tag{7.17}$$

with an I-point discrete Fourier transform

$$\begin{aligned}\hat{D}(k) &= \sum_{n=0}^{I-1} \hat{d}(n) e^{-j\frac{2\pi}{N}kn}, \quad k = 0,1,\ldots,I-1 \\ &= |\hat{D}(k)| e^{j\theta(k)}.\end{aligned} \tag{7.18}$$

For evaluation purposes the magnitude and the phase are considered separately as follows:

1. *Magnitude deviation* is defined here as the standard deviation of the equalized magnitude response [28]

$$\sigma = \sqrt{\frac{1}{I}\sum_{k=0}^{I-1}\left(10\log_{10}|\hat{D}(k)| - \bar{D}\right)^2}, \tag{7.19}$$

   with

$$\bar{D} = \frac{1}{I}\sum_{k=0}^{I-1} 10\log_{10}|\hat{D}(k)|.$$

   This measure is scale independent and equals zero for exact equalization.
2. *Linear phase deviation* is defined as the deviation of the unwrapped phase from a linear fit to its values and is defined here as

$$\Delta = \sqrt{\frac{1}{I}\sum_{k=0}^{I-1}\left(\theta(k) - \bar{\theta}(k)\right)^2}, \tag{7.20}$$

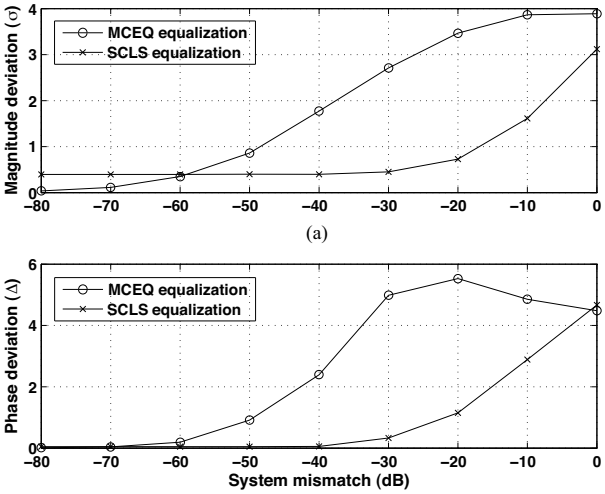


Fig. 7.2 (a) Magnitude deviation and (b) phase deviation *vs.* system mismatch for exact equalization with MCEQ from (7.11) (*cirlcles*) and approximate equalization with SCLS from (7.5) (*crosses*)

where $\bar{\theta}(k)$ is the least squares linear approximation to the phase at frequency bin k.

Two key effects regarding equalization filter design from inexact impulse responses are to be demonstrated: the performance degradation caused by increased system mismatch and the performance degradation caused by increased system length L for a fixed system mismatch.

7.3.1 Effects of System Mismatch

An illustrative comparison experiment was performed using an arbitrary system with $M = 2$ channels of length $L = 64$ taps. The taps of the impulse responses, $\mathbf{h}_m,\ m = 1,2$, were generated using random sequences drawn from a zero mean, unit variance Gaussian distribution. System mismatch ranging from 0 to -80 dB was modelled using (7.16). For each case, the impulse response was equalized using the MCEQ method in (7.11) with $L_\mathrm{i} = L-1$, $\tau = 0$ and with the SCLS method in (7.5) with $L_\mathrm{i} = 15L$, $\tau = L/2$. The results, averaged over 100 different channel realizations, are displayed in Fig. 7.2. It is seen that equalization using the MCEQ method introduces significant spectral deviation for $\mathcal{M}_m > -60$ dB, a level of system mismatch which is the capability of current (blind or non-blind) ATF estimation

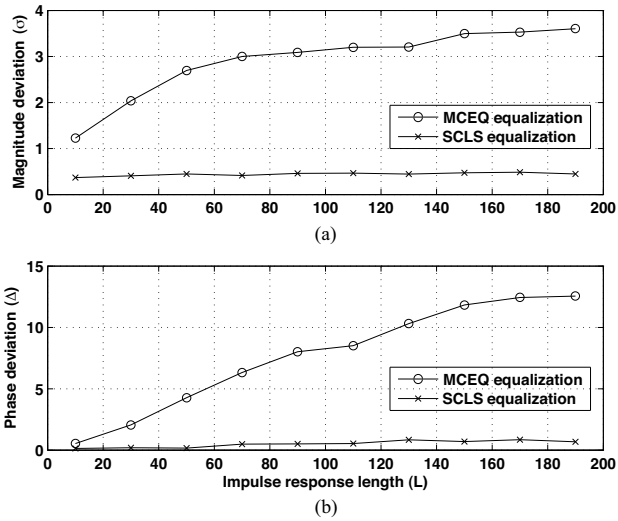


Fig. 7.3 (a) Magnitude deviation and (b) phase deviation *vs.* impulse response length for a system mismatch $\mathcal{M}_m = -30$ dB using exact equalization with MCEQ from (7.11) (*cirlcles*) and approximate equalization with SCLS from (7.5) (*crosses*)

techniques. In contrast, the single channel SCLS equalizer degrades much more gracefully, although equalization filters of very high orders are required. Furthermore, it is observed that for $\mathcal{M}_m < -70$ dB the multichannel method results in exact equalization while the single channel counterpart reaches a performance bound in the magnitude deviation. These observations are also in accordance with the results reported in [28, 31], where the authors studied equalization of ATFs measured at a different location to that at the point of processing.

7.3.2 Effects of System Length

We examine next the interrelation between system mismatch, impulse response length and equalization accuracy. We consider an arbitrary system with two randomly generated channels (as in Sect. 7.3.1) $\mathbf{h}_m,\ m = 1,2$ with length L varied in the range 10 to 190 taps and system mismatch $\mathcal{M}_m = -30$ dB. The lengths of the inverse filters were set to $L_\mathrm{i} = L - 1$ and $L_\mathrm{i} = 15L$ for the MCEQ and SCLS equalizers, respectively. Figure 7.3 shows the resulting magnitude and phase deviation for the different channel lengths as an average of 100 different random channel realizations. It can be seen that the exact equalization with MCEQ considerably decreases

in performance compared with the single channel SCLS, which appears more or less constant.

In summary, we have seen that exact multichannel equalization with inverse filters obtained from inexactly estimated systems gives worse results than approximate single channel equalization. However, SCLS inverse filter length of the order $15L$ is not suitable for realistic applications involving acoustic impulse responses and the achieved equalization is limited even when the system mismatch is low. In addition, the deteriorating effects of exact multichannel equalization, for a fixed system mismatch, were seen to increase with increased channel length. These observations lead us to the conclusion that when equalization filters are designed from inexact system estimates, approximate solutions are preferable and the system length should be kept as short as possible. This, consequently, motivates the development of a multichannel subband equalizer, where shorter system length and approximate equalization are inherent features.

7.4 Subband Multichannel Equalization

Now we derive the Subband Multichannel Equalizer (SB-MCEQ). A conceptual system diagram of the process considered in this derivation is depicted in Fig. 7.4; the SB-MCEQ uses a subband filtering model of the reverberation process and applies the multichannel equalizer depicted in Fig. 7.1 to each subband. Thus, there are three key factors to consider:

1. Choice of filter-bank structure
2. Determination of the subband equivalent filters of the full-band ATFs
3. Subband equalization filter design

Multirate processing [33] has been applied successfully in acoustic signal processing problems such as, for example, acoustic echo cancellation where considerable improvement in convergence rates has been demonstrated using subband adaptive filters [22, 29, 35, 36]. A subband version of MINT was first investigated in [37]. This approach uses a critically decimated filter-bank. The subband transfer functions to be equalized are estimated using a least squares estimate using the observation of a known reference signal. A different multichannel subband method was proposed by Wang and Itakura [34] for a critically decimated filter-bank. A single channel least squares equalizer is applied to each subband and each microphone signal and the full-band signal is reconstructed using the best microphone signal in each subband; the best microphone signal is selected for each subband using a normalized estimation error criterion from the estimation of the SCLS filters. Hopgood and Rayner [13] take a rigorous approach to subband equalization and study the relation between full-band and subband filters utilizing an autoregressive model of the acoustic impulse response. An adaptive method for multichannel equalization in oversampled subbands was proposed in [35] and was shown to provide significant improvement over the full-band counterpart.

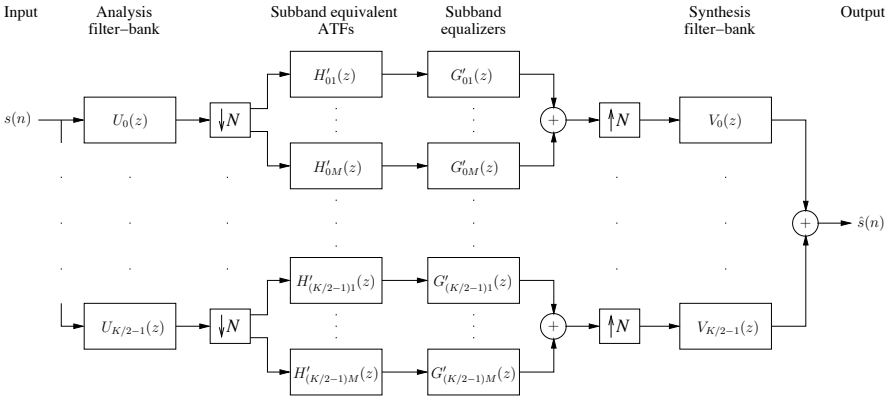


Fig. 7.4 Conceptual subband multichannel equalization system: both the ATFs and the corresponding inverse filters are applied to the decimated subband signals

The relation between full-band and subband filtering was studied, for example, by Lanciani *et al.* [16] for filtering of MPEG audio signals and by Reilly *et al.* [29] with applications to acoustic echo cancellation. The former authors derive the relations between the full-band and subband filters for critically decimated cosine modulated filter-banks [33], which is shown to require cross-subband filtering. On the other hand, Reilly *et al.* [29] show that good approximations can be obtained with a diagonal filtering matrix, involving only one filter per subband, for complex oversampled filter-banks because these suppress aliasing in adjacent subbands [35] sufficiently. We extend this approach to the multichannel case with application to ATF equalization. This method differs from the previously proposed methods in that it uses oversampled subbands in conjunction with explicit relations between the full-band and the subband ATFs.

7.4.1 Oversampled Filter-banks

The Generalized Discrete Fourier Transform (GDFT) filter-bank [36] is employed in the subsequent development work. The advantages of this filter-bank include straightforward implementation of fractional oversampling and computationally efficient implementations [36]. The oversampling is of importance because it facilitates the use of a single equalization filter per subband without the requirement for cross-filters. Within the framework of the GDFT filter-bank, the k^{th} subband analysis filters, $u_{k,i}$, are calculated from a single prototype filter, p_i, with bandwidth $\frac{2\pi}{K}$ according to the relation

$$u_{k,i} = p_i e^{j\frac{2\pi}{K}(k+k_0)(i+i_0)}, \tag{7.21}$$

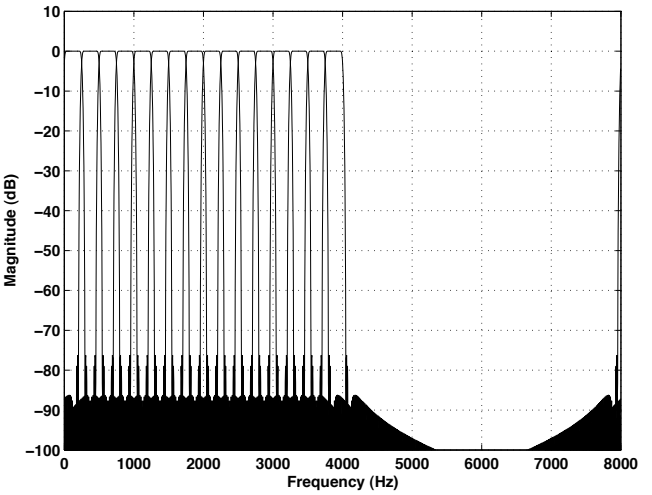


Fig. 7.5 Subband filters of length $L_{\text{pr}} = 512$ taps for $K = 32$ subbands, decimated by $N = 24$

where the properties of the frequency and time offset terms, k_0 and i_0, are discussed in, for example, [36]; we set these to $i_0 = 0$ and $k_0 = 1/2$ as in [29]. It has been shown [36] that a corresponding set of synthesis filters satisfying near perfect reconstruction can be obtained from the time-reversed, conjugated version of the analysis filters

$$v_{k,i} = u^*_{k,L_{\text{pr}}-i-1}, \tag{7.22}$$

where L_{pr} is the length of the prototype filter and, consequently, the length of all analysis and synthesis filters of the filter-bank. Although this filter design results in complex subband signals, for K even, only $K/2$ subbands need to be processed since the remaining subbands are straightforward complex conjugates of these.

The choice of decimation factor, N, and number of subbands, K, has several consequences. The use of a large number of subbands requires a long prototype filter to suppress aliasing efficiently. On the other hand, if too few subbands are used, the benefit of shorter subband filters is reduced. The choice of oversampling ratio (N/K) affects the performance of the equivalent subband filters as will be discussed in Sect. 7.4.2. The filter-bank used for the illustrative experiments in this chapter uses $K = 32$ subbands and decimation factor, $N = 24$. An $L_{\text{pr}} = 512$-tap prototype filter was designed using the iterative least squares method [36], giving an estimated aliasing suppression of 82 dB. The magnitude response of the analysis filters is shown in Fig. 7.5.

From the properties of the GDFT filter-bank outlined here, the following two properties can be assumed to be valid:

P1: Aliasing is sufficiently suppressed in the subbands

$$U_k(zW_N^i)V_k(z) \approx 0, \quad i > 0, \forall k, \tag{7.23}$$

where $W_N = e^{-j2\pi/N}$ and $U_k(zW_N^i)$ are the z-transform alias components arising from the decimation of the subband analysis filters, $u_{k,i}$.

P2: Magnitude distortion of the filter-bank is negligible

$$\sum_{k=0}^{K/2-1} U_k(z)V_k(z) \approx \kappa z^{-\tau}, \tag{7.24}$$

where $U_k(z)$ and $V_k(z)$ are the z-transforms of the subband analysis and synthesis filters from (7.21) and (7.22), respectively.

7.4.2 Subband Decomposition

Consider the K subband, M microphone system in Fig. 7.4. It is clear that in order to design the subband equalizers $G'_{km}(z)$, the subband ATFs $H'_{km}(z)$ must be found using, for example, complex subband decomposition [29]. The objective of the subband decomposition is to find a set of subband filters, $H'_{km}(z)$, $k = 0, 1, \ldots, K/2-1$, given the full-band filter $H_m(z)$, such that the total transfer function of the filter-bank, $F_m(z)$, is equivalent to the that of the full-band filter up to an arbitrary scale factor, κ, and an arbitrary delay, τ. This can be written as

$$F_m(z) = \kappa z^{-\tau} H_m(z), \quad \forall m. \tag{7.25}$$

The total transfer function of the filter-bank for the m^{th} channel is given by

$$F_m(z) = \frac{1}{N} \sum_{k=0}^{K/2-1} \sum_{i=0}^{N-1} U_k(zW_N^i)H'_{km}(z^N)V_k(z), \tag{7.26}$$

Evoking property P1 in (7.23), the aliasing components in (7.26) can be discarded, and the filter-bank transfer function reduces to

$$F_m(z) \approx \frac{1}{N} \sum_{k=0}^{K/2-1} U_k(z)H'_{km}(z^N)V_k(z). \tag{7.27}$$

In contrast to (7.26), the expression in (7.27) facilitates the use of a single filter $H'_{km}(z)$ in each subband.

Next, following the approach in [29], we choose the filters in each subband, $H'_{km}(z)$, such that they satisfy the relation

$$U_k(z)H'_{km}(z^N) = U_k(z)H_m(z), \quad \forall k. \tag{7.28}$$

Substituting (7.28) into (7.27) we obtain

$$F_m(z) \approx H_m(z)\frac{1}{N}\sum_{k=0}^{K/2-1} U_k(z)V_k(z). \tag{7.29}$$

Finally, due to property P2 in (7.24), we find that the overall filter-bank transfer function is

$$F_m(z) \approx \frac{\kappa}{N} z^{-\tau} H_m(z), \tag{7.30}$$

which is the desired result. Thus, the remaining problem is to solve for $H'_{km}(z)$ in (7.28).

Decimating (7.28) by a factor of N, the following approximation can be formed

$$\frac{1}{N}\sum_{i=0}^{N-1} U_k(z^{1/N}W_N^i)H'_{km}(z) \approx \frac{1}{N}\sum_{i=0}^{N-1} U_k(z^{1/N}W_N^i)H_m(z^{1/N}W_N^i), \tag{7.31}$$

which in the time domain is written, equivalently, as

$$\mathbf{U}_{N,k}\mathbf{h}'_{km} \approx \mathbf{r}_{N,km}, \tag{7.32}$$

where

$$\mathbf{h}'_{km} = [h'_{km,0}\ h'_{km,1}\ \ldots\ h'_{km,L'-1}]^T$$

is the L'-tap subband impulse response of the m^{th} microphone (note that L' is the same for all K subbands),

$$\mathbf{r}_{N,km} = [r_{km,0}\ r_{km,N}\ \ldots\ r_{km,N(L-1)}]^T$$

is a $\lceil (L+L_{\text{pr}}-1)/N \rceil \times 1$ vector with

$$r_{km,i} = h_{m,i} * u_{k,i}, \tag{7.33}$$

and

$$\mathbf{U}_{N,k} = \begin{bmatrix} u_{k,0} & 0 & \ldots & 0 \\ u_{k,N} & u_{k,0} & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u_{k,L_{\text{pr}}-1} & \ldots & \vdots & 0 \\ 0 & u_{k,L_{\text{pr}}-1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & u_{k,L_{\text{pr}}-1} \end{bmatrix},$$

where $*$ denotes convolution and L_{pr} is the length of the prototype filter (and hence of the analysis and synthesis filters). The convolution on the left-hand side of (7.32) is of length $\lceil L_{\text{pr}}/N \rceil + L' - 1$, and consequently, the length of the subband filters is

$$L' = \left\lceil \frac{L + L_{\text{pr}} - 1}{N} \right\rceil - \left\lceil \frac{L_{\text{pr}}}{N} \right\rceil + 1. \tag{7.34}$$

The estimates of the subband filters $\hat{\mathbf{h}}'_{km}$ are then found by solving the following optimization problem [29]

$$\hat{\mathbf{h}}'_{km} = \arg\min_{\mathbf{h}'_{km}} \|\mathbf{U}_{N,k}\mathbf{h}'_{km} - \mathbf{r}_{N,km}\|_2^2. \tag{7.35}$$

The k^{th} subband, m^{th} channel optimal (in the least squares sense) filters are calculated according to

$$\hat{\mathbf{h}}'_{km} = \mathbf{U}^{+}_{N,k}\mathbf{r}_{N,km}, \tag{7.36}$$

where $\mathbf{U}^{+}_{N,k}$ is the pseudo-inverse $\mathbf{U}^{+}_{N,k}$ of the matrix $\mathbf{U}_{N,k}$.

In summary, given a full-band ATF, $H_m(z)$, and $K/2$-band filter-bank satisfying perfect reconstruction and aliasing suppression in the subbands, a set of subband filters, $H'_{km}(z)$, of the order L/N, can be found such that the overall subband transfer function is equivalent to the full-band filter response. We now aim to exploit this significant order reduction in the subband filters of the very long full-band acoustic impulse responses to design the equalizing filters $G'_{km}(z)$.

7.4.3 Subband Multichannel Equalization

The multichannel equalization filters, $G'_{km}(z)$, can be calculated for each subband using the filters $\hat{H}'_{km}(z)$ obtained from (7.36). Here, this is done utilizing the multichannel equalization filter design from (7.11), which now becomes

$$\hat{\mathbf{g}}'_k = \hat{\mathbf{H}}'^{+}_k \mathbf{d}, \quad k = 0, 1, \ldots, \frac{K}{2} - 1, \tag{7.37}$$

such that for each subband

$$\sum_{m=1}^{M} \hat{G}'_{km}(z)\hat{H}'_{km}(z) = 1, \quad k = 1, 2, \ldots, \frac{K}{2} - 1, \tag{7.38}$$

where $\hat{G}'_{km}(z)$ is the minimum norm estimate of $G'_{km}(z)$. The equivalent subband ATFs are shorter than the full-band ATFs. Therefore, the length of the inverse filters, L'_{i}, is shortened and is given by

$$L'_{\text{i}} = \left\lceil \frac{L' - 1}{M - 1} \right\rceil, \quad M \geq 2, \tag{7.39}$$

where L' is defined in (7.34).

Equalization is achieved by applying the inverse filters, $\hat{\mathbf{g}}'_k$, to the subband signals of the reverberant observations in each subband k, $\forall k$ and an equalized full-band

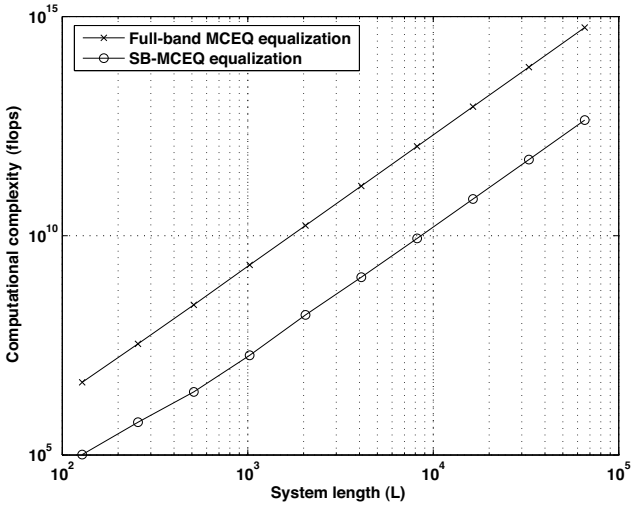


Fig. 7.6 Floating point operation count *vs.* system length for the full-band (*crosses*) and subband (*circles*) equalizers

signal is constructed. Assuming that exact equalization is achieved in each subband, the accuracy of the final result will depend on the reconstruction properties of the filter-bank, the accuracy of aliasing suppression and, *ergo*, on the design of the prototype filter. Consequently, the overall equalization of the subband method will not be exact in practice, which can be beneficial as discussed in Sect. 7.3. These dependencies will be explained through illustrative simulations in Sect. 7.7.

7.5 Computational Complexity

One of the several issues discussed in Sects. 7.1–7.3 was the large order of the acoustic impulse responses, which result in highly computationally intensive calculation of multichannel equalization filters. In some cases the computational complexity renders the MCEQ equalization infeasible [11]. A subband implementation, such as the SB-MCEQ, is expected to reduce this computational complexity.

In this section, we present a comparative analysis of the computations required for the solution of the full-band MCEQ equalizer design and the SB-MCEQ equalizer design (including the cost of the subband decomposition). The comparison is made in terms of floating point operations (flops), where one flop is defined as either one real multiplication or one real addition [4].

Consider the generic optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2, \tag{7.40}$$

which has a solution

$$\hat{\mathbf{x}} = \mathbf{A}^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{b}, \tag{7.41}$$

where $\mathbf{A}$ is an arbitrary real valued $p \times q$ matrix and $\mathbf{b}$ is a real valued $p \times 1$ vector. The number of flops required to solve this problem using the normal equations is given by [4]

$$pq^2 + \frac{q^3}{3}. \tag{7.42}$$

From the dimensions of the full-band equalization filter calculation in (7.11), the number of flops required for the MCEQ design is

$$(ML_\mathrm{i})^2(L + L_\mathrm{i} - 1) + \frac{(ML_\mathrm{i})^3}{3}. \tag{7.43}$$

The subband equalization filter design takes into consideration two separate calculations for each of the $K/2$ subbands: the cost of the subband inverse filter computation in (7.37) and the cost of the subband decomposition in (7.36). The data for these calculations is complex where, generally, one complex multiplication requires four real multiplications and two real additions and one complex addition requires two real additions. Under the assumption that an equal number of complex multiplications and complex additions are required to solve the system of equations considered here, we multiply the expression in (7.42) by a factor of four. The total flops required for the subband inverse filter design can be expressed as

$$2K\left((ML'_\mathrm{i})^2(L' + L'_\mathrm{i} - 1) + L_\mathbf{r}(L')^2 + \frac{(ML'_\mathrm{i})^3 + (L')^3}{3}\right), \tag{7.44}$$

where $L_\mathbf{r} = \lceil (L + L_\mathrm{pr} - 1)/N \rceil$. The key factor of the computational complexity is the system length and thus, the improvement achieved by the subband method will depend on the number of subbands and on the decimation ratio. An example is given in Fig. 7.6 where the computational complexity is calculated with (7.43) and (7.44) respectively. The subband implementation for this example is that presented in Section 7.4.1 with $K = 32$ subbands decimated by $N = 24$ and for $M = 8$ microphones. On average over all system lengths, the subband approach reduces the required flops by a factor of 106.

7.6 Application to Speech Dereverberation

Having derived the subband adaptive filters, we are now equipped with all the tools necessary to perform speech dereverberation in subbands. We consider a speech signal $s(n)$ produced in a reverberant room and observed by an M microphone array. The reverberant observation at the m^{th} microphone is written as

$$x_m(n) = \mathbf{h}_m^T\mathbf{s}(n) + v_m(n), \tag{7.45}$$

where

$$\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L+1)]^T,$$

and $v_m(n)$ is measurement noise. For the purpose of this work, we assume that $v_m(n) = 0$ in order to study explicitly the effects of inaccuracies in the ATF estimates. Assuming that a measurement or an estimate of the acoustic impulse responses $\hat{\mathbf{h}}_m$ is available, the subband dereverberation algorithm can be summarized as follows:

1. The subband equivalent filters are calculated with (7.36); these impulse responses are then used to design the subband equalization filters $\hat{\mathbf{g}}'_k$ with (7.37).
2. The decimated subband signals of the reverberant observations are obtained by processing each channel's observation data, $x'_{km}(n)$, with the analysis filterbank, followed by decimation such that

$$\begin{aligned} x'_{km}(n) &= y_{km}(Nn), \quad k = 0,1,\ldots K/2-1,\ m = 1,2,\ldots,M, \\ y_{km}(n) &= \mathbf{u}_k^T \mathbf{x}_m(n), \end{aligned} \tag{7.46}$$

where

$$\mathbf{u}_k = [u_{k,0}\ u_{k,1}\ \ldots\ u_{k,L_{\mathrm{pr}}-1}]^T$$

is the analysis filter with $u_{k,i}$ calculated from (7.21) and

$$\mathbf{x}_m(n) = [x_m(n)\ x_m(n-1)\ \ldots\ x_m(n-L_{\mathrm{pr}}+1)]^T.$$

3. The equalization filters are applied to each subband signal

$$\hat{s}_k(n) = \hat{\mathbf{G}}'_k \mathbf{x}'_{km}(n), \tag{7.47}$$

with

$$\hat{\mathbf{G}}'_k = [\hat{\mathbf{G}}'_{k1}\ \hat{\mathbf{G}}'_{k2}\ \ldots\ \hat{\mathbf{G}}'_{kM}],$$

where

$$\hat{\mathbf{G}}'_{km} = \mathrm{diag}\{\mathbf{g}'_{km}\ \cdots\ \mathbf{g}'_{km}\} = \begin{bmatrix} g'_{km,0} & 0 & \ldots & 0 \\ g'_{km,1} & g'_{km,0} & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ g'_{km,L'_{\mathrm{i}}-1} & \ldots & \vdots & 0 \\ 0 & g'_{km,L'_{\mathrm{i}}-1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & g'_{km,L'_{\mathrm{i}}-1} \end{bmatrix}$$

is a convolution matrix, constructed in a similar manner to $\mathbf{H}_m$ in (7.2) and

$$\mathbf{x}'_{km}(n) = [x'_{km}(n)\ x'_{km}(n-1)\ \ldots\ x'_{km}(n-L_{\mathrm{pr}}+1)]^T.$$

4. An equalized full-band signal is reconstructed by upsampling of the subband signals $s'_k(n)$ followed by the synthesis filter-bank according to

$$\hat{s}(n) = \sum_{k=0}^{K/2-1} \mathbf{v}_k^T \hat{\mathbf{s}}'_{N,k}(n), \tag{7.48}$$

where

$$\hat{\mathbf{s}}'_{N,k}(n) = [\hat{s}'_{N,k}(n)\ \hat{s}'_{N,k}(n-1)\ \ldots\ \hat{s}'_{N,k}(n-L_{\text{pr}}+1)]^T,$$
$$\hat{s}'_{N,k}(n) = \begin{cases} \hat{s}'_k(\frac{n}{N}), & \text{if } \frac{n}{N} \text{ is an integer,} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathbf{v}_k = [v_{k,0}\ v_{k,1}\ \ldots\ v_{k,L_{\text{pr}}-1}]^T$$

is the synthesis filter with $v_{k,i}$ calculated from (7.22).

This procedure designs the subband equalizer based on a given measurement of $\mathbf{h}_m$ in a computationally efficient manner and then applies the equalizer in the subbands before finally generating the full-band dereverberated signal.

7.7 Simulations and Results

The following simulation results are presented to demonstrate the performance of the SB-MCEQ equalization method. Four experiments were performed to show:

(i) The properties of the complex subband decomposition in relation to the downsampling factor N
(ii) A comparative performance evaluation with the full-band MCEQ using randomly generated channels
(iii) The application of the SB-MCEQ to simulated acoustic impulse responses with illustrative examples of the algorithm
(iv) The performance in speech dereverberation

7.7.1 Experiment 1: Complex Subband Decomposition

First, we provide some experimental results using the complex subband decomposition with focus on the effects of the downsampling factor N. The results, shown in Table 7.1, were obtained with a filter-bank with $K = 32$ subbands, a prototype filter of length $L_{\text{pr}} = 512$ and downsampling factors $N = 14, 16, \ldots, 32$; this variation of N represents a range from critical sampling ($N = K$) to 43.75% oversampling. A

Table 7.1 Complex subband decomposition for varying downsampling factor N

N	Aliasing suppression ξ_{aliasing} (dB)	NPM (dB)	Subband filtering accuracy, ξ_{accuracy} (dB)	Computational savings factor
32	18.3	−7.2	−4.0	228
30	31.5	−26.4	−20.8	196
28	33.5	−33.5	−28.0	159
26	51.5	−48.8	−44.7	128
24	82.2	−70.3	−54.0	106
22	94.5	−85.8	−58.2	80
20	103.1	−105.7	−70.8	61
18	99.8	−107.7	−74.3	45
16	146.1	−130.1	−87.9	30
14	206.0	−150.8	−100.2	21

simulated impulse response, shown in Fig. 7.8, was used and the subband equivalent filters were calculated using (7.36) and the following quantities were measured:

- *SNR due to in-band aliasing* (ξ_{aliasing}) is calculated from the prototype filter and is defined as the ratio of the energy of its passband to the energy of its stopband [36]

$$\xi_{\text{aliasing}} = 10\log_{10}\left(\frac{\sum_{k=0}^{\lfloor I/N \rfloor}|P(k)|^2}{\sum_{k=\lceil I/N \rceil+1}^{N-1}|P(k)|^2}\right)\ \text{dB}, \tag{7.49}$$

  where $P(k)$ is the I-point DFT of the prototype filter p_i. This measure quantifies how much of the aliasing between adjacent subbands is suppressed. Furthermore, it is related to the validity of property P1 in (7.23), which is a key property in the development of the subband decomposition. The larger the value of ξ_{aliasing}, the greater the accuracy of P1.
- *Normalized projection misalignment (NPM)* measures the similarity between two impulse responses, ignoring any scale factors. It was defined in Chap. 2. It is applied here by using a unit impulse as the input to the filter-bank. The resulting impulse response is then compared to the full-band impulse response. In this way, the accuracy of the desired overall transfer function of the filter-bank given in (7.25) is examined.
- *Subband filtering approximation accuracy* (ξ_{accuracy}) compares the accuracy of a signal filtered with the subband equivalent filters and the same signal at the output of the full-band filter. This measure is defined as in [29] to be

$$\xi_{\text{accuracy}} = 20\log_{10}\left(\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}\right)\ \text{dB}, \tag{7.50}$$

  where $\mathbf{x}$ is the output signal from the full-band filter and $\hat{\mathbf{x}}$ is the output signal from the filter-bank with the subband equivalent filters. The results are obtained with a zero mean white Gaussian noise input sequence for both the full-band and the subband filters, resulting in $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively.

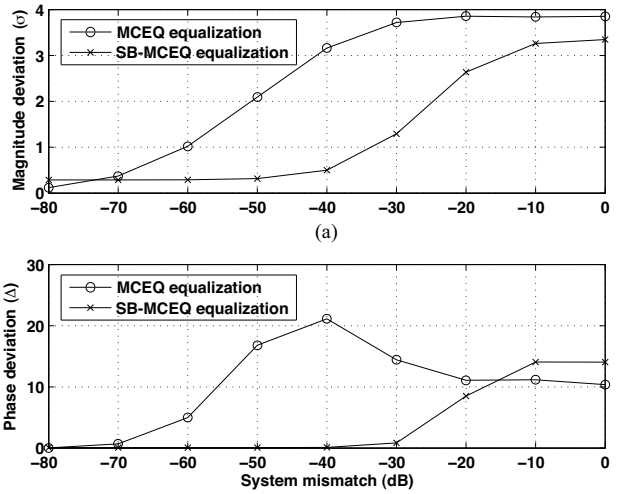


Fig. 7.7 (a) Magnitude deviation and (b) phase deviation *vs.* system mismatch for full-band equalization with MCEQ (*circles*) and subband equalization with SB-MCEQ (*crosses*) using randomly generated impulse responses

- *Computational savings factor* is the ratio between the number of computations, in flops, required to calculate the full-band MCEQ equalizing filters according to (7.43) and the computations required for the SB-MCEQ equalizing filter design according to (7.44).

The results of this experiment highlight the inversely proportional relationship of the computational gain and subband filtering accuracy, a trade-off that is controlled by N. Greatest computational gain is achieved at critical sampling (a savings factor of 228 in this case). However, in this extreme case the aliasing suppression is low and, consequently, the subband filtering accuracy is poor. On the other hand, when the oversampling is large ($N = 14$), the aliasing effects are reduced to very low levels ($\xi_{\text{aliasing}} = 206$ dB), resulting in very accurate subband filtering, however, with a relatively low computational gain (a savings factor of 21 in this case). Our choice of $N = 24$ for the following experiments represents a good trade-off between these two factors.

7.7.2 Experiment 2: Random Channels

This experiment demonstrates the performance of the SB-MCEQ equalizer, compared with full-band MCEQ. A system with $M = 5$ randomly generated $L = 512$-tap

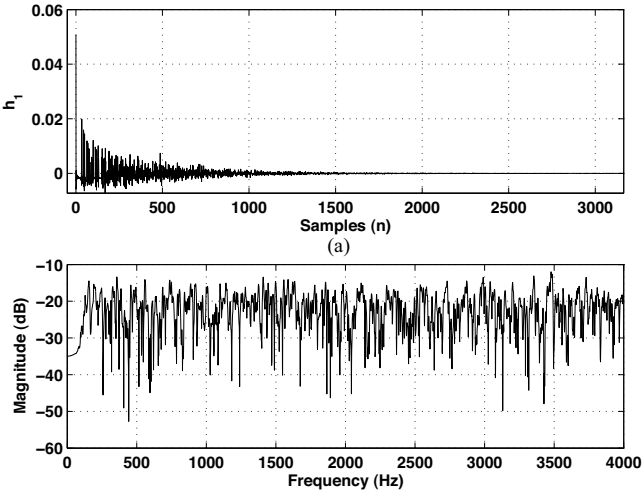


Fig. 7.8 Example of (a) a simulated room impulse response and (b) its corresponding magnitude response at one of the microphones in the array, generated using the source-image method

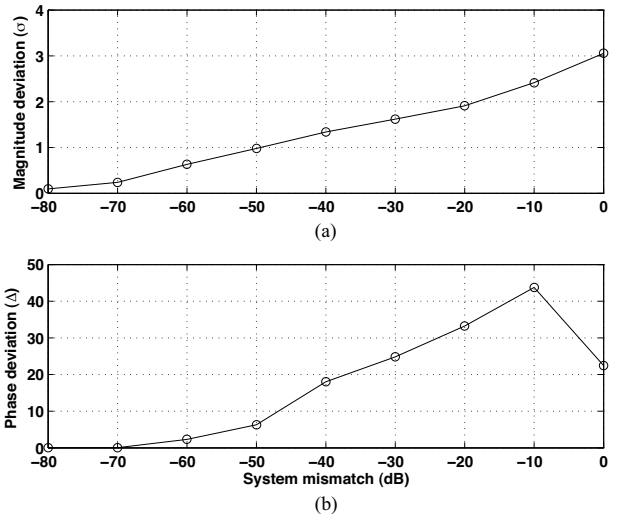


Fig. 7.9 (a) Magnitude deviation and (b) phase deviation *vs.* system mismatch for SB-MCEQ equalization of simulated acoustic impulse responses

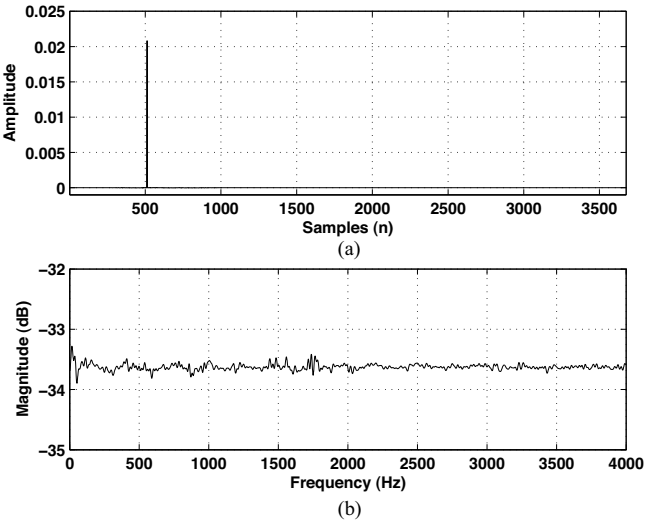


Fig. 7.10 Equalized (a) time domain impulse response and (b) magnitude response, using the SB-MCEQ method $\mathcal{M}_m = -\infty$ dB. The magnitude deviation is $\sigma = 0.025$. (Note that the magnitude scaling of the equalized impulse response is of no significance.)

channels was used and system misalignment, $\mathcal{M}_m$ varying between 0 and -80 dB was simulated with (7.16). The taps of the channel impulse responses were generated using random sequences drawn from a zero mean, unit variance Gaussian distribution. The results, which are an average over 100 different channel realizations, are shown in Fig. 7.7 for the full-band MCEQ (*circles*) and for the proposed subband implementation (*crosses*). Notably, the SB-MCEQ exhibits much gentler performance degradation with increased misalignment in comparison with the full-band MCEQ and with a similar behaviour as the single channel SCLS equalizer shown in Fig. 7.2. Thus, the SB-MCEQ method is shown in these results to be less sensitive to inexact impulse responses, while benefiting from the shorter filters of multichannel inversion. This improvement is a consequence of the reduced filter length in the subbands, which in Section 7.3.2 was shown to improve the MCEQ equalizer performance.

7.7.3 Experiment 3: Simulated Room Impulse Responses

We now demonstrate the performance of the SB-MCEQ equalizer for simulated ATFs. A linear array of $M = 8$ uniformly distributed microphones with 0.05 m separation between adjacent sensors was simulated using the source-image

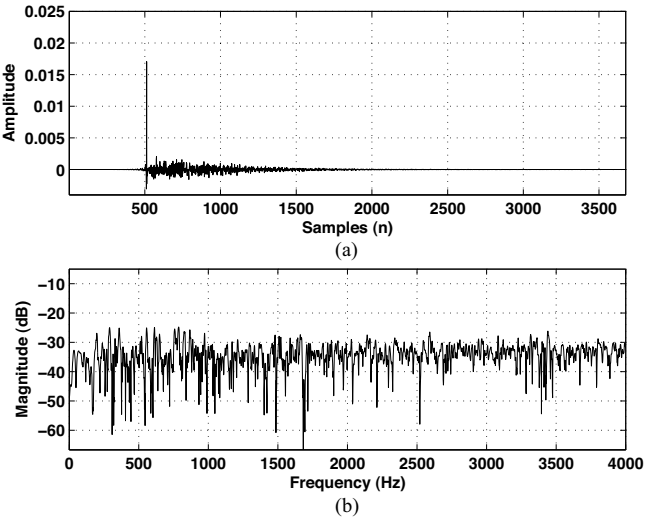


Fig. 7.11 Equalized (a) time domain impulse response and (b) magnitude response, using the SB-MCEQ method for $\mathcal{M}_m = -10$ dB. The magnitude deviation is $\sigma = 2.25$. (Note that the magnitude scaling of the equalized impulse response is of no significance but the relative scaling between Figs. 7.10a and 7.11a is of interest.)

method [1, 25] for a room with dimensions $5 \times 4 \times 3$. The impulse response at one of the microphones, $\mathbf{h}_1$, is depicted in Fig. 7.8. The channel lengths are $L = 3200$ taps, which is equivalent to $T_{60} = 0.4$ s at $f_s = 8$ kHz sampling frequency. Moreover, keeping the source-microphone configuration fixed, ATFs were simulated at 20 different locations in the room. System misalignment, $\mathcal{M}_m$ varying between 0 and -80 dB was simulated with (7.16). Figure 7.9 shows the results in terms of magnitude and phase deviation, as an average of the 10 measurement locations. This again shows a similar pattern to the single channel case as in the previous experiment. Thus, with the subband equalizer, we are able to achieve the lower sensitivity to system estimate inaccuracies of the approximate SCLS but with the filter lengths of the MCEQ. In addition, nearly perfect equalization is achieved with the SB-MCEQ method for $\mathcal{M}_m \leq -40$ dB.

Finally, we provide two characteristic examples of the subband equalizer output for the simulated ATFs. Figure 7.10a shows a typical outcome of the equalized acoustic impulse response in the time domain and Fig. 7.10b shows the corresponding magnitude response for $\mathcal{M}_m = -\infty$ dB. It can be seen that near perfect equalization is achieved with only small spectral deviation ($\sigma = 0.025$); this deviation results from the approximations in the subband filter decomposition and in the filter-bank reconstruction. Thus, the accuracy depends on the ability of the prototype filter to suppress aliasing and on the oversampling ratio. The delay in the equalized impulse in Fig. 7.10a is due to the filter-bank and is governed by the order of the prototype

filter L_{pr}. As a further illustration for a less accurate ATF estimation, a characteristic outcome for $\mathcal{M}_m = -10$ dB is shown in Fig. 7.11, where the spectral deviation is more notable ($\sigma = 2.26$), which is due to inaccuracies in the acoustic impulse response.

7.7.4 Experiment 4: Speech Dereverberation

The subband equalization method is now applied to speech dereverberation. Test data comprising the sentence "George made the girl measure a good blue vase." uttered by a a male talker was drawn from the APLAWD database [17]. The sampling frequency was set to $f_{\text{s}} = 8$ kHz. A room with dimensions $5 \times 4 \times 3$ m was simulated using the source-image method [1, 25]. An eight-element linear microphone array with 0.05 m spacing between adjacent microphones was modelled and the reverberation time was varied in the range $T_{60} = \{0.2 - 0.8\}$ s in incremental steps of 0.05 s. Various levels of misalignment were simulated ranging from 0 to $-\infty$. A delay-and-sum beamformer was used as a baseline algorithm. The segmental Signal to Reverberation Ratio (SRR) and Bark Spectral Distortion (BSD), defined in Chap. 2, were applied using 30 ms frames with 50% overlap to quantify the improvement of the processed speech.

The results in terms of Segmental SRR *vs.* reverberation time, T_{60}, are shown in Fig. 7.12 for reverberant speech, speech at the output of the delay-and-sum beamformer and speech inverse filtered with the SB-MCEQ and with various levels of misalignment in the impulse responses ranging from 0 to $-\infty$. The legend in the figure indicates the different plots. In all cases the inverse filtering approach provides significant improvements over the delay-and-sum beamformer, but the improvement degrades as the misalignment increases, as could be expected. Figure 7.13 shows the results evaluated in terms of the Bark spectral distortion for reverberant speech, speech at the output of the delay-and-sum beamformer and speech inverse filtered with SB-MCEQ and with various levels of misalignment in the impulse responses ranging from 0 to $-\infty$. It can be seen that for $\mathcal{M}_m < -30$ dB, the dereverberated speech signals are effectively equivalent (in terms of BSD) to the clean speech for all reverberation times considered. This was also confirmed with informal listening tests. The listening tests also indicate that, in the case of $\mathcal{M}_m = 0$ dB, there is an audible residual echo in the processed speech signal, despite the apparent improvement in terms of the error measures used. This result, however, is not surprising and corresponds accurately to the expected outcome based on the results presented in Figs. 7.7 and 7.9, where a relatively large spectral deviation is observed at $\mathcal{M}_m = 0$ dB.

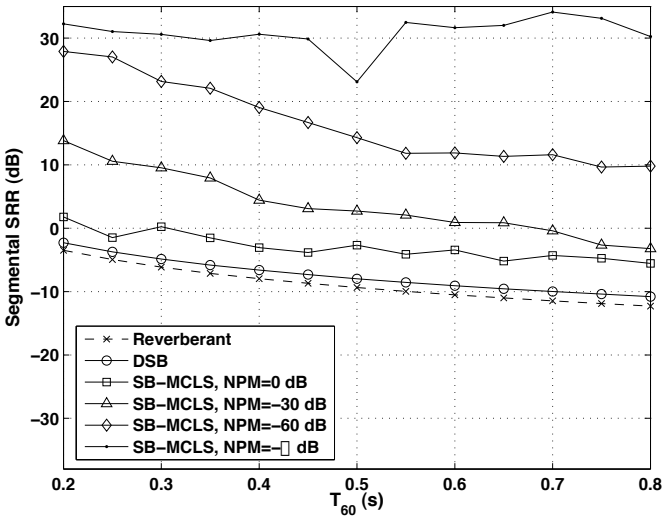


Fig. 7.12 Segmental SRR for reverberant speech at the microphone closest to the talker, speech at the output of a DSB and speech processed with subband inverse filtering for $\mathcal{M}_m = \{0, -30, -60, -\infty\}$ dB

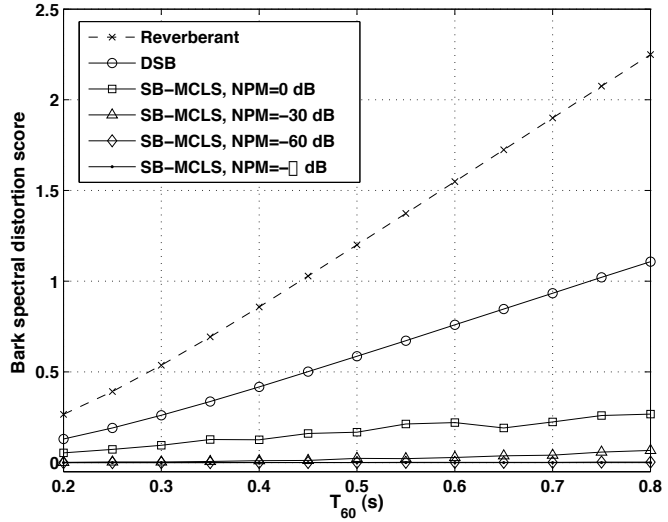


Fig. 7.13 Bark spectral distortion score for reverberant speech at the microphone closest to the talker, speech at the output of a DSB and speech processed with subband inverse filtering for $\mathcal{M}_m = \{0, -30, -60, -\infty\}$ dB

7.8 Summary

Equalization of acoustic impulse responses has been discussed both for single and multiple microphones. Single microphone approaches can provide only approximate equalization, require very long inverse filters and result in a long processing delay due to the non-minimum phase property of the ATFs. On the other hand, exact equalization with no delay and with inverse filters of similar order to the acoustic impulse responses is possible in the multi-microphone case. However, multichannel methods are very sensitive to inaccuracies in the estimated systems to be equalized, causing significant distortions to the equalized signal.

Consequently, a new algorithm was derived operating on decimated oversampled subband signals, where the full-band impulse response is decomposed into equivalent filters in the subbands and multichannel equalization is applied to each subband. It was shown that this method results in substantial computational savings at the cost of very small spectral distortion due to the filter-bank. Simulation results were presented to evaluate the performance of this method and equalization of channels of several thousand taps was demonstrated. Most importantly, experimental results indicated that the new method is more robust to errors in the impulse responses of the channels to be equalized, which is due to a combination of shorter filters and approximation of the filtering in the subbands. Thus, the proposed subband multichannel equalization benefits from the reduced sensitivity to channel estimation errors, shorter equalization filters, no delay due to the equalization (the delay due to the filter-bank is less than 32 ms in our examples), giving significant advantages over existing single and multichannel techniques. Finally, the subband inverse filtering method was applied in the context of speech dereverberation, where the results show that near perfect dereverberation can be achieved with impulse responses with estimation errors of $\mathcal{M}_m < -30$ dB.

References

1. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
2. Bharitkar, S., Hilmes, P., Kyriakakis, C.: Robustness of spatial average equalization: A statistical reverberation model approach. J. Acoust. Soc. Am. **116**(6), 3491–3497 (2004)
3. Cho, J.H., Morgan, D.R., Benesty, J.: An objective technique for evaluating doubletalk detectors in acoustic echo cancelers. IEEE Trans. Speech Audio Process. **7**(7), 718–724 (1999)
4. Golub, G.H., van Loan, C.F.: Matrix computations, 3 edn. John Hopkins Series in the Mathematical Sciences. John Hopkins University Press (1996)
5. Haneda, Y., Makino, S., Kaneda, Y.: Common acoustical pole and zero modeling of room transfer functions. IEEE Trans. Speech Audio Process. **2**(2), 320–328 (1994)
6. Haneda, Y., Makino, S., Kaneda, Y.: Multiple-point equalization of room transfer functions by using common acoustical poles. IEEE Trans. Speech Audio Process. **5**(4), 325–333 (1997)
7. Harikumar, G., Bresler, Y.: FIR Perfect Signal Reconstruction from Muliple Convolutions: Minimum Deconvolver Orders. IEEE Trans. Signal Process. **46**(1), 215–218 (1998)
8. Hikichi, T., Delcroix, M., Miyoshi, M.: Inverse filtering for speech dereverberation less sensitive to noise. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4 (2006)

9. Hikichi, T., Delcroix, M., Miyoshi, M.: On robust inverse filter design for room transfer function fluctuations. In: Proc. European Signal Processing Conf. (EUSIPCO) (2006)
10. Hikichi, T., Delcroix, M., Miyoshi, M.: Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. EURASIP J. Advances in Signal Process. **2007**, 1–12 (2007)
11. Hofbauer, M.: Optimal linear separation and deconvolution of acoustical convolutive mixtures. Ph.D. thesis, The Swiss Federal Institute of Technology (ETH), Zürich (2005)
12. Hofbauer, M., Loeliger, H.: Limitations for FIR multi-microphone speech dereverberation in the low-delay case. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 103–106 (2003)
13. Hopgood, J.R., Rayner, P.J.W.: A probabilistic framework for subband autoregressive models applied to room acoustics. In: Proc. IEEE Workshop Statistical Signal Processing, pp. 492 – 495 (2001)
14. Huang, Y., Benesty, J., Chen, J.: A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. IEEE Trans. Speech Audio Process. **13**(5), 882–895 (2005)
15. Kuttruff, H.: Room acoustics, 4 edn. Taylor & Francis (2000)
16. Lanciani, C.A., Schafer, R.W.: Subband-domain filtering of MPEG audio signals. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 917–920 (1999)
17. Lindsey, G., Breen, A., Nevard, S.: SPAR's archivable actual-word databases. Tech. rep., University College London (1987)
18. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
19. Mourjopoulos, J., Clarkson, P., Hammond, J.: A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 7, pp. 1858–1861 (1982)
20. Mourjopoulos, J.N.: Digital equalization of room acoustics. J. Audio Eng. Soc. **42**(11), 884–900 (1994)
21. Naylor, P.A., Gaubitch, N.D.: Speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Eindhoven, The Netherlands (2005)
22. Naylor, P.A., Tanrikulu, O., Constantinides, A.G.: Subband adaptive filtering for acoustic echo control using allpass polyphase IIR filterbanks. IEEE Trans. Speech Audio Process. **6**(2), 143–155 (1998)
23. Neely, S.T., Allen, J.B.: Invertibility of a room impulse response. J. Acoust. Soc. Am. **66**(1), 165–169 (1979)
24. Nelson, P.A., Orduña-Brustamante, F., Hamada, H.: Inverse filter design and equalization zones in multichannel sound reproduction. IEEE Trans. Speech Audio Process. **3**(3), 185–192 (1995)
25. Peterson, P.M.: Simulating the response of multiple microphones to a single acoustic source in a reverberant room. J. Acoust. Soc. Am. **80**(5), 1527–1529 (1986)
26. Proakis, J.G., Manolakis, D.G.: Digital signal processing, 3 edn. Prentice Hall (1996)
27. Radlović, B.D., Kennedy, R.A.: Nonminimum-phase equalization and its subjective importance in room acoustics. IEEE Trans. Speech Audio Process. **8**(6), 728–737 (2000)
28. Radlović, B.D., Williamson, R.C., Kennedy, R.A.: Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Acoust., Speech, Signal Process. **8**(3), 311–319 (2000)
29. Reilly, J.P., Wilbur, M., Seibert, M., Ahmadvand, N.: The complex subband decomposition and its application to the decimation of large adaptive filtering problems. IEEE Trans. Signal Process. **50**(11), 2730–2743 (2002)
30. Schroeder, M..R.: Statistical parameters of the frequency response curves of large rooms. J. Audio Eng. Soc. **35**(5), 299–305 (1987)
31. Talantzis, F., Ward, D.B.: Robustness of multi-channel equalization in an acoustic reverberant environment. J. Acoust. Soc. Am. **114**(2), 833–841 (2003)

32. Tohyama, M., Lyon, R.H., Koike, T.: Source waveform recovery in a reverberant space by cepstrum dereverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 157–160 (1993)
33. Vaidyanathan, P.P.: Multirate systems and filter banks. Prentice Hall (1993)
34. Wang, H., Itakura, F.: Realization of acoustic inverse filtering through multi-microphone subband processing. IEICE Trans. Fundamentals **E75-A**(11), 1474–1483 (1992)
35. Weiss, S., Rice, G.W., Stewart, R.W.: Multichannel equalization in subbands. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 203–206 (1999)
36. Weiss, S., Stewart, R.W.: On adaptive filtering in oversampled subbands. Shaker Verlag (1998)
37. Yamada, K., Wang, J., Itakura, F.: Recovering of broad band reverberant speech signal by subband MINT method. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 969–972 (1991)

Chapter 8
Bayesian Single Channel Blind Dereverberation of Speech from a Moving Talker

James R. Hopgood, Christine Evers, and Steven Fortune

Abstract This chapter discusses a model-based framework for single-channel blind dereverberation of speech, in which parametric models are used to represent both the unknown source and the unknown acoustic channel. The parameters of the entire model are estimated using the Bayesian paradigm, and an estimate of the source signal is found by either inverse filtering of the observed signal with the estimated channel coefficients, or directly within a sequential framework. Model-based approaches fundamentally rely on the availability of realistic and tractable models that reflect the underlying speech process and acoustic systems. The choice of these models is extremely important and is discussed in detail, with a focus on spatially varying room impulse responses. The mathematical framework and methodology for parameter estimation and dereverberation is also discussed. Some examples of the proposed approaches are presented with results.

8.1 Introduction and Overview

Acoustic dereverberation arises when an audio signal is radiated in a confined acoustic space. Blind dereverberation is an important and challenging signal processing problem, which is required when this audio signal is acquired by a sensor placed away from the source by a distance greater than the reverberation distance [25]. This problem differs from Acoustic Echo Cancellation (AEC) found in, for example, teleconferencing applications, where a known source signal emitted from a loudspeaker is distorted by acoustic reflections (or *system echoes*), and results in a feedback path to the microphone sensor. AEC is generally a non-blind deconvolution problem and is typically solved using well-known adaptive filtering algorithms. In blind dereverberation, however, the source signal,[1] the source location, and con-

University of Edinburgh, UK

[1] The source is necessarily unknown since, if it were known, there would be no need for signal enhancement.

sequently the Room Impulse Response (RIR) between the source and sensor, are all assumed unknown. If an estimate of the RIR were available, the effect of reverberation could be removed by filtering the observed signal with the inverse of the RIR. However, in practice, the RIR is unknown since it is not possible to measure the specific source-sensor Room Transfer Function (RTF) between any two arbitrary positions using a fixed measurement geometry. Although it might be possible to estimate the common-acoustic component of the response from a measurement between two other positions in the room, this is only useful with additional geometry-specific information.[2]

With only the observations available, the blind deconvolution problem is underdetermined, i.e., more unknowns than observations must be estimated from a single realisation of the measurement process at each time instance. Prior knowledge of the statistical properties of the source and channel is essential for solving this problem, and can be incorporated through a model-based approach to blind dereverberation. The rest of this section is organised as follows: an overview of a model-based approach to blind dereverberation and the numerical methods involved is presented in Sect. 8.1.1; a discussion of practical issues that occur in blind dereverberation is given in Sect. 8.1.2; the organisation of the remainder of the chapter is outlined in Sect. 8.1.3.

8.1.1 Model-based Framework

In a model-based approach to blind dereverberation, the source and acoustics are represented by parametric models. The parameters of this system model are estimated from the observed data, and subsequently used to reconstruct the source signal. The problem of blind dereverberation is thus transformed into an exercise in parameter estimation and inference. If all the parameters and observable variables in the source and channel models are regarded as unknown stochastic quantities, the system model can be rephrased in a statistical context using Probability Density Functions (PDFs). There is a plethora of statistical parameter estimation techniques available, including maximum likelihood methods such as the Expectation Maximization (EM) algorithm. However, a robust and consistent way of exploiting and manipulating these PDFs is by using Bayes's theorem to infer a degree of belief of an unknown hypothesis. More specifically, the Bayesian framework provides a learning procedure where knowledge of the system is inferred from prior belief and updated through the availability of new data.

In this chapter, Bayesian inference and associated numerical optimisation methods are used for parameter estimation. Monte Carlo approaches are used to obtain empirical estimates of the resulting target distributions by drawing a large number of samples from a (potentially different) hypothesis or sampling distribution. Parame-

[2] Such a measurement of such a common-acoustical component could, for example, be incorporated in self-calibrating teleconferencing applications.

ter estimates are then obtained from averaging the drawn variates. These algorithms are generally divided into offline batch methods and online sequential approaches.

8.1.1.1 Online *vs.* Offline Numerical Methods

Online methods assume the signal is presented in a stream and can be processed sequentially and immediately as each sample is observed. Batch methods, on the other hand, assume that the observed signal samples become available only as soon as all the data has been measured. Based on this collective information, batch methods *explore* the system using the knowledge inferred from all observations. In contrast, online methods are adaptive approaches that *track* a system model with each processed sample. Online methods thus facilitate real-time processing and can be used where data sets are not fixed, i.e., where new data constantly becomes available. However, in order to build a realistic hypothesis from one sample only, online methods often require more complex approaches than batch methods and can hence be more computationally expensive and complicated to implement. Implementations of online methods are based on Sequential Monte Carlo (SMC) techniques in the Bayesian framework, whereas batch methods are frequently implemented using Markov Chain Monte Carlo (MCMC) techniques, for example the Gibbs sampler.

The choice of whether an application operates sequentially or in a batch mode not only depends on the nature of the availability of data, but is closely tied to the choice of methodologies and models that can actually facilitate either online or offline estimation. Each methodology and model carries its own advantages and drawbacks that need to be weighed carefully in order to decide between sequential and batch processing. This is discussed further below, while a comparison of the numerical methods for online and offline approaches is given in Sect. 8.2.3, and a comparison of results for dereverberation of speech from a stationary talker is presented in Sect. 8.7.3.

8.1.1.2 Parametric Estimation and Optimal Filtering methods

In addition to the choice of using either batch or sequential processing, there is the choice of two distinct approaches to the inference problem:

1. Estimate the room impulse response and obtain an estimate of the source signal by inverse filtering the observed signal with the estimated channel coefficients. In general, a static parametric model is used for the RIR, so this is an exercise in offline parameter estimation using batch methods.
2. Estimate the source signal directly as though it were an unknown parameter – this is an exercise in optimal filtering, and therefore is solved in a sequential manner using online methods.

Each of these approaches fundamentally rely on the availability of realistic and tractable models that reflect the underlying speech processes and acoustic systems:

model selection is therefore extremely important. Generation of speech through the vocal tract as well as the effect of the reverberation process on audio signals should motivate the choice of a particular model. The nature of room acoustics is investigated in Sect. 8.3. Based on these findings, two different channel models are proposed in Sects. 8.4.6 and 8.4.7. The time-varying nature of speech signals and the rationale for the proposed speech production models are discussed in Sect. 8.6.

8.1.2 Practical Blind Dereverberation Scenarios

Blind dereverberation has recently received much attention in the literature, but often a number of key assumptions about the application setup are made. The first is in the use of multi-microphone techniques, and the second is in solutions that assume time-invariance of the acoustic channel. Neither of these assumptions is always appropriate in practice as outlined below.

8.1.2.1 Single-sensor Applications

Spatial diversity of acoustic channels can be constructively exploited by multiple sensor blind dereverberation techniques [28] in order to obtain an estimate of the remote speech signal. Nevertheless, despite the usefulness and power of spatial diversity, there are numerous applications where only a single measurement of the reverberant signal is available. Single-sensor blind dereverberation is utilised in applications where numerous microphones prove infeasible or ineffectual due to the physical size of arrays. Examples of applications with commercial appeal include hearing aids, hands-free telephony, and automatic speech recognition. For these reasons, this chapter considers the single-sensor problem of blind dereverberation, although Bayesian approaches to the multi-sensor case have been considered in [10, 15, 17].

8.1.2.2 Time-varying Acoustic Channels

Signal processing in acoustic environments is often approached with the assumption that the room impulse response is time-invariant. This is appropriate in scenarios where the source-sensor geometry is not rapidly varying, for example, a hands-free kit in a car cabin, in which the driver and the microphone are approximately fixed relative to one another, or in a work environment where a user is seated in front of a computer terminal. However, there are many applications where the source-sensor geometry is subject to change; the wearer of a hearing-aid typically wishes to move around a room, as might users of hands-free conference telephony equipment. A talker moving in a room at 1 m/s covers a distance of 50 mm in 50 ms. This distance might be enough for the room impulse response to vary sufficiently that any assumption of a time-invariant acoustic channel is no longer valid (see Sect. 8.4.5).

An implicit assumption often made is that the physical properties giving rise to the acoustics of the room are time-invariant; thus, it is assumed that it is the variable source-sensor geometry that leads to the changing RIR. However, it is not beyond possibility that the room acoustics may vary: the changing state of doors, windows, or items being moved in the room will influence the room dynamics.

Although there is some limited recent work dealing with time-varying acoustic channels [4, 31], generally the problem of single-channel blind dereverberation of speech from a moving talker has to date received little attention from the signal processing community. This is in part because the case of a stationary talker has not yet been solved satisfactorily. Nevertheless, the problem is of growing interest, and in itself can give insight to the simpler Linear Time-Invariant (LTI) problem. This chapter specifically attempts to bridge this gap by considering Linear Time-Variant (LTV) channels for blind dereverberation of speech from moving talkers.

8.1.3 Chapter Organisation

The remainder of this chapter is organised as follows: Section 8.2 introduces a mathematical formulation of the blind dereverberation problem including model ambiguities. Sect. 8.2.1 revises the Bayesian framework used for blind dereverberation. The nature of room acoustics is considered in Sect. 8.3, which provides motivation for the parametric channel models in Sect. 8.4. Noise and source models are outlined in Sects. 8.5 and 8.6, respectively. Details of several offline and online blind dereverberation algorithms are then given in Sect. 8.7, while some brief conclusions are found in Sect. 8.8.

8.2 Mathematical Problem Formulation

Typically, in single-channel blind deconvolution, the degraded observation, $x(n)$,[3] is modelled as the linear convolution of the unknown source signal, $s(n)$, and a room impulse response, $h_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n)$, in additive noise, $v(n)$, as indicated in Fig. 8.1. This model assumes the noise within an acoustic environment is an additive common signal unaffected by the acoustics of a room. Moreover, as discussed in Sects. 8.3 and 8.4, the RIR is dependent on the source and observer positions, $\mathbf{q}_{\text{src}}$ and $\mathbf{q}_{\text{mic}}$, respectively. If the source and sensor positions vary with time, such that $\mathbf{q}_{\text{src}} = \mathbf{q}_{\text{src}}(n)$ and $\mathbf{q}_{\text{mic}} = \mathbf{q}_{\text{mic}}(n)$ are functions of time, then the spatially varying nature of the RIR corresponds to a time-varying impulse response function. This response is denoted by $h_{(\mathbf{q}_{\text{src}}(\ell),\mathbf{q}_{\text{mic}}(\ell))}(n) = h(n,\ell)$, and represents the RIR at time index n to an impulse applied to the system at time index ℓ. Consequently, the discrete-time

[3] All signals are assumed to be defined over the range $n \in \mathcal{N} = \{0,\ldots,N-1\}, N \in \mathbb{Z}^+$ is a positive integer. In all other cases, unless stated otherwise, the following set notation is used for simplicity: $\mathcal{U} = \{1,\ldots,U\} \subset \mathbb{Z}^{+U}$.

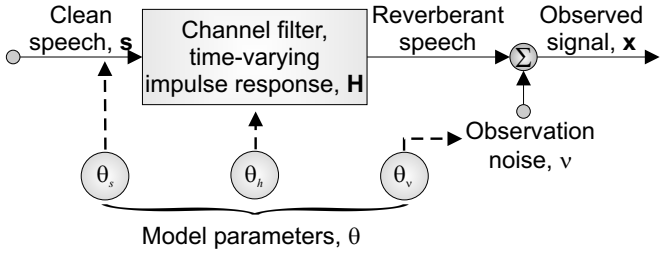


Fig. 8.1 General additive noise system model

model is written as:[4]

$$x(n) = \sum_{\ell \in \mathcal{L}} h(n,\ell)\, s(\ell) + \nu(n). \tag{8.1}$$

The characteristics of the noise term, $\nu(n)$, are discussed in depth in Sect. 8.5. Often, however, this observation error is used to encompass all other background noise sources in the acoustic environment; application of the central limit theorem is used to argue that the sum of all background noise is Gaussian and unaffected by the acoustics of the room. Additionally, some noise sources might lead to a diffuse sound field and, since they have unknown statistics, again it is reasonable to model their superposition as Gaussian. Thus, $\nu(n)$, is typically assumed to be White Gaussian Noise (WGN) with variance σ_ν^2, uncorrelated with both the RIR and the source signal, such that:

$$\nu(n) \sim \mathcal{N}\left(0, \sigma_\nu^2\right). \tag{8.2}$$

The convolution in (8.1) may be written in matrix-vector form by defining the vectors $[\mathbf{x}]_i = x(i)$, $[\mathbf{s}]_i = s(i)$, $[\nu]_i = \nu(i)$, $i \in \mathcal{N}$, and the matrix $[\mathbf{H}]_{i,j} = h(i,j)$, $\{i,j\} \in \mathcal{N} \times \mathcal{N}$, such that:

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \nu. \tag{8.3}$$

If the source and observer have a fixed spatial geometry, such that $\mathbf{q}_{\text{src}}$ and $\mathbf{q}_{\text{mic}}$ are time-invariant, then the RIR is also time-invariant due to its dependency on the fixed values of $\mathbf{q}_{\text{src}}$ and $\mathbf{q}_{\text{mic}}$. By writing $h_{(\mathbf{q}_{\text{src}}(\ell),\mathbf{q}_{\text{mic}}(\ell))}(n,\ell) \equiv h_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n-\ell) \triangleq h(n-\ell)$, (8.1) reduces to the standard LTI convolution:

$$x(n) = \sum_{\ell \in \mathcal{L}} h(n-\ell)\, s(\ell) + \nu(n) \equiv h(n) * s(n) + \nu(n), \tag{8.4}$$

and the matrix $\mathbf{H}$ of (8.3) becomes Toeplitz. The general objective of blind dereverberation is to estimate the source signal, $\mathbf{s}$, or the matrix of room impulse responses, $\mathbf{H}$, based on prior knowledge about $\mathbf{s}$, the noise ν, and $\mathbf{H}$. An inference framework is required to estimate the unknowns $\mathbf{s}$ and $\mathbf{H}$. As outlined in Sect. 8.1.1, the approach presented in this chapter is to parametrically model these unknowns and estimate

[4] Thus, if $s(n) = \delta(n-\ell)$ represents an impulse applied at time ℓ, the convolution of (8.1) gives the output $x(n) = h(n,\ell)$ as required.

the model parameters using the Bayesian paradigm, as described in the following section.

8.2.1 Bayesian Framework for Blind Dereverberation

Bayesian methods use probability density functions to quantify degrees of belief in an uncertain hypothesis, and utilise the rules of probability as the calculus for operating on those degrees of belief. Thus, a fundamental principle of the Bayesian philosophy is to regard all parameters and observable variables as unknown stochastic quantities. Two key characteristics of the Bayesian framework include the *consistency* of its inductive inference, and the utilisation of the *marginalisation operator*. Bayesian approaches are consistent since the calculus of probability is consistent: any valid use of the rules of probability will lead to a unique conclusion. Marginalisation is a powerful inferential tool that facilitates the reduction of the number of parameters appearing in the PDFs by the so-called *elimination of nuisance parameters*. Consider a data model, $\mathcal{M}$, with unknown parameters, $\boldsymbol{\theta}_{\mathcal{M}}$, for the N samples of observed data, $\mathbf{x} = \{x(n), n \in \mathcal{N}\}$. The posterior probability, $p(\boldsymbol{\theta}_{\mathcal{M}} \,|\, \mathbf{x}, \mathcal{M})$, for the unknown parameters is defined by Bayes's theorem as

$$p(\boldsymbol{\theta}_{\mathcal{M}} \,|\, \mathbf{x}, \mathcal{M}) = \frac{p(\mathbf{x} |\, \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M})\, p(\boldsymbol{\theta}_{\mathcal{M}} \,|\, \mathcal{M})}{p(\mathbf{x} |\, \mathcal{M})}, \tag{8.5}$$

where $p(\mathbf{x} |\, \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M})$ is the likelihood, $p(\boldsymbol{\theta}_{\mathcal{M}} \,|\, \mathcal{M})$ is the prior PDF on $\boldsymbol{\theta}_{\mathcal{M}}$. The term $p(\mathbf{x} |\, \mathcal{M})$ is called the evidence, and is usually regarded as a normalising constant. Given the likelihood and the prior distributions, Bayesian methods aim to estimate the unknown parameters from the posterior distribution.

In the most general case of single-channel blind dereverberation, the system is expressed by (8.3) where the original source signal, $\mathbf{s}$, the room impulse response, $\mathbf{H}$, and the noise, $\boldsymbol{\nu}$, are all considered as random vectors or matrices. Each of these random quantities possesses a corresponding PDF that models knowledge of the speech production process, the nature of reverberation, and the nature of any observation noise, respectively. Moreover, each of $\mathbf{s}$, $\mathbf{H}$, and $\boldsymbol{\nu}$, depends on a set of parameters denoted by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_h, \boldsymbol{\theta}_\nu\}$, respectively. Thus, a direct application of Bayes's theorem in (8.5) yields the joint PDF of all the unknown parameters given the observations $\mathbf{x}$:

$$p(\mathbf{s}, \mathbf{H}, \boldsymbol{\nu}, \boldsymbol{\theta} \,|\, \mathbf{x}) = \frac{p(\mathbf{x} |\, \mathbf{s}, \mathbf{H}, \boldsymbol{\nu}, \boldsymbol{\theta})\, p_S(\mathbf{s} |\, \boldsymbol{\theta}_s)\, p_H(\mathbf{H} |\, \boldsymbol{\theta}_h)\, p_\nu(\boldsymbol{\nu} |\, \boldsymbol{\theta}_\nu)\, p_\Theta(\boldsymbol{\theta})}{p_X(\mathbf{x})}, \tag{8.6}$$

where it is assumed that $\mathbf{s}$, $\mathbf{H}$ and $\boldsymbol{\nu}$ are *a priori* conditionally independent given the system parameters $\boldsymbol{\theta}$.[5] The denominator $p_X(\mathbf{x})$ is independent of the unknown vectors and can therefore be considered as a normalising constant, except in the case

[5] The subscripts denoting the variable which defines a PDF are omitted from the terms $p(\mathbf{s}, \mathbf{H}, \boldsymbol{\nu}, \boldsymbol{\theta} \,|\, \mathbf{x})$ and $p(\mathbf{x} |\, \mathbf{s}, \mathbf{H}, \boldsymbol{\nu}, \boldsymbol{\theta})$, in (8.6) and onwards, for clarity.

of model selection. The term $p_\Theta(\theta)$ contains all *a priori* knowledge, i.e., it reflects knowledge about the parameters before the data is observed. By means of prior densities, the posterior, $p(\mathbf{s}, \mathbf{H}, \nu, \theta \mid \mathbf{x})$, can therefore be manipulated by inferring any required statistic, leading to a fully interpretable PDF. If no prior knowledge is available, the prior PDF should be broad and flat compared to the likelihood. Such priors are known as non-informative and convey ignorance of the values of the parameters before observing the data.

If $\mathbf{s}$, $\mathbf{H}$, ν, and θ, are all known then the value of the observation vector $\mathbf{x} = \mathbf{H}\mathbf{s} + \nu$ is unique. Therefore, it directly follows that:

$$p(\mathbf{x} \mid \mathbf{s}, \mathbf{H}, \nu, \theta) = \delta(\mathbf{x} - [\mathbf{H}\mathbf{s} + \nu]).$$

Consequently, since the observations $\mathbf{x}$ are known, when any two of the three random vectors, $\{\mathbf{s}, \mathbf{H}, \nu\}$, in (8.6) are known, the solution of the third is trivial. Since the noise model in Fig. 8.1 is additive, ν is commonly considered as the determined random vector, and (8.6) simplifies to:

$$p(\mathbf{s}, \mathbf{H}, \theta \mid \mathbf{x}) \propto p_S(\mathbf{s} \mid \theta_s)\, p_H(\mathbf{H} \mid \theta_h)\, p_\nu(\mathbf{x} - \mathbf{H}\mathbf{s} \mid \theta_\nu)\, p_\Theta(\theta), \tag{8.7}$$

where $p_\nu(\cdot \mid \theta_\nu)$ is the noise PDF. As mentioned in Sect. 8.2, the objective is to estimate the source signal, $\mathbf{s}$, or the room impulse responses, $\mathbf{H}$. These are obtained from (8.7) using the *marginalisation operator*. By marginalising the RIRs, the source signal can be expressed directly, thus bypassing the estimation of the system response. The PDF of $\mathbf{s}$ is thus found by:

$$p(\mathbf{s} \mid \mathbf{x}) = \iint p(\mathbf{s}, \mathbf{H}, \theta \mid \mathbf{x})\, \mathrm{d}\mathbf{H}\, \mathrm{d}\theta, \tag{8.8a}$$

where the integrals are over all the elements of $\mathbf{H}$ and θ. If it is desired to obtain a source signal estimate by inverse-filtering the observations with the RIR, the source signal should be marginalised. The PDF of the room impulse response is thus found as:

$$p(\mathbf{H} \mid \mathbf{x}) = \iint p(\mathbf{s}, \mathbf{H}, \theta \mid \mathbf{x})\, \mathrm{d}\mathbf{s}\, \mathrm{d}\theta. \tag{8.8b}$$

In practice, the calculations involved in the marginalisation of either the source signal in (8.8a) or the channel response in (8.8b) are typically implicitly performed with appropriate dereverberation algorithms; there is little difference in the implementation of these marginalisation calculations. Moreover, the marginalisations are often performed numerically, as discussed in Sect. 8.2.3, so frequently the joint PDF, $p(\mathbf{s}, \mathbf{H}, \theta \mid \mathbf{x})$, of (8.7) is estimated.

8.2.2 Classification of Blind Dereverberation Formulations

The joint PDF in (8.7) of the source, channel, and model parameters, completely encapsulates the full system model shown in (8.1) and (8.3). Unfortunately, the length of the impulse responses and source are typically very long. Therefore, if the source signal, $\mathbf{s}$, and the channel, $\mathbf{H}$, are simply considered as unknown parameters, the dimension of the joint PDF will be extremely high. This will make estimation of the full parameter set difficult. However, some special cases and simplifications are considered, as follows:

Stochastic channel model The term $p_H(\mathbf{H}|\,\theta_h)$ in (8.7) allows for a stochastic channel model, inasmuch as the impulse response functions are still random processes given knowledge of the channel parameters, θ_h. While $\mathbf{H}$ is stochastic in nature given the parameters θ_h, often $p_H(\mathbf{H}|\,\theta_h)$ takes on a standard distribution, such as Gaussian, such that $\mathbf{H}$ is frequently amenable to the marginalisation in (8.8a). Some examples of stochastic channel models are discussed in Sect. 8.4.7.

Static parametric channel model If a static parametric model is used for the RIR, the channel model parameters, θ_h, completely determine $\mathbf{H}$. Hence, if $\mathbf{H} = \mathbf{G}(\theta_h)$ for some matrix $\mathbf{G}$ of functions, the channel PDF simplifies to $p_H(\mathbf{H}|\,\theta_h) = \delta(\mathbf{H} - \mathbf{G}(\theta_h))$. Therefore, Bayes's theorem in (8.7) reduces to:

$$p(\mathbf{s}, \theta \,|\, \mathbf{x}) \propto p_S(\mathbf{s}|\,\theta_s)\, p_V(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s}|\,\theta_V)\, p_\Theta(\theta), \tag{8.9}$$

where $\theta = \{\theta_h, \theta_s\}$ is the reduced parameter set. The observation likelihood in this expression, $p_V(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s}|\,\theta_V)$, is still determined by the observation noise PDF. However, since $p_V(\cdot|\,\theta_V)$ and $p_S(\mathbf{s}|\,\theta_s)$ are often Gaussian, it is straightforward to marginalise $\mathbf{s}$ in (8.8b):

$$p(\theta_s, \theta_h \,|\, \mathbf{x}) = \int p(\mathbf{s}, \theta \,|\, \mathbf{x})\, \mathrm{d}\mathbf{s}. \tag{8.10}$$

Unfortunately, such a marginalisation can then make removal of the nuisance parameters, θ_s, difficult. Static parametric channel models are discussed in detail in Sect. 8.4.6.

Zero observation noise with stochastic channel model In the case of no observation noise:

$$p_V(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s}|\,\theta_V) = \delta(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s}),$$

and so assuming a stochastic channel model, (8.7) simplifies to:

$$p(\mathbf{H}, \theta \,|\, \mathbf{x}) \propto p_S(\mathbf{s}|\,\theta_s)|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}\, p_H(\mathbf{H}|\,\theta_h)\, p_\Theta(\theta), \tag{8.11}$$

where the PDF $p(\mathbf{s}|\,\theta_s)|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}$ requires an appropriate probability transformation from $\mathbf{x}$ to $\mathbf{s}$ given θ_h to correctly determine its form.

Zero observation noise with static channel model Similarly, in the case of a static channel model and no observation noise, (8.9) simplifies to:

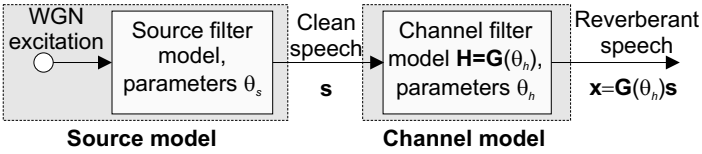


Fig. 8.2 General noiseless parametric model

$$p(\theta_s, \theta_h \mid \mathbf{x}) \propto p_S(\mathbf{s} \mid \theta_s)\big|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}\, p_\Theta(\theta_s, \theta_h). \tag{8.12}$$

Note that, in this context, the likelihood is $p_X(\mathbf{x} \mid \theta) = p_S(\mathbf{s} \mid \theta_s)\big|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}$. The interesting form of the simplified Bayes's expression in (8.12) is that the joint PDF is now just in terms of the model parameters. Therefore, assuming that the number of model parameters is substantially fewer than the length of the source signal and RIRs, this reduced parameter space should be simpler to estimate. Moreover, unlike the case in (8.10), the source model parameters, θ_s, can usually be marginalised, to leave the marginal PDF for the channel parameters:

$$p(\theta_h \mid \mathbf{x}) = \int p(\theta_s, \theta_h \mid \mathbf{x})\, \mathrm{d}\theta_s. \tag{8.13}$$

The optimal channel parameter, $\hat{\theta}_h$, estimates can then be used to recover the source signal from the reverberant observations using the relation $\mathbf{s} = \mathbf{G}^{-1}(\hat{\theta}_h)\mathbf{x}$. Figure 8.2 shows a graphical representation of the general parametric system model with zero observation noise.

8.2.3 Numerical Bayesian Methods

As discussed in Sect. 8.1.1, blind dereverberation can be approached either as an offline batch parameter estimation, or as an online optimal filtering problem. Offline estimation generally uses batch approaches such as MCMC methods, whereas online approaches use SMC methods.

8.2.3.1 Markov Chain Monte Carlo

In the batch approach, a Maximum Marginal *a Posteriori* (MMAP) estimate of the channel parameters is found by solving, for example, (8.13):

$$\hat{\theta}_{h,\mathrm{MMAP}} = \arg\max_{\theta_h} p(\theta_h \mid \mathbf{x}) = \arg\max_{\theta_h} \int p(\theta_s, \theta_h \mid \mathbf{x})\, \mathrm{d}\theta_s, \tag{8.14}$$

where $\mathbf{x}$ denotes all available data. The MMAP estimate, $\hat{\theta}_{h,\mathrm{MMAP}}$, is then used to inverse-filter the noise-free observed signal in (8.3) with the room transfer function

Algorithm 8.1 Generic two-component Gibbs sampler

for $i = 1, \ldots, I-1$ **do**
Sample $\theta_s^{(i+1)} \sim p\left(\theta_s \mid \theta_h^{(i)}, \mathbf{x}\right)$.
Sample $\theta_h^{(i+1)} \sim p\left(\theta_h \mid \theta_s^{(i+1)}, \mathbf{x}\right)$.
end for
Discard samples $\{\theta_s^{(i)}, \theta_h^{(i)}\}$ for $i = \{0, \ldots, I_{\text{burnin}} - 1\}$.
–
Note that the conditionals take the form:

$$p(\theta_s \mid \theta_h, \mathbf{x}) \propto p(\mathbf{x} \mid \theta_s, \theta_h)\, p(\theta_s), \tag{8.16a}$$
$$p(\theta_h \mid \theta_s, \mathbf{x}) \propto p(\mathbf{x} \mid \theta_s, \theta_h)\, p(\theta_h), \tag{8.16b}$$

where the measurement likelihood is given from (8.12) as:

$$p(\mathbf{x} \mid \theta_s, \theta_h) = p_S(\mathbf{s} \mid \theta_s)|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}. \tag{8.16c}$$

in order to reconstruct the speech signal:

$$\mathbf{s}_{\text{MMAP}} = \mathbf{G}^{-1}\left(\hat{\theta}_{h,\text{MMAP}}\right)\mathbf{x}. \tag{8.15}$$

Although deterministic optimisation methods could be used for directly determining the MMAP estimate, $\hat{\theta}_{h,\text{MMAP}}$, in practice it is difficult to find since the *a posteriori* PDF in (8.13) and (8.14) is usually multi-modal and subject to rapid fluctuation with variations in the parameter space. Instead, iterative stochastic sampling schemes can be used: MCMC methods can be utilised to sample from the joint PDF of the channel and source parameters, θ_h and θ_s, respectively. MCMC methods are based on constructing a Markov chain that has the desired distribution as its invariant distribution. Gibbs sampling [6, 9] is a MCMC method that approximates the joint PDF of the unknown model parameters by iteratively drawing random variates from the conditional densities in order to sample from their joint PDF. A generic form of a simple two-component Gibbs sampler is given in Algorithm 8.1. Independent of the initial distribution, the probabilities of the chain are guaranteed to converge to the invariant distribution, i.e., the joint PDF, after a sufficiently long burn-in period. A Minimum Mean Square Error (MMSE) estimate of the channel parameters is then obtained through numerical marginalisation of the nuisance parameters, which is achieved simply by computing the expected value of only the variates of interest:

$$\hat{\theta}_{h,\text{MMSE}} = \frac{1}{I - I_{\text{burnin}}} \sum_{i=I_{\text{burnin}}}^{I-1} \theta_h^{(i)}, \tag{8.17}$$

where $\theta_h^{(i)}$ are the samples drawn at iteration i, I is the total number of iterations and I_{burnin} is the number of samples discarded in the burn-in period. Often, it is assumed that the MMSE estimate of the channel parameters approximately corresponds to

Algorithm 8.2 Generic particle filter using importance sampling

for $n = 1, \ldots,$ number of samples **do**
  for $i = 1, \ldots,$ number of particles **do**
    Sample $\theta_n^{(i)} \sim \pi\left(\theta_n^{(i)} \,\middle|\, \mathbf{x}_{1:n}, \theta_{0:n-1}^{(i)}\right)$.
    Evaluate $w_n^{(i)} \propto \dfrac{p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_n^{(i)}\right) p\left(\theta_n^{(i)} \,\middle|\, \theta_{0:n-1}^{(i)}\right)}{\pi\left(\theta_n^{(i)} \,\middle|\, \mathbf{x}_{1:n}, \theta_{0:n-1}^{(i)}\right)}$.
  end for
  Normalisation of importance weights $w_n^{(i)} \rightarrow \frac{w_n^{(i)}}{\sum_i w_n^{(i)}}$.
  Resampling step (see, e.g., [40]).
end for

the MMAP channel estimate, $\hat{\theta}_{h,\text{MMSE}} \approx \hat{\theta}_{h,\text{MMAP}}$ [7]. An estimate of the source signal is then obtained by the inverse-filtering operation in (8.15).

8.2.3.2 Sequential Monte Carlo

SMC methods or Particle Filter (PF)s [40] facilitate direct estimation of the source signal, thus avoiding issues caused by inversion of non-minimum phase channels (see Sect. 8.3.3). It is desired to find the PDFs for the unknown signal states and parameters, $p(\mathbf{s}, \theta \,|\, \mathbf{x})$, for example, as given by (8.9), in a sequential online manner. Thus, the objective is to actually estimate, at time index n, $p(\mathbf{s}_{0:n}, \theta_{0:n} \,|\, \mathbf{x}_{0:n})$,[6] where $\theta \triangleq \{\theta_n\}$ is now assumed to consist of a sequence of parameters, and therefore $\theta_{0:n}$ is the sequence of parameters until time n. This posterior PDF is approximated at each time instance by a cloud of random variates, also called particles. Since the posterior PDF is usually difficult to sample from directly, these particles are drawn from an importance distribution, $\pi(\theta_n \,|\, \mathbf{x}_{1:n}, \theta_{0:n-1})$, which is straightforward to sample from. The resulting random variates are assigned weights to apportion their contribution to the empirical PDF appropriately. The posterior can then be updated on a per-sample basis by recursively updating the locations of the particles, and rejuvenating the particle cloud by resampling those particles that contribute most to the empirical PDF. The generic form of a particle filter is summarized in Algorithm 8.2. MMSE parameter estimates can be obtained from a sample mean of the particles, similar to (8.17). The aim is to obtain a direct estimate of the joint PDF of the source signal, and ideally as a byproduct, the model parameters.

8.2.3.3 General Comments

A comparison of online and offline methods is summarized in Table 8.1. One particular difference involves the inverse channel filtering implicitly used in the MCMC

[6] Note that in a sequential framework, the following notation is used to represent a sequence: $\mathbf{u}_{a:b} \triangleq \{u(a), u(a+1), \ldots, u(b)\}$.

Table 8.1 Comparison of online and offline methods

	Online	Offline
Method:	SMC	MCMC
Exploration by:	tracking/updating estimates	searching parameter space
Enhancement via:	direct source signal estimation	channel inversion
Results:	available in real-time	delayed
System model:	stochastic	static
Noise model:	flexible noise model	WGN or no noise
Estimated posterior PDF:	$p(\mathbf{s}_{0:n}, \boldsymbol{\theta}_{0:n} \mid \mathbf{x}_{1:n})$	model parameters (usually) $p(\boldsymbol{\theta} \mid \mathbf{x})$
Model advantages:	flexible system models	requires model selection

method [7], but avoided in the SMC approach since the latter estimates the source signal directly. As discussed in Sect. 8.3.3, channel inversion introduces several difficulties that can potentially increase the distortion in the enhanced signal. The discussion thus far has assumed that there is some *optimal estimate* of either the source signal, or model parameters. Since blind dereverberation is an inherently underdetermined problem, in that there are more unknowns than observations, this is a strong assumption. The choice of parametric models in, for example, Fig. 8.2, might lead to multiple modes in the joint PDF of (8.11) and (8.12), and therefore multiple *optimal solutions*. To ensure a unique solution, it is required to consider the system identifiability.

8.2.4 Identifiability

Single-channel blind dereverberation is an inherently under-determined problem. A characteristic of blind deconvolution is that the source signal and RIR must be *irreducible* for unambiguous deconvolution [24]. An irreducible signal is one in which the z-transform polynomial representation cannot be expressed as a product of at least two non-trivial factors over a given set.[7] This corresponds to saying that an irreducible signal is one that cannot be expressed as a time-invariant convolution of two or more signal components. Thus, a reducible signal, $h(n)$, is one which can be expressed as $h(n) = h_1(n) * h_2(n)$.

In the noiseless linear time-invariant case, as given by (8.4) with $v(n) = 0$, the observed signal may be expressed as $x(n) = h(n) * s(n)$. Hence, if $h(n)$ is *reducible* such that $h(n) = h_1(n) * h_2(n)$, the observed signal is given by $s(n) = h_1(n) * h_2(n) * s(n)$. Consequently, there are multiple solutions to the deconvolution problem, $\{\hat{h}(n), \hat{s}(n)\}$, as shown in Table 8.2. It is impossible to decide which of the solutions in Table 8.2 is the correct solution without additional knowledge.

[7] This is on the understanding that the delta function corresponds to a trivial factor, and is therefore not a signal component.

Table 8.2 Possible solutions, $\{\hat{h}(n), \hat{s}(n)\}$, to blind dereverberation of a stationary talker when the LTI channel, $h(t) = h_1(n) * h_2(n)$, is reducible

$\hat{h}(n)$	$\hat{s}(n)$
1	$h_1(n) * h_2(n) * s(n)$
$h_1(n)$	$h_2(n) * s(n)$
$h_2(n)$	$h_1(n) * s(n)$
$s(n)$	$h_1(n) * h_2(n)$
$h_1(n) * h_2(n)$	$s(n)$
$h_1(n) * s(n)$	$h_2(n)$
$h_2(n) * s(n)$	$h_1(n)$
$h_1(n) * h_2(n) * s(n)$	1

By realising that many linear systems are reducible when the signals are considered stationary and the system time-invariant, it is clear that *blind deconvolution* is impossible in such cases. If, however, $s(n)$ and $h(n)$ are quasi-stationary and quasi-time-invariant, respectively, then while the system is *locally reducible*, $s(n)$ and $h(n)$ are not *globally reducible*. This is provided that $s(n)$ and $h(n)$ possess different *rates* of *global* time-variation. In such a case, therefore, blind deconvolution is possible.

Several examples shall reiterate this point:

1. If, for example, the source is modelled as a stationary Autoregressive (AR) process and the channel as an LTI all-pole filter (see Sect. 8.4.3), the observed signal is also a stationary AR process. Consequently, it is not possible to attribute a particular pole estimated from the observed signal to either the source or channel; there is an identifiability ambiguity and the system is reducible. This source-channel ambiguity can be avoided by, for example, modelling the acoustic source as a Time-Varying AR (TVAR) process (see Sect. 8.6.2), and the channel by an LTI Finite Impulse Response (FIR) filter. The observed signal is then a Time-Varying ARMA (TVARMA) process, in which the poles belong to the source model and zeros to the channel; in this case, the system is *irreducible* given prior knowledge that the source has poles only, and the channel has zeros only. There appears to be no ambiguity in distinguishing between the parameters associated with each, and this model is used in [4] for the case of separating and recovering convolutively mixed signals. However, this TVAR-FIR source-channel model is of course not always realistic, as it cannot be ascertained that the source only has poles and no zeros, and the channel only has zeros and no poles.
2. In an alternative approach to single-channel blind dereverberation focusing on stationary talkers [21], the locally-stationary nature of the source and the *assumed* time-invariance of the channel are utilised to provide sufficient information to distinguish between the two models. In this approach it is argued that the statistics of speech signals remain quasi-stationary for around 20–50 ms. The source signal is modelled by a Block Stationary AR (BSAR) process (see Sect. 8.6.3), while the Acoustic Impulse Response (AIR) is modelled by an LTI

all-pole filter.[8] These models allow the AIR to be uniquely identified up to a scaling ambiguity, since essentially any common poles estimated from different blocks of the observed data must belong to the channel.

The issue of system identifiability is clearly determined by assumptions regarding the characteristics of the source signal and the acoustic impulse response. These characteristics must be appropriately reflected in the parametric models used, and it must be determined whether the proposed system model is identifiable. This, however, does not address the question of whether the underlying physical system is identifiable only from the observations. In blind dereverberation, this is an open question and readily in need of more investigation [34]. With these identifiability issues in mind, the following sections discuss appropriate channel (Sect. 8.4) and source models (Sect. 8.6).

8.3 Nature of Room Acoustics

The Bayesian paradigm suggests the use of either stochastic or static parametric channel models. This section considers the nature of room acoustics from a perspective relevant to the justification of commonly used models in blind dereverberation. The most general form of a room impulse response in continuous time, $h_{(\mathbf{q}_{\mathrm{src}}(\tau),\mathbf{q}_{\mathrm{mic}}(\tau))}(t)$, resulting from an impulse applied at time τ between a sound source and observer at positions $\mathbf{q}_{\mathrm{src}}(\tau)$ and $\mathbf{q}_{\mathrm{mic}}(\tau)$, respectively (see (8.1)), results from solving the acoustic wave equation. For clarity, the dependence on τ will subsequently be dropped, since τ is essentially characterised by the source-sensor geometry $(\mathbf{q}_{\mathrm{src}}, \mathbf{q}_{\mathrm{mic}})$. The solution is expressed in continuous-time as a linear combination of damped harmonics:

$$h_{(\mathbf{q}_{\mathrm{src}},\mathbf{q}_{\mathrm{mic}})}(t) = \begin{cases} 0 & \text{for } t < 0, \\ \sum_k \tilde{A}_k e^{-\tilde{\delta}_k t} \cos\left(\tilde{\omega}_k t + \tilde{\theta}_k\right) & \text{for } t \geq 0. \end{cases} \tag{8.18}$$

The amplitude coefficients, $\tilde{A}_k$, implicitly contain the locations of the source and sensor, $\mathbf{q}_{\mathrm{src}}$ and $\mathbf{q}_{\mathrm{mic}}$. On the other hand, the damping factors, $\tilde{\delta}_k$, corresponding to the quality-factor (Q-factor), the undamped natural frequencies, $\tilde{\omega}_k$, and phase terms, $\tilde{\theta}_k$, are *independent* of the source and receiver positions. Their values are determined by the room size, wall reflection coefficient, and room shape. While the general parametric model in (8.18) completely characterises the room impulse response, it is intractable for many estimation problems in signal processing and does not easily lead to an analytical solution in the Bayesian framework for blind

[8] In this chapter, the terms RIR and RTF specifically refer to any impulse response or transfer function, respectively, associated with room reverberation, whereas the terms acoustic impulse response and acoustic transfer function are used to refer to the response of an acoustic environment other than a room. In [21] and later in this chapter, results are presented for an acoustic gramophone horn, and therefore it is referred to by an acoustic rather than room response.

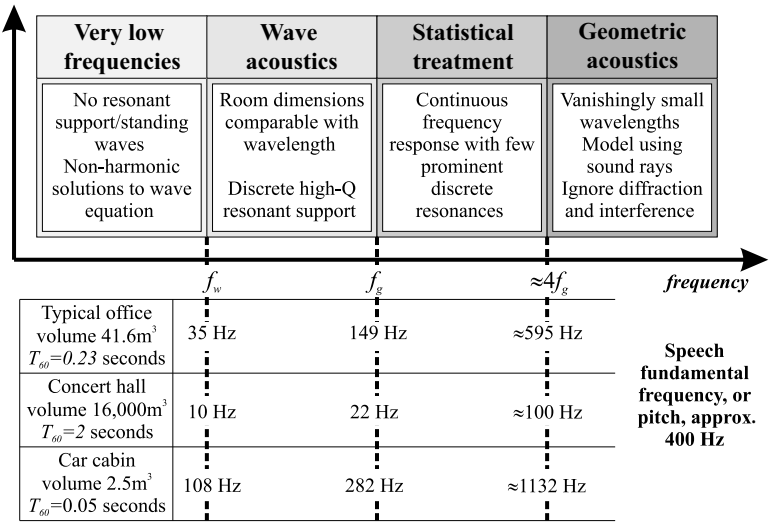


Fig. 8.3 Different regions of acoustic modelling

dereverberation. Even though numerical Bayesian methods (see Sect. 8.2.3) can be used to circumvent the lack of closed form solutions, (8.18) does not necessarily lead to a parsimonious representation, and therefore alternative models should be considered.

Moreover, while there are many other techniques for modelling an RIR, not all lend themselves to algorithms for straightforward parameter estimation. In general, each model applies to a different frequency range of the audible spectrum and, from a signal processing perspective, there is no single practical generative model for the entire audible frequency range [25].

8.3.1 Regions of the Audible Spectrum

Generally, the audible spectrum can be divided into four distinct regions, as summarised in Fig. 8.3. In the following, consider a typical shoebox shaped office environment with dimensions $2.78 \times 4.68 \times 3.2\,\text{m}$, volume $V = 41.6\,\text{m}^3$, and reverberation time of $T_{60} = 0.23$ seconds. This room is denoted by $\mathcal{R}$. A single-tone source of frequency f is assumed in the discussion, with the argument extending to wideband sources by using linear superposition.

Very Low Frequencies and Wave Acoustics At very low frequencies, $f < f_w = \frac{c}{2L}$, where c is the speed of sound, and L is the largest dimension of the acoustic environment, there is no resonant support. Typically, f_w is around 35 Hz for room $\mathcal{R}$. The so called wave-acoustics region corresponds to frequencies where the source wavelength is comparable to the room dimensions. It spans the lowest resonant

mode, approximately given by f_w, to the Schroeder frequency $f_g \approx 2000\sqrt{T_{60}/V}$ (Hz). Distinct resonants occur in which the Q-factor is sufficiently large that the average spacing of resonant frequencies is substantially larger than the average *half-width* of the resonant mode. For this room, distinct resonances occur between $f_w = 35$ Hz and $f_g = 149$ Hz.

In practice, however, the very low frequency and wave acoustic regions are generally irrelevant for speech dereverberation since electro-acoustic systems have a limited bandwidth at low frequencies. Analytical tools are thus utilised only for the high sound frequency and geometric acoustic regions.

High Sound Frequencies and Geometric Acoustics Above f_g, there is such a strong model overlap that the concept of a resonant mode becomes meaningless. However, below a frequency of around $4f_g$, the wavelengths are too long for the application of *geometric acoustics*. Thus, in this *transition region*, a statistical treatment is generally employed. For the room above, statistical theory is relevant from $f_g = 149$ Hz to $4f_g = 595$ Hz.

Above $4f_g$, *geometrical room acoustics* applies and assumes the limiting case of vanishingly small wavelengths. This assumption is valid if the dimensions of the room and its walls are large compared with the wavelength of sound: this condition is met for a wide-range of audio frequencies in standard rooms. In this frequency range, specular reflections and the *sound ray* approach to acoustics prevail. Geometrical acoustics usually neglect wave related effects such as diffraction and interference. The image method [1] for simulated AIRs is valid only in this frequency range.

8.3.2 The Room Transfer Function

Parametric modelling is often justified by considering the Room Transfer Function (RTF) between a sound source in an *enclosed space* and a receiver, rather than the time-domain representation in (8.18). The RTF is derived directly from (8.18) by taking Laplace transforms as:

$$H_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(s) = \sum_{k\in\mathcal{K}} \frac{\alpha_k + \beta_k s}{\tilde{\omega}_k^2 + (\tilde{\delta}_k + s)^2} \equiv \prod_{k\in\mathcal{K}} \frac{D_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(s)}{(s - s_k)(s + s_k)}, \tag{8.19}$$

where ω is angular frequency, $s_k = -\tilde{\delta}_k + j\tilde{\omega}_k$, the constants $\{\alpha_k, \beta_k\}$ and the polynomial $D_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(s)$ are functions of $\{\tilde{A}_k, \tilde{\delta}_k, \tilde{\theta}_k\}$ and consequently dependent on the source-sensor geometry.[9] Thus, the frequency response is:

$$H_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(j\omega) = \sum_{k\in\mathcal{K}} \frac{\alpha_k + j\beta_k\omega}{\tilde{\omega}_k^2 + \tilde{\delta}_k^2 - 2j\tilde{\delta}_k\omega - \omega^2}. \tag{8.20}$$

[9] It is easily shown that $\alpha_k = \tilde{A}_k\left(\tilde{\delta}_k\cos\tilde{\theta}_k - \tilde{\omega}_k\sin\tilde{\theta}_k\right)$ and $\beta_k = \tilde{A}_k\tilde{\omega}_k\cos\tilde{\theta}_k$.

When $\omega \approx \tilde{\omega}_k$, the associated term in (8.20) assumes a high absolute value. As such, $\tilde{\omega}_k$ is sometimes called an *eigenfrequency* of the room [25], or a *resonant frequency* due to the resonances occurring in the vicinity of $\tilde{\omega}_k$.

8.3.3 Issues with Modelling Room Transfer Functions

Audio signal processing in acoustic environments is a notoriously difficult and challenging field, and blind dereverberation is no exception. The difficulty arises due to the complexity of the room acoustics. There are a number of problems encountered in this application when dealing with AIRs, such as in (8.18), and RTFs of (8.19) [34].

Long and Non-minimum Phase AIRs

In general, RIRs are long and, for instance, a Finite Impulse Response (FIR) implementation would typically require $n_s = T_{60} f_\mathrm{s}$ coefficients, where f_s is the sampling frequency. For example, if $T_{60} = 0.5$ s and $f_\mathrm{s} = 10$ kHz, the length of the RIR is around $n_s = 5000$ coefficients. This can render modelling and parameter estimation difficult. Moreover, RIRs are often non-minimum phase, leading to difficulties with channel modelling and inversion. The non-minimum phase contribution to the perception of reverberation is significant [22, 33].

Robustness to Estimation Error and Variation of Inverse of the AIR

Any small error in an RIR estimate leads to a significant error in the inverse of the RIR. Thus, inversion can increase distortion in the enhanced signal compared to the reverberant signal. Any deviation from the true RIR means that attempts to equalise high-Q resonances can still leave high-Q resonances in the equalised response degrading the intelligibility of the restored signal. Similarly, a small change in source-sensor geometry might give rise to a small change in the RIR, so again the corresponding changes in the inverse of an RIR can sometimes be large.

Subband and Frequency-zooming Solutions

Since the proposed channel estimation techniques and source recovery methods discussed in this chapter implicitly use inverse-filtering methods, these issues are particularly pertinent. Some of these problems cannot be alleviated by either attempting to process the full frequency range of the source, nor by attempting to invert the *full-band* RTF using a single filter. In problems with long channels, it is better to utilise subband methods that attempt to enhance the reverberant signal by invert-

ing the channel response over a number of separate frequency ranges. Modelling each frequency band independently can lead to a parsimonious approximation of the RTF, lower model orders, and an overall reduction in the total number of parameters needed to approximate the acoustic channel. Moreover, there may be only a few bands that have high-Q resonances, which need careful equalisation, whereas other frequency bands have lower-Q factors, so less care is required.

An additional advantage of using subband models is that subbands possessing minimum phase characteristics can be inverted, despite the AIRs being non-minimum phase over the full frequency range. Hence, in the case of a non-minimum phase response, where a causal inverse does not exist, methods for detecting and equalising the minimum phase subbands should be developed: this follows the approaches in [45, 46]. Details of a subband all-pole model and methodology are discussed in Sect. 8.4.4.

8.4 Parametric Channel Models

This section discusses a variety of parametric models, both static and stochastic, that can be used tractably within a Bayesian framework. Rational parametric models are introduced, but it is important to note that it is the characteristic of the model parameters that determines whether the model is static or stochastic; this is discussed in Sect. 8.4.5.

8.4.1 Pole-zero and All-zero Models

The RTF in (8.19) is rational and can therefore, in principle, be modelled by a conventional pole-zero model [30]. From a physical point of view, poles represent resonances, and zeros represent time delays and anti-resonances. Two common simplifications of (8.19) are the all-zero and all-pole models, each with their own advantages and disadvantages.

There are several main limitations imposed by the nature of room acoustics of the resulting FIR filters given by all-zero models [29, 30]. Firstly, as discussed in Sect. 8.3.3, RIRs are, in general, very long and an all-zero filter typically requires as many taps as the length of the RIR. Secondly, the resulting FIR filter may be effective only for a limited spatial combination of source and receiver positions, $(\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$, as all-zero models lead to large variations in the RTF for small changes in source-observer positions [29, 30]. A further disadvantage of the pole-zero and all-zero models for the *single channel case* is that estimation of the zeros requires solving a set of non-linear equations.

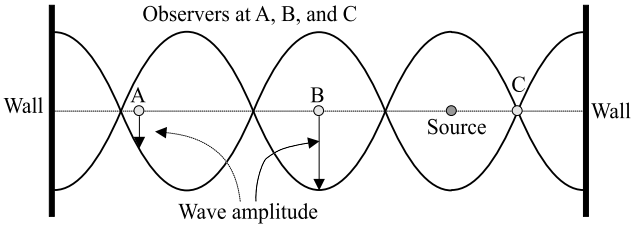


Fig. 8.4 Resonant standing waves for a 1-D room can be observed at any point except node points, such as point *C*. Since this standing wave occurs independently of the source location and can be observed at all observation points, the acoustical poles that reflect the information of the resonant frequencies are independent of source-sensor locations

8.4.2 The Common-acoustical Pole and Zero Model

The poles of the room transfer function on the right-hand side of (8.19) are functions of the damping factors, $\tilde{\delta}_k$, and undamped natural frequencies, $\tilde{\omega}_k$, and are, therefore, approximately independent of the source and sensor positions $(\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$. Consequently, the poles encapsulate all the information pertaining to the resonants of a room; *standing waves* occur independently of the source location and can be observed at any point in the room, except at node points, as depicted in the 1-D case shown in Fig. 8.4. Naturally, the amplitude of the standing wave varies depending on the sensor positions, as seen in Fig. 8.4, and this variation is reflected in the zeros of the RTF [14]. This leads to the Common-Acoustical Pole and Zero (CAPZ) model of an RTF, which was first introduced by Haneda *et al.* [13, 14]. It should be noted the acoustical argument used above for the justification of the CAPZ model is simplistic, and other investigations on the fluctuations of AIRs within reverberant environments suggest that this assumption may not be strictly true [34].

Nevertheless, the CAPZ model is particularly useful in applications where multiple room transfer functions from different source-observer positions are modelled, which could have applications in, for example, multi-channel blind source separation [10], or blind dereverberation from a moving talker. Like the general pole-zero model, the CAPZ model still suffers from the problem that it is not possible to write an input–output equation that is Linear-In-The-Parameters (LITP), which thereby complicates parameter estimation.

8.4.3 The All-pole Model

An LITP model that lends itself to straightforward parameter estimation is the all-pole model, which is widely used in many fields to approximate rational transfer functions. In discrete-time, its transfer function is given by:

$$H_{\mathbf{q}}(z) = G_{\mathbf{q}} \prod_{k \in \mathcal{P}} \frac{1}{1 - p_{\mathbf{q},k} z^{-1}} \equiv \frac{G_{\mathbf{q}}}{1 + \sum_{k \in \mathcal{P}} a_{\mathbf{q},k} z^{-k}}, \tag{8.21}$$

where $\mathbf{q} = (\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$ is the set of source and sensor positions, $G_{\mathbf{q}}$ is a gain term, $\{p_{\mathbf{q},k}\}_{k=1}^{P}$ denote the P poles, and $\{a_{\mathbf{q},k}\}_{k=1}^{P}$ denote the P all-pole parameters. It is claimed that typical all-pole model orders required for approximating RIRs with reverberation times $T_{60} \approx 0.5$ s are in the range $50 \leq P \leq 500$ [30], although this depends on the frequency range of the acoustic spectrum considered. In fact, practical experience seems to indicate this is a relatively conservative estimate, although it obviously depends on how much data is available for model order estimation. Mourjopoulos and Paraskevas [30] conclude that in many signal processing applications dealing with room acoustics, it may be both sufficient and more efficient to manipulate all-pole model coefficients rather than high order all-zero models. All-pole models are particularly useful for modelling resonances in the wave acoustics and high sound frequency regions.

Despite the dependence of the model parameters on the source-sensor positions, $\mathbf{q} = (\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$, a purported advantage of the all-pole over the all-zero model is its lower sensitivity to changes in $\mathbf{q}$ [30]. While the CAPZ model contributes to this argument, it is still the case that a subset of poles in the all-pole model must account for the variations in the RTF with source-sensor geometry, even if it is less sensitive than the all-zero model.

In the time-domain, suppose a signal, $s(n)$, is filtered through a room impulse response between a source position that varies as a function of time, $\mathbf{q}_{\text{src}}(n)$, and a fixed observation position $\mathbf{q}_{\text{mic}}$. As the source-sensor geometry varies as a function of time, the parameters that define the RIR also vary as a function of time. If the acoustic channel is modelled by an all-pole filter of order P, the observed signal, $x(n)$, received at the sensor, is expressed as

$$x(n) = -\sum_{k=1}^{P} a_k(n) x(n-k) + s(n), \tag{8.22}$$

where the all-pole coefficients, $\{a_{\mathbf{q},k}\}_{k=1}^{P}$, are now considered as functions of time and are denoted by $\{a_k(n)\}_{k=1}^{P}$. The nature of the parameter variations is discussed in Sect. 8.4.5.

8.4.4 Subband All-pole Modelling

The all-pole model in Sect. 8.4.3 will be referred to as the *full-band all-pole model*, since it essentially attempts to fit the entire frequency range simultaneously. The full-band all-pole model can result in a high number of parameters, the estimation of which will require a large computational load that can be unacceptable in computationally intensive algorithms such as blind dereverberation. The modelling of complicated room transfer functions requires a highly flexible and scalable para-

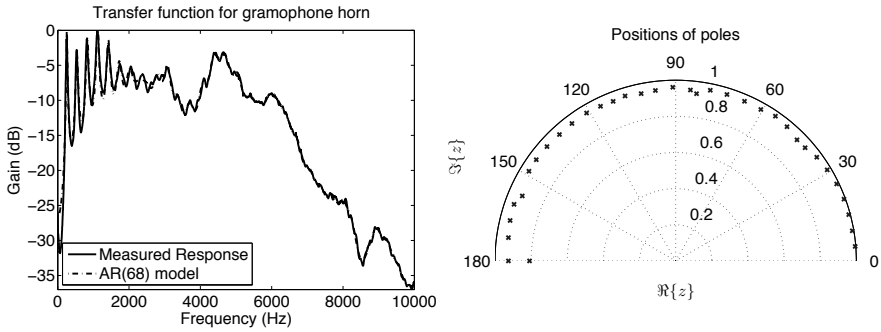


Fig. 8.5 (a) Transfer function of an acoustic gramophone horn [41] with the corresponding AR model and (b) poles corresponding to response in (a). The unit semi-circle maps to frequency range $0 \rightarrow 10$ kHz

metric model. As discussed in Sect. 8.3.3, a subband approach can resolve a number of modelling issues.

An intuitive rationale for why high model orders result in the full-band all-pole model is as follows: consider a transfer function that is highly resonant in a low frequency band, and much less resonant in a higher band, as shown in Fig. 8.5(a). Spencer [41] shows that this response can be accurately modelled by an all-pole model with 68 parameters. As shown in Fig. 8.5(b), these poles seem uniformly distributed around the edge of the unit circle. In the low frequency band, up to approximately 2 kHz, there are a number of closely spaced high-Q resonances; these can be modelled using approximately 12 poles. The response due to each pole-pair rolls-off at 40 dB per decade. Since the low-frequency poles are closely spaced with high spectral peaks, a large number of poles are needed at high-frequencies to counteract the roll-off effect of having a large number of low-frequency high-Q poles, while simultaneously attempting to model a relatively smooth frequency response. Thus, in essence, the full-band channel model requires many parameters because it attempts to fit the entire frequency range simultaneously, even though it may fit some regions in the frequency space better than others. Consequently, it is preferable to simply model a particular frequency band of the acoustic channel's spectrum by an all-pole filter, leading to lower model orders. Subband linear prediction was first considered in [27] and developed in [16–20, 38, 43]. The so-called *unconstrained subband all-pole model* is discussed, which attempts to fit different frequency bands independently, leading to a parsimonious approximation of the rational transfer functions and lower model orders. It is shown in [20] that the response in Fig. 8.5 (a), when using three subbands, can be modelled using just 51 parameters: a 25% reduction in parameters.

The subband all-pole model is more flexible for channel modelling than a single full-band. Makhoul [27] suggests a similar model when analysing speech using

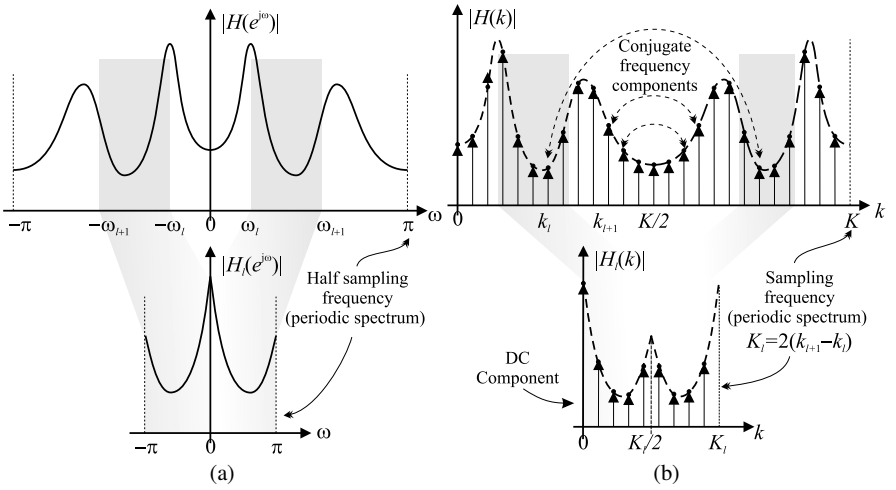


Fig. 8.6 Subband modelling – (a) continuous spectrum, (b) discrete spectrum and indices mapping

linear prediction. Consider a discrete-time representation of the system with B subbands; in subband $b \in \mathcal{B}$ the frequency response of the RTF, $H_{\mathbf{q}}\left(e^{j\omega}\right)$, is modelled by an all-pole spectrum in the region $[\omega_b, \omega_{b+1})$ obtained from (8.21) through the mapping graphically shown in Fig. 8.6(a):

$$\omega \rightarrow \pi \frac{\omega - \omega_b}{\omega_{b+1} - \omega_b}. \tag{8.23}$$

Thus, in the b^{th} subband, the mapped frequency response is given by:

$$H_{\mathbf{q}}^{(b)}\left(e^{j\omega}\right) = \frac{G_b}{1 + \sum\limits_{k \in \mathcal{P}_b} a_{b,k}\, e^{-j\omega k}}, \quad \omega \in [-\pi, \pi),$$

where $\mathbf{a}_b = \{a_{b,k}\}_{k=1}^{P_b}$ and $G_b \in \mathbb{R}^+$ denote model parameters in subband b. These parameters are implicitly conditional on $\mathbf{q} = (\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$, although this dependence has been dropped for clarity. The gain term, G_b, allows a further degree of freedom in the model, although to avoid scaling ambiguities, $G_0 \triangleq 1$. Hence, the total RTF is modelled for $\omega \in [-\pi, \pi)$ as:[10]

[10] Since the energy in subband b must be equivalent to the energy in the mapped frequency response, the scaling term γ_b in (8.24) is required:

$$\int_{\omega_b}^{\omega_{b+1}} \left|H_{\mathbf{q}}\left(e^{j\omega}\right)\right|^2 \mathrm{d}\omega = \frac{\omega_{b+1} - \omega_b}{\pi} \int_0^{\pi} \left|H_{\mathbf{q}}\left(e^{j\pi \frac{\omega - \omega_b}{\omega_{b+1} - \omega_b}}\right)\right|^2 \mathrm{d}\omega.$$

$$H_{\mathbf{q}}\left(e^{j\omega}\right) = \sum_{b=1}^{B} \underbrace{\left(\frac{\omega_{b+1}-\omega_b}{\pi}\right)^{\frac{1}{2}}}_{\gamma_b} H_{\mathbf{q}}^{(b)}\left(e^{j\pi\frac{\omega-\omega_b}{\omega_{b+1}-\omega_b}}\right)\mathbb{I}_{[\omega_b,\,\omega_{b+1})}(\omega), \tag{8.24}$$

where the indicator function is defined as $\mathbb{I}_{\mathcal{A}}(a) = 1$ if $a \in \mathcal{A}$ and zero otherwise. When the spectrum is sampled, the mapping in (8.23) is adjusted accordingly as indicated graphically in Fig. 8.6(b). Thus, each subband $b \in \mathcal{B}$ covers a total of $K_b = 2(k_{b+1} - k_b)$ frequency bins, namely $k \in \{k_b, \ldots, k_{b+1} - 1\}$ and the corresponding complementary frequency bins (see Fig. 8.6(b)). The subband boundaries are defined by $\{k_b, b \in \mathcal{B}\}$, with $k_0 \triangleq 0$ and $k_B \triangleq K$, where K is the total number of frequency bins. The frequency bin closest to the half sampling frequency is given by $k_{f_{s/2}} = \lfloor K/2 \rfloor$. The transfer function in a particular subband is obtained using the mapping $k \rightarrow \frac{k-k_b}{K_b}$ for $kK \leq 2$. This results in a sampled transfer function that is essentially identical to (8.24) with ω_b replaced by k_b.

A significant problem with this subband model as presented, however, is that the transfer function being modelled in each subband is no longer smooth, as indicated in the magnitude responses shown in Fig. 8.6(a). Moreover, due to the asymmetry of the phases, the subband phase response will be discontinuous and non-zero at the boundaries. Yet, the phase response of the subband all-pole model at the subband boundaries is zero. Techniques for dealing with this phase modelling problem are discussed in [19]. Despite this, the subband model is assumed throughout the rest of this chapter in order to reduce the complexity of the channel model.

8.4.5 The Nature of Time-varying All-pole Models

As argued in Sect. 8.4.3, a time-varying source-sensor geometry leads to a Time-Varying All-Pole (TVAP) model, as defined by (8.22). The subband all-pole model discussed in Sect. 8.4.4 is used in practice to model the complete RTF, and therefore discussions henceforth apply to a limited spectral region.

Consider again the interpretation of (8.22). While the poles in the CAPZ model discussed in Sect. 8.4.2 are invariant to changes in source-sensor positions, some of the poles in the all-pole model of (8.22) are not. The problem of modelling the RIR between a spatially varying source and sensor reduces to determining an appropriate model for the time-varying all-pole parameters, $\{a_k(n)\}_{k=1}^{P}$. Determining such a model is complicated, in part an open question, and is often constrained by the availability of suitable and tractable parameter estimation techniques. Appropriate models are discussed in Sects. 8.4.6 and 8.4.7. In the meantime, the spatially-varying nature of RIRs and the variation of the all-pole model parameters with spatial position is investigated. Simulated and measured RIRs are obtained for the acoustic set-up illustrated in Fig. 8.7 for a small office of size $2.78 \times 4.68 \times 3.2\,\mathrm{m}$ (length $\times$ width $\times$ height); this room matches room $\mathcal{R}$ discussed in Sect. 8.3.1. An acoustic source remains fixed while the microphone sensor is moved in 2 mm increments.

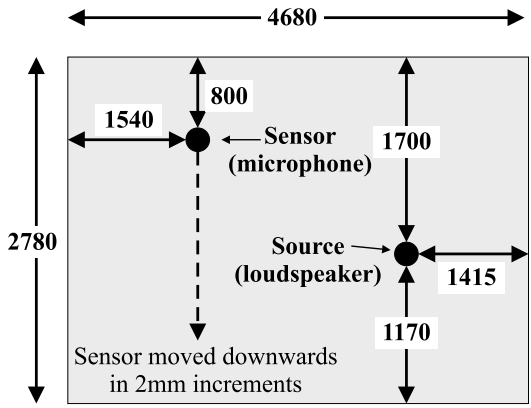


Fig. 8.7 Source and sensor locations in experimental set-up; all measurements in millimeters. Source and sensor elevation is 845 mm, room height of 3200 mm. The sensor is moved from its initial position in 2 mm increments

This experimental set-up mimics the spatially-varying nature of the RIR for moving sources.

The simulated RIRs are generated using the image method [1] with the reflection coefficient chosen to give a reverberation time of $T_{60} = 0.23$ s. This choice corresponds to the measured reverberation time of the real office. As the image model assumes geometric room acoustics, the simulated responses only apply above four times the Schroeder frequency, f_g, as discussed in Sect. 8.3.1, and in this case $4f_g = 595$ Hz. Using the simulated RIRs, the RTF is modelled in the frequency range between 600 to 1200 Hz by a 16^{th}-order subband all-pole model as discussed in Sect. 8.4.4. The variation of the resulting pole positions from the initial sensor position to a final offset of 400 mm is plotted in Fig. 8.8(a). The results indicate smooth pole variation and, consequently, the TVAP parameters of the RIR vary relatively smoothly with sensor spatial displacement. This can be confirmed by measures of the changes in the RIR, e.g., normalised projection misalignment.

For verification of these results using real data, 910 RIRs were measured in a real office by moving a 26-microphone linear array in small increments over a distance of 70 mm. To obtain comparable results to the simulated data, the pole variations are again acquired by modelling the RTF as a 16^{th}-order subband Autoregressive (AR) model in the range 600 to 1200 Hz. The poles for real RIRs are subject to larger variation than those for the simulated RIRs; they cover a wider region within the unit circle, and intersect the trajectories of neighbouring poles. To avoid cluttered pole trajectory plots, only a subset of the pole variations from the microphone array for several microphones (labelled mics. 7 and 8) are displayed in Figs. 8.8(c) and 8.8(d). This corresponds to offsets from 432 to 502 mm for microphone 7 and from 504 to 574 mm for microphone 8. For comparison with equivalent results for simulated data, see Fig. 8.8(b). The pole variations from the measured data clearly exhibit reasonably smooth trajectories, validating the simulated results.

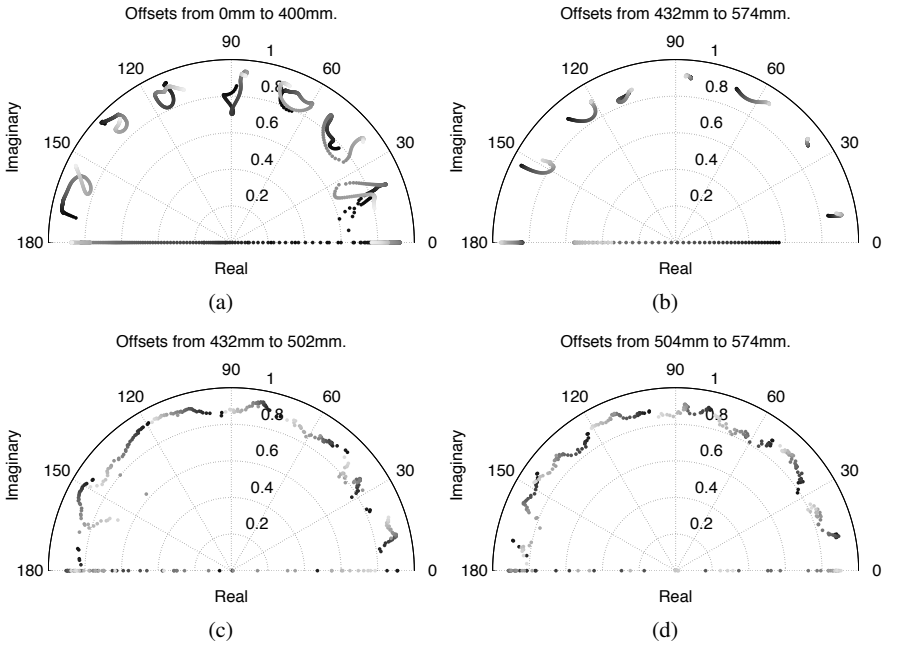


Fig. 8.8 Simulated and experimental results for spatiotemporal variation of the poles in all-pole modelling of RIRs; pole trajectories illustrated through colour map from *black* (starting point) to *light grey* (ending point). Model order: 16. (a) Simulated: $0 \rightarrow 400$ mm. (b) Simulated: $432 \rightarrow 574$ mm. (c) Measured: $432 \rightarrow 502$ mm (d) Measured: $504 \rightarrow 574$ mm

An in-depth discussion of the variability of room acoustics is beyond the scope of this chapter and requires considerably more investigation than the results presented in this section. Nonetheless, the results presented in Fig. 8.8 give useful insight into the possibilities for modelling the parameters $\{a_k(n)\}_{k=1}^{P}$ of the TVAP model in (8.22).

8.4.6 Static Modelling of TVAP Parameters

The smooth variations of the poles with changing position in Fig. 8.8 suggests that a suitable static model of the TVAP model parameters in (8.22) could be a deterministic function with unknown but fixed parameters. Such a function could be decomposed as a linear combination of basis functions. A similar decomposition will be used for modelling speech and this is discussed in Sect. 8.6.3 (see also (8.29) and (8.31)). Hence, the TVAP are modelled as:

$$a_p(n) = \sum_{k \in \mathcal{G}} a_{p,k} g_k(n-p), \tag{8.25}$$

where $\{a_{p,k}, p \in \mathcal{P}, k \in \mathcal{G}\}$ are the G unknown *static* time-invariant basis coefficients, $\{g_k(n)\}_{k \in \mathcal{G}}$ are the known time-varying basis functions. Note this model is assumed to apply over the full length of the source signal.

As the basis functions span the vector space to which the underlying time-varying all-pole parameters are mapped, they define the scope of their variation. Thus, their choice is essential. Unfortunately, no general rules for choosing these functions exist. The choice of basis is therefore dependent on the prior belief of the variation of the parameters. Amongst the wide range of basis functions that have been investigated [3, 11, 12, 39], standard choices include Fourier functions, Legendre polynomials and discrete prolate spheroidal sequences. These classes tend to assume smooth parameter behaviour and respond to abrupt changes as a low-pass filter [12]. Hence, for abrupt changes in the RIR with position (and therefore time), the parameters are not modelled correctly. A discontinuous basis like the step function can capture abrupt changes well, but cannot handle smooth variations [12]. Modelling rapid parameter variation is theoretically possible by utilising an infinite number of basis functions. However, this leads to over-parameterised coefficients since the model would have as many degrees of freedom as the RIR itself [12, 36].

8.4.7 Stochastic Modelling of Acoustic Channels

It might be argued that the variation of poles in Fig. 8.8, and therefore the corresponding parameters, is more stochastic in nature than a smooth predictable deterministic function. The simplest stochastic model for the TVAP parameters is the random walk:

$$a_p(n) = a_p(n-1) + w_{a_p}(n), \quad w_{a_p}(n) \sim \mathcal{N}\left(0, \sigma^2_{a_p}\right),$$

where $w_{a_p}(n)$ is a WGN process. In actuality, the TVAP coefficients are likely to be composed of a predictable deterministic variation or trend, which can be modelled by a linear combination of basis functions, and an unpredictable stochastic element that might be modelled by a random walk.

Alternatively, and inspired by models used for communication channels in the literature, it might be that the coefficients of the RIR in (8.1) are themselves modelled as a random walk:

$$h(n,\ell) \triangleq h_{\mathbf{q}(\ell)}(n) = h_{\mathbf{q}(\ell)}(n-1) + w_h(n),$$

where again, $\mathbf{q}(\ell) = (\mathbf{q}_{\text{src}}(\ell), \mathbf{q}_{\text{mic}}(\ell))$ denotes the source-sensor geometry, and $w_h(n)$ is WGN with variance σ^2_h. Perhaps a more structured approach is to model the RIR, $h_{\mathbf{q}}(n)$, as the product of a WGN process and a damping exponential decay as described in (2.28) in Chap. 2. Despite the fact that the process in (2.28) is

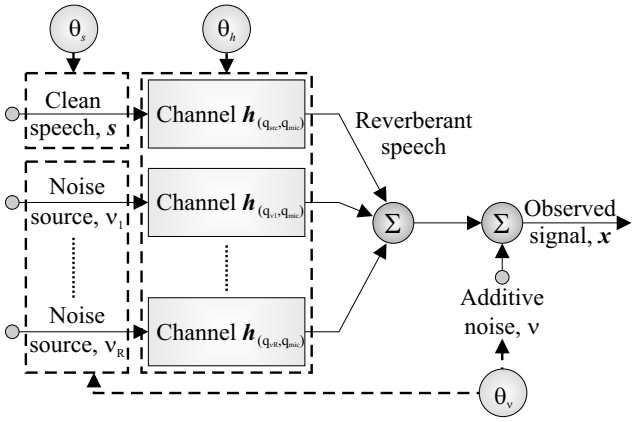


Fig. 8.9 Clean speech in a reverberant environment with remote noise signals

stochastic in nature given the variance of the WGN process and the damping factor, it is amenable to marginalisation due to its simple structure and the small parameter space (see the discussion in Sect. 8.2.2). These models are yet to receive substantial attention in the research literature, but have good potential for online or sequential algorithms (see Sect. 8.2.1). In the rest of this chapter, the static parametric model of Sect. 8.4.6 is used.

8.5 Noise and System Model

In the general problem formulation of Sect. 8.2, the noise was modelled as an additive measurement error at the microphone, as shown in Fig. 8.1. This was based on the argument that the observation noise is the superposition of all undesired sound sources in the room and therefore, by a central limit theorem argument, it will be WGN and unaffected by the room acoustics.

However, it is equally valid to argue that the underlying sources of noise arise from distinct localised positions; for example, the humming of computer fans, air conditioning units, or general distant traffic noise. Consider, then, the more general model shown in Fig. 8.9 in which spatially separated noise sources are each observed after they have propagated through the acoustic system; each noise source-sensor path has a distinct room impulse response. The receiver thus observes a noise contribution that is the linear combination of noise source signals filtered by separate channels due to the different AIRs associated with each noise-sensor geometry. Assuming that the noise sources are spatially-stationary, the model in (8.1) is written as:

$$x(n) = h_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n,\ell) * s(n) + \sum_{r=1}^{R} h_{(\mathbf{q}_{\nu r},\mathbf{q}_{\text{mic}})}(n) * \nu_r(n) + \nu(n), \tag{8.26}$$

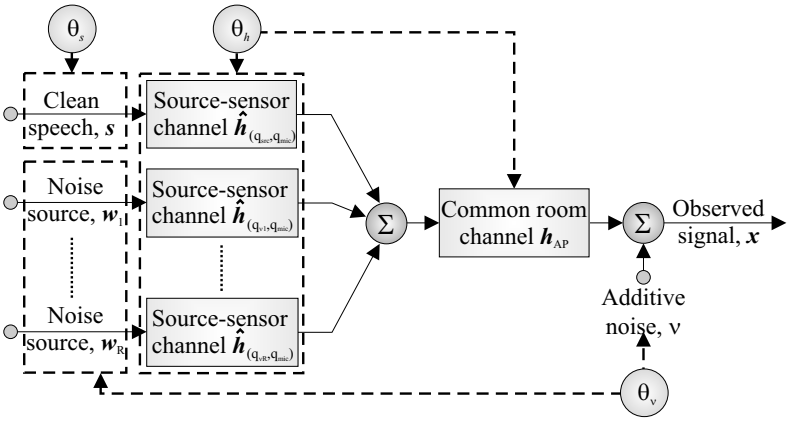


Fig. 8.10 Clean speech with remote noise signals in a reverberant environment which can be decomposed using the CAPZ model

where $h_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n,\ell)$ is the source-sensor RIR, $h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\text{mic}})}(n)$ is the RIR between the r^{th} noise source, $\nu_r(n)$, and the sensor, R denotes the number of noise sources, and $*$ represents either LTV or LTI convolution depending on the context. Although such a noise model is idealistic, it is also overly complicated, making it difficult to estimate all the relevant system parameters. Moreover, due to the lack of knowledge of the noise statistics, it might also be over-determined. Nevertheless, it is interesting to note that the model in Fig. 8.9 can be simplified by using the notion of common-acoustical poles as described in Sect. 8.4.2. Recall that each individual channel response can be decomposed into a combination of two components: one that is dependent on the source-sensor geometry, and one that is acoustically common to all source-sensor arrangements [14]. Using the CAPZ model, each RIR can be decomposed into a path-independent all-pole model, $h_{AP}(n)$, and a path-dependent pole-zero model, as shown in Fig. 8.10. Hence, (8.26) may be rewritten as:

$$x(n) = \left\{ \hat{h}_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n,\ell) * s(n) + \sum_{r=1}^{R} \hat{h}_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\text{mic}})}(n) * \nu_r(n) \right\} * h_{AP}(n) + \nu(n), \quad (8.27)$$

where $h_{\mathbf{q}}(n,\ell) = \hat{h}_{\mathbf{q}}(n,\ell) * h_{AP}(n)$. The modified coloured noise term

$$\nu_d(n) = \sum_{r=1}^{R} \hat{h}_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\text{mic}})}(n) * \nu_r(n)$$

is extremely difficult to model, and it can be argued that since $\nu_r(n)$ has undergone less filtering through $\hat{h}_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\text{mic}})}(n)$ than through $h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\text{mic}})}(n)$, $\nu_d(n)$ will be more Gaussian than $\sum_r h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\text{mic}})}(n) * \nu_r(n)$. Hence, $\nu_d(n)$ is modelled as WGN such that the overall model reduces to that shown in Fig. 8.11, and (8.27) reduces further to:

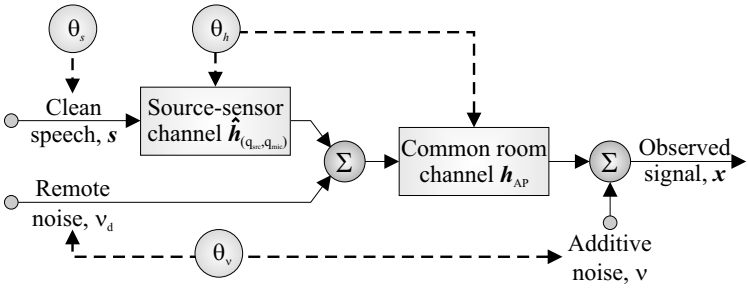


Fig. 8.11 Noise model simplification using CAPZ model

$$x(n) = \left\{\hat{h}_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n,\ell) * s(n) + \nu_d(n)\right\} * h_{AP}(n) + \nu(n), \tag{8.28}$$

where $\nu_d(n) \sim \mathcal{N}\left(0, \sigma^2_{\nu_d}\right)$ is the distant or remote WGN source. Moreover, it is possible to omit the observation or measurement noise term $\nu(n)$ by essentially combining it with $\nu_d(n)$ to obtain an even more simplified model. In essence, the model in (8.28) states that any remote noise sources that are affected by reverberation should not be modelled as white, but rather as WGN filtered by a common component of the room acoustics. It turns out that the shifting of the position of this noise term can help simplify the methodology used for source estimation, as described in Sect. 8.7.2.

8.6 Source Model

8.6.1 Speech Production

Speech sounds can be divided into three classes depending on the mode of excitation [32]. *Voiced sounds* are produced by vibrating vocal cords producing a periodic series of glottal pulses. The sound is quasi-periodic with a spectrum of rich harmonics at multiples of the fundamental or pitch frequency, f_0, as shown in Fig. 8.12. *Unvoiced sounds*, on the other hand, do not have a vibrating source: they are produced by turbulent flow, leading to a wideband noise source. *Plosive sounds*, with an impulsive source, also exist, but are transient and are considered less important in this model.

These different modes of excitation can be combined into the binary source-filter model of speech production, as shown in Fig. 8.12. One of two source excitations is selected, then filtered by the vocal tract, which is assumed to include the filtering effect of the mouth. The binary source-filter model is, of course, an over-simplification of the rather complicated speech production process. Although extended models do exist, the simple source-filter model is commonly used in the speech processing literature and gives adequate model performance [32]. Generally, linear time-variant

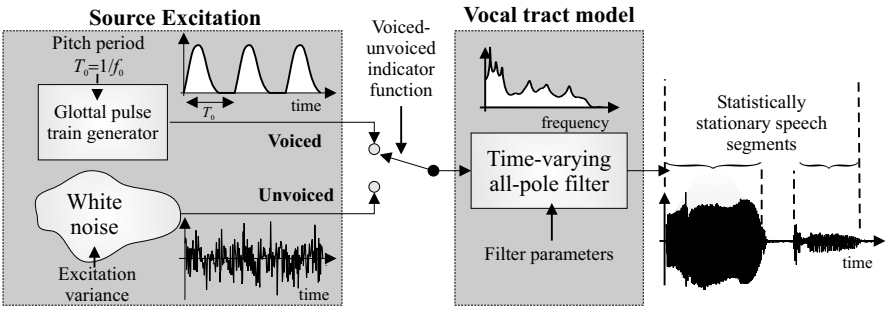


Fig. 8.12 Source-filter speech model, including typical time-domain waveforms for the voiced and unvoiced source excitation, a typical frequency response of the vocal tract and the resulting waveform

pole-zero filters and all-pole filters in particular are a popular approach for modelling the vocal tract of a talker due to their ability to accurately model the continuous short-term spectrum of speech [32]. Physically, the resonances (formants) of speech correspond to the poles of the vocal tract transfer function, while sounds that are generated through a coupling between oral and nasal tracts, for example French nasals, have anti-resonances and therefore are better modelled if the transfer function includes zeros. Thus, nasal and fricative sounds must be represented by pole-zero pairs but not by pole-only models. Nevertheless, pole-zero speech models generally require non-linear methods for estimating their parameters [27], and all-pole models are normally used instead.

8.6.2 Time-varying AR Modelling of Unvoiced Speech

According to the source-filter model for speech, unvoiced sounds correspond to a WGN excitation passing through a time-varying all-pole filter representing the vocal tract, as shown in Fig. 8.12. Hence, unvoiced speech is modelled as a TVAR process [11, 12, 27], which is defined as:

$$s(n) = -\sum_{q=1}^{Q_n} b_q(n)s(n-q) + \sigma_e(n)\,e(n), \qquad e(n) \sim \mathcal{N}(0,1)\,, \tag{8.29}$$

where $e(n)$ is the time-varying zero-mean WGN with unit variance, $\sigma_e^2(n)$ represents the variance of the excitation sequence $\hat{e}(n) = \sigma_e(n)\,e(n)$, $s(n)$ is the source signal, Q_n is the time-varying model order at time n and $\{b_q(n)\}_{q=1}^{Q_n}$ are the Time-Varying AR (TVAR) coefficients. Non-coincidentally, the TVAR process in (8.29) is of the same form as the TVAP channel model (8.22) in Sect. 8.4.3, except that the input is white. Thus, as discussed in Sect. 8.4.5, the problem of modelling unvoiced speech using this representation reduces to finding an appropriate model for

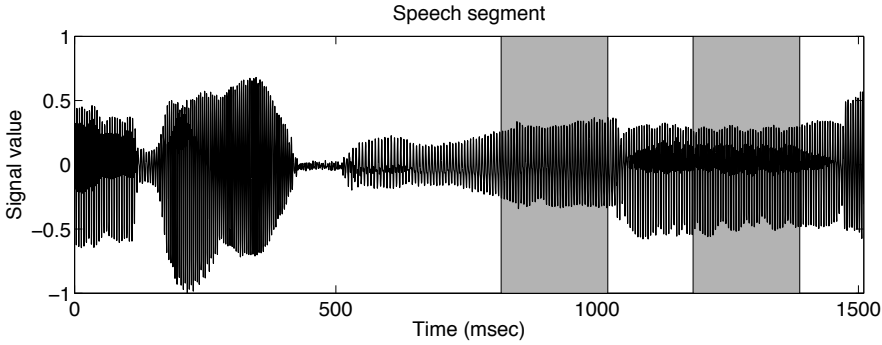


Fig. 8.13 Speech segment; shaded areas are of length 204 ms or 500 samples at a sampling frequency of $f_s = 2.45$ kHz

the TVAR parameters, $\{b_q(n)\}$. However, as discussed previously in Sect. 8.2.2, the model for the parameters is often determined by the methodology used for their estimation.

The most general variation of the parameters, $\{b_q(n)\}$, in (8.29) is when the parameters are completely uncorrelated at each sample. In this case, each sample of the signal is represented by more than one unknown coefficient. This over-determined parameterisation results in numerical problems as there is not enough data from a single realisation of a process to allow accurate parameter estimation. Therefore, it is necessary to introduce correlation into the parameter variations, and two distinct approaches are discussed in Sects. 8.6.3 and 8.6.4: namely static and stochastic source models.

8.6.2.1 Statistical Nature of Speech Parameter Variation

As explained above, it is difficult to estimate all the parameters $\{b_q(n)\}$ from (8.29) at each time step without access to the ensemble statistics. Hence, the precise statistical nature of the speech parameter variation for the TVAR model in (8.29) is essentially hidden; any estimation method is limited by prior assumptions on the statistical nature of the problem. Despite this, an illustration of the time-varying characteristics of the parameter variation can be given by taking a sliding window of block length M over a segment of speech; the window moves by one sample in each of S steps. In each window, the AR coefficients are estimated assuming the model within that block is stationary. The coefficients are computed by solving the standard Yule–Walker equations [23], and the corresponding poles are the roots of the characteristic equation. For the two segments of speech shown in the grey regions in Fig. 8.13, the corresponding pole variations introduced by the sliding window are shown in Fig. 8.14(a) and Fig. 8.14(b). The poles exhibit smooth variation over these segments of speech; this characteristic of pole movements is discussed, for

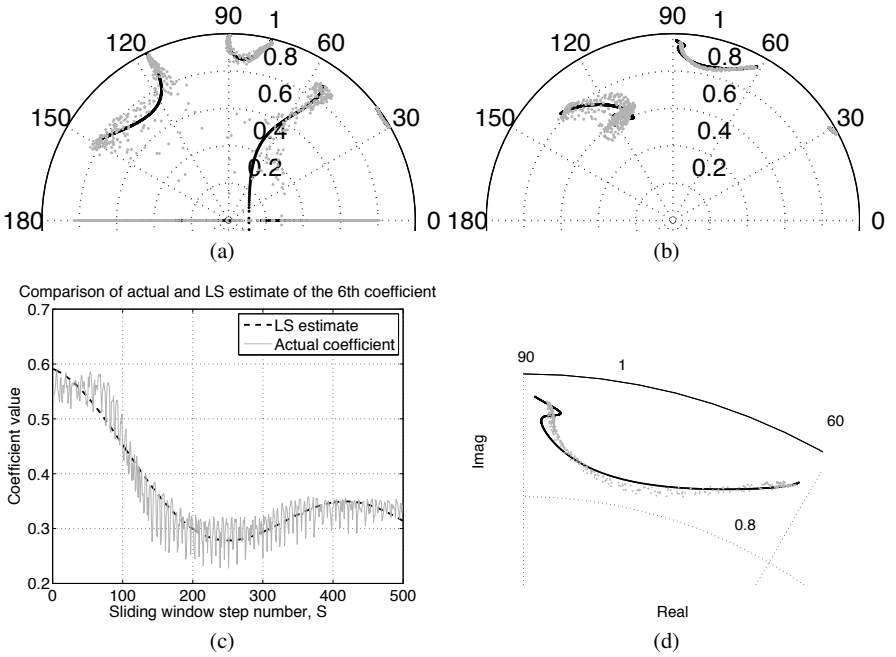


Fig. 8.14 (a) Birth and death of true poles (•) and LSE (•) for *left shaded area in Fig. 8.13*; model order: $Q = 8$. (b) True poles (•) and LSE (•) for speech segment in *right shaded area of Fig. 8.13*; model order: $Q = 6$. (c) Smooth pole variation (*Fig. 8.14 (b)*) corresponds to relatively smooth parameter variation, (d) Close-up of *Fig. 8.14(b)* showing LSE (•) matching true poles (•)

example, in [12]. Smooth pole variation often leads to relatively smooth parameter variation, as shown in Fig. 8.14(c).

8.6.3 Static Block-based Modelling of TVAR Parameters

Many statistical estimation methods impose stationarity on the model of the signal primarily to constructively exploit ergodicity. Since within the speech production process, the vocal tract is continually changing with time, sometimes slowly, sometimes rapidly as, for example, during plosive sounds and speech transitions, the assumption of stationarity is a limitation that results in poor modelling [44]. In order to reconcile partially the global non-stationarity while utilising the advantages of local ergodicity in estimation methods, a compromise is to model speech as a block-stationary process: the signal is divided into short segments or frames where the statistics of the signal are assumed to be *locally* stationary within blocks, but *globally* time-varying.

Thus, the signal $s(n)$ is partitioned into K contiguous disjoint blocks. Block $k \in \mathcal{K}$ begins at sample n_k with length $N_k = n_{k+1} - n_k$. In this block, the signal is represented by a stationary AR model of order Q_k. Using (8.29), this is equivalent to setting $Q_n = Q_k$, $\{b_q(n) = b_{k,q}, q \in \mathcal{Q}_k\}$, $\sigma_e(n) = \sigma_{e,k}$, $\forall n \in \mathcal{N}_k = \{n_k, \ldots, n_{k+1} - 1\} \subset \mathbb{Z}^{N_k}$, such that

$$s(n) = -\sum_{q=1}^{Q_k} b_{k,q} s(n-q) + \sigma_{e,k} e(n), \tag{8.30}$$

where $\{b_{k,q}\}_{q=1}^{Q_k}$ are the Block Stationary AR (BSAR) coefficients in block $k \in \mathcal{K}$ that are stationary within each block but vary over different blocks k. For continuous sounds such as vowels, the TVAR parameters change slowly, such that the BSAR model works well. With transient sounds such as plosives and stops, the BSAR model is not as good but still adequate [32]. In general, however, it is clear that even local stationarity prohibits the estimation of the full variation of the signal within that block, which is essential for accurate modelling of a time series.

8.6.3.1 Basis Function Representation

As an alternative to the BSAR model, correlation can be introduced into the parameter variations of $\{b_q(n)\}$ in (8.29) by a transformation of the non-stationary signal to a space where it can be analysed as an LTI process [3, 11, 12, 26, 35–37]. This corresponds to modelling the parameters, $\{b_q(n)\}$, as a linear combination of basis functions, and this is the same approach as used for modelling the channel in Sect. 8.4.6. To ensure that the correct number of basis functions and AR model orders are chosen, model order selection procedures should be implemented; [36] proposes such an algorithm based on the discrete Karhunen–Loève transform.

Ideally, the pole locations rather than the parameter variation are represented as a function of time by a parametric model. However, this is difficult as the relationship between poles and parameters is non-linear and a closed-form expression for the pole positions for high order models cannot be derived. If the TVAR coefficients can be represented by a linear combination of basis functions, (8.29) can be formulated as [11, 37]:

$$s(n) = -\sum_{q=1}^{Q} \underbrace{\left\{ \sum_{m=1}^{F} b_{q,m} f_m(n-q) \right\}}_{b_q(n)} s(n-q) + \sigma_e e(n), \tag{8.31}$$

where F is the number of basis functions, $\mathbf{b} = \{b_{q,m}\}_{q=1,m=1}^{Q\ \ M}$ are the *unknown* time-invariant basis coefficients, and $\{f_m(n)\}_{m=1}^{F}$ are the *known* time-varying basis functions. To demonstrate that the speech pole movements can be approximated by the model in (8.31), a Least Squares Estimate (LSE) fit to the AR parameters corresponding to the speech pole movements in Fig. 8.14(a) and Fig. 8.14(b) is performed using the trigonometric Fourier basis set

$$f_m(n) = \left\{ \sin\left(m\omega_0 \frac{n}{N}\right), \cos\left(m\omega_0 \frac{n}{N}\right) \right\} \quad \text{for} \quad m \in \{0,1,2\}, \tag{8.32}$$

with fundamental frequency $\omega_0 = 2\pi\frac{5}{9}$ rad/s. Due to the linearity of the source model in (8.31), the basis coefficients, $\mathbf{b}$, are obtained as the linear least squares estimate [23]. The full TVAR coefficients, $\{b_q(n)\}$, are then estimated by multiplication of the basis functions with the linear LSE estimate of the basis coefficients using the decomposition in (8.31). The estimates of the TVAR parameters are depicted in Fig. 8.14(a) and Fig. 8.14(b) in black dots, and show a good match to the actual poles (Fig. 8.14(d)). This and the results in [3, 11, 12, 26, 37] lead to the conclusion that a model based on the transformation from an LTV process to an LTI one through a set of basis functions can capture appropriately the time-variation of short segments of speech.

8.6.3.2 Choice of Basis Functions

The difficulties of choosing the basis functions are the same as those discussed in Sect. 8.4.6. A comparison of modelling speech signals using Fourier, Legendre and other basis sets is detailed in Charbonnier *et al.* [3]. It is often assumed for simplicity that the true speech parameters can be approximated by sinusoidal functions (Fourier basis), since these are seen to be a good model of the source parameter variations as depicted in Fig. 8.14(c).

The difficulty of abrupt parameter variations is seen in Fig. 8.14(a), where some of the speech poles evolve towards the origin and then abruptly jump away from it. Since the frequency response of poles approaching the origin becomes increasingly flat, this pole behaviour corresponds to a birth–death process. This effect does not occur for the same experiment using a lower order due to a more parsimonious representation. In other words, the death and birth of poles is an artefact introduced through the over-parameterisation of the model. Ideally, the system should have a time-varying model order so as to capture poles that contribute to the frequency response of the speech signal, and adjust the model order when poles become redundant. Thus, the model order, Q, and the block-length, N (see (8.33) in the next section) are in principle also random variables and could be allowed to vary with the block index. While this would capture any births or deaths of poles, the estimation techniques required, such as reversible-jump MCMC methods, greatly increase the computational burden and implementation complexity.

8.6.3.3 Block-based Time-varying Approach

An alternative approach to address the issue of abrupt parameter variations while using a limited set of basis functions is proposed, which relies on a block-based time-varying model. Here, the signal is segmented into shorter blocks that are modelled as locally time-varying, as well as globally time-varying. Instead of utilising one set of parameters coping with rapid global variation, several sets of param-

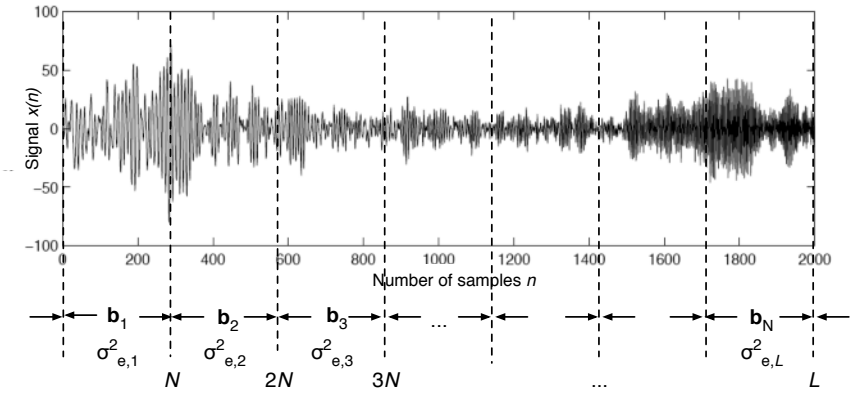


Fig. 8.15 Block-based time-varying AR speech production model

eters are introduced that capture the local variation within each block. For sufficiently short blocks, the time variation of the signal will be smooth and parameters can be estimated accurately using a standard choice of basis functions. This model thus attempts to incorporate the time-varying nature of the signal both locally as well as globally. In the block-based TVAR model, the source signal is expressed for a block of data, indexed by k and of length $N_k = n_{k+1} - n_k$, for samples $n \in \mathcal{N}_k = \{n_k, \ldots, n_{n+1} - 1\}$ as:

$$s(n) = -\sum_{q=1}^{Q} \underbrace{\left\{ \sum_{m=1}^{F} b_{kqm} f_m(n - n_k + Q - q) \right\}}_{b_q(n), n \in \mathcal{N}_k} s(n-q) + \sigma_{e,k} e(n), \tag{8.33}$$

where $e(n) \sim \mathcal{N}(0,1)$ and the block boundaries are specified by n_k and n_{k+1} in block $k \in \mathcal{K}$. This model is illustrated in Fig. 8.15 and reduces to the TVAR model (8.31) in the case of a single block. Note that this model implicitly assumes unvoiced speech segments because it uses a white excitation. An issue for further research is whether the model also works effectively for voiced speech.

8.6.4 Stochastic Modelling of TVAR Parameters

The parameter models of Sect. 8.6.3 are *static* in that once the parameters of the model are known, the speech production process is determined. Furthermore, the TVAR processes of (8.30) and (8.31) are *singly stochastic*, inasmuch as there is a single stochastic excitation to the system. If the parameters $\{b_q(n)\}$ of the general TVAR model of (8.29) are themselves allowed to evolve stochastically, then the process becomes *doubly stochastic*. Such a speech production model is used by

Vermaak *et al.* [44] who varied the parameters in (8.29) as a simple random walk given by:

$$\left.\begin{aligned} b_q(n) &= b_q(n-1) + \sigma_{b_q} w_b(n) \\ \phi_e(n) &= \phi_e(n-1) + \sigma_{\phi_e} w_{\phi_e}(n) \end{aligned}\right\} \quad \{w_b(n), w_{\phi_e}(n)\} \sim \mathcal{N}(0,1), \tag{8.34}$$

where $\phi_e(n) = \log \sigma_e^2(n)$ and $q \in \mathcal{Q}$.[11] A fixed model order is assumed for simplicity. Stability constraints can be enforced by only allowing the parameter set $\{b_q(n)\}$ to take on values in the *admissible region*, $\mathcal{B}_Q$, which corresponds to the instantaneous poles being inside the unit circle. Hence, defining the vector of TVAR coefficients at time n as $\mathbf{b}(n) = [b_1(n), \ldots, b_Q(n)]^T$, the source parameter variation in (8.34) can be written as the conditional PDFs [12]

$$p(\mathbf{b}(n) \,|\, \mathbf{b}(n-1)) \propto \mathcal{N}\left(\mathbf{b}(n) \,\middle|\, \mathbf{b}(t-1), \Delta_{\mathbf{b}}\right) \mathbb{I}_{\mathcal{B}_Q}(\mathbf{b}(n)), \tag{8.35a}$$

$$p(\phi_e(n) \,|\, \phi_e(n-1)) = \mathcal{N}\left(\phi_e(n) \,\middle|\, \phi_e(n-1), \delta_e^2\right), \tag{8.35b}$$

where $\phi_e(n) = \ln \sigma_e^2(n)$ and $\mathbb{I}_{\mathcal{B}_Q}(\mathbf{b}(n))$ is the indicator function defining the region of support, $\mathcal{B}_Q$, of $\mathbf{b}(n)$. The initial states are given defined by $p(\mathbf{b}(0)) \propto \mathcal{N}\left(\mathbf{b}(0) \,\middle|\, \mathbf{0}_{Q\times 1}, \Delta_{\mathbf{b},\mathbf{0}}\right) \mathbb{I}_{\mathcal{B}_Q}(\mathbf{b}(0))$ and $p(\phi_e(0)) \triangleq \mathcal{N}\left(\phi_e(0) \,\middle|\, 0, \delta_{e,0}^2\right)$.

Alternatively, the model can be reparameterised in terms of time-varying reflection coefficients or partial correlation coefficients [8]. If the reflection coefficients all have a magnitude of less than 1, the system is guaranteed to be stable. The key to utilising models in which the parameters $\{b_q(n)\}$ vary in a stochastic nature is to use a numerical Bayesian methodology that provides a natural environment for dealing with evolutionary or sequential problems. SMC (see Sect. 8.2.3) is particularly apt at tracking the unknown signal, $s(n)$, from the observations, $x(n)$, given in (8.1).

Nevertheless, it is still important to ensure that the motivation for a particular speech model does not become skewed by the desire to use a particular methodology. What motivates the model of (8.34): the sequential online numerical Bayesian methodology, or the "goodness" of the speech model? As discussed in Sect. 8.6.2, if it is assumed that the parameters vary slowly, a BSAR process might be more appropriate than the doubly stochastic model formed from (8.29) and (8.34). The parameters of a BSAR process, since they are time-invariant, can be estimated using a batch method such as MCMC. Thus, what really motivates the use of a BSAR model? It is apparent that the particular methodology utilised influences the choice of model.

Using the channel models in Sect. 8.4, the noise model in Sect. 8.5 and the speech models in this section, the Bayesian framework of Sect. 8.2.1 leads to Bayesian blind dereverberation algorithms as discussed in the next section.

---

[11] Variance terms are, by definition, positive, such that $\sigma_e^2(n) \in \mathbb{R}^+$; allowing the log-variance to vary as a random walk ensures this constraint is met.

[12] The set of Markov parameters $\left\{\Delta_{\mathbf{b}}, \Delta_{\mathbf{b},\mathbf{0}}, \delta_e^2, \delta_{e,0}^2\right\}$ are usually assumed known.

8.7 Bayesian Blind Dereverberation Algorithms

8.7.1 Offline Processing Using MCMC

In the offline approach to blind dereverberation, it is sought to find an analytical expression for the marginal PDF in (8.8b):

$$p(\mathbf{H} \mid \mathbf{x}) = \iint p(\mathbf{s}, \mathbf{H}, \theta \mid \mathbf{x}) \, \mathrm{d}\mathbf{s} \mathrm{d}\theta.$$

An MMAP estimate can be found either through deterministic or stochastic optimisation methods. The most straightforward situation in which an analytic solution to (8.8b) is possible is when appropriate static parametric models for the source signal and channel are used, and when it is assumed there is no observation noise. Thus, the Bayesian formulation reduces to (8.12) and the channel can be estimated using (8.13).

The static block-based TVAR model discussed in Sect. 8.6.3 is utilised for the speech signal, and an LTI all-pole filter for the channel model, such that the observed reverberant signal, $x(n)$, is given by (8.22). Given an estimate of the channel parameters, θ_h, the source, $s(n)$, can easily be recovered through a rearrangement of (8.22), in what is essentially an inverse filtering operation. Although it is possible to perform the marginalisation in (8.13) analytically, the resulting posterior PDF is complicated to optimise, and in practice the Gibbs sampler described in Sect. 8.2.3 is utilised. The Gibbs sampler implementation requires conditional densities. As indicated in (8.16) of Algorithm 8.1, these rely on the complete likelihood and the priors. Thus, the likelihood term and the choice of priors are described below.

8.7.1.1 Likelihood for Source Signal

It can be shown that the likelihood for all the source data across K blocks, each of size $N_k = n_{k+1} - n_k$, is given by

$$p_{\mathbf{S}}(\mathbf{s} \mid \theta_s) = p_{\mathbf{S}_0}(\mathbf{s}_0 \mid \mathcal{M}_s) \prod_{k \in \mathcal{K}} \frac{1}{\left(2\pi\sigma_{e,k}^2\right)^{N_k/2}} \exp\left\{ -\frac{\|\mathbf{s}_k + \mathbf{U}_k \mathbf{b}_k\|_2^2}{2\sigma_{e,k}^2} \right\}, \tag{8.36}$$

where the source parameter vector is defined by $\theta_s = \{\mathbf{b}, \boldsymbol{\sigma}_e\}$, with $\boldsymbol{\sigma}_e$ containing the excitation variances and $\mathbf{b} = \{\mathbf{b}_k, k \in \mathcal{K}\}$ containing the basis parameter coefficients. Thus, in block k:

- $[\boldsymbol{\sigma}_e]_k = \sigma_{e,k}^2 \in \mathbb{R}^+$ is the excitation variance, and $\mathbf{b}_k \triangleq [\mathbf{b}_{k,1}^T \ \ldots \ \mathbf{b}_{k,Q}^T]^T \in \mathbb{R}^{FQ \times 1}$, with $\left[\mathbf{b}_{k,q}\right]_i = b_{kqi}$ the basis function coefficients.
- The vector of source samples is $\mathbf{s}_k = [s(n_k) \ \ldots \ s(n_{k+1}-1)]^T \in \mathbb{R}^{N_k \times 1}$, and $\mathbf{S}_{k,q} = \mathrm{diag}\{s(n_k - q) \ \ldots \ s(n_{k+1} - 1 - q)\} \in \mathbb{R}^{N_k \times N_k}$ is a diagonal matrix of shifted source signal samples.

- $\mathbf{F}_{k,q} \in \mathbb{R}^{N_k \times F}$ is a matrix whose columns contains the basis functions such that the $(i,j)^{\text{th}}$ element of $\mathbf{F}_{k,q}$ is $[\mathbf{F}_{k,q}]_{ij} = f_i(j+Q-q)$.
- $\mathbf{U}_k \triangleq [\mathbf{U}_{k,1} \ \ldots \ \mathbf{U}_{k,Q}] \in \mathbb{R}^{N_k \times FQ}$, where $\mathbf{U}_{k,q} = \mathbf{S}_{k,q}\mathbf{F}_{k,q}$.

The vector containing *all* the source data is denoted $\mathbf{s} = [\mathbf{s}_0^T \ \ldots \ \mathbf{s}_K^T]^T$, $\mathbf{s}_0$ is the initial data for the first block and $\mathcal{M}_s$ is the data model.

8.7.1.2 Complete Likelihood for Observations

The complete likelihood can be expressed by writing (8.22) as $\mathbf{s} = \mathrm{A}\mathbf{x}$, where the vector of observation samples $\mathbf{x} = [x(0) \ \ldots \ x(N-1)]^T \in \mathbb{R}^{N\times 1}$, the vector of the source samples is $\mathbf{s} \in \mathbb{R}^{N-P\times 1}$ is as in (8.36) and $\mathbf{A} \in \mathbb{R}^{N-P\times N}$ is the matrix containing the TVAP channel coefficients:

$$\mathbf{A} = \begin{bmatrix} a_P(P) & \cdots & a_1(P) & 1 & 0 & \ldots & 0 \\ 0 & a_P(P+1) & \cdots & a_1(P+1) & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & & \ddots & \\ 0 & \cdots & 0 & a_P(N-1) & \cdots & a_1(N-1) & 1 \end{bmatrix}.$$

From (8.36), the likelihood of the observations given the source parameters, θ_s, and the channel coefficients, $\theta_h = \mathbf{a}$, is given by (see (8.12)):

$$\begin{aligned} p_X(\mathbf{x}\,|\,\theta) &= p_S(\mathbf{s}\,|\,\theta_s)\big|_{\mathbf{s}=\mathbf{A}\mathbf{x}} \\ &\approx \prod_{k\in\mathcal{K}} \frac{1}{\left(2\pi\sigma_{e,k}^2\right)^{\frac{N_k}{2}}} \exp\left\{-\frac{\|\mathbf{s}_k + \mathbf{U}_k\mathbf{b}_k\|_2^2}{2\sigma_{e,k}^2}\right\}\Bigg|_{\mathbf{s}=\mathbf{A}\mathbf{x}}, \end{aligned} \tag{8.37}$$

where the vectors $\{\mathbf{s}_k\}$ and matrices $\{\mathbf{U}_k\}$ are functions of the channel parameters and observations *via* the relationship $\mathbf{s} = \mathbf{A}\mathbf{x}$, and it has been assumed that $p_{\mathbf{S}_0}(\mathbf{s}_0\,|\,\mathcal{M}_s) \approxeq \text{const}$. The TVAP parameters in $\mathbf{A}$ are evaluated from the channel basis weighting coefficients, $\mathbf{a}$, through (8.25).

8.7.1.3 Prior Distributions of Source, Channel and Error Residual

The prior in (8.12) can be factorised assuming that the source parameters are independent between blocks and also independent of the channel parameters:

$$\begin{aligned} p_\Theta(\theta\,|\,\psi) &= p_{\Theta_h}(\theta_h\,|\,\psi_h)\, p_{\Theta_s}(\theta_s\,|\,\psi_s) \\ &= p\left(\mathbf{a}\,|\,\sigma_{\mathbf{a}}^2\right) p\left(\sigma_{\mathbf{a}}^2\,|\,\alpha_{\mathbf{a}},\beta_{\mathbf{a}}\right) \prod_{k\in\mathcal{K}} p\left(\mathbf{b}_k\,|\,\sigma_{\mathbf{b}_k}^2\right) p\left(\sigma_{\mathbf{b}_k}^2\right) p\left(\sigma_{e,k}^2\right), \end{aligned} \tag{8.38}$$

where $\psi = \{\psi_s, \psi_h\}$ are the hyper-parameters and hyper-hyperparameters. Note that $\sigma_{\mathbf{a}}^2$ and $\sigma_{\mathbf{b}_k}^2$ are the channel and source hyperparameters and that all the hyper-

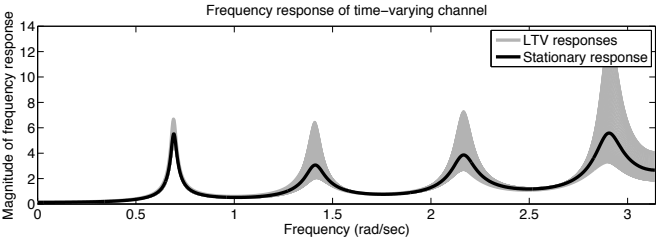


Fig. 8.16 Equivalent frequency response variation of the LTV all-pole channel

.

hyperparameters are assumed known (and therefore not shown in (8.38)). The terms in the likelihood for AR parameters usually take the form of a Gaussian [2]. Thus, to maintain analytical tractability, Gaussian priors are imposed on the channel and source parameters, i.e., $p\left(\mathbf{a}\,|\,\sigma_{\mathbf{a}}^2\right) = \mathcal{N}\left(\mathbf{a}\,|\,\mathbf{0}, \sigma_{\mathbf{a}}^2 \mathrm{I}_P\right)$ and $p\left(\mathbf{b}_k\,|\,\sigma_{\mathbf{b}_k}^2\right) = \mathcal{N}\left(\mathbf{b}_k\,|\,\mathbf{0}, \sigma_{\mathbf{b}_k}^2 \mathrm{I}_Q\right)$.[13] A standard prior for scale parameters, such as variances, is the inverse-Gamma density.[14] The prior distribution on the excitation variance, and the hyperparameters on the source and channel are therefore assigned as: $p\left(\sigma_{e,k}^2\right) = \mathcal{IG}\left(\sigma_{e,k}^2\,|\,\alpha_{e,k}, \beta_{e,k}\right)$, $p\left(\sigma_{\mathbf{b}_k}^2\right) = \mathcal{IG}\left(\sigma_{\mathbf{b}_k}^2\,|\,\alpha_{\mathbf{b}_k}, \beta_{\mathbf{b}_k}\right)$ and $p\left(\sigma_{\mathbf{a}}^2\right) = \mathcal{IG}\left(\sigma_{\mathbf{a}}^2\,|\,\alpha_{\mathbf{a}}, \beta_{\mathbf{a}}\right)$; $\{\alpha_{\{\mathbf{a},\mathbf{b}_k,e_k\}}, \beta_{\{\mathbf{a},\mathbf{b}_k,e_k\}}\}$ are the known hyper-hyperparameters. Thus, $\psi \triangleq \{\sigma_{\{\mathbf{a},\mathbf{b}_k\}}^2, \alpha_{\{\mathbf{a},\mathbf{b}_k,e_k\}}, \beta_{\{\mathbf{a},\mathbf{b}_k,e_k\}}\}$.

8.7.1.4 Posterior Distribution of the Channel Parameters

The joint-posterior PDF is found using Bayes's theorem in (8.13):

$$p\left(\mathbf{a},\mathbf{b},\sigma_e\,|\,\mathbf{x},\psi\right) \propto p\left(\mathbf{x}\,|\,\mathbf{a},\mathbf{b},\sigma_e\right) p\left(\mathbf{a},\mathbf{b},\sigma_e\,|\,\psi\right). \tag{8.39}$$

Using the relationships in (8.37) and (8.38), and the marginalisation of (8.13), the nuisance parameters $\mathbf{b}$ and σ_e can be marginalised out to form the marginal *a posteriori* PDF. As shown in [7], this evaluates to:

$$p\left(\mathbf{a}\,|\,\mathbf{x},\,\psi\right) \propto \exp\left\{-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma_{\mathbf{a}}^2}\right\} \prod_{k\in\mathcal{K}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} E_k^{-\left(\frac{N_k}{2}+\alpha_{e,k}\right)}, \tag{8.40a}$$

with
$$E_j = 2\beta_{e,j} + \mathbf{s}_j^T\mathbf{s}_j - \mathbf{s}_j^T\mathbf{U}_j\boldsymbol{\Sigma}_j^{-1}\mathbf{U}_j^T\mathbf{s}_j \tag{8.40b}$$

and
$$\boldsymbol{\Sigma}_j = \mathbf{U}_j^T\mathbf{U}_j + \delta_{\mathbf{b}_j}^{-2}\mathrm{I}_{FQ}, \tag{8.40c}$$

[13] $p\left(x\,|\,\mu,\sigma^2\right) = \mathcal{N}\left(x\,|\,\mu,\sigma\right)$ denotes a Gaussian PDF whereas $x \sim \mathcal{N}(\mu,\sigma)$ denotes that x is a Gaussian sample; I_K is the identity matrix of size $K \times K$.

[14] The inverse-Gamma PDF is $\mathcal{IG}\left(x\,|\,\alpha,\beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left\{-\frac{\beta}{x}\right\}$.

where $j \in \mathcal{K}$, $\delta_{\mathbf{b}_j}$ is a hyperparameter defined for analytical tractability as $\sigma^2_{\mathbf{b}_j} \triangleq \delta^2_{\mathbf{b}_j}\sigma^2_{e,j}$. Similarly to (8.37), it is understood in (8.40) that $\mathbf{s}_j$ and $\mathbf{U}_j$ are functions of the parameters $\mathbf{a}$ and the observed data $\mathbf{x}$. The MMAP estimate is found by solving $\hat{\mathbf{a}}_{\text{MMAP}} = \arg\max_{\mathbf{a}} p(\mathbf{a}\,|\,\mathbf{x}, \psi)$. This MMAP estimate is most easily found using Gibbs sampling (see Algorithm 8.1):

$$
\begin{aligned}
\mathbf{a}^{(i+1)} &\sim p\left(\mathbf{a}\,|\,\mathbf{b}^{(i)}, \sigma_e^{(i)}, \sigma_{\mathbf{a}}^{2(i)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\mathbf{b}_\ell^{(i+1)} &\sim p\left(\mathbf{b}\,|\,\mathbf{a}^{(i+1)}, \{\mathbf{b}_k\}_{k=1:\ell-1}^{(i+1)}, \{\mathbf{b}_k\}_{k=\ell+1:L}^{(i)}, \sigma_e^{(i)}, \sigma_{\mathbf{a}}^{2(i)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\sigma_{e,\ell}^{2(i+1)} &\sim p\left(\sigma^2_{e,\ell}\,\middle|\,\mathbf{a}^{(i+1)}, \mathbf{b}^{(i+1)}, \left\{\sigma^2_{e,k}\right\}_{k=1:\ell-1}^{(i+1)}, \left\{\sigma^2_{e,k}\right\}_{k=\ell+1:L}^{(i)}, \sigma_{\mathbf{a}}^{2(i)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\sigma_{\mathbf{a}}^{2(i+1)} &\sim p\left(\sigma^2_{\mathbf{a}}\,\middle|\,\mathbf{a}^{(i+1)}, \mathbf{b}^{(i+1)}, \sigma_e^{(i+1)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\sigma_{\mathbf{b}_\ell}^{2(i+1)} &\sim p\left(\sigma^2_{\mathbf{b}_\ell}\,\middle|\,\mathbf{a}^{(i+1)}, \mathbf{b}^{(i+1)}, \sigma_e^{(i+1)}, \sigma_{\mathbf{a}}^{2(i+1)}, \left\{\sigma^2_{\mathbf{b}_k}\right\}_{k=1:\ell-1}^{(i+1)}, \left\{\sigma^2_{\mathbf{b}_k}\right\}_{k=\ell+1:L}^{(i)}\right),
\end{aligned}
$$

where each of the conditional PDFs are also dependent on the observations, $\mathbf{x}$, and known hyper-hyperparameters. These conditionals take the form:

$$
\begin{aligned}
p(\mathbf{a}\,|\,\theta_{-\mathbf{a}}) &\propto p(\mathbf{x}\,|\,\theta_h, \theta_s)\, p\left(\mathbf{a}\,|\,\sigma^2_{\mathbf{a}}\right),\\
p\left(\mathbf{b}_\ell\,|\,\theta_{-\mathbf{b}_\ell}\right) &\propto p(\mathbf{x}\,|\,\theta_h, \theta_s)\, p\left(\mathbf{b}_\ell\,|\,\sigma^2_{\mathbf{b}_\ell}\right),\\
p\left(\sigma^2_{e,\ell}\,\middle|\,\theta_{-\sigma^2_{e,\ell}}\right) &\propto p(\mathbf{x}\,|\,\theta_h, \theta_s)\, p\left(\sigma^2_{e,\ell}\,\middle|\,\alpha_{e,\ell}, \beta_{e,\ell}\right),\\
p\left(\sigma^2_{\mathbf{a}}\,\middle|\,\theta_{-\sigma^2_{\mathbf{a}}}\right) &\propto p\left(\mathbf{a}\,|\,\sigma^2_{\mathbf{a}}\right) p\left(\sigma^2_{\mathbf{a}}\,\middle|\,\alpha_{\mathbf{a}}, \beta_{\mathbf{a}}\right),\\
p\left(\sigma^2_{\mathbf{b}_\ell}\,\middle|\,\theta_{-\sigma^2_{\mathbf{b}_\ell}}\right) &\propto p\left(\mathbf{b}_\ell\,|\,\sigma^2_{\mathbf{b}_\ell}\right) p\left(\sigma^2_{\mathbf{b}_\ell}\,\middle|\,\alpha_{\mathbf{b}_\ell}, \beta_{\mathbf{b}_\ell}\right),
\end{aligned}
$$

where $\theta = \{\theta_s, \theta_h\} = \{\mathbf{a}, \mathbf{b}, \sigma_e, \sigma^2_{\mathbf{a}}, \sigma_{\mathbf{b}}\}$ and $\theta_{-\alpha}$ denotes θ with element α removed. Full details of the form of these conditions can be found in [7].

8.7.1.5 Experimental Results

Results demonstrating the performance of this offline Bayesian inference problem are shown in Evers and Hopgood [7]. A single experimental result is presented in this section to summarise the performance of the algorithm. An acoustic channel is based on perturbations of an actual acoustic gramophone horn response up to a frequency of 1225 Hz [21]. This range matches that of the investigations in Sect. 8.4.5. Full-band signal enhancement could be achieved using subband methods as discussed in Sect. 8.4.4. The magnitude frequency response of the original time-invariant channel has four resonant modes which introduces a reasonable and noticeable amount of acoustic distortion into a signal passed through the filter. A time-varying response is obtained by perturbing each of the original channel poles in a circle of small radius. Despite there being a highly non-linear relationship be-

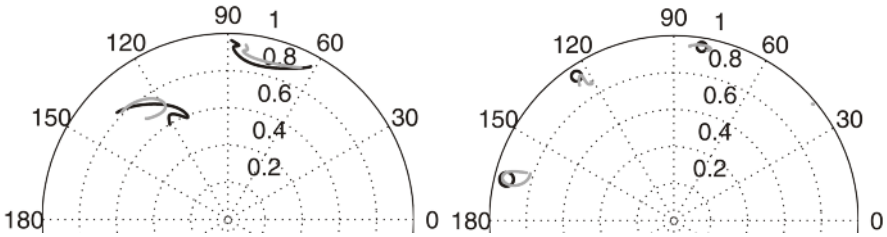


Fig. 8.17 Actual poles (•) *vs.* Gibbs estimates (◦) for (a) the source and (b) the channel

tween the poles and filter parameters, it is possible to model the parameter variation accurately using the sinusoidal basis set:

$$\{g_\ell(n)\} = \{1, \sin(2\pi n/N), \cos(2\pi n/N), \sin(2.5\pi n/N), \cos(2.5\pi n/N)\},$$

where N is the total number of samples. The variability of the channel is shown as grey lines in Fig. 8.16. Here, the magnitude frequency response of the acoustic impulse response is plotted at each time instance, assuming the parameters represent an equivalent LTI system. The frequency response of the original unperturbed channel corresponds to the black line; the actual pole variations are shown in Fig. 8.17(b).

The experiment presented considers globally modelling the source using a single-block TVAR. A synthetic fourth-order TVAR process is presented to the input of the eighth-order channel. The source is generated with time-varying parameters that reflect the statistical nature and pole variations of real speech. The parameter variations are chosen to give the LSE approximations of the two left-most pole trajectories shown in Fig. 8.14(b); these trajectories are reproduced in Fig. 8.17(a). The basis set used for the source corresponds to the Fourier set $\{f_m(n)\} = \{\sin(m\omega_0 n/N), \cos(m\omega_0 n/N)\}_{m=0}^{2}$ with fundamental frequency $\omega_0 = 2\pi\frac{5}{9}$ rad/s. The total number of source samples used is $N = 2000$, and is chosen to give sufficient data that the channel estimates have low variance. With regards to (8.33), $K = 1$, $n_1 = 4$ and $n_2 = N$, where n_k are the change-points, i.e., n_1 is the index of the first sample in the block and n_2 is the index of the last sample in the block. The Gibbs sampler is executed for 5000 iterations with a burn-in period of 500 (10%) samples, although the estimates tend to converge within a few hundred samples. A Monte Carlo experiment with 100 runs is executed to ensure that the performance is consistent. The averaged estimated pole trajectories are shown in Fig. 8.17(a) and Fig. 8.17(b); any individual run gives very similar results to the averaged performance.

The single-block TVAR model will not adequately capture the full time-varying nature of a real speech signal and therefore, as discussed in Sect. 8.6.3.3, a multi-block-based model is more robust and flexible. Results demonstrating the perfor-

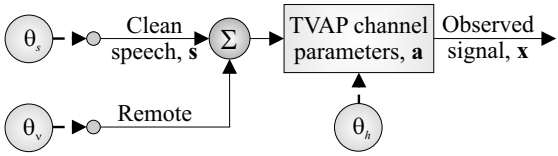


Fig. 8.18 Simplified system model for online dereverberation algorithm

mance of the MCMC algorithm for the block-based TVAR for both synthetic and real speech signals are presented in [7].

8.7.2 Online Processing Using Sequential Monte Carlo

Online or sequential estimation facilitates online processing of the signal, which is of particular interest for applications such as security surveillance systems where results should become available as soon as a signal sample is measured, i.e., where offline batch methods are impractical. Particle filters (or SMC methods) represent a target distribution by a large number of random variates from a hypothesis distribution. Incorporation of knowledge about the current and past measured samples allows for correction and evolution of the particles in time. Particle filters were shown to effectively enhance systems distorted by WGN [44] and for reverberant all-zero channels [4]. This section describes an extension of this work to reverberant all-pole channels (see Sect. 8.4.3) and spatially distinct noise sources (see Sect. 8.5).

8.7.2.1 Source and Channel Model

Various system and noise models were discussed in Sect. 8.5. The CAPZ channel model simplified the full system model in Fig. 8.9 to that shown in Fig. 8.11. Although the model in Fig. 8.11 is of great interest, the presence of the general RTFs dependent on source-sensor geometries leads to difficulties in uniquely modelling and blindly identifying the source signal. Additional identifiability results are required before it can be determined whether this model leads to unique solutions. As a compromise, a more simple model is used to facilitate online estimation; this model is shown in Fig. 8.18.[15] In this model, the source signal, $s(n)$, is distorted by WGN, $v(n)$, with variance $\sigma_v^2(n)$. This noisy speech signal is then filtered through the channel, which is modelled as a P^{th} order time-varying all-pole filter. The observations are thus given by:

[15] Although the noise and signal are assumed independent, a channel gain in Fig. 8.18 is unnecessary since there is an inherent scaling ambiguity.

$$x(n) = -\sum_{p=1}^{P} a_p(n)x(n-p) + s(n) + \sigma_v(n)v(n), \quad v(n) \sim \mathcal{N}(0,1). \tag{8.41}$$

It is important to note that this model differs from simply adding noise to (8.22); in other words, it differs from the model $\hat{x}(n) = x(n) + s(n)$ with $x(n)$ given by (8.22). The source signal, $s(n)$, results from (8.29), where the parameters vary stochastically as described in Sect. 8.6.4. In particular, the conditional PDFs for the parameter variation are given by (8.35). The measurement noise is assumed to have a similar variation as the excitation noise in (8.35b). Thus, $v(n)$ has a log-variance that follows a random walk:

$$p(\phi_v(n) \,|\, \phi_v(n-1)) \triangleq \mathcal{N}\left(\phi_v(n) \,\middle|\, \phi_v(n-1), \delta_v^2\right), \tag{8.42}$$

where $\phi_v(n) = \ln \sigma_v^2(n)$. The initial state is $p(\phi_v(0)) \triangleq \mathcal{N}\left(\phi_v(0) \,\middle|\, 0, \delta_{v0}^2\right)$. The hyperparameters $\{\delta_v^2, \delta_{v0}^2\}$ are assumed known.

8.7.2.2 Conditionally Gaussian State Space

Assuming known source and channel parameters, θ_s and θ_h respectively, the source model, (8.29) and measurement equation in (8.41) can be written in the linear state-space form:

$$\mathbf{s}(n) = \mathbf{B}(n)\mathbf{s}(n-1) + \sigma_e(n)\,\mathbf{c}\,e(n), \tag{8.43a}$$

$$x(n) = -\mathbf{a}^T(n)\mathbf{x}_{n-1:n-P} + \mathbf{c}^T\mathbf{s}(n) + \sigma_v(n)v(n), \tag{8.43b}$$

for $n > 0$. The state vector, $\mathbf{s}(n)$, and state transition matrix, $\mathbf{B}(n)$, are:

$$\mathbf{s}(n) = [s(n) \;\ldots\; s(n-P+1)]^T, \quad \mathbf{B}(n) \triangleq \begin{bmatrix} & \mathbf{b}(n)^T \\ \mathrm{I}_{Q-1} & \mathbf{0}_{Q-1\times 1} \end{bmatrix}.$$

Moreover, $\mathbf{c}^T \triangleq [1\;\mathbf{0}_{\times 1}Q-1]$, the TVAP channel parameters are contained in $\mathbf{a}(n) = [a_1(n) \;\ldots\; a_P(n)]^T$, while $\mathbf{x}_{n-1:n-P} = [x(n-1) \;\cdots\; x(t-P)]^T$ contains the P previous observations. The set of model parameters, $\theta_{0:n}$, defines the system parameters $\theta_n = \{\mathbf{b}(n),\, \mathbf{a}(n),\, \sigma_e^2(n),\, \sigma_v^2(n)\}$. Assuming $\theta_{0:n}$ are known, since the source excitation, $e(n)$, and the measurement noise, $v(n)$, are both WGN, (8.43) is a Conditionally Gaussian State Space (CGSS) system, and the optimal estimate of the state-vector, $\mathbf{s}(n)$, can be found using the Kalman Filter (KF). The KF recursion relationships [40] at time step n are shown in Algorithm 8.3.[16] However, by the very nature of blind deconvolution, the set of parameters, $\theta_{0:n}$, is unknown and therefore a direct application of the KF is not possible. Instead, the KF can be incorporated within

[16] Due to the presence of the linear combination of past observations, $-\mathbf{a}^T(n)\mathbf{x}_{n-1:n-P}$, in the observation equation, (8.43b), the standard KF equations are modified slightly; namely the predicted observation, (8.44c), and as a result the corrected state estimate, (8.44d).

Algorithm 8.3 Kalman filter recursion relationships

$$\mu(n|n-1) = \mathbf{B}(n)\mu(n-1|n-1) \quad \text{(prediction)}, \tag{8.44a}$$
$$\mathbf{P}(n|n-1) = \sigma_e^2(n)\mathbf{c}\mathbf{c}^T + \mathbf{B}(n)\mathbf{P}(n-1|n-1)\mathbf{B}^T(n), \tag{8.44b}$$
$$x(n|n-1) = -\mathbf{a}^T(n)\mathbf{x}_{n-1:n-P} + \mathbf{c}^T\mu(n|n-1), \tag{8.44c}$$
$$\mu(n|n) = \mu(n|n-1) + \mathbf{k}(n)\left(x(n) - x(n|n-1)\right) \quad \text{(correction)}, \tag{8.44d}$$
$$\mathbf{P}(n|n) = \left(\mathrm{I}_q - \mathbf{k}(n)\mathbf{c}^T\right)\mathbf{P}(n|n-1). \tag{8.44e}$$

The optimal Kalman gain, $\mathbf{k}(n)$, and measurement residual variance, $\sigma_z^2(n)$, are:

$$\mathbf{k}(n) = \frac{1}{\sigma_z^2(n)}\mathbf{P}(n|n-1)\mathbf{c}, \text{ with } \sigma_z^2(n) = \mathbf{c}^T\mathbf{P}(n|n-1)\mathbf{c} + \sigma_v^2(n). \tag{8.45}$$

Two important distributions are the conditional likelihood of the current observation given past observations, and the PDF of the state estimate:

$$p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n}\right) = \mathcal{N}\left(x(n) \,\middle|\, x(n|n-1), \sigma_z^2(n)\right), \tag{8.46}$$
$$p\left(\mathbf{s}(n) \,|\, \theta(n), \mathbf{x}_{1:n}\right) = \mathcal{N}\left(\mathbf{s}_{0:n} \,\middle|\, \mu(n|n), \mathbf{P}(n|n)\right). \tag{8.47}$$

a sequential Monte Carlo framework where at each time step, (8.44) is evaluated using an estimate of the parameters, $\theta_{0:n}$.

8.7.2.3 Methodology

The aim is to directly reconstruct the source signal, $\mathbf{s}_{0:n} = [s(0) \ \ldots \ s(n)]$, and the set of parameters, $\theta_{0:n}$, given only the distorted signal, $\mathbf{x}_{1:n}$. This can be achieved by sampling from the posterior distribution of the source signal and unknown parameters. Since the source signal is dependent on the model parameters and observations, the joint posterior can be written as

$$p\left(\mathbf{s}_{0:n}, \theta_{0:n} \,|\, \mathbf{x}_{1:n}\right) = p\left(\mathbf{s}_{0:n} \,|\, \theta_{0:n}, \mathbf{x}_{1:n}\right) p\left(\theta_{0:n} \,|\, \mathbf{x}_{1:n}\right). \tag{8.48}$$

The joint posterior often has a complicated functional form that cannot be sampled from directly. Instead, estimates of the source signal and model parameters can be obtained by drawing samples from the conditional densities in (8.48) separately. Given $\theta_{0:n}$, since the system in (8.43) is CGSS, the likelihood of the clean signal, $p\left(\mathbf{s}_{0:n} \,|\, \theta_{0:n}, \mathbf{x}_{1:n}\right)$, can be estimated using the KF equations (8.47) in Algorithm 8.3 [4, 44]. Hence, estimation of the joint posterior in (8.48) reduces to the estimation of $p\left(\theta_{0:n} \,|\, \mathbf{x}_{1:n}\right)$. In the simplest of particle filters, namely the Sequential Importance Resampling (SIR) PF, the hypothesis (or proposal) distribution is the prior density; thus, $\pi\left(\theta_n \,|\, \mathbf{x}_{1:n}, \theta_{0:n-1}\right) = p\left(\theta_n \,|\, \theta_{0:n-1}\right)$, and the weights are therefore given by $w_n \propto p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_n\right)$ (see Algorithm 8.2). The Kalman filter is then bombarded with these particles and particle resampling is performed to ensure

that only statistically significant particles are retained. The resampling method aims to keep particles corresponding to regions of high likelihood, as given by (8.46). The estimate of the source signal corresponds to the mean of the state estimates, $\mu(n|n)$, over all particles. In the SIR PF in Algorithm 8.4, particles are drawn from the priors in (8.35a), (8.35b) and (8.42), and the importance weights reduce to (8.46) [44]. The sampling of the channel parameters, however, requires special attention.

8.7.2.4 Channel Estimation Using Bayesian Channel Updates

Various approaches for modelling the TVAP parameter variations are given in Sects. 8.4.6 and 8.4.7. The static model describing $\{a_p(n)\}$ as a linear combination of basis functions, as given by (8.25), allows for smooth parameter variation. The model is also linear-in-the-parameters, so that (8.43b) can be written in the form:

$$x(n) = -\mathbf{a}^T \tilde{\mathbf{x}}_{n-1:n-P} + \mathbf{c}^T \mathbf{s}(n) + \sigma_v(n) w(n), \tag{8.49}$$

where $\tilde{\mathbf{x}}_{n-1:n-P}$ is a function of past samples of the observations and the channel basis functions, $g_\ell(n)$. The channel coefficients $\mathbf{a}$ are static parameters.

Particle filters implicitly assume that all unknown parameters are dynamic and, therefore, work well with time-varying parameters. Thus, the models in Sect. 8.4.7 are particularly suited for the PF framework. However, these models perhaps need more justification, and the static models are preferred. The static models also have the advantage of being able to model linear time-invariant channels. However, with static parameters, such as the channels in (8.25) and (8.49), the non-dynamics in the particles makes them degenerate into a few different values [42]. Various approaches for circumventing this problem exist, but a simple approach for linear Gaussian systems is a straightforward Bayesian update. Using Bayes's theorem, the channel posterior is,[17]

$$p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n}, \theta_{0:n}^{(-\mathbf{a})}\right) = \frac{p\left(x(n), \theta_n^{(-\mathbf{a})} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}, \mathbf{a}\right) p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right)}{p\left(x(n), \theta_n^{(-\mathbf{a})} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right)}.$$

Using the basic probability factorisation

$$p\left(x(n), \theta_n^{(-\mathbf{a})} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}, \mathbf{a}\right) = p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n}\right) p\left(\theta_n^{(-\mathbf{a})} \,\middle|\, \theta_{0:n-1}^{(-\mathbf{a})}\right),$$

and ignoring any terms that are not functions of the unknown channel parameters, $\mathbf{a}$, a recursive update follows:

$$p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n}, \theta_{0:n}^{(-\mathbf{a})}\right) \propto p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n}\right) p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right). \tag{8.50}$$

[17] $\theta^{(-\mathbf{a})}$ denotes the parameter set θ with the channel parameters, $\mathbf{a}$, removed.

Algorithm 8.4 SIR particle filter for reverberant systems

for $n = 1, \ldots,$ number of samples **do**
  for $i = 1, \ldots,$ number of particles **do**
    Sample a proposal of $\theta_n^{(-\mathbf{a})}$ from (8.35a), (8.35b), (8.42).
    Prediction step of KF: (8.44a), (8.44b), from Algorithm 8.3.
    Evaluation of $\mathbf{k}(n)$, $\sigma_z^2(n)$: (8.45), from Algorithm 8.3.
    Bayesian update of channel parameters: (8.51b).
    MMAP estimation of channel: $\mathbf{a}_{\text{MMAP}} = \mu_{\mathbf{a},n}$
    Evaluation of importance weights with $\mathbf{a}_{\text{MMAP}}$: (8.46).
    Correction step of KF: (8.44d), (8.44e), from Algorithm 8.3.
  end for
  Normalisation of importance weights.
  Resampling step (see, e.g., [5]).
end for

Table 8.3 Markov parameters for synthesis and estimation

$\delta_{e_0}^2$	$\delta_{n_0}^2$	δ_e^2	δ_n^2	$\boldsymbol{\Delta}_{a_0}$	$\boldsymbol{\Delta}_a$
0.5	0.5	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$0.5\mathrm{I}_Q$	$5 \cdot 10^{-4}\mathrm{I}_Q$

Assuming a Gaussian distribution on $\mathbf{a}$ at time $n-1$ with mean, $\mu_{\mathbf{a},n-1}$, and covariance, $\mathbf{P}_{\mathbf{a},n-1}$, such that $p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right) \triangleq \mathcal{N}\left(\mathbf{a} \,|\, \mu_{\mathbf{a},n-1}, \mathbf{P}_{\mathbf{a},n-1}\right)$, since (8.46) is also Gaussian, from (8.50), so is:

$$p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n}, \theta_{0:n}^{(-\mathbf{a})}\right) \propto \mathcal{N}\left(\mathbf{a} \,|\, \mu_{\mathbf{a},n}, \mathbf{P}_{\mathbf{a},n}\right), \tag{8.51a}$$

with covariance and mean

$$\begin{aligned} \mathbf{P}_{\mathbf{a},n} &= \left(\mathbf{P}_{\mathbf{a},n-1}^{-1} + \frac{1}{\sigma_z^2(n)} \mathbf{x}_{n-1:n-P} \mathbf{x}_{n-1:n-P}^T\right)^{-1}, \\ \mu_{\mathbf{a},n} &= \mathbf{P}_{\mathbf{a},n} \left(\frac{\mathbf{x}_{n-1:n-P}}{\sigma_z^2(n)} \left[x(n) - \mathbf{c}^T \mu(n|n-1)\right] + \mathbf{P}_{\mathbf{a},n-1}^{-1} \mu_{\mathbf{a},n-1}\right). \end{aligned} \tag{8.51b}$$

The initial mean, $\mu_{\mathbf{a},0}$, and variance, $\mathbf{P}_{\mathbf{a},0}$ are assumed known. At time n, the MMAP estimate of the channel is $\mathbf{a}_{\text{MMAP}} = \mu_{\mathbf{a},n}$. This channel estimate is then used for the Kalman filter correction step, (8.44d), and evaluation of the weights, (8.46). The complete SIR PF is summarized in Algorithm 8.4.

8.7.2.5 Experimental Results

To demonstrate the performance of the online method, both synthetic sources and real speech signals are estimated from a reverberant noisy signal. The synthetic signal is used as a benchmark for the ground truth, since for real speech, the true parameter variations in the source model, (8.29), are hidden.

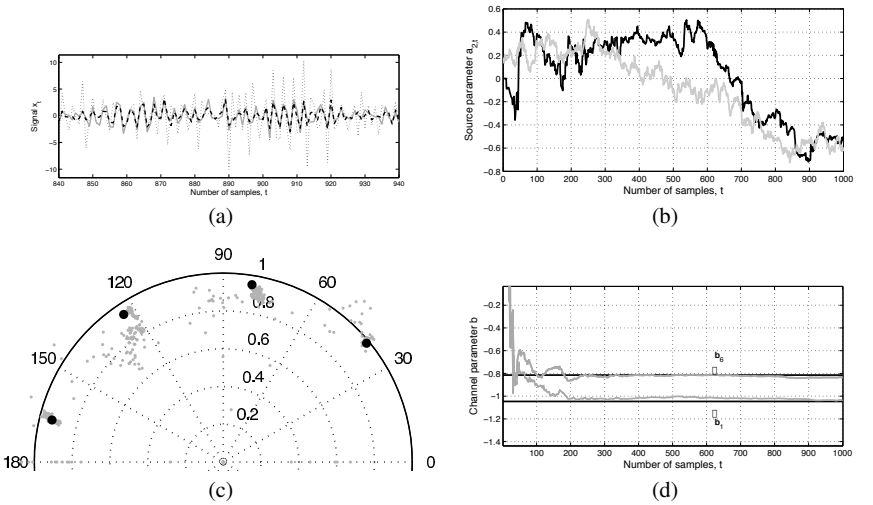


Fig. 8.19 (a) Synthetic data: estimate (▪▪▪), original (▬), observations (▥). (b) Estimated (▬) and actual source parameter (▬), $\mathbf{b}_{2,n}$. (c) Convergence of estimated (•) to actual channel poles (•). (d) Estimated (▬) and true (▬) channel parameters, $\mathbf{a}_{\{1,6\}}$

A fourth-order synthetic source signal is filtered through an eighth-order all-pole channel according to Fig. 8.18. The channel is, for simplicity, assumed to be stationary, and is identical to the initial channel parameter values used in Sect. 8.7.1. The noise level is such that the Signal Based Measure (SBM)[18] of the distorted signal is -6.15 dB. The Markov parameters are set to the values in Table 8.3 [44]. The particle filter is executed for 1000 samples and 800 particles, and $\mu_{\mathbf{a},0} = 0.5 \times \mathbf{1}_{P\times 1}$, $\mathbf{P}_{\mathbf{a},0} = \mathbf{0}_{Q\times 1}$. Even though the source parameter estimates appear inaccurate (Fig. 8.19(b)), the SBM of the enhanced signal is 4.42 dB, an improvement of 10.57 dB. The accuracy of the estimated signal compared to the clean signal and the observed signal is shown in Fig. 8.19(a). The evolution of the poles with time of the MMAP estimates of the stationary channel parameters are shown in Fig. 8.19(c). After few iterations, the estimates converge towards the actual channel poles. Likewise, the channel parameters converge after around 200 samples to the actual coefficients (Fig. 8.19(d)).

The words "The farmer's life must be arranged" uttered by a female talker sampled at 8 kHz are distorted by an eighth order acoustic horn channel [41] and noise with $\sigma_{\phi_{w_0}} = 0.5$ and constant $\sigma_{\phi_w} = 0.05$. The SBM of the observed signal is -5.73 dB. The SIR particle filter is run for 15,000 samples and 750 particles, estimating six source parameters, where $\sigma_{\phi_{\{w,v\}_0}} = 0.5$, $\sigma_{\phi_{\{w,v\}}} = 0.05$, $\boldsymbol{\Sigma}_{\{\mathbf{a}_0,\mathbf{a}\}} = \sigma_{\{\phi_{v_0},\phi_v\}}\mathrm{I}_Q$. The results are shown in Fig. 8.20. The particle filter removes low-

[18] $\mathrm{SBM}_{\mathrm{dB}} = 10\log_{10}\left(\frac{\|\mathbf{s}_{0:n-1}\|_2^2}{\|\bar{\mathbf{u}}_{0:n-1}-\mathbf{s}_{0:n-1}\|_2^2}\right)$, where $\bar{\mathbf{u}}$ is either the estimated, $\bar{\mathbf{s}}$, or the distorted, $\bar{\mathbf{x}}$, signal sequence.

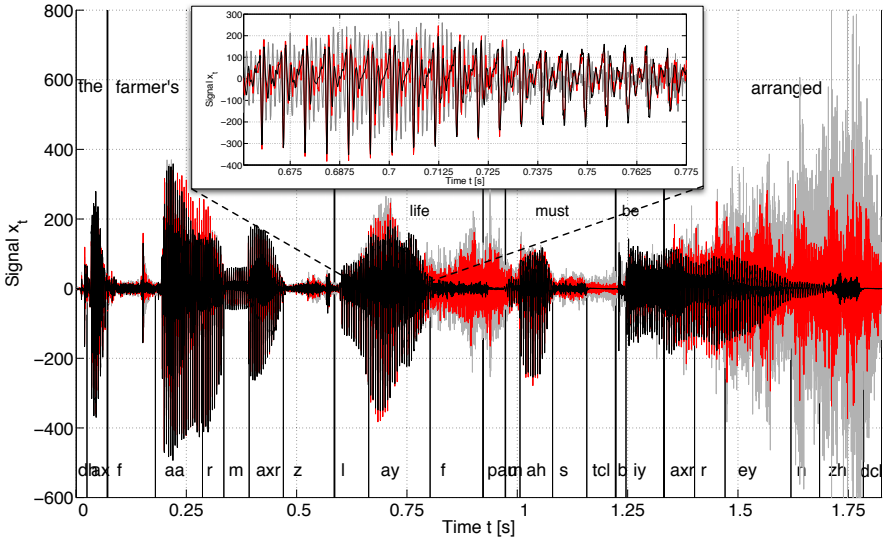


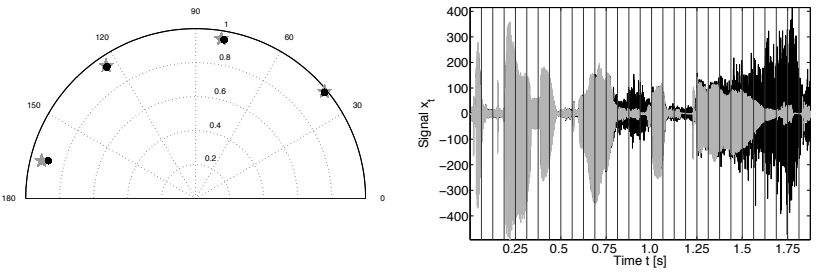
Fig. 8.20 Source signal (▬), its SIR estimate (▬) *vs.* observations (▬)

amplitude noise and the "metallic" sound effects generated by the channel. Between 0.8–0.97 s and 1.33–1.82 s, noise is dominant and the signal is not recovered. The SBM of the estimated signal is 1.950 dB, an improvement of 7.68 dB.

8.7.3 Comparison of Offline and Online Approaches

One particular difference involves the inverse channel filtering implicitly used in the MCMC method [7] but avoided in the SMC approach since the latter estimates the source signal directly. Channel inversion introduces several difficulties: (i) practical RIRs are non-minimum phase and thus difficult to invert, despite the phase being a major contributor to the perception of reverberation; (ii) any small error in the RIR estimate can lead to a significant error in its inversion since attempts to equalize high-Q resonances can still leave high-Q resonances in the equalized response. Both of these issues can potentially increase the distortion in the enhanced signal.

As a comparison with the real results presented in Sect. 8.7.2, a batch MCMC method is used for channel estimation. Although observation noise is not explicitly modelled by the approach in Sect. 8.7.1, the same observed data is used. The source model of (8.33) in Sect. 8.6.3 is again used, with $K = 30$ blocks of $N_k = 500$ samples length to match the number of samples used in Fig. 8.20. The source model order is $Q = 8$, and the basis functions are assumed to be piece-wise constant such that the model reduces to the BSAR process in (8.30). Hence, the model is equivalent to that used in [21]. The Gibbs sampler is run for 2000 iterations with a 10% burn-



(a) (b)

Fig. 8.21 (a) Actual channel poles (×) *vs.* Gibbs estimates (×) and (b) source signal (—) *vs.* Gibbs estimate (—)

in period. The channel estimate is shown in Fig. 8.21(a), and a comparison of the actual source and its estimate is shown in Fig. 8.21(b).

The SBM of the estimated source signal is 0.262 dB, an improvement of 6.02 dB. Notice that there is significant noise gain towards the end of the signal. The results can be improved by using a richer set of basis functions in the source model. Nevertheless, the results are comparable with the SMC method. Currently, the computational expense of the online SMC framework is greater, but in principle facilitates sequential estimation leading to real-time implementations.

8.8 Conclusions

This chapter has given an introduction to model-based Bayesian blind dereverberation. It has outlined the variety of source and channel models that can be used. Two key numerical methodologies have been discussed: offline batch methods and online sequential methods. There is a clear symbiosis between the methodologies available and the models that suit that methodological framework. The challenge that still remains for Bayesian blind dereverberation is to tackle the full acoustic spectrum simultaneously, as opposed to current implementations that deal with selected frequency bands independently.

References

1. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)

2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time series analysis: Forecasting and control. Holden–Day (1994)
3. Charbonnier, R., Barlaud, M., Alengrin, G., Menez, J.: Results on AR-modelling of nonstationary signals. Signal Processing **12**(2), 143–151 (1987)
4. Daly, M., Reilly, J.P., Manton, J.: A Bayesian approach to blind source recovery. In: Proc. Asilomar Conf. on Signals, Systems and Computers. Asilomar, Pacific Grove, CA (2004)
5. Doucet, A., de Freitas, J.F.G., Gordon, N.J. (eds.): Sequential Monte Carlo methods in practice. Springer (2000)
6. Doucet, A., Wang, X.: Monte carlo methods for signal processing: a review in the statistical signal processing context. IEEE Signal Process. Mag. **22**(6), 152–170 (2005)
7. Evers, C., Hopgood, J.R.: Parametric modelling for single-channel blind dereverberation of speech from a moving speaker. IET Signal Processing (2008)
8. Fong, W., Godsill, S.J., Doucet, A., West, M.: Monte Carlo smoothing with application to audio signal enhancement. IEEE Trans. Signal Process. **50**(2), 438–449 (2002)
9. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
10. Godsill, S.J., Andrieu, C.: Bayesian separation and recovery of convolutively mixed autoregressive sources. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1733–1736. Phoenix, Arizona (1999)
11. Grenier, Y.: Time-dependent ARMA modeling of nonstationary signals. IEEE Trans. Acoust., Speech, Signal Process. **31**, 899–011 (1983)
12. Hall, M.G., Oppenheim, A.V., Willsky, A.S.: Time-varying parametric modeling of speech. Signal Processing **5**(3), 267–285 (1978)
13. Haneda, Y., Kaneda, Y., Kitawaki, N.: Common-Acoustical-Pole and Residue model and its application to spatial interpolation and extropolation of a room transfer function. IEEE Trans. Speech Audio Process. **7**(6), 709–717 (1999)
14. Haneda, Y., Makino, S., Kaneda, Y.: Common acoustical pole and zero modelling of room transfer functions. IEEE Trans. Speech Audio Process. **2**(2), 320–328 (1994)
15. Hopgood, J.R.: Bayesian blind MIMO deconvolution of nonstationary autoregressive sources mixed through all-pole channels. In: Proc. IEEE Workshop Statistical Signal Processing (2003)
16. Hopgood, J.R.: Models for blind speech dereverberation: A subband all-pole filtered block stationary autoregressive process. In: European Signal Processing Conference. Antalya, Turkey (2005)
17. Hopgood, J.R.: A subband modelling approach to the enhancement of speech captured in reverberant acoustic environments: MIMO case. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Mohonk Mountain House, New York (2005)
18. Hopgood, J.R., Hill, S.I.: An exact solution for incorporating boundary continuity constraints in subband all-pole modelling. In: Proc. IEEE Workshop Statistical Signal Processing. Bordeaux, France (2005)
19. Hopgood, J.R., Rayner, P.J.W.: A probabilistic framework for subband autoregressive models applied to room acoustics. In: Proc. IEEE Workshop Statistical Signal Processing, pp. 492–494 (2001)
20. Hopgood, J.R., Rayner, P.J.W.: Bayesian formulation of subband autoregressive modelling with boundary continuity constraints. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Hong Kong (2003)
21. Hopgood, J.R., Rayner, P.J.W.: Blind single channel deconvolution using nonstationary signal processing. IEEE Trans. Speech Audio Process. **11**(5), 476–488 (2003)
22. Johansen, L.G., Rubak, P.: The excess phase in loudspeaker/room transfer functions: Can it be ignored in equalization tasks? J. Audio Eng. Soc. (1996). Preprint 4181
23. Kay, S.M.: Fundamentals of statistical signal processing, vol. 1. Prentice Hall Signal Processing Series (1993)
24. Kundur, D., Hatzinakos, D.: Blind image deconvolution. IEEE Signal Process. Mag. **13**(3), 43–64 (1996)

25. Kuttruff, H.: Room acoustics, 4th edn. Spon Press (2000)
26. Liporace, L.A.: Linear estimation of nonstationary signals. J. Acoust. Soc. Am. **58**(6), 1288–1295 (1976)
27. Makhoul, J.: Linear prediction: A tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
28. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
29. Mourjopoulos, J.N.: On the variation and invertibility of room impulse response functions. J. Sound Vib. **102**(2), 217–228 (1985)
30. Mourjopoulos, J.N., Paraskevas, M.A.: Pole and zero modeling of room transfer functions. J. Sound Vib. **146**(2), 281–302 (1991)
31. Nakatani, T., Kinoshita, K., Miyoshi, M.: Harmonicity-based blind dereverberation for single-channel speech signals. IEEE Trans. Audio, Speech, Lang. Process. **15**(1), 80–95 (2007)
32. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals. Prentice-Hall (1978)
33. Radlović, B.D., Kennedy, R.A.: Nonminimum-phase equalization and its subjective importance in room acoustics. IEEE Trans. Speech Audio Process. **8**(6), 728–737 (2000)
34. Radlović, B.D., Williamson, R.C., Kennedy, R.A.: Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Speech Audio Process. **8**(3), 311–319 (2000)
35. Rajan, J.J., Rayner, P.J.W.: Parameter estimation of time-varying autoregressive models using the Gibbs sampler. IEE Electronics Lett. **31**(13), 1035–1036 (1995)
36. Rajan, J.J., Rayner, P.J.W.: Generalized feature extraction for time-varying autoregressive models. IEEE Trans. Signal Process. **44**(10), 2498–2507 (1996)
37. Rajan, J.J., Rayner, P.J.W., Godsill, S.J.: Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler. IEE Proc.–Vis. Image Signal Process. **144**(4), 249–256 (1997)
38. Rao, S., Pearlman, W.A.: Analysis of linear prediction, coding, and spectral estimation from subbands. IEEE Trans. Inf. Theory **42**(4), 1160–1178 (1996)
39. Rao, T.S.: The fitting of nonstationary time-series models with time-dependent parameters. J. Royal Stat. Soc. B **32**(2), 312–322 (1970)
40. Ristic, B., Arulampalam, S., Gordon, N.: Beyond the Kalman filter – Particle filters for tracking applications. Artech House (2004)
41. Spencer, P.S.: System identification with application to the restoration of archived gramophone recordings. Ph. D. Thesis, University of Cambridge, UK (1990)
42. Storvik, G.: Particle filters for state-space models with the presence of unknown static parameters. IEEE Trans. Signal Process. **50**(2), 281–289 (2002)
43. Tan, S.L., Fischer, T.R.: Linear prediction of subband signals. IEEE J. Sel. Areas Commun. **12**(9), 1576–1583 (1994)
44. Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J.: Particle methods for Bayesian modeling and enhancement of speech signals. IEEE Trans. Speech Audio Process. **10**(3), 173–185 (2002)
45. Wang, H., Itakura, F.: Dereverberation of speech signals based on sub-band envelope estimation. IEICE Trans. Fund. Elec. Comms. Comp. Sci. **E74**(11), 3576–3583 (1991)
46. Wang, H., Itakura, F.: Realization of acoustic inverse filtering through multi-microphone subband processing. IEICE Trans. Fund. Elec. Comms. Comp. Sci. **E75-A**(11), 1474–1483 (1992)

Chapter 9
Inverse Filtering for Speech Dereverberation Without the Use of Room Acoustics Information

Masato Miyoshi, Marc Delcroix, Keisuke Kinoshita, Takuya Yoshioka, Tomohiro Nakatani, and Takafumi Hikichi

Abstract This chapter discusses multi-microphone inverse filtering, which does not use *a priori* information of room acoustics, such as room impulse responses between the target speaker and the microphones. One major problem as regards achieving this type of processing is the degradation of the recovered speech caused by excessive equalization of the speech characteristics. To overcome this problem, several approaches have been studied based on a multichannel linear prediction framework, since the framework may be able to perform speech dereverberation as well as noise attenuation. Here, we first discuss the relationship between optimal filtering and linear prediction. Then, we review our four approaches, which differ in terms of their treatment of the statistical properties of a speech signal.

9.1 Introduction

The inverse filtering of room acoustics is useful in various applications such as sound reproduction, sound-field equalization, and speech dereverberation. Usually an impulse response between a sound source and a microphone in an enclosure is modeled as a polynomial with a finite order, which is called an Acoustic Transfer Function (ATF), and an inverse filter is designed to remove the reverberation effect of the polynomial. Such inverse filter design may be roughly classified into two groups: one is to calculate an inverse of the replica of an ATF, which may be measured or estimated *a priori*. The other is to calculate the inverse directly from reverberant signals observed at microphones.

As regards the latter, a major problem that we must solve is the degradation of the recovered speech caused by excessive equalization of the speech characteristics [19]. This problem would not occur if speech were considered to be independent and identically distributed (i.i.d) [20].

NTT Communication Science Laboratories, Japan

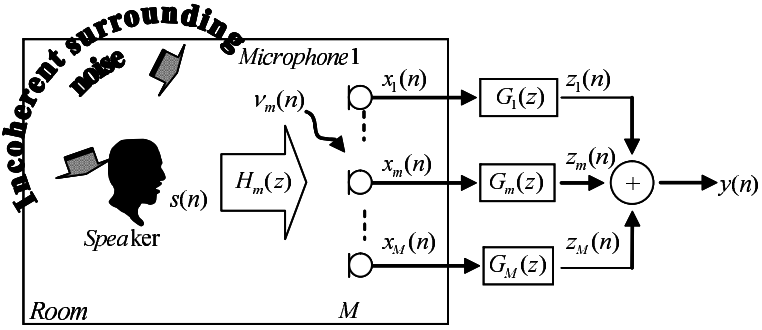


Fig. 9.1 Speech capture model with multiple microphones

In this chapter, we review our four approaches to solving this problem [10, 24, 30, 40]. These approaches are a type of multichannel inverse filtering without the use of ATF replicas as mentioned above. In order to calculate inverse filters, we have commonly used a framework of multichannel Linear Prediction (LP) [14, 35], since this framework can be considered similar to optimal inverse filtering, which may achieve speech dereverberation as well as noise attenuation.

The rest of this chapter is organized as follows. In the next section, we first model a speech capture system with multiple microphones in a noisy and reverberant enclosure. Then, we present an optimal inverse filtering framework followed by multichannel LP as a possible approximation of the framework. The excessive speech-characteristic equalization mentioned above, namely "over-whitening of the target speech", is described in detail. In Sect. 9.3, we review four different approaches to this problem.

1. Precise compensation for over-whitening of the target speech
2. Late reflection removal with multichannel multistep LP
3. Estimation of linear predictors and short-time speech characteristics
4. Probabilistic model based speech dereverberation

9.2 Inverse Filtering for Speech Dereverberation

In this section, we first define a simple acoustic model where a speech signal is captured with multiple microphones in a noisy and reverberant environment. Then, an optimal inverse filtering framework is reviewed in terms of the relation between noise and speech dereverberation performance. Next, we discuss whether multichannel LP can be used to approximate such inverse filtering. Finally, we describe the *over-whitening* effect, which degrades the speech recovered with the algorithm [19].

9.2.1 Speech Capture Model with Multiple Microphones

Let us consider the speech capture system shown in Fig. 9.1. A speech signal $s(n)$ at discrete time index n is captured with a microphone m ($m = 1,2,\ldots,M$) after being reverberated through an ATF $H_m(z)$ between the speaker and the microphone. The microphone simultaneously receives incoherent noise $\nu_m(n)$. Here, incoherent noise is defined as noise for which the source cannot be localized in space. Hence, the sound signals observed at the microphone may be expressed as

$$x_m(n) = h_m(n) * s(n) + \nu_m(n), \tag{9.1}$$

where $*$ denotes linear convolution and impulse response $h_{m,n}$ ($n = 0,1,\ldots,L$) corresponds to the coefficients of ATF $H_m(z)$; $h_{m,0},\ h_{m,1},\ \ldots,$ and $h_{m,L}$. Then, microphone signal $x_m(n)$ is processed through an inverse filter $G_m(z)$ with an impulse response given as $g_{m,n}$ ($n = 0,1,\ldots,p$). This impulse response, like $h_{m,n}$, corresponds to the filter coefficients, $g_{m,0},\ g_{m,1},\ldots,$ and $g_{m,p}$. Thus, the filter output signal may be given as follows:

$$\begin{aligned} z_m(n) &= g_m(n) * x(n) \\ &= g_m(n) * (h_m(n) * s(n) + \nu_m(n)) \\ &= s(n) * h_m(n) * g_m(n) + \nu_m(n) * g_m(n), \end{aligned} \tag{9.2}$$

or, equivalently, in matrix form as

$$\begin{aligned} z_m(n) &= \mathbf{g}_m{}^T \mathbf{x}_n^{(m)} \\ &= \mathbf{s}_n{}^T \mathbf{H}_m \mathbf{g}_m + (\mathbf{v}_n^{(m)})^T \mathbf{g}_m, \end{aligned} \tag{9.3}$$

where symbol T stands for the matrix transpose, and

$$\begin{aligned} \mathbf{x}_n^{(m)} &= [x_m(n), x_m(n-1), \ldots, x_m(n-p)]^T, \\ \mathbf{g}_m &= [g_{m,0}, g_{m,1}, \ldots, g_{m,p}]^T, \\ \mathbf{s}_n &= [s(n), s(n-1), \ldots, s(n-(L+p))]^T, \\ &\quad \leftarrow \quad p+1 \quad \rightarrow \\ \mathbf{H}_m &= \begin{bmatrix} \mathbf{h}_m & & & \\ & \mathbf{h}_m & 0 & \\ & & \ddots & \\ & 0 & & \mathbf{h}_m \end{bmatrix} \begin{matrix} \uparrow \\ L+p+1 \\ \downarrow \end{matrix}, \\ \mathbf{h}_m &= [h_{m,0}, h_{m,1}, \ldots, h_{m,L}]^T, \\ \mathbf{v}_n^{(m)} &= [\nu_m(n), \nu_m(n-1), \ldots, \nu_m(n-p)]^T. \end{aligned}$$

Finally, the sum of the filter output signals is calculated to recover the target speech, $s(n)$:

$$y(n) = \sum_{m=1}^{M} z_m(n) = \sum_{m=1}^{M} \mathbf{g}_m{}^T \mathbf{x}_n^{(m)} = \mathbf{g}^T \mathbf{x}_n, \tag{9.4}$$

or, equivalently,

$$y(n) = \sum_{m=1}^{M} z_m(n) = \sum_{m=1}^{M} \left(\mathbf{s}_n{}^T \mathbf{H}_m \mathbf{g}_m + (\mathbf{v}_n^{(m)})^T \mathbf{g}_m \right) = \mathbf{s}_n{}^T \mathbf{H}\mathbf{g} + \mathbf{v}_n{}^T \mathbf{g}, \tag{9.5}$$

where

$$\begin{aligned}
\mathbf{g} &= [\mathbf{g}_1{}^T, \mathbf{g}_2{}^T, \ldots, \mathbf{g}_M{}^T]^T, \\
\mathbf{x}_n &= [(\mathbf{x}_n^{(1)})^T, (\mathbf{x}_n^{(2)})^T, \ldots, (\mathbf{x}_n^{(M)})^T]^T, \\
\mathbf{H} &= [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_M], \\
\mathbf{v}_n &= [(\mathbf{v}_n^{(1)})^T, (\mathbf{v}_n^{(2)})^T, \ldots, (\mathbf{v}_n^{(M)})^T]^T.
\end{aligned}$$

9.2.2 Optimal Inverse Filtering

Let us consider the following cost function:

$$\begin{aligned}
f_{\text{cost}}[\mathbf{g}] &= E\{|s(n) - y(n)|^2\} \\
&= E\{|s(n) - \mathbf{g}^T \mathbf{x}_n|^2\} \\
&= E\{|s(n) - (\mathbf{s}_n{}^T \mathbf{H}\mathbf{g} + \mathbf{v}_n{}^T \mathbf{g})|^2\} \\
&= E\{|s(n) - \mathbf{s}_n{}^T \mathbf{H}\mathbf{g}|^2\} + E\{|\mathbf{v}_n{}^T \mathbf{g}|^2\},
\end{aligned} \tag{9.6}$$

where the mathematical expectation $E\{a\}$ is interpreted as the time average of a value a, $|a|$ denotes the absolute value, and the covariance matrix $E\{\mathbf{s}_n \mathbf{v}_n{}^T\}$ is assumed to be zero to simplify the last two expressions. In the final expression, the first term on the right-hand side corresponds to the speech dereverberation accuracy achieved with inverse filters $\mathbf{g}$. The second term shows the mean energy value of the incoherent noise processed with the filters. Considering the homogeneous equation obtained by differentiating the cost function with respect to filters $\mathbf{g}$, we can derive optimal inverse filters $\mathbf{g}_{\text{o}}$ as a solution that minimizes the cost function [27] as

$$\begin{aligned}\mathbf{g}_{\mathrm{o}} &= (E\{\mathbf{x}_n\mathbf{x}_n{}^T\})^+E\{\mathbf{x}_n s(n)\}\\ &= (\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^+\mathbf{H}^T E\{\mathbf{s}_n s(n)\}\\ &\approx ((\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+E\{\mathbf{v}_n\mathbf{v}_n{}^T\})+\delta^2\mathbf{I})^{-1}\mathbf{H}^T E\{\mathbf{s}_n s(n)\}\\ &= ((\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})+E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^{-1}\mathbf{H}^T E\{\mathbf{s}_n s(n)\}\\ &= (\mathbf{I}+(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^{-1}\mathbf{g}_{\mathrm{a}}, \end{aligned} \tag{9.7}$$

where $+$ denotes the Moore–Penrose inverse and $\mathbf{I}$ stands for an identity matrix. The approximation from the second expression of (9.7) to the third is valid only when δ is a small positive number [5]. Hereafter, $\mathbf{g}_{\mathrm{a}}$ are called accurate inverse filters. $\mathbf{g}_{\mathrm{a}}$ correspond to the inverse filters derived in the noise-free case and are represented as

$$\begin{aligned}\mathbf{g}_{\mathrm{a}} &= (\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}\mathbf{H}^T E\{\mathbf{s}_n s(n)\}\\ &= (\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n\}\mathbf{1}, \end{aligned} \tag{9.8}$$

where $\mathbf{1}$ stands for a column vector whose elements are zeros except for the first element, which is unity. Inverse filters $\mathbf{g}_{\mathrm{a}}$ may be further simplified as [5]

$$\begin{aligned}\mathbf{g}_{\mathrm{a}} &\approx (\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H})^+\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{1}\\ &= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}(E\{\mathbf{s}_n\mathbf{s}_n{}^T\})^{-1}E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{1}\\ &= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{1}, \end{aligned} \tag{9.9}$$

on condition that the Sylvester matrix $\mathbf{H}$ has a full row rank [18], namely the rank of $\mathbf{H}$ is $L+p+1$. This condition is interpreted as the following conditions based on the Multiple-input/output INverse Theorem (MINT) [29, 32]:

- ATFs $\mathbf{H}_m(z)$ have no common zero
- The inverse filter order is set to satisfy the relation:

$$p \geq \left\lceil \frac{L}{M-1} - 1 \right\rceil,$$

where $\lceil a \rceil$ rounds a number a up to the nearest integer.

Applying observed speech $\mathbf{s}_n{}^T\mathbf{H}$ to accurate inverse filters $\mathbf{g}_{\mathrm{a}}$, we may find that target speech $s(n)$ is recovered as follows:

$$y_{\mathrm{a}}(n) = \mathbf{s}_n{}^T\mathbf{H}\mathbf{g}_{\mathrm{a}} \approx \mathbf{s}_n{}^T\mathbf{H}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{1} = \mathbf{s}_n{}^T\mathbf{1} = s(n). \tag{9.10}$$

According to (9.7), the optimal inverse filters, $\mathbf{g}_{\mathrm{o}}$, may lose their accuracy in inverse filtering (and therefore speech dereverberation) compared with the accurate filters, $\mathbf{g}_{\mathrm{a}}$, because of the incoherent noise power $E\{\mathbf{v}_n\mathbf{v}_n{}^T\}$. On the other hand, the noise may be less amplified with optimal filters $\mathbf{g}_{\mathrm{o}}$ than with accurate filters $\mathbf{g}_{\mathrm{a}}$. Hereafter, we evaluate the effect of the noise on the performance of optimal filters $\mathbf{g}_{\mathrm{o}}$. To simplify the evaluation, the following condition is assumed:

$$r_{\mathrm{NS}} = \|(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\|_2 < 1, \tag{9.11}$$

where $\|\mathbf{A}\|_2$ stands for the spectral norm of a matrix $\mathbf{A}$, which is defined as the maximum eigenvalue of the matrix, $\kappa_{\max}[\mathbf{A}]$. Then, we may evaluate the performance degradation of optimal filters $\mathbf{g}_{\mathrm{o}}$ by comparison with accurate filters $\mathbf{g}_{\mathrm{a}}$ [8];

$$\begin{aligned}
R_g &= \frac{\|\mathbf{g}_{\mathrm{a}}-\mathbf{g}_{\mathrm{o}}\|_2^2}{\|\mathbf{g}_{\mathrm{a}}\|_2^2} \\
&= \frac{\|\{\mathbf{I}-(\mathbf{I}+(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^{-1}\}\mathbf{g}_{\mathrm{a}}\|_2^2}{\|\mathbf{g}_{\mathrm{a}}\|_2^2} \\
&\le \|\mathbf{I}-(\mathbf{I}+(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^{-1}\|_2^2 \\
&\le \left|\frac{\kappa_{\max}\left[(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\right]}{1+\kappa_{\min}\left[(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\right]}\right|^2 \\
&= \left|\frac{r_{\mathrm{NS}}}{1+\kappa_{\min}\left[(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\right]}\right|^2, \qquad (9.12)
\end{aligned}$$

where $\kappa_{\min}[\mathbf{A}]$ denotes the minimum eigenvalue of a square matrix $\mathbf{A}$. Next, the noise amplification with filters $\mathbf{g}_{\mathrm{o}}$ may be evaluated as

$$\begin{aligned}
R_v &= \frac{E\{|\mathbf{v}_n{}^T\mathbf{g}_{\mathrm{o}}|^2\}}{E\{|\mathbf{v}_n{}^T\mathbf{g}_{\mathrm{a}}|^2\}} \\
&= \frac{E\{|\mathbf{v}_n{}^T\{\mathbf{I}+(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\}^{-1}\mathbf{g}_{\mathrm{a}}|^2\}}{E\{|\mathbf{v}_n{}^T\mathbf{g}_{\mathrm{a}}|^2\}} \\
&\le \|(\mathbf{I}+(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^{-1}\|_2^2 \\
&\qquad\qquad \times \|(E\{\mathbf{v}_n\mathbf{v}_n{}^T\})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\|_2 \\
&= \left|\frac{1}{1+\kappa_{\min}\left[(\mathbf{H}^T E\{\mathbf{s}_n\mathbf{s}_n{}^T\}\mathbf{H}+\delta^2\mathbf{I})^{-1}E\{\mathbf{v}_n\mathbf{v}_n{}^T\}\right]}\right|^2 \quad (\le 1). \qquad (9.13)
\end{aligned}$$

Here, covariance matrix $E\{\mathbf{v}_n\mathbf{v}_n{}^T\}$ is assumed to be positive, and the following relation is utilized for simplifying the third expressions [18]:

$$\frac{\mathbf{x}^T\mathbf{A}^T\mathbf{D}\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{B}\mathbf{x}} \le \|\mathbf{A}^T\mathbf{D}\mathbf{A}\mathbf{B}^{-1}\|_2 \le \|\mathbf{A}^T\|_2\|\mathbf{A}\mathbf{B}^{-1}\mathbf{D}\|_2 \le \|\mathbf{A}^T\|_2^2\|\mathbf{B}^{-1}\mathbf{D}\|_2,$$

where $\mathbf{x}$ denotes a column vector, $\mathbf{A}$, $\mathbf{D}$ are Hermitian matrices, and $\mathbf{B}$ should be a positive Hermitian matrix.

As shown in (9.13), optimal inverse filters $\mathbf{g}_{\mathrm{o}}$ may be less sensitive to incoherent noise than the accurate filters $\mathbf{g}_{\mathrm{a}}$ (see also Appendix A), although the inverse filtering (and hence speech dereverberation) performance of $\mathbf{g}_{\mathrm{o}}$ may be inferior to that of $\mathbf{g}_{\mathrm{a}}$, as shown in (9.12). To increase the dereverberation accuracy of the optimal filters, we need to reduce incoherent noise prior to optimal inverse filtering.

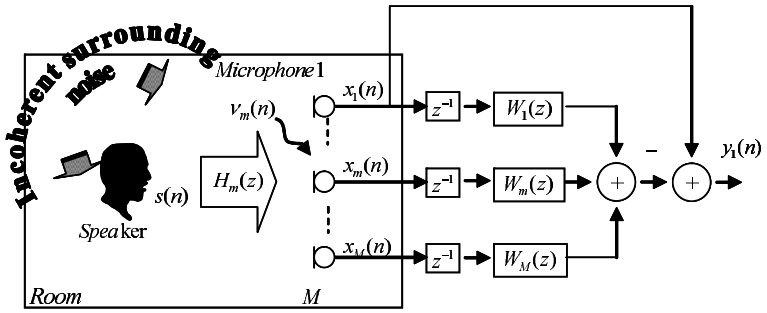


Fig. 9.2 Schematic diagram of an inverse filtering system with M microphones based on the multichannel LP

9.2.3 Unsupervised Algorithm to Approximate Optimal Processing

To achieve dereverberation with optimal inverse filters $\mathbf{g}_{\mathrm{o}}$ given as (9.7), we would need to calculate such filters using *a priori* knowledge of the speech $s(n)$. However, there are many situations where $s(n)$ is unknown and therefore we should approximate the optimal filters from only the signals observed at the microphones. Hereafter, we examine whether multichannel linear prediction [14, 35] has the potential to be used as an algorithm for such an approximation.

Figure 9.2 is a schematic diagram of an inverse filtering system with M microphones based on multichannel LP. A microphone signal $x_m(n)$ is delayed by one sample and processed through prediction filter $W_m(z)$ whose impulse response is given as $w_{m,n}$ $(n = 0, 1, \ldots, p)$. Then, the sum of the output signals of M predictors is subtracted from the non-delayed version of microphone signal $x_1(n)$. Here, microphone 1 is assumed to be the closest microphone to the speaker. Hence, the system output signal may be expressed as

$$\begin{aligned} y_1(n) &= x_1(n) - \sum_{m=1}^{M} w_m(n) * x_m(n-1) \\ &= x_1(n) - \sum_{m=1}^{M} \left(\mathbf{s}_{n-1}{}^T \mathbf{H}_m \mathbf{w}_m + (\mathbf{v}_{n-1}^{(m)})^T \mathbf{w}_m \right) \\ &= \left(\mathbf{s}_n{}^T \mathbf{h}_1 + v_1(n) \right) - \left(\mathbf{s}_{n-1}{}^T \mathbf{H}\mathbf{w} + \mathbf{v}_{n-1}{}^T \mathbf{w} \right), \end{aligned} \tag{9.14}$$

and

$$\begin{aligned} \mathbf{w}_m &= [w_{m,0}, w_{m,1}, \ldots, w_{m,p}]^T, \\ \mathbf{w} &= [\mathbf{w}_1{}^T, \mathbf{w}_2{}^T, \ldots, \mathbf{w}_M{}^T]^T. \end{aligned}$$

Since covariance matrix $E\{\mathbf{s}_n\mathbf{v}_n{}^T\}$ is assumed to be zero (see (9.6)), the mean energy of the system output signal may be expressed as

$$\begin{aligned} f_{\text{cost}}[\mathbf{w}] &= E\{|y_1(n)|^2\} \\ &= E\{|(\mathbf{s}_n{}^T\mathbf{h}_1 + v_1(n)) - (\mathbf{s}_{n-1}{}^T\mathbf{H}\mathbf{w} + \mathbf{v}_{n-1}{}^T\mathbf{w})|^2\} \\ &= E\{|\mathbf{s}_n{}^T\mathbf{h}_1 - \mathbf{s}_{n-1}{}^T\mathbf{H}\mathbf{w}|^2\} + E\{|v_1(n) - \mathbf{v}_{n-1}{}^T\mathbf{w}|^2\} \\ &= E\{|\mathbf{s}_n{}^T\mathbf{h}_1 - \mathbf{s}_{n-1}{}^T\mathbf{H}\mathbf{w}|^2\} + E\{|v_1(n)|^2\} + E\{|\mathbf{v}_{n-1}{}^T\mathbf{w}|^2\}, \end{aligned} \tag{9.15}$$

where covariance vector $E\{v_1(n)\mathbf{v}_{n-1}{}^T\}$ is also assumed to be zero. Here, the first term corresponds to the cost function of multichannel LP, which is intended to estimate the observed speech at microphone 1, $\mathbf{s}_n{}^T\mathbf{h}_1$, from all the observed speech signals $\mathbf{s}_{n-1}{}^T\mathbf{H}$. The second term shows the mean energy value of incoherent noise processed with predictors $\mathbf{w}$. Thus, by minimizing cost function $f_{\text{cost}}[\mathbf{w}]$, the optimal predictors may be calculated as

$$\begin{aligned} \mathbf{w}_{\text{o}} &= (E\{\mathbf{x}_{n-1}\mathbf{x}_{n-1}{}^T\})^+ E\{\mathbf{x}_{n-1}x(n)\} \\ &= (\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\}\mathbf{H} + E\{\mathbf{v}_{n-1}\mathbf{v}_{n-1}{}^T\})^+ \\ &\qquad \times \mathbf{H}^T E\{\mathbf{s}_{n-1}(\mathbf{s}_n{}^T\mathbf{h}_1 + v(n))\} \\ &\approx ((\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\}\mathbf{H} + \delta^2\mathbf{I}) + E\{\mathbf{v}_{n-1}\mathbf{v}_{n-1}{}^T\})^{-1} \\ &\qquad \times \mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_n{}^T\}\mathbf{h}_1 \\ &= (\mathbf{I} + (\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\}\mathbf{H} + \delta^2\mathbf{I})^{-1} E\{\mathbf{v}_{n-1}\mathbf{v}_{n-1}{}^T\})^{-1}\mathbf{w}_{\text{a}}, \end{aligned} \tag{9.16}$$

where

$$\mathbf{w}_{\text{a}} = (\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\}\mathbf{H} + \delta^2\mathbf{I})^{-1}\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_n{}^T\}\mathbf{h}_1. \tag{9.17}$$

Here, $\mathbf{w}_{\text{a}}$ represents a set of accurate predictors that minimizes only the first term of cost function $f_{\text{cost}}[\mathbf{w}]$. Comparing $\mathbf{w}_{\text{o}}$ with optimal inverse filters $\mathbf{g}_{\text{o}}$ shown in (9.7), we notice that the effect of incoherent noise on $\mathbf{w}_{\text{a}}$ is similar to that on accurate inverse filters $\mathbf{g}_{\text{a}}$.

Next, let us evaluate the speech dereverberation performance of accurate predictors $\mathbf{w}_{\text{a}}$. Assuming the full row-rank condition of the Sylvester matrix $\mathbf{H}$ (MINT condition), we may simplify the predictors to (see (9.9))

$$\begin{aligned} \mathbf{w}_{\text{a}} &\approx \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}(E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\})^{-1}E\{\mathbf{s}_{n-1}\mathbf{s}_n{}^T\}\mathbf{h}_1 \\ &= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{h}_1, \end{aligned} \tag{9.18}$$

where

$$\mathbf{C} = (E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\})^{-1}E\{\mathbf{s}_{n-1}\mathbf{s}_n{}^T\}$$
$$= \begin{bmatrix} c_1 & 1 & 0 & \cdots & 0 \\ c_2 & 0 & 1 & & \vdots \\ \vdots & \vdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ c_{L+p+1} & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{matrix} \uparrow \\ \\ L+p+1 \\ \\ \downarrow \end{matrix}, \tag{9.19}$$
$$\leftarrow \quad L+p+1 \quad \rightarrow$$

c_j denotes the coefficients of Autoregressive (AR) polynomial [22] given as

$$C(z) = 1 - \{c_1 z^{-1} + c_2 z^{-2} + \ldots + c_{L+p+1} z^{-(L+p+1)}\}$$
$$= \sum_{j=0}^{L+p+1} c_j z^{-j}, \;\; c_0 = 1. \tag{9.20}$$

Applying accurate predictors $\mathbf{w}_\mathrm{a}$ to one-sample delayed speech signals, $\mathbf{s}_{n-1}{}^T\mathbf{H}$, and then subtracting the resulting signal from observed speech $\mathbf{s}_n{}^T\mathbf{h}_1$, we may obtain prediction residual $y_1(n)$ as follows:

$$\begin{aligned} y_1(n) &= \mathbf{s}_n{}^T\mathbf{h}_1 - \mathbf{s}_{n-1}{}^T\mathbf{H}\mathbf{w} \\ &= \mathbf{s}_n{}^T\mathbf{h}_1 - \mathbf{s}_n{}^T\mathbf{H}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{h}_1 \\ &= (\mathbf{s}_n{}^T - \mathbf{s}_n{}^T\mathbf{C})\mathbf{h}_1 \\ &= \left(s(n) - \sum_{j=1}^{L+p+1} c_j s(n-j)\right) h_0^{(1)} \\ &= h_0^{(1)} \sum_{j=0}^{L+p+1} c_j s(n-j) \\ &\propto c(n) * s(n). \end{aligned} \tag{9.21}$$

This relation shows that the effect of reverberation on the speech observed at microphone 1, which is caused by ATF $H_1(z)$, may be precisely removed, but simultaneously the recovered speech may be degraded by its own autocorrelation. This degradation is known as *over-whitening*, and may occur in other algorithms [2, 36].

9.3 Approaches to Solving the Over-whitening of the Recovered Speech

As discussed above, multichannel LP may deal with incoherent noise in almost the same manner as the optimal inverse filtering described in Sect. 9.2.2. However, the recovered speech will suffer from over-whitening effects caused by the autocorrelation of the speech signal. To adopt multichannel LP as an approximation of the optimal inverse filtering, we need to find a way to reduce such harmful effects on the recovered speech.

In this section, we describe four different approaches to solving the effects of over-whitening on the recovered speech with multichannel LP [10, 24, 30, 40].

9.3.1 Precise Compensation for Over-whitening of Target Speech

If we could estimate the AR polynomial, $C(z)$, shown in (9.20) solely from observed speech signals $\mathbf{s}_{n-1}^T\mathbf{H}$, we would be able to recover the target speech by applying the inverse of the estimated AR polynomial, $1/\hat{C}(z)$, to the prediction residual, $y_1(n)$. There have been a few attempts to calculate such an estimate as a characteristic common to the observed speech signals [28, 33]. For example, the inverse AR polynomial, $1/C(z)$, may be estimated as the greatest common divisor (GCD) of the signal transmission channels between the speaker and the microphones. It is shown in [33] that such GCD could be obtained from the covariance matrix of the observed speech signals based on the subspace method.

In this section, we review another recently proposed approach [10, 28].

9.3.1.1 Principle

Let us consider a prediction matrix $\mathbf{Q}$ defined as follows [10, 28].

$$\begin{aligned}\mathbf{Q} &= (\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\}\mathbf{H} + \delta^2\mathbf{I})^{-1}\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_n{}^T\}\mathbf{H} \\ &\approx \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}(E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}{}^T\})^{-1}E\{\mathbf{s}_{n-1}\mathbf{s}_n{}^T\}\mathbf{H} \\ &= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{H}.\end{aligned} \tag{9.22}$$

This definition is obtained by replacing the speech signal observed at microphone 1, $\mathbf{s}_{n-1}^T\mathbf{H}_1$, with observed speech signals $\mathbf{s}_{n-1}^T\mathbf{H}$ in (9.18). Hence, the first column vector of $\mathbf{Q}$ is equivalent to accurate predictors $\mathbf{w}_a$. Then, we may find the following relation between non-zero eigenvalues κ_j of matrices $\mathbf{Q}$ and $\mathbf{C}$ [18]:

$$\begin{aligned}\kappa_j[\mathbf{Q}] &= \kappa_j[\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{H}] \\ &= \kappa_j[\mathbf{H}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}] \\ &= \kappa_j[\mathbf{C}]. \end{aligned} \tag{9.23}$$

Here, we deduce from this relation that the characteristic polynomials of $\mathbf{Q}$ and $\mathbf{C}$ are equivalent. The characteristic polynomial of companion matrix $\mathbf{C}$ may be calculated as

$$\begin{aligned} f_{\text{char}}[\mathbf{Q}] &= f_{\text{char}}[\mathbf{C}] \\ &= \det[\kappa\mathbf{I} - \mathbf{C}] \\ &= \kappa^{L+p+1} - (c_1\kappa^{L+p} + c_2\kappa^{L+p-1} + \ldots + c_{L+p+1}) \\ &= \kappa^{L+p+1}\sum_{j=0}^{L+p+1} c_j\kappa^{-j}. \end{aligned} \tag{9.24}$$

Comparing this relation with (9.20), we find that the coefficients of the AR polynomial $C(z)$ are equivalent to those of the characteristic polynomial of $\mathbf{C}$. Hence, we may deduce that an estimated AR polynomial $\hat{C}(z)$ can be obtained from the characteristic polynomial of prediction matrix $\mathbf{Q}$. Applying the inverse of such an estimated AR polynomial, $1/\hat{C}(z)$, to prediction residual $y_1(n)$, we recover the target speech signal as

$$\begin{aligned}\hat{s}(n) &= [1/\hat{C}(z)]y_1(n) \\ &= [1/\hat{C}(z)][C(z)]s(n) \\ &\approx s(n). \end{aligned} \tag{9.25}$$

Figure 9.3 is a schematic diagram of the multichannel inverse filtering system based on the procedure described above. The whole algorithm may be summarized as follows:

1. First, prediction matrix $\mathbf{Q}$ is calculated according to (9.22).
2. Prediction residual $y_1(n)$ is calculated by using (9.21), where accurate predictors $\mathbf{w}_a$ are given as the first column vector of $\mathbf{Q}$.
3. Simultaneously, estimated AR polynomial $C(z)$ is obtained from the characteristic polynomial of $\mathbf{Q}$ [34].
4. The target speech is recovered by applying the inverse of the estimated polynomial, $1/\hat{C}(z)$, to $y_1(n)$.

Below, we discuss three issues related to the proposed algorithm, which can be summarized as follows:

1. Close to perfect dereverberation
2. Dereverberation and coherent noise reduction
3. Sensitivity to incoherent noise

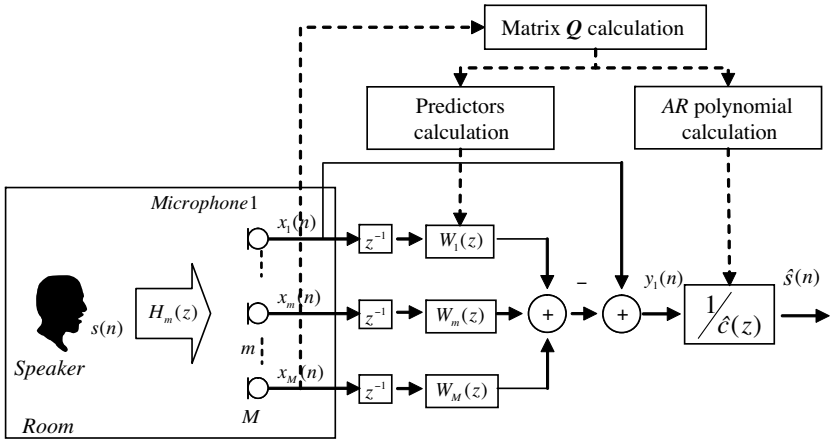


Fig. 9.3 Schematic diagram of proposed algorithm

9.3.1.2 Close to Perfect Dereverberation

A speech dereverberation experiment was conducted to confirm the high performance of the algorithm by using ATFs measured in an experimental chamber with a volume of about 40 m^3 and a reverberation time (T_{60}) of about 0.5 s. Up to 6 microphones were used. The distance between the loudspeaker and the microphones was set at about 4 m. The microphone signals were simulated by convolving the ATFs with speech taken from the ATR database [1]. The sampling rate was 8 kHz. The experimental results are shown in Figs. 9.4 and 9.5.

Figure 9.4 plots the energy density curves of an ATF and corresponding equalized ATFs using the proposed algorithm with three, four and six microphones. The original reverberation energy is attenuated by more than 20 dB. Comparing the curves for three, four and six microphones, we notice that the algorithm benefits from spatial information provided by increasing the number of microphones. The worse performance obtained with three microphones may be explained by the presence of overlapping zeros among all the ATFs. These zeros cannot be compensated by the algorithm and are therefore responsible for the remaining distortions. Increasing the spatial information may reduce the probability of overlapping zeros, and therefore the dereverberation performance improves. The effect of spatial information on the algorithm is discussed in more detail in [10].

Figures 9.5 plots spectrograms of the target clean speech, the observed reverberant speech, and the recovered speech when six microphones were used. We can see that the reverberation effect is completely removed and the recovered speech is very close to the target speech. These experimental results demonstrate that the proposed algorithm may achieve close to perfect speech dereverberation as suggested by (9.25).

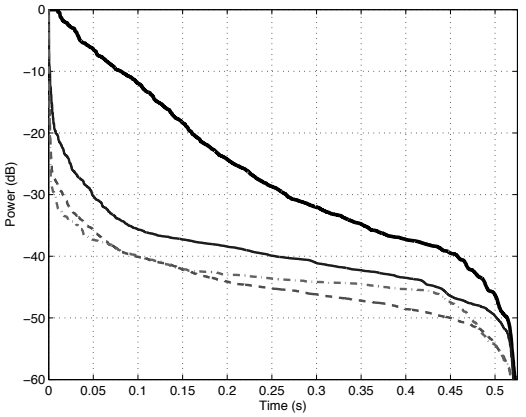


Fig. 9.4 Energy decay curve of the original ATF (thick solid line) and dereverberated ATF using three (*thin solid line*), four (*dashed line*), and six microphones (*dash-dotted line*)

9.3.1.3 Dereverberation and Coherent Noise Reduction

Let us consider an acoustic system with M ($M \geq 3$) microphones that receive coherent noise in addition to the target speech. Here, we define coherent noise as an undesired signal whose source can be localized in space.[1] We assume that there is at least one microphone closer to the speaker than to the noise source and choose as a reference the microphone that is the closest to the speaker but not the closest to the noise source. This reference microphone is hereafter called microphone 1. Figure 9.6 shows an example of the disposition of sources and microphones.

As above, the target speech is denoted as $s(n)$. The coherent noise signal is denoted as $\nu_c(n)$. We denote the ATFs between the speaker and the microphones as $H_{s,m}(z)$ ($m = 1,2,\ldots,M$), and the AR polynomial associated with the target speech as $C_s(z)$. Similarly, we denote the ATFs between the noise source and the microphones as $H_{\nu,m}(z)$, and the AR polynomial associated with noise signal $\nu_c(n)$ as $C_n(z)$. We can generalize the expression for the convolution matrix $\mathbf{H}$ for 2 sources as $\mathbf{H}^{2\text{src}} = \begin{bmatrix} \mathbf{H}_{s,1} & \cdots & \mathbf{H}_{s,M} \\ \mathbf{H}_{\nu,1} & \cdots & \mathbf{H}_{\nu,M} \end{bmatrix}$, and define a source signal vector as

$$\begin{aligned} \mathbf{s}_n^{2\text{src}} &= [s(n),\ldots,s(n-(L+p)),\nu_c(n),\ldots,\nu_c(n-(L+p))]^T \\ &= [\mathbf{s}_n^T,\mathbf{v}_{c,n}^T]^T. \end{aligned}$$

We can derive a prediction matrix $\mathbf{Q}^{2\text{src}}$ simply by replacing $\mathbf{H}$ and $\mathbf{s}$ with $\mathbf{H}^{2\text{src}}$ and $\mathbf{s}^{2\text{src}}$, respectively, in (9.22) [9]. Then, we can derive the following residual, $y_1(n)$,

[1] Here we consider one coherent source, although the discussion could be extended to more noise sources.

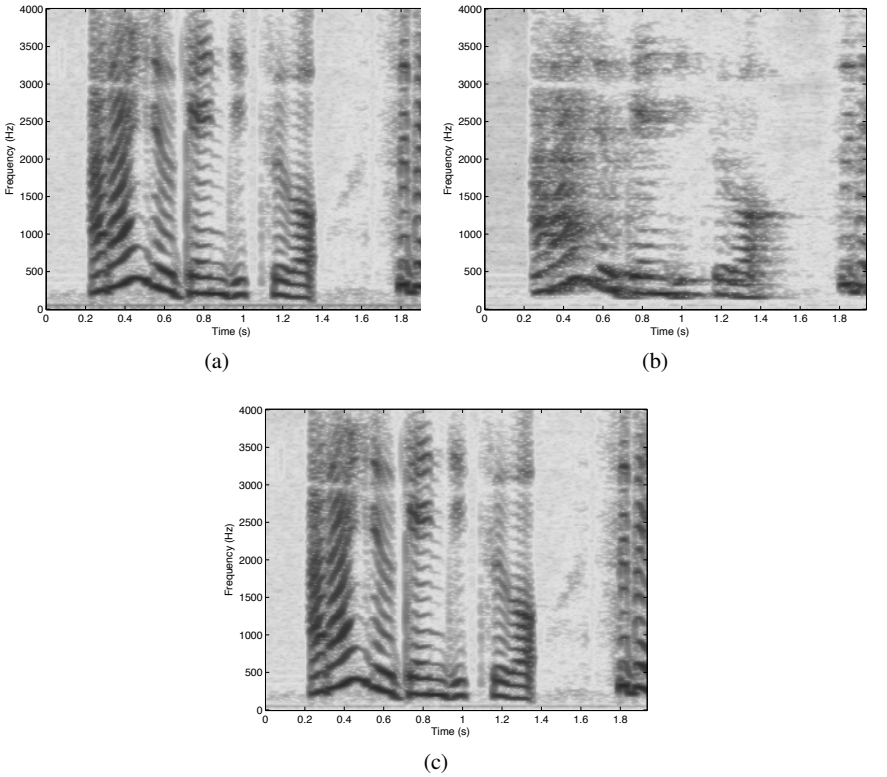


Fig. 9.5 Spectrograms of (a) the target signal, (b) the reverberant signal and (c) the recovered signal using six microphones. The impulse responses were measured in a real room with a reverberation time of 0.5 s

referring to (9.21):

$$\begin{aligned} y_1(n) &= (\mathbf{s}_n^{2\mathrm{src}})^T \mathbf{h}_1^{2\mathrm{src}} - \mathbf{s}_{n-1}^{2\mathrm{src},T} \mathbf{H}^{2\mathrm{src}} \mathbf{w}^{2\mathrm{src}} \\ &= ((\mathbf{s}_n^{2\mathrm{src}})^T - \mathbf{s}_{n-1}^{2\mathrm{src},T} \mathbf{C}^{2\mathrm{src}}) \mathbf{h}_1 \\ &= h_{s1}^{(0)}[C_s(z)]s(n) + h_{v1}^{(0)}[C_v(z)]v_c(n) \\ &\approx h_{s1}^{(0)}[C_s(z)]s(n), \end{aligned} \tag{9.26}$$

where $\mathbf{C}^{2\mathrm{src}} = (E\{\mathbf{s}_{n-1}^{2\mathrm{src}}(\mathbf{s}_{n-1}^{2\mathrm{src}})^T\})^{-1} E\{\mathbf{s}_{n-1}^{2\mathrm{src}}(\mathbf{s}_n^{2\mathrm{src}})^T\}$, and assuming that the noise and the target speech are uncorrelated, matrix $\mathbf{C}^{2\mathrm{src}}$ may be expressed as:

$$\mathbf{C}^{2\mathrm{src}} = \begin{bmatrix} \mathbf{C}_s & 0 \\ 0 & \mathbf{C}_v \end{bmatrix}, \tag{9.27}$$

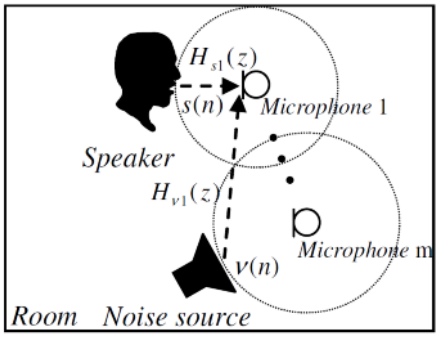


Fig. 9.6 Example of the disposition of the sources and microphones in the room

where

$$\mathbf{C}_s = (E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}^T\})^{-1}E\{\mathbf{s}_{n-1}\mathbf{s}_n^T\}$$
$$\mathbf{C}_v = (E\{\mathbf{v}_{c,n-1}\mathbf{v}_{c,n-1}^T\})^{-1}E\{\mathbf{v}_{c,n-1}\mathbf{v}_{c,n}^T\}.$$

According to the hypothesis that microphone 1 is closer to the speaker than the noise source, noise $v_c(n)$ will arrive at microphone 1 after target speech $s(n)$, and therefore we can consider that $h_{v1}^{(0)} = 0$. Here, the noise reduction shown in the simplification from the second to the third line of (9.26) may be intuitively understood. According to the above assumption regarding the microphone and noise source disposition, microphone 1 is not the microphone closest to the noise source, and therefore noise $v_c(n)$ arrives first at other microphones. Thus, the predictors $\mathbf{w}^{2\text{src}}$ may use the observations at these microphones to produce a replica of the noise included in the signal observed at microphone 1 (based on MINT). Then, the noise is cancelled out by subtracting this replica from the signal at microphone 1.

The predictors suppress the effect of room reverberation and remove the interference emanating from the noise source. We therefore observe that, based on the assumptions regarding the source and microphone positions, the presence of a noise source does not affect the prediction residual. As before, the target speech could be recovered by filtering the residual with the inverse of the estimated target source AR polynomial, $1/\hat{C}_s(z)$.

Let us look at the effect of the noise source on the estimated AR polynomial obtained from the characteristic polynomial of the prediction matrix as shown in (9.24). It is easy to show that the characteristic polynomial of $\mathbf{Q}^{2\text{src}}$ becomes the product of both source AR polynomials as [18]:

$$f_{\text{char}}[\mathbf{Q}^{2\text{src}}] = f_{\text{char}}[\mathbf{C}^{2\text{src}}] = C_s(z)C_v(z). \tag{9.28}$$

As seen in (9.28), the estimated AR polynomial is affected by the noise AR polynomial. If such an AR polynomial were used, the recovered signal would be degraded.

Note that if the noise source is white, $C_v(z) = 1$, the estimated AR polynomial is equivalent to the AR polynomial of the target source, $C_s(z)$. If the noise source is colored, we need to eliminate the effect of the noise AR polynomial. Here we assume that the noise is stationary and that there are periods when only the noise source is active. These periods could be determined by using a voice activity detection method [26]. During one of these periods, we can estimate the AR polynomial of the noise source, $\hat{C}_v(z)$ by using the proposed algorithm for a single source. Then by pre-filtering the observed microphone signals with $\hat{C}_v(z)$, the noise is pre-whitened and therefore, the effect of $C_v(z)$ is removed from the estimated AR polynomial given by the characteristic polynomial of the prediction matrix [9].

We used the proposed method for dereverberating speech in the presence of colored noise. The room impulse responses were generated by the image method [3]. The simulated room was designed with a reverberation time of 0.5 s. We truncated the impulse responses to 0.2 s in order to simplify the computation involved in the experiments. The computational complexity of this simulation was about the same as that involved in performing single-source dereverberation in a room with a room impulse response duration of 0.4 s. The colored noise was generated by applying a 30 tap long AR process to white noise. The speech signals consisted of 4 s utterances by a male and a female speaker drawn from the ATR database [1]. The sampling rate was 8 kHz.

To measure both the dereverberation and noise reduction performance we decompose the output of the algorithm as follows:

$$\hat{s}(n) = \hat{s}_D(n) + \hat{s}_I(n), \tag{9.29}$$

where $\hat{s}_D(n)$ corresponds to the dereverberated speech and $\hat{s}_I(n)$ corresponds to the remaining interference. In the experiments, $\hat{s}_D(n)$ is obtained by applying the predictors and estimated AR process only to the signals emanating from the target source, $x_{s,m}(n)$ $(m = 1, \ldots, M)$. Similarly, $\hat{s}_I(n)$ is obtained by using only the signals emanating from the noise source, $x_{v,m}(n)$ $(m = 1, \ldots, M)$. We use the input and output Signal to Distortion Ratio (SDR) to evaluate the dereverberation performance:

$$\mathrm{SDR_{In}} = 10\log_{10}\left(\frac{\sum |s(n)|^2}{\sum |s(n) - x_{s,m}(n)|^2}\right), \tag{9.30}$$

$$\mathrm{SDR_{Out}} = 10\log_{10}\left(\frac{\sum |s(n)|^2}{\sum |s(n) - \hat{s}_D(n)|^2}\right). \tag{9.31}$$

The noise reduction performance is measured by the input and output Signal to Noise Ratio (SNR):

$$\mathrm{SNR_{In}} = 10\log_{10}\left(\frac{\sum |x_{s,m}(n)|^2}{\sum |x_{v,m}(n)|^2}\right), \tag{9.32}$$

$$\mathrm{SNR_{Out}} = 10\log_{10}\left(\frac{\sum |\hat{s}_D(n)|^2}{\sum |\hat{s}_I(n)|^2}\right). \tag{9.33}$$

Table 9.1 Results using for female and male speakers for 4 seconds of speech

			Proposed		DSB	
	SDR_{In}	SNR_{In}	SDR_{Out}	SNR_{Out}	SDR_{Out}	SNR_{Out}
Female	1 dB	0 dB	22 dB	11 dB	1 dB	5 dB
	1 dB	10 dB	22 dB	11 dB	1 dB	15 dB
Male	0 dB	0 dB	22 dB	11 dB	2 dB	5 dB
	0 dB	10 dB	22 dB	11 dB	2 dB	15 dB

Table 9.1 shows the results obtained for a female and a male speaker for input SDRs of 0 and 10 dB using the proposed method, and a conventional Delay-and-sum Beamformer (DSB) [19]. The second column of Table 9.1 shows that with the proposed method, the predictors successfully reduce both the reverberation and the noise. Interestingly, the same output SNR and SDR values were obtained when the input SDR was in the -5 to 15 dB range [9]. The noise reduction performance can be greatly improved by using longer observation data. For example, using 10 s of observation data, the output SNR was increased to 16 dB. For comparison, we also show the results we obtained with a conventional DSB. DSB is one approach frequently used to remove spatially localized noise. We observe that the DSB reduces the noise by around 5 dB but has little effect on reducing reverberation. To show the optimum performance of the DSB, we assumed that the time delays were known beforehand. In practice, time-delay estimation under such noisy and reverberant conditions may be difficult and poorer performance would be expected. Note that with the proposed algorithm, time-delay estimation is not needed.

This experiment proves that the proposed algorithm could achieve both the dereverberation and reduction of coherent noise.

9.3.1.4 Sensitivity to Incoherent Noise

We demonstrated the robustness of the proposed method to coherent noise sources. Here, we discuss the effect of incoherent noise on the algorithm. The proposed algorithm relies on the computation of the predictors and speech AR polynomial. The calculation of the predictors is relatively robust as regards incoherent noise for input SNRs higher than 20 dB. Incoherent noise mainly affects the accuracy of the computation of the AR polynomial. Recall that the AR polynomial can be estimated from the characteristic polynomial of prediction matrix $\mathbf{Q}$. This result was demonstrated theoretically, and it relies on the fact that the covariance matrix of the observed signals is rank deficient. If there is no incoherent noise, the covariance matrix can be expressed as:

$$\mathbf{R} = \mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}^T\}\mathbf{H}. \tag{9.34}$$

Since $\mathbf{H}$ has more columns than rows, $\mathbf{R}$ is rank deficient. In the presence of incoherent noise, the covariance matrix becomes the sum of the covariance matrices of the reverberant source signals and the incoherent noise:

$$\mathbf{R} = \mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}^T\}\mathbf{H} + E\{\mathbf{v}_{n-1}\mathbf{v}_{n-1}^T\}, \tag{9.35}$$

where $E\{\mathbf{v}_{n-1}\mathbf{v}_{n-1}^T\}$ is the noise covariance matrix. In this case, it can be easily seen that $\mathbf{R}$ has a full rank. For example, when much observation data are available, the covariance matrix of the noise tends to be a unit matrix multiplied by a scalar , σ^2, equivalent to the noise variance. One of the effects of the noise is to add the scalar, σ^2, to the eigenvalues of the covariance matrix of the observed signals, which therefore becomes full rank. In this case, the AR polynomial may not be obtained accurately. In theory, σ^2 can be estimated as the smallest eigenvalue of the covariance matrix of the observed signals. Consequently, a prediction matrix that would not be affected by the noise could be calculated as:

$$\begin{aligned}\mathbf{Q} = &(\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_{n-1}^T\}\mathbf{H} + E\{\mathbf{v}_{n-1}\mathbf{v}_{n-1}^T\} - \sigma^2\mathbf{I})^+ \\ &(\mathbf{H}^T E\{\mathbf{s}_{n-1}\mathbf{s}_n^T\}\mathbf{H} + E\{\mathbf{v}_{n-1}\mathbf{v}_n^T\} - \sigma^2\boldsymbol{\Upsilon}),\end{aligned} \tag{9.36}$$

where $\boldsymbol{\Upsilon}$ corresponds to a block shifting matrix defined as:

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \mathbf{D} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{D} \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \vdots \\ \vdots & \vdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Using (9.36) the AR polynomial was obtained precisely. Such an approach may require much observation data in order to approximate the noise covariance matrix with a unit matrix. The estimation of the covariance matrix of incoherent noise with limited observation data remains an issue.

9.3.2 Late Reflection Removal with Multichannel Multistep LP

Here we introduce another approach to solving the over-whitening effects. If we could assume a certain time lag, τ_o, after which the autocorrelation of the target speech becomes fairly small, we would be able to reduce the effects of over-whitening on the direct sound and subsequent early reflections by replacing unit delays with τ-sample ($\tau \geq \tau_o + 1$) delays as shown in Fig. 9.2. With this idea, some coloration caused by the early reflections may characterize the prediction residual as recovered speech, but severe reverberation arising from the late reflections may be cancelled out.

In Automatic Speech Recognition (ASR) [12], for example, such severe reverberation is considered more problematic rather than the coloration caused by early reflections [15]. Hence, there have been a number of studies aimed at mitigating the reverberant effect from late reflections [17, 37]. In most of these studies, the late re-

flections are estimated based on the assumption that the reverberation energy curve in a room decays exponentially, and then subtracted from the observed speech in the power spectrum domain. However, since late reflection energy may not be well estimated solely with such an exponential model, sufficient improvement in the ASR performance has yet to be achieved.

In this section, we review a dereverberation algorithm based on multichannel LP with τ-sample delay units, which may achieve a better estimate of late reflections [24].

9.3.2.1 Principle

Let us consider a multichannel LP system with M microphones. The locations of the microphones and speaker are assumed to be the same as those shown in Fig. 9.2. The observed speech at microphone 1, $\mathbf{s}_n{}^T\mathbf{h}_1$, is predicted with the observed speech signals, $\mathbf{s}_n{}^T\mathbf{H}$, processed with the abovementioned τ-sample delay units. Thus, referring to (9.18) to (9.21), we may express τ-*step* predictors $\mathbf{w}_\tau$ as [25]

$$\begin{aligned}\mathbf{w}_\tau &= (\mathbf{H}^T E\{\mathbf{s}_{n-\tau}\mathbf{s}_{n-\tau}{}^T\}\mathbf{H} + \delta^2\mathbf{I})^{-1}\mathbf{H}^T E\{\mathbf{s}_{n-\tau}\mathbf{s}_n{}^T\}\mathbf{h}_1 \\ &\approx (\mathbf{H}^T E\{\mathbf{s}_{n-\tau}\mathbf{s}_{n-\tau}{}^T\}\mathbf{H})^{+}\mathbf{H}^T E\{\mathbf{s}_{n-\tau}\mathbf{s}_n{}^T\}\mathbf{h}_1. \end{aligned} \tag{9.37}$$

If we could assume that the Sylvester matrix $\mathbf{H}$ has a full row rank (see (9.8) and (9.9)), predictors $\mathbf{w}_\tau$ would be simplified as

$$\begin{aligned}\mathbf{w}_\tau &\approx \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}E\{\mathbf{s}_{n-\tau}\mathbf{s}_{n-\tau}{}^T\}^{-1}E\{\mathbf{s}_{n-\tau}\mathbf{s}_n{}^T\}\mathbf{h}_1 \\ &= \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}_\tau\mathbf{h}_1, \end{aligned} \tag{9.38}$$

where

$$\mathbf{C}_\tau = \begin{array}{c} \begin{array}{ccccc} \leftarrow & & L+p+1 & & \rightarrow \\ \leftarrow & & \tau & \rightarrow & \end{array} \\ \left[\begin{array}{ccccc|c} \cdots & 0 & c^{(\tau)}_{\tau_o,1} & \cdots & c^{(\tau)}_{1,1} & \\ & \vdots & 0 & \ddots & \vdots & \mathbf{I} \\ & & \vdots & \ddots & c^{(\tau)}_{1,\tau_o} & \\ & & & & 0 & \hline \\ & & & & \vdots & 0 \end{array}\right] \end{array} \begin{array}{c} \uparrow \\ \\ L+p+1 \\ \\ \downarrow \end{array} . \tag{9.39}$$

$c^{(\tau)}_{j,k}$ denotes the coefficients of a j-step AR polynomial given as

$$C_j^{(\tau)}(z) = 1 - \sum_{k=1}^{L+p+1} c_{j,k}^{(\tau)} z^{-(j+k-1)}. \tag{9.40}$$

We assume that $c_{j,k}^{(\tau)}$ can be neglected when $\tau_o + 1 < j + k$, thus denoted as 0 in (9.39). Applying τ-*step* predictors $\mathbf{w}_\tau$ to τ-sample delayed speech signals ${\mathbf{s}_{n-\tau}}^T\mathbf{H}$, and then subtracting the resulting signal from observed speech ${\mathbf{s}_n}^T\mathbf{h}_1$, we obtain the τ-*step* prediction residual $y_\tau^{(1)}(n)$ as follows.

$$\begin{aligned}
y_\tau^{(1)}(n) &= {\mathbf{s}_n}^T\mathbf{h}_1 - {\mathbf{s}_{n-\tau}}^T\mathbf{H}\mathbf{w}_\tau \\
&= {\mathbf{s}_n}^T\mathbf{h}_1 - {\mathbf{s}_{n-\tau}}^T\mathbf{H}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^T\mathbf{C}_\tau\mathbf{h}_1 \\
&= ({\mathbf{s}_n}^T - {\mathbf{s}_{n-\tau}}^T\mathbf{C}_\tau)\mathbf{h}_1 \\
&= \sum_{j=0}^{\tau-1}\left(s(n-j) - \sum_{k=0}^{L+p} s(n-\tau-j)c_{\tau-j,k+1}^{(\tau)}\right)h_0^{(1)} \\
&\approx \sum_{j=0}^{\tau-\tau_o-1} s(n-j)h_j^{(1)} \\
&\qquad + \sum_{j=\tau-\tau_o}^{\tau-1}\left(s(n-j) - \sum_{k=0}^{L+p} s(n-\tau-k)c_{\tau-j,k+1}^{(\tau)}\right)h_j^{(1)}.
\end{aligned} \tag{9.41}$$

Here, the abovementioned assumption, $c_{j,k}^{(\tau)} \approx 0$ $(j+k > \tau_o + 1)$, is reflected in the simplification of the last two expressions. First, we notice that the late reflections caused by the terms of ATF $h_1(z)$ higher than τ^{th} are cancelled out. As regards the first term in the final expression, it consists of the direct sound and a few following early reflections that are not over-whitened. The second term consists of early reflections that suffer from the over-whitening effect caused by the target-speech autocorrelation within a time lag of τ. Note that τ is set at around 30 ms in our experimental algorithm.

As for this over-whitening problem, we have experimentally confirmed that pre-whitening of the observed speech signals should be performed before the abovementioned τ-*step* LP to reduce the effects of AR coefficients $c_{j,k}^{(\tau)}$ $(j+k \leq \tau_o + 1)$. In our experimental algorithm, the pre-whitening was performed by using conventional LP with a predictor order of around 20. Moreover, in ASR applications, Cepstral Mean Subtraction (CMS) [4] may be utilized to reduce such early reflections in the first term as well as the remaining over-whitening effects in the second term.

In our actual implementation of the proposed algorithm, we employed spectral subtraction [7] to subtract the estimated late reflections ${\mathbf{s}_{n-\tau}}\mathbf{H}\mathbf{w}_\tau$ from the observed speech signal, ${\mathbf{s}_n}^T\mathbf{h}_1$. We assumed that spectral subtraction would be insensitive to the phase differences between the late reflections included in the actual observed speech and estimation ones. Thus, at the expense of some artifacts in the resultant speech, this procedure may work robustly compared with the simple time-domain subtraction shown in (9.41). The implementation of the proposed algorithm is summarized as follows:

Summary of the proposed algorithm

1. First, pre-whitening is applied to the observed speech signals, $\mathbf{s}_n{}^T\mathbf{H}$.
2. Next, τ-*step* predictors $\mathbf{w}_\tau$ are calculated by using those pre-whitened signals instead of the observed signals shown in (9.37).
3. Then, the late reflections, $\mathbf{s}_{n-\tau}\mathbf{H}\mathbf{w}_\tau$, are estimated by applying $\mathbf{w}_\tau$ to the τ-sample delayed version of the observed speech, $\mathbf{s}_{n-\tau}{}^T\mathbf{H}$.
4. Next, the estimated late reflections and the observed speech signal at microphone 1, $\mathbf{s}_n{}^T\mathbf{h}_1$, are both divided into short-time frames with Hamming windows, and their power spectra are calculated with the short-time Fourier transform (STFT).
5. Then, the power spectrum of the estimated late reflections is subtracted from that of the observed speech.
6. Finally, the resulting spectrum is converted back to a time-domain signal with the inverse STFT and the overlap-add technique. To synthesize the signal, the phase of the observed signal at microphone 1 is applied.

Moreover, this algorithm may be repeated for all the microphone signals if the differences in the arrival times of the target speech at the microphones are much smaller than τ samples. Then, to benefit from the spatial diversity provided by the multiple microphones, the concept of the DSB [11, 38] may be incorporated. In this procedure, we should adjust the delays among the M τ-*step* prediction residuals and calculate the sum of the residuals as the resultant signal. The delays may be estimated, for example, based on the cross-correlation of the residuals.

Below, we discuss three issues related to the proposed algorithm, which can be summarized as follows:

1. Speech dereverberation performance in terms of ASR score
2. Speech dereverberation in a noisy environment
3. Dereverberation of multiple sound source signals

9.3.2.2 Speech Dereverberation Performance in Terms of ASR Score

We conducted a speech dereverberation experiment to test the proposed algorithm. Four microphones and a loudspeaker were positioned in a room with a volume of about 40 m^3 and a reverberation time of around 0.5 s. The microphones were equally spaced at 0.2 m. We recorded the reverberant speech for four different loudspeaker positions, with distances of 0.5, 1.0, 1.5 and 2.0 m between the microphones and the loudspeaker. The SNRs of the recordings were about 15 to 20 dB. The SNRs were improved to around 30 dB by high-pass filtering with a cutoff frequency of 200 Hz. One hundred utterances taken from the Japanese Newspaper Article Sentences (JNAS) corpus were used as the target speech signals. The sampling frequency was 12 kHz.

ASR performance was evaluated in terms of the Word Error Rate (WER) averaged over genders and speakers. In the acoustic model, we used the following

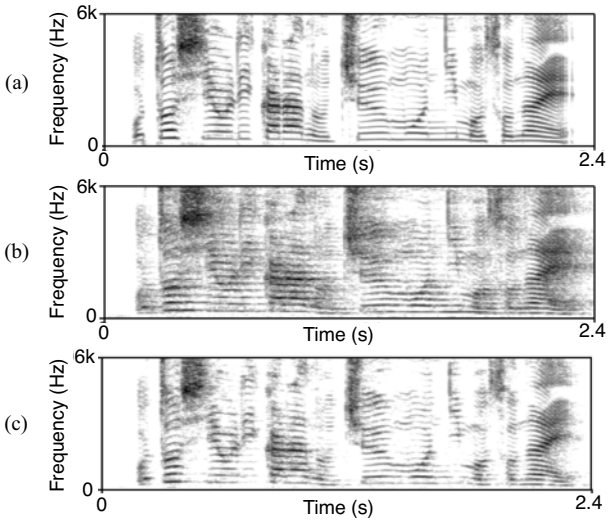


Fig. 9.7 Spectrograms in a real reverberant environment when the distance between the microphones and speaker was set to 1.5 m: (a) clean speech, (b) recorded reverberant speech, and (c) speech processed with the proposed algorithm

parameters: 12 order MFCCs + energy, their Δ and $\Delta\Delta$, 3 state HMMs, and 16 mixture Gaussian distributions [12]. The model was trained on clean speech processed with CMS. The language model was a standard trigram model trained on Japanese newspaper articles written over a ten-year period. The average duration of the test data was about 6 s.

As regards the algorithm parameters, the number of predictor taps and the delay τ were set at 750 and 30 ms (360 samples), respectively. To pre-whiten the observed speech signals, we used a 20th-order predictor calculated similarly to the approach described in [13]. No special parameters such as over-subtraction parameters or smoothing parameters were used for the spectral subtraction shown in procedure step 5 above. The length of the Hamming window for the short-time Fourier transform (STFT) was 360 samples, and the frame overlap factor was 1/8. The speech dereverberation was performed utterance by utterance, which means that the amount of observed speech data used to calculate predictors $\mathbf{w}_\tau$ is equivalent to the duration of each input utterance.

Figure 9.7 shows spectrograms of (a) target clean speech, (b) reverberant speech observed 1.5 m away from the loudspeaker, and (c) dereverberated speech with the proposed algorithm using four microphones. All speech signals were processed with the CMS. We can clearly see that the effects of reverberation on the target speech are greatly reduced with the proposed algorithm.

Table 9.2 shows the relation between the WER and the distance from the loudspeaker to the microphones. In this table, "No proc" corresponds to the observed speech processed with the CMS, and "Proposed" to the speech dereverberated with

Table 9.2 Speech dereverberation evaluated with the WER

Dist.	0.5 m	1.0 m	1.5 m	2.0 m
No proc.	21%	45%	57%	66%
Proposed	8%	11%	13%	14%

Dist.: distance between loudspeaker and microphones
No proc.: observed speech
Proposed: dereverberated speech with proposed algorithm
Baseline performance of ASR system was 5%

the proposed algorithm using four microphones. In this experiment, the baseline performance was 5%, which is the WER obtained with recordings made under a non-reverberant condition. The proposed algorithm achieved an excellent and stable dereverberation performance for all reverberant conditions.

These results show that the proposed algorithm works well even in a severely reverberant environment.

9.3.2.3 Speech Dereverberation in a Noisy Environment

In the previous experiment, the proposed algorithm performed very well with respect to the dereverberation of recorded speech with a relatively high SNR of 30 dB. Here, as pre-processing of the dereverberation algorithm, we examine spectral subtraction for its potential to reduce incoherent noise. This processing consists of similar procedures to those in steps 4–6 of the above algorithm. Each microphone signal, which consists of an observed speech signal $\mathbf{s}_n{}^T\mathbf{H}_m$ and incoherent noise $v_m(n)$, is first converted into its power spectrum. Next, the power spectrum of the noise is estimated in speech-absent segments, and this power spectrum is subtracted from that of the microphone signal. Then, the resulting power spectrum is converted back into a time-domain signal. Each time-domain signal is used for the input of the proposed dereverberation algorithm.

To test the combination of the spectral subtraction and the proposed dereverberation algorithm (hereafter called the comb-algorithm), we conducted a simulation using model impulse responses calculated with the image method [3]. The model room was designed to have the same dimensions as the room used in the previous experiment. The reverberation time was set to at around 0.65 s. One hundred utterances taken from the JNAS corpus were used as the target speech, and convolved with the model impulse responses to simulate reverberant observed speech signals. To simulate a noisy environment, pink noise was artificially generated and added to the reverberant speech with an SNR of 10 or 20 dB. The SNR is defined as the ratio of the reverberant speech and additive noise. All other conditions were the same as those of the previous experiment.

Figure 9.8 shows the spectrograms of (a) clean speech, (b) noisy and reverberant speech observed at a distance of 1.5 m from the loudspeaker with an SNR of 20 dB,

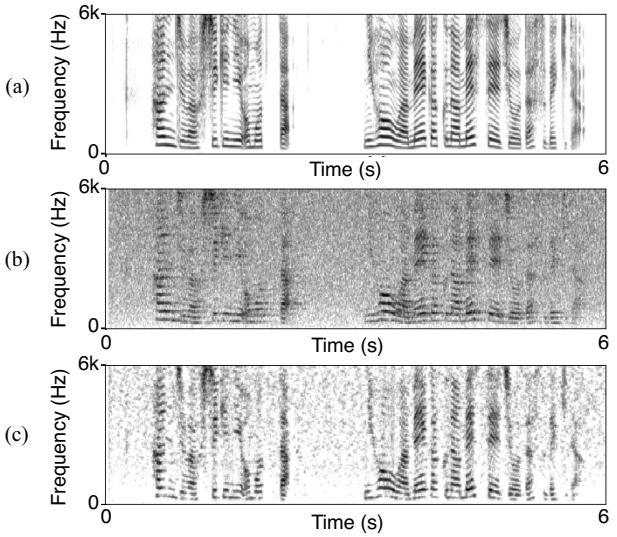


Fig. 9.8 Spectrograms: (a) clean speech, (b) noisy reverberant speech, and (c) processed speech

Table 9.3 Dereverberation and denoising evaluated with the WER

Dist.	0.5 m	1.0 m	1.5 m	2.0 m
Denoise (10 dB)	20%	42%	55%	65%
Denoise (20 dB)	12%	42%	55%	63%
Comb. (10 dB)	11%	12%	12%	15%
Comb. (20 dB)	5%	6%	8%	10%

Dist: distance between loudspeaker and microphones
Denoise: reverberant speech processed only with spectral subtraction and CMS
Comb.: dereverberated speech with comb-algorithm
Baseline performance of the ASR system was 5% for SNR = 20 dB, or 12% for SNR = 10 dB

and (c) speech processed with the proposed comb-algorithm. We can clearly see the dereverberation and denoising effects of the comb-algorithm.

Table 9.3 shows the relation between the WER and the distance from the loudspeaker to the microphones. "Denoise" represents the WER of reverberant speech processed only with the spectral subtraction and CMS. "Comb" corresponds to the WER achieved by using the comb-algorithm with four microphones. The SNR is given in parentheses. The baseline WER was 5% for non-reverberant speech with an SNR of 20 dB, and 12% for an SNR of 10 dB. These speech signals were processed only with the spectral subtraction and CMS. As shown in the table, the WER achieved with the comb-algorithm was excellent and close to the baseline performance.

These results demonstrate that the comb-algorithm works well even in a noisy reverberant environment.

9.3.2.4 Dereverberation of Multiple Sound Source Signals

Suppose here that there is an additional sound source in the acoustic situation considered in (9.37) to (9.41). Then, we may obtain a similar result to (9.41), where the late reflections associated with the additional source signal observed at microphone m ($m = 1, 2, \ldots, M$; $M \geq 3$) as well as those associated with the target speech are removed. This relation may be expressed as follows:

$$\tilde{y}_{\tau}^{(m)}(n) = \sum_{j=0}^{\tau-1} \left(s(n-j) - \sum_{k=0}^{L+p} s(n-\tau-k) c_{\tau-j,k+1}^{(\tau)} \right) h_j^{(m)} + \sum_{j=0}^{\tau-1} \left(\tilde{s}(n-j) - \sum_{k=0}^{L+p} \tilde{s}(n-\tau-k) \tilde{c}_{\tau-j,k+1}^{(\tau)} \right) \tilde{h}_j^{(m)}, \tag{9.42}$$

where $\tilde{s}(n)$, $\tilde{h}(n)$ and $\tilde{c}_{j,k}^{(\tau)}$ stand for an additional source signal, the ATF between the additional source and microphone m, and the coefficients of *j-step* AR polynomial defined similarly to (9.39) and (9.40).

Looking at this relation, we may notice that the arrival time difference between the target speech and additional source sound at each microphone is preserved. This means that we can utilize microphone-array technology such as null beamforming [11, 38] or blind source separation [6] in a reverberant enclosure, where such technology would not otherwise work well to emphasize, for example, the target speech by suppressing the additional source sound that remains after the proposed dereverberation algorithm (see the second term of (9.42)).

To test the abovementioned idea, we conducted a simulation using model impulse responses calculated with the image method. The simulated room had the same dimensions as the previous simulation, and its reverberation time was set to 0.6 s. The target speaker and pink-noise source were placed 1.5 m away and at a $\pm$ 45° angle to the microphone array, which consisted of four equally spaced microphones. The SNR was set at 5 dB at the center microphone. The sampling frequency, the number of predictor taps and τ were set at 8 kHz, 4,000 and 50, respectively.

Figure 9.9 shows the impulse responses between each sound source and a microphone before and after the dereverberation achieved with the proposed algorithm. We can see that the late reflections were greatly reduced for both sound sources. The energy ratio of the late reflections to the whole impulse response was reduced by 65% for the target speaker and by 62% for the pink-noise source.

These results demonstrate that the proposed algorithm may effectively reduce the late reflection energy even when there are multiple sound sources.

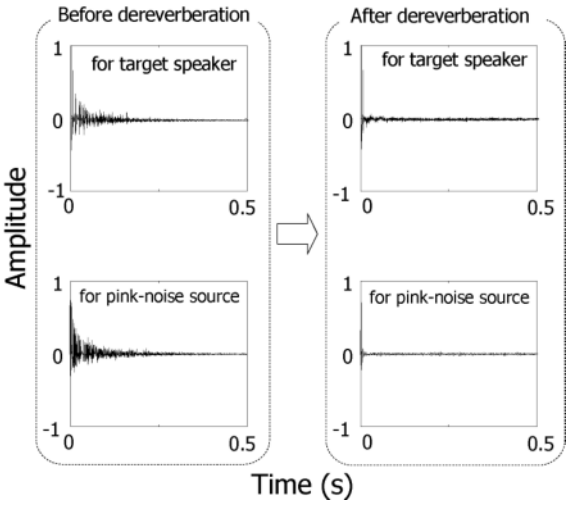


Fig. 9.9 The impulse responses between each sound source and a microphone before and after the dereverberation achieved with the proposed algorithm

9.3.3 Joint Estimation of Linear Predictors and Short-time Speech Characteristics

Another approach to addressing the over-whitening problem is to jointly estimate the linear predictors and the short-time characteristics of speech signals. We next review this approach.

9.3.3.1 Background

Let us assume that companion matrix $\mathbf{C}$ in (9.19) is a one-tap shift matrix. Then, accurate linear predictors $\mathbf{w}_{\mathrm{a}}$ do not cause over-whitening. The companion matrix $\mathbf{C}$ is identical to a one-tap shift matrix if and only if speech signal $s(n)$ follows a stationary white process. Therefore, if the signals observed at the microphones, $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$, are pre-processed so that $s(n)$ can be considered a stationary white process, the optimal predictors $\mathbf{w}_{\mathrm{o}}$ will closely approximate the optimal inverse filters $\mathbf{g}_{\mathrm{o}}$.

Then, what kind of pre-processing is suitable for this purpose? To investigate this question, it is useful to focus on the difference between a speech signal and a stationary white process. A speech signal differs from a stationary white process in the following two respects:

1. A speech signal is colored because it is physically produced by a time-variant articulatory filter.

2. The variance of the (assumed) white signal that is the input to the articulatory filter, which is hereafter called an innovation variance, is non-stationary.

Therefore, if the observed signals in each short-time frame are processed by the inverse articulatory filter of the corresponding frame and then scaled by the reciprocal of the square root of the innovation variance of the corresponding frame, we can consider $s(n)$ to follow the stationary white process. In the following, we refer to the inverse of the articulatory filter as an inverse articulatory filter.

The most fundamental problem when pre-processing the observed signals according to the above idea lies in the fact that the true inverse articulatory filter and the innovation variance of each frame are unknown in advance. Hence, they also need to be estimated using the observed signals. The inverse articulatory filter of each frame may be estimated as the prediction error filter that is obtained from a linear prediction analysis of the corresponding short-time observed signals. The innovation variance of that frame may then be estimated as the variance of the short-time observed signals filtered with the estimate of the inverse articulatory filter. This is the most convenient method for estimating the inverse articulatory filter and the innovation variance. Note that a small value is chosen for the order of the linear prediction analysis so that the resultant prediction error filter can approximate the true inverse articulatory filter. This kind of pre-processing was first proposed by Gillespie *et al.* [16]. However, the method described in [16] is not based on multichannel LP. Gaubitch *et al.* [13] also confirmed the effectiveness of this kind of pre-processing experimentally.

Although the above method for estimating the inverse articulatory filter and the innovation variance is easy to implement, the dereverberation performance achieved with this method is sometimes unsatisfactory. Furthermore, its mathematical background is unclear. Intuitively, the dereverberation performance will be improved by repeating the estimation of the predictors and the estimation of the inverse articulatory filter and the innovation variance alternately. We can expect this cyclic estimation to lead to a more accurate estimate of the inverse articulatory filter and the innovation variance and hence to an improvement in the dereverberation performance.

Yoshioka *et al.* [40] showed that such heuristically-derived pre-processing methods can be justified from the information theoretic viewpoint. The important point is that now we estimate the predictors, the inverse articulatory filter, and the innovation variance jointly. In the remainder of this subsection, we describe the principle and algorithms based on the concept of joint estimation.

9.3.3.2 Principle

First, let us introduce the model of a speech signal based on the above two characteristics of speech. We assume that:

1. $s(n)$ follows a time-variant AR process of order q. Hence, we have

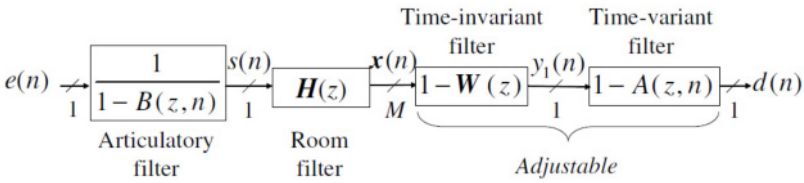


Fig. 9.10 System diagram

$$s(n) = \sum_{k=1}^{q} b(k,n)s(n-k) + e(n), \tag{9.43}$$

where $b(k,n)$ denotes the k^{th} regression coefficients at time index n and $e(n)$ denotes the innovations process. (9.43) is equivalent to filtering the innovations process $e(n)$ with an articulatory filter of the form $1/(1-B(z,n))$ with

$$B(z,n) = \sum_{k=1}^{q} b(k,n)z^{-k}. \tag{9.44}$$

2. The innovations $\{e(n)\}_{n\in\mathbb{Z}}$ consist of zero-mean uncorrelated random variables. The variances of $\{e(n)\}_{n\in\mathbb{Z}}$ are not necessarily identical.
3. The articulatory filter $1/(1-B(z,n))$ has no time-invariant poles. Thus, we have

$$\mathrm{GCD}\{\ldots, 1-B(z,0), 1-B(z,1), \ldots\} = 1, \tag{9.45}$$

where $\mathrm{GCD}\{P_1(z),\ldots,P_n(z)\}$ represents the greatest common divisor of polynomials $P_1(z),\ldots,P_n(z)$.

Now, let us consider filtering the residual of the multichannel LP, $y_1(n)$, with time-variant filter $1-A(z,n)$ as shown in Fig. 9.10. The output of the time-variant filter, $d(n)$, is given as follows:

$$d(n) = y_1(n) - \sum_{k=1}^{q} a(k,n)y_1(n-k). \tag{9.46}$$

Similarly to (9.43), (9.46) is equivalent to filtering $y_1(n)$ with

$$A(z,n) = \sum_{k=1}^{q} a(k,n)z^{-k}. \tag{9.47}$$

The adjustable parameters are $\{w_m(k)\}_{1\le m\le M,\ 1\le k\le p}$ and $\{a(k,n)\}_{1\le k\le q,\ 1\le n\le N}$, where N is the number of samples.

We have the following theorem under the above assumptions [40].

Theorem 9.1. *Assume that the output $d(n)$ is equal to the innovations process $e(n)$, and that $1-A(z,n)$ has no time-invariant zero, i.e.,*

$$d(n) = e(n), \tag{9.48}$$

$$\mathrm{GCD}\{1 - A(z,1), \ldots, 1 - A(z,N)\} = 1, \tag{9.49}$$

then, the residual of the multichannel LP, $y_1(n)$, is equal to $s(n)$.

Accordingly, we simply have to set up the tap weights[2] $\{w_m(k)\}$ and $\{a(k,n)\}$ jointly so that $d(n)$ is made equal to $e(n)$.

Since the innovations process $e(n)$ is inaccessible in reality, we have to develop criteria defined solely by using $d(n)$. To develop such criteria, we focus on the fact that output $d(n)$ is an estimate of innovations process $e(n)$. Therefore, it would be natural to set up $\{w_m(k)\}$ and $\{a(k,n)\}$ so that the outputs $\{d(n)\}_{1 \le n \le N}$ are uncorrelated.

Let $\mathcal{K}(\xi_1, \ldots, \xi_n)$ denote a suitable measure of correlation between random variables $\xi_1, \ldots, \xi_n$. Then, the task is mathematically formulated as

$$\underset{\{a(k,n)\},\ \{w_m(k)\}}{\text{minimize}} \quad \mathcal{K}(d(1), \ldots, d(N)). \tag{9.50}$$

A reasonable definition of $\mathcal{K}(\cdot)$ is

$$\mathcal{K}(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \log \upsilon(\xi_i) - \log |\det \Sigma(\xi)|, \tag{9.51}$$

$$\xi = [\xi_1, \ldots, \xi_n]^T, \tag{9.52}$$

where $\upsilon(\xi_1), \ldots, \upsilon(\xi_n)$, respectively, represent the variances of random variables $\xi_1, \ldots, \xi_n$, and $\Sigma(\xi)$ denotes the covariance matrix of ξ.

Then, we try to minimize

$$\mathcal{K}(d(1), \ldots, d(N)) = \sum_{n=1}^{N} \log \upsilon(d(n)) - \log |\det \Sigma(\mathbf{d})|, \tag{9.53}$$

$$\mathbf{d} = [d(1), \ldots, d(N)]^T, \tag{9.54}$$

with respect to $\{a(k,n)\}$ and $\{w_m(k)\}$, where N is the number of observed samples. This loss function can be further simplified as

$$\mathcal{K}(d(1), \ldots, d(N)) = \sum_{n=1}^{N} \log \upsilon(d(n)) + \text{constant}. \tag{9.55}$$

Hence, (9.50) is finally reduced to

$$\underset{\{a(k,n)\},\ \{w_m(k)\}}{\text{minimize}} \quad \sum_{n=1}^{N} \log \upsilon(d(n)). \tag{9.56}$$

[2] Hereafter, we omit the range of indices unless required.

Therefore, we have to set up the tap weights $\{a(k,n)\}$ and $\{w_m(k)\}$ to minimize the logarithmic mean of the variances of outputs $\{d(n)\}$. It is noteworthy that the solution for (9.56) is proven to give $d(n) = e(n)$ [40].

9.3.3.3 Algorithms

Here we derive algorithms for accomplishing (9.56). Before proceeding, we introduce an approximation of time-variant filter $1 - A(z,n)$. Since a speech signal within a short-time frame of several tens of milliseconds is almost stationary, we approximate $1 - A(z,n)$ by using a filter that is globally time-variant but locally time-invariant as

$$1 - A(z,n) = 1 - A_t(z),\ t = \left\lfloor \frac{n-1}{W} + 1 \right\rfloor, \tag{9.57}$$

where W is the frame size and $\lfloor \cdot \rfloor$ represents the floor function. Furthermore, if sample index n is in the range of the t^{th} frame, we estimate the variance of $d(n)$ by $\frac{1}{W}\sum_{n=1}^{W} d(W(t-1)+n)^2$. By using these approximations, task (9.56) is reformulated as

$$\underset{\Theta_a,\,\Theta_w}{\text{minimize}} \sum_{t=1}^{T} \log\Big(\sum_{n=1}^{W} d(W(t-1)+n)^2\Big), \tag{9.58}$$

where $\Theta_a = \{a_t(k)\}_{1\le t\le T, 1\le k\le q}$ and $\Theta_w = \{w_m(k)\}_{1\le m\le M, 1\le k\le p}$.

Task (9.58) may be solved by using the gradient descent method as described in [40]. Although the gradient descent method provides an accurate solution for (9.58), its convergence to an optimum value is very slow. Furthermore, the gradient descent method requires the step-size to be adjusted well. Therefore, we describe here an alternative method that uses no additional parameters such as step-size. This method is novel and published here for the first time.

Let us represent the cost function of (9.58) by

$$\mathcal{L}(\Theta_a, \Theta_w) = \sum_{t=1}^{T} \log\Big(\sum_{n=1}^{W} d(W(t-1)+n)^2\Big). \tag{9.59}$$

The difficulty in minimizing $\mathcal{L}$ arises from the term $\log(\sum_{n=1}^{W} d(W(t-1)+n)^2)$. To avoid this difficulty, we capitalize on the following inequality:

$$\log x \le \frac{x}{\lambda} - 1 + \log\lambda, \tag{9.60}$$

where the equality holds if and only if $x = \lambda$. By using (9.60), we have

$$\mathcal{L}(\Theta_a, \Theta_w) \le \sum_{t=1}^{T} \Big(\frac{1}{\lambda_t}\sum_{n=1}^{W} d(W(t-1)+n)^2 - 1 + \log\lambda_t\Big), \tag{9.61}$$

Algorithm 9.1 Summary of the algorithm

1. Initialize Θ_a and Λ by $\Theta_a^{(0)}$ and $\Lambda^{(0)}$, respectively. Iteration index i is set at 0.
2. Update the estimate of Θ_w to $\Theta_w^{(i+1)}$ to minimize $\mathcal{L}^+(\Theta_a^{(i)}, \Theta_w^{(i+1)}, \Lambda^{(i)})$. Since we have

$$\mathcal{L}(\Theta_a^{(i)}, \Theta_w^{(i+1)}) \leq \mathcal{L}^+(\Theta_a^{(i)}, \Theta_w^{(i+1)}, \Lambda^{(i)}) \leq \mathcal{L}^+(\Theta_a^{(i)}, \Theta_w^{(i)}, \Lambda^{(i)}) = \mathcal{L}(\Theta_a^{(i)}, \Theta_w^{(i)}), \tag{9.62}$$

   the value of cost function $\mathcal{L}$ is also reduced.
3. Update the estimate of Θ_a to $\Theta_a^{(i+1)}$ to minimize $\mathcal{L}(\Theta_a^{(i+1)}, \Theta_w^{(i+1)})$. Consequently, we obtain

$$\mathcal{L}(\Theta_a^{(i+1)}, \Theta_w^{(i+1)}) \leq \mathcal{L}(\Theta_a^{(i)}, \Theta_w^{(i+1)}). \tag{9.63}$$

4. Update the value of Λ to $\Lambda^{(i+1)}$ to minimize $\mathcal{L}^+(\Theta_a^{(i+1)}, \Theta_w^{(i+1)}, \Lambda^{(i+1)})$. Then we have

$$\mathcal{L}^+(\Theta_a^{(i+1)}, \Theta_w^{(i+1)}, \Lambda^{(i+1)}) = \mathcal{L}(\Theta_a^{(i+1)}, \Theta_w^{(i+1)}). \tag{9.64}$$

5. Increment i and return to Step 2, unless convergence is reached.

where we denote the right-hand side of (9.61) by $\mathcal{L}^+(\Theta_a, \Theta_w, \Lambda)$ with $\Lambda = \{\lambda_t\}_{1 \leq t \leq T}$.

By using the function $\mathcal{L}^+$ introduced above, we obtain an algorithm for achieving (9.58) based on the coordinate descent method as shown in Algorithm 9.1. We easily find that this algorithm monotonically reduces the value of cost function $\mathcal{L}$. How the algorithm in Algorithm 9.1 reduces the cost is illustrated in Fig. 9.11. This kind of algorithm is essentially identical to the expectation maximization (EM) algorithm and was first introduced in [23].

The remaining issues are the specific realization of Steps 2, 3, and 4. As regards Step 3, we find that, for each t, $\{a_t^{(i+1)}(k)\}_{1 \leq k \leq P}$ is calculated by applying linear predictive analysis to $\{y_1(n)\}_{(t-1)W < n \leq (tW)}$, where $y_1(n)$ is the tentative estimate of the speech signal $s(n)$. It is also found that Step 4 is accomplished by

$$\lambda_t^{(i+1)} = \sum_{n=1}^{W} d((t-1)W + n)^2. \tag{9.65}$$

Note that $\Theta_a^{(i+1)}$ and $\Lambda^{(i+1)}$ are the tentative estimates of the (inverse) articulatory filter and the innovation variances, respectively.

Finally, the update rule for Θ_w in Step 2 is derived as follows. By setting the differential of $\mathcal{L}^+$ with respect to $\mathbf{w}$ at zero, we obtain the following linear equation relating to $\mathbf{w}^{(i+1)}$:

$$\left(\sum_{t=1}^{T} \sum_{n=(t-1)W+1}^{tW} \mathbf{u}_t(n-1)\mathbf{u}_t(n-1)^T \right) \mathbf{w}^{(i+1)} = \sum_{t=1}^{T} \sum_{n=(t-1)W+1}^{tW} u_{t,1}(n)\mathbf{u}_t(n-1), \tag{9.66}$$

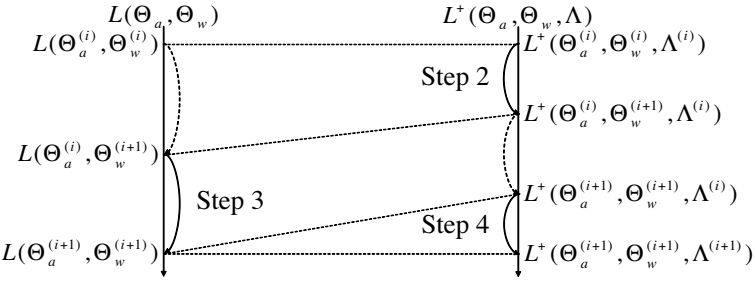


Fig. 9.11 Schematic diagram of the algorithm

where

$$\mathbf{u}_t(n) = \frac{1}{\sqrt{\lambda_t^{(i)}}} \left(\mathbf{x}(n) - \sum_{k=1}^{q} a_t(k) \mathbf{x}(n-k) \right) \tag{9.67}$$

and

$$u_{t,1}(n) = \frac{1}{\sqrt{\lambda_t^{(i)}}} \left(x_1(n) - \sum_{k=1}^{q} a_t(k) x_1(n-k) \right) \tag{9.68}$$

are the observed signals pre-processed by the tentative estimates of the inverse articulatory filter and the innovation variances.

In conclusion, the dereverberation method that iterates the estimation of the inverse filter set and the estimation of the inverse articulatory filter and the innovation variances is completely justified from the perspective of information theory.

9.3.4 Probabilistic Model Based Speech Dereverberation

This section describes a different approach to overcoming the over-whitening problem. The probabilistic model based approach is introduced into multichannel LP based speech dereverberation. When the signal of interest is speech that manifests certain inherent characteristics, we can consider the dereverberation as a speech enhancement problem (as opposed to a problem of inverting the room impulse response), where reverberant components in the microphone signal are to be suppressed according to the source characteristics [31, 39]. The use of source characteristics has led to a new formulation of multichannel LP based on a probabilistic speech model, in which the objective is not to estimate a filter that whitens the observed signal as in (9.15), but to design one that would turn reverberant speech into a signal that is probabilistically more like clean speech [30]. The over-whitening problem does not arise with this mechanism provided that the probabilistic speech

model appropriately represents the spectral features of speech. The resultant optimization problem can be solved by employing maximum likelihood estimation.

9.3.4.1 Probabilistic Speech Model

With this approach, (9.14) is used as a model for representing the relationship between the observed signal $x_m(n)$ and the residual of the multichannel LP $\hat{y}(n)$, where the residual is taken as the signal to be obtained, namely a speech signal $s(n)$. For simplicity, we consider a noise-free case in this section, where $v_m(n) = 0$, and we set $\hat{y}(n) = s(n)$. (It is easy to show the noise robust property of this approach in a similar way to that of the multichannel LP shown by (9.15).) In addition, because short-time segments of the signals are dealt with as objects to be processed with this approach, we rewrite (9.14) so that it includes the short-time segments explicitly as

$$\begin{aligned}\bar{\mathbf{x}}_1(n) &= \sum_{m=1}^{M} \bar{\mathbf{X}}_m(n-p:n-1)\mathbf{w}_m + \bar{\mathbf{s}}(n), \\ &= \bar{\mathbf{X}}(n-p:n-1)\mathbf{w} + \bar{\mathbf{s}}(n),\end{aligned} \tag{9.69}$$

where p is the length of the prediction filter $\mathbf{w}_m$ in each channel, $\bar{\mathbf{s}}(n)$ and $\bar{\mathbf{x}}_m(n)$ are vectors of short-time segments of length K for $s(n)$ and $x_m(n)$ defined as[3]

$$\begin{aligned}\bar{\mathbf{s}}(n) &= [s(n), s(n-1), \ldots, s(n-K+1)]^T, \\ \bar{\mathbf{x}}_m(n) &= [x_m(n), x_m(n-1), \ldots, x_m(n-K+1)]^T,\end{aligned}$$

and $\bar{\mathbf{X}}(n_1 : n_2)$ is a matrix that contains a time series of $\bar{\mathbf{x}}_m(n)$ for all microphones m from time n_1 to n_2, defined as

$$\begin{aligned}\bar{\mathbf{X}}_m(n_1 : n_2) &= [\bar{\mathbf{x}}_m(n_2), \bar{\mathbf{x}}_m(n_2-1), \ldots, \bar{\mathbf{x}}_m(n_1)], \\ \bar{\mathbf{X}}(n_1 : n_2) &= [\bar{\mathbf{X}}_1(n_1 : n_2), \bar{\mathbf{X}}_2(n_1 : n_2), \ldots, \bar{\mathbf{X}}_M(n_1 : n_2)].\end{aligned}$$

A probabilistic speech model is introduced as a criterion for evaluating the degree to which the residual is likely to be clean speech in a probabilistic sense, and used to determine the prediction filter coefficients. A Time-Varying Gaussian Source Model (TVGSM) has been shown to be suitable as a general probabilistic speech model for multichannel LP. With this model, the following assumptions are introduced.

1. Each short-time speech segment of the order of tens of milliseconds is a realization of a stationary multivariate Gaussian random process with a zero mean and a covariance matrix of $\mathbf{R}_s(n) = E\{\bar{\mathbf{s}}(n)\bar{\mathbf{s}}(n)^T\}$, where $E\{\cdot\}$ represents the expectation function. The Probability Density Function (PDF) of the segment, $f_s(\bar{\mathbf{s}}(n))$, is defined as

[3] In this section, symbols with bars such as $\bar{\mathbf{x}}(n)$ and $\bar{\mathbf{X}}$, respectively, denote a vector of a short-time segment and a matrix of a time series of the vector.

$$f_s(\bar{\mathbf{s}}(n)) = \mathcal{N}(\bar{\mathbf{s}}(n); 0, \mathbf{R}_s(n)), \tag{9.70}$$

where $\mathcal{N}(\bar{\mathbf{s}}; \mu, \Sigma)$ denotes a PDF of a multivariate Gaussian random process with a mean vector μ and a covariance matrix Σ, which is defined as

$$\mathcal{N}(\bar{\mathbf{s}}; \mu, \Sigma) = (2\pi)^{-K/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}|\bar{\mathbf{s}} - \mu|^2_{\Sigma^{-1}}\right),$$
$$|\mathbf{s}|^2_{\Sigma^{-1}} = \mathbf{s}^T\Sigma^{-1}\mathbf{s}.$$

Note that the covariance matrix $\mathbf{R}_s(n)$ can be employed as an autocorrelation matrix[4] in the above definition because $\bar{\mathbf{s}}(n)$ is assumed to be stationary.

2. The autocorrelation matrices in the above PDF may vary over different short-time segments.

With TVGSM, the characteristics of a short-time speech segment are represented by the autocorrelation matrix of the segment, or equivalently by the autocorrelation function. Because the autocorrelation function of a segment contains information that is equivalent to its power spectrum, it can represent the characteristics of the power, the spectral envelope, and other finer spectral structures such as the harmonicity of the segment. The properties of an autocorrelation matrix can be more clearly determined when we further assume the short-time segment to be an autoregressive process excited by white Gaussian noise. Then, the autocorrelation matrix of the segment can be parameterized using the AR process parameters [21] as

$$\mathbf{R}_s(n) = \sigma^2(n)(\mathbf{A}(n)^T\mathbf{A}(n))^{-1}, \tag{9.71}$$

where $\mathbf{A}(n)$ is an upper triangular Toeplitz matrix whose first row contains the AR coefficients and $\sigma^2(n)$ is the average energy of the AR residual [21]. With this model, $\sigma^2(n)$ and $\mathbf{A}(n)$, respectively, correspond to the power and spectral shape information of the segment. The order of the AR process determines how finely the model represents the spectral features. By contrast, the time-varying nature of speech can be represented by appropriately changing the values of the autocorrelation functions over different short-time segments with TVGSM.

Note that TVGSM has been adopted for many useful speech enhancement techniques, such as Wiener filtering and source and channel estimation [41]. The autoregressive hidden Markov model [21] is also a class of TVGSM.

9.3.4.2 Likelihood Function for Multichannel LP

A likelihood function is needed as a criterion for determining the prediction filter with the probabilistic model based formulation. We define it as

$$\mathcal{L}(\Theta) = \log f_x(\bar{\mathbf{X}}(1:N); \Theta), \tag{9.72}$$

[4] An autocorrelation matrix is defined as a symmetric Toeplitz matrix that contains an autocorrelation function of a segment in its first column.

where $f_x(\bar{\mathbf{X}})$ is the Probability Density Function (PDF) of the $\bar{\mathbf{X}}$ being observed and Θ is a set of parameters to be estimated. Θ should contain the prediction filter coefficients $\mathbf{w}$ and a time series of the autocorrelation matrices of the short-time speech segments $\mathbf{R}_s(n)$. Dereverberation is defined as a problem of finding the parameter set that maximizes the likelihood function as

$$\hat{\Theta} = \arg\max_{\Theta} \mathcal{L}(\Theta). \tag{9.73}$$

It is easy to rewrite the likelihood function by expanding the PDF along with the time sequence of $\bar{\mathbf{X}}(1:N)$ as

$$\begin{aligned}\mathcal{L}(\Theta) &= \log f_x(\bar{\mathbf{X}}(1:1);\Theta) \\ &+ \sum_{n=1}^{N} \log f_{x|X}(\bar{\mathbf{x}}_2(n),\bar{\mathbf{x}}_3(n),\ldots,\bar{\mathbf{x}}_M(n)|\bar{\mathbf{x}}_1(n),\bar{\mathbf{X}}(1:n-1);\Theta) \\ &+ \sum_{n=1}^{N} \log f_{x|X}(\bar{\mathbf{x}}_1(n)|\bar{\mathbf{X}}(1:n-1);\Theta).\end{aligned}$$

In the above equation, the first two terms on the right-hand side are not terms of interest when determining the prediction filter coefficients that predict the first channel as in (9.69), and thus we disregard them. By contrast, according to (9.69), $\bar{\mathbf{x}}_1(n)$ only depends on $\bar{\mathbf{X}}(n-p:n-1)$ and $\bar{\mathbf{s}}(n)$. Therefore, the third term can be rewritten as

$$\mathcal{L}(\Theta) = \sum_{n=1}^{N} \log f_{x|X}(\bar{\mathbf{x}}_1(n)|\bar{\mathbf{X}}(n-p:n-1);\Theta). \tag{9.74}$$

Finally, the likelihood function can further be rewritten based on (9.69) and (9.70) as

$$\begin{aligned}\mathcal{L}(\Theta) &= \sum_{n=1}^{N} \log f_s(\bar{\mathbf{s}}(n) = \bar{\mathbf{x}}_1(n) - \bar{\mathbf{X}}(n-p:n-1)\mathbf{w};\Theta), \\ &= \sum_{n=1}^{N} \log \mathcal{N}(\bar{\mathbf{x}}_1(n);\bar{\mathbf{X}}(n-p:n-1)\mathbf{w},\mathbf{R}_s(n)), \\ &= -\frac{1}{2}\sum_{n=1}^{N} |\bar{\mathbf{x}}_1(n) - \bar{\mathbf{X}}(n-p:n-1)\mathbf{w}|^2_{\mathbf{R}_s(n)^{-1}} \\ &- \frac{1}{2}\sum_{n=1}^{N} \log\det\mathbf{R}_s(n) + \text{const.}\end{aligned} \tag{9.75}$$

As a consequence, dereverberation is achieved by finding the parameter set $\Theta = \{\mathbf{w}, \bar{\mathbf{R}}_s\}$ that maximizes (9.75), where $\bar{\mathbf{R}}_s = [\mathbf{R}_s(0), \mathbf{R}_s(1), \ldots, \mathbf{R}_s(N)]$.

It is important to note that various dereverberation algorithms can be derived by introducing more specific source assumptions into TVGSM. For example, when we assume $\bar{\mathbf{s}}(n)$ to be stationary white Gaussian noise, we can set $\mathbf{R}_s(n) = \mathbf{I}$, and the

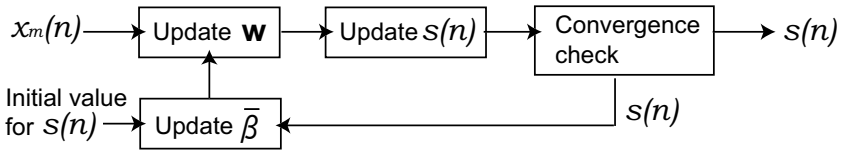


Fig. 9.12 Processing flow of autocorrelation codebook based speech dereverberation

above likelihood function becomes equivalent to the cost function of the multichannel LP shown by (9.15). By contrast, when we assume that $\bar{\mathbf{s}}(n)$ is a stationary AR process, we can set $\mathbf{R}_s(n) = (\mathbf{A}^T\mathbf{A})^{-1}$ based on (9.71). Then, (9.75) can further be rewritten by disregarding the constant terms as

$$\mathcal{L}(\Theta) = -\frac{1}{2}\sum_{n=1}^{N} |\mathbf{A}\bar{\mathbf{x}}_1(n) - \mathbf{A}\bar{\mathbf{X}}(n-p:n-1)\mathbf{w}|^2.$$

According to the above likelihood function, we can determine the prediction filter **w** using multichannel LP assuming the source signal to be a stationary white Gaussian process once the observed signal is pre-whitened by **A**. This corresponds to the pre-processing used for multichannel multistep LP based speech dereverberation discussed in Sect. 9.3.2. In addition, when we assume the source to be a time-varying AR process, we can derive a dereverberation algorithm similar to that described in Sect. 9.3.3 by setting $\mathbf{R}_s(n) = \sigma^2(n)(\mathbf{A}(n)^T\mathbf{A}(n))^{-1}$.

9.3.4.3 Autocorrelation Codebook-based Speech Dereverberation

As an example, we detail a dereverberation method based on the autocorrelation codebook [30], in which *a priori* knowledge of the speech signals is introduced as a PDF of the signal represented by a set of autocorrelation functions, referred to as the autocorrelation codebook. With this method, the following assumption is introduced:

> Each short-time speech segment $\bar{\mathbf{s}}(n)$ can be categorized as one of a finite number of states, β, where $1 \le \beta \le N_\beta$. In each state β, $\bar{\mathbf{s}}(n)$ is modeled by a multivariate Gaussian random process with an autocorrelation matrix $\tilde{\mathbf{R}}_s(\beta)$ determined by the state, that is, $f_s(\bar{\mathbf{s}}(n);\beta) = \mathcal{N}(\bar{\mathbf{s}}(n);0,\tilde{\mathbf{R}}_s(\beta))$.

According to this assumption, the PDF of the short-time speech segments can be represented by a finite number of autocorrelation matrices, referred to as codewords of the autocorrelation codebook. The autocorrelation codebook has to be prepared in advance based on this approach. It can be generated, for example, based on a certain clustering method using clean speech data.

With autocorrelation codebook based multichannel LP, we need to include the state sequence $\bar{\beta} = [\beta(1),\beta(2),\ldots,\beta(N)]^T$ over successive short-time segments in

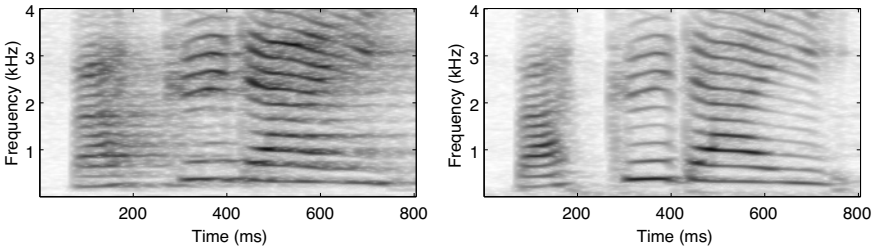


Fig. 9.13 Spectrograms of (a) a reverberated signal and (b) a signal dereverberated with the autocorrelation codebook using an observed signal that contains a five-word sequence uttered by a female talker ($T_{60} = 0.5$ s)

the parameter set Θ instead of the autocorrelation matrix sequence $\bar{\mathbf{R}}_s$. (See [21] for an example of parameter estimation with a finite state speech model.) The likelihood function is then rewritten by disregarding constant terms as

$$\mathcal{L}(\Theta) = -\frac{1}{2}\sum_{n=1}^{N} |\bar{\mathbf{x}}_1(n) - \bar{\mathbf{X}}(n-p:n-1)\mathbf{w}|^2_{\tilde{\mathbf{R}}_s(\beta(n))^{-1}} - \frac{1}{2}\sum_{n=1}^{N} \log\det\tilde{\mathbf{R}}_s(\beta(n)). \tag{9.76}$$

It is difficult to give the closed-form solution that maximizes (9.76). Instead, a repetitive estimation method can be derived for this maximization, where the likelihood function can be maximized up to a stationary point by iteratively updating the state sequence and the prediction filter coefficients in turn from certain initial values. They are summarized as follows:

1. Set the initial values as $\bar{\mathbf{s}}^{(0)}(n) = \bar{\mathbf{x}}_1(n)$ and the iteration counter as $i = 1$.
2. Update $\beta^{(i)}(n)$, $\mathbf{w}^{(i)}$ and $\bar{\mathbf{s}}^{(i)}(n)$ in turn to obtain a value that maximizes the likelihood function on the variables as

$$\begin{aligned}
\beta^{(i)}(n) &= \arg\max_{\beta} p(\bar{\mathbf{s}}^{(i-1)}(n);\beta) \ \text{ for all } n,\\
\mathbf{w}^{(i)} &= \Big(\sum_{n=1}^{N} \bar{\mathbf{X}}(n-p:n-1)^T\tilde{\mathbf{R}}_s(\beta^{(i)}(n))^{-1}\bar{\mathbf{X}}(n-p:n-1)\Big)^{-1}\\
&\quad\times\sum_{n=1}^{N} \bar{\mathbf{X}}(n-p:n-1)^T\tilde{\mathbf{R}}_s(\beta^{(i)}(n))^{-1}\bar{\mathbf{x}}_1(n),\\
\bar{\mathbf{s}}^{(i)}(n) &= \bar{\mathbf{x}}_1(n) - \bar{\mathbf{X}}(n-p:n-1)\mathbf{w}^{(i)} \ \text{ for all } n.
\end{aligned}$$

3. If the iteration converges, terminate the iteration and take $\bar{\mathbf{s}}^{(i)}(n)$ as the dereverberated signal. Otherwise, set $i = i+1$, and return to step 2.

It is guaranteed that the likelihood function monotonically increases according to this iteration. Figure 9.12 summarizes the processing flow of this optimization method. It has been shown by simulation experiments that autocorrelation codebook based speech dereverberation can effectively recover the quality of speech signals based on a few seconds of observance of reverberant signals [30]. Figure 9.13 shows example spectrograms obtained before and after dereverberation when the observed signals were captured by two microphones with a reverberation time (T_{60}) of 0.5 s.

9.4 Concluding Remarks

We have demonstrated that there are at least four solutions to the problem of overwhitening target speech, which is an essential drawback of inverse filtering based on multichannel LP. We are currently pursuing research to improve the noise-robustness of these solutions.

Appendix A

Here, we derive a sufficient condition on which optimal inverse filters $\mathbf{g}_{\mathrm{o}}$ are considered to be less sensitive to noise as well as fluctuation in ATFs than accurate filters $\mathbf{g}_{\mathrm{a}}$ by using the condition numbers related to these two filter-sets. The condition numbers are, respectively, given as [22]:

$$\text{for } \mathbf{g}_{\mathrm{o}}\text{: condNo}[\mathbf{g}_{\mathrm{o}}] = \frac{\kappa_{\max}[(\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}) + E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]}{\kappa_{\min}[(\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}) + E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]},$$

$$\text{for } \mathbf{g}_{\mathrm{a}}\text{: condNo}[\mathbf{g}_{\mathrm{a}}] = \frac{\kappa_{\max}[\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}]}{\kappa_{\min}[\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}]}.$$

As regards condition number condNo$[\mathbf{g}_{\mathrm{o}}]$, the following inequality holds.

$$\text{condNo}[\mathbf{g}_{\mathrm{o}}] \leq \frac{\kappa_{\max}[(\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I})] + \kappa_{\max}[E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]}{\kappa_{\min}[(\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I})] + \kappa_{\min}[E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]}.$$

Here, the following relation is considered [18]:

$$\kappa_{\min}[\mathbf{A}] + \kappa_{\min}[\mathbf{B}] \leq \kappa_j[\mathbf{A}+\mathbf{B}] \leq \kappa_{\max}[\mathbf{A}] + \kappa_{\max}[\mathbf{B}],$$

where $\mathbf{A}$ and $\mathbf{B}$ are $J \times J$ (J: natural number) Hermitian matrices, and

$$\kappa_{\max}[\mathbf{A}+\mathbf{B}] = \kappa_1[\mathbf{A}+\mathbf{B}] \geq \ldots \geq \kappa_j[\mathbf{A}+\mathbf{B}] \geq \ldots \geq \kappa_J[\mathbf{A}+\mathbf{B}] = \kappa_{\min}[\mathbf{A}+\mathbf{B}].$$

Hence, the sufficient condition may be given as follows:

$$\frac{\kappa_{\max}[(\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I})] + \kappa_{\max}[E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]}{\kappa_{\min}[(\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I})] + \kappa_{\min}[E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]}$$
$$\leq \frac{\kappa_{\max}[\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}]}{\kappa_{\min}[\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}]},$$
$$\Leftrightarrow \frac{\kappa_{\max}[E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]}{\kappa_{\min}[E\{\mathbf{v}_n \mathbf{v}_n{}^T\}]} \leq \frac{\kappa_{\max}[\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}]}{\kappa_{\min}[\mathbf{H}^T E\{\mathbf{s}_n \mathbf{s}_n{}^T\}\mathbf{H} + \delta^2\mathbf{I}]}. \tag{9.77}$$

This relation could be interpreted to mean that incoherent noise $\mathbf{v}_n$ should be *flatter* (and more spatially uncorrelated) than observed speech signals $\mathbf{s}_n{}^T\mathbf{H}$.

References

1. ATR International Speech database. Online (in Japanese). URL `http://www.red.atr.co.jp/database_page/digdb.html`
2. Aichner, R., Araki, S., Makino, S., Nishikawa, T., Saruwatari, H.: Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming. In: Proc. IEEE Int. Workshop on Neural Networks for Signal Processing, pp. 445–454 (2002)
3. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
4. Atal, B.S.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. **55(6)**, 1304–1312 (1974)
5. Ben-Israel, A., Greville, T.N.E.: Generalized inverses: theory and applications. Springer (1974)
6. Benesty, J., Makino, S., Chen, J.: Speech enhancement. Springer (2005)
7. Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Speech Audio Process. **27(2)**, 113–120 (1979)
8. Campbell, S.L., Jr., C.D.M.: Generalized inverses of linear transformations. Dover Publications (1979)
9. Delcroix, M., Hikichi, T., Miyoshi, M.: Dereverberation and denoising using multichannel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15**(6), 1791–1801 (2007)
10. Delcroix, M., Hikichi, T., Miyoshi, M.: Precise dereverberation using multi-channel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15**(2), 430–440 (2007)
11. Flanagan, J.L.: Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. **78(11)**, 1508–1518 (1985)
12. Furui, S.: Digital speech processing, synthesis, and recognition. Marcel Dekker (2001)
13. Gaubitch, N.D., Naylor, P.A., Ward, D.B.: On the use of linear prediction for dereverberation of speech. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), vol. 1, pp. 99–102 (2003)
14. Giannakis, G.B., Hua, Y., Stoica, P., Tong, L.: Signal processing advances in wireless and mobile communications. Prentice–Hall (2001)
15. Gillespie, B.W., Atlas, L.E.: Acoustic diversity for improved speech recognition in reverberant environments. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 557–600 (2002)
16. Gillespie, B.W., Malvar, H.S., Florêncio, D.A.F.: Speech dereverberation via maximum-kurtosis subband adaptive filtering. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 3701–3704 (2001)
17. Habets, E.A.P.: Multi-channel speech dereverberation based on a statistical model of late reverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 173–176 (2005)

18. Harville, D.A.: Matrix algebra from a statistician's perspective. Springer (1997)
19. Haykin, S.: Adaptive filter theory, 3rd edn. Prentice–Hall (1996)
20. Haykin, S.: Unsupervised adaptive filtering: blind source separation. Wiley Interscience (2000)
21. Juang, B., Rabiner, L.: Mixture autoregressive hidden Markov models for speech signals. IEEE Trans. Acoust., Speech, Signal Process. **ASSP-33**(6), 1404–1413 (1985)
22. Kailath, T., Sayed, A.H., Hassibi, B.: Linear estimation. Prentice–Hall (2000)
23. Kameoka, H.: Statistical approach to multipitch analysis. Ph.D. thesis, The University of Tokyo (2007)
24. Kinoshita, K., Delcroix, M., Nakatani, T., Miyoshi, M.: A linear prediction-based microphone array for speech dereverberation in a realistic sound field. In: Proc. of Audio Engineering Society 13th Regional Convention (2007)
25. Kinoshita, K., Nakatani, T., Miyoshi, M.: Dereverberation of highly reverberant convolutive mixtures based on multi-step linear prediction. In: Proc. Int. Symp. on Circuits and Systems (2008)
26. Li, K., Swamy, M.N.S., Ahmad, M.O.: An improved voice activity detection using higher order statistics. IEEE Trans. Speech Audio Process. **13**(5), 965–974 (2005)
27. Mitra, S.K.: Optimal inverse of a matrix. Sankhya **37(A)**, 550–563 (1975)
28. Miyoshi, M.: Estimating AR parameter-sets for linear-recurrent signals in convolutive mixtures. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 585–589 (2003)
29. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Speech Audio Process. **36(2)**, 145–152 (1988)
30. Nakatani, T., Juang, B., Hikichi, T., Yoshioka, T., Kinoshita, K., Delcroix, M., Miyoshi, M.: Study on speech dereverberation with autocorrelation codebook. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) pp. 193–197 (2007)
31. Nakatani, T., Kinoshita, K., Miyoshi, M.: Harmonicity based blind dereverberation for single-channel speech signals. IEEE Trans. Audio, Speech, Lang. Process. **15**(1), 80–95 (2007)
32. Nelson, P.A., Orduña-Bustamante, F., Hamada, H.: Multichannel signal processing techniques in the reproduction of sound. J. Audio Eng. Soc. **44(11)**, 973–989 (1996)
33. Qiu, W., Hua, Y., Abed-Meraim, K.: A subspace method for the computation of the GCD of polynomials. Automatica **33**(4), 741–743 (1997)
34. Rombouts, S., Heyde, K.: An accurate and efficient algorithm for the computation of the characteristic polynomial of a general square matrix. J. Comput. Phys. **140**, 453–458 (1998)
35. Slock, D.T.M.: Blind fractionally-spaced equalization, perfect-reconstruction filter banks and multichannel lineawr prediction. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. IV, pp. 585–588 (1994)
36. Sun, X., Douglas, S.: A natural gradient convolutive blind source separation algorithm for speech mixtures. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 59–64 (2001)
37. Tashev, I., Allred, D.: Reverberation reduction for improved speech recognition. In: Proc. Hands-Free Communication and Microphone Arrays (2005)
38. van Trees, H.L.: Optimum array processing. Wiley Interscience (2002)
39. Yegnanarayana, B., Murthy, P.S.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Speech Audio Process. **8**(3), 267–281 (2000)
40. Yoshioka, T., Hikichi, T., Miyoshi, M.: Dereverberation by using time-variant nature of speech production system. EURASIP J. Advances in Signal Process. **2007**(Article ID 65698), doi:10.1155/2007/65698 (2007)
41. Zhao, Y.: An EM algorithm for linear distortion channel estimation based on observations from a mixture of Gaussian sources. IEEE Trans. Speech Audio Process. **7**(4), 400–413 (1999)

Chapter 10
TRINICON for Dereverberation of Speech and Audio Signals

Herbert Buchner[1] and Walter Kellermann[2]

Abstract In this chapter, we develop an analytical top-down approach to the problem of blind dereverberation of speech and audio signals based on TRINICON (TRIple-N Independent component analysis for CONvolutive mixtures), a general framework for broadband adaptive Multi-Input Multi-Output (MIMO) signal processing. Two fundamentally different approaches to the dereverberation problem for realistic scenarios can be distinguished: The "identification-and-inversion approach", which results in a two-step procedure consisting of blind identification of the acoustic MIMO mixing system, followed by an inversion of the identified system. As an alternative, the "direct-inverse approach" blindly estimates the inverse of the acoustic mixing system directly. As shown in this chapter, for both cases TRINICON yields the information-theoretically optimum estimation procedures in a unified way and allows for a direct comparison between the approaches, paves the way to synergies, and yields various useful insights for practical realizations. This chapter also relates other known algorithms, and presents novel improved algorithms as special cases of the generic concept.

10.1 Introduction

Blind signal processing of convolutive mixtures of unknown time series is an important building block in modern systems involving broadband signal acquisition by sensor arrays in multipath or convolutive environments. A challenging and important example for such environments is given by 'natural' acoustic human/machine interfaces using multiple microphones to support sound signal acquisition so that the users may be untethered and mobile in real rooms. To obtain the desired source signals, the signal processing generally has to cope with two fundamental problems due to the distance between the sources and the sensors: (i) the presence of

[1] Deutsche Telekom Laboratories, Berlin University of Technology, Germany
[2] University of Erlangen-Nuremberg, Germany

additive noise and interferers, e.g., competing speakers, and (ii) the disturbing effect of reflections and scattering of the desired source signals in the recordings. In this chapter we tackle these problems by blind adaptive Multi-Input Multi-Output (MIMO) filtering.

In this introductory section, we first formulate the fundamental adaptive filtering problems and distinguish 'direct' and 'inverse' problems in Sect. 10.1.1. Moreover, we introduce a classification into two different generic approaches to blind deconvolution that are fundamental to the dereverberation approaches for speech and audio signals. In Sect. 10.1.2 we introduce a compact matrix notation, which we will use throughout this chapter. Section 10.1.3 provides an overview of our analysis of the two generic approaches to blind deconvolution as useful for blind dereverberation.

10.1.1 Generic Tasks for Blind Adaptive MIMO Filtering

The signal acquisition scenario mentioned above is modeled such that the original source signals $s_q(n)$, $q = 1,\ldots,Q$ are filtered by a linear MIMO system before they are picked up by the sensors yielding the sensor signals $x_p(n)$, $p = 1,\ldots,P$. In this chapter, we describe this MIMO mixing system by length-M Finite Impulse Response (FIR) filters, i.e.,

$$x_p(n) = \sum_{q=1}^{Q} \sum_{\kappa=0}^{M-1} h_{qp,\kappa} s_q(n-\kappa), \tag{10.1}$$

where $h_{qp,\kappa}$, $\kappa = 0,\ldots,M-1$ denote the coefficients of the FIR filter model from the q^{th} source signal $s_q(n)$ to the p^{th} sensor signal $x_p(n)$ according to Fig. 10.1. Throughout this chapter, we assume that the number Q of sources is less or equal to the number P of sensors. The cases $Q < P$ and $Q = P$ are of particular interest as detailed below and they are commonly known as *overdetermined* and *(fully) determined*, respectively. Note that in general, the sources $s_q(n)$ may or may not be all simultaneously active at a particular instant of time.

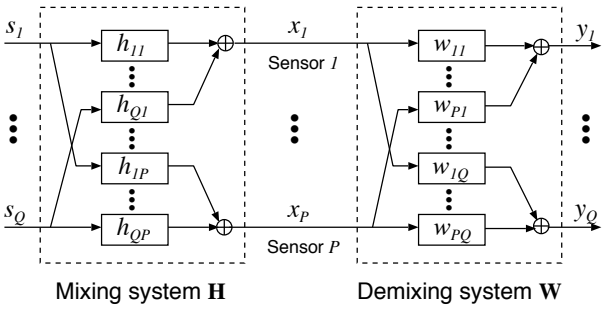


Fig. 10.1 Setup for blind MIMO signal processing

Obviously, since only the sensor signals, i.e., the output signals of the mixing system, are assumed to be accessible to the blind signal processing, *any* type of linear blind adaptive MIMO signal processing may be described by the structure shown in Fig. 10.1. Thus, with respect to a yet undefined optimization criterion, we are interested in finding a corresponding demixing system by the blind adaptive signal processing whose output signals $y_q(n)$ are described by

$$y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq,\kappa} x_p(n-\kappa), \tag{10.2}$$

and where the parameter L denotes the FIR filter length of the demixing filters with coefficients $w_{pq,\kappa}$.

Depending on the optimization criterion for determining the coefficients $w_{pq,\kappa}$, we distinguish two general classes of blind signal processing problems as summarized in Table 10.1 along with the corresponding supervised problems[1,2].

- *Direct blind adaptive filtering problems*: This class summarizes here Blind System Identification (BSI) and Blind Source Separation (BSS)/blind interference cancellation for convolutive mixtures.
  In the BSS approach, we want to determine a MIMO FIR demixing filter that separates the signals *up to an – in general arbitrary – filtering and permutation ambiguity* by forcing the output signals to be mutually independent. Traditionally, and perhaps somewhat misleadingly, BSS has often been considered to be an inverse problem in the literature, e.g., [32, 51]. In another interpretation, BSS may be considered as a set of *blind beamformers* [6, 25] under certain restricting conditions, most notably the fulfillment of the spatial sampling theorem by the microphone array. Furthermore, under the farfield assumption, the directions of arrival can be extracted from the corresponding array patterns, which in turn can be calculated from the BSS filter coefficients, e.g., [63].
  In this chapter (Sect. 10.3) we will see that, more generally, a properly designed broadband BSS system actually performs blind MIMO system identification (which is independent of the spatial sampling theorem). The general broadband approach presented here unifies the BSS and BSI concepts and provides various algorithmic synergy effects and new applications. One important and particularly illustrative application of the general broadband approach to MIMO BSI is the acoustic localization of multiple simultaneously active sources even in reverberant environments as detailed in [19, 21]. In this chapter, we utilize the gen-

[1] Note that in supervised adaptive filtering one may distinguish the analogous general classes of problems. There, we classify system identification and interference cancellation after [45] as (there may be others, or at least other terms) "direct supervised adaptive filtering problems", whereas inverse modeling and linear prediction after [45] may be classified as "inverse supervised adaptive filtering problems".

[2] The TRINICON framework for broadband adaptive MIMO filtering presented in Sect. 10.4 is applicable to all of the problems listed in Table 10.1 and yields corresponding generic adaptation algorithms.

eral MIMO BSI approach for deconvolution and especially to dereverberation of acoustic signals (see below) as another new application.

- *Inverse blind adaptive filtering problems*: This class stands here for MultiChannel Blind Deconvolution (MCBD) and so-called MultiChannel Blind Partial Deconvolution (MCBPD)[3] with respect to the mixing system $\mathbf{H}$ and forms the main part of this chapter. Furthermore, the linear prediction problem as known from the literature on supervised adaptive filtering may also be considered as an inverse blind adaptive filtering problem, as we show in this chapter. The relation between linear prediction and MCBD/MCBPD will also be shown later in this chapter.

  The goal of any *blind deconvolution* approach is to recover the original signals *up to an arbitrary* (frequency-independent) *scaling and possibly a time shift*. In the general MIMO case, i.e., for multiple simultaneously active sources, blind deconvolution also includes separation of the source signals (up to a permutation ambiguity). MCBD and MCBPD provide adaptive methods to the blind deconvolution problem for independent identically distributed (i.i.d.) sources and for general nonwhite sources, respectively.

  For the intended acoustic applications, i.e., for speech and audio source signals, the problem of blind deconvolution means that we want to *dereverberate* the signals by inverting the effect of the convolutive mixture matrix $\mathbf{H}$. In this case, blind deconvolution is denoted by *blind dereverberation*. Furthermore, for blind dereverberation, i.e., in acoustic applications, we typically have to deal with nonwhite sources. Hence, for a direct adaptive approach to blind dereverberation the more general MCBPD method has to be used, as we will discuss later in more detail.

  In terms of the MIMO system description, for the task of blind deconvolution/blind dereverberation, strictly speaking, an inversion of (long and usually nonminimum-phase) room impulse responses is necessary. However, using the Multiple-input/output INverse Theorem (MINT) [68], any MIMO FIR system $\mathbf{H}$ can exactly be inverted by a MIMO FIR system $\mathbf{W}$ if P, Q, and L are suitably chosen, and if the impulse responses $h_{qp}\ \forall\ p \in \{1,\ldots,P\}$ do not have common zeros in the z-plane. Therefore, in principle, there is a general solution to the MCBD problem by using multiple sensors. In this chapter we present adaptive blind deconvolution algorithms that should ideally converge to the ideal MINT solution.

From the two classes of blind adaptive filtering problems shown in Table 10.1, it becomes obvious that two different fundamental approaches to effective blind deconvolution – and thus to dereverberation – are conceivable.

One approach is to perform blind MIMO system identification as mentioned above, followed by a (MINT-based) inversion of the estimated mixing system,

[3] Later in Sect. 10.6 we will see that in practical systems for the blind deconvolution tasks it is important to take the spectral characteristics of the source signals into account. The method of multichannel blind *partial* deconvolution, introduced in Sect. 10.6 to address this issue, also belongs to the class of inverse blind adaptive filtering problems.

Table 10.1 Classification of the linear adaptive filtering problems

	Supervised adaptive filtering problems (after [45])	Blind adaptive filtering problems (treated in this chapter)
"Direct adaptive filtering problems"	System identification	Blind system identification
	Interference cancellation	Blind source separation/ blind interference cancellation
"Inverse adaptive filtering problems"	Inverse modeling/equalization	Blind (partial) deconvolution
	Linear prediction	Linear prediction

e.g., [36, 43]. In this chapter we refer to this approach as the *Identification-and-Inversion approach* (II approach) to blind deconvolution.

The other, theoretically equivalent but, as we will see later, in practice often more reliable approach is to perform directly a blind estimation of the actual inverse of the MIMO mixing system, e.g., [4, 16, 28, 40]. In this chapter we refer to this approach as the *Direct-Inverse approach* (DI approach) to blind deconvolution. Note that for blind *dereverberation*, the DI approach implies the application of MCBPD for nonwhite signals.

10.1.2 A Compact Matrix Formulation for MIMO Filtering Problems

To compactly formulate and analyze the blind adaptive MIMO filtering problems in Sects. 10.2 and 10.3, respectively, we introduce the following matrix formulation of the overall system in Fig. 10.1 consisting of the mixing and demixing systems. This matrix formulation is also used in the TRINICON (TRIple-N Independent component analysis for CONvolutive mixtures) framework described later in Sect. 10.4 in order to blindly estimate the adaptive demixing filter coefficients.

For capturing the mixing system with coefficients $h_{qp,\kappa}$, $\kappa = 0,\ldots,M-1$ and the demixing system with coefficients $w_{pq,\kappa}$, $\kappa = 0,\ldots,L-1$, $p = 1,\ldots,P$, $q = 1,\ldots,Q$, we form the $QM \times P$ mixing coefficient matrix

$$\breve{\mathbf{H}} = \begin{bmatrix} \mathbf{h}_{11} & \cdots & \mathbf{h}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{h}_{Q1} & \cdots & \mathbf{h}_{QP} \end{bmatrix} \tag{10.3}$$

and the $PL \times Q$ demixing coefficient matrix

$$\check{\mathbf{W}} = \begin{bmatrix} \mathbf{w}_{11} & \cdots & \mathbf{w}_{1Q} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{P1} & \cdots & \mathbf{w}_{PQ} \end{bmatrix}, \tag{10.4}$$

respectively, where

$$\mathbf{h}_{qp} = \left[h_{qp,0}, \ldots, h_{qp,M-1}\right]^{\mathrm{T}}, \tag{10.5}$$

$$\mathbf{w}_{pq} = \left[w_{pq,0}, \ldots, w_{pq,L-1}\right]^{\mathrm{T}} \tag{10.6}$$

denote the coefficient vectors of the individual FIR filters of the MIMO systems, and where superscript $^{\mathrm{T}}$ denotes transposition of a vector or a matrix. The downwards pointing hat symbol ('check') on top of $\mathbf{H}$ and $\mathbf{W}$ in (10.3) and (10.4) serves to distinguish these *condensed* matrices from the corresponding larger matrix structures as introduced below in (10.10). Although seemingly a merely formal peculiarity, the rigorous distinction between these different matrix structures is an essential tool for the development of the general TRINICON framework, as shown later.

Analogously, the coefficients $c_{qr,\kappa}$, $q = 1, \ldots, Q$, $r = 1, \ldots, Q$, $\kappa = 0, \ldots, M+L-2$ of the overall system of length $M+L-1$ from the sources to the demixing filter outputs are combined into the $Q(M+L-1) \times Q$ matrix,

$$\check{\mathbf{C}} = \begin{bmatrix} \mathbf{c}_{11} & \cdots & \mathbf{c}_{1Q} \\ \vdots & \ddots & \vdots \\ \mathbf{c}_{Q1} & \cdots & \mathbf{c}_{QQ} \end{bmatrix}, \tag{10.7}$$

where

$$\mathbf{c}_{qr} = \left[c_{qr,0}, \ldots, c_{qr,M+L-2}\right]^{\mathrm{T}}. \tag{10.8}$$

All these subfilter coefficients $c_{qr,\kappa}$ are obtained by convolving the mixing filter coefficients with the demixing filter coefficients. In general, a convolution of two such finite-length sequences can also be written as a matrix-vector product so that the coefficient vector for the model from the q^{th} source to the r^{th} output here reads

$$\mathbf{c}_{qr} = \sum_{p=1}^{P} \mathbf{H}_{qp,[L]} \mathbf{w}_{pr}. \tag{10.9}$$

The so-called *convolution matrix* or *Sylvester matrix* $\mathbf{H}_{qp,[L]}$ of size $M+L-1 \times L$ in this equation exhibits a special structure, containing M filter taps in each column,

$$\mathbf{H}_{qp,[L]} = \begin{bmatrix} h_{qp,0} & 0 & \cdots & 0 \\ h_{qp,1} & h_{qp,0} & \ddots & \vdots \\ \vdots & h_{qp,1} & \ddots & 0 \\ h_{qp,M-1} & \vdots & \ddots & h_{qp,0} \\ 0 & h_{qp,M-1} & \ddots & h_{qp,1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & h_{qp,M-1} \end{bmatrix}. \tag{10.10}$$

The additional third index in brackets denotes the width of the Sylvester matrix, which has to correspond to the length of the column vector $\mathbf{w}_{pr}$ in (10.9) so that the matrix-vector product is equivalent to a linear convolution. The brackets serve to emphasize this fact and to clearly distinguish the meaning of this index from the meaning of the third index of the individual matrix elements, e.g., i of $h_{qp,i}$ in (10.10).

We may now compactly express the overall system matrix $\check{\mathbf{C}}$ after (10.7) using this Sylvester matrix formulation to finally obtain

$$\check{\mathbf{C}} = \mathbf{H}_{[L]}\check{\mathbf{W}}, \tag{10.11}$$

where $\mathbf{H}_{[L]}$ denotes the $Q(M+L-1)\times PL$ MIMO block Sylvester matrix combining all channels,

$$\mathbf{H}_{[L]} = \begin{bmatrix} \mathbf{H}_{11,[L]} & \cdots & \mathbf{H}_{1P,[L]} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{Q1,[L]} & \cdots & \mathbf{H}_{QP,[L]} \end{bmatrix}. \tag{10.12}$$

Based on this matrix formulation, we are now able to compactly formulate the blind adaptive MIMO filtering problems in the coming Sects. 10.2 and 10.3 and to discuss the corresponding ideal solutions, regardless of how the adaptation is actually performed in practice (note that this also implies that the results are valid for both blind and supervised adaptation). The blind adaptation of the demixing filter coefficients towards these ideal solutions will be treated later in Sects. 10.4–10.6.

10.1.3 Overview of this Chapter

This chapter consists of three parts. Based on the matrix notation in Sect. 10.1.2, we formulate and analyze both the above-mentioned inverse and the direct blind adaptive MIMO filtering problems in Sects. 10.2 and 10.3, respectively, and we relate these categories of adaptive MIMO filtering problems to the two fundamental approaches to blind deconvolution, i.e., the DI approach and the II approach. As it turns out, the explicit formulation and analysis of the theoretically ideal solution of the direct filtering problems is somewhat more involved and less well known than

that of the inverse filtering problem. Accordingly, Sect. 10.3 gives a detailed review of a recent comprehensive treatment [19] of the direct filtering problems. Thereby, a fundamental relation between BSI and BSS for convolutive mixtures is of particular practical importance. The resulting practical scheme for BSI serves as a basis for the identification-and-inversion approach to blind deconvolution in the general MIMO case. In this respect, Sect. 10.3 follows the ideas first outlined in [18, 21].

Section 10.4 constitutes the second major part of this chapter and is devoted to the adaptation of the MIMO demixing system towards the ideal solutions discussed in Sects. 10.2 and 10.3. Our considerations are based on TRINICON, a previously introduced versatile framework for broadband adaptive MIMO signal processing [13, 15–17], which is especially well suited for speech and audio signals. The general information-theoretic optimization criterion of TRINICON allows us to exploit all fundamental properties of the excitation signals, such as their nonstationarity, their spectral characteristics (nonwhiteness), and their probability densities (nongaussianity). Moreover, in addition to the inherent broadband structure necessary for a proper system identification and deconvolution, the top-down, i.e., *deductive* approach of the TRINICON framework also allows us to present relations to both already known and new efficient algorithms. So far, this deductive approach has already led to various new insights into the several classes of adaptive filtering problems shown in Table 10.1, most notably blind source separation [15, 19], blind system identification including a generic framework for source localization [19], and the corresponding supervised adaptive problems [23]. Based on the ideas first outlined in [16], the aim of this chapter is to consider TRINICON for inverse blind adaptive problems in more detail.

In the third part of this chapter we first apply TRINICON to BSS and the identification-and-inversion approach to blind deconvolution/blind dereverberation in Sect. 10.5, followed by the application to the direct-inverse approach in Sect. 10.6. As in the previously studied classes of adaptive filtering problems, we will see that the general framework again allows us to relate various known and seemingly different algorithms for dereverberation, and it also yields improvements beyond the current state of the art. Section 10.7 presents results for both the II approach and the DI approach.

10.2 Ideal Inversion Solution and the Direct-inverse Approach to Blind Deconvolution

This section presents a concise summary on the ideal inversion solution for MIMO FIR systems. This inversion solution represents the ideal solution of the DI approach to blind deconvolution. Hence, its discussion also yields important guidelines for the design of the adaptive system based on the DI approach.

As mentioned above, the aim of the inverse adaptive filtering problem is to recover the original signals $s_q(n)$, $q = 1, \ldots, Q$, as shown in Fig. 10.1, up to an arbitrary frequency-independent scaling, time shift, and possibly a permutation of the

demixing filter outputs. Disregarding the potential permutation among the output signals,[4] this condition may be expressed in terms of an *ideal* $Q(M+L-1)\times Q$ *overall system matrix*

$$\check{\mathbf{C}}_{\text{ideal,inv}} = \text{Bdiag}\left\{[0,\ldots,0,1,0,\ldots,0]^{\text{T}},\ldots,[0,\ldots,0,1,0,\ldots,0]^{\text{T}}\right\}\boldsymbol{\Lambda}_{\alpha}, \quad (10.13)$$

where the $\text{Bdiag}\{\cdot\}$ operator describes a block-diagonal matrix containing the listed vectors on the main diagonal. Here, these target vectors, i.e., the ideal overall impulse responses, represent pure delays. The diagonal matrix $\boldsymbol{\Lambda}_{\alpha} = \text{Diag}\left\{[\alpha_1,\ldots,\alpha_Q]^{\text{T}}\right\}$ accounts for the scaling ambiguity. The *condition for the ideal inversion solution* thus reads as

$$\mathbf{H}_{[L]}\check{\mathbf{W}} = \check{\mathbf{C}}_{\text{ideal,inv}}. \quad (10.14)$$

This system of linear equations may generally be solved exactly or approximately by the Moore–Penrose pseudoinverse (e.g., [44]), denoted by $^{+}$, so that

$$\begin{aligned}\check{\mathbf{W}}_{\text{LS,inv}} &= \mathbf{H}_{[L]}^{+}\check{\mathbf{C}}_{\text{ideal,inv}}\\ &= \left[\mathbf{H}_{[L]}^{\text{T}}\mathbf{H}_{[L]}\right]^{-1}\mathbf{H}_{[L]}^{\text{T}}\check{\mathbf{C}}_{\text{ideal,inv}}. \end{aligned} \quad (10.15)$$

Note that this expression corresponds to the least-squares (LS) solution

$$\check{\mathbf{W}}_{\text{LS,inv}} = \arg\min_{\check{\mathbf{W}}} \|\mathbf{H}_{[L]}\check{\mathbf{W}} - \check{\mathbf{C}}_{\text{ideal,inv}}\|_2^2. \quad (10.16)$$

It can be shown that under certain conditions, which can be fulfilled in practice and are described below, this solution becomes the ideal inversion solution, i.e., the pseudoinverse in (10.15) turns into the true matrix inverse,

$$\check{\mathbf{W}}_{\text{ideal,inv}} = \mathbf{H}_{[L]}^{-1}\check{\mathbf{C}}_{\text{ideal,inv}}. \quad (10.17)$$

The principle to calculate the exact inverse using (10.17) is known as MINT [68] and is applicable even for mixing systems with nonminimum phase. The basic requirement for $\mathbf{H}_{[L]}$ in order to be invertible is that it is of full row rank. This assumption can be interpreted such that the FIR acoustic impulse responses contained in $\mathbf{H}_{[L]}$ do not possess any common zeros in the z-domain, which usually holds in practice for a sufficient number of sensors [68]. Another requirement for invertibility of $\mathbf{H}_{[L]}$ is that the number of its rows equals the number of its columns, i.e., $Q(M+L-1) = PL$ according to the dimensions noted above in conjunction with (10.12). From this condition, we immediately obtain the *optimum filter length for inversion* [35]:

[4] This could formally be described by an additional permutation matrix in the ideal solution. However, since in many practical cases this ambiguity may be resolved by a signal classification approach or other prior information, we renounced this formal treatment for clarity.

$$L_{\text{opt,inv}} = \frac{Q}{P-Q}(M-1). \tag{10.18}$$

As an important consequence the MIMO mixing system can be inverted exactly even with a finite-length MIMO demixing system, as long as $P > Q$, i.e., the number of sensors is greater than the number of sources. Note that P, Q, M must be such that $L_{\text{opt,inv}}$ is an integer number in order to allow the matrix inversion in (10.17). Otherwise, we have to resort to the general LS approximation (10.15) with $L_{\text{opt,inv}} = \lceil Q(M-1)/(P-Q) \rceil$.

Based on the generic TRINICON framework for adaptive MIMO filtering in Sect. 10.4, we will present in Sect. 10.6 a coherent overview of blind deconvolution algorithms which aim at the ideal inversion solution (10.15) or the general LS solution (10.17) for a suitable choice of parameters, respectively.

10.3 Ideal Solution of Direct Adaptive Filtering Problems and the Identification-and-inversion Approach to Blind Deconvolution

As an alternative deconvolution approach, the "identification-and-inversion approach" to blind deconvolution is based on a two-step procedure: first, the acoustic MIMO mixing system is blindly identified, and then the identified system is inverted in a separate step. Obviously, for the latter step the results of the previous section can be applied, preferably the MINT solution. In this section, we therefore concentrate on the ideal solution of the system identification step. As we shall see, the relation between source separation and MIMO system identification is of fundamental importance for the practical realization of blind system identification.

In contrast to the inversion problem, the goal of any separation algorithm, such as BSS or conventional beamforming, is to eliminate only the crosstalk between the different sources $s_q(n)$, $q = 1, \ldots, Q$ in the output signals $y_q(n)$, $q = 1, \ldots, Q$ of the demixing system (see Fig. 10.1). Disregarding again a potential permutation among the output signals, this condition may be expressed in terms of the overall system matrix $\check{\mathbf{C}}$ as

$$\check{\mathbf{C}} - \text{bdiag}\left\{\check{\mathbf{C}}\right\} = \text{boff}\left\{\check{\mathbf{C}}\right\} = \mathbf{0}. \tag{10.19}$$

Here, the operator $\text{bdiag}\{\cdot\}$ applied to a block matrix consisting of several submatrices or vectors sets all submatrices or vectors on the off-diagonals to zero. Analogously, the $\text{boff}\{\cdot\}$ operation sets all submatrices or vectors on the diagonal to zero.

With the overall system matrix (10.11), the condition for the ideal separation is expressed as

$$\text{boff}\left\{\mathbf{H}_{[L]}\check{\mathbf{W}}\right\} = \mathbf{0}. \tag{10.20}$$

This relation for the ideal solution of *direct blind adaptive filtering problems* is the analogous expression to the relation (10.14) for the ideal solution of the inverse blind adaptive filtering problems.

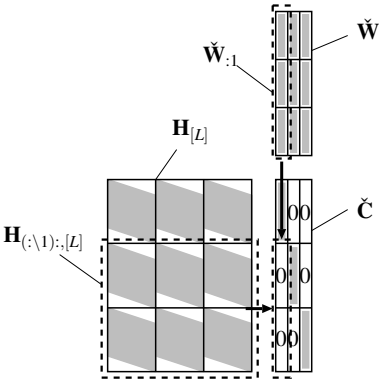


Fig. 10.2 Overall system $\check{\mathbf{C}}$ for the ideal separation, illustrated for $P = Q = 3$

As we will see in this section, the relation (10.20) allows us

- To derive an explicit expression of the ideal separation solution analogously to (10.17).
- To establish a link between BSS and BSI, which will serve as an important basis to the identification-and-inversion approach to blind dereverberation in the general MIMO case
- To establish the conditions for ideal BSI.
- To derive the optimum separation FIR filter length $L_{\text{opt,sep}}$ analogously to (10.18), for which the ideal separation solution (10.19) can be achieved.

If we are only interested in separation with certain other constraints on the output signals, but not in system identification, we may impose further explicit conditions to the block-diagonal elements of $\mathbf{H}_{[L]}\check{\mathbf{W}}$ in addition to the condition (10.20) on the block-offdiagonals. For instance, the so-called *minimum distortion principle* after [67] can, in fact, be regarded as such an additional condition. However, since this is not within the scope of system identification we will not discuss these conditions further in this chapter.

Traditionally, BSS has often been considered as an inverse problem (e.g., [32, 51]). In this section we show that the theoretically ideal convolutive (blind) source separation solution corresponds to blind MIMO system identification. By choosing an appropriate filter length L we show that for broadband algorithms the well-known filtering ambiguity (e.g., [64]) can be avoided. In the following, we consider the ideal broadband solution of mere MIMO separation approaches and relate it to the known blind system identification approach based on single-input multiple-output (SIMO) models [8, 36, 43]. This section follows the ideas outlined in [18, 21]. Some of these ideas were also developed independently in [48] in a slightly different way.

This section discusses the ideal separation condition $\text{boff}\left\{\mathbf{H}_{[L]}\check{\mathbf{W}}\right\} = \mathbf{0}$ as illustrated in Fig. 10.2 for the case $Q = P = 3$. Since in this equation we impose explicit constraints only on the block-offdiagonal elements of $\check{\mathbf{C}}$, this is equivalent to establishing a set of homogeneous systems of linear equations

$$\mathbf{H}_{(:\backslash q):,[L]}\check{\mathbf{W}}_{:q} = \mathbf{0},\ q = 1,\ldots,Q \tag{10.21}$$

to be solved. Each of these systems of equations results from the constraints on one column of $\check{\mathbf{C}}$, as illustrated in Fig. 10.2 for the first column. The notation in the indices in (10.21) indicates that for the q^{th} column $\check{\mathbf{W}}_{:q}$ of the demixing filter matrix $\check{\mathbf{W}}$, we form a submatrix $\mathbf{H}_{(:\backslash q):,[L]}$ of $\mathbf{H}_{[L]}$ by removing the q^{th} row $\mathbf{H}_{q:,[L]}$ of Sylvester submatrices of the original matrix $\mathbf{H}_{[L]}$.

For homogeneous systems of linear equations such as (10.21) it is known that nontrivial solutions $\check{\mathbf{W}}_{:q} \not\equiv \mathbf{0}$ are indeed obtained if the rank of $\mathbf{H}_{(:\backslash q):,[L]}$ is smaller than the number of elements of $\check{\mathbf{W}}_{:q}$. Based on this and later in this section, we will also derive an expression of the optimum separation filter length $L_{\text{opt,sep}}$ for an arbitrary number of sensors and sources analogously to the optimum inversion filter length $L_{\text{opt,inv}}$ in (10.18).

In the following sections, we first discuss the solution of (10.21) for the case $P = Q = 2$ and then generalize the results to more than two sources and sensors.

10.3.1 Ideal Separation Solution for Two Sources and Two Sensors

For the case $Q = P = 2$, the set of homogeneous linear systems of equations (10.21) reads

$$\mathbf{H}_{11,[L]}\mathbf{w}_{12} + \mathbf{H}_{12,[L]}\mathbf{w}_{22} = \mathbf{0}, \tag{10.22a}$$

$$\mathbf{H}_{21,[L]}\mathbf{w}_{11} + \mathbf{H}_{22,[L]}\mathbf{w}_{21} = \mathbf{0}. \tag{10.22b}$$

Since the matrix-vector products in these equations represent convolutions of FIR filters they can equivalently be written as a multiplication in the z-domain:

$$H_{11}(z)W_{12}(z) + H_{12}(z)W_{22}(z) = 0, \tag{10.23a}$$

$$H_{21}(z)W_{11}(z) + H_{22}(z)W_{21}(z) = 0. \tag{10.23b}$$

Due to the FIR filter structure the z-domain representations can be expressed by the zeros $z_{0H_{qp},\nu}$, $z_{0W_{pq},\mu}$ and the gains $A_{H_{qp}}$, $A_{H_{pq}}$ of the filters $H_{qp}(z)$ and $W_{pq}(z)$, respectively:

$$A_{H_{11}} \prod_{\nu=1}^{M-1} (z - z_{0H_{11},\nu}) \cdot A_{W_{12}} \prod_{\mu=1}^{L-1} (z - z_{0W_{12},\mu}) =$$
$$- A_{H_{12}} \prod_{\nu=1}^{M-1} (z - z_{0H_{12},\nu}) \cdot A_{W_{22}} \prod_{\mu=1}^{L-1} (z - z_{0W_{22},\mu}), \tag{10.24a}$$
$$A_{H_{21}} \prod_{\nu=1}^{M-1} (z - z_{0H_{21},\nu}) \cdot A_{W_{11}} \prod_{\mu=1}^{L-1} (z - z_{0W_{11},\mu}) =$$
$$- A_{H_{22}} \prod_{\nu=1}^{M-1} (z - z_{0H_{22},\nu}) \cdot A_{W_{21}} \prod_{\mu=1}^{L-1} (z - z_{0W_{21},\mu}). \tag{10.24b}$$

Analogously to the case of MINT [68] described in the previous section, we assume that the impulse responses contained in $\mathbf{H}_{(:\backslash q):,[L]}$, i.e., $H_{11}(z)$ and $H_{12}(z)$ in (10.24a) and $H_{21}(z)$ and $H_{22}(z)$ in (10.24b), respectively, do not share common zeros. If no common zeros exist and if we choose the *optimum*[5] *filter length for the case* $Q = P = 2$ as $L_{\text{opt,sep}} = M$, then the equality in (10.24a) can only hold if the zeros of the demixing filters are chosen as $z_{0W_{12},\mu} = z_{0H_{12},\mu}$ and $z_{0W_{22},\mu} = z_{0H_{11},\mu}$ for $\mu = 1,\ldots,M-1$. Analogously, the equality in (10.24b) can only hold if $z_{0W_{11},\mu} = z_{0H_{22},\mu}$ and $z_{0W_{21},\mu} = z_{0H_{21},\mu}$ for $\mu = 1,\ldots,M-1$. Additionally, to fulfill the equality, the gains of the demixing filters in (10.24a) have to be chosen as $A_{W_{22}} = \alpha_2 A_{H_{11}}$ and $A_{W_{12}} = -\alpha_2 A_{H_{12}}$, where α_2 is an arbitrary scalar constant. Thus, the demixing filters are only determined up to a scalar factor α_2. Analogously, for the equality (10.24b) the gains of the demixing filters are given as $A_{W_{11}} = \alpha_1 A_{H_{22}}$ and $A_{W_{21}} = -\alpha_1 A_{H_{21}}$ with the scalar constant α_1.

In summary, this leads to the ideal separation filter matrix $\check{\mathbf{W}}_{\text{ideal,sep}}$ given in the time domain as

$$\check{\mathbf{W}}_{\text{ideal,sep}} = \begin{bmatrix} \alpha_1 \mathbf{h}_{22} & -\alpha_2 \mathbf{h}_{12} \\ -\alpha_1 \mathbf{h}_{21} & \alpha_2 \mathbf{h}_{11} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{22} & -\mathbf{h}_{12} \\ -\mathbf{h}_{21} & \mathbf{h}_{11} \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}, \tag{10.25}$$

where due to the scaling ambiguity each column is multiplied by an unknown scalar α_q.

From (10.25) we see that under the conditions put on the zeros of the mixing system in the z-domain, and for $L = L_{\text{opt,sep}}$, this *ideal separation solution corresponds to a MIMO system identification up to an arbitrary scalar constant*. Thus, a suitable algorithm that is able to perform *broadband* BSS *under these conditions* can be used for blind MIMO system identification (if the source signals provide sufficient spectral support for exciting the mixing system). In Sect. 10.4, a suitable algorithmic framework for this task will be presented. Moreover, as we will see in the following section, this approach can be seen as a generalization of the state-of-the-art method for the blind identification of SIMO systems.

[5] Note that for $L < L_{\text{opt,sep}} = M$ it is obviously not possible to compensate all zeros of $H_{11}(z)$ and $H_{12}(z)$ by $W_{22}(z)$ and $W_{12}(z)$, respectively. On the other hand, in the case $L > L_{\text{opt,sep}} = M$, the filters $W_{12}(z)$ and $W_{22}(z)$ will exhibit $L - M$ arbitrary common zeros, which are undesired. We will consider the practically important issue of order-overestimation in Sect. 10.3.5.

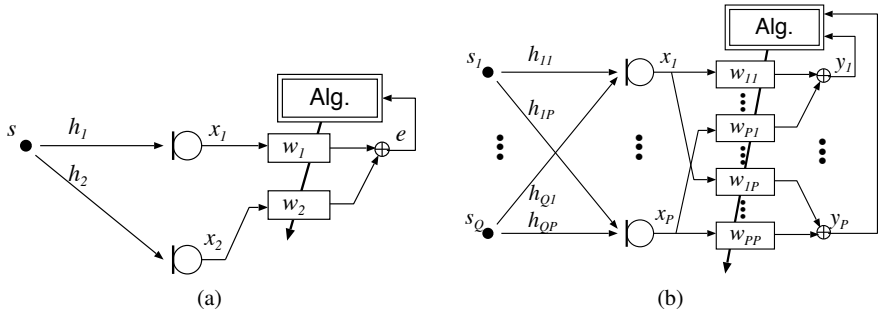


Fig. 10.3 Blind system identification based on (a) SIMO and (b) MIMO models

In practice, the difficulty of finding the correct filter length $L_{\text{opt,sep}}$ is obviously another important issue since the length M of the mixing system is generally unknown. In Sect. 10.3.5 we will address this problem and the consequences of overestimation and underestimation, respectively.

10.3.2 Relation to MIMO and SIMO System Identification

From a system-theoretic point of view, the BSS approach aiming at the ideal solution (10.25) can be interpreted as a generalization of the popular class of blind SIMO system identification approaches, e.g., [36, 43, 61], as illustrated in Fig. 10.3(a).

The main reason for the popularity of this SIMO approach is that the optimum filters can be found as the result of a relatively simple least-squares error minimization. From Fig. 10.3(a) and for $e(n) = 0$ it follows for sufficient excitation $s(n)$ that

$$h_1(n) * w_1(n) = -h_2(n) * w_2(n). \tag{10.26}$$

This can be expressed in the z-domain as $H_1(z)W_1(z) = -H_2(z)W_2(z)$. Comparing this error cancelling condition with the ideal separation conditions (10.23a) and (10.23b), we immediately see that the SIMO-based approach does indeed correspond exactly to one of the separation conditions, and for deriving the ideal solution, we may apply exactly the same reasoning as in the MIMO case above. Thus, assuming that $H_1(z)$ and $H_2(z)$ have no common zeros, the equality of (10.26) can only hold if the filter length is chosen again as $L = M$. Then, this leads to the ideal cancellation filters $W_1(z) = \alpha H_2(z)$ and $W_2(z) = -\alpha H_1(z)$, which can be determined up to an arbitrary scaling by the factor α as in the MIMO case. For $L > M$, the scaling ambiguity would result in arbitrary *filtering*. For the SIMO case, this scaling ambiguity was derived similarly in [36].

Note that the SIMO case may also be interpreted as a special 2×2 MIMO case according to Fig. 10.3(b) with the specialization being that one of the sources is

always identical to zero so that the BSS output corresponding to this (virtual) source must also be identical to zero, whereas the other BSS output signal is not of interest in this case. This again leads to the cancellation condition (10.26) and illustrates that the relation between broadband BSS and SIMO-based BSI will also hold from an algorithmic point of view, i.e., known adaptive solutions for SIMO BSI can also be derived as special cases of the algorithmic framework for the MIMO case.

Adaptive algorithms performing the error minimization mentioned above for the SIMO structure have been proposed in the context of blind deconvolution, e.g., in [36, 43], and blind system identification for passive source localization, e.g., in [8, 27]. In the latter case, this algorithm is also known as the Adaptive Eigenvalue Decomposition (AED) algorithm, which points to the fact that, in the SIMO case, the homogeneous system of equations (10.21) may be reformulated as an analogous signal-dependent homogeneous system of equations containing the sensor-signal correlation matrix instead of the mixing filter matrix. The solution vector (in the SIMO case the matrix $\check{\mathbf{W}}$ reduces to a vector) of the homogeneous system can then be interpreted as the *eigenvector corresponding to the zero-valued (or smallest) eigenvalue* of the sensor correlation matrix. In [27, 43] this SIMO approach, i.e., the single-source case, was also generalized to more than $P = 2$ microphone channels. In Sect. 10.5 we will present how – from an algorithmic point of view – the AED does indeed directly follow from the general TRINICON framework for broadband adaptive MIMO filtering. Moreover, this will lead to a generalization of the original least-squares-based AED algorithm so that it is able to additionally exploit higher-order statistics and also contains an inherent adaptation control. This algorithmic link between the SIMO and MIMO cases will also lead to important insights for the direct-inverse approach to blind deconvolution later in Sect. 10.6.

10.3.3 Ideal Separation Solution and Optimum Separation Filter Length for an Arbitrary Number of Sources and Sensors

As mentioned above, for homogeneous systems of linear equations such as the ideal separation conditions (10.21) it is known that nontrivial solutions $\check{\mathbf{W}}_{:q} \not\equiv \mathbf{0}$ are obtained if the rank of $\mathbf{H}_{(:\backslash q):,[L]}$ is smaller than the number of elements of $\check{\mathbf{W}}_{:q}$. Additionally, as in the case of MINT [68] described in the previous section, we assume that the impulse responses contained in $\mathbf{H}_{(:\backslash q):,[L]}$ do not share common zeros in the z-domain so that $\mathbf{H}_{(:\backslash q):,[L]}$ is assumed to have full row rank. Thus, combining these conditions leads to the requirement that the matrix $\mathbf{H}_{(:\backslash q):,[L]}$ is *wide*, i.e., the number PL of its columns must be greater than the number $(Q-1)(M+L-1)$ of its rows to obtain non-trivial solutions, i.e., $PL > (Q-1)(M+L-1)$. Solving this inequality for L yields the lower bound for the separation filter length as

$$L_{\text{sep}} > \frac{Q-1}{P-Q+1}(M-1). \tag{10.27}$$

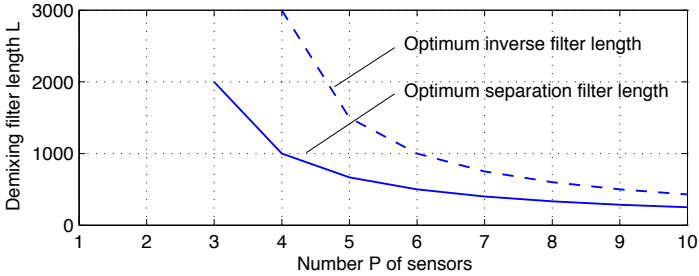


Fig. 10.4 Comparison of the optimum filter lengths for inversion and separation for $M = 1000$ and $Q = 3$

The difference between the number of columns of $\mathbf{H}_{(:\backslash q):,[L]}$ and the number of rows further specifies the dimension of the space of possible non-trivial solutions $\check{\mathbf{W}}_{:q}$, i.e., the number of linearly independent solutions spanning the solution space. Obviously, due to the bound derived above, the best choice we can make to narrow down the solutions is a one-dimensional solution space, i.e., $PL = (Q-1)(M+L-1)+1$. Now solving this *equality* for L and choosing the integer value to be strictly larger than the above bound finally results in the *optimum separation filter length* as

$$L_{\text{opt,sep}} = \frac{(Q-1)(M-1)+1}{P-Q+1}. \tag{10.28}$$

Note that narrowing down the solution space to a one-dimensional space by this choice of filter length means precisely that in this case the *filtering ambiguity of BSS reduces to an arbitrary scaling*. These considerations show that this is possible even for an arbitrary number P of sensors and an arbitrary number Q of sources, where $P \geq Q$. However, the parameters P, Q, M must be such that $L_{\text{opt,sep}}$ is an integer number in order to allow the ideal separation solution. Otherwise, we have to resort to approximations by choosing, e.g., the next higher integer, i.e., $L_{\text{opt,sep}} = \lceil [(Q-1)(M-1)+1]/(P-Q+1) \rceil$.

To actually obtain the ideal separation solution $\check{\mathbf{W}}_{\text{ideal,sep}}$ with (10.28) for the general, i.e., not necessarily square case $P \geq Q$, we again consider the original set of homogeneous systems of linear equations (10.21). For the choice $L = L_{\text{opt,sep}}$, we may easily augment the matrix $\mathbf{H}_{(:\backslash q):,[L]}$ to a square matrix $\tilde{\mathbf{H}}_{(:\backslash q):,[L]}$ by adding one row of zeros on both sides of (10.21). The corresponding augmented set of linear systems of equations

$$\tilde{\mathbf{H}}_{(:\backslash q):,[L]} \check{\mathbf{W}}_{:q} = \mathbf{0}, \; q = 1, \ldots, Q \tag{10.29}$$

is equivalent to the original set (10.21). However, we may now interpret the *general solution vector* $\check{\mathbf{W}}_{:q}$ *of (10.21) for the* q^{th} *column of* $\check{\mathbf{W}}$ as *the eigenvector corresponding to the zero-valued eigenvalue of the augmented matrix* $\tilde{\mathbf{H}}_{(:\backslash q):,[L]}$.

The general equation (10.28) for the optimum separation filter length plays the same role for BSI as (10.18) for inversion. Comparing these two equations, we can verify that in contrast to the inversion, which requires $P > Q$ for the ideal solution

using FIR filters, the ideal separation condition can be met for $P = Q$. Moreover, for the special case $P = Q = 2$, the general expression (10.28) also confirms the choice $L_{\text{opt,BSS}} = M$ as already obtained in Sect. 10.3.1. Figure 10.4 compares the different optimum filter lengths through an example.

10.3.4 General Scheme for Blind System Identification

In Sects. 10.3.1 and 10.3.2 we have explicitly shown the relation between the ideal separation solution and the mixing system for the two-sensor cases. These considerations also resulted in a link to the well-known SIMO-based system identification method (note that for BSI with more than two sensors, a simple approach is to apply several of these schemes in parallel, e.g., [49]), and also showed that the MIMO case with two simultaneously active sources is a generalization of the SIMO system identification method. In the case of more than two sources we cannot directly extract the estimated mixing system coefficients $h_{qp,\kappa}$ from the separation solution $\check{\mathbf{W}}$. The previous Sect. 10.3.3 generalized the considerations on the two-sensor cases for the *separation* task. In this section, we now outline the generalization of the two-sensor cases in Sects. 10.3.1 and 10.3.2 for the *identification* task which is the first step of the identification-and-inversion approach to blind deconvolution, as detailed in Sect. 10.3.5. The considerations so far suggest the following generic *two-step BSI scheme for an arbitrary number of sources* (where $P \geq Q$):

(1) Based on the available sensor signals, perform a properly designed broadband BSS (see Sect. 10.4) resulting in an estimate of the demixing system matrix.
(2) Analogously to the relation (10.21) between the mixing and demixing systems, and the associated considerations in Sect. 10.3.3 for the separation task, determine an *estimate of the mixing system matrix* using the estimated demixing system from the first step.

In general, to perform step (2) for more than two sources, some further considerations are required. First, an equivalent reformulation of the homogeneous system of equations (10.21) is necessary so that now the *demixing system matrix* instead of the mixing system matrix is formulated *as a blockwise Sylvester matrix*. Note that this corresponds to a *block-transposition* (which we denote here by superscript $^{\text{bT}}$) of (10.21), i.e.,

$$\left(\mathbf{W}^{\text{bT}}\right)_{(:\backslash q):,[M]} \left(\check{\mathbf{H}}^{\text{bT}}\right)_{:q} = \mathbf{0},\ q = 1,\ldots,Q. \tag{10.30}$$

The block-transposition is an extension of the conventional matrix transposition. It means that we keep the original form of the channel-wise submatrices but we may change the order of the mixing and demixing subfilters by exploiting the commutativity of the convolutions. Note that the commutativity property does not hold for the MIMO system matrices as a whole, i.e., $\mathbf{W}_{(:\backslash q):,[M]}$ and $\check{\mathbf{H}}_{:q}$, so that they have to be block-transposed to change their order.

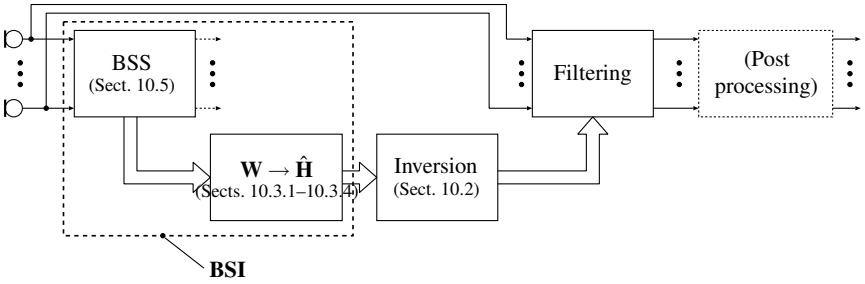


Fig. 10.5 Identification-and-inversion approach to blind dereverberation

Similarly to Sect. 10.3.3, we may then calculate the corresponding estimate of the mixing system in terms of eigenvectors using the complementary form (10.30) of the homogeneous system of equations. Based on this system of equations, we can devise various powerful strategies for BSI in the general MIMO case.

10.3.5 Application of Blind System Identification to Blind Deconvolution

In order to obtain a complete blind dereverberation system after the identification-and-inversion approach, the considerations in the previous sections suggest the structure shown in Fig. 10.5. As discussed above, the acoustic MIMO mixing system can be blindly identified by means of an adaptive broadband BSS algorithm. Algorithmic solutions will be detailed in Sect. 10.5 based on the TRINICON framework outlined in Sect. 10.4. For the subsequent inversion of the estimated mixing system we refer to Sect. 10.2.

Attractive features of the identification-and-inversion approach to blind dereverberation are that (1) it is relatively easy to deal with an increased number of microphone channels (the so-called overdetermined case for blind adaptive filtering) by simple parallelization of BSI algorithms, and (2) the approach is applicable for nearly arbitrary audio source signals, as long as they exhibit sufficient spectral support.

Based on the blind SIMO system identification mentioned in Sect. 10.3.2 (i.e., the estimate of the channel impulse responses is the eigenvector corresponding to the minimum eigenvalue of the correlation matrix), the identification-and-inversion approach to blind dereverberation was proposed, e.g., in [36, 43], for one acoustic source signal.

Using the general scheme for blind MIMO system identification from the previous Sects. 10.3.1–10.3.4 and the TRINICON framework shown below, we are now in a position to generalize the identification-and-inversion approach to multiple simultaneously active sources, i.e., to the MIMO case. Note that the MINT in Sect. 10.2 is already capable of handling the general MIMO case for $P < Q$. As in

the SIMO case, the blind MIMO system identification approach has already been successfully applied in the context of passive source localization in reverberant environments, e.g., in [19, 21].

Note that previously, in [49], the identification-and-inversion approach was discussed for the MIMO case under the assumption that from time to time each source signal occupies a time interval exclusively. Then, during every single-talk interval, a SIMO system was blindly identified and its channel impulse responses were saved for later dereverberation when more than one source was active. Obviously, in practice, the applicability of this approach will be very limited in time-varying environments and with increasing numbers of independent sources (consider, e.g., a cocktail party scenario). In addition, a sophisticated multichannel sound source detection algorithm that distinguishes single and multiple speaker activity would be needed in practice. Such a required multichannel adaptation control is inherently available in TRINICON-based BSS/BSI algorithms for the general MIMO case.

However, both in the SIMO case and in the general MIMO case, there are still some fundamental challenges in the context of this dereverberation approach:

- The channel impulse responses must not exhibit common zeros in the z-domain (both for the system identification (see Sects. 10.3.1 and 10.3.3) and also for the subsequent system inversion (see Sect. 10.2)).
- The filter length must be known exactly (both for the system identification (see Sects. 10.3.1 and 10.3.3) and for the subsequent system inversion (see Sect. 10.2)).

The first problem can be mitigated in practice by increasing the number of microphones so that the probability for common zeros is reduced [68]. Hence, the choice of the correct filter length $L_{\text{opt,sep}}$ is the major remaining difficulty in this approach.[6]

The consequences of overestimation and underestimation of the filter order can be seen, e.g., from (10.24a) and (10.24b). In the case of *underestimation*, i.e., for $L < L_{\text{opt,sep}} = M$ it is obviously not possible to compensate all zeros of $H_{11}(z)$ and $H_{12}(z)$ by $W_{22}(z)$ and $W_{12}(z)$, respectively. The case of *overestimation*, i.e., $L > L_{\text{opt,sep}} = M$, is by far more problematic. In this case, the filters $W_{12}(z)$ and $W_{22}(z)$ will exhibit $L-M$ arbitrary common zeros, which are undesired. This corresponds to the requirement to narrow down the solution space addressed in Sect. 10.3.3, by avoiding an overestimation of the filter length in order to prevent a filtering ambiguity. In other words, in the overestimated case, the ideal blind identification solution $\hat{H}_1(z) = \alpha H_1(z)$ and $\hat{H}_2(z) = \alpha H_2(z)$ turns into $\hat{H}_1(z) = C_{\min}(z)H_1(z)$ and $\hat{H}_2(z) = C_{\min}(z)H_2(z)$ with the *common polynomial* $C_{\min}(z)$ corresponding to an arbitrary filtering. Consequently, after the inverse filtering in Fig. 10.5, the overestimation of the filter length would result in a remaining filtering $1/C_{\min}(z)$ of the original source signals.

[6] Note that in some other applications of blind adaptive filtering we do not require a complete identification of the mixing system. For instance, for acoustic source localization only the positions of the dominant components are required. Fortunately, this is in line with the requirement to avoid an overestimation of the filter length. Thus, in these applications the choice $L \leq L_{\text{opt,sep}}$ is preferable in practice.

Various ways exist to solve the filtering ambiguity problem caused by the overestimation of the filter order. The transfer function order could be obtained if the dimension of the null space in the autocorrelation matrix of the observed signals is precisely calculated [38, 43], i.e., by counting the number of very small eigenvalues. Another way to find the optimum order is to use a suitable cost function, e.g., [9, 36, 77]. Unfortunately, these blind system order estimation approaches are often unreliable (particularly in noisy environments) and computationally too complex (especially the latter ones, i.e., [9, 36, 77]). An alternative approach proposed, e.g., in [46] is to *compensate* for the remaining filtering $1/C_{\min}(z)$ using a post filter (Fig. 10.5) by estimating the common polynomial with a multichannel linear prediction scheme. This approach seems to be numerically very sensitive for large filter lengths. Note also that this latter approach slightly limits the application domain by assuming sources that can be modeled by AR processes, such as speech signals.

A fundamentally different alternative to the identification-and-inversion approach to blind dereverberation is the direct-inverse approach. Here, the aim is to directly estimate the inverse MIMO filter after Sect. 10.2 based on a dereverberation cost function. It is therefore inherently more robust to the order-overestimation problem. However, as we will see later in this chapter, this comes at the cost of the requirement for a more precise stochastic modeling of the source signals, which again specializes the application domain, e.g., to speech signals. Moreover, the direct-inverse approach requires that all microphone channels be taken into account at once, which renders the adaptation more complex.

Similar to the adaptation aspects of the identification-and-inversion approach in Sect. 10.5, we will treat the algorithmic aspects of the direct-inverse approach in Sect. 10.6. Both approaches are presented in a unified way based on TRINICON as outlined next in Sect. 10.4. The unified treatment also allows for an illuminating comparison.

10.4 TRINICON – A General Framework for Adaptive MIMO Signal Processing and Application to Blind Adaptation Problems

For the blind estimation of the coefficients corresponding to the desired solutions discussed in the previous section, we have to consider and exploit the properties of the excitation signals, such as their nonstationarity, their spectral characteristics, and their probability densities.

In the existing literature, the known algorithms for blind system identification, blind source separation, and blind deconvolution were introduced independently. The BSS problem has mostly been addressed for instantaneous mixtures or by narrowband approaches in the frequency domain, which adapt the coefficients independently in each Discrete Fourier Transform (DFT) bin, e.g., [34, 51, 83]. On the other hand, in the case of MCBD, many approaches either aim at whitening the output signals as they are based on an i.i.d. model of the source signals (e.g., [4, 28]), which

is undesirable for generally nonwhite speech and audio signals, as these should not be whitened, or are rather heuristically motivated, e.g., [40].

The aim of this section is to present an overview of the algorithmic part of broadband blind adaptive MIMO filtering based on TRINICON, a generic concept for adaptive MIMO filtering that takes the signal properties of speech and audio signals (nonwhiteness, nonstationarity, and nongaussianity) into account, and allows a unified treatment of broadband BSS (as needed for a proper BSI) and MCBD algorithms as applicable to speech and audio signals in real acoustic environments [13, 15–17]. This framework generally uses multivariate stochastic signal models in the cost function to describe the temporal structure of the source signals and thereby provides a powerful cost function for both, BSS/BSI and MCBD, and, for the latter, also leads to improved algorithms for speech dereverberation.

Although both time-domain and equivalent broadband frequency-domain formulations of TRINICON have been developed with the corresponding multivariate models in both the time domain and the frequency domain [15, 17], in this chapter we mainly consider the time-domain formulation. Furthermore, we restrict ourselves here to gradient-based coefficient updates and disregard Newton-type adaptation algorithms for clarity and brevity. The algorithmic TRINICON framework is directly based on the matrix notation developed above.

Throughout this section, we regard the symmetric case where the number Q of *maximum simultaneously active source signals* $s_q(n)$ is equal to the number of sensor signals $x_p(n)$, i.e., $Q = P$. However, it should be noted that in contrast to other blind algorithms in the Independent Component Analysis (ICA) literature, we do not assume prior knowledge about the exact number of active sources. Thus, even if the algorithms will be derived for $Q = P$, the number of simultaneously active sources may change throughout the application of the TRINICON-based algorithm and only the condition $Q \leq P$ has to be fulfilled.

10.4.1 Matrix Notation for Convolutive Mixtures

To introduce an algorithm for broadband processing of convolutive mixtures, we first need to formulate the convolution of the FIR demixing system of length L in the following matrix form [17]:

$$\mathbf{y}^{\mathrm{T}}(n) = \mathbf{x}^{\mathrm{T}}(n)\mathbf{W}, \tag{10.31}$$

where n denotes the time index, and

$$\mathbf{x}^{\mathrm{T}}(n) = [\mathbf{x}_1^{\mathrm{T}}(n), \ldots, \mathbf{x}_P^{\mathrm{T}}(n)], \tag{10.32}$$

$$\mathbf{y}^{\mathrm{T}}(n) = [\mathbf{y}_1^{\mathrm{T}}(n), \ldots, \mathbf{y}_P^{\mathrm{T}}(n)], \tag{10.33}$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{P1} & \cdots & \mathbf{W}_{PP} \end{bmatrix}, \tag{10.34}$$

$$\mathbf{x}_p^{\mathrm{T}}(n) = [x_p(n), \ldots, x_p(n-2L+1)], \tag{10.35}$$

$$\mathbf{y}_q^{\mathrm{T}}(n) = [y_q(n), \ldots, y_q(n-D+1)] \tag{10.36}$$

$$= \sum_{p=1}^{P} \mathbf{x}_p^{\mathrm{T}}(n)\mathbf{W}_{pq}. \tag{10.37}$$

The parameter D in (10.36), $1 \leq D < L$, denotes the number of lags taken into account to exploit the nonwhiteness of the source signals as shown below. $\mathbf{W}_{pq}$, $p = 1, \ldots, P$, $q = 1, \ldots, P$ denote $2L \times D$ Sylvester matrices that contain all coefficients of the respective filters:

$$\mathbf{W}_{pq} = \begin{bmatrix} w_{pq,0} & 0 & \cdots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,L-1} & \ddots & w_{pq,1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & w_{pq,L-1} \\ 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{10.38}$$

Note that for $D = 1$, (10.31) simplifies to the well-known vector formulation of a convolution, as it is used extensively in the literature on supervised adaptive filtering, e.g., [45].

10.4.2 Optimization Criterion

Various approaches exist to blindly estimate the demixing matrix $\mathbf{W}$ for the above-mentioned tasks by utilizing the following source signal properties [51] which we all combine into an efficient and versatile algorithmic framework [13, 15, 16]:
(i) *Nongaussianity* is exploited by using higher-order statistics for ICA. ICA approaches can be divided into several classes. Although they all lead to similar update

rules, the minimization of the mutual information among the output channels can be regarded as the most general approach to solve direct adaptive filtering problems according to Table 10.1, such as source separation [15, 51] and system identification [19, 23]. To obtain an even more versatile estimator not only allowing spatial separation but also temporal separation for dereverberation and inverse adaptive filtering problems in general, we use the Kullback–Leibler Divergence (KLD) [29] between a certain *desired* joint PDF (essentially representing a hypothesized stochastic source model) and the joint PDF of the actually estimated output signals [16]. Note that the mutual information is a special case of KLD [29]. The desired PDF in the KLD is factorized with respect to the different sources (for the direct adaptive filtering problems, such as source separation) and possibly also with respect to certain temporal dependencies (for inverse adaptive filtering problems, such as dereverberation) as shown below. The KLD is guaranteed to be positive [29], which is a necessary condition for a useful cost function.

(ii) *Nonwhiteness* is exploited by simultaneous minimization of output cross-relations over multiple time-lags. We therefore consider multivariate PDFs, i.e., 'densities including D time-lags'.

(iii) *Nonstationarity* is exploited by simultaneous minimization of output cross-relations at different time-instants. We assume ergodicity within blocks of length N, so that the ensemble average is replaced by time averages over these blocks.

Based on the KLD, we now define the following general cost function taking into account all three fundamental signal properties (i)-(iii):

$$\mathcal{J}(m,\mathbf{W}) = -\sum_{i=0}^{\infty}\beta(i,m)\frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\left\{\log(\hat{p}_{s,PD}(\mathbf{y}(j))) - \log(\hat{p}_{y,PD}(\mathbf{y}(j)))\right\}, \tag{10.39}$$

where $\hat{p}_{s,PD}(\cdot)$ and $\hat{p}_{y,PD}(\cdot)$ are the assumed or estimated PD-variate source model (i.e., desired) PDF and output PDF, respectively. In this chapter we assume that these PDFs are generally described by certain data-dependent parameterizations, so that we can write in more detail

$$\hat{p}_{s,PD} = \hat{p}_{s,PD}\left(\mathbf{y}, \boldsymbol{\mathcal{Q}}_s^{(1)}, \boldsymbol{\mathcal{Q}}_s^{(2)}, \ldots\right) \tag{10.40a}$$

and

$$\hat{p}_{y,PD} = \hat{p}_{y,PD}\left(\mathbf{y}, \boldsymbol{\mathcal{Q}}_y^{(1)}, \boldsymbol{\mathcal{Q}}_y^{(2)}, \ldots\right), \tag{10.40b}$$

respectively. We further assume that the model parameter estimates are given by the generic form

$$\boldsymbol{\mathcal{Q}}_s^{(r)}(i) = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\left\{\boldsymbol{\mathcal{G}}_s^{(r)}(\mathbf{y}(j))\right\}, \quad r = 1,2,\ldots, \tag{10.41a}$$

$$\mathcal{Q}_y^{(r)}(i) = \frac{1}{N} \sum_{j=iN_L}^{iN_L+N-1} \left\{ \mathcal{G}_y^{(r)}(\mathbf{y}(j)) \right\}, \quad r = 1,2,\ldots, \tag{10.41b}$$

where $\mathcal{G}_s^{(r)}$ and $\mathcal{G}_y^{(r)}$ are suitable functions of the observation vectors $\mathbf{y}$, and $\mathcal{Q}_s^{(r)}$ and $\mathcal{Q}_y^{(r)}$ represent block-averages of $\mathcal{G}_s^{(r)}(\mathbf{y})$ and $\mathcal{G}_y^{(r)}(\mathbf{y})$, respectively. In general, the bold calligraphic symbols denote multidimensional arrays, or in other words, tensorial quantities. The elements of $\mathcal{Q}_s^{(r)}$, $\mathcal{Q}_y^{(r)}$, $\mathcal{G}_s^{(r)}$, and $\mathcal{G}_y^{(r)}$ are denoted by $\mathcal{Q}_{s,i_1,i_2,\ldots}^{(r)}$, $\mathcal{Q}_{y,i_1,i_2,\ldots}^{(r)}$, $\mathcal{G}_{s,i_1,i_2,\ldots}^{(r)}$, and $\mathcal{G}_{y,i_1,i_2,\ldots}^{(r)}$, respectively, where $i_1, i_2, \ldots$ are the indices in the corresponding tensor dimensions. Well-known special cases of such parameterizations are estimates of the variance $\hat{\sigma}_y^2(i) = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1} \left\{ y^2(j) \right\}$ and the correlation matrix $\mathbf{R}_{\mathbf{yy}}(i) = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1} \left\{ \mathbf{y}(j)\mathbf{y}^T(j) \right\}$ in the multivariate case $PD > 1$. The index m denotes the block time index for a block of N output samples shifted by N_L samples relatively to the previous block. Furthermore, D is the memory length, i.e., the number of time-lags to model the nonwhiteness of the P signals as above. β is a window function with finite support that is normalized so that $\sum_{i=0}^{m} \beta(i,m) = 1$, allowing for online, offline, and block-online algorithms [3, 15].

10.4.3 Gradient-based Coefficient Update

In this chapter we concentrate on iterative gradient-based block-online coefficient updates, which can be written in the general form

$$\check{\mathbf{W}}^0(m) := \check{\mathbf{W}}(m-1), \tag{10.42a}$$

$$\check{\mathbf{W}}^\ell(m) = \check{\mathbf{W}}^{\ell-1}(m) - \mu\Delta\check{\mathbf{W}}^\ell(m), \ \ell = 1,\ldots,\ell_{\max}, \tag{10.42b}$$

$$\check{\mathbf{W}}(m) := \check{\mathbf{W}}^{\ell_{\max}}(m), \tag{10.42c}$$

where μ is a step-size parameter, and the superscript index ℓ denotes an iteration parameter to allow for multiple iterations ($\ell = 1,\ldots,\ell_{\max}$) within each block m. The $LP \times P$ coefficient matrix $\check{\mathbf{W}}$ (defined in (10.4)) to be optimized is smaller than the $2LP \times DP$ Sylvester matrix $\mathbf{W}$ used above for the formulation of the cost function, and it contains only the non-redundant elements of $\mathbf{W}$.

Obviously, when calculating the gradient of $\mathcal{J}(m,\mathbf{W})$ with respect to $\check{\mathbf{W}}$ explicitly, we are confronted with the problem of the different matrix formulations $\mathbf{W}$ and $\check{\mathbf{W}}$. The larger dimensions of $\mathbf{W}$ (see, e.g., (10.38)) are a direct consequence of taking into account the nonwhiteness signal property by choosing $D > 1$. As noted above, the rigorous distinction between these different matrix structures is an essential aspect of the general TRINICON framework and leads to an important building block whose actual implementation is fundamental to the properties of the resulting algorithm, the so-called *Sylvester constraint* ($\mathcal{SC}$) on the coefficient update, formally introduced in [15, 17]. Using the Sylvester constraint operator the gradient descent update can be written as

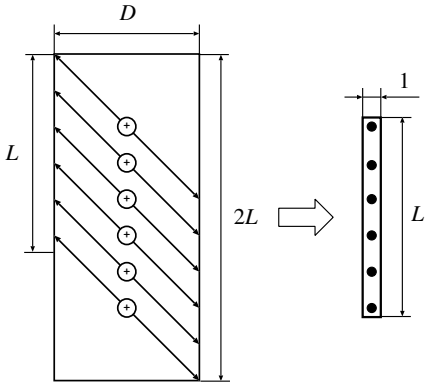


Fig. 10.6 Illustration of the generic Sylvester constraint ($\mathcal{SC}$), after [19] for one channel

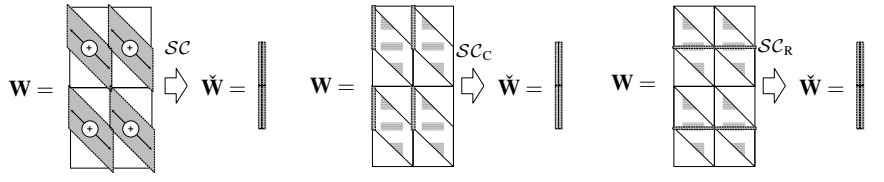
$$\Delta\check{\mathbf{W}}^{\ell}(m) = \mathcal{SC}\left\{\nabla_{\mathbf{W}}\mathcal{J}(m,\mathbf{W})\right\}|_{\mathbf{W}=\mathbf{W}^{\ell}(m)}\,. \tag{10.43}$$

Depending on the particular realization of ($\mathcal{SC}$), we are able to select both, well-known and novel improved adaptation algorithms [3]. As discussed in [3] there are two particularly simple and popular realizations of ($\mathcal{SC}$) leading to two different classes of algorithms (see Fig. 10.7):

1. Computing only the *first column* of each channel of the update matrix to obtain the new coefficient matrix $\check{\mathbf{W}}$. This method is denoted as ($\mathcal{SC}_{\mathrm{C}}$).
2. Computing only the L^{th} *row* of each channel of the update matrix to obtain the new coefficient matrix $\check{\mathbf{W}}$. This method is denoted as ($\mathcal{SC}_{\mathrm{R}}$).

It can be shown that in both cases the update process is significantly simplified [3]. However, in general, both choices require some tradeoff regarding algorithm performance. While $\mathcal{SC}_{\mathrm{C}}$ may provide a potentially more robust convergence behavior, it will not work for arbitrary source positions, which is in contrast to the more versatile $\mathcal{SC}_{\mathrm{R}}$ [3]. Specifically, $\mathcal{SC}_{\mathrm{C}}$ allows us to adapt only *causal* demixing systems. In geometrical terms this means that in the case of separating two sources using $\mathcal{SC}_{\mathrm{C}}$, they are required to be located in different half-planes with respect to the orientation of the microphone array [3]. For separating sources located in the same half-plane, or for more than two sources, noncausal demixing filters are required. With $\mathcal{SC}_{\mathrm{R}}$ it is possible to initialize $\check{\mathbf{W}}_{pp}$, $p = 1,\ldots,P$ with *shifted* unit impulses to allow noncausal filter taps [3]. Since acoustic scenarios exhibit nonminimum phase impulse responses, the need for noncausal demixing filters is further amplified in the dereverberation application.

In [19] an explicit formulation of a *generic* Sylvester constraint was derived to further formalize and clarify this concept, and to combine the versatility of $\mathcal{SC}_{\mathrm{R}}$ with the robust performance of $\mathcal{SC}_{\mathrm{C}}$ [20]. It turns out that the generic Sylvester constraint corresponds – up to the constant D denoting the width of the submatrices – to a *channel-wise arithmetic averaging* of elements according to Fig. 10.6.



(a) (b) (c)

Fig. 10.7 Illustration of two efficient approximations of (a) the generic Sylvester constraint $\mathcal{SC}$, (b) the column Sylvester constraint $\mathcal{SC}_{\mathrm{C}}$, and (c) the row Sylvester constraint $\mathcal{SC}_{\mathrm{R}}$

Note that the previously introduced approaches, classified by the choice of ($\mathcal{SC}_{\mathrm{C}}$) or ($\mathcal{SC}_{\mathrm{R}}$) as mentioned above, thus correspond to approximations of ($\mathcal{SC}$) by neglecting most of the elements within this averaging process, as illustrated in Fig. 10.7. In Sect. 10.6, we will see that by choosing the different Sylvester constraints, we are also able to establish relations to various known multichannel blind deconvolution algorithms from the literature.

It can be shown (see Appendix A) that by taking the gradient of $\mathcal{J}(m)$ with respect to the demixing filter matrix $\breve{\mathbf{W}}(m)$ according to (10.43), we obtain the following generic gradient descent-based TRINICON update rule:

$$\Delta\breve{\mathbf{W}}^{\ell}(m) = \frac{1}{N}\sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\mathbf{x}(j)\left[\boldsymbol{\Phi}_{s,PD}^{\mathrm{T}}(\mathbf{y}(j)) - \boldsymbol{\Phi}_{y,PD}^{\mathrm{T}}(\mathbf{y}(j))\right]\right\}, \tag{10.44a}$$

with the *desired* generalized score function

$$\boldsymbol{\Phi}_{s,PD}(\mathbf{y}(j)) = -\frac{\partial \log \hat{p}_{s,PD}(\mathbf{y}(j))}{\partial \mathbf{y}(j)} - \frac{1}{N}\sum_{r}\sum_{i_1,i_2,\ldots}\frac{\partial \mathcal{G}_{s,i_1,i_2,\ldots}^{(r)}}{\partial \mathbf{y}}\sum_{j=iN_L}^{iN_L+N-1}\frac{\partial \hat{p}_{s,PD}}{\partial \mathcal{Q}_{s,i_1,i_2,\ldots}^{(r)}}, \tag{10.44b}$$

resulting from the hypothesized source model $\hat{p}_{s,PD}$, and the actual generalized score function

$$\boldsymbol{\Phi}_{y,PD}(\mathbf{y}(j)) = -\frac{\partial \log \hat{p}_{y,PD}(\mathbf{y}(j))}{\partial \mathbf{y}(j)} - \frac{1}{N}\sum_{r}\sum_{i_1,i_2,\ldots}\frac{\partial \mathcal{G}_{y,i_1,i_2,\ldots}^{(r)}}{\partial \mathbf{y}}\sum_{j=iN_L}^{iN_L+N-1}\frac{\partial \hat{p}_{y,PD}}{\partial \mathcal{Q}_{y,i_1,i_2,\ldots}^{(r)}}, \tag{10.44c}$$

where the stochastic model parameters are given by (10.41), and $\mathcal{G}_{s,i_1,i_2,\ldots}^{(r)}$, $\mathcal{G}_{y,i_1,i_2,\ldots}^{(r)}$, $\mathcal{Q}_{s,i_1,i_2,\ldots}^{(r)}$, and $\mathcal{Q}_{y,i_1,i_2,\ldots}^{(r)}$ are the elements of $\boldsymbol{\mathcal{G}}_s^{(r)}$, $\boldsymbol{\mathcal{G}}_y^{(r)}$, $\boldsymbol{\mathcal{Q}}_s^{(r)}$, and $\boldsymbol{\mathcal{Q}}_y^{(r)}$, respectively, as explained below (10.41). The form of the coefficient update (10.44a) with the

generalized score functions (10.44b) and (10.44c) also fits well into the theory of so-called estimating functions [5].

The hypothesized source model $\hat{p}_{s,PD}(\cdot)$ in (10.44b) is chosen according to the class of signal processing problem to be solved (see Table 10.1). For instance, a factorization of $\hat{p}_{s,PD}(\cdot)$ among the sources yields BSS (or BSI via the scheme described in Sect. 10.3.4), i.e.,

$$\hat{p}_{s,PD}(\mathbf{y}(j)) \overset{\text{(BSS)}}{=} \prod_{q=1}^{P} \hat{p}_{y_q,D}(\mathbf{y}_q(j)), \tag{10.45a}$$

while a complete factorization leads to the traditional MCBD approach,

$$\hat{p}_{s,PD}(\mathbf{y}(j)) \overset{\text{(MCBD)}}{=} \prod_{q=1}^{P} \prod_{d=1}^{D} \hat{p}_{y_q,1}(y_q(j-d+1)). \tag{10.45b}$$

Additionally, in Sect. 10.6 we will introduce another, more general class, called the MultiChannel Blind Partial Deconvolution (MCBPD) approach.

10.4.3.1 Alternative Formulation of the Gradient-based Coefficient Update

Both for practical realizations and also for some theoretical considerations, an equivalent reformulation of the gradient-based update (10.44a) is often useful. This alternative formulation is obtained by transforming the output signal PDF $\hat{p}_{y,PD}(\mathbf{y})$ in the cost function into the PD-dimensional input signal PDF using $\mathbf{W}$ as a mapping matrix for this linear transformation. The relation (10.134) in Appendix B shows this PDF transformation. (Note that the result of Appendix B is needed again later in this chapter.) Gradient calculation as above leads to the alternative formulation of the gradient-based update,

$$\Delta\check{\mathbf{W}}^{\ell}(m) = \frac{1}{N}\sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\left[\mathbf{x}(j)\boldsymbol{\Phi}_{s,PD}^{\mathrm{T}}(\mathbf{y}(j)) - \mathbf{V}\left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\mathbf{V}\right)^{-1}\right]\right\}, \tag{10.46a}$$

with the window matrix

$$\mathbf{V} = \mathrm{Bdiag}\{\tilde{\mathbf{V}},\ldots,\tilde{\mathbf{V}}\}, \tag{10.46b}$$

$$\tilde{\mathbf{V}} = \left[\mathbf{I}_{D\times D},\, \mathbf{0}_{D\times(2L-D)}\right]^{\mathrm{T}}. \tag{10.46c}$$

10.4.4 Natural Gradient-based Coefficient Update

It is known that stochastic gradient descent generally suffers from slow convergence in many practical problems due to statistical dependencies in the data being processed. A modification of the ordinary gradient, which is especially popular in the field of ICA and BSS due to its computational efficiency, is the so-called *natural gradient* [51]. It can be shown that by taking the natural gradient of $\mathcal{J}(m)$ with respect to the demixing filter matrix $\mathbf{W}(m)$ [17],

$$\Delta\check{\mathbf{W}} \propto \mathcal{SC}\left\{\mathbf{W}\mathbf{W}^{\mathrm{T}}\frac{\partial\mathcal{J}}{\partial\mathbf{W}}\right\}, \tag{10.47}$$

we obtain the following generic TRINICON-based update rule:

$$\Delta\check{\mathbf{W}}^{\ell}(m) = \frac{1}{N}\sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\mathbf{W}^{\ell}(i)\mathbf{y}(j)\left[\boldsymbol{\Phi}^{\mathrm{T}}_{s,PD}(\mathbf{y}(j)) - \boldsymbol{\Phi}^{\mathrm{T}}_{y,PD}(\mathbf{y}(j))\right]\right\}. \tag{10.48}$$

Moreover, from (10.46a) we obtain an alternative formulation of (10.48):

$$\Delta\check{\mathbf{W}}^{\ell}(m) = \sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\mathbf{W}^{\ell}(i)\left[\frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\mathbf{y}(j)\boldsymbol{\Phi}^{\mathrm{T}}_{s,PD}(\mathbf{y}(j)) - \mathbf{I}\right]\right\}, \tag{10.49}$$

which exhibits an especially simple – and thus computationally efficient – structure. An important feature of this natural gradient update is that its adaptation performance is largely independent of the conditioning of the acoustic mixing system matrix [17].

10.4.5 Incorporation of Stochastic Source Models

The general update equations (10.42) with (10.44), (10.46), (10.48) and (10.49) offer the possibility to account for all the available information on the statistical properties of the desired source signals. To apply this general approach in a real-world scenario, appropriate multivariate score functions $\boldsymbol{\Phi}^{\mathrm{T}}_{s,PD}(\mathbf{y})$ (and $\boldsymbol{\Phi}^{\mathrm{T}}_{y,PD}(\mathbf{y})$ where required) in the update equations have to be determined, based on appropriate multivarate stochastic signal models.

The selection of the stochastic signal models is based on several different considerations. As already illustrated by (10.45a) and (10.45b), the design of the signal model is instrumental in defining the class of the adaptive filtering problem according to Table 10.1. This aspect will be detailed in Sects. 10.5 and 10.6. Another

important aspect is that many of the different adaptation techniques in the literature represent different approximations of the probability density functions.

For estimating PDFs a distinction between parametric and non-parametric techniques is common (see, e.g., [33]).

A *parametric* technique defines a family of density functions in terms of a set of parameters as in (10.40a) and (10.40b). The parameters are then optimized so that the density function corresponds to the observed samples. In the context of ICA different parametric representations have been used. Examples include Gaussian models in the simplest case, Gaussian mixture models, and generalized Gaussian models. The important class of spherically-invariant random processes, as detailed below, may also be understood as a parametric approach. Other parametric techniques are based on higher moments [56], e.g., Gram–Charlier expansion, Parson densities, or on higher cumulants [56], e.g., the Edgeworth expansion. As an important representative of these techniques, we consider the Gram–Charlier expansion for TRINICON, as detailed below.

The *non-parametric* techniques usually define the estimated density directly in terms of the observed samples. The best known non-parametric estimate is the histogram, which is very data intensive. Somewhat less data is required by the Parzen windows method [33]. Note that sometimes the above-mentioned techniques based on series with higher moments are also classified as non-parametric in the literature [56]. Obviously, the incorporation of various assumptions about the densities by truncating these series expansions in practice provides a smooth transition to powerful parametric techniques that require less data than the simpler non-parametric techniques.

Another important aspect in the choice of stochastic models is their robustness. According to [50], robustness denotes insensitivity to a certain amount of deviations from the statistical modeling assumptions due to some fraction of outliers with some arbitrary probability density. Unfortunately, many of the traditional estimation techniques, such as least-squares estimation, or the higher-order techniques mentioned above turn out to be fairly sensitive in this sense. The theory of *robust statistics* [50] provides a systematic framework to robustify the various techniques and it has been very successfully applied to adaptive filtering, e.g., [39]. In [23] the theory of multivariate robust statistics was introduced in TRINICON. Although we will not consider the robustness extensions in detail in this chapter, it is important to note that they fit well into the general class of spherically-invariant random processes detailed below.

Finally, it should be noted that in addition to the model selection the choice of estimation procedure for the corresponding *stochastic model parameters* (e.g., correlation matrices in (10.50) below, higher-order moments, scaling parameter for robust statistics in [23], etc.), in other words, the practical realization of (10.41), is another important design consideration. The estimation of the stochastic model parameters and the TRINICON-based updates of the adaptive filter coefficients are performed in an alternating way.

Similar to the estimation of correlation matrices in linear prediction problems [66] in actual implementations we have to distinguish between the more accurate

so-called *covariance method* and the approximative *correlation method* leading to a lower complexity, e.g., [3]. As we will see later in this chapter, based on these different estimation methods for the correlation matrices and on the above-mentioned approximations $\mathcal{SC}_{\mathrm{R}}\{\cdot\}$ and $\mathcal{SC}_{\mathrm{C}}\{\cdot\}$ of the Sylvester constraint $\mathcal{SC}\{\cdot\}$ we can establish an illustrative classification scheme for BSI and deconvolution algorithms.

10.4.5.1 Spherically Invariant Random Processes as Signal Model

An efficient and fairly general solution to the problem of determining the high-dimensional score functions in broadband adaptive MIMO filtering is to assume so-called *spherically invariant random processes* (SIRPs), e.g., [11, 42, 85], as proposed in [13, 15]. The general form of correlated SIRPs of D^{th} order is given with a properly chosen function $f_{p,D}(\cdot)$ for the p^{th} output channel of the MIMO system by

$$\hat{p}_{y_p,D}(\mathbf{y}_p(j)) = \frac{1}{\sqrt{\pi^D \det(\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}(i))}} f_{p,D}\left(\mathbf{y}_p^T(j)\mathbf{R}^{-1}_{\mathbf{y}_p\mathbf{y}_p}(i)\mathbf{y}_p(j)\right), \tag{10.50}$$

where $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}$ denotes the corresponding $D \times D$ autocorrelation matrix with the corresponding number of lags. These models are representative for a wide class of stochastic processes. Speech signals in particular can be represented by SIRPs very accurately [11]. A major advantage arising from the SIRP model is that multivariate PDFs can be derived analytically from the corresponding univariate PDF together with the (lagged) correlation matrices. The function $f_{p,D}(\cdot)$ can thus be calculated from the well-known univariate models for speech, e.g., the Laplacian density. Using the chain rule, the corresponding score function, e.g., (10.44b) can be derived from (10.50), as detailed in [13, 15].

To calculate the score function for SIRPs in general, we employ the chain rule to (10.50) so that the first term in (10.44b) reads

$$-\frac{\partial \log \hat{p}_{y_p,D}(\mathbf{y}_p)}{\partial \mathbf{y}_p} = -\frac{\frac{\partial \hat{p}_{y_p,D}(\mathbf{y}_p)}{\partial \mathbf{y}_p}}{\hat{p}_{y_p,D}(\mathbf{y}_p)} = 2\underbrace{\left[-\frac{1}{f_{p,D}(u_p)}\frac{\partial f_{p,D}(u_p)}{\partial u_p}\right]}_{:=\phi_{y_p,D}(u_p)}\mathbf{R}^{-1}_{\mathbf{y}_p\mathbf{y}_p}(i)\mathbf{y}_p(j), \tag{10.51}$$

where $u_p = \mathbf{y}_p^T\mathbf{R}^{-1}_{\mathbf{y}_p\mathbf{y}_p}\mathbf{y}_p$. For convenience, we call the scalar function $\phi_{y_p,D}(u_p)$ the *SIRP score*. It can be shown (after a somewhat tedious but straightforward derivation) that for SIRPs in general, the second term in (10.44b) is equal to zero so that the general score function is given by the simple expression (10.51). A great advantage of SIRPs is that the required function $f_D(u)$ can actually be derived analytically from the corresponding *univariate* PDF [11]. As a practical important example, following the procedure in [11], we obtain, e.g., as the *optimum SIRP score for univariate Laplacian PDFs* [13]:

$$\phi_{y_q,D}(u_q) = -\frac{1}{D - \sqrt{2u_q}\frac{K_{D/2+1}(\sqrt{2u_q})}{K_{D/2}(\sqrt{2u_q})}}, \tag{10.52}$$

where $K_\nu(\cdot)$ denotes the ν^{th} order modified Bessel function of the second kind.

10.4.5.2 Multivariate Gaussians as Signal Model: Second-order Statistics

To see the link to adaptation algorithms that are based purely on second-order statistics (SOS), we use the model of *multivariate Gaussian* PDFs

$$\hat{p}_{y_p,D}(\mathbf{y}_p(j)) = \frac{1}{\sqrt{(2\pi)^D \det \mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}(i)}} \mathrm{e}^{-\frac{1}{2}\mathbf{y}_p^T(j)\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}^{-1}(i)\mathbf{y}_p(j)} \tag{10.53}$$

as a special case of a SIRP with $f_{q,D}(u_q) = \frac{1}{\sqrt{2^D}} \exp(-\frac{1}{2}u_q)$. Hence, the score function for the generic SOS case is obtained straightforwardly from (10.51) for the constant SIRP score $\phi_{y_p,D}(u_p) = 1/2$, and it can be shown that most of the popular SOS-based adaptation algorithms represent special cases of the corresponding algorithms based on SIRPs, e.g., [13, 15, 16, 23]. Moreover, by transforming the model into the DFT domain, this relation also carries over to various links to novel and existing popular frequency-domain algorithms [15, 19].

It is interesting to note that the generic SOS-based update was originally obtained independently in [17] (first for the BSS application) as a generalization of the cost function of [55]:

$$\mathcal{J}_{\text{SOS}}(m, \mathbf{W}) = \sum_{i=0}^{\infty} \beta(i,m) \left\{ \log\det \mathbf{R}_{\mathbf{ss}}(i) - \log\det \mathbf{R}_{\mathbf{yy}}(i) \right\}. \tag{10.54}$$

This cost function can be interpreted as a distance measure between the actual time-varying output-correlation matrix $\mathbf{R}_{\mathbf{yy}}$ and a certain *desired* output-correlation matrix $\mathbf{R}_{\mathbf{ss}}$.

10.4.5.3 Nearly Gaussian Densities as Signal Model

Two different expansions are commonly used to obtain a parameterized representation of probability density functions that only slightly deviate from the Gaussian density (often called *nearly Gaussian densities*): the Edgeworth and the Gram–Charlier expansions, e.g., [51]. They lead to very similar approximations, so here we only consider the Gram–Charlier expansion. As explained in Appendix C, these expansions are based on the so-called Chebyshev–Hermite polynomials $P_{\text{H},n}(\cdot)$.

We first illustrate the idea in the univariate case. A fourth-order expansion of a univariate, zero-mean, and nearly Gaussian PDF is given in (10.140) in Appendix C with the estimates of *skewness* $\hat{\kappa}_3 = \hat{E}\left\{y^3\right\}$ and the *kurtosis* $\hat{\kappa}_4 = \hat{E}\left\{y^4\right\} - 3\hat{\sigma}^4$,

the latter one being the most important higher-order statistical quantity in our context. Generally, speech signals exhibit supergaussian densities whose third-order cumulants are negligible compared to its fourth-order cumulants. Therefore, we are particularly interested in the approximation

$$\hat{p}(y) \approx \frac{1}{\sqrt{2\pi}\hat{\sigma}} \mathrm{e}^{-\frac{y^2}{2\hat{\sigma}^2}} \left(1 + \frac{\hat{\kappa}_4}{4!\,\hat{\sigma}^4} P_{\mathrm{H},4}\left(\frac{y}{\hat{\sigma}}\right)\right). \tag{10.55}$$

Similar to the specialization (10.54) of the TRINICON optimization criterion for the case of SOS, the Gram–Charlier-based model also allows an interesting illustration of the criterion. By exploiting the near-gaussianity by the approximation $\log(1+\varepsilon) \approx \varepsilon$ for $\log\left(1 + \frac{\hat{\kappa}_4}{4!\,\hat{\sigma}^4} P_{\mathrm{H},4}\left(\frac{y}{\hat{\sigma}}\right)\right)$ in the logarithmized respresentation of (10.55), and noting that $P_{\mathrm{H},4}\left(\frac{y}{\hat{\sigma}}\right) = \left(\frac{y}{\hat{\sigma}}\right)^4 - 6\left(\frac{y}{\hat{\sigma}}\right)^2 + 3$ we can develop the following expression appearing in the TRINICON criterion (10.39):

$$\begin{aligned}
&\frac{1}{N} \sum_{j=iN_L}^{iN_L+N-1} \log \hat{p}(y) \\
&\qquad \approx \frac{1}{N} \left(\sum_{j=iN_L}^{iN_L+N-1} \log \frac{1}{\sqrt{2\pi}\hat{\sigma}} \mathrm{e}^{-\frac{y^2}{2\hat{\sigma}^2}} \right) + \frac{1}{N} \left(\sum_{j=iN_L}^{iN_L+N-1} \frac{\hat{\kappa}_4}{4!\,\hat{\sigma}^4} P_{\mathrm{H},4}\left(\frac{y}{\hat{\sigma}}\right) \right) \\
&\qquad = \frac{1}{N} \left(\sum_{j=iN_L}^{iN_L+N-1} \log \frac{1}{\sqrt{2\pi}\hat{\sigma}} \mathrm{e}^{-\frac{y^2}{2\hat{\sigma}^2}} \right) + \frac{\hat{\kappa}_4^2}{4!\,(\hat{\sigma}^2)^4},
\end{aligned} \tag{10.56}$$

where $\hat{\kappa}_4 = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1} y^4 - 3\hat{\sigma}^4$ represents an estimate for the kurtosis based on block averaging. As we can see, in addition to the SOS, the optimization is directly based on the normalized kurtosis, which is a widely-used measure of *nongaussianity*. This additive representation will play a particularly important role in the application to the direct-inverse approach to blind dereverberation in Sect. 10.6.

To obtain general coefficient update rules based on this representation, we finally consider the multivariate formulation of the Gram–Charlier expansion after (10.146a) in Appendix C. To calculate the multivariate Chebyshev–Hermite polynomials, we apply the relation

$$P_{\mathrm{H},\mathbf{n}}(\mathbf{y}_p) = \prod_{d=1}^{D} P_{\mathrm{H},n_d}(y_{d,p}) \tag{10.57}$$

from (10.144) so that

$$\begin{aligned}
\hat{p}_{y_p,D}(\mathbf{y}_p(j)) &= \frac{1}{\sqrt{(2\pi)^D \det \mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}(i)}} \mathrm{e}^{-\frac{1}{2}\mathbf{y}_p^T(j)\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}^{-1}(i)\mathbf{y}_p(j)} \\
\times \sum_{n_1=0}^{\infty} \cdots \sum_{n_D=0}^{\infty} & a_{n_1\ldots n_D,p}\, P_{\mathrm{H},n_1}\left(\left[\mathbf{L}_p^{-1}(i)\mathbf{y}_p(j)\right]_1\right) \ldots P_{\mathrm{H},n_D}\left(\left[\mathbf{L}_p^{-1}(i)\mathbf{y}_p(j)\right]_D\right),
\end{aligned}$$

with the coefficients according to (10.146b),

$$a_{n_1\ldots n_D,p} = \frac{\hat{E}\left\{P_{\mathrm{H},n_1}\left(\left[\mathbf{L}_p^{-1}(i)\mathbf{y}_p(j)\right]_1\right)\ldots P_{\mathrm{H},n_D}\left(\left[\mathbf{L}_p^{-1}(i)\mathbf{y}_p(j)\right]_D\right)\right\}}{n_1!\ldots n_D!}. \quad (10.58)$$

Multivariate generalizations of the skewness and the kurtosis were introduced by Mardia in [65]. In our context the corresponding multivariate generalization of the kurtosis can be written as

$$\hat{\kappa}_{4,\mathrm{norm}}^{(D)} = \hat{E}\left\{\left[\mathbf{y}_p^T(j)\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}^{-1}(i)\mathbf{y}_p(j)\right]^2\right\} - D(D+2). \quad (10.59)$$

Similar to the univariate case, this quantity can be related to our formulation of the multivariate probability density. Note that for $D=1$ it corresponds to the traditional normalized kurtosis $\hat{\kappa}_4/\hat{\sigma}^4 = \hat{E}\{y_p^4\}/\hat{\sigma}^4 - 3$, as it appears in, e.g., (10.55).

In this chapter, we further consider an important special case of this general multivariate model, which is particularly useful for speech processing. In this case, the inverse covariance matrix $\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}^{-1} = (\mathbf{L}_p^T\mathbf{L}_p)^{-1}$ is first factorized as [62]

$$\mathbf{R}_{\mathbf{y}_p\mathbf{y}_p}^{-1}(i) = \mathbf{A}_p(i)\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}_p\tilde{\mathbf{y}}_p}^{-1}(i)\mathbf{A}_p^T(i), \quad (10.60)$$

where $\mathbf{A}_p(i)$ and $\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}_p\tilde{\mathbf{y}}_p}(i)$ denote a $D\times D$ unit lower triangular matrix (i.e., its elements on the main diagonal are equal to 1) and a diagonal matrix, respectively [62]. The $D\times D$ unit lower triangular matrix $\mathbf{A}_p(i)$ can be interpreted as a (time-varying) convolution matrix of a whitening filter. It is therefore convenient for computational reasons to model the signal y_p as an autoregressive (AR) process of order $n_A = D-1$, with time-varying AR coefficients $a_{p,k}(n)$, and residual signal $\tilde{y}_p(n)$, i.e.,

$$y_p(n) = -\sum_{k=1}^{D-1} a_{p,k}(n)y_p(n-k) + \tilde{y}_p(n). \quad (10.61)$$

The matrices $\mathbf{A}_p$ and $\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}_p\tilde{\mathbf{y}}_p}$ can then be written as

$$\mathbf{A}_p = \begin{bmatrix} 1 & a_{p,1}(n) & a_{p,2}(n) & \cdots & \cdots & \cdots & \cdots & a_{p,D-1}(n) \\ 0 & 1 & a_{p,1}(n-1) & \cdots & \cdots & \cdots & \cdots & a_{p,D-2}(n-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}^T \quad (10.62)$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}_p\tilde{\mathbf{y}}_p} &= \mathrm{Diag}\left\{\hat{\sigma}_{\tilde{y}_p}^2(n),\ldots,\hat{\sigma}_{\tilde{y}_p}^2(n-D+1)\right\} \\ &= \hat{E}\left\{\begin{bmatrix}\tilde{y}_p(n)\\ \vdots \\ \tilde{y}_p(n-D+1)\end{bmatrix}[\tilde{y}_p(n),\ldots,\tilde{y}_p(n-D+1)]\right\}. \end{aligned} \quad (10.63)$$

Now, the multivariate stochastic signal model can be rewritten by shifting the *prefiltering matrix* $\mathbf{A}_p$ into the data terms, i.e.,

$$\tilde{\mathbf{y}}_p := \mathbf{A}_p^T \mathbf{y}_p = [\tilde{y}_p(n), \tilde{y}_p(n-1), \ldots, \tilde{y}_p(n-D+1)]^T . \tag{10.64}$$

Moreover, by assuming the whitened elements of vector $\tilde{\mathbf{y}}_p$ to be i.i.d. (which in practice is a widely used assumption in AR modeling), so that the expansion coefficients $a_{n_1 \cdots n_D, p}$ are factorized, due to (10.57) with $\mathbf{L}_p(i) = \mathrm{Diag}\left\{\frac{1}{\hat{\sigma}_{\tilde{y}_p}(j)}, \ldots, \frac{1}{\hat{\sigma}_{\tilde{y}_p}(j-D+1)}\right\} \mathbf{A}^T(i)$ and (10.64) we obtain the following model representation:

$$\hat{p}_{y_p,D}(\mathbf{y}_p(j)) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\, \hat{\sigma}^2_{\tilde{y}_p}(j-d+1)}} \mathrm{e}^{-\frac{\tilde{y}_p^2(j-d+1)}{2\hat{\sigma}^2_{\tilde{y}_p}(j-d+1)}} \times \sum_{n_d=0}^{\infty} \frac{\hat{E}\left\{P_{\mathrm{H},n_d}\left(\frac{\tilde{y}_p(j-d+1)}{\hat{\sigma}_{\tilde{y}_p}(j-d+1)}\right)\right\}}{n_d!} P_{\mathrm{H},n_d}\left(\frac{\tilde{y}_p(j-d+1)}{\hat{\sigma}_{\tilde{y}_p}(j-d+1)}\right).$$

By considering only the fourth-order term in addition to SOS again, i.e.,

$$\hat{p}_{y_p,D}(\mathbf{y}_p(j)) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\, \hat{\sigma}^2_{\tilde{y}_p}(j-d+1)}} \mathrm{e}^{-\frac{\tilde{y}_p^2(j-d+1)}{2\hat{\sigma}^2_{\tilde{y}_p}(j-d+1)}} \times \left(1 + \frac{\hat{\kappa}_{4,\tilde{y}_p}}{4!\sigma^4_{\tilde{y}_p}(j-d+1)} P_{\mathrm{H},n_d}\left(\frac{\tilde{y}_p(j-d+1)}{\hat{\sigma}_{\tilde{y}_p}(j-d+1)}\right)\right),$$

and by exploiting the near-gaussianity using the approximation $\log(1+\varepsilon) \approx \varepsilon$, after a straightforward calculation we obtain the following expression for the score function (10.44c):

$$\boldsymbol{\Phi}_{y,PD}(\mathbf{y}(j)) = \mathbf{A}(i)\left[\frac{\tilde{y}_p(j-d+1)}{2\hat{\sigma}^2_{\tilde{y}_p}(j-d+1)} - \left(\frac{\sum_{j=iN_L}^{iN_L+N-1} \tilde{y}_p^4(j-d+1)}{3\left(\sum_{j=iN_L}^{iN_L+N-1} \tilde{y}_p^2(j-d+1)\right)^2} - 1\right) \times \left(\frac{\tilde{y}_p^3(j-d+1)}{\hat{\sigma}^4_{\tilde{y}_p}(j-d+1)} - \frac{\tilde{y}_p(j-d+1)\sum_{j=iN_L}^{iN_L+N-1} \tilde{y}_p^4(j-d+1)}{\hat{\sigma}^6_{\tilde{y}_p}(j-d+1)}\right)\right], \tag{10.65}$$

where the expression in brackets denotes a column vector composed of the elements for $d = 1, \ldots, D$ and $p = 1, \ldots, P$, and $\mathbf{A}(i) = [\mathbf{A}_1(i), \ldots, \mathbf{A}_P(i)]$ after (10.62). Note that the first term corresponds to the SOS as in (10.51), while the second term is related to the multivariate normalized kurtosis. This expression will play an important role in Sect. 10.6.

10.5 Application of TRINICON to Blind System Identification and the Identification-and-inversion Approach to Blind Deconvolution

In Sect. 10.3 we developed the identification-and-inversion approach to blind deconvolution from a system-theoretic point of view. We have seen that in the general MIMO case its practical (i.e., adaptive) realization can be traced back to the problem of blind source separation for convolutive mixtures with appropriately chosen filter length L and subsequent inversion, e.g., using MINT (Fig. 10.5). Both signal separation and system identification belong to the class of direct adaptive filtering problems according to Table 10.1. On the other hand, it was shown that in the SIMO case this approach leads to a well-known class of realizations for which the AED algorithm in its various versions is known from the literature. Hence, as the two main aspects in this section

- We discuss the specialization of the TRINICON framework to practical algorithms that are suitable for adaptive MIMO BSI. Various different BSS algorithms have been proposed in recent years (e.g., [64]), and many of them can be related to TRINICON [15, 19]. However, of special importance for BSI and the identification-and-inversion approach to dereverberation are efficient realizations of *broadband* BSS algorithms.
- We develop the relation to the SIMO case explicitly from an algorithmic point of view. This will lead to various new insights and also to some generalizations of the AED.

Both of these main aspects will also serve as important starting points for the developments in Sect. 10.6. An experimental comparison of the identification-and-inversion approach with the direct-inverse approach to blind dereverberation also follows in Sect. 10.6.

10.5.1 Generic Gradient-based Algorithm for Direct Adaptive Filtering Problems

To begin with, we specialize TRINICON to the case of direct adaptive filtering problems, i.e., signal separation and system identification. Again, for simplicity of the presentation, we concentrate here on iterative Euclidean gradient-based and natural gradient-based block-online coefficient updates. As mentioned in Sect. 10.4, the class of signal separation and system identification algorithms is specified by the factorization of the hypothesized source model $\hat{p}_{s,PD}(\cdot)$ among the sources according to (10.45a). The desired multivariate score function then becomes the partitioned vector

$$\boldsymbol{\Phi}_{s,PD}(\mathbf{y}(j)) = \left[\boldsymbol{\Phi}^{\mathrm{T}}_{y_1,D}(\mathbf{y}_1(j)),\ldots,\boldsymbol{\Phi}^{\mathrm{T}}_{y_P,D}(\mathbf{y}_P(j))\right]^{\mathrm{T}}, \tag{10.66a}$$

$$\boldsymbol{\Phi}_{y_p,D}(\mathbf{y}_p(j)) = -\frac{\partial \log \hat{p}_{y_p,D}(\mathbf{y}_p(j))}{\partial \mathbf{y}_p(j)}. \tag{10.66b}$$

The corresponding generic coefficient update rules are then directly given by (10.44a), (10.46a), (10.48), and (10.49).

In this section, our considerations are based on the SIRP model (including SOS as a special case). Accordingly, each partition of the vector (10.66a) is given by (10.51). The resulting general class of broadband BSS algorithms was first presented in [13] and has led to various efficient realizations so far (see Sect. 10.5.3). The idea of using a SIRP model was also adopted, e.g., in the approximate DFT-domain realizations [47, 57].

10.5.1.1 Illustration for Second-order Statistics

By setting the SIRP scores $\phi_{y_p,D}(\cdot) = 1/2$, $p = 1,\ldots,P$, we obtain the particularly illustrative case of SOS-based adaptation algorithms. Here, the source models are simplified to multivariate Gaussian functions described by $PD \times PD$ correlation matrices $\mathbf{R}$ estimated from the length N signal blocks, so that the update rules (10.44a) and (10.48) lead to [16]

$$\Delta\check{\mathbf{W}}(m) = \sum_{i=0}^{\infty} \beta(i,m)\, \mathcal{SC}\left\{\mathbf{R}_{\mathbf{xy}}(i)\left[\mathbf{R}_{\mathbf{ss}}^{-1}(i) - \mathbf{R}_{\mathbf{yy}}^{-1}(i)\right]\right\} \tag{10.67}$$

and

$$\begin{aligned}\Delta\check{\mathbf{W}}(m) &= \sum_{i=0}^{\infty} \beta(i,m)\, \mathcal{SC}\left\{\mathbf{W}(i)\mathbf{R}_{\mathbf{yy}}(i)\left[\mathbf{R}_{\mathbf{ss}}^{-1}(i) - \mathbf{R}_{\mathbf{yy}}^{-1}(i)\right]\right\} \\ &= \sum_{i=0}^{\infty} \beta(i,m)\, \mathcal{SC}\left\{\mathbf{W}(i)\left[\mathbf{R}_{\mathbf{yy}}(i) - \mathbf{R}_{\mathbf{ss}}(i)\right]\mathbf{R}_{\mathbf{ss}}^{-1}(i)\right\},\end{aligned} \tag{10.68}$$

respectively. The BSS versions of these generic SOS natural gradient updates follow immediately by setting

$$\mathbf{R}_{\mathbf{ss}}(i) = \mathrm{bdiag}\,\mathbf{R}_{\mathbf{yy}}(i). \tag{10.69}$$

The update (10.68) together with (10.69) was originally obtained independently in [17] from the cost function (10.54). The mechanism of (10.68) based on the model (10.69) is illustrated in Fig. 10.8. By minimizing $\mathcal{J}_{\mathrm{SOS}}(m)$, all cross-correlations for D time-lags are reduced and will ideally vanish, while the auto-correlations are untouched to preserve the structure of the individual signals.

A very important feature of the TRINICON-based coefficient updates is the inherent normalization by the auto-correlation matrices, reflected by the inverse of $\mathbf{R}_{\mathbf{ss}}(i) = \mathrm{bdiag}\,\mathbf{R}_{\mathbf{yy}}(i)$ in (10.68). As we will see in Sect. 10.5.2, this normalization can in fact be interpreted as an *adaptive step-size control*. In fact, as was shown

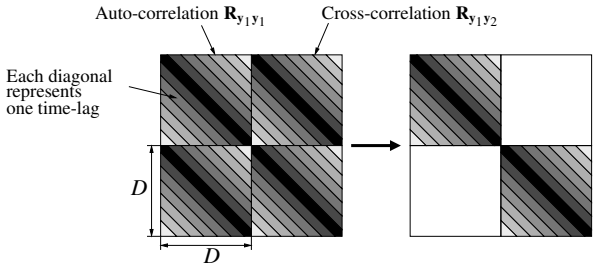


Fig. 10.8 Illustration of SOS-based broadband BSS

in [15], the update equations of another very popular subclass of second-order BSS algorithms, based on a cost function using the Frobenius norm[7] $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_{i,j} a_{ij}^2$ of a matrix $\mathbf{A} = (a_{ij})$, e.g., [26, 51, 52, 69, 74, 79], differ from the more general TRINICON-based updates mainly in the inherent normalization. The gradient-based update resulting from the Frobenius norm can be regarded as an analogon to the traditional Least Mean Square (LMS) algorithm [45] in supervised adaptive filtering without step-size control. Indeed, many simulation results have shown that for large filter lengths L, these Frobenius-based updates are prone to instability, while the properly normalized updates show a very robust convergence behaviour even for hundreds or thousands of filter coefficients for the application in real acoustic environments, e.g., [17]. As we will see in Sect. 10.6, an analogous consideration concerning the inherent normalization is also possible for dereverberation algorithms of the direct-inverse-type.

The realization of this normalization is also an important aspect in various efficient approximations of generic broadband algorithms, e.g., [2, 3, 72], with a reduced computational complexity for real-time operation. Moreover, a close link has been established [15, 17] to various popular frequency-domain algorithms, as we discuss in more detail in Sect. 10.5.3.

In Sect. 10.5.2 we show that taking into account the nongaussianity (in addition to the SOS) can be regarded as a further improvement of the inherent adaptation control.

10.5.2 Realizations for the SIMO Case

As mentioned in Sect. 10.3.5, most of the existing literature on the identification-and-inversion approach to blind deconvolution is based on the SIMO mixing model, e.g., [9, 36, 38, 43, 46, 49, 77]. Using the TRINICON framework, the approach has been developed rigorously for the more general MIMO case based on first principles.

[7] Analogously to the TRINICON-based $\mathcal{J}_{\mathrm{SOS}}$ this approach may be generalized for convolutive mixtures to $\mathcal{J}_{\mathrm{F}}(m) = \sum_{i=0}^{\infty} \beta(i,m) \|\mathbf{R}_{\mathbf{yy}}(i) - \mathrm{bdiag}\,\mathbf{R}_{\mathbf{yy}}(i)\|_{\mathrm{F}}^2$.

In this section we show how to deduce the class of SIMO-based algorithms from TRINICON. Besides a generalization of these algorithms, this consideration will also serve as an important background for the later developments in Sect. 10.6.

As a starting point, we consider the gradient-based update (10.46a) of the MIMO demixing system $\check{\mathbf{W}}$ with the specialized score function (10.66) for separation and identification problems.

The ideal separation filter matrix $\check{\mathbf{W}}_{\text{ideal,sep}}$ in the 2×2 case is given by (10.25), i.e.,

$$\check{\mathbf{W}}_{\text{ideal,sep}} = \begin{bmatrix} \mathbf{h}_{22} & -\mathbf{h}_{12} \\ -\mathbf{h}_{21} & \mathbf{h}_{11} \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}, \tag{10.70}$$

where due to the scaling ambiguity (in blind problems) each column is multiplied by an unknown scalar α_q. For $L = L_{\text{opt,sep}} = M$, this ideal separation solution corresponds to a MIMO system identification up to an arbitrary scalar constant (independently of the adaptation method and the possible prior knowledge).

We now consider the SIMO mixing model in Fig. 10.3(a) as a specialization of the MIMO mixing model in Fig. 10.3(b), i.e., $\mathbf{h}_{11} \to \mathbf{h}_1, \mathbf{h}_{12} \to \mathbf{h}_2, \mathbf{h}_{21} \to \mathbf{0}, \mathbf{h}_{22} \to \mathbf{0}$.

According to the right-hand side of (10.70) the corresponding ideal *demixing system* taking into account this prior knowledge reads as

$$\begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{21} & \mathbf{w}_{22} \end{bmatrix} = \alpha \begin{bmatrix} \mathbf{0} & -\mathbf{h}_2 \\ \mathbf{0} & \mathbf{h}_1 \end{bmatrix}. \tag{10.71}$$

By comparing both sides of this equation, we immediately obtain the corresponding demixing system structure shown on the right-hand side in Fig. 10.3(a). This is indeed the well-known SIMO BSI/AED approach, which in this way follows rigorously from the general equation (10.70) together with prior knowledge on the specialized mixing system. Moreover, we can see that only the second column of the demixing matrix is relevant for the adaptation process. The elements of the first column can be regarded as *don't cares*.

We now consider the *second term* of the coefficient update (10.46a). From the relation (10.134) in Appendix B it immediately follows that

$$\log \hat{p}_{\mathbf{y},PD}(\mathbf{y}(n)) = \text{const.}\ \forall\, \mathbf{W} \Rightarrow \log \left|\det\left\{\mathbf{V}^{\mathrm{T}}\mathbf{W}\right\}\right| = \text{const.}\ \forall\, \mathbf{W}. \tag{10.72}$$

Specifically, in the case of SOS (e.g., (10.54)) this leads to

$$\log |\det \mathbf{R}_{\mathbf{yy}}| = \text{const.}\ \forall\, \mathbf{W} \Rightarrow \log \left|\det\left\{\mathbf{V}^{\mathrm{T}}\mathbf{W}\right\}\right| = \text{const.}\ \forall\, \mathbf{W}. \tag{10.73}$$

As the second term in the update (10.46a) respresents the *gradient* of the expression $\log \left|\det\left\{\mathbf{V}^{\mathrm{T}}\mathbf{W}\right\}\right|$ with respect to $\mathbf{W}$, we conclude that *the second term in the coefficient update is equal to zero if* $\det \mathbf{R}_{\mathbf{yy}}$ *is independent of* $\mathbf{W}$. We therefore now consider the dependence of $\det \mathbf{R}_{\mathbf{yy}}$ on $\mathbf{W}$ in more detail. Since $\mathbf{R}_{\mathbf{yy}} = \hat{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{W}^T\mathbf{H}^T\mathbf{R}_{\mathbf{ss}}\mathbf{H}\mathbf{W}$, we have

$$\log |\det \mathbf{R}_{\mathbf{yy}}| = \underbrace{\log |\det \mathbf{R}_{\mathbf{ss}}|}_{=\text{const.}\ \forall\, \mathbf{W}} + 2 \log \left|\det\{\mathbf{W}^T\mathbf{H}^T\}\right|. \tag{10.74}$$

Now let $\mathbf{W} = \left[\mathbf{W}_1^T, \ldots, \mathbf{W}_P^T\right]^T$ and $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_P]$ be MISO and SIMO, respectively, as special case of the MIMO definition (10.12). In this special case, the input–output relation of the overall system reads as

$$\mathbf{y} = \mathbf{W}^T\mathbf{H}^T\mathbf{s} = \left(\sum_{p=1}^{P} \mathbf{W}_p^T\mathbf{H}_p^T\right)\mathbf{s}, \tag{10.75}$$

and $\sum_{p=1}^{P} \mathbf{W}_p^T\mathbf{H}_p^T$ represents an upper triangular matrix with diagonal elements $\sum_{p=1}^{P} w_{p,0}h_{p,0}$. Hence, in the SIMO case, (10.74) simplifies to

$$\log|\det\mathbf{R}_{\mathbf{yy}}| = \text{const.} + 2N\log\left|\sum_{p=1}^{P} w_{p,0}h_{p,0}\right|. \tag{10.76}$$

Again, in the special case of only one active source, we can formulate an interesting statement concerning the first taps $w_{p,0}$ of the demixing subfilters. As the demixing subfilters ideally compensate for the individual time-differences of arrival at the microphones, only the subfilter $\mathbf{w}_{p_{\text{far}}}$ connected to the microphone that has the greatest distance to the source, may exhibit a nonzero value at its first tap weight, i.e.,

$$w_{p,0} = \alpha \cdot \delta_{p,p_{\text{far}}}, \tag{10.77}$$

where δ_{ij} denotes the Kronecker symbol. Introducing this property finally leads to

$$\begin{aligned}\log|\det\mathbf{R}_{\mathbf{yy}}| &= \text{const.} + 2N\log\left|\alpha h_{p_{\text{far}},0}\right| \\ &= \text{const.}\end{aligned} \tag{10.78}$$

Hence, together with (10.73), we can draw the conclusion that *in the SIMO case, the second term of the coefficient update (10.46a) disappears without loss of generality*.

Next, we consider the *first term* $\mathbf{x}(j)\boldsymbol{\Phi}_{s,PD}^T(\mathbf{y}(j))$ in the coefficient update (10.46a) for the SIMO case and note that its second (block) column reads as $\mathbf{x}(j)\boldsymbol{\Phi}_{y_2,D}^T(\mathbf{y_2}(j))$. We now perform the following formal substitutions in order to be in accordance with the literature on blind SIMO identification and supervised adaptive filtering, e.g., [45] (see Figs. 10.3(a) and (b)):

$$\mathbf{y}_2 \rightarrow \mathbf{e}, \quad \begin{bmatrix} \mathbf{w}_{12} \\ \mathbf{w}_{22} \end{bmatrix} = \begin{bmatrix} -\hat{\mathbf{h}}_2 \\ \hat{\mathbf{h}}_1 \end{bmatrix} \rightarrow \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}. \tag{10.79}$$

Hence, the second column of the first term of the coefficient update is finally expressed as $\mathbf{x}(j)\boldsymbol{\Phi}_{e,D}^T(\mathbf{e}(j))$. Note that the substitution of the coefficient notation in (10.79) is justified by (10.71).

Thus, we obtain the following sub-matrix of the specialized gradient-based TRINICON update:

$$\mathbf{w}^{\ell}(m) = \mathbf{w}^{\ell-1}(m) + \frac{\mu}{N}\sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\mathbf{x}(j)\boldsymbol{\Phi}_{e,D}^{T}(\mathbf{e}(j))\right\}. \tag{10.80}$$

This formally represents the *triple-N-generalization of the LMS algorithm* from supervised adaptive filtering theory (see also [23]), which in its well-known original form exhibits the simple update [45]

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu\,\check{\mathbf{x}}(n)e(n), \tag{10.81}$$

where the length-L vector $\check{\mathbf{x}}$ is a truncated version of $\mathbf{x}$ (formally, this truncation is obtained by $(\mathcal{SC})$ for $D = 1$, see Fig. 10.6). Although not shown in this chapter, it is possible to analogously derive the corresponding generalizations of other supervised algorithms (NLMS, RLS, etc., which may essentially be seen as special cases of a Newton-type update, e.g., [22]) by choosing a Newton-type TRINICON coefficient update instead of the gradient descent-type update.

From the generalized LMS update (10.80) above we can make the following observations in comparison with the simple case (10.81): Due to the generalized approach, we inherently obtain

- block online adaptation, possibly with multiple iterations ℓ to speed up the convergence [15]
- block averaging by $N > 1$ for a more uniform convergence
- an error nonlinearity to take into account the *nongaussianity* of the signals (by a proper choice of $\boldsymbol{\Phi}_{e,D}^{T}(\cdot)$)
- multivariate error $\mathbf{e}$ to take into account the *nonwhiteness* of the signals (by choosing $D > 1$).

Note that, in various ways, the RLS algorithm can be seen as the optimal supervised adaptation algorithm. However, the RLS is optimum only in the case of a Gaussian source signal and Gaussian additive noise on the microphones, with the noise being additionally stationary and white. The general update resulting from TRINICON does not have these restrictions.

10.5.2.1 Coefficient Initialization

The general relation between MIMO BSI and SIMO BSI also leads to an important guideline for the initialization of the filter coefficients. In particular, we consider the question whether the algorithm can converge to the (undesired) trivial solution $\mathbf{w} = \mathbf{0}$. As we will show, the answer is no, as long as the initialization $\mathbf{w}(0)$ is not orthogonal to the ideal solution $\mathbf{w}_{\text{ideal}} = \left[-\mathbf{h}_2^T\ \mathbf{h}_1^T\right]^T$.

To prove this condition, we pre-multiply the update (10.80) with $\mathbf{w}_{\text{ideal}}^T$ on both sides of the update equation:

$$\mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell}(m) = \mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell-1}(m) + \frac{\mu}{N}\sum_{i=0}^{\infty}\beta(i,m) \times \left[-\mathbf{h}_2^T\ \mathbf{h}_1^T\right]\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\begin{bmatrix}\mathbf{x}_1(j)\\ \mathbf{x}_2(j)\end{bmatrix}\boldsymbol{\Phi}_{e,D}^T(\mathbf{e}(j))\right\}, \tag{10.82}$$

$$\mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell}(m) = \mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell-1}(m) + \frac{\mu}{N}\sum_{i=0}^{\infty}\beta(i,m) \times \sum_{j=iN_L}^{iN_L+N-1}\left(\mathbf{h}_1^T\mathcal{SC}\left\{\mathbf{x}_2(j)\boldsymbol{\Phi}_{e,D}^T(\mathbf{e}(j))\right\} - \mathbf{h}_2^T\mathcal{SC}\left\{\mathbf{x}_1(j)\boldsymbol{\Phi}_{e,D}^T(\mathbf{e}(j))\right\}\right). \tag{10.83}$$

With (10.148) from Appendix D this expression can be expanded to

$$\mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell}(m) = \mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell-1}(m) + \frac{\mu}{N}\sum_{i=0}^{\infty}\beta(i,m) \times \sum_{j=iN_L}^{iN_L+N-1}\sum_{l=1}^{D}\left(\mathbf{h}_1^T\check{\mathbf{x}}_2(j-l+1) - \mathbf{h}_2^T\check{\mathbf{x}}_1(j-l+1)\right)\Phi_{e,l}(\mathbf{e}(j)). \tag{10.84}$$

Since $\mathbf{h}_1^T\check{\mathbf{x}}_2(\cdot) - \mathbf{h}_2^T\check{\mathbf{x}}_1(\cdot) \equiv 0$ is fixed due to the acoustic model, we have $\mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell}(m) = \mathbf{w}_{\text{ideal}}^T\mathbf{w}^{\ell-1}(m) = \text{const.}$, i.e., *provided that* $\mathbf{w}_{\text{ideal}}^T\mathbf{w}(0) \neq 0$, *the coefficient vector* $\mathbf{w}$ *will not converge to zero.*

10.5.2.2 Efficient Implementation of the Sylvester Constraint for the Special Case of SIMO Models

As already explained for the general MIMO case, we also further specialize the generalized LMS update (10.80) by incorporating the SIRP model. Introducing the score function (10.51) immediately leads to a SIRPs-based generalized LMS update analogously to [23]

$$\mathbf{w}^{\ell}(m) = \mathbf{w}^{\ell-1}(m) + \frac{2\mu}{N}\sum_{i=0}^{\infty}\beta(i,m) \times \sum_{j=iN_L}^{iN_L+N-1}\mathcal{SC}\left\{\mathbf{x}(j)\mathbf{e}^T(j)\mathbf{R}_{\mathbf{ee}}^{-1}(i)\right\}\phi_{e,D}\left(\mathbf{e}^T(j)\mathbf{R}_{\mathbf{ee}}^{-1}(i)\mathbf{e}(j)\right). \tag{10.85}$$

As in the general MIMO case, we can see that the SIRP model leads to an inherent normalization by the auto-correlation matrix. Note that the SOS case follows for $\phi_{e,D}(\cdot) = 1/2$. In both the SOS case and for general SIRPs the normalization by the correlation matrix in conjunction with $N > 1$ may be interpreted as an *inherent step-size control*. (It also illustrates why BSS does not require a separate double-talk detector, such as traditional supervised algorithms do, e.g., for acoustic echo cancel-

lation or adaptive beamforming.) Moreover, in [23] it was shown that for a suitable choice of parameters, the general SIRP-based update (10.85) can be interpreted as a multivariate, i.e., *triple-N generalization of the robust LMS algorithm* based on robust statistics [50], as mentioned in Sect. 10.4.5.

To further simplify the realization, we next study the expression

$$\mathcal{SC}\left\{\mathbf{x}(j)\mathbf{e}^T(j)\mathbf{R}_{\mathbf{ee}}^{-1}(i)\right\} \tag{10.86}$$

appearing in (10.85). According to the structure of the generic Sylvester constraint in Fig. 10.6 and [19] (see also Appendix D), the l^{th} element of the p^{th} subvector (contributing to the p^{th} channel impulse response) can be expanded to

$$\sum_{d=1}^{D}\left[\mathbf{x}_p(j)\right]_{l+d-1}\left[\mathbf{R}_{\mathbf{ee}}^{-1}(i)\mathbf{e}(j)\right]_d = \check{\mathbf{x}}_{p,D}^T(j-l+1)\mathbf{R}_{\mathbf{ee}}^{-1}(i)\mathbf{e}(j), \tag{10.87}$$

where $\check{\mathbf{x}}_{p,D}$ denotes the length-D vector

$$\check{\mathbf{x}}_{p,D}(n) = \left[x_p(n), x_p(n-1), \ldots, x_p(n-D+1)\right]^T. \tag{10.88}$$

With this expansion, (10.86) reads as

$$\mathcal{SC}\left\{\mathbf{x}(j)\mathbf{e}^T(j)\mathbf{R}_{\mathbf{ee}}^{-1}(i)\right\} = \begin{bmatrix} \check{\mathbf{x}}_{1,D}^T(j) \\ \vdots \\ \check{\mathbf{x}}_{1,D}^T(j-L+1) \\ \check{\mathbf{x}}_{2,D}^T(j) \\ \vdots \\ \check{\mathbf{x}}_{2,D}^T(j-L+1) \end{bmatrix} \mathbf{R}_{\mathbf{ee}}^{-1}(i)\mathbf{e}(j). \tag{10.89}$$

In the same way as shown in Sect. 10.4.5 in the context of nearly Gaussian source models, we now factorize the inverse covariance matrix $\mathbf{R}_{\mathbf{ee}}^{-1}$ as [62]

$$\mathbf{R}_{\mathbf{ee}}^{-1}(i) = \mathbf{A}(i)\boldsymbol{\Sigma}_{\tilde{\mathbf{e}}\tilde{\mathbf{e}}}^{-1}(i)\mathbf{A}^T(i), \tag{10.90}$$

where $\mathbf{A}(i)$ and $\boldsymbol{\Sigma}_{\tilde{\mathbf{e}}\tilde{\mathbf{e}}}(i)$ denote again a $D \times D$ unit lower triangular matrix and a diagonal matrix, respectively [62].

By interpreting $\mathbf{A}(i)$ as a time-varying convolution matrix of a whitening filter, we model the signal e as an AR process of order $D-1$, with time-varying AR coefficients $a_k(n)$, and residual signal $\tilde{e}(n)$, i.e.,

$$e(n) = -\sum_{k=1}^{D-1} a_k(n)e(n-k) + \tilde{e}(n). \tag{10.91}$$

Now, (10.89) can be rewritten by shifting the *prefiltering matrix* $\mathbf{A}$ into the data terms, i.e.,

$$\tilde{\mathbf{e}} := \mathbf{A}^T \mathbf{e} = [\tilde{e}(n), \tilde{e}(n-1), \ldots, \tilde{e}(n-D+1)]^T, \tag{10.92}$$

$$\check{\tilde{\mathbf{x}}}_{p,D} := \mathbf{A}^T \check{\mathbf{x}}_{p,D} = [\tilde{x}_p(n), \tilde{x}_p(n-1), \ldots, \tilde{x}_p(n-D+1)]^T, \tag{10.93}$$

so that

$$\begin{aligned}
\mathcal{SC}\left\{\mathbf{x}(j)\mathbf{e}^T(j)\mathbf{R}_{\mathbf{ee}}^{-1}(i)\right\} &= \begin{bmatrix} \check{\tilde{\mathbf{x}}}_{1,D}^T(j) \\ \vdots \\ \check{\tilde{\mathbf{x}}}_{1,D}^T(j-L+1) \\ \check{\tilde{\mathbf{x}}}_{2,D}^T(j) \\ \vdots \\ \check{\tilde{\mathbf{x}}}_{2,D}^T(j-L+1) \end{bmatrix} \boldsymbol{\Sigma}_{\tilde{\mathbf{e}}\tilde{\mathbf{e}}}^{-1}(i)\tilde{\mathbf{e}}(j) \\
&= \left[\check{\tilde{\mathbf{x}}}(j), \ldots, \check{\tilde{\mathbf{x}}}(j-D+1)\right] \begin{bmatrix} \frac{\tilde{e}(j)}{\sigma_{\tilde{e}}^2(j)} \\ \vdots \\ \frac{\tilde{e}(j-D+1)}{\sigma_{\tilde{e}}^2(j-D+1)} \end{bmatrix} \\
&= \sum_{d=0}^{D-1} \check{\tilde{\mathbf{x}}}(j-d)\frac{\tilde{e}(j-d)}{\sigma_{\tilde{e}}^2(j-d)}.
\end{aligned} \tag{10.94}$$

Finally, (10.85) becomes

$$\begin{aligned}
\mathbf{w}^{\ell}(m) = \mathbf{w}^{\ell-1}(m) + \frac{2\mu}{N}\sum_{i=0}^{\infty}\beta(i,m) \\
\times \sum_{j=iN_L}^{iN_L+N-1}\sum_{d=0}^{D-1} \check{\tilde{\mathbf{x}}}(j-d)\frac{\tilde{e}(j-d)}{\sigma_{\tilde{e}}^2(j-d)}\phi_{e,D}\left(\tilde{\mathbf{e}}^T(j)\boldsymbol{\Sigma}_{\tilde{\mathbf{e}}\tilde{\mathbf{e}}}^{-1}(i)\tilde{\mathbf{e}}(j)\right).
\end{aligned} \tag{10.95}$$

Note that this formulation provides a computationally efficient realization of the generic Sylvester constraint.

Moreover, it is interesting to note that both the error signal e and the input (i.e., microphone) signal vector $\check{\mathbf{x}}$ appear as filtered versions in the update. After interpreting $\mathbf{A}$ in (10.90) as a whitening filter, this adaptation algorithm can in fact be interpreted as a so-called *filtered-x*-type algorithm [24]. As shown in Fig. 10.9, this type of algorithm typically appears whenever there is another filter between the adaptive filter and the position of the error calculation. This cascade structure will also be of fundamental importance in the direct-inverse approach in Sect. 10.6.

10.5.3 Efficient Frequency-domain Realizations for the MIMO Case

For convolutive mixtures, the classical approach of frequency-domain BSS appears to be an attractive alternative where all techniques originally developed for instan-

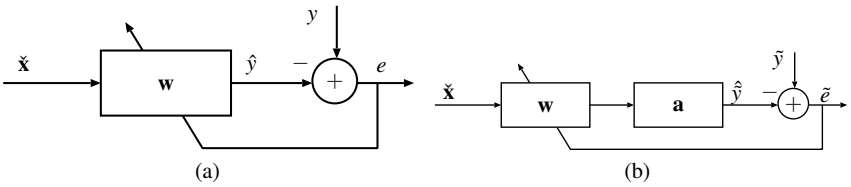


Fig. 10.9 Supervised adaptive filtering in (a) conventional and (b) filtered-x configuration

taneous BSS are typically applied independently in each frequency bin, e.g., [51]. However, this traditional narrowband approach exhibits several limitations as identified in, e.g., [7, 53, 78]. In particular, the permutation problem, which is inherent to BSS, may then also appear independently in each frequency bin so that extra repair measures are needed to address this *internal* permutation. Problems caused by circular convolution effects due to the narrowband approximation are reported in, e.g., [78].

In [15] it is shown how the equations of the TRINICON framework can be transformed into the frequency domain in a rigorous way (i.e., without any approximations) in order to avoid the above-mentioned problems. As in the case of the time-domain algorithms, the resulting generic DFT-domain BSS may serve both as a unifying framework for existing algorithms, and also as a guideline for developing new improved algorithms by certain suitable *selective* approximations as shown in, e.g., [15] or [2]. Figure 10.10 gives an overview on the most important classes of DFT-domain BSS algorithms known so far. A very important observation from this framework using multivariate PDFs is that, in general, all frequency components are linked together so that the internal permutation problem is avoided (the following elements are reflected in Fig. 10.10 by different approximations of the generic SIRP-based BSS):

1. Constraint matrices appearing in the generic frequency-domain formulation (see, e.g., [15]) describe the inter-frequency correlation between DFT components.
2. The multivariate score function, derived from the multivariate PDF is a broadband score function. As an example, for SIRPs the argument of the multivariate score function (which is a nonlinear function in the higher-order case) is $\mathbf{y}_p^{\mathrm{T}}(j)\mathbf{R}^{-1}_{\mathbf{y}_p\mathbf{y}_p}(i)\mathbf{y}_p(j)$ according to (10.50). Even for the simple case $\mathbf{R}^{-1}_{\mathbf{y}_p\mathbf{y}_p}(i) = \mathbf{I}$ where we have $\mathbf{y}_p^{\mathrm{T}}(j)\mathbf{y}_p(j) = \|\mathbf{y}_p(j)\|^2$, i.e., the quadratic norm, and – due to Parseval's theorem – the same in the frequency domain, i.e., the quadratic norm over all DFT components, we immediately see that all DFT-bins are taken into account simultaneously so that the internal permutation problem is avoided. Note that the traditional narrowband approach (with the internal permutation problem) would result as a special case if we assumed all DFT components to be statistically independent from each other (which is of course not the case for real-world broadband signals such as speech and audio signals). In contrast to

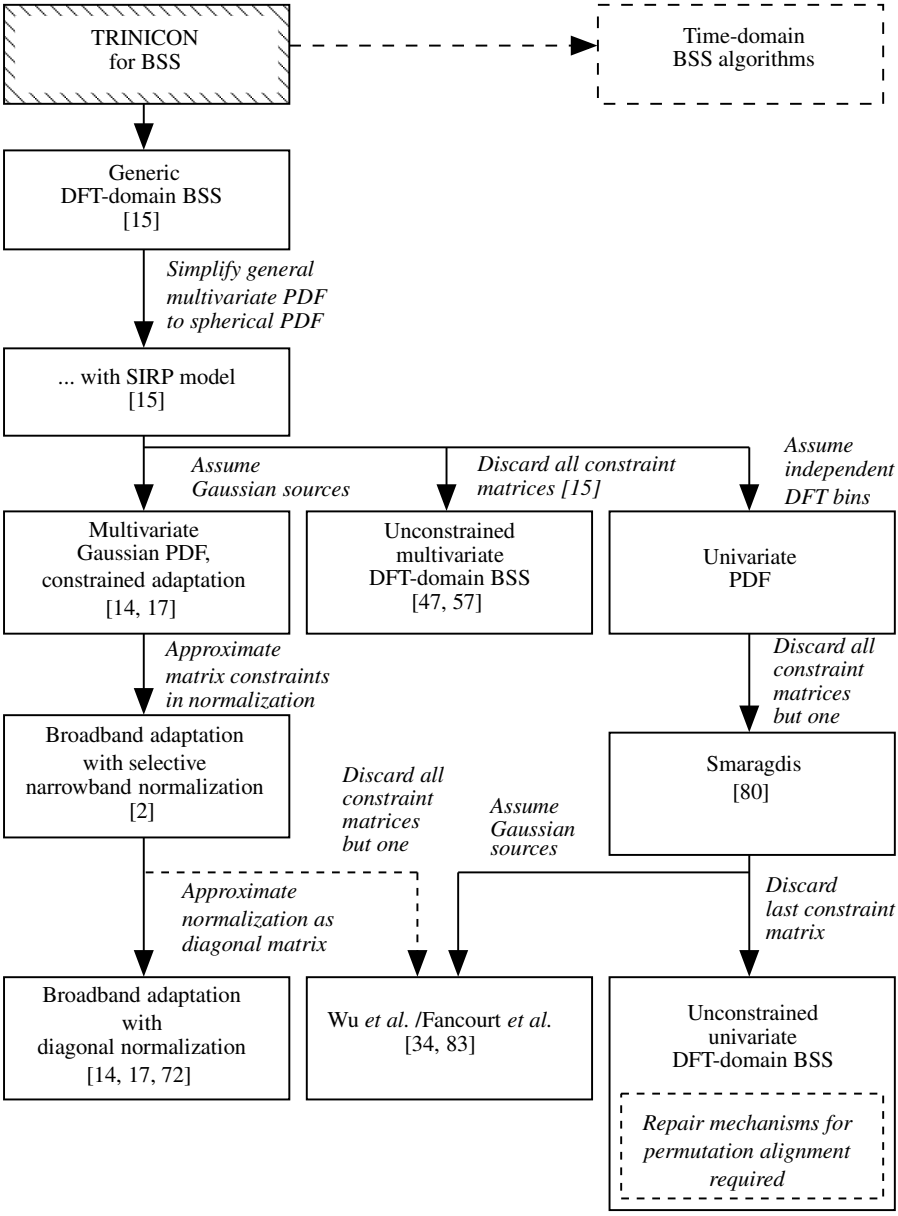


Fig. 10.10 Overview of BSS algorithms in the DFT domain. Note that the broadband algorithms in the left-hand column are also suitable for BSI, and thus, for the identification-and-inversion approach to blind deconvolution/blind dereverberation

this independence approximation the dependencies among all frequency components (including higher-order dependencies) are inherently taken into account in TRINICON in an optimal way by considering the joint densities as the most comprehensive description of random signals. Actually, in the traditional narrowband approach, the additionally required repair mechanisms for permutation alignment try to exploit such inter-frequency dependencies.

From the viewpoint of blind system identification, *broadband algorithms with constraint matrices* (i.e., the algorithms represented in the first column of Fig. 10.10) are of particular interest. Among these algorithms, the system described in [2] has turned out to be very efficient in this context. A pseudo-code of this algorithm is also included in [2].

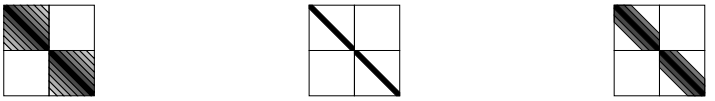
Another important consideration for the practical implementation of BSI is the proper choice of the Sylvester constraint. Since the *column constraint* $\mathcal{SC}_{\mathrm{C}}$ is not suited for arbitrary source configurations, it is generally *not appropriate* for BSI and deconvolution. Thus, for the implementations discussed in this chapter the *row constraint* $\mathcal{SC}_{\mathrm{R}}$ is used.

10.6 Application of TRINICON to the Direct-inverse Approach to Blind Deconvolution

In this section we discuss multichannel blind adaptation algorithms with the aim to solve the inverse adaptive filtering problem (see Table 10.1) directly without BSI as an intermediate step. This section mainly follows and extends the concept first presented in [16].

The two main aspects in this section are as follows:

- First, we briefly discuss traditional ICA-based Multichannel Blind Deconvolution (MCBD) algorithms from the literature. Unfortunately, as we will see, these algorithms are not well suited for speech and audio signals. However, our considerations lead to various insights and to a classification scheme that is also useful for both the pure separation/identification algorithms from the previous section and also to the MultiChannel Blind Partial Deconvolution (MCBPD) algorithms considered afterwards.
- A discussion of the MCBPD algorithms is also given. These algorithms can be regarded as advanced versions of MCBD so that they are also suitable for speech and audio signals. As already mentioned at the end of Sect. 10.3.5, these algorithms are not just based on the spatial diversity and the statistical independence of the different source signals, but they require more precise stochastic source models. Based on the results of Sect. 10.4, and to some extent of Sect. 10.5, we present a general framework which unifies the treatment of many of the known algorithms for the direct-inverse approach to blind dereverberation of speech signals, and also leads to various new algorithms.



(a) (b) (c)

Fig. 10.11 Desired correlation matrices $\mathbf{R}_{\mathbf{ss}}$ for (a) BSS (Sect. 10.5), (b) MCBD (Sect. 10.6.1), and (c) MCBPD (Sect. 10.6.2) with TRINICON in the SOS case

10.6.1 Multichannel Blind Deconvolution

Analogously to the Sect. 10.5.1, we now specialize TRINICON to the case of traditional MCBD algorithms. As shown by (10.45b), this class of algorithms is specified by a complete factorization of the hypothesized source model $\hat{p}_{s,PD}(\cdot)$, i.e., traditionally, ICA-based MCBD algorithms assume i.i.d. source models, e.g., [4, 28]. In other words, in addition to the separation of statistically independent sources, MCBD algorithms also temporally whiten the output signals; thus this approach is not directly suitable for audio signals. Nevertheless, studying these algorithms leads to some important insights, because in contrast to some BSS algorithms they are inherently broadband algorithms. Their popularity results from the fact that due to the complete factorization of the source model, they only require univariate PDFs. Thereby, the multivariate score function (10.44b) reduces to a vector of univariate score functions each representing a scalar nonlinearity. As, additionally, the second term in (10.44b) is commonly neglected in most of these algorithms, the scalar nonlinearity reads

$$\Phi_{y_p,1}(y_p(j-d+1)) = -\frac{\partial \log \hat{p}_{y_p,1}(y_p(j-d+1))}{\partial y_p(j-d+1)}. \tag{10.96}$$

The corresponding generic coefficient update rules are then given by (10.44a), (10.46a), (10.48), and (10.49).

In the SOS case, analogously to the representation in Sect. 10.5.1, the complete factorization of the output PDF corresponds to the desired correlation matrix $\mathbf{R}_{\mathbf{ss}} = \operatorname{diag} \mathbf{R}_{\mathbf{yy}}$, as illustrated in Fig. 10.11(b).

Using (10.96) several relationships between the generic HOS natural gradient update rule (10.49) and well-known MCBD algorithms in the literature can be established [1]. As noted in Sect. 10.4.5, these links are obtained by the application of different implementations of the Sylvester constraint $\mathcal{SC}$, the distinction between the correlation and covariance method [66] for the estimation of the cross-relation

$$\mathbf{R}_{\mathbf{y}\boldsymbol{\Phi}(\mathbf{y})}(i) = \frac{1}{N} \sum_{j=iN_L}^{iN_L+N-1} \mathbf{y}(j)\boldsymbol{\Phi}_{s,PD}^{\mathrm{T}}(\mathbf{y}(j)) \tag{10.97}$$

in (10.49), and the different approximations of the multivariate PDFs. This altogether spans a whole tree of algorithms as depicted in Fig. 10.12. Here, the most

general algorithm is given as the generic HOS natural gradient algorithm (10.49), which is based on multivariate PDFs. A distinction with respect to the implementation of the Sylvester constraint $\mathcal{SC}$ leads to two branches, which can again be split up with respect to the method used for the estimation of the cross-relation matrices. Approximating the multivariate PDFs by univariate ones, neglecting the nonstationarity, and using the Sylvester constraint $\mathcal{SC}_{\mathrm{R}}$ yields the two block-based MCBD algorithms presented in [30, 54]. By changing the block-based adaptation to a sample-by-sample algorithm, a link to the popular MCBD algorithm in [4] and [31] can be established. (It should be noted that also the so-called nonholonomic extension [15] of [4] presented in [28] can be derived from the framework.) By using the Sylvester constraint $\mathcal{SC}_{\mathrm{C}}$ a link to the MCBD algorithm in [88] is obtained. However, it should be remembered that algorithms based on $\mathcal{SC}_{\mathrm{C}}$ are less general as only causal filters can be adapted and thus for MCBD algorithms only minimum-phase systems can be treated, as was pointed out in [88].

Note that by using the general Sylvester constraint without approximations, a performance gain both over $\mathcal{SC}_{\mathrm{R}}$ and $\mathcal{SC}_{\mathrm{C}}$ is possible [20].

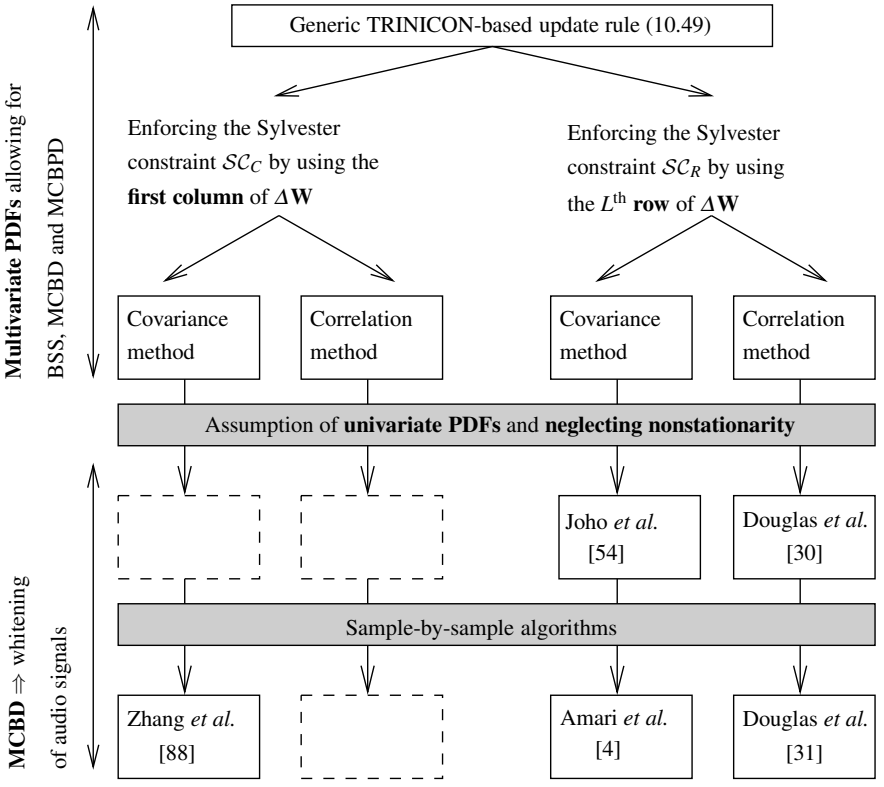


Fig. 10.12 Overview of links between the generic algorithm (10.49) and existing MCBD algorithms after [1]

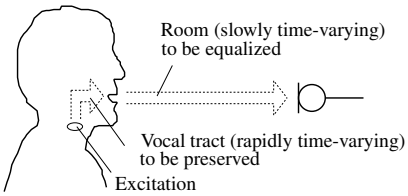


Fig. 10.13 Illustration of speech dereverberation as an MCBPD application (after [16])

10.6.2 Multichannel Blind Partial Deconvolution

Signal sources that are non i.i.d. should not become i.i.d. at the output of the blind adaptive filtering stage. Therefore, their statistical dependencies should be preserved. In other words, the adaptation algorithm has to distinguish between the statistical dependencies within the source signals, and the statistical dependencies introduced by the mixing system $\check{\mathbf{H}}$, i.e., the reverberant room. We denote the corresponding generalization of the traditional MCBD technique as MCBPD [16]. Equations (10.44)–(10.49) inherently contain a statistical source model (signal properties (i)–(iii) in Sect. 10.4.2), expressed by the multivariate densities, and thus provide all necessary requirements for the MCBPD approach.

Ideally, only the influence of the room acoustics should be minimized. A typical example for MCBPD applications is speech dereverberation, which is especially important for distant-talking automatic speech recognition (ASR), where there is a strong demand for speech dereverberation without introducing artifacts to the signals. In this application, MCBPD allows us to distinguish between the actual speech production system, i.e., the vocal tract, and the reverberant room (Fig. 10.13).

For the distinction between the production system of the source signals (e.g., the speech production system) and the room acoustics we can again exploit all three fundamental signal properties already mentioned in Sect. 10.4.2:

(i) *Nonwhiteness.* The auto-correlation structure of the speech signals can be taken into account, as illustrated in Fig. 10.11(c). While the room acoustics influences all off-diagonals, the effect of the vocal tract is concentrated in the first few off-diagonals around the main diagonal. In the simplest case, these first Z off-diagonals of $\mathbf{R}_{\mathbf{yy}}$ are now taken over into the banded matrix

$$\mathbf{R}_{\mathbf{ss}} = \text{bandbdiag}_Z \mathbf{R}_{\mathbf{yy}}, \tag{10.98}$$

as illustrated in Fig. 10.11(c). Note that there is a close link to linear prediction techniques as detailed below which gives guidelines for the number of lags to be preserved.

(ii) *Nonstationarity.* The speech production system and the room acoustics also differ in their time-variance according to Fig. 10.13. While the room acoustics is assumed to be constant during the adaptation process, the speech signal is only short-time stationary [66], modeled by the time-varying speech pro-

duction model. Typically, the duration of the stationarity intervals is assumed to be approximately 20 ms [66]. We therefore adjust the block length N and in practice preferably also the block shift N_L in the criterion (10.39) with the model parameter estimates (10.41) and in the corresponding updates (10.44)–(10.49) to the assumed duration of the stationarity interval. Note that for a block-based adaptation (typically performed by exploiting the efficiency of the FFT, cf. Sect. 10.5.3 for the case of BSS) and $N = N_L < L$, this corresponds to a *partitioned* block formulation as known from supervised adaptive filtering, e.g., [22].

(iii) *Nongaussianity.* Speech is a well-known example for supergaussian signals. Due to a convolutive sum – in our application describing the filtering by the room acoustics – the PDFs of the recorded sensor/microphone signals tend to be somewhat closer to Gaussians. Hence, another strategy is to maximize the nongaussianity of the output signals of the demixing system (as far as possible by the MIMO FIR filters), e.g., [12, 41, 60, 82]. This strategy is addressed, e.g., using the kurtosis as a widely-used distance measure of nongaussianity as in the second term in (10.56). It can be shown that this second term can indeed be identified as an estimate of the so-called *negentropy*, which is an information-theoretic distance measure to the Gaussian [51].

Formally, the above-mentioned exploitation of the nonwhiteness to distinguish between the coloration of the sources and the mixing system is achieved by decoupling the prediction order n_A in (10.61) from the dimension D of the correlation matrix $\mathbf{R}_{\mathbf{yy}}$, i.e.,

$$\tilde{y}_p(n) = \sum_{k=0}^{n_A} a_{p,k}(n) y_p(n-k) \tag{10.99}$$

with $0 \leq n_A \leq D-1$ and $a_{p,0}(n) \equiv 1$. This corresponds to a generalization of the upper triangular matrix structure (10.62) in the factorization (10.60) to the *banded* matrix

$$\mathbf{A}_p = \begin{bmatrix} 1 & a_{p,1}(n) & a_{p,2}(n) & \cdots & a_{p,n_A}(n) & 0 & \cdots & 0 \\ 0 & 1 & a_{p,1}(n-1) & \cdots & a_{p,n_A-1}(n-1) & a_{p,n_A}(n-1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}^T \tag{10.100}$$

so that we can again apply the compact notation

$$\tilde{\mathbf{y}}_q = \mathbf{A}_q^T \mathbf{y}_q = [\tilde{y}_q(n), \tilde{y}_q(n-1), \ldots, \tilde{y}_q(n-D+1)]^T, \tag{10.101}$$

$$\tilde{\check{\mathbf{x}}}_{p,D}^{(q)} = \mathbf{A}_q^T \check{\mathbf{x}}_{p,D} = \left[\tilde{x}_p^{(q)}(n), \tilde{x}_p^{(q)}(n-1), \ldots, \tilde{x}_p^{(q)}(n-D+1)\right]^T. \tag{10.102}$$

Hence, the resulting formulation of the generalized score function (10.65) carries over to MCBPD, as well as to the traditional MCBD and to broadband BSS/BSI, depending on the parameter n_A. In other words, the different modes in Fig. 10.11 are selected by certain choices of the order n_A. This is further illustrated in Fig. 10.14.

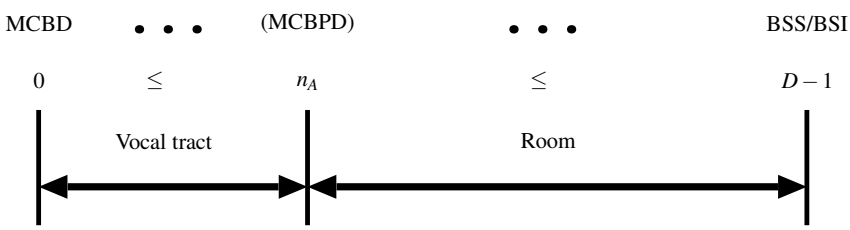


Fig. 10.14 Illustration of the parameter n_A

The corresponding general gradient descent-based coefficient update for nearly Gaussian sources is then obtained by introducing the score function (10.65) into the generic update (10.46a). Note that for an efficient implementation of the Sylvester constraint of the first term in (10.46a) we can apply the same procedure as demonstrated in (10.87) and (10.89). With (10.102) we then obtain

$$\begin{aligned}\check{\mathbf{w}}_{pq}^{\ell}(m) = \check{\mathbf{w}}_{pq}^{\ell-1}(m) - \frac{\mu}{N}\sum_{i=0}^{\infty}\beta(i,m)\sum_{j=iN_L}^{iN_L+N-1}\sum_{d=0}^{D-1}\check{\tilde{\mathbf{x}}}_p^{(q)}(j-d)\\ \times\left[\frac{\tilde{y}_q(j-d)}{2\hat{\sigma}_{\tilde{y}_q}^2(j-d)} - \left(\frac{\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}_q^4(j-d)}{3\hat{\sigma}_{\tilde{y}_q}^4(j-d)} - 1\right)\right.\\ \left.\times\left(\frac{\tilde{y}_q^3(j-d)}{\hat{\sigma}_{\tilde{y}_q}^4(j-d)} - \frac{\tilde{y}_q(j-d)\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}_q^4(j-d)}{\hat{\sigma}_{\tilde{y}_q}^6(j-d)}\right)\right]\\ +\mu\sum_{i=0}^{\infty}\beta(i,m)\mathcal{SC}\left[\mathbf{V}\left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\mathbf{V}\right)^{-1}\right]_{pq}.\end{aligned} \tag{10.103}$$

This general TRINICON-based MIMO coefficient update for nearly Gaussian sources leads both to blind separation and dereverberation of the signals.

Analogously to the considerations at the end of Sect. 10.5.2 we see that this update rule can again be interpreted as a so-called *filtered-x*-type algorithm since both the input (i.e., microphone) signal vector and the output signals appear as filtered versions in the update. Analogously to Fig. 10.9 we immediately obtain Fig. 10.15 for the dereverberation application as a consequence of this filtered-x interpretation. While $\mathbf{W}$, driven by the filtered-x-type coefficient update, ideally inverts the room acoustic mixing system $\mathbf{H}$, the (set of) linear prediction filter(s) $\mathbf{A}$ from the stochastic source model ideally inverts the (set of) speech production system(s) of the source(s). The coefficient updates of $\mathbf{W}$ and the estimation of $\mathbf{A}$ are carried out in an alternating fashion like the estimation of the other stochastic model parameters, as mentioned in Sect. 10.4.5. Note that (in accordance with the known filtered-x concept) the filtered input vector $\check{\tilde{\mathbf{x}}}$ in (10.103) is obtained using the filter coefficients from the Linear Prediction (LP) analysis of the *output* signals y_p. In other words, the

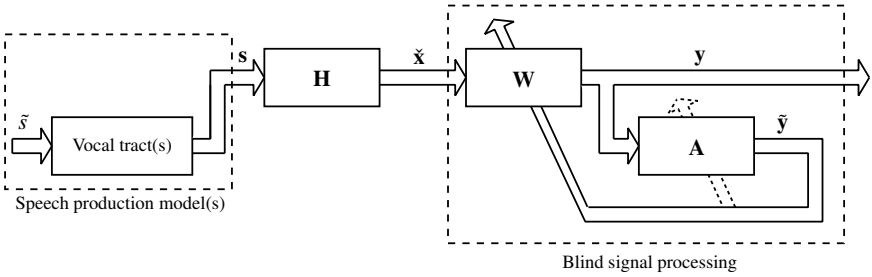


Fig. 10.15 Inversion of the speech production models within the blind signal processing and filtered-x-type interpretation

coefficients of the output LP analysis filters are copied to the input transformation filters according to (10.102).

It should be mentioned that the linear prediction is also classified as a (blind) inverse adaptive filtering problem in Table 10.1, and hence, the estimate of the prediction coefficients can also be obtained directly from the TRINICON optimization criterion (10.39). Assuming a single-source scenario and SOS-based estimation of the prediction coefficients for this inverse adaptive filtering problem, as a special case of (10.39) according to (10.54) and the considerations in Sect. 10.5.2 for the single-source case, we obtain

$$\mathcal{J}_{\text{pred}}(m,\mathbf{A}) = \sum_{i=0}^{\infty} \beta(i,m) \log\det \operatorname{diag} \mathbf{R}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}(i) \propto \sum_{i=0}^{\infty} \beta(i,m) \log \hat{\sigma}^2_{\tilde{y},i}. \tag{10.104}$$

Furthermore, assuming stationarity, this criterion is equivalent to the traditional least-squares-based estimate $\mathcal{J}_{\text{pred,LS}}(m,\mathbf{A}) \propto \hat{\sigma}^2_{\tilde{y},m}$ due to the monotonicity of the logarithm, while for non-stationary signals, it is more general. Nevertheless, for the practical experiments in Sect. 10.7 we will apply the Levinson–Durbin algorithm as an efficient realization of the LS-based estimation using the so-called correlation method [66].

10.6.3 Special Cases and Links to Known Algoritms

According to Fig. 10.14, all of the previously discussed algorithms from the various classes according to Table 10.1 can be regarded as special cases of the MCBPD concept. In this section, we only discuss algorithms that are specifically designed for dereverberation using the direct-inverse approach. Moreover, we focus here on algorithms based on the Gram–Charlier model, i.e., we discuss special cases of (10.103) and relations to some known algorithms.

10.6.3.1 SIMO vs. MIMO Mixing Systems

Similar to the considerations in Sect. 10.5.2 for SIMO-based BSI, we deduce now the specialized coefficient update for the case of SIMO mixing systems, i.e., for the case of only one source signal. Again, we first consider the last term of the generic gradient-based update (10.103). According to the corresponding steps of the derivation in Sect. 10.5.2 (Eqs. (10.72)-(10.78)) we can see that in the same way the last term also disappears for MCBPD in the SIMO case. Next, we pick the filter coefficients of interest for the SIMO case. Assuming the active source signal will appear on the first output of the demixing filter, it is straightforward to pick $\mathbf{w}$ as the first column of the general MIMO coefficient matrix $\check{\mathbf{W}}$. This way we immediately obtain

$$\begin{aligned}\mathbf{w}^{\ell}(m) = \mathbf{w}^{\ell-1}(m) - \frac{\mu}{N}\sum_{i=0}^{\infty}\beta(i,m)\sum_{j=iN_L}^{iN_L+N-1}\sum_{d=0}^{D-1}\check{\tilde{\mathbf{x}}}(j-d)\\ \times\left(\frac{\tilde{y}(j-d)}{2\hat{\sigma}^2_{\tilde{y}}(j-d)} - \left(\frac{\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}^4(j-d)}{3\hat{\sigma}^4_{\tilde{y}}(j-d)} - 1\right)\right.\\ \left.\times\left(\frac{\tilde{y}^3(j-d)}{\hat{\sigma}^4_{\tilde{y}}(j-d)} - \frac{\tilde{y}(j-d)\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}^4(j-d)}{\hat{\sigma}^6_{\tilde{y}}(j-d)}\right)\right).\end{aligned} \tag{10.105}$$

Note that the structure of the resulting algorithm is very similar to the one of the generalized AED (10.95) in Sect. 10.5.2. The main differences are the different sign of the update term and the fact that we now pick the *first* column of $\check{\mathbf{W}}$, since we are now interested in obtaining the enhanced signal rather than in minimizing an error signal for the signal cancellation in the AED.

10.6.3.2 Efficient Implementation Using the Correlation Method

An efficient implementation that still exploits all three fundamental signal properties as discussed in Sect. 10.6.2 is obtained by assuming a global nonstationarity of the source signals but short-time stationarity in each block as known from linear prediction. As a first step to obtain a simplified update equation, we integrate the sum over d into the sum over j. Next, we replace the time-varying output prediction error variances by blockwise constant values $\hat{\sigma}_{\tilde{y}_p,i}$ for the i^{th} block. This finally allows us to move the sum over j into the numerators in the brackets in order to obtain the compact expression

$$\begin{aligned}\check{\mathbf{w}}_{pq}^{\ell}(m) = \check{\mathbf{w}}_{pq}^{\ell-1}(m) - \frac{\mu}{N}\sum_{i=0}^{\infty}\beta'(i,m) \\ \times\left[\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\check{\mathbf{x}}_p^{(q)}(j)\tilde{y}_q(j)}{2\hat{\sigma}_{\tilde{y}_{q,i}}^2} - \left(\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\tilde{y}_q^4(j)}{3\hat{\sigma}_{\tilde{y}_{q,i}}^4} - 1\right)\right. \\ \left.\times\left(\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\check{\mathbf{x}}_p^{(q)}(j)\tilde{y}_q^3(j)}{\hat{\sigma}_{\tilde{y}_{q,i}}^4} - \frac{\sum_{j=iN_L'}^{iN_L'+N-1}\check{\mathbf{x}}_p^{(q)}(j)\tilde{y}_q(j)\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}_q^4(j)}{\hat{\sigma}_{\tilde{y}_{q,i}}^6}\right)\right] \\ +\mu\sum_{i=0}^{\infty}\beta(i,m)\mathcal{SC}\left[\mathbf{V}\left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\mathbf{V}\right)^{-1}\right]_{pq}. \end{aligned} \tag{10.106}$$

Note that for the SIMO case this expression is simplified in a straightforward way, as mentioned in the previous paragraph, so that the last term again disappears. This efficient version is also used for the experiments in Sect. 10.7.

10.6.3.3 Relations to Some Known HOS Approaches

As has already been mentioned in Sect. 10.6.2 most of the HOS-based blind deconvolution approaches aim at finding deconvolution filters that render the output signals as nongaussian as possible [12, 60, 82] with kurtosis being the most common measure for nongaussianity.

In [41] an approach to speech dereverberation by kurtosis maximization was presented. It is based on the idea of performing the whole adaptation and filtering procedure on LP residuals as a heuristic extension of the ideas in [10, 86]. Hence, the main structural difference of this approach to the general TRINICON-based update rule is that the LP analysis is carried out using the microphone signals, i.e., the input signals of the blind adaptive filter rather than on its output signals as in the above-mentioned and systematically obtained filtered-x structure. Nevertheless, the resulting algorithm also exhibits several remarkable similarities to the generic update. The adaptation rule in [41] is based directly on the kurtosis, i.e., the *square root* of only the second term in (10.56). The update therefore structurally corresponds to the part in the second parentheses of the second term in the brackets in (10.103). (The first term in (10.103) results from the SOS and the expression in the first parentheses in the second term results from the application of the chain rule due to the *square* of the kurtosis in the Gram–Charlier expansion.)

The same approximate expression of the update rule, i.e., the gradient descent directly based on the kurtosis is also used in [87]. Note that these approaches are based on the acoustic SIMO model.

10.6.3.4 Relations to Some Known SOS Approaches

It is known that linear filtering of a source signal increases the temporal predictability of the observed signal. A deconvolution filter that makes its output less predictable may thus be able to recover the source signal. This observation is the key to most SOS-based linear deconvolution methods, i.e., in essence they aim at finding deconvolution filters that minimize a measure of predictability of the output signal, e.g., [81]. Hence, in a certain sense, blind deconvolution may also be interpreted as the application of a very long linear prediction error filter. Note that this is also reflected by the symmetric structure in Fig. 10.15.

As a simple approach, the optimization criterion in [81] is directly based on the variance of the long-term prediction error at the output of the deconvolution filter. In order to avoid trivial solutions and to preserve some of the temporal structure of the source signals, this long-term prediction error variance is normalized by a short-term prediction error variance, and finally the logarithm of this ratio is taken. Although this approach does not explicitly exploit the nonstationarity of the signals in the sense as outlined in Sect. 10.6.2, this logarithm of the ratio between the prediction error variances – which can be expressed as a difference between two logarithmic prediction error variances – can still roughly be related to the generic SOS-based optimization criterion (10.54) considering the link with linear prediction at the end of Sect. 10.6.2, and the short-term prediction error variance in the normalization as a special case of the *desired* correlation matrix $\mathbf{R}_{\mathbf{ss}}$.

Another related approach to preserve the temporal structure of the original source signal is called correlation shaping in [40]. The heuristically introduced optimization criterion after Gillespie and Atlas in [40] for the SIMO case reads

$$\mathcal{J}_{\mathrm{GA}} = \sum_{\kappa} \gamma(\kappa) \left(r_{yy}(\kappa) - r_{ss}(\kappa) \right)^2 , \tag{10.107}$$

where κ denotes the lag of the output correlation sequence $r_{yy}(\kappa)$ and a certain desired correlation sequence $r_{ss}(\kappa)$. The factor $\gamma(\kappa)$ allows for an individual weighting of the lags. As a preferred embodiment of this concept, in [40] it is proposed to choose $\gamma(\kappa)$ and $r_{ss}(\kappa)$ such that $r_{yy}(\kappa)$ is minimized for all lags outside of the *don't care* region $-Z \leq \kappa \leq Z$. Obviously, this approach is equivalent to the minimization of the Frobenius norm $\mathcal{J}_{\mathrm{F,GA}} = \|\mathbf{R}_{\mathbf{yy}} - \mathbf{R}_{\mathbf{ss}}\|_{\mathrm{F}}^2$ with the banded matrix $\mathbf{R}_{\mathbf{ss}} = \mathrm{bandbdiag}_Z \mathbf{R}_{\mathbf{yy}}$ after (10.98) and Fig. 10.11(c) if the so-called *correlation method* is used for the estimation of $\mathbf{R}_{\mathbf{yy}}$ (i.e., this matrix is assumed to be Toeplitz). Hence, in the context of dereverberation the approach [40] can be seen directly as an analogon to the Frobenius-based approaches for BSS/BSI mentioned in Sect. 10.5.1 (e.g., [26, 51, 52, 69, 74, 79]). The main differences between [40] and the generic SOS-based MCBPD are:

(i) The criterion (10.107) does not exploit the nonstationarity of the signals in the sense as outlined in Sect. 10.6.2.

(ii) As already explained in Sect. 10.5.1, in contrast to the generic SOS criterion (10.54) the minimization of the Frobenius-based criterion does not lead to the inherent normalization of the coefficient update, which can be interpreted as an in-

herent step-size control according to Sect. 10.5.2, and hence is an important feature for a robust adaptation performance. Similar to the BSS/BSI case, many simulation results have shown that for large filter lengths L, the Frobenius-based adaptation is prone to instability, while the generic MCBPD adaptation shows a very robust convergence behavior for real acoustic environments, as we will see in Sect. 10.7.

In [35, 37] a third related SOS-based approach was presented. As in the previously described SOS-based algorithms, this approach distinguishes between the speech production system and the room acoustics by exploiting only the nonwhiteness. It explicitly takes into account an estimate of the *long-term* power spectral density of the speech signal. Moreover, an interesting aspect of this approach is that it was originally derived directly from MINT (see Sect. 10.2) describing the ideal inversion solution at the equilibrium of the adaptation. Indeed, it can be shown (analogously to the analysis of the equilibria for BSS in [17] in the SOS case) that ideally the equilibrium of the SOS-based update (10.67) in the case of MCBPD with (10.98) corresponds to the MINT solution according to Sect. 10.2. We now show how this approach can be derived rigorously from the TRINICON-based coefficient update (10.67). Under the stationarity assumption we have in the equilibrium

$$\Delta\mathbf{W} = \mathbf{R}_{\mathbf{xy}}\left[\mathbf{R}_{\mathbf{ss}}^{-1} - \mathbf{R}_{\mathbf{yy}}^{-1}\right] = \mathbf{0}, \tag{10.108}$$

i.e.,

$$\mathbf{R}_{\mathbf{xy}} = \mathbf{R}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{ss}}. \tag{10.109}$$

Developing the left-hand side of this equation as $\mathbf{R}_{\mathbf{xx}}\mathbf{W}$ and the right-hand side of this equation using Sylvester matrices and corresponding data matrices $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{S}$ of compatible dimensions as in [17] as $\mathbf{R}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{ss}} = \mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{S}^T\mathbf{S} = \mathbf{X}^T(\mathbf{Y}^T)^+\mathbf{S}^T\mathbf{S} = \mathbf{X}^T(\mathbf{S}^T)^+(\mathbf{C}^T)^+\mathbf{S}^T\mathbf{S} = \mathbf{X}^T\mathbf{S} = \mathbf{R}_{\mathbf{xs}}$, where $\cdot^+$ denotes the Moore–Penrose pseudoinverse, we obtain

$$\mathbf{R}_{\mathbf{xx}}\mathbf{W} = \mathbf{R}_{\mathbf{xs}}. \tag{10.110}$$

Note that this relation is in fact the Wiener–Hopf equation for the inverse filtering configuration. (This again reflects the equivalence to the traditional LS approach for inverse adaptive filtering problems in the stationary case, as mentioned at the end of Sect. 10.6.2 for the linear prediction problem.) Next, a filter $\mathbf{B}$ in the Sylvester structure modeling the vocal tract is introduced so that $\mathbf{S} = \mathbf{S}_0\mathbf{B}$, where $\mathbf{S}_0$ denotes a corresponding data matrix of the i.i.d. excitation signal. Hence

$$\mathbf{R}_{\mathbf{ss}} = \mathbf{S}^T\mathbf{S} = \mathbf{B}^T\mathbf{R}_{\mathbf{s_0s_0}}\mathbf{B} = \mathbf{B}^T\mathbf{B}. \tag{10.111}$$

Using this model, we can rewrite (10.110) as

$$\mathbf{R}_{\mathbf{xx}}\mathbf{W} = \mathbf{H}^T\mathbf{R}_{\mathbf{ss}} = \mathbf{H}^T\mathbf{B}^T\mathbf{B}. \tag{10.112}$$

Multiplication by the pseudoinverse of $\mathbf{B}$ on both sides, and exploiting the commutation property of the convolution ($\mathbf{B}$ denotes a SISO system), we can write

$$\mathbf{R}_{\mathbf{xx}}\mathbf{B}^{+}\mathbf{W} = \mathbf{B}^{T}\mathbf{H}^{T}, \tag{10.113}$$

or

$$\left(\mathbf{B}^{+}\right)^{T}\mathbf{R}_{\mathbf{xx}}\mathbf{B}^{+}\mathbf{W} = \mathbf{H}^{T}. \tag{10.114}$$

Let us denote the inverse filter of the vocal tract similarly as in the previous sections as $\mathbf{A} := \mathbf{B}^{+}$. Using this filter the correlation matrix $\mathbf{R}_{\mathbf{xx}}$ is transformed into $\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{A}^{T}\mathbf{R}_{\mathbf{xx}}\mathbf{A}$ so that

$$\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\mathbf{W} = \mathbf{H}^{T}. \tag{10.115}$$

We now pick only the first columns of the Sylvester matrices for the SIMO case on both sides. Moreover, it is important to assume that the first microphone is the one that is closest to the source [37]. Using this assumption we finally obtain

$$\mathbf{w} = h_{1,0}\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\mathbf{1}, \tag{10.116}$$

where $\mathbf{1} = [1,0,\ldots,0]^{T}$ and $h_{1,0}$ denotes the first coefficient of the acoustic model from the source to the first microphone, which acts as an arbitrary scaling factor. This expression exactly corresponds to the algorithm presented in [37] including the whitening procedure, originally introduced in a heuristic way. We can see from this derivation that this algorithm indeed follows from TRINICON for the SOS case and stationarity assumption. Moreover, we see that in contrast to the previously presented approaches, this algorithm requires some prior knowledge of the source position. In other words, it may in fact be regarded as a *semi-blind* deconvolution algorithm. Furthermore, it becomes obvious that extending this approach to the general MIMO case raises the problem of estimating the relative positions of multiple simultaneously active sound sources.

10.7 Experiments

In this section, we evaluate the dereverberation performance for both the SIMO case (i.e., one source) and the MIMO case (two sources) using measured data. In the first set of experiments in the SIMO case, we compare the convergence properties based on the exploitation of the different stochastic signal properties (SOS, HOS) for the ideal demixing filter length. We then compare the DI approach with the II approach and investigate the sensitivity of both approaches with respect to the overestimation of the filter lengths. Finally, by extending the scenario to the MIMO case, we consider both the separation performance and the dereverberation performance. For illustration, we also compare the results in the MIMO case with the corresponding results of pure separation algorithms.

10.7.1 The SIMO Case

The experiments were conducted using speech data convolved with impulse responses of length $M = 9000$ of a real room with a reverberation time $T_{60} \approx 700$ ms and a sampling frequency of 16 kHz. To begin with, we consider an acoustic SIMO scenario, i.e., there is only $Q = 1$ active sound source in the room. A linear four-element microphone array ($P = 4$) with an inter-element spacing of 16 cm was used. Preliminary experiments using MINT (see Sect. 10.2) applied to the measured impulse responses showed that for the choice $P = 4$ the ideal inversion solution indeed exists for the given acoustic scenario, i.e., the mixing system is invertible according to Sect. 10.2. The speech signal arrived from 24° relative to the normal plane of the array axis and the distance between the speaker and the center of the microphone array was 165 cm.

As has already been mentioned, according to MINT the overdetermined scenario $P > Q$ is required for dereverberation. From a practical point of view it is thus interesting to consider the required degrees of freedom depending on the number of sensors. The total number of filter coefficients is $C := LP$. According to (10.18), we obtain as the optimal number of filter coefficients in the SIMO case

$$C = P \cdot \frac{M-1}{P-1} = \frac{P}{P-1} \cdot (M-1). \tag{10.117}$$

We see that for the minimum number $P = 2$ of sensors we require $C = 2 \cdot (M-1)$ coefficients. For $P \to \infty$ it follows $C \to M-1$. It turns out that the total number of required filter coefficients *decreases* with an increasing number of microphones. Hence, the framework is well suitable and efficient for the overdetermined case.

To evaluate our simulation results there are various possible quality measures for dereverberation of speech and audio signals (e.g., [58, 59, 75, 76]), such as the reverberation time (T_{60}), the definition (D_{50}), the clarity index (C_{80}), the (Rapid) Speech Transmission Index (STI/RASTI), or Spectral Distortion (SD). While the first three quantities are system-based and are defined in the context of room acoustics, the latter two are signal-dependent distortion measures. Another signal-dependent quantity which is commonly used in the signal processing literature for the evaluation of dereverberation approaches is the *Signal-to-Reverberant Ratio* (SRR, see, e.g., [70]). Similarly to the quantities D_{50} and C_{80} it measures the power ratio between the direct sound and the contribution by the reverberation. However, since the SRR is signal-based, it also takes into account the excitation of the adaptation algorithm. It is measured in decibels (dB) and is defined for a signal s_q at a sensor with signal x_p as

$$\mathrm{SRR}_{p,s_q} = 10\log \frac{\sum_n \left(\sum_{\kappa=0}^{n_\Delta} h_{qp,\kappa} s_q(n-\kappa)\right)^2}{\sum_n \left(\sum_{\kappa=n_\Delta}^{M-1} h_{qp,\kappa} s_q(n-\kappa)\right)^2}\mathrm{dB}, \tag{10.118}$$

where n_Δ is a discrete-time index defining the boundary between the direct signal path and the contribution by the reverberation. Note that usually, in the case of

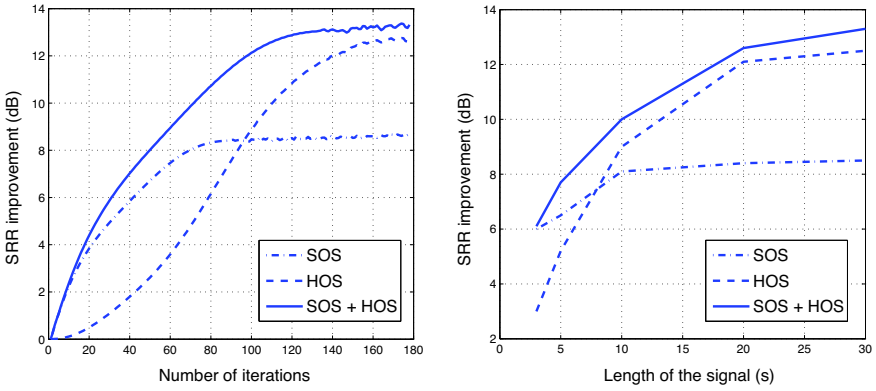


Fig. 10.16 SRR performance of SIMO-based MCBPD for (a) increasing number of offline-iterations, (b) different overall signal lengths

speech signals, the first 50 ms after the main peak of the impulse responses are also added to the contribution of the direct path, i.e., n_Δ is replaced by the so-called critical delay time n_{50}, which is known to contribute to the speech intelligibility [59]. In the following simulation results this perceptual effect is taken into account. The SRR after (10.118) also forms the basis for the definition of the so-called *segmental SRR* (e.g., [70]), which is usually preferred in practice due to the nonstationarity of speech and audio signals and the higher correlation to the quality perceived by auditory measurements. The segmental SRR is based on time-varying local SRR estimates which are obtained by decomposing the signals into K_S segments of length N_S, i.e., the averaging in (10.118) is performed only over these short intervals. The segmental SRR is then defined as the average of the local SRR estimates over the K_S segments. In our simulations, we use $N_S = 320$. This corresponds to the typical stationarity interval for speech (20 ms for a sampling rate of 16 kHz).

Furthermore, in the context of adaptive signal processing, another interesting aspect of the SRR is that formally it corresponds directly to the definition of the so-called *Signal-to-Interference Ratio* (SIR), which is usually used in the literature for the evaluation of signal separation approaches, such as BSS. If we consider the MCBPD optimization criterion, which can also be regarded as a *contrast function* for signal separation and dereverberation, we may hypothesize that in practice, the potential SRR improvement will generally be upper-bounded by the potential SIR improvement in the MIMO case. The same consideration also applies to the segmental SRR and the segmental SIR.

We first consider the direct-inverse approach to SIMO-based dereverberation. Our simulations are based on the coefficient update (10.106) (without the last term in the SIMO case) using the correlation method. We chose $L = 3000$ according to (10.18), the block length $N = N'_L = 320$ corresponding to a stationarity interval of 20 ms, and $n_A = 32$. Figure 10.16 shows the SRR improvement for offline (batch)

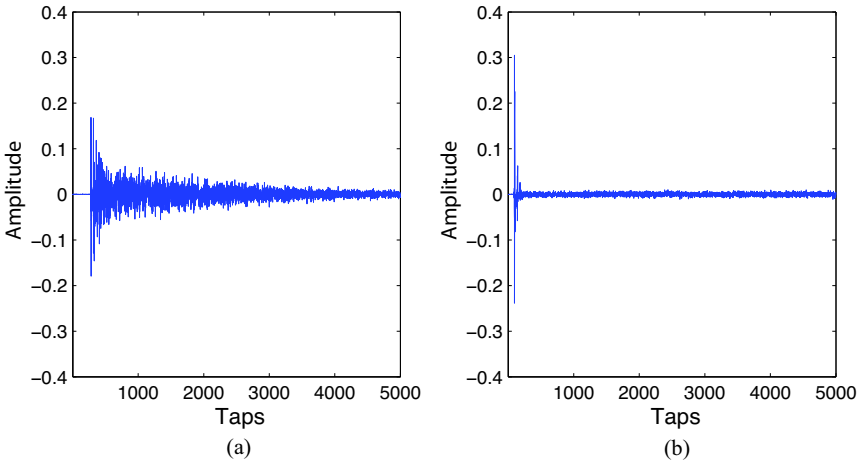


Fig. 10.17 First 5000 taps of (a) one of the measured room impulse responses of the mixing system **H** and (b) impulse response of the overall system **C** after convergence

adaptation, i.e., $\beta(i,m) = \beta(i)$ in (10.39) (and thus $\beta'(i,m) = \beta'(i)$ in (10.106)) corresponds to a rectangular window function over the entire available signal length, and the outer sum in (10.39) and (10.106) turns into a summation of the contributions from all blocks with equal weights. Figure 10.16(a) illustrates the convergence over the number of iterations. We see that the optimization based purely on second-order statistics (SOS, dash-dot line, only the first term in the brackets in (10.106) was used) exhibits a rapid initial convergence, while the kurtosis-based approach (HOS, dashed line, only the second term in the brackets in (10.106) was used) finally achieves a higher level of SRR improvement at the cost of a slower initial convergence. By exploiting all the available statistical signal properties (SOS+HOS, solid line, both terms in the brackets in (10.106) were used), the TRINICON framework combines the advantages of the former two approaches. The higher data requirement for HOS-based estimation is also reflected in Fig. 10.16(b). Here, we performed the offline adaptation for various overall signal lengths. It can be seen that the SOS-based contribution of the optimization already provides reasonable performance for relatively short signal lengths. Hence, in practice, where online adaptation is required due to potential changes of the room impulse responses, the synergy effects provided by TRINICON appear to be attractive.

Figure 10.17 shows the first 5000 taps of one of the room impulse responses of the measured mixing system and of the overall system (i.e., between the source and the MCBPD output) after dereverberation, based on the combined (SOS+HOS) TRINICON approach and 180 iterations with a signal length of 30s (see Fig. 10.16). The same parameters were used for the spectrograms for the first three seconds of the signals in Fig. 10.18. Both representations illustrate a significant enhancement of the speech signals. The spectrograms were computed as sequences of DFTs of windowed data segments. In this example, the Hamming window length was chosen

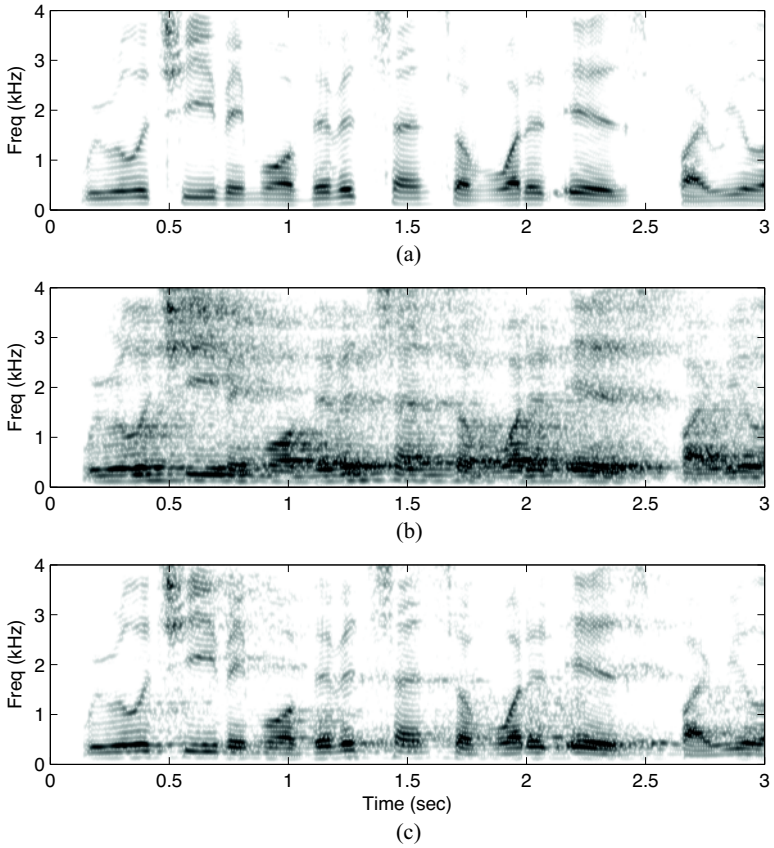


Fig. 10.18 Spectrograms for 0...4 kHz of the first 3 s of (a) original source signal $s(n)$ (b) received signal $x_1(n)$ at microphone 1 and (c) output signal $y(n)$ after convergence

to be 20 ms, as it is typical in speech analysis. This is short enough so that any single 20 ms frame will typically contain data from only one phoneme, yet long enough that it will include at least two periods of the fundamental frequency during voiced speech assuming the lowest voiced pitch to be around 100 Hz.

As mentioned in Sects. 10.2 and 10.3, the correct choice of the filter length is an important issue in blind dereverberation, especially in the application of the identification-and-inversion approach. Hence, we now compare the DI and II approaches with respect to the sensitivity of overestimation of the filter lengths. Note that formally, according to Sect. 10.5.2, the TRINICON-based adaptation algorithm for blind system identification differs only slightly from the corresponding MCBPD algorithm (e.g., (10.105)): the sign of the update term is changed and the relation between the filter coefficients and the estimates of the mixing system, i.e., (10.79), has to be taken into account. Moreover, in the II approach to dereverberation, ad-

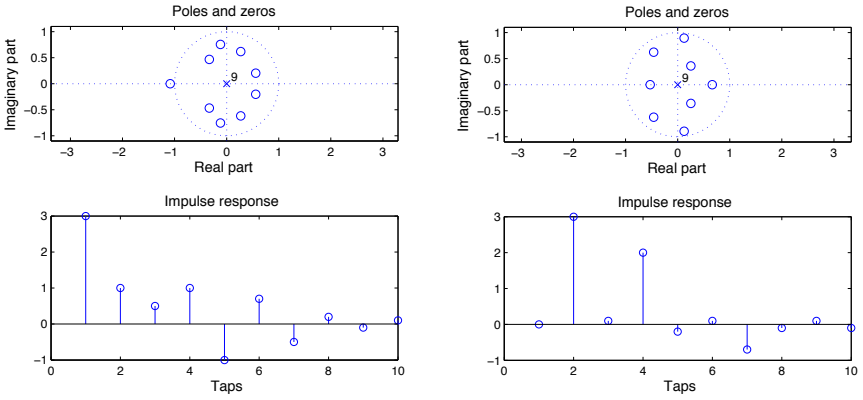


Fig. 10.19 Poles and zeros in the z-domain of subfilters (a) $\mathbf{h}_1$ and (b) $\mathbf{h}_2$ of a simple SIMO mixing system without common zeros

ditionally, the application of MINT (10.17) is required to calculate the demixing system based on the estimated mixing system. These modifications were made in (10.106) for our next experiment comparing the II approach with the DI approach (using (10.106) without modifications).

To allow for a fair comparison between the two different approaches, we assumed the same mixing system with only two sensor channels in both cases. For this experiment, the mixing system was composed of two very simple artificially created impulse responses in order to guarantee the avoidance of common zeros (or even near common zeros), as shown in Fig. 10.19. Hence, as long as the optimal filter length is chosen, this SIMO system is guaranteed to be invertible, which we also confirmed by applying MINT in a supervised manner. Table 10.2 shows the results of the blind estimation in terms of SRR improvement for both the DI and II approaches for different demixing filter lengths, and without any of the additional repair measures mentioned in Sect. 10.3.5. Note that in this experiment we chose n_Δ in the above SRR definition equal to the delay of the main peaks of the impulse responses due to their short lengths. Obviously, the numerical results confirm that with both approaches the best performance is obtained by choosing the optimal filter length according to Sects. 10.2 and 10.3. Moreover, the results clearly show that the direct-inverse approach is significantly more robust to overestimation of the filter length. On the other hand, however, we have to note that the potential applicability of the identification-and-inversion approach is more general because the distinction between the speech production system and the room acoustics is not required in this case.

Table 10.2 Comparison of the DI approach to blind dereverberation with the II approach with respect to the sensitivity of overestimation of the filter length for the simple example $M = 10$, $P = 2$, $L_{\mathrm{DI,opt}} = 9$, $L_{\mathrm{II,opt}} = 10$

	$L \approx 80\% L_{\mathrm{opt}}$	$L = L_{\mathrm{opt}}$	$L \approx 120\% L_{\mathrm{opt}}$	$L \approx 140\% L_{\mathrm{opt}}$	$L \approx 150\% L_{\mathrm{opt}}$
DI	29.8 dB	31.2 dB	27.3 dB	24.1 dB	22.4 dB
II	22.0 dB	25.4 dB	9.6 dB	4.5 dB	0.2 dB

10.7.2 The MIMO Case

Finally, we extend the investigation of MCBPD for the direct-inverse approach to the MIMO case. We again consider the same acoustic scenario with $T_{60} \approx 700$ ms, as described above for the SIMO case. In the following experiment there are two active speakers (one male speaker and one female speaker). The configuration is symmetric with respect to to the linear microphone array. We again apply the coefficient update (10.106) using the correlation method and the same parameter settings as described for the SIMO case. Figure 10.20 shows both the improvement of the signal-to-interference ratio (i.e., source separation at the ouputs) and the improvement of the signal-to-reverberation ratio. The SIR and SRR curves were averaged between the contributions from the two sources. Similar to the SIMO case, TRINICON provides synergies between the SOS-based adaptation and the HOS-based adaptation. This advantage can be seen in both the separation and the dereverberation performances. We also confirm that the SRR improvement is generally upper bounded by the SIR improvement. It is remarkable that the SRR improvements in the MIMO case are only slightly lower than those in the SIMO case. As a reference, we also included the SIR convergence curve of the popular narrowband BSS algorithm after Fancourt and Parra [34], which is based on SOS (see also Sect. 10.5.3). We see that the initial convergence of the rigorously derived broadband approach is well comparable with that of the narrowband algorithm, while the final SIR performance is significantly higher. The reference curve for a pure separation algorithm based on SOS ([17] as a special case of (10.106) with $n_A = L - 1$ according to Fig. 10.14, $N = L$, and using only the first term in the brackets) in the SRR plot, and the comparison with a conventional delay-and-sum beamformer confirms the high efficiency of the MCBPD extension presented in this chapter.

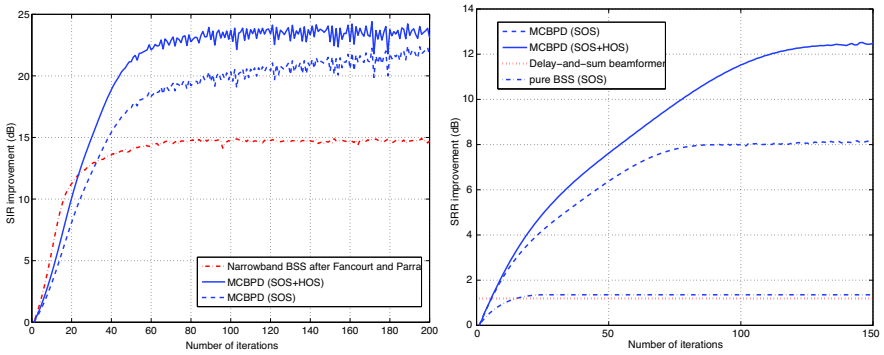


Fig. 10.20 (a) SIR and (b) SRR performance of MIMO-based MCBPD

10.8 Conclusions

Based on the TRINICON framework for broadband adaptive MIMO filtering, in this chapter we developed a strictly analytical top-down approach to the problem of blind dereverberation of speech and audio signals. It was shown that this provides both a common framework for various existing and novel powerful blind dereverberation algorithms and allows for a direct comparison between the various algorithms and the different existing approaches to blind dereverberation.

Comparing the two fundamental approaches to blind dereverberation, i.e., the identification-and-inversion approach and the direct-inverse approach, we can summarize that in principle the II approach is suitable for arbitrary audio signals, however, on the downside, this flexibility with respect to the source signals implies a high sensitivity to overestimation of the optimum filter length and common zeros in z-domain representation of the mixing system paths, so that additional repair mechanisms are necessary. Moreover, the explicit MINT-based inversion of the estimated mixing matrix in the II approach increases the computational complexity. On the other hand, the direct-inverse approach avoids the two-step procedure and the related problems of the II approach, but requires more stringent stochastic model assumptions on the source signals in order to avoid whitening effects. Fortunately, the TRINICON framework inherently allows the incorporation of powerful source models leading to a high separation and dereverberation performance without distortions for signals like speech.

Appendix A: Compact Derivation of the Gradient-based Coefficient Update

For the following compact derivation, we formulate the TRINICON coefficient optimization criterion (10.39) in the following way:

$$\mathcal{J} = \hat{E}_{\text{long}} \left\{ \hat{E}_{\text{block}} \left\{ f\left(\mathbf{y}, \boldsymbol{\mathcal{Q}}^{(1)}, \boldsymbol{\mathcal{Q}}^{(2)}, \ldots\right) \right\} \right\}, \tag{10.119}$$

with

$$f = -\left(\log \hat{p}_{\text{s},PD}(\mathbf{y}) - \log \hat{p}_{\text{y},PD}(\mathbf{y})\right), \tag{10.120}$$

and the operators $\hat{E}_{\text{block}}\{a\} = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1} a(j)$ for averaging within each block, and $\hat{E}_{\text{long}}\{b\} = \sum_{i=0}^{\infty} \beta(i,m) \cdot b(i)$ over multiple blocks depending on the choice of the function β. The set of quantities

$$\boldsymbol{\mathcal{Q}}^{(r)} = \hat{E}_{\text{block}} \left\{ \boldsymbol{\mathcal{G}}^{(r)}(\mathbf{y}) \right\}, \; r = 1, 2, \ldots, \tag{10.121}$$

(where $\boldsymbol{\mathcal{G}}^{(r)}$ are suitable functions of the observation vectors $\mathbf{y}$) contains all stochastic model parameters $\boldsymbol{\mathcal{Q}}_s^{(\cdot)}$ and $\boldsymbol{\mathcal{Q}}_y^{(\cdot)}$ according to (10.41) determining $\hat{p}_{s,PD}(\cdot)$ and $\hat{p}_{y,PD}(\cdot)$, respectively.

The gradient of (10.119) with respect to $\check{\mathbf{W}}$ reads according to (10.43) (omitting the iteration index here for simplicity) as:

$$\Delta\check{\mathbf{W}} = \hat{E}_{\text{long}} \left\{ \mathcal{SC} \left\{ \hat{E}_{\text{block}} \left\{ \frac{\partial}{\partial \mathbf{W}} f\left(\mathbf{y}, \boldsymbol{\mathcal{Q}}^{(1)}, \boldsymbol{\mathcal{Q}}^{(2)}, \ldots\right) \right\} \right\} \right\}. \tag{10.122}$$

We now apply the general multivariate chain rule:

$$\frac{\partial}{\partial \mathbf{W}} f\left(\mathbf{y}, \boldsymbol{\mathcal{Q}}^{(1)}, \boldsymbol{\mathcal{Q}}^{(2)}, \ldots\right) = \sum_i \frac{\partial [\mathbf{y}]_i}{\partial \mathbf{W}} \frac{\partial f}{\partial [\mathbf{y}]_i} + \sum_r \sum_{i_1, i_2, \ldots} \frac{\partial \mathcal{Q}_{i_1,i_2,\ldots}^{(r)}}{\partial \mathbf{W}} \frac{\partial f}{\partial \mathcal{Q}_{i_1,i_2,\ldots}^{(r)}}, \tag{10.123}$$

where $\mathcal{Q}_{i_1,i_2,\ldots}^{(r)}$ denote the elements of $\boldsymbol{\mathcal{Q}}^{(r)}$. Analogously $\mathcal{G}_{i_1,i_2,\ldots}^{(r)}$ denote the elements of $\boldsymbol{\mathcal{G}}^{(r)}$. The derivatives in the second term with respect to $\mathbf{W}$ can be expressed as

$$\frac{\partial \mathcal{Q}_{i_1,i_2,\ldots}^{(r)}}{\partial \mathbf{W}} = \hat{E}_{\text{block}} \left\{ \frac{\partial}{\partial \mathbf{W}} \mathcal{G}_{i_1,i_2,\ldots}^{(r)}(\mathbf{y}) \right\} = \hat{E}_{\text{block}} \left\{ \sum_i \frac{\partial \mathcal{G}_{i_1,i_2,\ldots}^{(r)}}{\partial [\mathbf{y}]_i} \frac{\partial [\mathbf{y}]_i}{\partial \mathbf{W}} \right\}. \tag{10.124}$$

With the MIMO relation $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ and with (10.124) we obtain[8] from (10.123)

$$\frac{\partial}{\partial \mathbf{W}} f\left(\mathbf{y}, \boldsymbol{\mathcal{Q}}^{(1)}, \boldsymbol{\mathcal{Q}}^{(2)}, \ldots\right) = \mathbf{x}\frac{\partial f}{\partial \mathbf{y}^T} + \sum_r \sum_{i_1, i_2, \ldots} \hat{E}_{\text{block}} \left\{ \mathbf{x} \frac{\partial \mathcal{G}_{i_1,i_2,\ldots}^{(r)}}{\partial \mathbf{y}^T} \right\} \frac{\partial f}{\partial \mathcal{Q}_{i_1,i_2,\ldots}^{(r)}}. \tag{10.125}$$

By introducing this equation into (10.122), we obtain

[8] Since in element-wise formulation, $[\mathbf{y}]_i = \sum_\ell [\mathbf{x}]_\ell [\mathbf{W}]_{\ell i}$, we obtain $\frac{\partial [\mathbf{y}]_i}{\partial [\mathbf{W}]_{jk}} = \sum_\ell [\mathbf{x}]_\ell \delta_{j\ell}\delta_{ki} = [\mathbf{x}]_j \delta_{ki}$, and thus $\left[\sum_i \frac{\partial [\mathbf{y}]_i}{\partial [\mathbf{W}]_{jk}} \frac{\partial f}{\partial [\mathbf{y}]_i}\right] = \left[\sum_i [\mathbf{x}]_j \delta_{ki} \frac{\partial f}{\partial [\mathbf{y}]_i}\right] = \left[[\mathbf{x}]_j \frac{\partial f}{\partial [\mathbf{y}]_k}\right] = \mathbf{x}\frac{\partial f}{\partial \mathbf{y}^T}$.

$$\begin{aligned}\Delta\breve{\mathbf{W}} &= \hat{E}_{\text{long}}\left\{\mathcal{SC}\left\{\hat{E}_{\text{block}}\left\{\mathbf{x}\frac{\partial f}{\partial \mathbf{y}^T}\right\}\right.\right.\\ &\quad\left.\left.+\sum_r\sum_{i_1,i_2,\ldots}\hat{E}_{\text{block}}\left\{\mathbf{x}\frac{\partial \mathcal{G}^{(r)}_{i_1,i_2,\ldots}}{\partial \mathbf{y}^T}\right\}\hat{E}_{\text{block}}\left\{\frac{\partial f}{\partial \mathcal{Q}^{(r)}_{i_1,i_2,\ldots}}\right\}\right\}\right\}\\ &= \hat{E}_{\text{long}}\left\{\mathcal{SC}\left\{\hat{E}_{\text{block}}\left\{\mathbf{x}\left(\frac{\partial f}{\partial \mathbf{y}^T}\right.\right.\right.\right.\\ &\quad\left.\left.\left.\left.+\sum_r\sum_{i_1,i_2,\ldots}\frac{\partial \mathcal{G}^{(r)}_{i_1,i_2,\ldots}}{\partial \mathbf{y}^T}\hat{E}_{\text{block}}\left\{\frac{\partial f}{\partial \mathcal{Q}^{(r)}_{i_1,i_2,\ldots}}\right\}\right)\right\}\right\}\right\}. \end{aligned} \tag{10.126}$$

With (10.120) the last expression finally leads to the gradient-based update (10.44).

Appendix B: Transformation of the Multivariate Output Signal PDF in (10.39) by Blockwise Sylvester Matrix

Due to the linear MIMO relation

$$\mathbf{y}^{\mathrm{T}}(n) = \mathbf{x}^{\mathrm{T}}(n)\mathbf{W}(n), \tag{10.127}$$

from (10.31) we express the PD-variate output log-likelihood $\log(\hat{p}_{\mathbf{y},PD}(\mathbf{y}(n)))$ in (10.39) in terms of the $2PL\times PD$ MIMO coefficient matrix $\mathbf{W}$ and the corresponding multivariate input PDF.

Since in general, $\mathbf{W}$ is not quadratic ($D \leq L$), we cannot immediately apply the well-known relation between the PDFs of two linearly related vectors via the determinant of a quadratic mapping matrix [73]. However, in our case $2PL > PD$, i.e., for 'tall' matrices $\mathbf{W}$ we can form a joint PDF $\hat{p}_{\mathbf{y}\tilde{\mathbf{x}},2LP}(\mathbf{y}(n),\tilde{\mathbf{x}}(n))$ of the output vector $\mathbf{y}$ and certain elements $\tilde{\mathbf{x}}$ of the input vector $\mathbf{x}$ so that this joint PDF exhibits the same dimensionality as the input PDF $\hat{p}_{\mathbf{x},2LP}(\mathbf{x}(n))$. Then, after the transformation

$$\hat{p}_{\mathbf{y}\tilde{\mathbf{x}},2LP}(\mathbf{y}(n),\tilde{\mathbf{x}}(n)) = \frac{\hat{p}_{\mathbf{x},2LP}(\mathbf{x}(n))}{\left|\det\tilde{\mathbf{W}}\right|}, \tag{10.128}$$

with a quadratic $2LP\times 2LP$ matrix $\tilde{\mathbf{W}}$, the desired multivariate output PDF $\hat{p}_{\mathbf{y},PD}(\mathbf{y}(n))$ is obtained without loss of generality as a marginal density by integration for $\tilde{\mathbf{x}}(n)$ [73].

In our application a *channel-wise* extension of matrix $\mathbf{W}$ is desirable so that the MIMO relation (10.127)

$$\left[\mathbf{y}_1^{\mathrm{T}},\ldots,\mathbf{y}_P^{\mathrm{T}}\right] = \left[\mathbf{x}_1^{\mathrm{T}},\ldots,\mathbf{x}_P^{\mathrm{T}}\right]\begin{bmatrix}\mathbf{W}_{11} & \cdots & \mathbf{W}_{1P}\\ \vdots & \ddots & \vdots\\ \mathbf{W}_{P1} & \cdots & \mathbf{W}_{PP}\end{bmatrix}$$

may be extended to

$$\left[\mathbf{y}_1^{\mathrm{T}}, \tilde{\mathbf{x}}_1^{\mathrm{T}}, \ldots, \mathbf{y}_P^{\mathrm{T}}, \tilde{\mathbf{x}}_P^{\mathrm{T}}\right] = \left[\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_P^{\mathrm{T}}\right] \tilde{\mathbf{W}}, \tag{10.129}$$

where $\tilde{\mathbf{x}}_p$, $p = 1, \ldots, P$ denote vectors containing the $2L-D$ last elements of $\mathbf{x}_p$ and

$$\tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W}_{11} & \begin{bmatrix} \mathbf{0}_{D\times 2L-D} \\ \mathbf{I}_{2L-D\times 2L-D} \end{bmatrix} & \cdots & \mathbf{W}_{1P} & \mathbf{0}_{2L\times 2L-D} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{W}_{P1} & \mathbf{0}_{2L\times 2L-D} & \cdots & \mathbf{W}_{PP} & \begin{bmatrix} \mathbf{0}_{D\times 2L-D} \\ \mathbf{I}_{2L-D\times 2L-D} \end{bmatrix} \end{bmatrix}. \tag{10.130}$$

With (10.128) we obtain

$$\begin{aligned} \hat{p}_{\mathbf{y},PD}(\mathbf{y}(n)) &= \frac{1}{\left|\det\tilde{\mathbf{W}}\right|} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{p}_{\mathbf{x},2LP}(\mathbf{x}(n)) \mathrm{d}\tilde{\mathbf{x}}_1 \cdot \ldots \cdot \mathrm{d}\tilde{\mathbf{x}}_P \\ &= \frac{1}{\left|\det\tilde{\mathbf{W}}\right|} \hat{p}_{\mathbf{x}_{PD},PD}(\mathbf{x}_{PD}(n)), \end{aligned} \tag{10.131}$$

which leads to the following simple expression for the desired log-likelihood:

$$\log \hat{p}_{\mathbf{y},PD}(\mathbf{y}(n)) = \log \hat{p}_{\mathbf{x}_{PD},PD}(\mathbf{x}_{PD}(n)) - \log\left|\det\tilde{\mathbf{W}}\right|. \tag{10.132}$$

Since the first term on the right hand-side of (10.132) does not depend on the filter coefficients, it does not need to be considered further for the gradient of the optimization criterion (10.39). To simplify the important second term in (10.132) together with $\tilde{\mathbf{W}}$ from (10.130) we exploit the fact that we can exchange colums or rows of $\tilde{\mathbf{W}}$ without changing the value of $|\det\tilde{\mathbf{W}}|$. Application of the general matrix relation

$$\det\begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{A}_2 & \mathbf{I} \end{bmatrix} = \det\mathbf{A}_1 \tag{10.133}$$

immediately leads then to the compact formulation

$$\log \hat{p}_{\mathbf{y},PD}(\mathbf{y}(n)) = \log \hat{p}_{\mathbf{x}_{PD},PD}(\mathbf{x}_{PD}(n)) - \log\left|\det\left\{\mathbf{V}^{\mathrm{T}}\mathbf{W}\right\}\right|, \tag{10.134}$$

with the window matrix $\mathbf{V}$ defined in (10.46). Note that $\mathbf{V}^{\mathrm{T}}\mathbf{W}$ is only of dimension $DP \times DP$.

Appendix C: Polynomial Expansions for Nearly Gaussian Probability Densities

Orthogonal Polynomials

Let I be a finite or infinite interval and $r(x)$ be a continuous and positive function (which we here call a *weighting function*) on the interval such that $\int_I f(x)r(x)\mathrm{d}x$ exists for every *polynomial* $f(x)$. Then there is a unique set of polynomials $P_n(x)$, $n = 0, 1, ...,$ of order n such that

$$\int_I P_k(x)P_n(x)r(x)\mathrm{d}x := \langle P_k, P_n\rangle_r = c_n\,\delta_{kn} \tag{10.135}$$

with a predefined constant c_n. These polynomials $P_n(x)$ are called *orthogonal polynomials*. The operation $\langle\cdot,\cdot\rangle_r$ denotes the *inner product* in the vector space of the polynomials.

An important class of orthogonal polynomials in our context are the so-called Chebyshev–Hermite polynomials $P_{\mathrm{H},n}(x)$, which are specified by $I = (-\infty,\infty)$, the weighting function $r(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$, and $c_n = n!$, e.g., [56].

For the orthogonal polynomials considered here there is an important *proposition* stating that they even form a basis in a Hilbert space so that any quadratically integrable function $f(x)$ with respect to $r(x)$ on I can be expressed by the expansion, e.g., [56]

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{c_n}\,\langle f, P_n\rangle_r\; P_n(x). \tag{10.136}$$

Polynomial Expansion for Univariate Densities

The two different expansions that are usually used to obtain a parameterized representation of nearly Gaussian probability density functions are the Edgeworth and the Gram–Charlier expansions, e.g., [51]. They lead to very similar approximations, so in this chapter we only consider the Gram–Charlier expansion. These expansions are based on the above-mentioned Chebyshev–Hermite polynomials $P_{\mathrm{H},n}(x)$.

Let $p(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{x^2}{2\sigma^2}}\,\tilde{p}\left(\frac{x}{\sigma}\right)$ represent an arbitrary univariate probability density, where $\tilde{p}(\cdot)$ contains the higher-order contributions. According to (10.136) the higher-order statistics contribution $\tilde{p}$ can readily be expanded as

$$\tilde{p}(x) = \sum_{n=0}^{\infty} a_n P_{\mathrm{H},n}(x), \tag{10.137a}$$

$$a_n = \frac{1}{n!}\int_{-\infty}^{\infty} \tilde{p}(x')P_{\mathrm{H},n}(x')\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x'^2/2}\mathrm{d}x'. \tag{10.137b}$$

Hence, the complete density function $p(x)$ is finally expressed as

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \sum_{n=0}^{\infty} a_n P_{\mathrm{H},n}\left(\frac{x}{\sigma}\right). \tag{10.138a}$$

The coefficients a_n after (10.137b) can be compactly written using the expectation operator:

$$a_n = \frac{1}{n!} E\left\{P_{\mathrm{H},n}\left(\frac{x}{\sigma}\right)\right\}. \tag{10.138b}$$

Example: Fourth-order Approximation for a Zero-mean Process

To obtain explicit expressions for the coefficients (10.138b), the Chebyshev–Hermite can be calculated using the derivatives of the standardized Gaussian probability density function (corresponding to the weighting function $r(x)$):

$$P_{\mathrm{H},n}(x) = (-1)^n \frac{1}{r(x)} \frac{\partial^n r(x)}{\partial x^n}, \tag{10.139}$$

so that $P_{\mathrm{H},0}(x) = 1$, $P_{\mathrm{H},1}(x) = x$, $P_{\mathrm{H},2}(x) = x^2 - 1$, $P_{\mathrm{H},3}(x) = x^3 - 3x$, $P_{\mathrm{H},4}(x) = x^4 - 6x^2 + 3$. The resulting expansion coefficients for zero-mean processes are $a_0 = 1$, $a_1 = a_2 = 0$, $a_3 = \frac{E\{x^3\}}{3!\sigma^3}$, $a_4 = \frac{1}{4!}\left(\frac{E\{x^4\}}{\sigma^4} - 3\right)$, so that

$$p(x) \approx \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \left(1 + \frac{\kappa_3}{3!\,\sigma^3} P_{\mathrm{H},3}\left(\frac{x}{\sigma}\right) + \frac{\kappa_4}{4!\,\sigma^4} P_{\mathrm{H},4}\left(\frac{x}{\sigma}\right)\right), \tag{10.140}$$

with [71] the *skewness* $\kappa_3 = E\{x^3\}$ and the *kurtosis* $\kappa_4 = E\{x^4\} - 3\sigma^4$. In the context of higher-order statistics-based estimation the kurtosis plays a particularly prominent role since it indicates whether a PDF is supergaussian ($\kappa_4 > 0$) or subgaussian ($\kappa_4 < 0$).

Multivariate Orthogonal Polynomials

Based on the previous section we may now generalize the Gram–Charlier expansion to multivariate probability density functions for a vector $\mathbf{x}$ of length D.

We formulate the orthogonality relation analogously to (10.135),

$$\int_{I^D} P_{\mathbf{k}}(\mathbf{x}) P_{\mathbf{n}}(\mathbf{x}) r(\mathbf{x}) \mathrm{d}\mathbf{x} = c_{\mathbf{n}}\, \delta_{\mathbf{kn}}, \tag{10.141}$$

and the inner product

$$\langle f, g \rangle_r := \int_{I^D} f(\mathbf{x}) g(\mathbf{x}) r(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{10.142}$$

The D-variate Chebyshev–Hermite polynomials are specified by the D-variate weighting function [84]

$$r(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D}} \mathrm{e}^{-\|\mathbf{x}\|_2^2/2} = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_i^2/2} = \prod_{i=1}^{D} r_1(x_i). \tag{10.143}$$

As we can see, in this case we have a *product weighting function*. It can be shown [84] that this has the very advantageous consequence that it also leads to corresponding *product polynomials*

$$P_{\mathbf{n}}(\mathbf{x}) = \prod_{i=1}^{D} P_{i,n_i}(x_i). \tag{10.144}$$

Note that $\mathbf{n}$ denotes a vector of indices n_i, $i = 1, \ldots, D$. The expansion of a multivariate function $f(\mathbf{x})$ is then given as

$$f(\mathbf{x}) = \sum_{\mathbf{n}=\mathbf{0}}^{\infty} \frac{1}{c_{\mathbf{n}}} \langle f, P_{\mathbf{n}} \rangle_r \, P_{\mathbf{n}}(\mathbf{x}). \tag{10.145}$$

Polynomial Expansion for Multivariate Densities

Let $p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D \det \mathbf{R}_{\mathbf{xx}}}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^T \mathbf{R}_{\mathbf{xx}}{}^{-1} \mathbf{x}} \tilde{p}\left(\mathbf{L}^{-1}\mathbf{x}\right)$ represent an arbitrary D-variate probability density, where $\tilde{p}(\cdot)$ again contains the higher-order contributions, and $\mathbf{L}$ is obtained by the Cholesky decomposition $\mathbf{R}_{\mathbf{xx}} = \mathbf{L}^T\mathbf{L}$ (note that $\sqrt{\mathbf{x}^T \mathbf{R}_{\mathbf{xx}}{}^{-1} \mathbf{x}} = \|\mathbf{L}^{-1}\mathbf{x}\|_2$).

In the same way as in the univariate case, we now obtain the following representation of a multivariate probability density function $p(\mathbf{x})$:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D \det \mathbf{R}_{\mathbf{xx}}}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^T \mathbf{R}_{\mathbf{xx}}{}^{-1} \mathbf{x}} \sum_{\mathbf{n}=\mathbf{0}}^{\infty} a_{\mathbf{n}} P_{\mathrm{H},\mathbf{n}}\left(\mathbf{L}^{-1}\mathbf{x}\right), \tag{10.146a}$$

with the coefficients

$$a_{\mathbf{n}} = \frac{1}{\prod_{i=1}^{D} n_i!} E\left\{P_{\mathrm{H},\mathbf{n}}(\mathbf{L}^{-1}\mathbf{x})\right\}. \tag{10.146b}$$

Note that $P_{\mathrm{H},\mathbf{n}}(\cdot)$ in (10.146a) and (10.146b) is given by (10.144).

Appendix D: Expansion of the Sylvester Constraints in (10.83)

We consider here an expression with the Sylvester Constraint for one channel of the form

$$\mathbf{a}^T \mathcal{SC}\left\{\mathbf{b}\mathbf{c}^T\right\},$$

where **a**, **b**, **c** denote column vectors of length L, $2L$, and D, respectively. With the explicit expression of the generic Sylvester constraint for one channel after Fig. 10.6 and [19],

$$[\mathbf{w}]_m = \sum_{k=1}^{2L} \sum_{\ell=1}^{D} [\mathbf{W}]_{k\ell}\, \delta_{k,(m+\ell-1)},$$

where δ_{ij} denotes the Kronecker symbol, the above expression reads as

$$\sum_{m=1}^{L} a_m \sum_{k=1}^{2L} \sum_{\ell=1}^{D} b_k c_\ell \delta_{k,(m+\ell-1)} = \sum_{\ell=1}^{D} \sum_{m=1}^{L} a_m b_{m+\ell-1} c_\ell. \tag{10.147}$$

From the linearity of the operations, we easily deduce

$$\begin{aligned} &\mathbf{a}_1^T \mathcal{SC}\left\{\mathbf{b}_1 \mathbf{c}^T\right\} + \mathbf{a}_2^T \mathcal{SC}\left\{\mathbf{b}_2 \mathbf{c}^T\right\} \\ &= \sum_{\ell=1}^{D} \left(\sum_{m=1}^{L} a_{1,m} b_{1,m+\ell-1} + \sum_{m=1}^{L} a_{2,m} b_{2,m+\ell-1} \right) c_\ell. \end{aligned} \tag{10.148}$$

References

1. Aichner, R., Buchner, H., Kellermann, W.: On the causality problem in time-domain blind source separation and deconvolution algorithms. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 5, pp. 181–184. Philadelphia, PA, USA (2005)
2. Aichner, R., Buchner, H., Kellermann, W.: Exploiting narrowband efficiency for broadband convolutive blind source separation. EURASIP J. on App. Signal Process. **2007**(Article ID 16381) (2007). DOI doi:10.1155/2007/16381
3. Aichner, R., Buchner, H., Yan, F., Kellermann, W.: A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. Signal Processing **86**(6), 1260–1277 (2006)
4. Amari, S., Douglas, S.C., Cichocki, A., Yang, H.H.: Multichannel blind deconvolution and equalization using the natural gradient. In: Proc. IEEE Int. Workshop Signal Processing Advances in Wireless Communications, pp. 101–107 (1997)
5. Amari, S., Kawanabe, M.: Information geometry of estimating functions in semiparametric statistical models. Bernoulli **2**(3), 29–54 (1996)
6. Araki, S., Makino, S., Mukai, R., Hinamoto, Y., Nishikawa, T., Saruwatari, H.: Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 1785–1788. Orlando, USA (2002)
7. Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency-domain blind source separation for convolutive mixtures of speech. IEEE Trans. Speech Audio Process. **11**(2), 109–116 (2003)
8. Benesty, J.: Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J. Acoust. Soc. Am. **107**(1), 384–391 (2000)
9. Bobillet, W., Grivel, E., Guidorzi, R., Najim, M.: Cancelling convolutive and additive coloured noises for speeech enhancement. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 777–780. Montreal, Canada (2004)
10. Brandstein, M.S.: On the use of explicit speech modeling in microphone array applications. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 6, pp. 3613–3616. Seattle, WA, USA (1998)

11. Brehm, H., Stammler, W.: Description and generation of spherically invariant speech-model signals. Signal Processing **12**(2), 119–141 (1987)
12. Broadhead, M.K., Pflug, L.A.: Performance of some sparseness criterion blind deconvolution methods in the presence of noise. J. Acoust. Soc. Am. **102**(2), 885–893 (2000)
13. Buchner, H., Aichner, R., Kellermann, W.: Blind source separation for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 223–226. Kyoto, Japan (2003)
14. Buchner, H., Aichner, R., Kellermann, W.: A generalization of a class of blind source separation algorithms for convolutive mixtures. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA). Nara, Japan (2003)
15. Buchner, H., Aichner, R., Kellermann, W.: Blind source separation for convolutive mixtures: A unified treatment. In: Y. Huang, J. Benesty (eds.) Audio signal processing for next-generation multimedia communication systems. Kluwer Academic Publishers (2004)
16. Buchner, H., Aichner, R., Kellermann, W.: TRINICON: A versatile framework for multichannel blind signal processing. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 889–892. Montreal, Canada (2004)
17. Buchner, H., Aichner, R., Kellermann, W.: A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. Speech Audio Process. **13**(1), 120–134 (2005)
18. Buchner, H., Aichner, R., Kellermann, W.: Relation between blind system identification and convolutive blind source separation. In: Proc. Workshop Hands-Free Speech Communication and Microphone Arrays (HSCMA), pp. d–3–d–4. Piscataway, NJ, USA (2005)
19. Buchner, H., Aichner, R., Kellermann, W.: TRINICON-based blind system identification with application to multiple-source localization and separation. In: S. Makino, T.W. Lee, S. Sawada (eds.) Blind speech separation, pp. 101–147. Springer (2007)
20. Buchner, H., Aichner, R., Kellermann, W.: The TRINICON framework for adaptive MIMO signal processing with focus on the generic Sylvester constraint. In: Proc. ITG Conf. on Speech Communication. Aachen, Germany (2008)
21. Buchner, H., Aichner, R., Stenglein, J., Teutsch, H., Kellermann, W.: Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), III-97–III-100. Philadelphia, USA (2005)
22. Buchner, H., Benesty, J., Gänsler, T., Kellermann, W.: Robust extended multidelay filter and double-talk detector for acoustic echo cancellation. IEEE Trans. Audio, Speech, Lang. Process. **14**(5), 1633–1644 (2006)
23. Buchner, H., Kellermann, W.: A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation. In: Proc. Workshop Hands-Free Speech Communication and Microphone Arrays (HSCMA), pp. 17–20. Trento, Italy (2008)
24. Burgess, J.C.: Active adaptive sound control in a duct: A computer simulation. J. Acoust. Soc. Am. **70**(3), 715–726 (1981)
25. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non gaussian signals. IEE Proc.-F **140**, 362–370 (1993)
26. Cardoso, J.F., Souloumiac, A.: Jacobi angles for simultaneous diagonalization. SIAM J. Mat. Anal. Appl. **17**(1), 161–164 (1996)
27. Chen, J., Huang, Y., Benesty, J.: Time delay estimation. In: Y. Huang, J. Benesty (eds.) Audio signal processing for next-generation multimedia communication systems, pp. 197–227. Kluwer Academic Publishers (2004)
28. Choi, S., Amari, S., Cichocki, A., Liu, R.: Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 371–376. Aussois, France (1999)
29. Cover, T., Thomas, J.: Elements of information theory. Wiley & Sons (1991)
30. Douglas, S., Sawada, H., Makino, S.: A causal frequency-domain implementation of a natural gradient multichannel blind deconvolution and source separation algorithms. In: Proc. Int. Congr. on Acoustics, vol. 1, pp. 85–88. Kyoto, Japan (2004)

31. Douglas, S., Sawada, H., Makino, S.: Natural gradient multichannel blind deconvolution and source separation using causal FIR filters. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 5, pp. 447–480. Montreal, Canada (2004)
32. Douglas, S.C.: Blind separation of acoustic signals. In: M.S. Brandstein, D.B. Ward (eds.) Microphone arrays: Signal processing techniques and applications, pp. 355–380. Springer (2001)
33. Duda, R.O., Hart, P.E.: Pattern classification and scene analysis, 2nd edn. Wiley & Sons, New York (1973)
34. Fancourt, C.L., Parra, L.: The coherence function in blind source separation of convolutive mixtures of nonstationary signals. In: Proc. Int. Workshop Neural Networks Signal Processing (NNSP), pp. 303–312 (2001)
35. Furuya, K.: Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse-filtering theorem (MINT). In: Proc. Int. Workshop Hands-Free Speech Communication (HSC), pp. 59–62. Kyoto, Japan (2001)
36. Furuya, K., Kaneda, Y.: Two-channel blind deconvolution of nonminimum phase FIR systems. IEICE Trans. Fundamentals **E80-A**(5), 804–808 (1997)
37. Furuya, K., Kataoka, A.: Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. IEEE Trans. Audio, Speech, Lang. Process. **15**(5), 1579–1591 (2007)
38. Gannot, S., Moonen, M.: Subspace methods for multimicrophone speech dereverberation. EURASIP J. on App. Signal Process. **2003**(11), 1074–1090 (2003)
39. Gänsler, T., Gay, S.L., Sondhi, M.M., Benesty, J.: Double-talk robust fast converging algorithms for network echo cancellation. IEEE Trans. Audio, Speech, Lang. Process. **8**(6), 656–663 (2000)
40. Gillespie, B., Atlas, L.: Strategies for improving audible quality and speech recognition accuracy of reverberant speech. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. I–676–I–679. Hongkong, China (2003)
41. Gillespie, B.W., Malvar, H.S., Florêncio, D.A.F.: Speech dereverberation via maximum-kurtosis subband adaptive filtering. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 6, pp. 3701–3704. Salt Lake City, UT, USA (2001)
42. Goldman, J.: Detection in the presence of spherically symmetric random vectors. IEEE Trans. Inf. Theory **22**(1), 52–59 (1976)
43. Gürelli, M.I., Nikias, C.L.: EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. IEEE Trans. Signal Process. **43**(1), 134–149 (1995)
44. Harville, D.A.: Matrix algebra from a statistician's perspective. Springer (1997)
45. Haykin, S.: Adaptive filter theory, fourth edn. Prentice–Hall (2002)
46. Hikichi, T., Miyoshi, M.: Blind algorithm for calculating the common poles based on linear prediction. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 89–92. Montreal, Canada (2004)
47. Hiroe, A.: Solution of permutation problem in frequency domain ICA using multivariate probability density functions. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 601–608 (2006)
48. Hofbauer, M.: Optimal linear separation and deconvolution of acoustical convolutive mixtures. Ph.D. thesis, Swiss Federal Institute of Technology (2005)
49. Huang, Y., Benesty, J., Chen, J.: Separation and dereverberation of speech signals with multiple microphones. In: J. Benesty, S. Makino, J. Chen (eds.) Speech Enhancement, pp. 271–298. Springer (2005)
50. Huber, P.J.: Robust statistics. Wiley (1981)
51. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley-Interscience (2001)
52. Ikeda, S., Murata, N.: An approach to blind source separation of speech signals. In: Proc. Int. Symp. on Nonlinear Theory and its Applications. Crans-Montana, Switzerland (1998)
53. Ikram, M.Z., Morgan, D.R.: Exploring permutation inconsistency in blind separation of speech signals in a reverberant environments. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 1041–1044. Istanbul, Turkey (2000)

54. Joho, M., Schniter, P.: Frequency domain realization of a multichannel blind deconvolution algorithms based on the natural gradient. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 15–26. Nara, Japan (2003)
55. Kawamoto, M., Matsuoka, K., Ohnishi, N.: A method of blind separation for convolved non-stationary signals. Neurocomputing **22**(1), 157–171 (1998)
56. Kendal, M.G., Stuart, A.: The Advanced Theory of Statistics, vol. 1, 2nd edn. Hafner Publishing Company (1963)
57. Kim, T., Eltoft, T., Lee, T.W.: Independent vector analysis: an extension of ICA to multivariate components. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 165–172 (2006)
58. Kleijn, W.B., Paliwal, K.K. (eds.): Speech coding and synthesis. Elsevier Science (1995)
59. Kuttruff, H.: Room acoustics, 4th edn. Spon Press (2000)
60. Lambert, R.H.: Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures. Ph.D. thesis, Univ. of Southern California, Los Angeles, CA, USA (1996)
61. Liu, H., Xu, G., Tong, L.: A deterministic approach to blind identification of multi-channel FIR systems. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 581–584 (1994)
62. Ljung, L.: System identification: Theory for the user. Prentice-Hall (1987)
63. Lombard, A., Rosenkranz, T., Buchner, H., Kellermann, W.: Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 233–236. Taipei, Taiwan (2009)
64. Makino, S., Lee, T.W., Sawada, S. (eds.): Blind speech separation. Springer (2007)
65. Mardia, K.: Measures of multivariate skewness and kurtosis with applications. Biometrika **57**(3), 519–530 (1970)
66. Markel, J.D., Gray, A.H.: Linear prediction of speech, 3rd edn. Springer (1976)
67. Matsuoka, K., Nakashima, S.: Minimal distortion principle for blind source separation. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 722–727. San Diego, CA, USA (2001)
68. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
69. Molgedey, L., Schuster, H.G.: Separation of a mixture of independent signals using time delayed correlations. Phys. Rev. Lett. **72**, 3634–3636 (1994)
70. Naylor, P.A., Gaubitch, N.D.: Speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC). Eindhoven, The Netherlands (2005)
71. Nikias, C.L., Mendel, J.M.: Signal processing with higher-order spectra. IEEE Signal Process. Mag. **10**(3), 10–37 (1993)
72. Nishikawa, T., Saruwatari, H., Shikano, K.: Comparison of time-domain ICA, frequency-domain ICA and multistage ICA for blind source separation. In: Proc. European Signal Processing Conf. (EUSIPCO), vol. 2, pp. 15–18 (2002)
73. Papoulis, A.: Probability, random variables, and stochastic processes, 3rd edn. McGraw–Hill (1991)
74. Parra, L., Spence, C.: Convolutive blind source separation of non-stationary sources. IEEE Trans. Speech Audio Process. **8**(3), 320–327 (2000)
75. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice–Hall (1993)
76. Reichardt, W., Alim, A., Schmidt, W.: Definition und Messgrundlage eines objektiven Masses zur Ermittlung der Grenze zwischen brauchbarer und unbrauchbarer Durchsichtigkeit bei Musikdarbietung. Acoustica **32**, 126–137 (1975)
77. Santamaria, I., Via, J., C.C.Gaudes: Robust blind identification of simo channels: a support vector regression approach. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 5, pp. 673–676. Montreal, Canada (2004)
78. Sawada, H., Mukai, R., de Ryhove, S.K., Araki, S., Makino, S.: Spectral smoothing for frequency-domain blind source separation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 311–314. Kyoto, Japan (2003)

79. Schobben, D.W.E., Sommen, P.C.W.: A frequency-domain blind signal separation method based on decorrelation. IEEE Trans. Signal Process. **50**(8), 1855–1865 (2002)
80. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neurocomputing **22**, 21–34 (1998)
81. Stone, J.V.: Blind deconvolution using temporal predictability. Neurocomputing **49**, 79–86 (2002)
82. Wiggins, R.A.: Minimum entropy deconvolution. Geoexploration **16**, 21–35 (1978)
83. Wu, H.C., Principe, J.C.: Simultaneous diagonalization in the frequency domain (SDIF) for source separation. In: Proc. Int. Sypm. on Independent Component Analysis and Blind Signal Separation (ICA), pp. 245–250 (1999)
84. Xu, Y.: Lecture notes on orthogonal polynomials of several variables. In: Advances in the theory of spectral functions and orthogonal polynomials, vol. 2, pp. 135–188. Nova Science Publishers, Hauppage, NY (2004)
85. Yao, K.: A representation theorem and its applications to spherically-invariant random processes. IEEE Trans. Inf. Theory **19**(5), 600–608 (1973)
86. Yegnanarayana, B., Murthy, P.S.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Speech Audio Process. **8**(3), 267–281 (2000)
87. Yoshioka, T., Hikichi, T., Miyoshi, M.: Dereverberation by using time-variant nature of speech production systems. EURASIP J. Advances in Signal Process. **2007** (2007)
88. Zhang, L.Q., Cichocki, A., Amari, S.I.: Geometrical structures of FIR manifold and their application to multichannel blind deconvolution. In: Proc. IEEE Int. Workshop Neural Networks for Signal Processing (NNSP), pp. 303–312. Madison, WI, USA (1999)

Index