Blind Identification of Acoustic Systems and Enhancement of Reverberant Speech

by
Nikolay Dian Gaubitch
MEng (Hons)

A Thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy of University of London and
Diploma of Imperial College

Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London
University of London
2007

All happenings are in the mind.

Whatever happens in all minds, truly happens.

- George Orwell (1903-1950), "Nineteen Eighty-Four"

Abstract

Speech signals obtained in rooms with microphones positioned at a distance from the talker are degraded in quality due to additive noise and reverberation, where the latter arises from multiple reflections from the surrounding walls and objects. Reverberation causes speech to sound distant and spectrally distorted and can also reduce intelligibility. Therefore, dereverberation is an important speech enhancement process for hands-free terminals.

In this thesis, dereverberation techniques are categorized into beamforming, speech enhancement and blind system identification/inversion. Two algorithms, one from each of the latter two categories, are proposed and evaluated. First, it is shown that the autoregressive coefficients of clean speech can be estimated accurately from reverberant multichannel observations and that reverberation mainly affects the predication residual. Consequently, a new method for processing the prediction residual of reverberant speech is derived, combining spatial averaging of the speech signals and temporal larynx cycle averaging. An enhanced speech signal with reduced reverberation is obtained with the processed residual.

Second, an adaptive blind SIMO system identification (BSI) algorithm is introduced and an optimal adaptation step-size is derived, which results in faster convergence and increased robustness to noise. Then, an adaptive common roots identification algorithm is developed and employed to demonstrate the degrading effects of common zeros on the BSI algorithm. Adaptive BSI in subbands is proposed for reduced channel length and thus, increased efficiency. Furthermore, exact and approximate acoustic impulse response equalization is discussed. A new subband multichannel least squares inverse filtering method is derived, which significantly reduces the computational complexity and is less

sensitive to inaccuracies in the channel estimates compared to its fullband counterpart. Substantial reduction in reverberation can be obtained using this method, even when the impulse response estimates contain errors. Finally, the thesis is concluded with a comparative discussion of the proposed methods and possible directions for prospective research in speech dereverberation.

Acknowledgements

I would like to express my gratitude to the following people:

First and foremost, my supervisor and mentor, Dr. Patrick A. Naylor, for years of devoted attention, fruitful discussions, knowledge-transfer, encouragement and inspiration. Without him, this thesis would have never seen the light.

Dr. Darren B. Ward, who co-supervised me for the first 18 months of my studies and gave me a great deal of support and guidance during these early stages.

The Engineering and Physical Sciences Research Council and the Department of the Electrical and Electronic Engineering at Imperial College for financial support.

Prof. Jan M. Hirsch from Uppsala University for many inspiring and insightful discussions regarding various aspects of research and academic life.

All my friends and colleagues around the world for moral support, patience, and great times.

Everyone in my family for their joint contribution to make me the person I am today and for all thinkable forms of support during all these years.

Ingrid for her love, encouragement, incredible patience and for always believing in me more than I do.

Last but not least, I am indebted to my mother, who has always supported and encouraged me wholeheartedly in every task that I have undertaken, both in good and bad times, and not least during the preparation of this thesis.

List of Publications

The following publications were produced during the time of this work:

1. N. D. Gaubitch, D. B. Ward and P. A. Naylor, "Statistical Analysis of the Autoregressive Modeling of Reverberant Speech," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4031-4039, Dec. 2006.

2. N. D. Gaubitch, Md. K. Hasan and P. A. Naylor, "Generalized Optimal Step-Size for Blind Multichannel LMS System Identification," *IEEE Signal Processing Letters*, vol. 13, no. 10, pp. 624-627, Oct. 2006.

3. N. D. Gaubitch and P. A. Naylor, "The Complex Multichannel LMS Algorithm for Adaptive Blind System Identification," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Paris, France, Sept. 2006.

4. S. Javidi, N. D. Gaubitch and P. A. Naylor, "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*, Paris, France, Sept. 2006.

5. J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt and P. A. Naylor, "Evaluation of Speech Dereverberation Algorithms using the MARDY Database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Paris, France, Sept. 2006.

6. X. Lin, N. D. Gaubitch and P. A. Naylor, "Two-Stage Blind Identification of SIMO Systems with Common Zeros," in *Proc. European Signal Processing Conf.*, Florence, Italy, Sept. 2006.

7. N. D. Gaubitch, Md. K. Hasan and P. A. Naylor, "Noise Robust Adaptive Blind Channel Identification Using Spectral Constraints," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Processing*, Toulouse, France, May 2006.

8. N. D. Gaubitch and P. A. Naylor, "Analysis of the dereverberation performance of microphone arrays," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Eindhoven, The Netherlands, Sept. 2005.

9. P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Eindhoven, The Netherlands, Sept. 2005.

10. N. D. Gaubitch, J. Benesty and P. A. Naylor, "Adaptive common root estimation and the common zeros problem in blind channel estimation," in *Proc. European Signal Processing Conf.*, Antalya, Turkey, Sept. 2005.

11. N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Mulit-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf.*, Vienna, Austria, Sept. 2004.

12. N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Kyoto, Japan, Sept. 2003.

Contents

List of Figures

List of Tables

Acronyms

AR	:	Auto Regressive
BSD	:	Bark Spectral Distortion
BSI	:	Blind System Identification
DFT	:	Discrete Fourier Transform
DRR	:	Direct-to-Reverberant Ratio
DSB	:	Delay-and-Sum Beamformer
FIR	:	Finite Impulse Response
GCI	:	Glottal Closure Instant
GDFT	:	Generalised Discrete Fourier Transform
LP	:	Linear Prediction
LS	:	Least Squares
MCLMS	:	Multichannel Least Mean Squares
NPM	:	Normalised Projection Misalignment
RTF	:	Room Transfer Function
SEGSRR	:	Segmental Signal-to-Reverberant Ratio
SMERSH	:	Spatiotemporal averaging Method for Enhancement of Reverberant Speech
SNR	:	Signal-to-Noise Ratio
SRA	:	Statistical Room Acoustics
SRR	:	Signal-to-Reverberant Ratio

Mathematical Notations

Here the notation used throughout the document is summarized.

General Notations

a	:	scalar quantity
$\boldsymbol{a}$	:	vector quantity
$\boldsymbol{A}$	:	matrix quantity
$a(n)$	:	function of a discrete variable n
a_n	:	function of a finite discrete variable n
$A(e^{j\omega})$	:	Fourier transform of a discrete function $a(n)$
$A(z)$	:	z-transform of a discrete function $a(n)$

Operators

$a * b$	:	linear convolution
$\boldsymbol{A}^T$	:	non-conjugate matrix transpose
$\boldsymbol{A}^H$	:	Hermitian (conjugate) matrix transpose
$\boldsymbol{A}^{-1}$	:	matrix inverse
$\boldsymbol{A}^+$	:	matrix pseudo inverse
$\mathscr{E}\{\cdot\}$	:	spatial expectation
$E\{\cdot\}$	:	mathematical expectation
a^*	:	complex conjugate
$\|\cdot\|$	:	Euclidean norm
$\deg[\cdot]$	:	degree of a polynomial
$\Im\{\cdot\}$	:	imaginary component of a complex number
$\Re\{\cdot\}$	:	real component of a complex number
$\lvert\cdot\rvert$	:	absolute value
$\lceil\cdot\rceil$	:	ceiling operator

Symbols and Variables

$\mathbf{0}$	:	null vector
A	:	room area (m^2)
$A(z)$	:	z-domain inverse prediction filter
$A(e^{j\omega})$	:	frequency domain inverse prediction filter
$\boldsymbol{a}$	:	AR coefficients from clean speech
$\boldsymbol{a}_{\mathrm{opt}}$	:	optimal AR coefficient estimates from clean speech
$\hat{\boldsymbol{a}}$	:	AR coefficient estimates

$B(e^{j\omega})$	:	frequency domain inverse prediction filter from reverberant speech
$\boldsymbol{b}$	:	AR coefficients obtained from a reverberant observation
$\boldsymbol{b}_{\text{opt}}$	:	optimum AR coefficient estimates obtained from a reverberant observation
$\hat{\boldsymbol{b}}_{\text{opt}}$	:	optimum AR coefficient estimates from multichannel linear prediction
$\bar{\boldsymbol{b}}$	:	AR coefficients from a beamformer output
$\bar{\boldsymbol{b}}_{\text{opt}}$	:	optimum AR coefficient estimates from a beamformer output
$\mathcal{B}$	:	Bark spectrum
b	:	Bark scale
c	:	speed of sound (m/s)
D_m	:	distance from source to microphone (meters)
$\boldsymbol{d}$	:	DFT vector
d_I	:	Itakura distance (dB)
$E(e^{j\omega})$	:	frequency domain prediction residual from clean speech
$\tilde{E}_m(e^{j\omega})$	:	frequency domain prediction residual from a reverberant observation
$E_{ml}(z)$	:	z-domain cross-relation error
$E_{ml,i}(z)$	:	subband z-domain cross-relation error
$\bar{E}_\ell(e^{j\omega})$	:	frequency domain larynx cycle at the output of a beamformer
$\hat{E}_\ell(e^{j\omega})$	:	frequency domain processed larynx cycle
$e(n)$	:	prediction residual from clean speech
$e_m(n)$	:	prediction residual from a reverberant observation
$\bar{e}(n)$	:	prediction residual at the output of a delay-and-sum beamformer
$\hat{e}(n)$	:	enhanced prediction residual
$\bar{e}_\ell(\ell)$	:	larynx cycle from a beamformer output
$\hat{e}_\ell(\ell)$	:	processed larynx cycle
$e_{ml}(n)$	:	cross-relation error
$e'_{ml}(n)$	:	cross-relation error for channels with no common zeros
$e_{c,m}(n)$	:	error for determining common zeros
F	:	frame length in frame based processing
f	:	frequency (Hz)
f_s	:	sampling frequency (Hz)
f_{Sch}	:	Schroeder frequency (Hz)
$G_m(z)$	:	z-domain inverse filter
$\hat{G}_m(z)$	:	z-domain inverse filter estimate
$\hat{G}_{\text{opt},m}(z)$	:	optimum z-domain inverse filter estimate
$G_{m,i}(z)$	:	subband inverse filter transfer function
$\hat{\boldsymbol{g}}$	:	concatenated inverse filter finite impulse responses estimate
$\boldsymbol{g}_\ell$	:	larynx cycle averaging equivalent inverse system
$\hat{\boldsymbol{g}}_\ell(n_\ell)$	:	larynx cycle inverse filter at the n_ℓth larynx cycle
$\boldsymbol{g}_m$	:	inverse filter finite impulse response
$\hat{\boldsymbol{g}}_m$	:	inverse filter finite impulse response estimate
$\hat{\boldsymbol{g}}_{\text{opt}}$	:	optimum concatenated inverse filter finite impulse responses estimate
$\bar{H}(e^{j\omega})$	:	RTF at the output of a delay-and-sum beamformer
$H_{\text{ap},m}(z)$	:	z-domain RTF all-pass component
$H_{\text{c}}(z)$	:	z-domain common zeros component of RTF

$H_{\mathrm{d},m}(e^{j\omega})$	:	direct component of the RTF
$H_{\mathrm{EQ}}(e^{j\omega})$	:	equalised RTF
$\bar{H}_{\mathrm{EQ}}(e^{j\omega})$	:	mean level of equalised RTF
$H_m(e^{j\omega})$	:	room transfer function
$H_m(z)$	:	z-domain RTF
$H'_m(z)$	:	z-domain characteristic zeros component of RTF
$\hat{H}'_m(z)$	:	estimate of the z-domain characteristic zeros component of RTF
$H_{\mathrm{mp},m}(z)$	:	z-domain RTF minimum phase component
$H_{\mathrm{r},m}(e^{j\omega})$	:	reverberant component of the RTF
$\boldsymbol{h}$	:	vector of concatenated finite RIRs
$\hat{\boldsymbol{h}}$	:	vector of concatenated RIR estimates
$\hat{\boldsymbol{h}}'$	:	concatenated estimates vector of the characteristic zeros RIR components
$\boldsymbol{h}_{\mathrm{ap},m}$	:	all-pass component of the RIR
$\boldsymbol{h}_{\mathrm{mp},m}$	:	minimum phase component of the RIR
$\boldsymbol{h}_{\mathrm{c}}$	:	common zeros RIR component vector
$\hat{\boldsymbol{h}}_{\mathrm{c}}$	:	estimate of the common zeros RIR component vector
$\hat{\boldsymbol{h}}_{\mathrm{c,opt}}$	:	optimum estimate of the common zeros RIR component vector
$\boldsymbol{h}_{\mathrm{d}}$	:	direct RIR component vector
$\boldsymbol{h}_m$	:	finite room impulse response vector
$\boldsymbol{h}'_m$	:	characteristic zeros RIR component vector
$\hat{\boldsymbol{h}}_m$	:	RIR estimate vector
$\hat{\boldsymbol{h}}'_m$	:	estimate of the characteristic zeros RIR component vector
$\hat{\boldsymbol{h}}_{\mathrm{opt}}$	:	vector of concatenated optimum RIR estimates
$\boldsymbol{I}$	:	identity matrix
$\mathcal{I}$	:	number of cycles in the inter-cycle averaging
i	:	subband index
J	:	cost function for AR coefficient calculation
J_M	:	cost function for multichannel AR coefficient calculation
$J_{\mathrm{c}}(n)$	:	cost function for common zeros component calculation at time n
$\tilde{J}(n)$	:	normalised cost function at time n for adaptive BSI
$J_\mu(n)$	:	cost function at time n for optimal step-size computaiton
K	:	number of subbands
k	:	wave number
L	:	RIR length
L'	:	characteristic RIR component length
$\tilde{L}$	:	decomposed subband RIR length
L_{c}	:	common zeros RIR component length
L_{i}	:	inverse filter length
$\tilde{L}_{\mathrm{i}}$	:	subband inverse filter length
L_{pr}	:	subband prototype filter length
$\mathcal{L}$	:	length of a larynx cycle
$\boldsymbol{\ell}_m$	:	microphone position vector
M	:	total number of microphones
m, l	:	channel (microphone) index

N	:	decimation factor
N_{FT}	:	number of frequency points
n	:	discrete time sample index
$\mathcal{N}$	:	number of simulation runs
n_ℓ	:	GCI position time
$\mathcal{P}_x$	:	total microphone signal power
$\mathcal{P}_\nu$	:	total observation noise power
p	:	linear prediction order
$\boldsymbol{p}$	:	cross-correlation vector between reverberant and enhanced larynx cycles
$\boldsymbol{p}_i$	:	subband analysis correlation vector
$\boldsymbol{Q}$	:	autocorrelation matrix from reverberant observations
$\hat{\boldsymbol{Q}}$	:	mean autocorrelation matrix from M micropohones
$\bar{\boldsymbol{Q}}$	:	autocorrelation matrix from a beamformer output
$\boldsymbol{q}$	:	autocorrelation vector from reverberant observations
$\hat{\boldsymbol{q}}$	:	mean autocorrelation vector from M micropohones
$\bar{\boldsymbol{q}}$	:	autocorrelation vector from a beamformer output
$\boldsymbol{R}$	:	autocorrelation matrix from clean speech
$\boldsymbol{r}$	:	autocorrelation vector from clean speech
$s(n)$	:	input speech sample at time n
$\boldsymbol{s}(n)$	:	input vector of speech signal at time n
$\hat{s}(n)$	:	enhanced speech signal sample at time n
$\hat{\boldsymbol{s}}(n)$	:	vector of enhanced speech samples
$s_{\mathrm{d}}(n)$	:	sample of direct path speech at time n
$\boldsymbol{s}_{\mathrm{d}}(n)$	:	vector of direct path speech signal samples at time n
$S(e^{j\omega})$	:	frequency domain input speech signal
$S(z)$	:	z-domain input speech signal
T_{60}	:	reverberation time (s)
$\boldsymbol{t}$	:	dsb AR coefficients offset vector
$\boldsymbol{\mathcal{T}}$	:	dsb AR coefficients scaling matrix
$\boldsymbol{U}_i$	:	subband analysis filter convolution matrix
$U_i(z)$	:	z-domain subband analysis filter
$\boldsymbol{u}$	:	desired output for inverse filter design
$u_i(n)$	:	subband analysis filter
V	:	room volume (m^3)
$V(z)$	:	z-domain linear prediction filer
$V(e^{j\omega})$	:	frequency domain linear prediction filer
$V_i(z)$	:	z-domain subband synthesis filter
v	:	Tukey window taper ratio
$v_i(n)$	:	subband synthesis filter
W_K	:	twiddle factor
w_m	:	beamformer channel weight
w_u	:	larynx cycle weight
$\bar{X}(e^{j\omega})$	:	frequency domain output of a delay-and-sum beamformer
$X_{\mathrm{c}}(z)$	:	z-domain reverberant speech, common zeros in the RTF
$X_m(e^{j\omega})$	:	frequency domain reverberant speech observation

$X_m(z)$	:	z-domain reverberant speech observation
$X'_m(z)$	:	z-domain reverberant speech, no common zeros in the RTF
$\hat{X}_m(z)$	:	z-domain subband filtered reverberant speech observation
$\bar{x}(n)$	:	output of a delay-and-sum beamformer at time n
$\boldsymbol{x}_c(n)$	:	vector of reverberant speech samples at time n, common zeros in the RTF
$x_m(n)$	:	reverberant speech observation sample at time n
$\boldsymbol{x}_m(n)$	:	vector of reverberant speech samples at time n
$x'_m(n)$	:	reverberant speech observation sample at time n, no common zeros in the RTF
$\boldsymbol{x}'_m(n)$	:	vector of reverberant speech samples at time n, no common zeros in the RTF
$\hat{x}_m(n)$	:	subband filtered reverberant speech observation sample at time n
$Y_{ml,i}(z)$	:	z-domain subband cross filtering output
$y_{ml,i}(n)$	:	subband cross filtering output at time n
$\boldsymbol{y}_i$	:	source position vector
z	:	z-transform variable
α	:	average wall absorption coefficient
β	:	forgetting factor
Γ	:	matrix with eigenvectors
γ	:	spatial expectation RTF
$\bar{\gamma}$	:	uncorrelated component of a beamformer RTF
$\boldsymbol{\gamma}$	:	DYPSA cost function vector
Δ	:	distance between zeros
Δ_{EQ}	:	equalised phase distortion
$\delta(n)$	:	unit impulse
Θ_i	:	random rotation matrix
θ_i	:	random translation vector
θ	:	arbitrary angle
κ	:	arbitrary scale factor
Λ	:	diagonal matrix with eigenvalues
λ	:	weight vector for DYPSA cost function
μ	:	adaptation step-size
μ_{opt}	:	optimal step-size
$\tilde{\mu}_{\mathrm{opt}}$	:	approximate optimal step-size
$\mathcal{N}_m(e^{j\omega})$	:	observation noise in the frequency domain
$\nu_m(n)$	:	observation noise
Ξ	:	correlated component in the autocorrelation matrix of a beamformer output
ξ	:	correlated component in the autocorrelation vector of a beamformer output
σ^2	:	variance
σ_{EQ}	:	standard deviation of the equalised magnitude response
τ	:	arbitrary delay
τ_m	:	time delay of arrival
Φ	:	larynx cycle autocorrelation matrix
$\psi(k)$	:	correlated component of a beamformer RTF
Ω	:	subset of GCI candidates

Chapter 1

Introduction

1.1 Context of Work

There is a growing demand for high quality hands-free speech input in various telecommunication applications [1, 2]. One driving force behind this development is the rapidly increasing use of portable devices such as mobile telephones, personal digital assistants (PDAs) and laptop computers [3]. Furthermore, there is a worldwide expansion of broadband internet access [4]. These facts have paved the way for several advanced speech applications such as voice over IP telephony, teleconferencing with automatic camera steering, automatic speech-to-text conversion, speaker identification and voice-controlled device operation and car interior communication systems [2]. Another important application where speech obtained from a distant talker is of interest is that of hearing aids [2]. In hands-free speech acquisition the talker would typically be located at a distance of $0.3-4$ m from the microphones. In such a scenario, the speech signal is affected by the user's surrounding environment, which results in the following three distinct effects [1, 2]:

(a) *Additive measurement noise* due to, for example, computer fans, other talkers, surrounding traffic and the sound acquisition equipment. When the noise level is of comparable strength to the speech signal, it is difficult for a listener to distinguish the desired speech signal from the noise and thus intelligibility is reduced.

(b) *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.

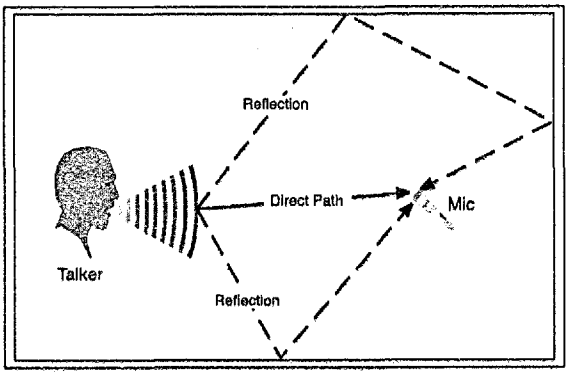

Figure 1.1: Reverberant room.

(c) *Reverberation* which 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 perceptual experience of the listener and the performance of applications such as speech recognizers [1]. Speech enhancement and acoustic echo cancellation are two widely researched fields which address problems (a) and (b) respectively. Several significant contributions have been made in these areas [5–7] and many algorithms have been implemented and are in use in commercial applications [8]. The problem of reverberation, on the other hand, has received much less attention in the literature and still remains largely unsolved. Nevertheless, finding solutions to this problem are essential for the future development of applications with hands-free speech acquisition. The focus of the forthcoming chapters in this thesis is on the analysis and enhancement of reverberant speech.

Although dereverberation is a relatively new area of research, historically, the effects of reflected sound have been apparent to thinkers for many centuries. This is evident, for example, in the notion of reflected speech occurring in Plato's Republic [9]: *"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?"*. More recently, pioneering work on sound and acoustics in the 19th century was undertaken by, for example, Rayleigh [10] and Sabine [11].

When speech signals are obtained in an enclosed space by one or several microphones positioned at a distance from the talker, the observed signal consists of a superposition of the direct path speech component and delayed and attenuated copies of it due to multiple reflections from the surrounding walls and other objects, as illustrated in Fig. 1.1. Reverberation alters the characteristics of the speech signal which can be extremely problematic in applications including speech recognition, source localization and speaker verification and significantly reduces the performance of algorithms developed without taking the room effects into consideration. The deleterious effects are further magnified as the distance between the talker and the microphones is increased. The perceptual effects of reverberation can be summarized as:

(i) *The box effect* – the reverberated speech signal can be viewed as the same source signal coming from several sources positioned at different locations and thus arriving at the observation point at different times and with different intensities [12]. This adds spaciousness to the sound [13] and makes the talker sound as if positioned "inside a box".

(ii) *The distant talker effect* - the perceived spaciousness explained in point (i) makes the talker sound far away from the microphone.

(iii) *Reduced intelligibility* - The signal observed at any given microphone consists of the sum of all these virtual sources, i.e. the reflections, which results in spectral smearing and can make the speech less intelligible.

At the time of writing, the problem of reverberation is often resolved by utilizing a headset, where the microphone is kept close to the mouth. Alternatively, a directional microphone positioned directly in front of the talker can be used. The advent of bluetooth technology [14] has also allowed for high quality wireless headsets which has further extended the usability of such devices. Nevertheless, these solutions impose restrictions on the flexibility and comfort of the user, which are the main desired features in the use of hands-free equipment [1] and in some applications, such as teleconferencing, where there are multiple participants, these solutions may not be practical. Therefore, a signal processing approach independent of the relative talker-microphone configuration is preferable.

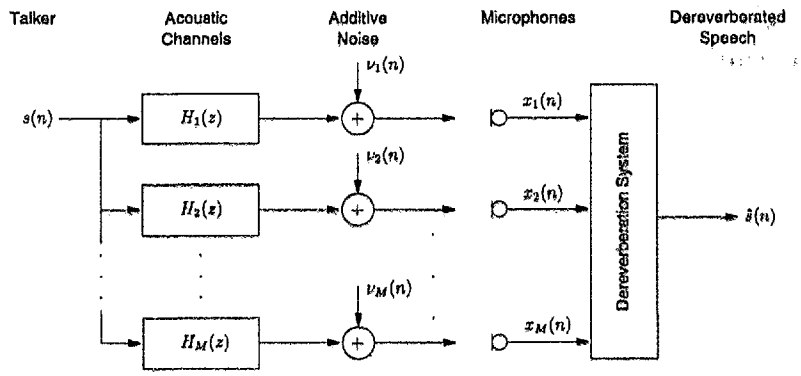

Figure 1.2: Generic multichannel reverberation-dereverberation system model.

A system diagram for the general problem of multichannel dereverberation (which will be treated with more rigour in Section 2.1) is shown in Fig. 1.2. The aim is to find a multiple-input-single-output dereverberation system which uses only the information from the M microphones to estimate the clean speech input $s(n)$. Recent efforts in acoustic signal processing research have produced several algorithms for reverberant speech enhancement, which can be divided broadly into three main categories:

(i) *Beamforming* – the 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. Beamforming is exclusive to multi-microphone systems.

(ii) *Speech enhancement* – the speech signals are modified so as to represent better some features of the clean speech signal according to *a priori* models of the speech waveform or spectrum.

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

1.2 Scope and Original Contribution

The focus of the research presented in this thesis is on the improvement of the perceived quality of reverberant speech for hands-free telephony applications such as teleconferencing, aimed for use in, for example, offices. The objective is to make hands-free speech sound as similar as possible to that of a close talking microphone.

The dereverberation problem is approached assuming availability of multiple microphone observations. This is motivated by research in the area of microphone arrays, which has demonstrated for a multitude of applications that using multiple microphones is advantageous and often superior to single microphone systems in the processing of acoustic signals [15]. In addition, low-cost eight-element microphone arrays have appeared on the market for improved hands-free speech recognition applications [16] and Microsoft Windows Vista incorporates support for multimicrophone systems [17]. Although a multimicrophone input is assumed, the number of sensors is restricted to be $M \leq 8$ in order to keep these studies realistic.

The development work in the thesis comprises two major areas. First, the effects of reverberation on the autoregressive (AR) modelling of reverberant speech are studied and a speech enhancement method using linear prediction residual processing is developed. The effectiveness of this algorithm is demonstrated with simulation experiments. An attractive feature of the method is that it does not require computationally burdensome room impulse response identification. Secondly, an adaptive blind system identification (BSI) algorithm and its application to acoustic systems is investigated. Strategies to improve the performance of this algorithm in practical environments are considered, including adaptation step-size control and subband processing for increased robustness to noise and for reduced complexity. A key conclusion is to identify the conditions under which exact channel estimates can be obtained and to show that these conditions carry significant restrictions in practice using current approaches. Finally, a new subband multichannel room impulse response inversion algorithm is derived, which is more computationally efficient and more robust to inaccuracies in the impulse response estimates compared to its fullband counterpart.

Statement of Originality

The following aspects of the thesis are believed to be original contributions:

- Derivation of a closed form theoretical expression for the expected dereverberation performance of a microphone array. (Chapter 2, Section 2.6)
- Statistical analysis of the effects of reverberation on the autoregressive modelling of speech for the cases of a single and multiple observations. (Chapter 3)
- Development of a novel method for enhancement of reverberant prediction residuals based on a combination of spatial averaging and a new approach to temporal averaging of neighbouring larynx cycles, with application to speech dereverberation. (Chapter 4, Section 4.2, Section 4.3)
- Evaluation of the DYPSA algorithm for glottal closure instants estimation from reverberant speech for a single microphone and the extension of the algorithm to multiple microphones. (Chapter 4, Section 4.2.1)
- Derivation of a complex multichannel LMS (MCLMS) algorithm for blind identification of SIMO systems. (Chapter 5, Section 5.3.3)
- Derivation and implementation of an optimal step-size for adaptive blind SIMO system identification using the multichannel LMS algorithm. (Chapter 5, Section 5.4)
- Derivation of an adaptive common roots detection and estimation algorithm and application of this algorithm to the study of blind system identification in the presence of common or near common zeros. (Chapter 5, Section 5.5)
- Study of the relation between the fullband cross-relation error and the subband cross-relation errors and an implementation of the MCLMS in oversampled, decimated subbands. (Chapter 5, Section 5.6.1)
- Derivation and implementation of a multichannel subband inverse filtering method for long acoustic impulse responses and application of this method to speech dereverberation (Chapter 6, Section 6.2).

1.3 Thesis Outline

The remaining chapters of this thesis are organized as follows.

Chapter 2 provides a formal introduction to the problem of speech dereverberation. The characteristics of the room impulse response and the room transfer function are discussed and the source-image method for simulation of room acoustics is presented. Next, objective metrics for evaluation of dereverberated speech are addressed and the measures utilized in the subsequent work are defined. An expression for the expected dereverberation performance of the delay-and-sum beamformer is derived. It is then used to demonstrate the improvement in direct-to-reverberant ratio that can be expected from what is considered the simplest speech enhancement technique. Finally, an overview of the speech dereverberation literature is provided, highlighting pros and cons of many existing methods.

Chapter 3 investigates the AR modelling of reverberant speech. Several results from statistical room acoustics are reviewed and employed to analyze the effects of reverberation on the AR modelling of speech both for single and for multiple observations. It is shown that, in terms of spatial expectation, the AR coefficients obtained from single and from multiple observations provide a close estimate of the clean speech AR coefficients, while AR coefficients obtained from a DSB result in worse estimates due to spatial aliasing. It is further demonstrated that multichannel estimates give the best approximation to the clean speech AR coefficients at any single source-microphone position and is the preferred method to use in reverberant environments. Consequently, the effect of the room impulse responses is found to reside primarily in the prediction residual.

Chapter 4 presents the development of a method for enhancing the prediction residuals from reverberant speech, using spatial averaging and a new approach to temporal averaging of neighbouring larynx cycles. The DYPSA algorithm is reviewed and evaluated for glottal closure instants (GCIs) identification from reverberant speech. It is demonstrated that the GCIs can be estimated accurately using a multimicrophone system, making the algorithm suitable for the application to larynx cycle

segmentation. Each processed larynx cycle is then used to update a slowly varying equalization filter. This filter, which is only updated in voiced speech segments, is applied in the enhancement of the reverberant prediction residual. An estimate of the clean speech signal is obtained with linear prediction synthesis using the processed prediction residual. The new method preserves the naturalness of speech while it attenuates reverberant effects both in voiced and unvoiced regions.

Chapter 5 addresses blind SIMO system identification for acoustic impulse response estimation. The complex multichannel LMS (MCLMS) algorithm for adaptive blind SIMO system identification is derived and evaluated. An optimal adaptation stepsize is derived and implemented and it is demonstrated to provide significantly improved convergence performance for both noisy and noise-free conditions. Subsequently, the MCLMS is used to develop an adaptive common roots detection and identification algorithm, which is used to investigate the effects of common zeros on adaptive BSI. Finally, the relation between the fullband error and the corresponding subband errors is derived and an oversampled filter bank implementation of the MCLMS is proposed including a discussion of the necessary building blocks.

Chapter 6 presents a study of the problem of inverse filter design for long, non-minimum phase impulse responses from approximate channel estimates. Considering the computational complexity, noise and inaccurate impulse responses, a subband design of a least squares multichannel inverse filtering algorithm is implemented and evaluated. It is shown that this method can significantly reduce the computational intensity for long impulse responses and it is also demonstrated to reduce the sensitivity to inaccuracies in the estimated impulse responses.

Chapter 7 summarizes the work presented in this thesis and provides a comparative summary of the main results obtained for speech dereverberation. Finally, the thesis is concluded with guidelines for further developments of the herein presented ideas and an outlook into the future of reverberant speech enhancement.

Chapter 2

Fundamentals of Reverberant Speech Enhancement and Literature Overview

THIS chapter introduces reverberant speech enhancement and the problem is formulated mathematically both in the time and in the frequency domains. The finite impulse response (FIR) model of the room acoustic paths used in this thesis is described and the source-image method for simulating room impulse responses is reviewed. Key characteristics of room impulse responses known from room acoustic studies are summarized and several consequences of these properties regarding speech dereverberation algorithms are deduced. Subsequently, qualitative measurement of reverberant speech is discussed and the metrics used in this thesis are motivated and defined. Finally, a comprehensive overview of the existing literature on speech dereverberation is presented, emphasizing the strengths and weaknesses of many state-of-the-art methods and providing an annotated bibliography.

2.1 Problem Formulation

Reverberation is the process of multipath propagation of an acoustic signal $s(n)$ from its source to one or more microphones. The observed signal, $x_m(n)$, at microphone m in an array of M microphones can be described as the superposition of the direct signal at time n, which is attenuated and delayed due to the direct path propagation from the talker to

the microphone, and an infinite series of attenuated components of that signal from earlier time instances [13]. This can be expressed mathematically as

$$x_m(n) = \sum_{i=0}^{\infty} h_{m,i} s(n-i), \tag{2.1}$$

where the coefficients $h_{m,i}$ encompass the attenuation and the propagation delay of the direct signal and all the reflected components.

Rooms are generally stable systems with the coefficients $h_{m,i}$ tending to zero and therefore, it is sufficient to consider only the first L coefficients. The choice of L is often linked to the reverberation time of the room, which is discussed in Section 2.2. Taking into account any additive noise sources as discussed in Section 1.1, the observed signal at the mth microphone can be written in a vector form

$$x_m(n) = \boldsymbol{h}_m^T \boldsymbol{s}(n) + \nu_m(n), \tag{2.2}$$

where $\boldsymbol{h}_m = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T$ is the L-tap impulse response of the acoustic channel from the source to microphone m, $\boldsymbol{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L+1)]^T$ is the speech signal vector and $\nu_m(n)$ is observation noise. In the frequency domain this can be equivalently expressed as

$$X_m(e^{j\omega}) = H_m(e^{j\omega}) S(e^{j\omega}) + \mathcal{N}_m(e^{j\omega}), \tag{2.3}$$

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

The aim of dereverberation is to process the reverberant observations, $x_m(n)$, $m = 1, \ldots, M$, so as to form $\hat{s}(n)$, an estimate of $s(n)$. This is a blind problem since, in most practical cases, neither the signal $s(n)$ nor the room impulse responses $\boldsymbol{h}_m$ are available. The overall reverberation-dereverberation procedure is illustrated in Fig. 1.2. Note that processing here does not necessarily imply deconvolution as will be explained in Section 2.5.

2.2 The Room Impulse Response

The room impulse response (RIR) is a key characteristic feature of the acoustics of a given enclosure and therefore, as is evident from (2.2), study of the RIR is a natural approach dereverberation. This section reviews some important characteristics of RIRs in terms of physical, psychoacoustical and signal processing properties and also a method for the simulation of room acoustics. For the remainder of the thesis, it is assumed explicitly that the RIRs are modelled as finite impulse response (FIR) filters. However, several other models have been considered in the literature, including both finite and infinite impulse response structures [18–22]. The choice of the RIR model will generally influence the algorithmic development. Further discussion on the consequences of the chosen RIR model is given in the concluding Chapter 7.

A common quantitative measure of the impulse response of a room is the reverberation time, originally introduced by Sabine [13]. The reverberation time, T_{60}, is defined as the time taken for the reverberant energy to decay by 60 dB from its original strength once the source has been abruptly shut off [13] and is governed by the room geometry and the reflectivity of the walls. The relation between these factors is often given by Eyring's reverberation formula [13, pp. 128]

$$T_{60} = -\frac{24V \ln 10}{cA \ln(1-\alpha)}, \tag{2.4}$$

where V is the room volume, A is the total room area, α is the average wall absorption coefficient and c is the speed of sound in air. The effect of air attenuation is not taken into account in this formulation; it is often considered negligible in small rooms [13]. The reverberation time naturally leads to a definition of the length L (samples) of the room impulse response which, for sampling frequency f_s, can be written

$$L = T_{60} f_s. \tag{2.5}$$

The reverberation time is approximately constant for a given room. However, the room impulse response is spatially variant and will vary as the talker, the microphones or other objects in the room change location [13]. A particular characteristic that varies

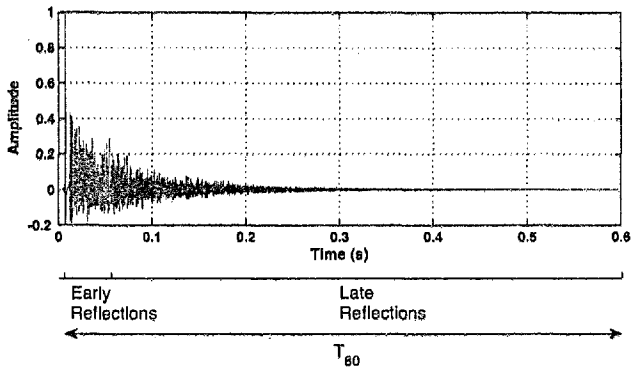

Figure 2.1: An example impulse response for a room with dimensions $(6.4 \times 5 \times 4)$ **m, reverberation time of** $T_{60} = 0.6$ **s and source-microphone separation of** $D_1 = 2.5$ **m.**

with the source-microphone separation is the relation between the energies of the direct path component and the reflected components. The critical reverberation distance is the distance such that these two energies are equal and is given by [13, pp. 136]

$$D_{\mathrm{R}} = \frac{1}{4}\sqrt{\frac{A}{\pi}} = \frac{1}{10}\sqrt{\frac{V}{\pi T_{60}}}. \tag{2.6}$$

Figure 2.1 shows a simulated room impulse response generated with the source-image method [12,23], which will be described in Section 2.3. In this example, a rectangular room of dimensions $(6.4 \times 5 \times 4)$ m is assumed, with reverberation time $T_{60} = 0.6$ s and source-microphone separation $D_1 = 2.5$ m. The early and the late reflections are indicated as two distinct regions of the RIR. The early reflections are often taken as the first 50 ms of the room impulse response [13], and constitute well defined impulses of large magnitude relative to the smaller magnitude and diffuse nature of the late reflections. The RIR early reflections cause spectral changes in the formants of speech and lead to a perceptual effect referred to as coloration [13]. In general, closely spaced echoes are not distinguished by the 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 of speech with a similar effect to increasing the strength of the direct path sound [1, 13, 24]. However, coloration can degrade the quality of recorded speech [13]. 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 that is referred to as the reverberation tail and provide the major contribution to what is generally perceived as reverberation [13].

In terms of spectral characteristics, the room transfer function is proportional to the sound pressure [13, 25] and has been studied extensively in the room acoustics literature where many properties have been established [13]. One property of interest in the context of dereverberation is the average distance between minimum and maximum spectral points which has been shown to extend beyond 10 dB [13, 26]. Moreover, since the room transfer function is changing depending on the location of the source and the microphone, it can be described as a random process [13]. The latter will be elaborated in Chapter 3. Finally, Neely and Allen [27] demonstrated that the RIRs in most real rooms will be non-minimum phase.

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

(i) Hand-free telephony users can be expected to move around in the room and so the RIRs will vary with time.

(ii) The use of measured impulse responses is not a feasible option due to the dependence on source-microphone position and on the room.

(iii) If the source-microphone separation is much smaller than the critical distance in (2.6), the effects of reverberation are negligible. Thus, dereverberation is of greatest importance when $D_m \geq D_{\mathrm{R}}$.

(iv) The reverberation time in office sized rooms can be expected to vary in the range $0.1 - 1$ s. Consequently, this involves FIR filters of several thousand taps, which also increases with increased sampling frequency.

(v) Although the full channel length does not need to be included, the late reflections are important in dereverberation and in particular in the case when $D_m > D_{\mathrm{R}}$.

(vi) The non-minimum phase property and the large dynamic range of the room transfer functions will be problematic in designing RIR equalization filters.

2.3 Room Acoustics Simulation

The results presented in this thesis are based on simulated environments in which computer generated RIRs are convolved with speech recordings made in anechoic conditions. There are several methods available for simulation of room acoustics based on ray tracing [13], digital waveguide models [28,29] and source-image models [12,13,23]. Here, the commonly employed source-image method, originally proposed by Allen and Berkley [12], is used and is summarized in this section.

The source-image method models reverberation with a set of source images such that, when the source is active, all image sources simultaneously transmit the same signal. Due to different distances from each image source to the measurement location, the signals arrive at different times and with different intensities. The finite reflection coefficient (assumed uniform for all six walls), ϕ, is applied to account for the sound reflected by the walls and is related to the wall absorption coefficient in (2.4) as $\alpha = 1 - \phi^2$. 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 ith RIR tap can be written [12]

$$h_{m,i} = \sum_{\epsilon=0}^{1} \sum_{\rho=-\infty}^{\infty} \frac{\delta(i - (|\boldsymbol{D}_\epsilon + \boldsymbol{D}_\rho|/c))}{4\pi|\boldsymbol{D}_\epsilon + \boldsymbol{D}_\rho|} \phi^{(|q-u|+|r-v|+|s-w|+|q|+|r|+|s|)}, \tag{2.7}$$

where $\boldsymbol{D}_\epsilon = (x - \tilde{x} + 2u\tilde{x}, y - \tilde{y} + 2v\tilde{y}, z - \tilde{z} + 2w\tilde{z})$ and $\boldsymbol{D}_\rho = (qL_x, rL_y, sL_z)$ such that $|\boldsymbol{D}_\epsilon + \boldsymbol{D}_\rho|$ defines the distance between the receiver and each source image. The triplets $\epsilon = \{u, v, w\}$ and $\rho = \{q, r, s\}$ indicate that each summation in fact consists of three different summations. Although ρ is given in the interval $[-\infty, \infty]$, in practice, the limits are finite and are governed by the chosen order of the source images. The expression in (2.7) follows from the solution of the wave equation for a rectangular, lossless enclosure [12].

In the original implementation of (2.7), the impulses calculated at fractional sampling delays are rounded to the nearest integer sample, which at lower sampling frequencies can introduce significant errors. Peterson [23] proposed to apply a low-pass filter to each impulse obtained in (2.7), which better satisfies the sampling theorem and provides a more accurate representation of the simulated impulse responses, in particular for multimicrophone scenarios. This is equivalent to a fractional delay [30] implementation of

each impulse. Consequently, a Hanning-windowed ideal low-pass filter is applied to each simulated image impulse according to [23]

$$w_{\mathrm{L},n} = \begin{cases} 0.5\,(1+\cos(2\pi n/L_{\mathrm{w}}))\,\mathrm{sinc}(2\pi f_{\mathrm{co}} n), & L_{\mathrm{w}}/2 < n < L_{\mathrm{w}}/2 \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

where f_{co} is the filter cut-off frequency, here taken as $f_{\mathrm{co}} = f_{\mathrm{s}}/2$ and L_{w} is the window length, here set to 8 ms, i.e. $L_{\mathrm{w}} = \frac{8 f_{\mathrm{s}}}{1000}$.

2.4 Measurement and Evaluation

Reliable quantitative measurement of reverberation in a speech signal is particularly difficult and an unanimously accepted methodology has yet to emerge. In this section, commonly used objective measures are reviewed. The metrics are considered in two classes: (i) measures using the RIR and (ii) measures comparing a reverberant speech signal to a clean reference speech signal.

2.4.1 Evaluation Using Room Impulse Responses

Many studies of speech intelligibility in reverberant rooms have emerged from room acoustics research [13, 24]. From the RIR, it is possible to measure the Direct-to-Reverberant Ratio (DRR) for an observed reverberant signal as the ratio of the power due to the direct acoustic path to the power due to the non-direct paths. Several variants of this measure exist [13]. The Speech Transmission Index [13], for example, is a measure of speech intelligibility in reverberant environments based on the reverberation time and the masking properties of the ear in different frequency bands. In this work only the most fundamental measure of this class is considered, the direct-to-reverberant ratio.

Direct-to-Reverberant Ratio

The direct-to-reverberant ratio measures the energy ratio between the direct component (which may or may not include the early reflections) and the total energy of the reflected signal components. The definition of the DRR is [13]

$$\mathrm{DRR} = 10\log_{10}\left[\frac{\|\boldsymbol{h}_{\mathrm{d}}\|^2}{\|\boldsymbol{h}-\boldsymbol{h}_{\mathrm{d}}\|^2}\right]\ \mathrm{dB} \tag{2.9}$$

where $\boldsymbol{h}_{\mathrm{d}} = [h_0\ h_1\ \ldots\ h_\tau\ 0\ \ldots\ 0]^T$ is the direct-path component of the impulse response $\boldsymbol{h}$ and τ is the number of samples to consider as a direct component. Here, h_0 corresponds to the sample position of the direct-path impulse. This formulation relies on the simplifying assumption that the propagation time of the direct-path is an integer multiple of the sampling period so that the impulse responses can be time-aligned accurately. The choice of τ can vary in the range of $0 - 80$ ms depending on how large a portion of the impulse response is included as early reflections. For example, a common measure is the C_{50} [13], where the early reflections include the first 50 ms as discussed in Section 2.2.

DRR requires the impulse response of the room and the resulting impulse response after the reverberation reduction processing in order to evaluate the achieved improvement. However, it is not always possible to obtain a processed impulse response, for example, in frame based speech enhancement algorithms, where the processing is not necessarily linear between frames. An example of one such algorithm will be given in Chapter 4. Therefore, the DRR and other RIR based evaluation methods have limited applicability in speech dereverberation as a general metric.

2.4.2 Evaluation Using Speech Signals

This class of evaluation methods, which is inherited from the speech enhancement and speech coding communities [31, 32], uses a processed speech signal and the clean speech signal as a reference to quantify the distortion in the speech signal under evaluation. Following on from the discussion in Section 2.4.1, this is a preferred evaluation approach with application to a broader range of dereverberation algorithms. Two objective measures will be considered: Segmental Signal-to-Reverberation Ratio (SEGSRR) and Bark Spectral Distortion (BSD).

Segmental Signal-to-Reverberation Ratio

The Segmental Signal-to-Reverberation Ratio resembles the DRR but is evaluated using the speech signals and is defined as

$$\mathrm{SEGSRR} = \frac{10}{\mathcal{I}} \sum_{i=0}^{\mathcal{I}-1} \log_{10} \left[\frac{\|s_{\mathrm{d}}(i)\|^2}{\|\hat{s}(i) - s_{\mathrm{d}}(i)\|^2} \right] \ \mathrm{dB}, \tag{2.10}$$

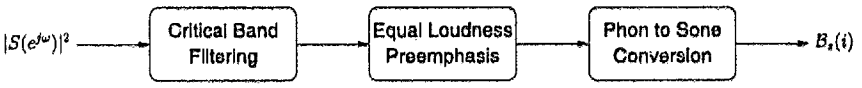

Figure 2.2: Bark Spectrum Calculation.

where $s_{\mathrm{d}}(i) = [s_{\mathrm{d}}(iF)\ s_{\mathrm{d}}(iF+1)\ \cdots\ s_{\mathrm{d}}(iF+F-1)]^T$ is a vector of samples from the ith frame of the clean speech signal with $s_{\mathrm{d}}(n) = \boldsymbol{h}_{\mathrm{d}}^T \boldsymbol{s}(n)$, $\hat{s}(i) = [\hat{s}(iF)\ \hat{s}(iF+1)\ \cdots\ \hat{s}(iF+F-1)]^T$ is a vector of samples from the ith frame of the enhanced speech signal, $\mathcal{I}$ is the total number of frames and F is the frame length in samples.

Normalized Bark Spectral Distortion

The normalized Bark Spectral Distortion is a perceptually motivated measure based on the difference between the Bark spectra of two signals and is defined as [31,33,34]

$$\mathrm{BSD} = \frac{\sum_{i=0}^{\mathcal{I}-1} \|\mathcal{B}_{\hat{s}}(i) - \mathcal{B}_{s_{\mathrm{d}}}(i)\|^2}{\sum_{i=0}^{\mathcal{I}-1} \|\mathcal{B}_{s_{\mathrm{d}}}(i)\|^2}, \tag{2.11}$$

where $\mathcal{B}_{s_{\mathrm{d}}}(i) = [\mathcal{B}_{s_{\mathrm{d}}}(iF)\ \mathcal{B}_{s_{\mathrm{d}}}(iF+1)\ \ldots\ \mathcal{B}_{s_{\mathrm{d}}}(iF+F-1)]^T$ and $\mathcal{B}_{\hat{s}}(i) = [\mathcal{B}_{\hat{s}}(iF)\ \mathcal{B}_{\hat{s}}(iF+1)\ \ldots\ \mathcal{B}_{\hat{s}}(iF+F-1)]^T$ are, respectively, the Bark spectra of the direct path speech signal $s_{\mathrm{d}}(n)$ and the processed speech signal $\hat{s}(n)$. The calculation of the Bark spectrum is illustrated in Fig. 2.2 and consists of the following three stages:

1. *Critical band filtering* - inspired by the physiology of the human peripheral auditory system, the critical bank filtering is performed in two steps. First the signal spectrum is converted from the Hertz scale to the Bark scale according to the relation [31,34]

$$b = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f}{7500}\right)^2. \tag{2.12}$$

   This is then followed by convolution of the transformed spectrum with the transformed transfer function of the critical band filter

$$10\log_{10} F(b) = 7 - 7.5(b-0.215) - 17.5\sqrt{(0.196 + (b-0.215)^2)}. \tag{2.13}$$

   Note that, in the Bark scale, the critical band filters are equivalent for each band in contrast to their frequency domain versions [31].

2. *Perceptual weighting* - is motivated by the frequency sensitivity profile of the ear. This can be represented by the equal loudness level contours [13]. An illustration of the equal loudness level contours generated in MATLAB [35] according to the ISO226 standard are shown in Fig. 2.3. Each curve indicates how the intensity level of a tone must be varied in order to maintain a constant level of perceived loudness. For example, a tone at 100 Hz may need to be up to 35 dB more intense than a tone of 1000 Hz for the two to be perceived as equally loud. The loudness level in phons is the perceived level in dB at 1 kHz, i.e. if the a sound is perceived to have a loudness of 40 dB it has a loudness of 40 phons. The acoustic level is measured at the listener's ear, relative to a standardized reference level set to the threshold of hearing at 1 kHz [31].

3. *Phon-to-Sone conversion* - is required since the increase in phons for a doubling of the subjective loudness is not constant. Therefore, a sone is defined as the increase in power which doubles the subjective loudness. The phon-to-sone conversion is given by [31]

$$\text{Sone} = \begin{cases} 2^{(\text{Phon}-40)/10}, & \text{Phon} \geq 40 \\ (\text{Phon}/40)^{2.642}, & \text{Phon} < 40 \end{cases} \tag{2.14}$$

   In practice it is difficult to determine the level in phons in the subject's ear. Therefore, it is often assumed [31, 34] that comfortable listening levels are about 60 dB above the average speech threshold, which is around 20 dB SPL and, consequently, only the upper part of 2.14 is used.

SEGSRR and BSD were chosen because they are known to give high correlation with subjective distortion measures for speech coders, with correlation factors (on a normalized scale between 0 and 1) of 0.77 and 0.9 respectively [31]. Preliminary results in [36] have shown that these measures successfully capture the distortion in speech caused by the reverberation tail, however, they are less sensitive to coloration. Consequently, Wen and Naylor [37] proposed an approach to evaluation of dereverberation algorithms, which considers the effects of coloration and the effects of the reverberation tail separately.

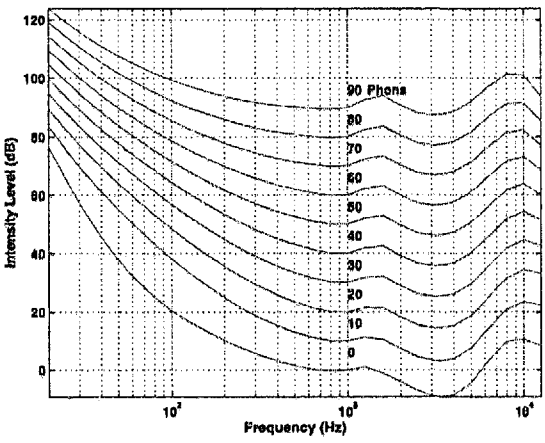

Figure 2.3: Equal loudness level contours generated with [35] according to the ISO226 standard. The numbers on the curves indicate loudness in phons.

2.5 Literature Overview

This section presents an overview of the existing literature which deals explicitly with the 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.

2.5.1 Beamforming

Beamforming techniques were among the first multichannel processing approaches for speech acquisition in noisy and reverberant environments with various developments [15]. The most fundamental of these techniques is the delay-and-sum beamformer (DSB). The observed microphone signals are delayed to compensate for different times of arrival and then weighted and summed [38,39]. The output of the DSB can be written

$$\bar{x}(n) = \sum_{m=1}^{M} w_m x_m(n - \tau_m), \tag{2.15}$$

where τ_m is the propagation delay in samples from the source to the mth sensor and w_m is the weighting applied to the mth sensor. In this way, the coherent components across channels, due to the direct paths, are added constructively while incoherent components,

due to reverberation, are attenuated [39]. Alternatively, the DSB can be interpreted as forming a beam in the direction of the desired source, which is referred to as spatial filtering [38, 39]. The design of the weights is the spatial equivalent to temporal FIR filter design; the number of microphones is related to the number of taps and the spacing between sensors is linked to the sampling frequency [38, 39]. Consequently, there is a spatial sampling criterion, analogous to the time domain Nyquist sampling theorem [40], which relates the distance between microphones to the frequency components in the signal such that spatial aliasing can be avoided. This is defined as [39]

$$\|\ell_m - \ell_{m+1}\| < \frac{c}{2f}, \tag{2.16}$$

where $\|\cdot\|$ denotes the Euclidean norm, ℓ_m is the three dimensional position vector of the mth microphone, f denotes frequency and c the speed of sound in air. Talantzis and Ward studied an alternative design of optimal weights in [41]. It can be seen from the expression in (2.16) that for broadband signals, such as speech, a linear array with uniformly spaced microphones may not be the optimal solution. Indeed, several designs have been proposed with three or four subarrays and with different distances between the microphones such that each of these subarrays covers a different bandwidth [15, 39]. The DSB can also be extended into a filter-and-sum beamformer where the simple element weights are replaced by FIR filters [39].

Several variants of the DSB exist. For example, in a frequency domain approach by Allen et al [42] the signal is processed in frequency subbands. The signals are cophased in each frequency band and the gain is adjusted according to the cross-correlation between the channels to remove incoherent components before the summation. A two-dimensional microphone array was proposed by Flanagan et al [43] which uses the 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 [44]. Adaptive beamforming approaches which automatically adjust the weights of the beamformer [39, 45] and which may also include constraints in the adapta-

tion rule [46] have been studied. Generally, beamformers have been found most efficient in the application to additive noise source suppression [15]. Reverberation can be partially reduced as will be shown in Section 2.6. However, since reverberant sound comes from all possible directions in a room [13] it will always enter the path of the beam.

Further improvements to the beamformers in reverberant environments can be achieved by multiple beamforming, 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 directions of the strong initial reflections [44, 47]. The additional reflections are treated as mirror sources in a similar way to the source-image method for simulation of room acoustics [12] described in Section 2.3. Another approach is the matched filter beamformer where the microphone signals are convolved with the time-reversed room impulse responses [44, 48–51]. However, both these methods require at least partial knowledge of the room impulse response and can be rather treated as an alternative to inverse filtering.

2.5.2 Speech Enhancement Approaches to Dereverberation

An early technique in the class of speech enhancement dereverberation was proposed by Schafer and Openheim [40, 52]. 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 this peak and attenuate it. An alternative to this was also considered where a low-pass weighting function is 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 [52].

A class of techniques emerged from the observation that the residual signal following linear prediction analysis contains the effects of reverberation, comprising peaks corresponding to excitation events in voiced speech together with additional peaks due to the reverberant channel [33, 53]. Several methods for prediction residual processing have been developed using established models of speech production. These aim to suppress the effects of reverberation without degrading the original characteristics of the residual such that the dereverberated speech can be synthesized using the processed residual and the all-pole filter resulting from prediction analysis on the reverberant speech. It is assumed

that the effect of reverberation on the AR coefficients is negligible [33].

An early idea based on the linear prediction processing was proposed in a patent by Allen [54] where the author suggested the generation of synthetic clean speech from reverberant speech by identifying the LP parameters from one or multiple reverberant observations. Griebel and Brandstein [5, 15, 34, 55] use wavelet extrema clustering to reconstruct an enhanced prediction residual. In [56] 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 [57] use time-aligned Hilbert envelopes to represent the strength of the peaks in the prediction residuals. The Hilbert envelopes are then summed and used as a weight vector which is applied to the prediction residual of one of the microphones. In [53] the authors derive a weighting function based on the signal-to-reverberant ratio in different regions of the prediction residual. Gillespie et al [58] 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 [59], where the authors 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 considerably reduce naturalness in the dereverberated speech.

A related method, proposed by Nakatani et al [60], assumes a sinusoidal speech model. First the fundamental frequency of the speech signal is identified from the reverberant observations, followed by the identification of the remaining sinusoidal components. Using the magnitudes 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 does tend to the RIR equalization filter, however, it is very long and takes over an hour of training [60].

Spectral subtraction has been widely applied, with some success, in noise reduction [6, 7] and was applied to dereverberation by Habets [61]. The author assumes a statistical model of the room impulse response constituting Gaussian noise modulated by a decaying exponential function. The decay rate of this function is governed by the rever-

beration time. It is then shown that, if the reverberation time can be blindly estimated, 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 [36, 61], provided that the assumed unknowns are available. Finally, a two stage method combining linear prediction and spectral subtraction was proposed in [59]. Here the authors use the kurtosis maximization adaptive filter to remove the early reflections and spectral subtraction for the reverberation tail.

In summary, several speech enhancement approaches to dereverberation have appeared in the literature. Most of them do not assume explicit knowledge of the room impulse response, however, they often require blind identification of other features, for example, related to the speech signal. Nevertheless, many of these methods are computationally efficient and suitable for real-time implementation.

2.5.3 Blind System Identification and Inversion for Dereverberation

The RIRs are generally unknown and, as discussed in Section 2.2, the impulse responses between the talker and the microphones change depending on the room and the talker-microphone configuration. Thus, *a priori* measurement of the impulse responses for a given room and source-microphone configuration is not a viable solution. The aim is therefore to perform blind system identification using the reverberant observations only.

Blind System Identification

Blind multichannel system identification using second order statistics is often based on the cross-relation of an observation-channel pair given by [62]: $x_1 * h_2 = (s * h_1) * h_2 = x_2 * h_1$, which leads to the system of equations $\boldsymbol{R}\boldsymbol{h} = \boldsymbol{0}$ where, in general for M channels, $\boldsymbol{R}$ is a correlation-like matrix [63] and $\boldsymbol{h} = [\boldsymbol{h}_1^T\ \boldsymbol{h}_2^T\ \ldots\ \boldsymbol{h}_M^T]^T$ is a vector of the concatenated impulse responses. It can be seen from this system of equations that the desired solution is the eigenvector corresponding to the zeroth eigenvalue in $\boldsymbol{R}$ or, in the presence of noise, the smallest eigenvalue. Several alternative solutions have been proposed. A least squares approach for solving this problem is given in [62]. An eigendecomposition method was proposed by Gürelli and Nikias [64] and also in adaptive mode. Gannot and Moonen [65] use eigendecomposition methods for blind system identification both in the fullband and

in the subband domains. Huang and Benesty proposed the use of the cross-relation as an error function and use it to derive multichannel LMS and Newton adaptive filters, both in the time [66–68] and in the frequency [63] domains.

Blind acoustic system identification of this type suffers from several limitations which are the subject of current research in the community. (i) Channels cannot be identified uniquely when they contain common zeros. (ii) The correlation matrix of the source signal $E\{\mathbf{s}(n)\mathbf{s}^T(n)\}$ must be full rank. (iii) Noise in the observations can cause the adaptive algorithms to misconverge. Some approaches have been developed to improve robustness [67, 69, 70]. (iv) Many approaches assume knowledge of the order of the unknown system. This issue has been addressed, for example, in [65] and [71]. (v) Solutions for $\boldsymbol{h}$ are normally found only to within a scale factor.

Other blind system identification approaches include Subramaniam et al [72] who proposed the use of the cepstrum for blind identification of 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 [72, 73]. One problem with this method is that it is sensitive to zeros close to the unit circle, which, as will be seen in Section 5.5.5, will often arise in acoustic systems. A method introduced by Triki and Slock [74] uses multichannel linear prediction to whiten the input signal and subsequent multichannel linear prediction which to identify the channels. Finally, in [21] an all-pole model of the channel impulse response is used (in contrast to the FIR model employed in the methods above). The RIR is assumed stationary while the source signal is assumed to be a locally stationary (but globally nonstationary) AR process. In this way, the all-pole channel parameters can be identified by observing several frames of the input signal and collecting information about 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. Nevertheless, these approaches also appear to suffer similar problems as the eigendecomposition methods, i.e. sensitivity to noise and channel order estimation.

Inverse Filtering

If the acoustic impulse responses, $\boldsymbol{h}_m$, are available, for example, from a blind system identification algorithm, dereverberation can be achieved in principle by an inverse system, $\boldsymbol{g}_m$, satisfying $\boldsymbol{h}_m^T \boldsymbol{g}_m = \kappa\delta(n-\tau)$, where κ and τ are arbitrary scale and delay factors. However, direct inversion of the acoustic channel is normally not feasible since it: (i) can be several thousand taps in length, (ii) has non-minimum phase [27] and (iii) may contain spectral nulls that 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 least squares (LS) inverse filters [75,76] can be designed by solving the optimization problem $\hat{\boldsymbol{g}}_m = \min_{\boldsymbol{g}_m} \|\boldsymbol{h}_m^T \boldsymbol{g}_m - \delta(n-\tau)\|^2$. This requires extremely long inverse filters (ideally of infinite length) and result in large processing delay. Homomorphic inverse filtering has also been investigated [12, 75, 77, 78], where the impulse response is decomposed into a minimum phase component, $\boldsymbol{h}_{\text{mp}}$ and an all-pass component, $\boldsymbol{h}_{\text{ap}}$, such that $\boldsymbol{h} = \boldsymbol{h}_{\text{mp}}^T \boldsymbol{h}_{\text{ap}}$. 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 [77,79]. An important result is that magnitude compensation only results in audible distortions in the processed speech signal [27, 77]. Only approximate equalization can be achieved with the single channel approaches.

In the multichannel case, exact inversion can be achieved with the multichannel least squares method [68, 80]. MINT was the first multichannel equalization method proposed by Miyoshi and Kaneda [80], which was implemented in a subband version [81]. Adaptive versions have also been considered [82]. If there are no common zeros between the two channel transfer functions, a pair of inverse filters, $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ can be found such that $\boldsymbol{h}_1^T \boldsymbol{g}_1 + \boldsymbol{h}_2^T \boldsymbol{g}_2 = \delta(n)$. Thus, exact equalization can be performed, with inverse filters of length similar to the channel length [68, 80]. This will be discussed in detail in Chapter 6. Undermodelled estimates of $\boldsymbol{h}_m$ are problematic 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 [83].

2.6 Dereverberation performance of the Delay-and-Sum Beamformer

The delay-and-sum beamformer attenuates some of the reflections, however, it captures all sound along the axis of the beam [44]. A second problem with the DSB is the localization of the source in highly reverberant environments [33]. Here the first issue is considered and exact knowledge of the source location is assumed. A closed form expression is presented for the expected improvement in direct-to-reverberation ratio, $\mathscr{E}\{\overline{\mathrm{DRR}}\}$, that can be achieved with the DSB compared to a single microphone, where $\mathscr{E}\{\cdot\}$ is the spatial expectation and represents the expected value over all source-microphone positions (see Section 3.2 for more detail). The following expression is evaluated

$$\mathscr{E}\{\overline{\mathrm{DRR}}\} = 10\log_{10}\left(\frac{\mathscr{E}\{\mathrm{DRR}_{\mathrm{DSB}}\}}{\mathscr{E}\{\mathrm{DRR}'\}}\right), \tag{2.17}$$

where DRR' is the DRR of the best microphone, defined as the microphone closest to the source and $\mathrm{DRR}_{\mathrm{DSB}}$ is the DRR of the RIR at the output of the delay-and-sum beamformer. The result of the analysis is summarized in Theorem 1.

Theorem 1. *The expected improvement in direct-to-reverberant ratio that can be achieved with a DSB is*

$$\mathscr{E}\{\overline{\mathrm{DRR}}\} = 10\log_{10}\left(\frac{D'^2\sum_{m=1}^{M}\sum_{l=1}^{M}\frac{1}{D_m D_l}}{\sum_{m=1}^{M}\sum_{l=1}^{M}\frac{\sin k\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_l\|}{k\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_{m+1}\|}\cos(k[D_m-D_l])},\right), \tag{2.18}$$

where D_m is the distance between the source and the mth microphone, $D' = \min_m(D_m)$ is the distance from the source to the closest microphone and $\boldsymbol{\ell}_m$ is the mth microphone three dimensional coordinate vector. The wave number is $k = 2\pi f/c$ with f denoting frequency and c being the speed of sound in air, which here is taken as $c = 344$ m/s at room temperature.

Proof. see Appendix A. □

The following observations can be made from the expression in (2.18): (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,

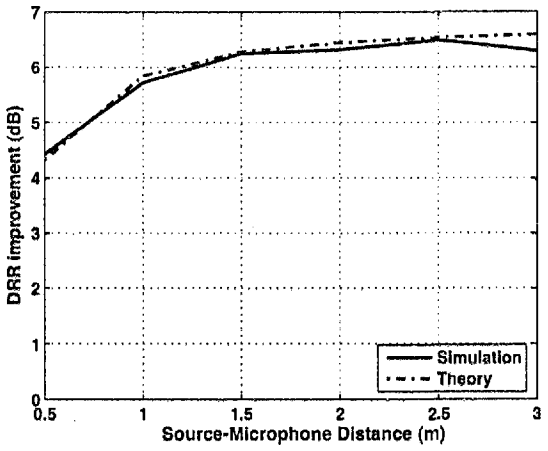

Figure 2.4: DRR improvement vs. source-microphone distance for an array of $M = 5$ microphones.

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 the improvement is infinite, i.e. perfect dereverberation is achieved.

2.6.1 Simulation Results: DSB performance

Two simulation results are presented to validate the theoretical expression in (2.18) and to gain some insight in the expected performance of the DSB for dereverberation. For these simulations, the source image method [12] was used to generate finite room impulse responses, $\boldsymbol{h}_m$. The room transfer function, $H_m(e^{j\omega})$, was then found by taking the Fourier transform of $\boldsymbol{h}_m$. A room with dimensions $(6.4 \times 5 \times 4)$ m was modelled with a linear array comprising M microphones and with uniform spacing between adjacent microphones set to $\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_{m+1}\| = 0.2$ m. The reverberation time was set to $T_{60} = 0.5$ s giving $\alpha = 0.2656$. Frequencies in the range $300 - 3400$ 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 (see Section 3.2).

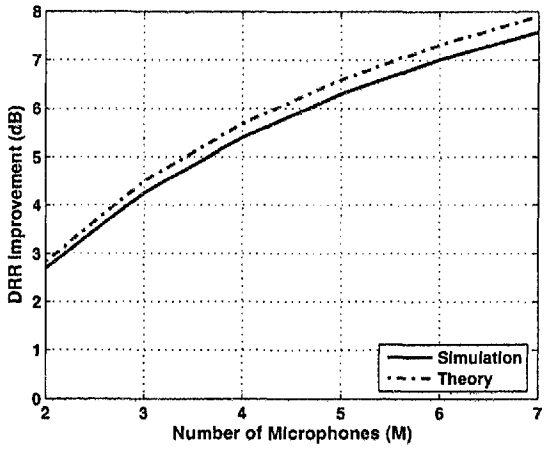

Figure 2.5: DRR improvement vs. number of microphones at a fixed distance of $D' = 2$ m.

Experiment 1: effect of source-microphone distance

In the first experiment, an array with $M = 5$ microphones is employed. The distance between the array and the source (defined here as the distance to the closest microphone) was varied from 0.5 to 3 m. The result is shown in Fig. 2.4, where the improvement in DRR calculated with the expression in (2.18) is plotted with a dashed line and the experimental result with a solid line. It is observed that the improvement increases with distance when the source is close to the array but flattens out for large distances. 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

Subsequently, the distance between the source and the array was kept fixed at 2 m while the number of microphones was increased. The result is shown in Fig. 2.5, where the improvement in DRR, calculated with the expression in (2.18), is plotted with a dashed line and the experimental result with a solid line. This result demonstrates the improvement proportionality to the number of microphones.

In summary, the delay-and-sum beamformer is a simple approach that can provide moderate reverberation reduction. Therefore, beamformers are often used as pre or post processing techniques and as a performance benchmark for new algorithms [15, 34, 58].

2.7 Related Fields of Research

The results presented in this thesis target mainly the enhancement of reverberant speech. However, there are related fields in speech and audio processing which may benefit from the provided material. These include, for example, blind source separation, virtual acoustic rendering and cross-talk cancellation, where knowledge of the room impulse responses and their inversion are useful.

2.8 Summary

This chapter has formulated the problem of reverberant speech enhancement. The properties of the room impulse responses were considered from physical, perceptual and signal processing points of view. The source-image method for room acoustic simulation was discussed. Subsequently, objective evaluation of dereverberated speech was addressed and objective measures were considered in two classes: those using the room impulse response, such as direct-to-reverberant ratio, and those using the recorded speech signals including signal-to-reverberant ratio and the Bark spectral distortion. It was concluded that the latter is preferable in the application to dereverberation since some processing approaches do not provide a processed impulse response for measurement. The existing literature on reverberant speech enhancement was reviewed categorizing existing methods into beamforming, speech enhancement and blind system identification/inversion. Beamforming approaches are often simple but with limited dereverberation capacity since reflections along the beam are not cancelled. The performance of the DSB was demonstrated with a closed form expression for the expected improvement in terms of DRR. Speech enhancement approaches, possibly in combination with a beamformer, improve the performance and are generally suitable for real-time applications. Finally, blind acoustic SIMO system identification and inversion approaches can, in principle, achieve perfect dereverberation. However, the effects of noise and long, unknown, non-minimum phase impulse responses in addition to high computational intensity constrain the practical applicability of such approaches.

Chapter 3

Statistical Analysis of the Autoregressive Modelling of Reverberant Speech

HANDS free speech input is required in many modern telecommunication applications that employ auto-regressive (AR) techniques such as linear predictive coding (LPC). When the hands-free input is obtained in enclosed reverberant spaces such as typical office rooms, the speech signal is distorted by the room transfer function. In this chapter, theoretical results from statistical room acoustics (SRA) are utilized to analyze the AR modelling of speech under these reverberant conditions. Three cases are considered: (i) AR coefficients calculated from a single observation, (ii) AR coefficients calculated jointly from a M-channel observation ($M > 1$) and (iii) AR coefficients calculated from the output of a delay-and-sum beamformer. The statistical analysis, with supporting simulations, shows that the spatial expectations of the AR coefficients for cases (i) and (ii) are approximately equal to those from the original speech, while for case (iii) there is a discrepancy due to spatial correlation between the microphones. It is subsequently demonstrated that at each individual source-microphone position (without spatial expectation), the M-channel AR coefficients from case (ii) provide the best approximation to the clean speech coefficients when microphones are closely spaced (< 0.3 m).

3.1 Background

Many hands-free telecommunication applications involving, for example, speech coding and speech enhancement, make use of autoregressive (AR) analysis techniques such as linear predictive coding (LPC). These applications are often employed in systems used inside rooms where the observed speech signal becomes reverberant due to the enclosed space. There is an interest in AR modelling of degraded speech and the properties of the AR coefficients have been studied in the context of parameter quantization noise and ambient acoustical noise [84–86]. Several dereverberation algorithms have been proposed which operate on the linear prediction (LP) residual under the explicit or implicit assumptions that the AR coefficients are not affected by reverberation [53, 54, 56–58]. These methods utilize known features of the LP residual of speech signals to attenuate components due to reverberation. Yegnanarayana and Satyanarayana [53] provided a comprehensive study on the effects of reverberation on the LP residual. This chapter presents an investigation of the effects of reverberation on the AR coefficients.

Mathematical tools from statistical room acoustics (SRA) theory [13, 83, 87] are used for the analysis of the relation between the sets of AR coefficients obtained from clean speech and those obtained from reverberant speech. SRA provides a means for describing the sound field in a room that is mathematically tractable compared to, for example, wave theory [13]. SRA has been shown useful for the analysis of signal processing techniques in reverberant environments and has recently been applied by several researchers. Radlović et al. [83], Talantzis et al. [41] and Bharitkar et al. [88] utilized SRA to investigate the robustness of channel equalization. Further, Talantzis et al. [89] investigated the performance of blind source separation, Gustafsson et al. [90] analyzed the performance of sound source localization and Ward [25] used SRA to measure the performance of acoustic crosstalk cancellation in reverberant environments.

In this study, three cases will be considered: (i) AR coefficients calculated from a single observation, (ii) AR coefficients jointly calculated from an M-channel observation ($M > 1$) and (iii) AR coefficients obtained from the output of a delay-and-sum beamformer. Extending the work in [91], it will be shown in terms of spatial expectation that the AR coefficients obtained from reverberant speech are approximately equal to those

from clean speech for cases (i) and (ii), while the AR coefficients obtained from the output of the delay-and-sum beamformer differ due to spatial correlation between the microphones. Furthermore, it will be demonstrated that the M-channel AR coefficients from (ii) provide the best estimate of the clean speech coefficients compared to the other two cases under consideration. It is believed that the results here also relate to and explain the following statement in [54]: *"...it has been recognized that any practical or typical room transfer function has certain properties that make it possible to accurately determine the speaker's vocal tract transfer function from the reverberative speech signal."* and *"...arrays of plural microphones can also be used to advantage..."* which continues *"For this case, each new microphone requires its own correlation computer. The new outputs from this computer $R'(\tau_1)$, $R'(\tau_2)$,..., $R'(\tau_{14})$ are added to the other $R(\tau)$'s of other microphones thus giving more accurate data for the coefficient computer."*

The remainder of this chapter is organized as follows. A review of the statistical room model including the conditions under which the theory is valid is presented in Section 3.2. Also, the simulation environment is defined. This is followed by a brief review of the AR modelling of speech in Section 3.3. In Section 3.4 an analysis of the effects of reverberation on the AR coefficients and on the residual signal is presented for the single channel case. Section 3.5 presents the analysis of the two multichannel AR modeling cases. Simulation results are presented in Section 3.6 and finally conclusions regarding AR modeling of reverberant speech are drawn Section 3.7.

3.2 Statistical Room Acoustics

In this Section, the statistical models of room reverberation and the conditions under which this is assumed valid are summarized. Within the framework of SRA, the amplitudes and phases of all reflected acoustic plane waves in a room are randomly distributed such that they form a uniform, diffuse sound field at any point in the room. Subsequently, it is assumed that the room transfer function (RTF) from the source to the mth microphone can be expressed as the sum of a direct component, $H_{\mathrm{d},m}(e^{j\omega})$ and a reverberant component,

$H_{\mathrm{r},m}(e^{j\omega})$, such that

$$H_m(e^{j\omega}) = H_{\mathrm{d},m}(e^{j\omega}) + H_{\mathrm{r},m}(e^{j\omega}), \quad m = 1, 2, \ldots, M. \tag{3.1}$$

The direct component is defined here as only the direct-path signal between the source and the microphone and, consequently, the reverberant component is due to all (early and late) reflections.

Under the conditions stated at the end of this section, and due to the different propagation directions and the random relation of the phases of the direct component and all the reflected waves, it can be assumed that the direct and the reverberant components are uncorrelated [13,87]. Hence, the spatial expectation of the cross terms of the squared magnitude of (3.1) is zero [83] and the spatially expected energy density spectrum of the RTF can be written

$$\mathcal{E}\{|H_m(e^{j\omega})|^2\} = |H_{\mathrm{d},m}(e^{j\omega})|^2 + \mathcal{E}\{|H_{\mathrm{r}}(e^{j\omega})|^2\}, \tag{3.2}$$

where $\mathcal{E}\{\cdot\}$ is the spatial expectation operator, with the spatial expectation defined over all allowed microphone-source positions in a room [87,90]. Only the reverberant component varies with position and its spatial expectation is independent of the microphone index m. The computation of $\mathcal{E}\{\cdot\}$ is described in Section 3.2.1. The direct component of the RTF is the free-space Green's function and is defined as [87]

$$H_{\mathrm{d},m}(e^{j\omega}) = \frac{e^{jkD_m}}{4\pi D_m}, \tag{3.3}$$

where D_m is the distance from the source to the mth microphone and $k = 2\pi f/c$ is the wave number with f denoting frequency and c the speed of sound in air, which here is taken at room temperature as $c = 344$ m/s. From SRA, the expected density spectrum of the reverberant component is given by [25,83]

$$\mathcal{E}\{|H_{\mathrm{r}}(e^{j\omega})|^2\} = \left(\frac{1-\alpha}{\pi A \alpha}\right), \tag{3.4}$$

where A is the total surface area of the room and α the average wall absorption coefficient.

The spatial cross-correlation of the reverberant paths between the mth and the lth acoustic channels has been shown to be [25]

$$\mathcal{E}\{H_{r,m}(e^{j\omega})H^*_{r,l}(e^{j\omega})\} = \left(\frac{1-\alpha}{\pi A\alpha}\right)\frac{\sin k\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_l\|}{k\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_l\|}, \tag{3.5}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\boldsymbol{\ell}_m$ is the three dimensional position vector of the mth microphone, with the origin at $(x, y, z) = (0, 0, 0)$. For the underlying principles in the derivation of (3.4) and (3.5) please refer to [25, 83] and the references therein.

These approximations are known to represent closely the acoustic properties of a room provided that the following conditions are satisfied [13, 83]:

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 defined as

$$f_{Sch} = 2000\sqrt{\frac{T_{60}}{V}}\ \text{Hz}, \tag{3.6}$$

   where T_{60} is the reverberation time and V is the volume of the room in m^3.

3. Speakers and microphones are situated in the room interior, at least a half-wavelength from the surrounding walls.

These conditions usually hold for most practical situations over the significant speech bandwidth. Also, image method [12] simulations and measured impulse responses of a real room have been demonstrated to coincide closely with SRA theory [90].

3.2.1 Experimental Environment

A room is considered with a single source and an array of microphones as shown in Fig. 3.1. All results presented in this chapter are based on computer simulations with the simulated environment defined as follows. The dimensions of the room were set to $6.4 \times 5 \times 4$ m. These dimensions were specifically chosen to conform with the ratio (1:1.25:1.6), as in [83], in order to obtain the best approximation of a diffuse sound field so as to satisfy the

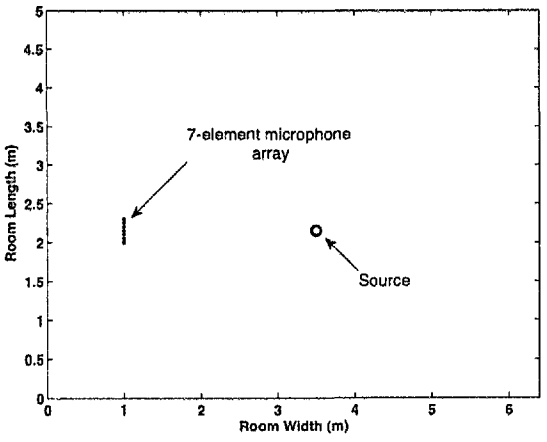

Figure 3.1: Plane view of the simulated room environment with the initial position of the microphones (·) and the source (○).

conditions above. The microphones were positioned in a linear array with the distance between adjacent microphones uniformly set to $\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_{m+1}\| = 0.05$ m. The source was at a distance, $D_4 = 2.5$ m, from the centre of the array. The source and the microphones were assumed omnidirectional and were always at least a half-wavelength from the surrounding walls, where the wavelength is taken with respect to the lowest frequency component in the signal. The source-image method [12, 23], was used to generate the RIRs, $\boldsymbol{h}_m$. The RTF, $H_m(e^{j\omega})$, is then found by taking the Fourier transform of $\boldsymbol{h}_m$. Anechoic speech samples were taken from the APLAWD database [92] and all the signals under consideration were bandlimited to $0.3 - 7$ kHz with sampling frequency $f_s = 16$ kHz.

To compute the spatial expectation, $\mathcal{E}\{\cdot\}$, the method used by Radlović et al [83] and Gustafsson et al [90] was utilized. An initial position for the source, $\boldsymbol{y}_0$, and for each of the microphones, $\boldsymbol{\ell}_{m,0}$, was selected. A random translation vector, $\boldsymbol{\theta}_i$, and a random rotation matrix, $\boldsymbol{\Theta}_i$, were generated and applied to the initial coordinates of the source-receiver configuration to obtain the ith realization coordinates $\boldsymbol{y}_i = \boldsymbol{\Theta}_i \boldsymbol{y}_0 + \boldsymbol{\theta}_i$ and $\boldsymbol{\ell}_{m,i} = \boldsymbol{\Theta}_i \boldsymbol{\ell}_{m,0} + \boldsymbol{\theta}_i$. In this way, the distance between the source and the microphones and between successive microphones is kept constant for all $i = 1, 2, \ldots, \mathcal{N}$. An estimate of $\mathcal{E}\{\cdot\}$ is obtained by averaging the outcomes for all $\mathcal{N}$ experiments.

3.3 AR Modelling of Speech

This section provides a brief review of the AR modelling of speech using linear prediction [93, 94]. A speech signal, $s(n)$, can be expressed as a linear combination of its p past samples using a linear predictor [93]

$$s(n) = -\boldsymbol{a}^T \boldsymbol{s}(n-1) + e(n), \tag{3.7}$$

where $\boldsymbol{a} = [a_1\ a_2\ \ldots\ a_p]^T$ is a $p \times 1$ vector of AR coefficients, $\boldsymbol{s}(n-1) = [s(n-1)\ s(n-2)\ \ldots\ s(n-p)]^T$ is a vector of observation samples at time n, $e(n)$ is the LP residual and p is the prediction order. The prediction error filter and the all-pole predictor are respectively

$$A(z) = 1 + \boldsymbol{a}^T \boldsymbol{z}, \tag{3.8}$$

and

$$V(z) = \frac{1}{A(z)}. \tag{3.9}$$

where $\boldsymbol{z} = [z^{-1}\ z^{-2}\ \ldots\ z^{-p}]$.

The AR coefficients can be obtained by minimizing the sum of the squared prediction error,

$$\begin{aligned} J &= \sum_{n=-\infty}^{\infty} e^2(n) \\ &= \sum_{n=-\infty}^{\infty} \left(s(n) + \boldsymbol{a}^T \boldsymbol{s}(n-1)\right)^2, \end{aligned} \tag{3.10}$$

with respect to each of the coefficients in $\boldsymbol{a}$. Equivalently, by Parseval's theorem a frequency domain formulation of the error in (3.10) can be expressed as [93]

$$\begin{aligned} J &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2\, d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |1 + \boldsymbol{a}^T \boldsymbol{d}|^2 |S(e^{j\omega})|^2\, d\omega, \end{aligned} \tag{3.11}$$

where $S(e^{j\omega})$ and $E(e^{j\omega})$ are the discrete-time Fourier transforms of $s(n)$ and $e(n)$ respectively and $\boldsymbol{d} = [e^{-j\omega}\ e^{-j2\omega} \ldots, e^{-jp\omega}]^T$ is a $p \times 1$ DFT vector. The optimum set of p AR

coefficients that minimize the error J is

$$\boldsymbol{a}_{\text{opt}} = \arg\min_{\boldsymbol{a}} J = -\boldsymbol{R}^{-1}\boldsymbol{r}, \tag{3.12}$$

where

$$\boldsymbol{R} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S(e^{j\omega})|^2 \boldsymbol{d}\boldsymbol{d}^H \, d\omega \tag{3.13}$$

is a $p \times p$ autocorrelation matrix and

$$\boldsymbol{r} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S(e^{j\omega})|^2 \boldsymbol{d} \, d\omega \tag{3.14}$$

is a $p \times 1$ vector of autocorrelation coefficients. In practice, the error signal is evaluated over finite windowed frames [93], however for the work presented in this chapter the effects of the window will not be considered.

3.4 Single-Channel AR Modeling of Reverberant Speech

In this section, AR modelling of reverberant speech using linear prediction is considered [93,94] and the analysis of the effects of reverberation on the AR coefficients obtained from a single channel is presented. Also, the consequences this has on the linear prediction residual are discussed.

3.4.1 Effect of Reverberation on the AR Coefficients

Consider a speech signal, $s(n)$, produced at a point in a noiseless, reverberant room. The observation by a single microphone positioned at some distance from the speaker is denoted

$$x(n) = \boldsymbol{h}^T \boldsymbol{s}(n), \tag{3.15}$$

where $\boldsymbol{h} = [h_0 \; h_1 \; \ldots \; h_{L-1}]^T$ is the L-tap impulse response of the acoustic channel from the source to the microphone and $\boldsymbol{s}(n) = [s(n) \; s(n-1) \; \ldots \; s(n-L+1)]^T$ is the input vector at time n. This is equivalent to (2.2) with $\nu_m(n) = 0$ and since a single channel is considered ($M = 1$) the subscript m is omitted. The relation between the AR coefficients obtained by linear prediction from $s(n)$ and those from $x(n)$ is summarized in Theorem 2.

Theorem 2. *Let* $\boldsymbol{a}_{opt} = [a_{opt,1}\ a_{opt,2}\ \ldots\ a_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the clean speech signal,* $s(n)$*, and let* $\boldsymbol{b}_{opt} = [b_{opt,1}\ b_{opt,2}\ \ldots\ b_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the reverberant speech signal,* $x(n)$*. The spatially expected values of the reverberant speech AR coefficients are approximately equal to those of the AR coefficients calculated from clean speech, i.e.,*

$$\mathscr{E}\{\boldsymbol{b}_{opt}\} \cong \boldsymbol{a}_{opt}. \tag{3.16}$$

Proof. LP analysis is applied on the reverberant speech signal, $x(n)$, to obtain the optimum set of AR coefficients

$$\boldsymbol{b}_{\text{opt}} = -\boldsymbol{Q}^{-1}\boldsymbol{q}, \tag{3.17}$$

with

$$\boldsymbol{Q} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{j\omega})|^2 \boldsymbol{d}\boldsymbol{d}^H\, d\omega, \tag{3.18}$$

$$\boldsymbol{q} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{j\omega})|^2 \boldsymbol{d}\, d\omega, \tag{3.19}$$

where $\boldsymbol{Q}$ is a $p \times p$ autocorrelation matrix and $\boldsymbol{q}$ is a $p \times 1$ vector with autocorrelation coefficients.

In order to study the LP coefficients of reverberant speech, the spatial expectation is taken on both sides of (3.17)

$$\mathscr{E}\{\boldsymbol{b}\} = -\mathscr{E}\{\boldsymbol{Q}^{-1}\boldsymbol{q}\}, \tag{3.20}$$

However, we would like to consider the expectation of each term of (3.20). Adopting the approach used in [83] and [41], consider a function, $f(x_1,\ x_2,\ \ldots,\ x_n)$, of random variables [95] with mean values $E\{x_i\} = \mu_i$, which is written here $f(x)$ for brevity. Letting $f'(x) = \partial(f(x))/\partial x_i|_{x=\mu}$, the Taylor series expansion of $f(x)$ about the mean, μ, is $f(x) = f(\mu)+\sum_{i=1}^{n} f'(\mu)(x_i-\mu_i)+\breve{f}(x)$, where $\breve{f}(x)$ are the second order terms and above. All the partial derivatives up to the first order vanish [83] at $(\mu_1,\ \mu_2,\ \ldots,\ \mu_n)$ and, consequently, it can be seen that $\mathscr{E}\{f(x)\} \cong f(\mathscr{E}\{x\})$ holds up to the zeroth order of approximation. Using this relationship, (3.20) can be written

$$\mathscr{E}\{\boldsymbol{b}\} \cong -\mathscr{E}\{\boldsymbol{Q}\}^{-1}\mathscr{E}\{\boldsymbol{q}\}. \tag{3.21}$$

This reduces the problem to studying the properties of the AR coefficients in terms of the autocorrelation functions. In practice, the accuracy of this approximation will depend on the estimation of the mean value of the random variables.

Now, consider the spatial expectation of the uth element of $\boldsymbol{q}$ in (3.19)

$$\begin{aligned}\mathcal{E}\{q_u\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{E}\{|X(e^{j\omega})|^2\}e^{-j\omega u}\,d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{E}\{|H(e^{j\omega})|^2\}|S(e^{j\omega})|^2e^{-j\omega u}\,d\omega, \end{aligned} \tag{3.22}$$

for $u = 1, 2, \ldots, p$. The term $S(e^{j\omega})$ is taken outside the spatial expectation since it is independent of the source-microphone position.

From (3.2), (3.3) and (3.4) the SRA expression for the expected energy density spectrum of the room transfer function is

$$\mathcal{E}\{|H(e^{j\omega})|^2\} = \frac{1}{(4\pi D)^2} + \left(\frac{1-\alpha}{\pi A\alpha}\right) = \gamma. \tag{3.23}$$

Since γ is independent of frequency, by substitution with (3.23) into (3.22) the spatial expectation of the uth element of $\boldsymbol{q}$ can be written

$$\begin{aligned}\mathcal{E}\{q_u\} &= \frac{\gamma}{2\pi}\int_{-\pi}^{\pi}|S(e^{j\omega})|^2e^{-j\omega u}\,d\omega, \quad u = 1, 2, \ldots, p \\ &= \gamma r_u, \end{aligned} \tag{3.24}$$

where r_u is the uth component of the clean speech autocorrelation vector $\boldsymbol{r}$. By similar reasoning the (u, v)th component of $\boldsymbol{Q}$ in (3.18) becomes

$$\mathcal{E}\{Q_{u,v}\} = \gamma R_{u,v}, \quad u, v = 1, 2, \ldots, p \tag{3.25}$$

where $R_{u,v}$ is the (u, v)th component of the clean speech autocorrelation matrix $\boldsymbol{R}$. Substituting the results in (3.24) and (3.25) into (3.21) gives (3.16). □

This result states that if LP analysis is applied to reverberant speech, the coefficients $\boldsymbol{a}_{\text{opt}}$ and $\boldsymbol{b}_{\text{opt}}$ are not necessarily equal at a single observation point. However, in terms of spatial expectation, the AR coefficients from reverberant speech are approximately equal to those from clean speech. The accuracy of the approximation depends on the accuracy of estimation of the spatial expectation of the autocorrelation function.

3.4.2 Effect of Reverberation on the Prediction Residual

Consider a frequency domain formulation of the source-filter model described in Section 3.3. The speech signal is expressed as

$$S(e^{j\omega}) = E(e^{j\omega})V(e^{j\omega}), \tag{3.26}$$

where $E(e^{j\omega})$ is the Fourier transform of the LP residual and $V(e^{j\omega})$ is the transfer function of the all-pole filter from (3.9) evaluated for $z = e^{j\omega}$.

Now consider the speech signal produced in a noiseless reverberant room defined in (2.3), which in the frequency domain leads to

$$\begin{aligned} X(e^{j\omega}) &= S(e^{j\omega})H(e^{j\omega}) \\ &= E(e^{j\omega})V(e^{j\omega})H(e^{j\omega}). \end{aligned} \tag{3.27}$$

Referring to (3.16), an inverse filter, $B(e^{j\omega}) = 1 + \boldsymbol{b}^T\boldsymbol{d}$, can be obtained such that $\mathscr{E}\{B(e^{j\omega})\} \cong A(e^{j\omega})$, where $A(e^{j\omega})$ is given by (3.8) for $z = e^{j\omega}$. Filtering the reverberant speech signal with this inverse filter, whose coefficients are obtained from the reverberant speech signal results in

$$\hat{E}(e^{j\omega}) \cong E(e^{j\omega})H(e^{j\omega}), \tag{3.28}$$

where $\hat{E}(e^{j\omega})$ is the Fourier transform of the LP residual, $\hat{e}(n)$, from the reverberant speech signal. Thus, in the time domain, the LP residual obtained from reverberant speech is approximately equal to the clean speech residual convolved with the room impulse response. The approximation in (3.28) arises from the AR modelling. Therefore, if the AR coefficients used were identical to those from clean speech, the approximation would be an equivalence.

In summary, it has been shown that the AR coefficients obtained from reverberant speech are approximately equal to those from clean speech in terms of spatial expectation. Furthermore, the LP residual obtained from a reverberated speech signal is approximately equal to the clean speech residual convolved with the room impulse response. This approximation depends on the accuracy of the estimation of the AR coefficients. Intuitively,

the result in (3.16) 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 AR coefficients than use of a single observation alone. This now motivates the study of multichannel AR modeling in the following section.

3.5 Multichannel AR Modeling of Reverberant Speech

Several microphone array techniques have been applied as preprocessing in speech applications, proving advantageous to single channel algorithms [15]. This Section investigates the use of a microphone array to obtain the AR coefficients and how these compare to the AR coefficients from clean speech. Two alternative approaches are considered. In the first alternative, the AR coefficients are obtained by formulating an estimation procedure that jointly minimizes the squared errors over all M channels. In the second alternative, the AR coefficients are obtained from the output of an M-channel array using delay-and-sum beamforming.

3.5.1 M-Channel AR Coefficients

The speech signal observed at the mth microphone in an array of M microphones can be expressed as

$$x_m(n) = \boldsymbol{h}_m^T \boldsymbol{s}(n). \tag{3.29}$$

where $\boldsymbol{h}_m = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T$ is the L-tap room impulse response from the source to the mth microphone.

In linear prediction terms, the observation at the mth sensor from (3.29) can be written

$$x_m(n) = -\boldsymbol{b}_m^T \boldsymbol{x}_m(n-1) + e_m(n), \quad m = 1, 2, \ldots, M \tag{3.30}$$

where $\boldsymbol{b}_m = [b_{m,1}\ b_{m,2}\ \ldots\ b_{m,p}]^T$ are the prediction coefficients, $\boldsymbol{x}_m(n-1) = [x(n-1)\ x(n-2)\ \ldots\ x(n-p)]^T$ is the mth microphone observation vector at time n and $e_m(n)$ is the prediction residual obtained from the mth microphone signal. From (3.30), a joint

M-channel error function can be formulated as [15]

$$J_M = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=-\infty}^{\infty} e_m^2(n). \tag{3.31}$$

The optimum set of coefficients that minimize this error, similarly to (3.12), are given by

$$\hat{b}_{\text{opt}} = -\hat{Q}^{-1}\hat{q} \tag{3.32}$$

with

$$\hat{Q} = \frac{1}{M}\sum_{m=1}^{M} Q_m \tag{3.33}$$

and

$$\hat{q} = \frac{1}{M}\sum_{m=1}^{M} q_m. \tag{3.34}$$

where $\hat{Q}$ and $\hat{q}$ are, respectively, the $p \times p$ mean autocorrelation matrix and the $p \times 1$ mean autocorrelation vector across the M microphones. The relation between the clean speech coefficients and the coefficients obtained using (3.32) is summarized in Corollary 1.

Corollary 1. *Replacing* (3.18) *and* (3.19) *with their averages considered over M microphones* (3.33) *and* (3.34) *and then following the steps of the proof of Theorem 2, it can be shown that the spatial expectation of the AR coefficients obtained from minimization of* (3.31) *is approximately equal to those from clean speech. That is*

$$\mathcal{E}\{\hat{b}_{opt}\} \cong a. \tag{3.35}$$

This result implies that the optimal AR coefficients obtained using a spatial expectation over M channels are equivalent to the spatial expectation of the AR coefficients in the single microphone case in (3.16). However at each individual position, the M-channel case provides a more accurate estimation of the clean speech AR coefficients than that obtained with a single reverberant channel, as will be shown by simulations in Section 3.6. This is because the averaging of the autocorrelation functions in (3.33) and (3.34) is equivalent in effect to the calculation of the spatial expectation operation in the single channel case (3.21).

3.5.2 AR Coefficients from a Beamformer Output

The output of a delay-and-sum beamformer can be written [38]

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n-\tau_m), \tag{3.36}$$

where τ_m is the propagation delay in samples from the source to the mth sensor. Assuming that the time-delays of arrival are known for all microphones, linear prediction can be performed on the beamformer output, $\bar{x}(n)$, following the approach in Section 3.4.1. This is summarized in Theorem 3.

Theorem 3. *Let* $\boldsymbol{a}_{opt} = [a_{opt,1}\ a_{opt,2}\ \ldots\ a_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the clean speech signal,* $s(n)$*, and let* $\bar{\boldsymbol{b}}_{opt} = [\bar{b}_{opt,1}\ \bar{b}_{opt,2}\ \ldots\ \bar{b}_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the DSB output,* $\bar{x}(n)$*. The spatial expectation of the AR coefficients calculated by linear prediction from the output of the DSB is*

$$\mathcal{E}\{\bar{\boldsymbol{b}}_{opt}\} \cong \boldsymbol{\mathcal{T}}\boldsymbol{a}_{opt} - \boldsymbol{t}, \tag{3.37}$$

with $\boldsymbol{\mathcal{T}} = \boldsymbol{I} - \frac{1}{\bar{\gamma}}\boldsymbol{R}^{-1}\boldsymbol{\Gamma}\left(\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Gamma}^H\frac{1}{\bar{\gamma}}\boldsymbol{R}^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^H$ *and* $\boldsymbol{t} = (\bar{\gamma}\boldsymbol{R}+\boldsymbol{\Xi})^{-1}\boldsymbol{\xi}$*, where these terms are defined in the following proof.*

Proof. Consider a speech signal, $s(n)$, observed using M microphones and combined using a DSB to give a signal $\bar{x}(n)$. In the frequency domain this can be expressed as

$$\begin{aligned}\bar{X}(e^{j\omega}) &= \left(\frac{1}{M}\sum_{m=1}^{M} H_m(e^{j\omega})e^{-j2\pi f\tau_m}\right)S(e^{j\omega}) \\ &= \bar{H}(e^{j\omega})S(e^{j\omega}),\end{aligned} \tag{3.38}$$

where $\bar{X}(e^{j\omega})$ is the Fourier transform of $\bar{x}(n)$, $S(e^{j\omega})$ is the Fourier transform of $s(n)$, $H_m(e^{j\omega})$ is the RTF with respect to the mth microphone and $\bar{H}(e^{j\omega})$ is the averaged RTF at the DSB output. The AR coefficients, $\bar{\boldsymbol{b}}_{\text{opt}}$, at the beamformer output are calculated as in Section 3.3

$$\bar{\boldsymbol{b}}_{\text{opt}} = -\bar{\boldsymbol{Q}}^{-1}\bar{\boldsymbol{q}}, \tag{3.39}$$

with

$$\bar{\boldsymbol{Q}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{H}(e^{j\omega})|^2|S(e^{j\omega})|^2\boldsymbol{d}\boldsymbol{d}^H\, d\omega \tag{3.40}$$

and

$$\bar{q} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{H}(e^{j\omega})|^2 |S(e^{j\omega})|^2 \boldsymbol{d}\, d\omega, \tag{3.41}$$

where $\bar{Q}$ is a $p \times p$ autocorrelation matrix and $\bar{q}$ is a $p \times 1$ autocorrelation vector.

As in the proof for Theorem 2, it is desired to find the spatial expectation of the AR coefficients obtained from the DSB,

$$\mathcal{E}\{\bar{b}_{\text{opt}}\} \cong -\mathcal{E}\{\bar{Q}\}^{-1}\mathcal{E}\{\bar{q}\} \tag{3.42}$$

and the expectation of uth component of $\bar{q}$ is considered

$$\mathcal{E}\{\bar{q}_u\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{E}\{|\bar{H}(e^{j\omega})|^2\} |S(e^{j\omega})|^2 e^{-j\omega u}\, d\omega, \tag{3.43}$$

The expected energy density spectrum of the averaged RTFs can be written

$$\begin{aligned}\mathcal{E}\{|\bar{H}(e^{j\omega})|^2\} &= \frac{1}{M^2}\left[\sum_{m=1}^{M} \mathcal{E}\{|H_m(e^{j\omega})|^2\}\right. \\ &+ \left.\sum_{m=1}^{M}\sum_{\substack{l=1 \\ l\neq m}}^{M} \mathcal{E}\{H_m(e^{j\omega})H_l^*(e^{j\omega})\} e^{-j2\pi f(\tau_m - \tau_l)}\right].\end{aligned} \tag{3.44}$$

From (3.2), (3.3) and (3.4), the expected energy density for the mth channel is

$$\mathcal{E}\{|H_m(e^{j\omega})|^2\} = \frac{1}{(4\pi D_m)^2} + \left(\frac{1-\alpha}{\pi A \alpha}\right), \tag{3.45}$$

and from (3.3) and (3.5), the expected cross-correlation between the mth and the ith microphones is

$$\mathcal{E}\{H_m(e^{j\omega})H_i^*(e^{j\omega})\} = \frac{e^{jk(D_m - D_n)}}{16\pi^2 D_m D_l} + \left(\frac{1-\alpha}{\pi A \alpha}\right)\frac{\sin k\|\ell_m - \ell_l\|}{k\|\ell_m - \ell_l\|}. \tag{3.46}$$

Next, without loss of generality, let $\tau_m = D_m/c$. Then, by substituting (3.45) and (3.46) into (3.44) and rearranging the terms, the following expression for the mean energy density

at the DSB output is obtained

$$\mathscr{E}\{|\bar{H}(e^{j\omega})|^2\} = \bar{\gamma} + \psi(k), \tag{3.47}$$

with

$$\bar{\gamma} = \frac{1}{(4\pi M)^2} \sum_{m=1}^{M} \sum_{l=1}^{M} \frac{1}{D_m D_l} + \left(\frac{1-\alpha}{M\pi A\alpha}\right) \tag{3.48}$$

and

$$\psi(k) = \left(\frac{1-\alpha}{M^2\pi A\alpha}\right) \sum_{m=1}^{M} \sum_{\substack{i=1 \\ i\neq m}}^{M} \frac{\sin k\|\ell_m - \ell_l\|}{k\|\ell_m - \ell_l\|} \cos(k[D_m - D_l]), \tag{3.49}$$

where $\bar{\gamma}$ is a frequency independent component and $\psi(k)$ is a component due to spatial correlation.

Now, let

$$\xi_u = \frac{1}{2\pi} \int_{-\pi}^{\pi} \psi(k)|S(e^{j\omega})|^2 e^{-j\omega u}\, d\omega \tag{3.50}$$

and

$$\Xi_{u,v} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \psi(k)|S(e^{j\omega})|^2 e^{-j\omega(u-v)}\, d\omega \tag{3.51}$$

be the uth element of a vector $\boldsymbol{\xi}$ and the (u,v)th element of a matrix $\boldsymbol{\Xi}$ respectively. The expected value of the uth element of $\bar{\boldsymbol{q}}$ from (3.41) then becomes

$$\mathscr{E}\{\bar{q}_u\} = \bar{\gamma} r_u + \xi_u, \quad u = 1, 2, \ldots, p, \tag{3.52}$$

where r_u is the uth element of the vector $\boldsymbol{r}$ in (3.14). Similarly, the expected value of the (u,v)th component of $\bar{\boldsymbol{Q}}$ from (3.40) is

$$\mathscr{E}\{\bar{Q}_{u,v}\} = \bar{\gamma} R_{u,v} + \Xi_{u,v}, \quad u, v = 1, 2, \ldots, p, \tag{3.53}$$

where $R_{u,v}$ is the (u,v)th element of the matrix $\boldsymbol{R}$ in (3.13).

The spatial expectation of the set of coefficients for the DSB output is therefore

$$\mathscr{E}\{\bar{\boldsymbol{b}}\} \cong -(\bar{\gamma}\boldsymbol{R} + \boldsymbol{\Xi})^{-1}(\bar{\gamma}\boldsymbol{r} + \boldsymbol{\xi}), \tag{3.54}$$

Since $\boldsymbol{\Xi}$ is a Hermitian symmetric matrix, it can be factored as

$$\boldsymbol{\Xi} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^H, \tag{3.55}$$

where $\mathbf{\Gamma}$ is a matrix of eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. Using the matrix inversion lemma [96] leads to the following relation

$$(\bar{\gamma}\boldsymbol{R}+\boldsymbol{\Xi})^{-1}=\frac{1}{\bar{\gamma}}\boldsymbol{R}^{-1}-\frac{1}{\bar{\gamma}^2}\boldsymbol{R}^{-1}\mathbf{\Gamma}\left(\mathbf{\Lambda}^{-1}-\mathbf{\Gamma}^H\frac{1}{\bar{\gamma}}\boldsymbol{R}^{-1}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^H\boldsymbol{R}^{-1}. \tag{3.56}$$

Finally, substituting the result from (3.56) into (3.54) gives the result in (3.37). □

Theorem 3 states that in terms of spatial expectation, the AR coefficients obtained by LP 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 (3.49) that the inter-channel correlation and its significance are governed by the reverberation time, the distance between adjacent microphones, the source-microphone separation and 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. From (3.49) it is evident that the term $\psi(k)$ 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 $\boldsymbol{\mathcal{T}}$ tends to the identity matrix $\boldsymbol{I}$ and the vector $\boldsymbol{t}$ tends to zero and the result in (3.37) tends to the result in (3.16). Furthermore, if estimates of $\boldsymbol{\mathcal{T}}$ and $\boldsymbol{t}$ were available, and since $\boldsymbol{\mathcal{T}}$ is a square matrix, the effects of the spatial cross-correlation could be compensated as $\boldsymbol{a}_{\text{opt}} \cong \boldsymbol{\mathcal{T}}^{-1}(\mathscr{E}\{\bar{\boldsymbol{b}}_{\text{opt}}\}+\boldsymbol{t})$. 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 $\psi(k)=0, \forall k$ and consequently $\boldsymbol{\Xi}$ and $\boldsymbol{\xi}$ of (3.50) and (3.51) are equal to zero. Thus, the matrix $\boldsymbol{\mathcal{T}}$ becomes the identity matrix $\boldsymbol{I}$ and the vector $\boldsymbol{t}$ is zero which results in the expression in (3.37) becoming equivalent to the single channel case in (3.16).

3.6 Simulations and Results

Having established the theoretical relationship between the AR 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,

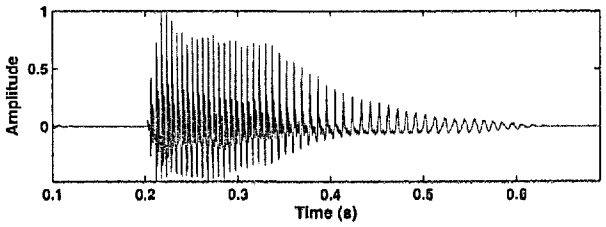

Figure 3.2: Speech sample used in the experiments comprising the time-domain waveform of the diphthong /eI/ as in the alphabet letter 'a' uttered by a male talker.

two specific points will be demonstrated: 1. On average over all positions in the room, the AR coefficients obtained from a single microphone as in case (i) and those calculated from M-microphones as in case (ii) are not affected by reverberation while the AR coefficients from the DSB become more dissimilar from the clean speech coefficients with increased reverberation time. 2. The M-channel AR coefficients are the most accurate estimates of the clean speech AR coefficients from the three cases studied.

The Itakura distance was used as a similarity measure between two sets of AR coefficients and is defined as [94]:

$$d_I = \log\left(\frac{\hat{\boldsymbol{a}}^T \boldsymbol{R} \hat{\boldsymbol{a}}}{\boldsymbol{a}^T \boldsymbol{R} \boldsymbol{a}}\right), \tag{3.57}$$

where $\boldsymbol{R}$ is the autocorrelation matrix of the speech signal defined in (3.13), $\boldsymbol{a}$ is the set of clean speech coefficients and $\hat{\boldsymbol{a}}$ are the 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 the clean speech and thus $d_I \geqslant 0$. For the experiments, the diphthong /eI/ as in the alphabet letter 'a' uttered by a male talker was used as an example and is depicted in Fig. 3.2. The LP analysis was performed using selective linear prediction [97] 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 [94]. 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 only in order to avoid errors due to bandlimiting filters.

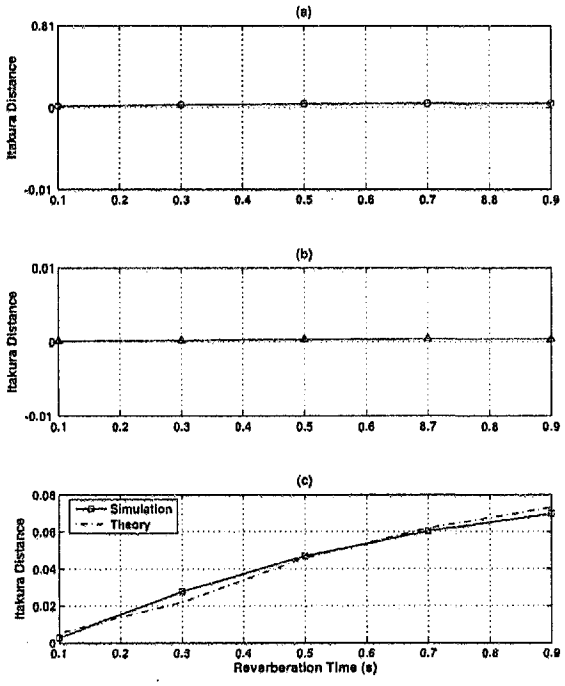

Figure 3.3: Itakura distance vs. reverberation time for the spatially expected AR coefficients of (a) a single channel, (b) M=7 channels and (c) DSB output simulation (square) and the theoretical expression for the DSB output (3.37) (dashed).

Experiment 1:

The spatial expectation was calculated from $\mathcal{N} = 200$ realizations of the source-array positions and an average autocorrelation function was calculated for each of the cases under consideration. This was repeated, varying the reverberation time, T_{60}, from 0.1 s to 0.9 s. For each case the Itakura distance was calculated for the spatial expectation of the coefficients. Figure 3.3 shows the Itakura distance of the spatially expected AR coefficients versus reverberation time for (a) a single channel, (b) $M = 7$ channels and (c) the DSB output simulation and the theoretical expression for the DSB output in (3.37) (dashed). 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 proportionally to the reverberation time.

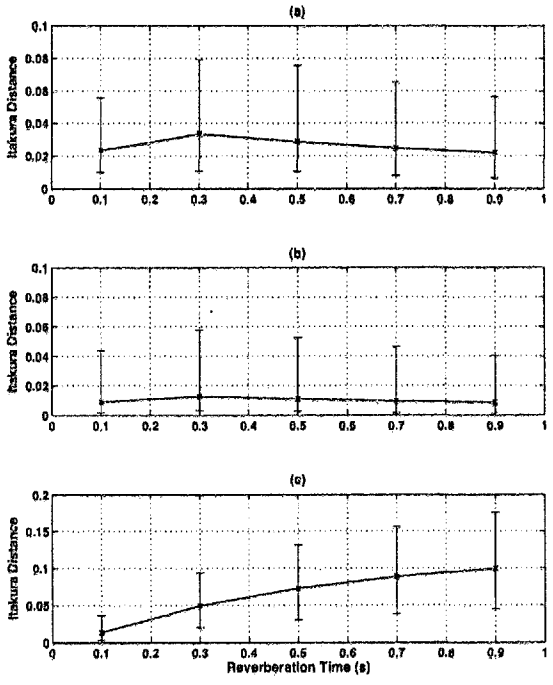

Figure 3.4: Itakura distance vs. reverberation time in terms of the AR coefficients for each individual outcome for (a) a single channel, (b) M-channels and (c) the DSB output. The error bars indicate the maximum and minimum error while crosses show the mean value.

Experiment 2:

This experiment illustrates the individual outcomes for the three cases at the $\mathcal{N} = 200$ different locations. Thus, using the same conditions as in Experiment 1, the AR coefficients were computed at each individual source-array position using (3.17), (3.32) and (3.39) and the Itakura distance was calculated. Figure 3.4 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 AR coefficients. It can also be seen that the estimation error for the AR coefficients obtained from the DSB output becomes greater with increasing reverberation

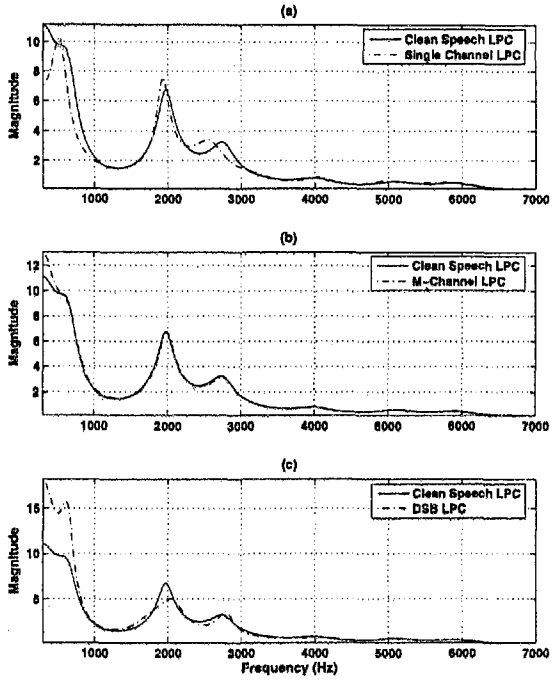

Figure 3.5: Spectral envelopes calculated from the AR coefficients of clean speech compared with spectral envelopes obtained from the AR coefficients of (a) a single channel, (b) $M = 7$ channels and (c) the DSB output.

time. This result may appear counterintuitive, however, it conforms with the theoretical expression in (3.37) and will be clarified further with the following experiment. Figure 3.5 shows examples of the spectral envelopes from the AR 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.

Experiment 3:

In line with the discussion in Section 3.5.2, the discrepancy in the estimated AR 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 AR coefficients obtained at

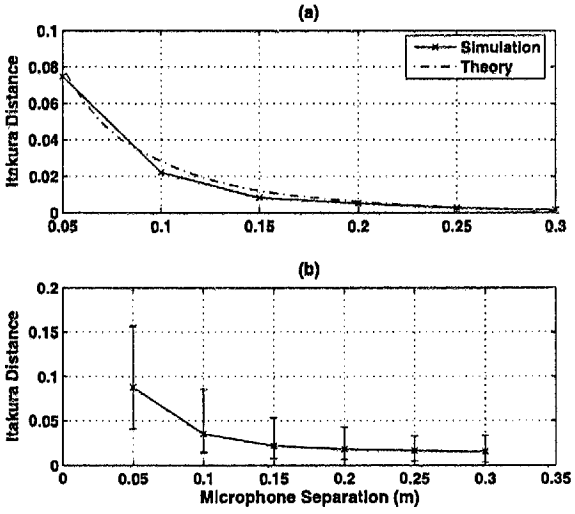

Figure 3.6: Itakura distance vs. microphone separation for (a) the theoretical results calculated with (3.37) (dashed) and the simulated results (crosses) for the spatially expected AR coefficients at the output of the DSB and (b) the AR coefficients for each individual outcome. Error bars indicate the maximum and the minimum errors while crosses show the mean value.

output of a DSB. All the parameters of the room, the source and the microphone array were kept fixed while the separation, $\|\ell_m - \ell_{m+1}\|$, between adjacent microphones in the linear array was increased from 0.05 m to 0.3 m in steps of 0.05 m. The results are shown in Figure 3.6 where the Itakura distance is plotted against microphone separation for (a) the theoretical results calculated with (3.37) (dashed) and the simulated results (crosses) for the spatially expected AR coefficients at the output of the DSB and (b) the AR 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 and at a microphone separation of $\|\ell_m - \ell_{m+1}\| = 0.3$ m the results are comparable to the M-channel case both in terms of spatial expectation and of the individual outcomes. This is due to the fact that the spatial correlation between microphones becomes negligible.

3.7 Summary

In this chapter, statistical room acoustic theory has been used for the analysis of the AR modeling of reverberant speech. Investigating three scenarios it has been shown that, in terms of spatial expectation, the AR 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 AR 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 arrangement. It was also demonstrated that AR coefficients calculated jointly 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 AR coefficients is the preferred option 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 most of these methods utilize microphone arrays for the residual processing, M-channel joint calculation of the AR coefficients should be deployed to ensure the validity of this assumption.

Chapter 4

Spatiotemporal Averaging Method for Enhancement of Reverberant Speech

ENHANCEMENT of reverberant speech by processing the prediction residual has emerged as an important dereverberation technique. Linear prediction of speech is related to the source-filter speech production model where the prediction residual approximately represents the glottal excitation and the AR coefficients constitute the vocal tract shaping filter [98]. The prediction residual comprises a pulse train for voiced speech and random noise for unvoiced speech [93, 98]. Therefore, it can be processed using this *a priori* information and an enhanced speech signal can be obtained without explicit knowledge of the room impulse response. In this chapter, a method for processing a reverberant prediction residual is presented, which uses a combination of spatial averaging and a novel approach of larynx cycle temporal averaging. The use of the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [99] to estimate glottal closure instants in reverberant speech is demonstrated and applied to larynx cycle segmentation of the prediction residual from spatially averaged speech. The enhanced larynx cycles are used to update a slowly time-varying filter which accommodates processing of both voiced and unvoiced speech. The approach is applied to speech dereverberation in the Spatiotemporal Averaging Method for Enhancement of Reverberant Speech (SMERSH). Finally, the efficiency of the new method is demonstrated with simulation results in terms of Bark spectral distortion score and segmental signal-to-reverberant ratio.

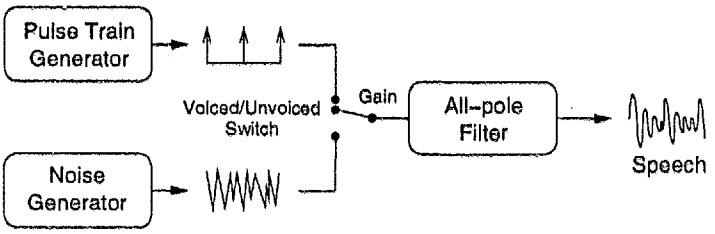

Figure 4.1: The speech production model based on linear prediction.

4.1 Prediction Residual Processing for Reverberant Speech Enhancement

Consider, the speech production model [94, 98] depicted in Fig. 4.1. This is a simplified model of the human speech production system, which broadly divides the speech generation process into two components: (i) an excitation sequence comprising a combination of quasi-periodic impulses for voiced speech and random noise for unvoiced speech and (ii) an all-pole filter which models the vocal tract, glottal pulse and lip radiation [94, 98]. In physical terms the excitation sequence describes approximately the properties of the vocal cords in voiced speech, while the all-pole filter models the frequency shaping of the throat, the mouth and the nose. The time between the excitation pulses defines the fundamental frequency [93, 98].

A speech signal, $s(n)$, can be decomposed into these two components using a linear predictor [93], which was defined in (3.7) and is reproduced here for convenience

$$s(n) = -\boldsymbol{a}^T \boldsymbol{s}(n-1) + e(n), \tag{4.1}$$

where the AR coefficients, $\boldsymbol{a}$, characterize the all-pole filter and the prediction residual, $e(n)$, represents the excitation sequence. Thus, the motivation for the prediction residual processing methods is twofold. First, it is assumed that reverberation mainly affects the excitation sequence and not the all-pole filter. The validity of this assumption was studied in Chapter 3 and it was concluded that if multichannel LPC is used, the clean speech AR coefficients can be estimated from reverberant observations with satisfactory accuracy. Consequently, the reverberant prediction residual in voiced speech contains the original impulses due to the excitation followed by several erroneous peaks due to multi-path

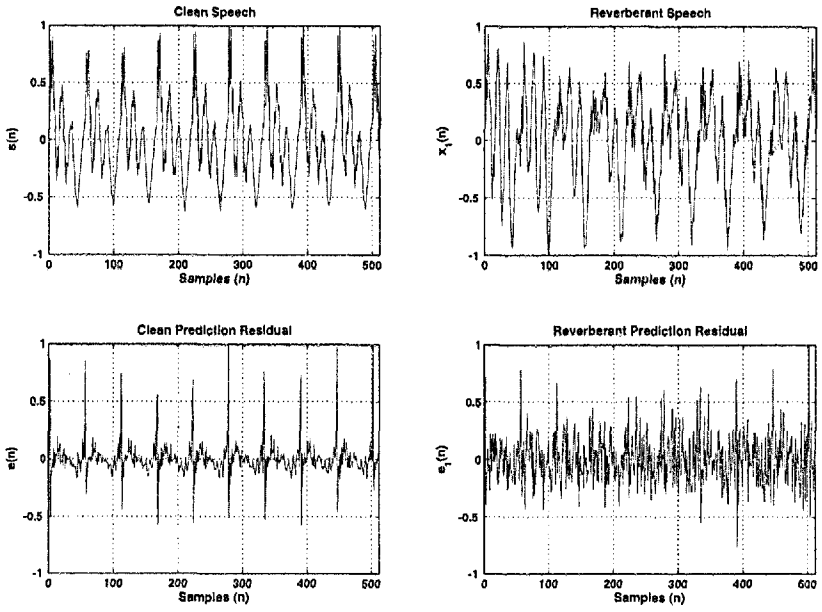

Figure 4.2: Clean and reverberant speech and the corresponding prediction residuals.

reflections from the surrounding walls. An example with portions of clean voiced speech, reverberant voiced speech and the corresponding prediction residuals is shown in Fig. 4.2. 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 resulting room impulse response is shown in Fig. 4.3. It can be seen in the bottom right plot of Fig. 4.2 that the reverberant residual contains several peaks of similar strength as the true excitation peaks. This leads to the definition of an erroneous peak adopted in the remainder of this chapter: *an erroneous peak* is a pulse in the prediction residual of reverberant speech, which 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.

Consequently, the *a priori* information about the structure of the prediction residual for voiced speech due to the source-filter model can be utilized to enhance the reverberant speech. This is achieved by attenuating the erroneous peaks in the prediction

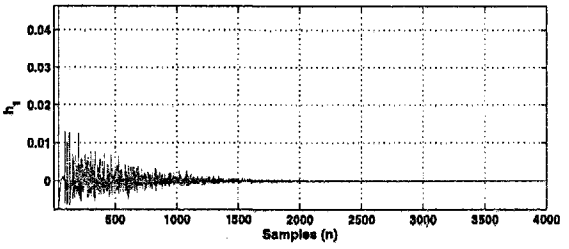

Figure 4.3: Simulated room impulse response used in the example of Fig. 4.2 for a room with dimensions $6.4 \times 5 \times 4$ m, $T_{60} = 0.5$ s, and source-microphone separation $D_1 = 1.5$ m.

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. Several methods for processing a reverberant prediction residual have been proposed by various authors [15,33,34,53,56–58]. The general procedure of the use of linear prediction for reverberant speech enhancement is shown in Fig. 4.4. Linear prediction analysis is performed on the reverberant observations, $x_m(n)$, $m = 1, 2, \ldots M$ in order to obtain the prediction residuals $e(n)$ and a set of AR coefficients, $\hat{\boldsymbol{a}}$, which are an estimate of the clean speech coefficients, $\boldsymbol{a}$. The prediction residuals are then processed to find an estimate of the clean speech residual, $\hat{e}(n)$. Finally, a clean speech estimate, $\hat{s}(n)$, is found by synthesis using $\hat{e}(n)$ and $\hat{\boldsymbol{a}}$ such that

$$\hat{s}(n) = -\hat{\boldsymbol{a}}^T \hat{\boldsymbol{s}}(n-1) + \hat{e}(n), \tag{4.2}$$

where $\hat{\boldsymbol{s}}(n-1) = [\hat{s}(n-1), \hat{s}(n-2), \ldots, \hat{s}(n-p)]^T$ is a vector of samples from the enhanced speech signal.

The most attractive feature of the prediction residual processing methods is that they can reduce the effects of reverberation without specific knowledge of the room transfer function, which is generally not available and difficult and computationally expensive to estimate and invert as will be discussed in the following chapters. Therefore, it makes these algorithms practical and suitable for online implementations.

In Chapter 3, a study of the AR coefficients from reverberant speech was presented. Here, the focus is on the processing of the prediction residual. In the remainder of this section a review of the existing methods for reverberant prediction residual enhancement

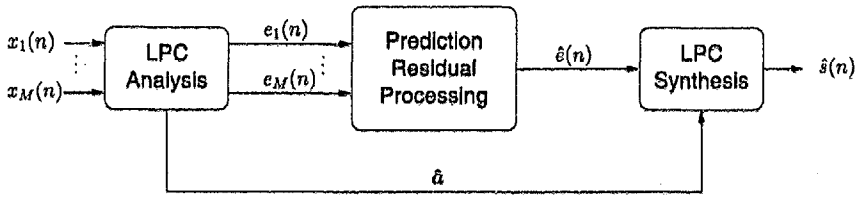

Figure 4.4: Reverberant speech enhancement via linear prediction.

is presented and some identified drawbacks with these methods are highlighted. In Section 4.2 a new method for prediction residual enhancement is developed which addresses some of the shortcomings of the existing algorithms. The new method is based on a spatial averaging and a new approach of temporal inter-cycle averaging. This method is then incorporated into a dereverberation algorithm, which is presented in Section 4.3 and is evaluated with comparative experiments in Section 4.4. Finally, a summary of the work presented in this chapter is provided in Section 4.5

4.1.1 Overview of Existing Methods for Prediction Residual Processing

Various methods for processing the prediction residual resulting from the linear prediction analysis of reverberant speech have been proposed in the literature and 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 room impulse response are not.

Yegnanarayana et al [53] provided a comprehensive study of the reverberant prediction residual. They demonstrate 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 only allows for small amounts of dereverberation, however, it is able to operate on a single channel.

Yegnanarayana and Satyanarayana [57] proposed to use *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 which emphasizes the positions of the true excitation peaks. This weighting function is applied to the prediction residual of one of the channels.

An approach proposed by Brandstein and Griebel [33, 34], 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 multi-path 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.

A different approach by Griebel and Brandstein [56] uses a *weight function based on coarse estimates of the room impulse responses.* The authors show that coarse estimates of the acoustic channel impulse responses can be obtained by averaging the phase transform version of the generalized cross-correlation [100] from the multiple microphones. These estimates are then applied in a matched filtering operation to obtain a weighting function for each of the M microphone's prediction residuals. 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 course channel estimates. The results in [56], for example, were generated using $M = 15$ microphones.

Finally, an adaptive algorithm was proposed by Gillespie et al [58] using a *kurtosis maximizing sub-band adaptive filter.* The authors demonstrate that the kurtosis of the prediction residual decreases as a function of increased reverberation, which was also suggested in [53]. They use this observation to derive an adaptive filter which maximizes the kurtosis of the prediction residual. The filter is applied directly to the observed signal rather than on the prediction residual and thus, avoids the LP synthesis stage. This avoids the dependence on the AR coefficients to some extent. The adaptive filter is implemented in a multi-channel subband framework for increased efficiency. This approach is an example of a hybrid method between speech enhancement and blind system identification and inversion. An extension to this method was presented in [59], where it was combined with spectral subtraction to remove residual reverberation due to late reflections.

All the described methods do achieve their goal in reducing the peaks in the prediction residual introduced by the RIR. However, they suffer from one common drawback. They tend to ignore the form of the prediction residual both for the peaks due to the original excitation and also the information between the glottal closure instants. Modifying the excitation peaks or excessively flattening the waveform between these may result in distortions in the reconstructed speech signal. In particular, the enhanced speech sounds less natural [101]. Furthermore, most of the methods do not consider the unvoiced/silent speech segments in the dereverberation process. In the following sections, a new method is proposed which overcomes these shortcomings.

4.2 Spatiotemporal Averaging for Prediction Residual Enhancement

In this section, a new technique is developed for multichannel enhancement of the prediction residual obtained from linear prediction of reverberant speech. A preliminary version of this approach was presented in [102]. In the first stage of the method, spatial averaging is applied to the time-aligned microphone observations. Next, the prediction residual processing is performed on the resulting signal. This spatially averaged signal is already enhanced to some extent and, therefore, the peaks due to the original excitation are emphasized. The motivation behind this is best explained with the aid of an example. Consider the characteristic illustration in Fig. 4.5 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 following observations can be made from the excitation sequences in this and in other examples:

(i) The prediction residual obtained from the output of the DSB (Fig. 4.5c) differs from that in clean speech by seemingly random peaks that are left unattenuated after the spatial averaging. These peaks appear uncorrelated among consecutive larynx-cycles.

(ii) In the case of clean speech (Fig. 4.5a), the main features of the prediction residual between consecutive larynx-cycles change slowly and show high inter-cycle correlation.

(iii) The strong periodic peaks in the DSB prediction residual appear to represent the excitation due to glottal closure seen in the clean speech.

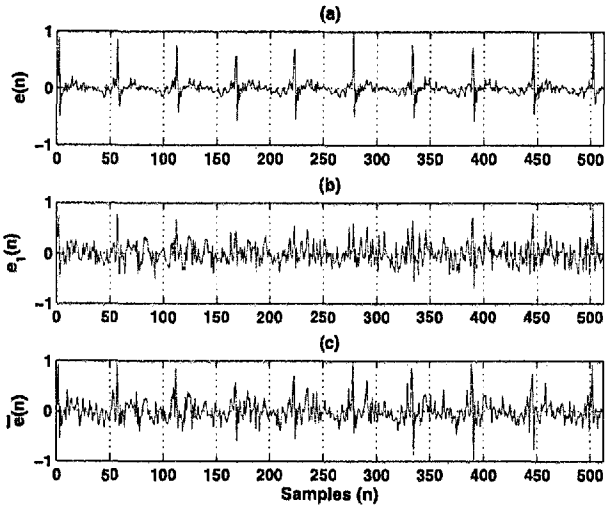

Figure 4.5: Prediction residuals obtained from (a) clean speech, $s(n)$, (b) reverberant speech $x_1(n)$ and (c) speech at the output of a DSB, $\bar{x}(n)$.

The second of these properties is well-known in speech processing and has been applied, for example, in the context of concatenation based text-to-speech synthesis such as the time domain pitch-synchronous overlap-add (TD-PSOLA) approach [103] and in the wavelet extrema clustering algorithm for prediction residual enhancement [15, 33, 34]. Motivated by these observations, it is proposed 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 issues to consider in such an averaging procedure. First, it is necessary to correctly identify the peaks that belong to the original excitation so as to segment the larynx cycles. Second, the main peak, attributed to glottal closure is very important to speech quality [101] and should therefore remain unchanged. Thus, it should be excluded from the averaging process. The implementation of these features of the algorithm are discussed in the following.

4.2.1 Estimation of Glottal Closure Instants in Reverberant Speech for Larynx Cycle Segmentation

If a Laryngograph (EGG) signal is available, the HQTx algorithm [104] can be used for identification of glottal closure instants (GCIs), which would give accurate information about the position of the true peaks in the original excitation sequence. However, an

EGG signal is generally not available in practice and an alternative approach is desirable, which is able to automatically identify the GCIs from the speech signal alone. One such algorithm is DYPSA which has been demonstrated to identify accurately the instances of glottal closure in voiced anechoic speech [99,105]. In this section, the DYPSA algorithm is summarized. Moreover, a study is presented of the performance of DYPSA on reverberant speech. This compares three cases: (i) clean speech, (ii) reverberant speech and (iii) speech at the output of a DSB. Experimental results are provided to demonstrate that the GCI's can be identified with a satisfactory accuracy at the output of a DSB.

The DYPSA algorithm

DYPSA consists of three main components:

(i) *The phase-slope function* - 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 positive zero-crossings of the phase-slope function.

(ii) *The phase-slope projection* - is introduced to generate GCI candidates when a local minimum is followed by a local maximum without crossing a zero. The midpoint between these is identified and projected onto the time axis with unit slope. In this way, GCIs whose positive going slope does not cross the zero point and are, consequently, missed by the phase-slope function are identified.

(iii) *Dynamic Programming* - uses known characteristics of voiced speech and forms a cost function to select a subset of the GCI candidates which are most likely to correspond to the true ones. The subset of candidates is selected according to the minimization problem defined as

$$\min_{\Omega} \sum_{i=1}^{|\Omega|} \boldsymbol{\lambda}^T \boldsymbol{\gamma}(i), \tag{4.3}$$

where Ω is a subset with GCI's of size $|\Omega|$ selected from all GCI candidates, $\boldsymbol{\lambda} = [\lambda_A\ \lambda_P\ \lambda_J\ \lambda_F\ \lambda_S]^T = [0.8\ 0.5\ 0.4\ 0.3\ 0.1]^T$ is a vector of weighting factors with the values taken here as in [99] and $\boldsymbol{\gamma}(i) = [\gamma_A(i)\ \gamma_P(i)\ \gamma_J(i)\ \gamma_F(i)\ \gamma_S(i)]^T$ is a vector of cost elements evaluated at the ith GCI of the subset. The cost vector elements are:

- *Speech waveform similarity*, $\gamma_A(i)$, between neighbouring candidates, where candidates not correlated with the previous candidate are penalized.

- *Pitch deviation*, $\gamma_P(i)$, between the current and the previous two candidates, where candidates with large deviation are penalized.
- *Projected candidate cost*, $\gamma_J(i)$, for the candidates from the phase-slope projection, which often arise from erroneous peaks.
- *Normalized energy*, $\gamma_F(i)$, which penalizes candidates that do not correspond to high energy in the speech signal.
- *Ideal phase-slope function deviation*, $\gamma_S(i)$, where candidates arising from zero-crossings with gradients close to unity are favoured.

Gathering the characteristics of the prediction residuals resulting from clean and reverberant speech and spatially averaged speech at the output of a DSB together with the properties of DYPSA, the following remarks can be made: First, the reverberant prediction residual contains many peaks due to the room impulse response, which are of strength comparable to the desired peaks in the clean speech residual. Consequently, the phase-slope function and the phase-slope projection are likely to produce many erroneous candidates. Secondly, peaks of similar strength as the true peaks from the clean prediction residual are likely to result in wrong candidates if they both occur in the same analysis frame for the short time Fourier transform. Thirdly, a voiced speech segment of weak energy which was preceded by a high energy component is likely to result in erroneous candidates due to the smearing effect of the room impulse response. Such segments occur, for example, at the end of voiced utterances. All of these effects are reduced considerably by the delay-and-sum beamformer. It can be seen from the dynamic programming criteria that DYPSA is robust to spurious peaks in the prediction residual. Evidently, this is an attractive feature for GCI identification in reverberant speech and can be expected to discriminate many of the erroneous candidates due to reverberation.

Experimental Results: DYPSA and reverberant speech

The results from this study are presented here for the three different cases: (i) clean speech (ii) reverberant speech for one microphone and (iii) speech at a DSB output. The APLAWD database [92] was used for evaluation. This database contains anechoic recordings comprising ten repetitions of five sentences uttered by five male and five female talkers. Each recording is accompanied by a corresponding Laryngograph signal, accommodating GCI identification with the HQTx algorithm [99, 104]. The effects of room reverberation

were simulated by convolution of FIR room impulse responses obtained from the source-image method [12] and the anechoic speech samples. The simulated room had dimensions of $5 \times 4 \times 3$ m and an eight element microphone array was used with 0.05 m separation between successive elements. The talker was positioned at a distance of 2.5 meters away from the centre of the microphone array. The reverberation time T_{60} was varied between 0.1 s and 0.5 s in steps of 0.05 s.

Furthermore, following the approach in [99], the GCI's obtained with HQTx were employed as the reference. Two metrics were applied for performance evaluation of the DYPSA algorithm on voiced speech. First, detection rate is used, which is the percentage of GCIs obtained with HQTx for which exactly one GCI is detected with DYPSA. The second measure was the identification accuracy, which measures the standard deviation of the distance between the reference GCIs and the GCIs identified with DYPSA.

Figure 4.6 shows a plot of the detection rate versus reverberation time, for clean speech, reverberant speech and speech pre-processed with the DSB, indicated with black, grey and white bars respectively. The clean speech results are the same as those in [99] showing a detection rate of 95.7% and identification accuracy of 0.71 ms. The detrimental effect of reverberation is apparent and a performance degradation of up to 40% is observed at reverberation time $T_{60} = 0.5$ s. On the contrary, when the DSB is used as a pre-processor, the detection rate performance is improved significantly with an average of 15% improvement over the single channel case. The corresponding accuracy of the identified GCI's is plotted in Fig. 4.7 for (a) reverberant speech, (b) DSB output and (c) clean speech, which also demonstrates a considerable performance improvement due to the DSB. Finally, it is interesting to note that DYPSA with the DSB as a pre-processor, provides $71.8\% - 93.9\%$ detection rate with accuracy ranging from $0.71 - 0.85$ ms. These results, although worse than for the clean speech case, are superior or at least comparable to other existing algorithms operating on clean speech as seen from the results presented in [99].

In summary, if a Laryngograph (EGG) signal is available, an accurate larynx cycle segmentation can be performed using the HQTx. For practical applications where the EGG signal is not available, DYPSA can be applied with satisfactory accuracy and with up to 93.9% detection rate, when applied on spatially averaged multichannel speech observations.

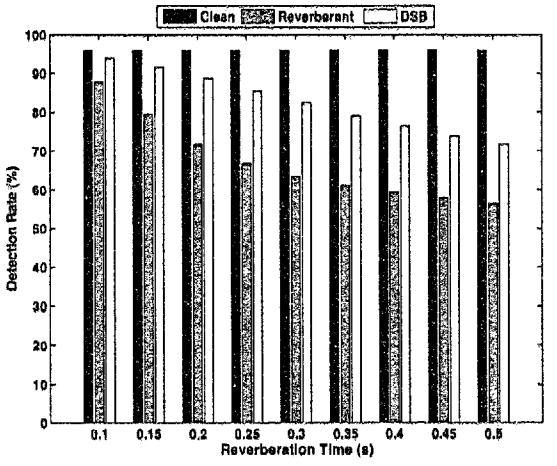

Figure 4.6: Detection rate vs. reverberation time.

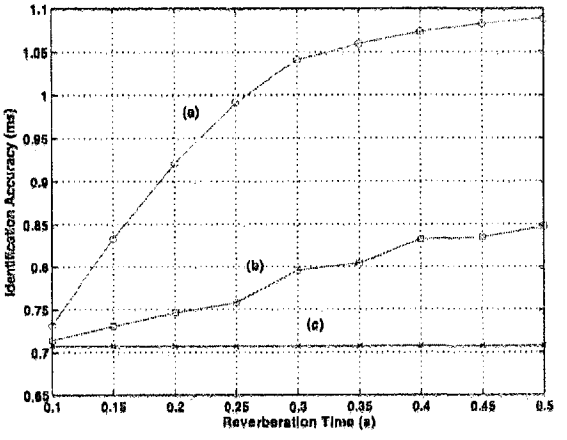

Figure 4.7: Identification accuracy vs. reverberation time for (a) reverberant speech, (b) speech at the output of a DSB and (c) clean speech.

4.2.2 Weighted Inter-cycle Averaging

Having identified the instants of glottal closure and consequently segmented the prediction residual into larynx cycles, the subsequent step is to perform the temporal inter-cycle averaging. As discussed in the introduction of Section 4.2, it is desirable to perform the averaging between the successive larynx-cycles only and to leave the glottal pulse

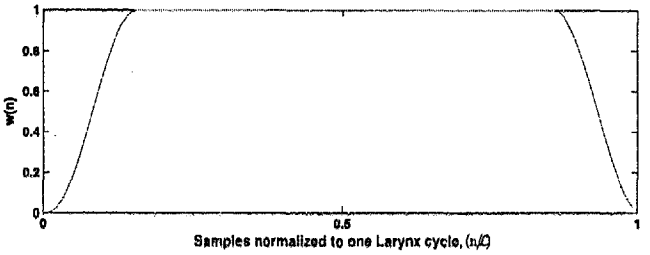

Figure 4.8: Larynx cycle weight function.

undisturbed. One way to satisfy this condition is by applying a weight function on each larynx frame prior to the averaging. This weight function should, ideally, exclude only the true glottal pulse. However, in practical situations, the position of the glottal pulse is identified to an uncertainty in the order of 1 ms as demonstrated in Section 4.2.1. Furthermore, the glottal pulse is not a true impulse but is spread in time [94]. In selecting the weight function, these variations must be taken into consideration and the weights have to be chosen such that as much as possible of the larynx-cycle is included in the averaging process.

One weight function which was found suitable in terms of meeting the set requirements and with a reasonable trade-off between the issues described above, is the time-domain Tukey window defined as [106]

$$w_u = \begin{cases} \frac{1+\cos\left(\frac{2\pi u}{v(\mathcal{L}-1)}-\pi\right)}{2}, & u < \frac{v\mathcal{L}}{2} \\ \frac{1+\cos\left(\frac{2\pi}{v}-\frac{2\pi u}{v(\mathcal{L}-1)}-\pi\right)}{2}, & u > \mathcal{L}-\frac{v\mathcal{L}}{2}-1 \\ 1.0, & \text{otherwise,} \end{cases} \tag{4.4}$$

where $\mathcal{L}$ is the number of samples in one larynx-cycle and v is the taper ratio of the window. An example of this weighting function with $v = 0.3$ is depicted in Fig. 4.8. The taper ratio constant 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 glottal closure instants identification algorithm. Following the averaging procedure, the applied weights need to be reversed so as to restore the original glottal impulse. This can be done by applying an inverse weighting, $1 - w_u$, to the larynx frame under consideration.

Thus, in the proposed method, each enhanced larynx cycle in a voiced speech segment is obtained by averaging the current weighted larynx cycle frame under consideration with $\mathcal{I}$ of its neighbouring weighted larynx cycles. Since the signal is changing slowly from one larynx cycle to the next, the larynx cycle frames are attenuated by a constant $1/(|i|+1)$, according to their offset from the larynx cycle under consideration, $i = 0$. The result is then added to the original larynx cycle weighted with the inverse weight function. The final expression for the enhanced larynx cycle becomes

$$\hat{\bar{e}}_\ell = (\boldsymbol{I} - \boldsymbol{W})\bar{e}_\ell + \frac{\sum_{i=-\mathcal{I}}^{\mathcal{I}} \frac{1}{(|i|+1)} \boldsymbol{W} \bar{e}_{\ell+i}}{\sum_{i=-\mathcal{I}}^{\mathcal{I}} \frac{1}{(|i|+1)}}, \tag{4.5}$$

with

$$\boldsymbol{W} = \begin{bmatrix} w_0 & 0 & \dots & 0 \\ 0 & w_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{\mathcal{L}-1} \end{bmatrix},$$

where $\bar{e}_\ell = [\bar{e}_\ell(n_\ell)\ \bar{e}_\ell(n_\ell+1)\ \dots\ \bar{e}_\ell(n_\ell+\mathcal{L}-1)]^T$ is the ℓth larynx-cycle at the output of the DSB with its GCI peak at time n_ℓ, $\hat{\bar{e}}_\ell = [\hat{\bar{e}}_\ell(n_\ell)\ \hat{\bar{e}}_\ell(n_\ell+1)\ \dots\ \hat{\bar{e}}_\ell(n_\ell+\mathcal{L}-1)]^T$ is the ℓth larynx cycle of the enhanced residual and the $\mathcal{L} \times \mathcal{L}$ matrix $\boldsymbol{W}$ contains weighting coefficients calculated from (4.4) along its main diagonal. Since the larynx-cycles are not strictly periodic but may vary within a few samples, $\mathcal{L}$ is set to equal the length of the larynx cycle being processed. Those larynx cycles of the $2\mathcal{I}+1$ 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 an important control factor in the inter-cycle averaging scheme. If too many cycles are included, the averaging will remove uncorrelated portions from the original excitation. If too few cycles are considered, then erroneous peaks due to reverberation will remain. In the experiments for the results presented here the number of cycles for averaging was set to $\mathcal{I} = 4$. Generally, it was found through several experiments that $\mathcal{I} > 4$ provides less accurate results. The duration of a larynx cycle is usually in the range of $2 - 20$ ms [94] and therefore including 9 cycles in the averaging (i.e. $\mathcal{I} = 4$) requires up to 180 ms of voiced speech which is reasonable in practice.

4.2.3 Equivalent Inverse System

The spatiotemporal averaging algorithm proposed in Sections 4.2.1 and 4.2.2 attenuates the reverberant components in the prediction residual. However, two unresolved and important issues remain. First is the fact that only voiced speech segments are processed, leaving reverberant effects clearly audible in the unvoiced speech and silence. Second, the algorithm does not take advantage of past correct larynx-cycle frames in case of erroneous larynx-cycle segmentation, which may occur due to inaccuracies in the identified GCIs.

These issues are addressed here by introducing a slowly varying filter which performs the equivalent inverse operation (in the least squares sense) of the inter-cycle averaging operation performed in (4.5), i.e.

$$\hat{e}_\ell(n_\ell) = \boldsymbol{g}_\ell^T \bar{\boldsymbol{e}}_\ell, \tag{4.6}$$

where $\boldsymbol{g}_\ell = [g_{\ell,0}\ g_{\ell,1}\ \ldots\ g_{\ell,L_\mathrm{i}-1}]^T$ is an L_i-tap FIR filter, $\bar{\boldsymbol{e}}_\ell = [\bar{e}_\ell(n_\ell)\ \bar{e}_\ell(n_\ell-1)\ \ldots\ \bar{e}_\ell(n_\ell-L_\mathrm{i}+1)]^T$ is the input vector of the ℓth larynx-cycle of the prediction residual obtained from the DSB and $\hat{e}_\ell(n_\ell)$ is the ℓth processed larynx-cycle.

From (4.6) an error can be formulated, $E_{LC} = \boldsymbol{g}_\ell^T \bar{\boldsymbol{e}}_\ell - \hat{e}_\ell(n_\ell)$, and an estimate, $\hat{\boldsymbol{g}}_\ell$, of the inverse system for the ℓth larynx-cycle, $\boldsymbol{g}_\ell$, can be obtained by minimizing the mean squared error $\|E_{LC}\|^2$ such that

$$\hat{\boldsymbol{g}}_\ell = \min_{\boldsymbol{g}_\ell} \|\boldsymbol{g}_\ell^T \bar{\boldsymbol{e}}_\ell - \hat{e}_\ell(n_\ell)\|^2. \tag{4.7}$$

This can be solved in the least squares sense as

$$\hat{\boldsymbol{g}}_\ell = \boldsymbol{\Phi}^{-1}\boldsymbol{p}, \tag{4.8}$$

with $\boldsymbol{\Phi}$ being the $L_\mathrm{i} \times L_\mathrm{i}$ correlation matrix of the DSB larynx-cycle input vector defined as

$$\boldsymbol{\Phi} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{E}_\ell(e^{j\omega})|^2 \boldsymbol{d}\boldsymbol{d}^H\, d\omega \tag{4.9}$$

and $\boldsymbol{p}$ being the $L_\mathrm{i} \times 1$ cross-correlation vector between the DSB larynx-cycle input vector and the enhanced larynx-cycle

$$\boldsymbol{p} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \bar{E}_\ell(e^{j\omega})\hat{E}_\ell^*(e^{j\omega})\boldsymbol{d}\, d\omega, \tag{4.10}$$

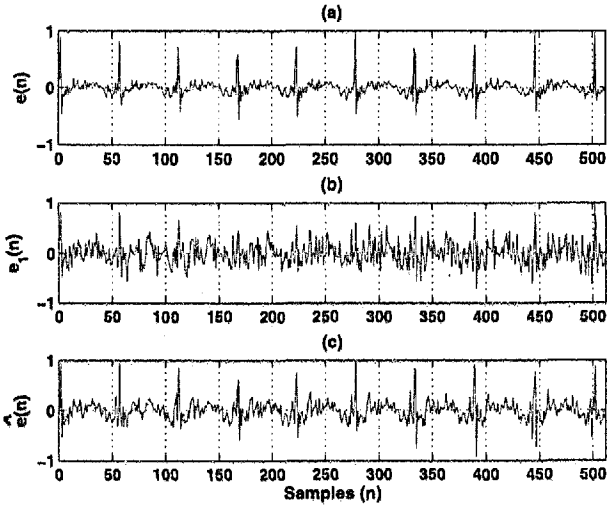

Figure 4.9: Prediction residuals from (a) clean speech, (b) reverberant speech and (c) prediction residual processed with the proposed method.

where $\bar{E}_\ell(e^{j\omega})$ is the Fourier transform of $\bar{e}_\ell(n_\ell)$, $\hat{E}_\ell(e^{j\omega})$ is the Fourier transform of $\hat{e}_\ell(n_\ell)$ and $\boldsymbol{d} = [e^{-j\omega}\ e^{-j2\omega} \ldots, e^{-jL_i\omega}]^T$ is a $L_i \times 1$ DFT vector.

Subsequently, $\hat{\boldsymbol{g}}_\ell$ is used to update a slowly varying filter, $\hat{\boldsymbol{g}}$, according to the equation

$$\hat{\boldsymbol{g}}(n_\ell) = \beta\hat{\boldsymbol{g}}(n_{\ell-1}) + (1-\beta)\hat{\boldsymbol{g}}_\ell, \tag{4.11}$$

where $0 \leqslant \beta \leqslant 1$ is a forgetting factor with typical values in the range $\{0.1 - 0.3\}$. The filter is initialized to $\hat{\boldsymbol{g}}(0) = [1\ 0\ 0 \ldots 0]^T$ the update of $\hat{\boldsymbol{g}}$ is performed only during voiced speech segments. In the unvoiced speech or silences, the filter is kept and run at its last update. This is motivated by the fact that $\hat{\boldsymbol{g}}$ is an inverse filter related to the room transfer function and that it will equalize the reverberation effects also in these periods. This is confirmed by the simulation results and by several informal listening tests.

Figure 4.9 shows an example of prediction residuals obtained from (a) clean speech, (b) reverberant speech and (c) a residual processed with the proposed algorithm. It can be seen that the enhanced residual in Fig. 4.9c closely resembles that of clean speech in Fig. 4.9a and that the waveform between the GCI's is restored. In the example of Fig. 4.9, the mean squared error between the clean and the reverberant residuals is -14.5 dB and the mean squared error of the residual processed with the proposed method -18.1 dB.

4.3 Application to Reverberant Speech Enhancement

Leveraging the results from Chapter 3 and the prediction residual enhancement method presented in Section 4.2, the use of AR modelling of reverberant speech in the context of multichannel speech dereverberation is now demonstrated. Based on the results from the analysis in Chapter 3 the coefficients calculated jointly from M microphones provide a close approximation to the clean speech coefficients. Consequently, using these M-channel coefficients in (3.8) to filter the output of a DSB then results in a prediction residual which is closer to the clean speech residual with the original excitation peaks clearly visible and allows processing according to the method described in Section 4.2, in order to obtain an estimate of the clean speech excitation. Finally, with the processed prediction residual and the M-channel AR coefficients, both being approximately equal to those obtained from clean speech, a speech signal with reduced reverberation can be constructed. The spatiotemporal averaging method for enhancement of reverberant speech is summarized in Table 4.1. In addition, a system diagram of the algorithm is given in Fig. 4.10. The numbers in the figure indicate various stages of the processed signal; an illustrative example, showing the speech signal at each of these stages, will be given in Section 4.4.

4.4 Simulations and Results

Simulation results are provided here to demonstrate the performance of the SMERSH algorithm. The experiments aim to demonstrate the following: 1. the outcome at the various stages of the SMERSH algorithm, 2. the performance of the algorithm using GCI's obtained from HQTx, 3. the performance of the algorithm with the GCI's obtained from DYPSA and hence to show that no loss in performance is caused by applying DYPSA instead of HQTx and 4. the improvement obtained with the proposed dereverberation algorithm over the delay-and-sum beamformer.

Test data from the APLAWD database comprising the sentence "George made the girl measure a good blue vase" uttered by five male and five female talkers was used as an example. The instants of glottal closure were found using either the HQTx algorithm [104] or DYPSA [99]. For the LPC analysis and synthesis, a prediction order of $p = 13$ and 30 ms Hamming window with 50% overlap was employed. The sampling frequency was $f_s = 8$ kHz. A room with dimensions $5 \times 4 \times 3$ m was simulated using the source-image

Table 4.1: SMERSH - Spatiotemporal Averaging Method for Enhancement of Reverberant Speech

1. Calculate the AR coefficients using M-channel LPC as in (3.32):

$$\hat{b}_{\text{opt}} = -\hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{q}}.$$

2. Apply the delay-and-sum beamformer to the M microphone signals to obtain $\bar{x}(n)$ (3.36) (assuming that the time delays are known):

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

3. Apply the filter from (3.8) with coefficients $\hat{b}_{\text{opt}}$ to the DSB output to obtain the prediction residual, $\bar{e}(n)$:

$$\bar{e}(n) = \hat{b}_{\text{opt}}^{T}\bar{\boldsymbol{x}}(n).$$

4. Use e.g. DYPSA or HQTx to identify the positions of the times of the position of the excitation peaks, n_ℓ in the prediction residual $\bar{e}(n)$, and segment the into larynx-cycles $\bar{\boldsymbol{e}}_\ell$.

For each larynx-cycle, $\ell = 1, 2, \ldots$:

5. Calculate the temporally averaged larynx-cycle according to (4.5):

$$\hat{\boldsymbol{e}}_\ell = (\boldsymbol{I} - \boldsymbol{W})\bar{\boldsymbol{e}}_\ell + \frac{\sum_{i=-\mathcal{I}}^{\mathcal{I}} \frac{1}{(|i|+1)}\boldsymbol{W}\bar{\boldsymbol{e}}_{\ell+i}}{\sum_{i=-\mathcal{I}}^{\mathcal{I}} \frac{1}{(|i|+1)}}.$$

6. Calculate the least squares inverse filter as in (4.8):

$$\hat{\boldsymbol{g}}_\ell = \boldsymbol{\Phi}^{-1}\boldsymbol{p}.$$

and update the time-varying filter $\hat{\boldsymbol{g}}$ according to (4.11):

$$\hat{\boldsymbol{g}}(n_\ell) = \beta\hat{\boldsymbol{g}}(n_{\ell-1}) + (1-\beta)\hat{\boldsymbol{g}}_\ell.$$

7. Apply filter on the prediction residual from the DSB output until the beginning of the next larynx cycle, $\ell + 1$:

$$\hat{e}(n) = \hat{\boldsymbol{g}}_\ell^{T}\bar{\boldsymbol{e}}. \tag{4.12}$$

8. Obtain an estimate of the clean speech signal, $\hat{s}(n)$, by synthesis using the enhanced residual $\hat{e}(n)$ and the filter from (3.9) with AR coefficients, $\hat{b}_{\text{opt}}$:

$$\hat{s}(n) = [\hat{b}_{\text{opt}}^{-1}]^{T}\hat{\boldsymbol{e}}(n),$$

where $[\hat{b}_{\text{opt}}^{-1}]$ represents the all-pole filter coefficients calculated from the AR coefficients $\hat{b}_{\text{opt}}$

method [12]. An eight element linear microphone array with 0.05 m spacing between adjacent microphones was assumed and the reverberation time was varied in the range $T_{60} = \{0.2 - 0.8\}$ s in incremental steps of 0.05 s. The algorithm was evaluated at ten different source-receiver positions in the room. The results presented here are averaged over all speakers and all different source-microphone positions. The delay-and-sum beamformer was used as a baseline algorithm. The segmental SRR and the Bark spectral distortion measures defined in Section 2.4 were applied to quantify the improvement of the processed speech.

A portion of a speech signal is shown in Fig. 4.11 to give a typical illustrative example of the processed signal through various stages of the SMERSH algorithm for $T_{60} = 0.6$ s. The numbers on the plots correspond to the numbers in Fig. 4.10 showing: 1. the clean speech signal, 2. the reverberant speech signal, which apparently differs from the clean speech signal, 3. the prediction residual from the spatially averaged observations, 4. the identified GCIs, 5. the temporally averaged prediction residual and finally, 6. the enhanced speech signal. The enhanced signal closely resembles the clean speech signal. However, it is not an exact reconstruction, which will also be observed in the following two experiments in terms of the utilized quantitative signal similarity measures.

Experiment 1: with HQTx

In this experiment, the HQTx algorithm was used to find the GCIs from the Laryngograph (EGG) signals available in the APLAWD database. Voiced speech segments were selected using the density of GCIs where any subsequent estimates with an interval of less than 20 ms were considered to be part of a voiced speech segment.

The result in terms of segmental SRR vs. increased reverberation time is shown in Fig. 4.12 for (a) reverberant speech at the microphone closest to the talker, (b) speech at the output of the DSB and (c) speech processed with the proposed method, SMERSH. The improvement over the reverberant speech in terms of SRR is shown in Fig. 4.13 where white bars indicate improvement due to the DSB and grey bars indicate improvement due to SMERSH. The improvement is calculated as the score difference between the reverberant and the processed speech. The results in terms of Bark spectral distortion are shown in Fig. 4.14 for (a) reverberant speech, (b) speech at the output of the DSB and (c) speech processed with the proposed method, SMERSH and also the improvement obtained with both methods is shown in Fig. 4.15 for speech processed with the DSB indicated with

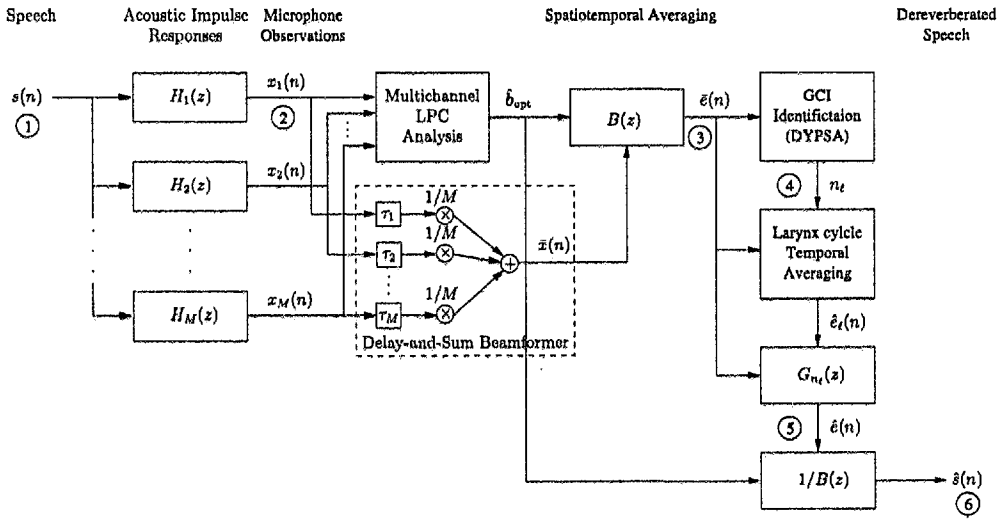

Figure 4.10: System diagram of SMERSH.

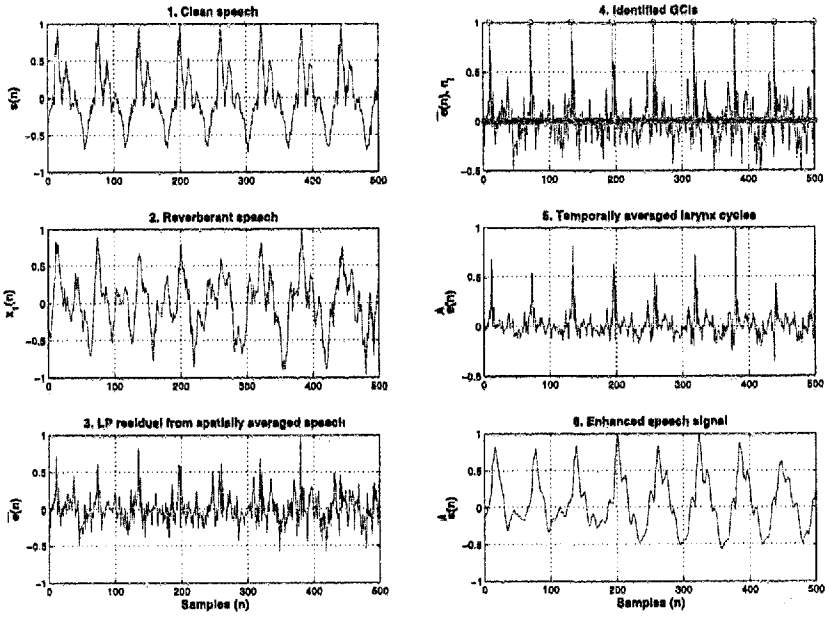

Figure 4.11: Illustration of the various stages of SMERSH at $T_{60} = 0.6$ s.

white bars and speech processed with SMERSH indicated with grey bars.

It can be seen from the plots, both in the temporal and spectral distortion measures, that the proposed method provides only small improvement over the DSB for short reverberation times (< 0.3 s). On the contrary, as the reverberation time increases the improvement due to SMERSH becomes more apparent and improvements up to 3.77 dB in segmental SRR are obtained at $T_{60} = 0.8$ s which corresponds to 47.8% improvement over the DSB. At that reverberation time a BSD score improvement of 1.85 is observed which is and represents a 26.5% improvement over the DSB.

Experiment 2: with DYPSA

The DYPSA algorithm was utilized to automatically extract the GCIs using the speech signals only. In practice, DYPSA requires a voiced speech detector in order to discard GCI estimates outside voiced regions. For this experiment, the GCIs from the HQTx algorithm were used as a voiced speech detector as in the previous experiment. Only those GCIs obtained with DYPSA that are within a voiced region were kept. However, within this voiced region, the GCIs were left as found by DYPSA, i.e. including also possible false or inaccurate estimates.

The equivalent set of results as in Experiment 1 are presented here for the outcome of Experiment 2. Segmental SRR vs. increased reverberation time for (a) reverberant speech, (b) DSB speech and (c) speech processed with SMERSH is shown in Fig 4.16. The respective improvement in terms of SRR are given in Fig. 4.17, where DSB processed speech is indicated with white bars and SMERSH processed speech is indicated with grey bars. The results in terms of Bark spectral distortion for the same three cases are plotted in Fig. 4.18 with the respective improvement in Fig. 4.19

It is seen from the resulting plots that the algorithm's performance when using DYPSA for automatic GCI identification is comparable to that with HQTx. Improvements of up to 4 dB in segmental SRR are obtained at $T_{60} = 0.8$ s which is 50.8% better than the DSB. At that reverberation time a BSD score improvement of 1.86 is observed which corresponds to 26.9% performance gain over the DSB. Finally, informal listening tests on the sentence used in the experiments and on other sentences have resulted in the following observations regarding perceptual quality of the processed speech: (i) the reverberant effects due to the room are reduced considerably, (ii) the speaker appears to be closer to the microphone, and (iii) no deleterious artefacts are introduced by the processing.

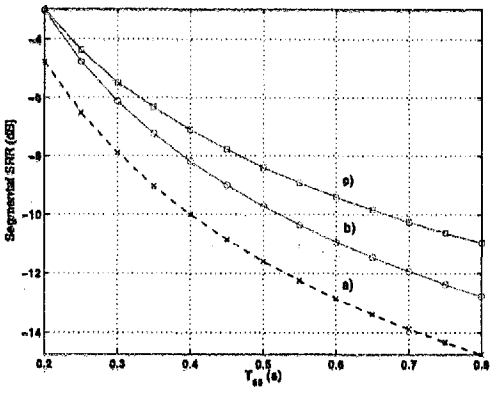

Figure 4.12: Segmental SRR for (a) reverberant speech at the microphone closest to the talker, (b) speech at the output of a DSB and (c) speech processed with SMERSH with HQTx for larynx cycle segmentation.

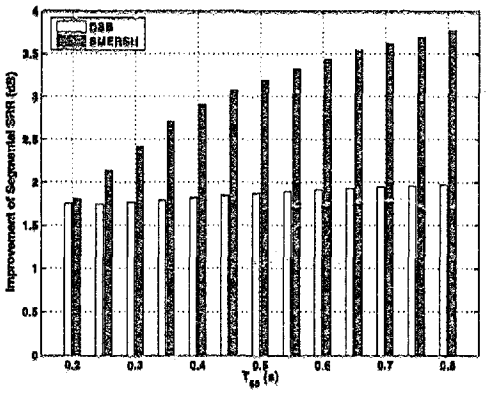

Figure 4.13: Improvement in Segmental SRR with DSB (white bars) and SMERSH with HQTx for larynx cycle segmentation (grey bars).

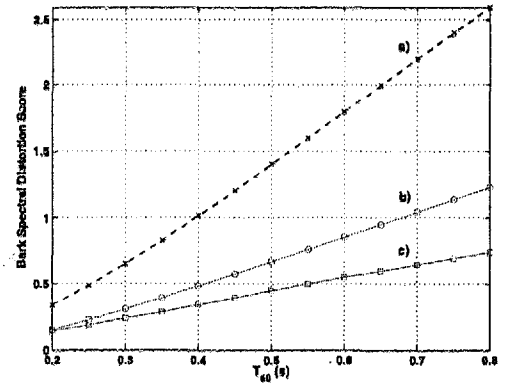

Figure 4.14: Bark spectral distortion score for (a) reverberant speech at the microphone closest to the talker, (b) speech at the output of a DSB and (c) speech processed with SMERSH with HQTx for larynx cycle segmentation.

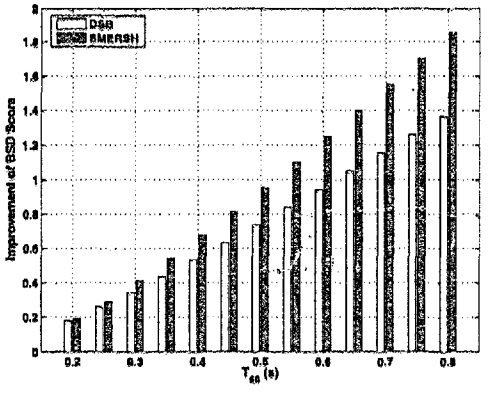

Figure 4.15: Improvement in Bark spectral distortion score with DSB (white bars) and SMERSH with HQTx for larynx cycle segmentation (grey bars).

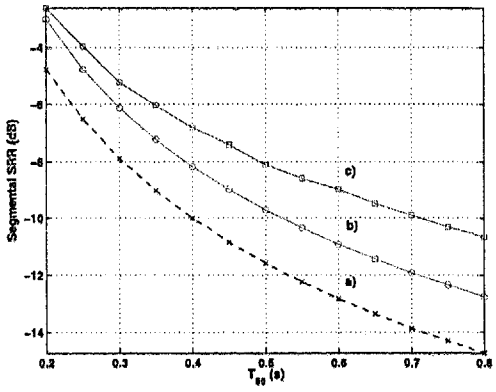

Figure 4.16: Segmental SRR for (a) reverberant speech at the microphone closest to the talker, (b) speech at the output of a DSB and (c) speech processed with SMERSH with DYPSA for larynx cycle segmentation.

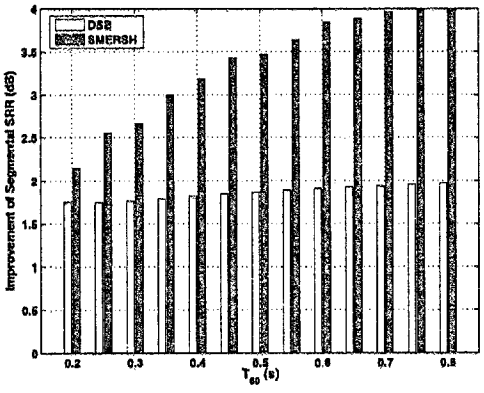

Figure 4.17: Improvement of Segmental SRR with DSB (white bars) and SMERSH with DYPSA for larynx cycle segmentation (grey bars).

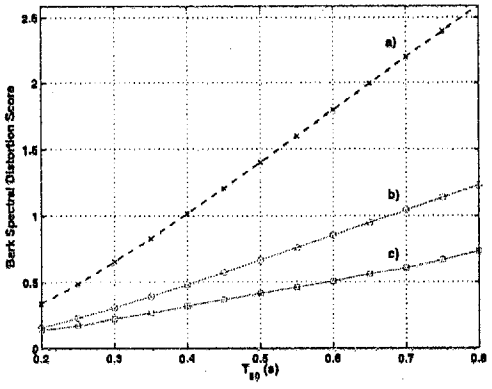

Figure 4.18: Bark spectral distortion score for (a) reverberant speech at the microphone closest to the talker, (b) speech at the output of a DSB and (c) speech processed with SMERSH with DYPSA for larynx cycle segmentation.

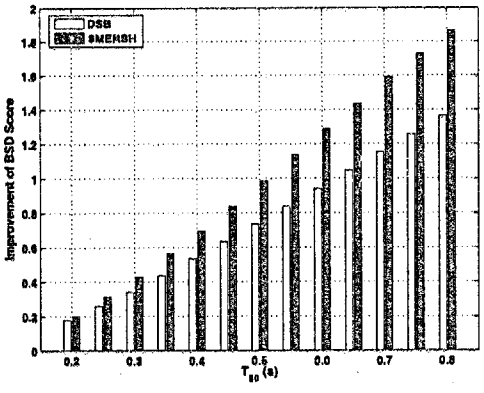

Figure 4.19: Improvement of Bark spectral distortion score with DSB (white bars) and SMERSH with DYPSA for larynx cycle segmentation (grey bars).

4.5 Summary

Processing of the linear prediction residual for reverberant speech enhancement has been discussed. The general concept of linear prediction in dereverberation was described and existing approaches for prediction residual enhancement were reviewed. Weaknesses of these existing methods were identified to be reduced naturalness in the processed speech and lack of processing in the unvoiced speech regions. Subsequently, these issues were addressed in the development of a novel multimicrophone method to prediction residual enhancement based on spatial averaging of the observed signals and a new approach of temporal averaging of neighbouring larynx cycles. The DYPSA algorithm was evaluated for automatic GCI identification in reverberant speech. It was demonstrated that, after spatial averaging of the speech signals, detection rates of up to 93.9% can be achieved, which thus makes larynx cycle segmentation feasible in practice. A slowly time-varying inverse filter was derived which is updated from every enhanced larynx cycle; this filter aids the processing in unvoiced speech regions and reduces sensitivity to inaccurate GCI identification. Finally, the performance of the algorithm was demonstrated through several experiments. The results demonstrate performance improvements over the DSB of up to 50.8% in terms of segmental SRR and up to 26.9% in terms of Bark spectral distortion.

Chapter 5

Blind Identification of Acoustic SIMO Systems

BLIND system identification (BSI) is an active area of research with applications in various fields of engineering. In the case of successful identification of the room impulse responses, dereverberation can be achieved with great accuracy. Adaptive algorithms are attractive for dereverberation applications, since the room impulse response varies with source-microphone position and consequently, with time.

In this chapter, a complex multichannel LMS (MCLMS) algorithm is derived from the cross-relation between channels, which operates both on complex and on real data. This method generally degrades in performance in the presence of noise, common or near-common zeros and for inaccurate choice of channel order. An optimal step-size for the MCLMS algorithm is derived and is shown to improve the identification performance in the presence of noise. Subsequently, the MCLMS is used to develop an adaptive method for the estimation of common roots in polynomials, which is applied to investigate the performance of adaptive BSI methods in the presence of common zeros. It is demonstrated that near-common zeros can significantly degrade the performance of such algorithms. A simulated example is provided to demonstrate that near-common zeros are likely to exist in random or acoustic systems as the order becomes large ($L > 200$). Consequently, a preliminary study towards adaptive blind system identification in subbands is presented as a strategy to reduce the adaptive filter lengths.

5.1 Background

BLIND system identification (BSI) has several applications in various fields of engineering [107], in particular where blind deconvolution or source separation is required. Examples include communications systems where the received signal must be equalized to obtain the transmitted signal [108], geophysics where the reflectivity of the earth layers is explored by extracting seismic signals from the sensor observations [109] and speech dereverberation where the acoustic impulse responses are estimated blindly from reverberant speech, and then deconvolution is performed to remove the effects of the room [65, 68, 72].

Several blind multichannel SIMO system identification algorithms have been reported in the literature and a review of many of these can be found in, for example [110]. Recently, a class of adaptive approaches based on the cross-relation between channels [62] has been proposed, with implementations in the time-domain [66] and in the frequency-domain [63]. The latter was also incorporated into a two-stage algorithm for blind source separation and dereverberation [68]. Adaptive algorithms are attractive for real-time applications due to their ability to track changes in the acoustic paths caused by source and/or receiver position fluctuations. The simplest of these methods is the multichannel LMS (MCLMS), which is chosen for the study presented in this chapter.

There are several problems inherent in most existing blind system identification algorithms:

(i) Sensitivity to channel order selection (in particular undermodelling) [46];

(ii) Even small amounts of noise (SNR < 30 dB) make blind system identification difficult [70];

(iii) The very long acoustic impulse responses of rooms introduce a high computational burden even for the adaptive methods [63];

(iv) Common zeros and near-common zeros pose a severe problem and are linked to the channel length [111].

In this chapter, the issues of noise, common zeros and lengthy impulse responses are investigated in the framework of multichannel adaptive blind system identification. It is shown, that, if the adaptation step-size is carefully controlled, the performance in noisy conditions can be improved with significant gains in terms of system mismatch.

Subsequently, it is demonstrated that the adaptation rate is severely degraded in the presence of common or near-common zeros and, also, that near-common zeros are likely to exist in acoustic systems. Finally, a preliminary study is presented, investigating the possibility of applying the BSI algorithms in conjunction with subband processing as a strategy to partition the problem. It is shown that, in principle, it is possible to perform the adaptation in the subbands, however, there are several issues that need to be resolved before such an approach can be applied in practice.

The remainder of this chapter is organized as follows. In Section 5.2, the problem of blind system identification is formulated and the sufficient identifiability conditions are stated. Adaptive BSI is presented in Section 5.3, where the complex MCLMS algorithm is derived and discussed. In Section 5.4, an optimal step-size is derived and its contribution to performance improvement in noisy and noise-free environments is demonstrated. Next, in Section 5.5 the MCLMS is used to develop an adaptive algorithm for detection and identification of common roots in polynomials, which is applied to investigate the sensitivity of the MCLMS to common and near-common zeros. In Section 5.6 a preliminary study of blind system identification in decimated subbands is presented. The penultimate Section 5.7 demonstrates the ability and the limitations of the MCLMS to identify acoustic impulse responses and finally, a conclusive summary is provided in Section 5.8.

5.2 The Blind SIMO System Identification Problem

In the SIMO system of Fig. 5.1, a signal, $s(n)$, is observed in a noisy multipath environment by an array of sensors at a distance from the source. The signal received at the mth sensor is

$$x_m(n) = \boldsymbol{h}_m^T \boldsymbol{s}(n) + \nu_m(n), \tag{5.1}$$

where $\boldsymbol{h}_m = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T$ is the L-tap impulse response of the channel between the source and the mth sensor, $\boldsymbol{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L+1)]^T$ is the source signal vector and $\nu_m(n)$ is measurement noise at the mth sensor.

The objective of a blind system identification algorithm is to form an estimate $\hat{\boldsymbol{h}}_m = [\hat{h}_{m,0}\ \hat{h}_{m,1}\ \ldots\ \hat{h}_{m,L-1}]^T$ of the impulse responses $\boldsymbol{h}_m$, using only the observations $x_m(n)$, $m = 1, 2, \ldots, M$. It can be seen that this problem is complementary to the problem of dereverberation formulated in Section 2.1, where an estimate of $s(n)$ is sought for. If $\boldsymbol{h}_m$ is available then $s(n)$ can be inferred.

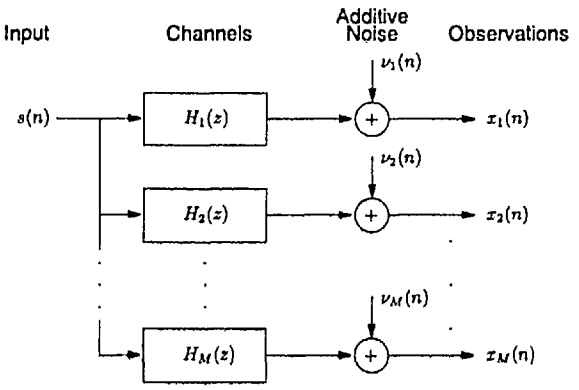

Figure 5.1: SIMO system block diagram.

5.2.1 Channel Identifiability

It has been shown that blind multichannel identification is possible provided that the following sufficient identifiability conditions are satisfied [62]:

(i) The polynomials formed from the channel transfer functions do not share any common zeros across all channels.

(ii) The autocorrelation matrix of the source signal is full rank.

The first condition highlights the importance of channel diversity. Multichannel identification exploits the spectral disparity between different channels, where an extreme counter example is when all zeros between all channels are identical. In this case, the multichannel system reduces to a single channel equivalent. The second condition sets a requirement on the input to have rich spectral excitation. Again, the practical significance of this condition can be illustrated with an extreme example of an input signal constituting a pure tone. In this case, the channel can be identified only at the particular frequency of the excitation tone.

5.2.2 Evaluation of Estimated Channels

An important issue in BSI is evaluation of the mismatch between the true and the estimated system responses. A comprehensive treatment on this topic was presented by Morgan et al [112], where the normalized projection misalignment (NPM) was defined as

$$\text{NPM} = 20\log_{10}\left(\frac{\|\boldsymbol{h} - \kappa\hat{\boldsymbol{h}}\|}{\|\boldsymbol{h}\|}\right) \text{ dB}, \tag{5.2}$$

with

$$\kappa = \frac{\boldsymbol{h}^T\hat{\boldsymbol{h}}}{\hat{\boldsymbol{h}}^T\hat{\boldsymbol{h}}}, \tag{5.3}$$

where κ is a scaling factor, $\boldsymbol{h}$ is the true impulse response and $\hat{\boldsymbol{h}}$ is the estimated impulse response. Substituting (5.3) into (5.2), the NPM expression becomes

$$\text{NPM} = 20\log_{10}\left(\frac{1}{\|\boldsymbol{h}\|}\left\|\boldsymbol{h} - \frac{\boldsymbol{h}^T\hat{\boldsymbol{h}}}{\hat{\boldsymbol{h}}^T\hat{\boldsymbol{h}}}\hat{\boldsymbol{h}}\right\|\right) \text{ dB}. \tag{5.4}$$

It can be seen from (5.4), that, the true channel vector is projected onto the estimated channel vector. In this way, only the misalignment is accounted for, ignoring the effect of any arbitrary scale factor [112]. This is important since many methods identify the channels correctly up to an arbitrary scale factor, as will be shown in Section 5.3. Finally, the estimated channel vector can be made time dependent to accommodate progressive evaluation of adaptive algorithms [66].

5.3 Adaptive Blind SIMO System Identification

In this section, an eigendecompostion method for blind system identification [46] is presented and an adaptive multichannel LMS solution to this problem is derived for the general case of complex signals, extending the method proposed in [66].

5.3.1 The Channel Cross-Relation

Many blind system identification algorithms [62–67, 113] are based on the cross-relation between two channels, which follows from the associative property of convolution [62]

$$x_1 * h_2 = (s * h_1) * h_2 = x_2 * h_1. \tag{5.5}$$

Consequently, in the noise-free case the following relationship is obtained

$$\boldsymbol{x}_m^T(n)\boldsymbol{h}_l = \boldsymbol{x}_l^T(n)\boldsymbol{h}_m, \quad m, l = 1, 2, \ldots, M, \; m \neq l, \tag{5.6}$$

where $\boldsymbol{x}_m(n) = [x_m(n)\ x_m(n-1)\ \ldots\ x_m(n-L+1)]$ is the mth observation vector.

5.3.2 The Eigendecomposition Method

It is now shown how the channel cross-relation can be used to identify the unknown system responses. Premultiplying (5.6) by $\boldsymbol{x}_m(n)$ and taking the expectation results in

$$\boldsymbol{R}_{x_m x_m}\boldsymbol{h}_l = \boldsymbol{R}_{x_m x_l}\boldsymbol{h}_m, \quad m, l = 1, 2, \ldots, M, \; m \neq l \tag{5.7}$$

where

$$\boldsymbol{R}_{x_m x_l} = E\{\boldsymbol{x}_m(n)\boldsymbol{x}_l^T(n)\} \tag{5.8}$$

is the correlation matrix and $E\{\cdot\}$ is the expectation operator. Adding together the cross relations for one particular channel, $\boldsymbol{h}_m$, gives

$$\sum_{m=1, m\neq l}^{M} \boldsymbol{R}_{x_m x_m}\boldsymbol{h}_l = \sum_{m=1, m\neq l}^{M} \boldsymbol{R}_{x_m x_l}\boldsymbol{h}_m, \quad l = 1, 2, \ldots, M, \tag{5.9}$$

and taking all sensor pair combinations into account, a system of equations can be written

$$\boldsymbol{R}\boldsymbol{h} = \boldsymbol{0}, \tag{5.10}$$

where $\boldsymbol{h} = [\boldsymbol{h}_1^T \; \boldsymbol{h}_2^T \; \ldots \; \boldsymbol{h}_M^T]^T$ is a vector of the concatenated impulse responses and

$$\boldsymbol{R} = \begin{bmatrix} \sum_{m\neq 1} \boldsymbol{R}_{x_m x_m} & -\boldsymbol{R}_{x_2 x_1} & \cdots & -\boldsymbol{R}_{x_M x_1} \\ -\boldsymbol{R}_{x_1 x_2} & \sum_{m\neq 2} \boldsymbol{R}_{x_m x_m} & \cdots & -\boldsymbol{R}_{x_M x_2} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{R}_{x_1 x_M} & -\boldsymbol{R}_{x_2 x_M} & \cdots & \sum_{m\neq M} \boldsymbol{R}_{x_m x_m} \end{bmatrix}. \tag{5.11}$$

It is seen from (5.10) that the channel response vector $\boldsymbol{h}$ is the eigenvector of $\boldsymbol{R}$ corresponding to the zeroth eigenvalue. Consequently, if the channel identifiability conditions from Section 5.2.1 are satisfied, the impulse responses of the unknown system can be identified uniquely up to a scale factor by finding the eigenvector corresponding to the zeroth eigenvalue of the correlation matrix $\boldsymbol{R}$ [62]. In the presence of white noise the smallest eigenvalue is sought, whose value is governed by the noise variance [65]. The problem in (5.10) can be solved, for example, using eigenvalue decomposition [65], adaptive methods [63,64,66] or singular value decomposition applied directly on the data matrix [62,65]. In this work, the focus will be on adaptive approaches.

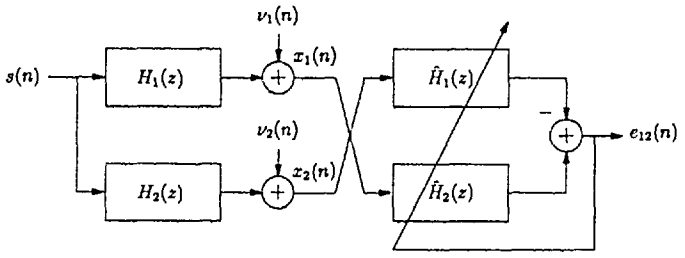

Figure 5.2: System diagram of the multichannel adaptive blind system identification based on the cross-relation between channels for $M = 2$ sensors.

5.3.3 The Complex Multichannel LMS for Blind SIMO system identification

A class of adaptive algorithms for blind identification of SIMO systems has been proposed with implementations both in the time domain [66] and in the frequency domain [63]. In the derivation of these algorithms, it is implicitly assumed that both the signals and the unknown systems are real valued. However, in some cases there is a need for complex adaptive filters. Such cases include blind system identification in subbands where the subband signals may be complex [114] and in particular for oversampled filter banks with an arbitrary oversampling factor [115]. There are also other applications of interest in, for example, communications [116, 117].

In this section, the approach by Widrow et al in [118] for the derivation of the complex LMS adaptive algorithm is employed to extend the work by Huang and Benesty in [66] and to derive the complex multichannel LMS algorithm for blind adaptive identification of SIMO systems.

In the presence of noise and with channel estimates $\hat{\boldsymbol{h}}_m$, the equality in (5.6) does not hold and an error function can be defined as

$$\begin{aligned} e_{ml}(n) &= \boldsymbol{x}_m^T(n)\hat{\boldsymbol{h}}_l - \boldsymbol{x}_l^T(n)\hat{\boldsymbol{h}}_m \\ &= \Re\{e_{ml}(n)\} + j\Im\{e_{ml}(n)\}, \quad m,l = 1,2,\ldots,M,\ m \neq l, \end{aligned} \tag{5.12}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote real and imaginary components respectively. The objective of the complex MCLMS is to adapt simultaneously both the real and the imaginary

components of $\hat{h}$ [118] and consequently, a cost function is formulated

$$\begin{aligned} J(n) &= \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} e_{ml}(n) e_{ml}^{*}(n) \\ &= \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \Re\{e_{ml}(n)\}^2 + \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \Im\{e_{ml}(n)\}^2, \end{aligned} \tag{5.13}$$

where $[\cdot]^*$ denotes complex conjugation.

The optimal estimate of the channels is found by minimizing $J(n)$ with respect to the channel estimates $\hat{h}$,

$$\hat{h}_{\text{opt}} = \arg\min_{\hat{h}} E\{J(n)\}, \quad \text{subject to } \|\hat{h}\| = 1, \tag{5.14}$$

where $E\{\cdot\}$ is the expectation operator and the unit norm constraint is introduced to avoid the trivial estimate $\hat{h} = 0$. In the case of a channel coefficient vector with complex entries, the constraint only affects the magnitude of the solution and not the phase. Enforcing the unit norm constraint at all times, the normalized cost function can be written

$$\tilde{J}(n) = \frac{J(n)}{\|\hat{h}\|^2}. \tag{5.15}$$

The LMS adaptive algorithm finds the desired solution iteratively with the coefficients being updated according to the relation [45, 118]

$$\hat{h}(n+1) = \hat{h}(n) - \mu \nabla \tilde{J}(n), \tag{5.16}$$

where ∇ is the gradient operator and μ is a positive step-size. The gradient estimate consists of a real and an imaginary component

$$\nabla \tilde{J}(n) = \Re\left\{\nabla \tilde{J}(n)\right\} + j\Im\left\{\nabla \tilde{J}(n)\right\}. \tag{5.17}$$

Next, each of these components is evaluated individually. The instantaneous gradient estimate at time n with respect to the real component of the channel vector, $\Re\{h\}$,

gives

$$\begin{aligned}\Re\left\{\nabla\tilde{J}(n)\right\} &= \frac{\partial}{\partial\Re\{\boldsymbol{h}\}}\left(\frac{J(n)}{\|\boldsymbol{h}\|^2}\right)\\ &= \frac{1}{\|\boldsymbol{h}\|^4}\left[\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}\}}\|\boldsymbol{h}\|^2 - 2J(n)\Re\{\boldsymbol{h}\}\right]\\ &= \frac{1}{\|\boldsymbol{h}\|^2}\left[\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}\}}\frac{\|\boldsymbol{h}\|^2}{\|\boldsymbol{h}\|^2} - 2\frac{J(n)}{\|\boldsymbol{h}\|^2}\Re\{\boldsymbol{h}\}\right]\\ &= \frac{1}{\|\boldsymbol{h}\|^2}\left[\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}\}} - 2\tilde{J}(n)\Re\{\boldsymbol{h}\}\right], \end{aligned} \tag{5.18}$$

where

$$\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}\}} = \begin{bmatrix} \left(\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}_1\}}\right)^T \\ \left(\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}_2\}}\right)^T \\ \vdots \\ \left(\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}_M\}}\right)^T \end{bmatrix}.$$

Evaluating the partial derivative of $J(n)$ with respect to the real coefficients of the mth channel only results in

$$\frac{\partial J(n)}{\partial\Re\{\boldsymbol{h}_m\}} = -\sum_{l=1}^{M}\left[\boldsymbol{x}_l^*(n)e_{ml}(n) + \boldsymbol{x}_l(n)e_{ml}^*(n)\right]. \tag{5.19}$$

Similarly, the instantaneous gradient estimate at time n with respect to the imaginary component of the channel vector, $\Im\{\boldsymbol{h}\}$, is

$$\begin{aligned}\Im\left\{\nabla\tilde{J}(n)\right\} &= \frac{\partial}{\partial\Im\{\boldsymbol{h}\}}\left(\frac{J(n)}{\|\boldsymbol{h}\|^2}\right)\\ &= \frac{1}{\|\boldsymbol{h}\|^2}\left[\frac{\partial J(n)}{\partial\Im\{\boldsymbol{h}\}} - 2\tilde{J}(n)\Im\{\boldsymbol{h}\}\right]\end{aligned} \tag{5.20}$$

with

$$\frac{\partial J(n)}{\partial\Im\{\boldsymbol{h}_m\}} = j\sum_{l=1}^{M}\left[\boldsymbol{x}_l^*(n)e_{ml}(n) - \boldsymbol{x}_l(n)e_{ml}^*(n)\right]. \tag{5.21}$$

Next, substituting the results from (5.18), (5.19), (5.20) and (5.21) into (5.17) the instantaneous estimate of the overall gradient is obtained

$$\nabla\tilde{J}(n) = \frac{1}{\|\hat{\boldsymbol{h}}\|^2}\left[2\tilde{\boldsymbol{R}}^*(n)\hat{\boldsymbol{h}}(n) - 2\tilde{J}(n)\hat{\boldsymbol{h}}(n)\right], \tag{5.22}$$

where

$$\tilde{\boldsymbol{R}}(n) = \begin{bmatrix} \sum_{m\neq 1} \tilde{\boldsymbol{R}}_{x_m x_m}(n) & -\tilde{\boldsymbol{R}}_{x_2 x_1}(n) & \cdots & -\tilde{\boldsymbol{R}}_{x_M x_1}(n) \\ -\tilde{\boldsymbol{R}}_{x_1 x_2}(n) & \sum_{m\neq 2} \tilde{\boldsymbol{R}}_{x_m x_m}(n) & \cdots & -\tilde{\boldsymbol{R}}_{x_M x_2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\boldsymbol{R}}_{x_1 x_M}(n) & -\tilde{\boldsymbol{R}}_{x_2 x_M}(n) & \cdots & \sum_{m\neq M} \tilde{\boldsymbol{R}}_{x_m x_m}(n) \end{bmatrix} \tag{5.23}$$

is the instantaneous estimate of the matrix $\boldsymbol{R}$ from (5.11) with

$$\tilde{\boldsymbol{R}}_{x_m x_l}(n) = \boldsymbol{x}_m(n)\boldsymbol{x}_l^H(n), \tag{5.24}$$

and $\hat{\boldsymbol{h}}(n) = [\hat{\boldsymbol{h}}_1^T(n)\ \hat{\boldsymbol{h}}_2^T(n)\ \ldots\ \hat{\boldsymbol{h}}_M^T(n)]^T$ is a vector of the concatenated channel estimates at time n.

Finally, substituting (5.22) into (5.16) and assuming that the channel estimates are normalized after each iteration, as in [66], the update equation for the complex MCLMS algorithm becomes

$$\hat{\boldsymbol{h}}(n+1) = \frac{\hat{\boldsymbol{h}}(n) - 2\mu[\tilde{\boldsymbol{R}}^*(n)\hat{\boldsymbol{h}}(n) - J(n)\hat{\boldsymbol{h}}(n)]}{\|\hat{\boldsymbol{h}}(n) - 2\mu[\tilde{\boldsymbol{R}}^*(n)\hat{\boldsymbol{h}}(n) - J(n)\hat{\boldsymbol{h}}(n)]\|}. \tag{5.25}$$

The result in (5.25) differs from the result presented in [66] in that the complex conjugation is applied to the correlation matrix, $\boldsymbol{R}$. This is also consistent with the result in [118] where complex conjugation is applied to the input data vector. In Section 5.4, the choice of step-size is discussed and an optimal step-size is derived for the MCLMS algorithm.

5.3.4 Effects of the Complex Factor

The solutions obtained from the eigenvalue decomposition methods are accurate up to an arbitrary constant, κ, as will be demonstrated in Section 5.3.5. The effects of this factor are now formulated and discussed. In the general case of complex data, this constant is essentially a complex number, $\kappa = |\kappa|e^{j\theta_\kappa}$, with a magnitude $|\kappa| = \sqrt{\Re\{\kappa\}^2 + \Im\{\kappa\}^2}$ and a phase $\theta_\kappa = \Im\{\ln(\kappa)\}$. Therefore, it introduces an arbitrary additive phase ambiguity in addition to the gain ambiguity occurring for real signals [62, 66]. At convergence, the uth coefficient of the mth estimated channel vector is related to the corresponding true

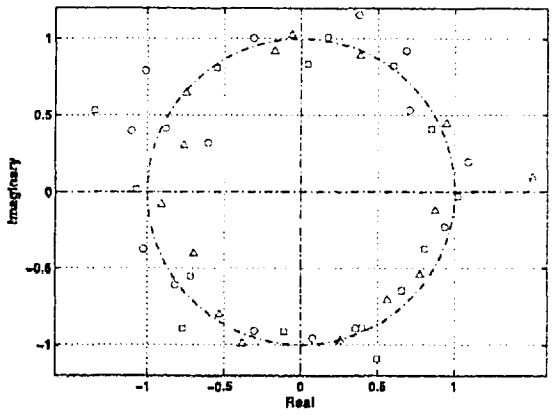

Figure 5.3: Channel zeros of a system of $M = 3$ for Channel 1 (squares), Channel 2 (circles) and Channel 3 (triangles).

channel coefficient according to

$$\hat{h}_{m,u} = |\kappa||h_{m,u}|e^{j(\theta_{h_m}(u)+\theta_\kappa)}, \tag{5.26}$$

where $\theta_{h_m}(u) = \Im\{\ln(h_{m,u})\}$ is the phase of $h_{m,u}$. This makes it difficult to use the estimates in, for example, equalization where the recovered signal would also have a phase error. Thus, it is desirable to compensate for the additive angle θ_m and ideally also for the gain ambiguity. A straightforward approach can be used if the phase of one true tap value is known. In practice, one such known value could be inferred from the direct path component of the fullband impulse responses [69, 119, 120].

Moreover, experiments have shown that if the adaptive algorithm is initialized with a vector whose uth value is set to $\kappa_0 = |\kappa_0|e^{j\theta_{\kappa_0}}$, such that $\hat{\mathbf{h}}(0) = [0 \ \ldots \ \kappa_0 \ \ldots \ 0]^T$, the algorithm will converge to a solution where the phase of that same uth value of the estimated channel is equal to θ_{κ_0}. Consequently, *a priori* knowledge of the phase of one true tap can be used in the initialization of the algorithm to avoid phase errors.

5.3.5 Simulation Examples of the Complex MCLMS

Simulation results are now presented in order to validate the performance of the algorithm in (5.25). These example experiments utilize a system of $M = 3$ random complex channels of length $L = 16$. The channel zeros are shown in Fig. 5.8 where squares, circles and

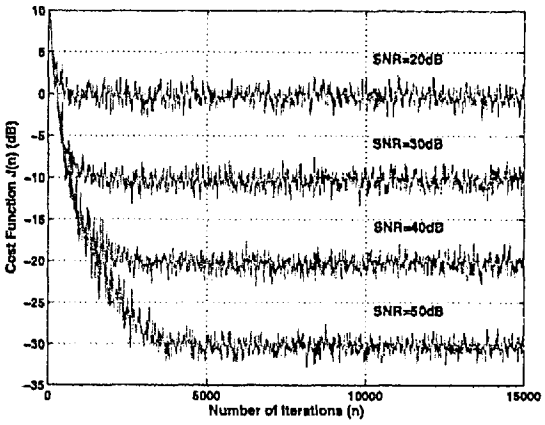

Figure 5.4: Cost function trajectory for the Complex MCLMS and for varying levels of SNR with $M = 3$ random channels of length $L = 16$.

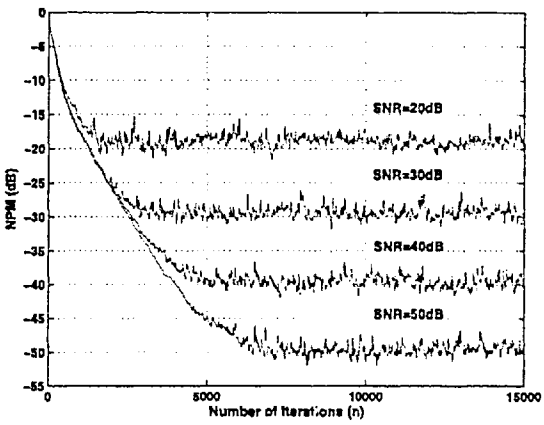

Figure 5.5: Normalized projection misalignment for the Complex MCLMS and for varying levels of SNR with $M = 3$ random channels of length $L = 16$.

triangles indicate each of the three distinct channels. It was assured that there are no common zeros between the channels and the input was complex white Gaussian noise so as to satisfy the identifiability conditions stated in Section 5.2.1. The SNR was varied in the range $20 - 50$ dB and the adaptation step-size was set to $\mu = 10^{-4}$.

Trajectories of the cost function for a typical run and for different noise levels are presented in Fig. 5.4. In Fig. 5.5 the plot shows the corresponding result in terms of NPM. It can be seen that the channels are identified correctly, up to a multiplicative complex factor and hence, it can be concluded that the complex MCLMS algorithm is operating correctly.

5.4 Optimal Step-Size for Blind Multichannel LMS System Identification

The choice of step-size in adaptive blind system identification using the multichannel least mean squares algorithm is critical and controls its convergence rate, stability and sensitivity to noise [45]. In this section, an optimal step-size for the MCLMS algorithm is derived and its properties are investigated. An implementation technique for the Wiener solution of this self-adaptive step-size is presented and it is shown that considerable performance improvements are achieved compared to existing approaches in the presence of noise.

Previous work on non-blind adaptive system identification has shown that employment of variable and optimal step-sizes results in performance exceeding that of a fixed step-size, e.g. [121] and [122]. An early approach for optimal variable step-size in blind channel identification was presented in [67] for the unconstrained MCLMS, where the optimal step-size is obtained at each iteration using the instantaneous values of the channel and the gradient estimates.

In the work presented here, a general optimal step-size is derived for the unconstrained MCLMS algorithm, extending the previous approach to both noisy and noise-free cases. In a noise-free environment, the optimal step-size in [67] becomes a special case of the herein proposed Wiener step-size. In the presence of noise, an underlying assumption in [67] is invalid and it is shown that, in such cases, the general optimal step-size can give a significant improvement.

5.4.1 The Unconstrained MCLMS

It was shown in [67] that the trivial solution of (5.10) can be avoided if the estimation vectors are initialized appropriately and consequently, the unit norm constraint can be excluded. This leads to a simplified version of the gradient defined in (5.22), which now becomes [67]

$$\nabla J(n) = 2\check{\boldsymbol{R}}(n)\hat{\boldsymbol{h}}(n), \tag{5.27}$$

and the update equation in (5.25) reduces to

$$\hat{\boldsymbol{h}}(n+1) = \hat{\boldsymbol{h}}(n) - 2\mu\tilde{\boldsymbol{R}}(n)\hat{\boldsymbol{h}}(n), \tag{5.28}$$

where $\tilde{\boldsymbol{R}}(n)$ is defined in (5.23). This unconstrained MCLMS is used for the development and the experiments of the optimal step-size [123].

5.4.2 Wiener Solution of the Self Adaptive Step-Size

Now the general optimal step-size is derived and also, the implementation issues regarding this step-size are discussed.

Step-size derivation

The concept of an optimal step-size is defined here as the step-size, $\mu(n)$, which minimizes the misalignment at every iteration, given the current channel estimate $\hat{\boldsymbol{h}}(n)$.

Consequently, a cost function is defined

$$J_\mu(n) = E\{\|\boldsymbol{h} - \kappa\hat{\boldsymbol{h}}(n+1)\|^2|\hat{\boldsymbol{h}}(n)\}, \tag{5.29}$$

where κ is the scaling constant inherent in blind system identification based on the cross-relation and is assumed, for now, to be known. Minimizing $J_\mu(n)$ with respect to μ gives the optimal step-size at time n,

$$\mu_{\text{opt}}(n) = \arg\min_{\mu} J_\mu(n). \tag{5.30}$$

Substituting (5.16) into (5.29) the cost function becomes

$$\begin{aligned} J_\mu(n) &= E\{\|\boldsymbol{h} - \kappa\hat{\boldsymbol{h}}(n) + \kappa\mu\nabla J(n)\|^2|\hat{\boldsymbol{h}}(n)\} \\ &= E\{\|\boldsymbol{h}\|^2 - 2\kappa\boldsymbol{h}^T\hat{\boldsymbol{h}}(n) + 2\kappa\mu\boldsymbol{h}^T\nabla J(n) \\ &\quad -2\kappa^2\mu\hat{\boldsymbol{h}}^T(n)\nabla J(n) + \kappa^2\mu^2\|\nabla J(n)\|^2 + \kappa^2\|\hat{\boldsymbol{h}}(n)\|^2|\hat{\boldsymbol{h}}(n)\}. \end{aligned} \tag{5.31}$$

The optimal step-size in the MMSE sense is obtained from

$$\frac{\partial J_\mu(n)}{\partial\mu} = 0 \tag{5.32}$$

Thus, using (5.31) and (5.32), the Wiener step-size can be written

$$\mu_{\text{opt}}(n) = \frac{\hat{\boldsymbol{h}}^T(n) - \frac{1}{\kappa}\boldsymbol{h}^T}{E\{\|\nabla J(n)\|^2|\hat{\boldsymbol{h}}(n)\}} E\{\nabla J(n)|\hat{\boldsymbol{h}}(n)\} \tag{5.33}$$

As in the derivation of the LMS algorithm [45], the expected values may be approximated by their instantaneous estimates. Let $\nabla J(n)$ and $\|\nabla J(n)\|^2$ be the instantaneous estimates of the conditional expectations $E\{\nabla J(n)|\hat{\boldsymbol{h}}(n)\}$ and $E\{\|\nabla J(n)\|^2|\hat{\boldsymbol{h}}(n)\}$ respectively. The optimal step-size is then given by

$$\tilde{\mu}_{\text{opt}}(n) = \frac{\hat{\boldsymbol{h}}^T(n)\nabla J(n)}{\|\nabla J(n)\|^2} - \gamma(n), \tag{5.34}$$

with

$$\gamma(n) = \frac{\boldsymbol{h}^T\nabla J(n)}{\kappa\|\nabla J(n)\|^2}. \tag{5.35}$$

Step-size implementation

In the noise-free case, $\nu_m(n) = 0$, $\forall m$ and it can be assumed that the gradient vector is orthogonal to the true solution at all times, such that $\hat{\boldsymbol{h}}^T \nabla J(n) \approx 0$, as discussed in [67]. In this case, the solution in (5.34) reduces to

$$\mu_1(n) = \frac{\hat{\boldsymbol{h}}^T(n)\nabla J(n)}{\|\nabla J(n)\|^2}, \tag{5.36}$$

which is the result presented in [67] and can be regarded as a special case of the Wiener step-size in (5.33). Figure 5.6 shows an example plot of the optimal step-size components a) $\mu_1(n)$ from (5.36) and b) $\gamma(n)$ from (5.35), which together form $\tilde{\mu}_{\text{opt}}(n)$ (5.34). The first 20 samples ($n = 0, 1, \ldots, 19$) in Fig. 5.6a have been excluded from the plot since they are generally large and shadow the fine detail which is of interest here. It is interesting to note that the step-size $\mu_1(n)$ varies in the range $0 - 0.05$ rather than approaching zero with convergence as might be expected. Also, it is seen that $\gamma(n)$ varies in a range very close to zero.

The assumption of orthogonality between the gradient and the true solution does not hold in the presence of noise, i.e. $\hat{\boldsymbol{h}}^T \nabla J(n) \neq 0$, and the approximation employed in (5.36) becomes inaccurate, even when small amounts of noise are introduced. Thus, $\gamma(n)$ is not zero and becomes more significant as the SNR decreases. This can be seen in Fig. 5.7b, where $\gamma(n)$ is plotted for SNR = 20 dB. The examples in Figs. 5.6 and 5.7 were generated using the simulation settings described in Section 5.4.3.

In order to implement the optimal step-size in noisy conditions it is necessary to estimate $\gamma(n)$ as it requires knowledge of the true impulse responses which are not available. Thus, an approximate expression $\hat{\gamma}(n) \approx \gamma(n)$ needs to be formed. It can be seen that the first term of (5.34) and $\gamma(n)$ will tend to equivalence as the channel estimates $\hat{\boldsymbol{h}}$ approach the true solution. This can also be observed in Fig. 5.7 where after approximately 5×10^3 iterations $\mu_1(n)$ is close to $\gamma(n)$. Therefore, it can be concluded that the purpose of $\gamma(n)$ is to attenuate the step-size, bringing it towards zero as the channel estimates improve. Using this as motivation, $\hat{\gamma}(n)$ can be written

$$\hat{\gamma}(n) = \beta(n)\frac{\hat{\boldsymbol{h}}^T(n)\nabla J(n)}{\|\nabla J(n)\|^2}, \tag{5.37}$$

where $\beta(n)$ is a weighting function which should slowly attenuate the step-size as the

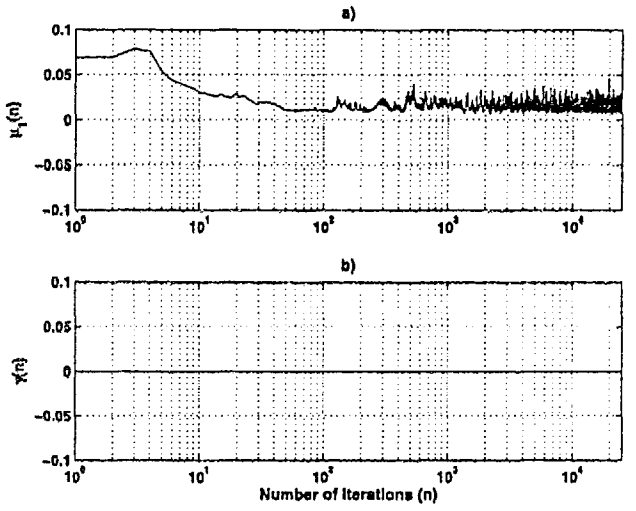

Figure 5.6: Step-size behaviour in the noise-free case for a) $\mu_1(n)$ and b) $\gamma(n)$, which together form $\tilde{\mu}_{\text{opt}}(n)$.

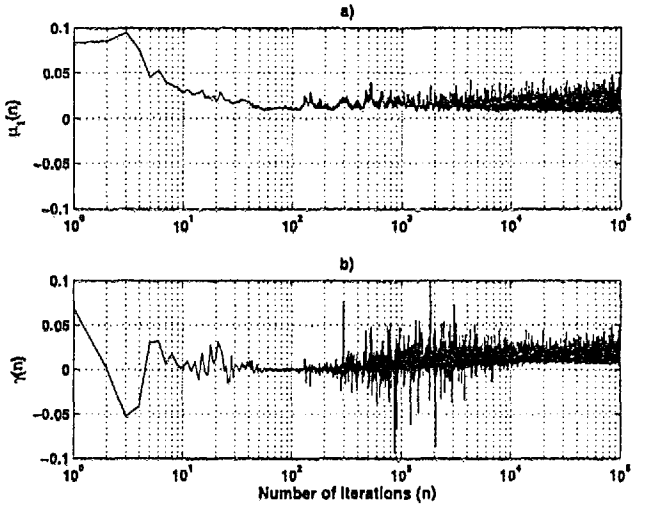

Figure 5.7: Step-size behaviour at SNR = 20 dB for a) $\mu_1(n)$ and b) $\gamma(n)$, which together form $\tilde{\mu}_{\text{opt}}(n)$.

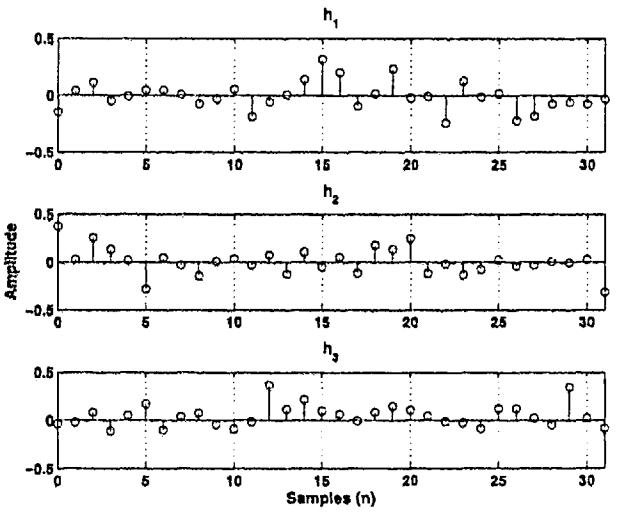

Figure 5.8: Three random channels of length $L = 32$ used in the simulations.

adaptive filter converges. Replacing $\gamma(n)$ in (5.34) with (5.37), an approximate Wiener step-size for noisy conditions is obtained

$$\mu_2(n) = (1 - \beta(n))\frac{\hat{h}^T(n)\nabla J(n)}{\|\nabla J(n)\|^2}. \tag{5.38}$$

A weighting function which was found to be suitable is

$$\beta(n) = \exp\left(\frac{-J(n)L\mathcal{P}_x}{\mathcal{P}_\nu}\right), \tag{5.39}$$

where $\mathcal{P}_x$ and $\mathcal{P}_\nu$ are the observed signal power and the noise power respectively. The expression in (5.39) highlights some of the desired features of the weighting function:

(i) it tends to unity as the error tends to zero;

(ii) it tends to zero as the SNR tends to infinity;

(iii) it is dependent on the error signal and therefore provides a means of tracking changes in the system.

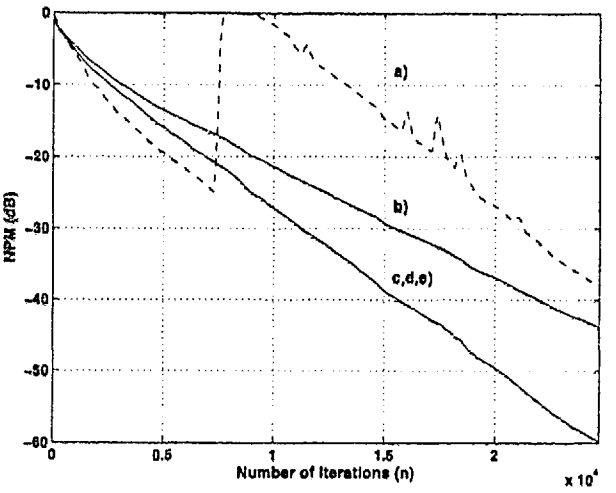

Figure 5.9: Normalized projection misalignment in the noise-free case for the unconstrained MCLMS with step-size a) $\mu = 0.02$, b) $\mu = 0.01$, c) $\tilde{\mu}_{\text{opt}}(n)$ d) $\mu_1(n)$ and e) $\mu_2(n)$.

5.4.3 Simulation Results: optimal step-size

Simulation results are presented to investigate the use of the step-size parameters $\tilde{\mu}_{\text{opt}}(n)$, $\mu_1(n)$ and $\mu_2(n)$ defined in Section 5.4. For the experiments three random channels of length $L = 32$ were used and are shown in Fig. 5.8. The excitation signal was white Gaussian noise and the sampling frequency was set to $f_s = 8$ kHz. In all cases the adaptive algorithms were initialized with $\hat{h}(0) = [1\ 1\ \ldots\ 1]^T$.

Experiment 1: a noise-free case

First, the noise-free case was considered. The unconstrained MCLMS with update equation according to (5.28) was executed with fixed and with variable optimal step-sizes. The results are shown in Fig. 5.9 where the NPM is plotted against time for a) a fixed step-size $\mu = 0.02$, b) a fixed step-size $\mu = 0.01$, c) the optimal step-size $\tilde{\mu}_{\text{opt}}(n)$ d) the optimal step-size from [67] $\mu_1(n)$ and e) the approximate optimal step-size $\mu_2(n)$ in (5.38). Figures 5.9a and 5.9b demonstrate the difficulty in choosing a fixed step-size since $\mu = 0.02$ results in instability. As expected for the noise-free case, all three implementations of the optimal step-size perform uniformly resulting in a monotonic convergence.

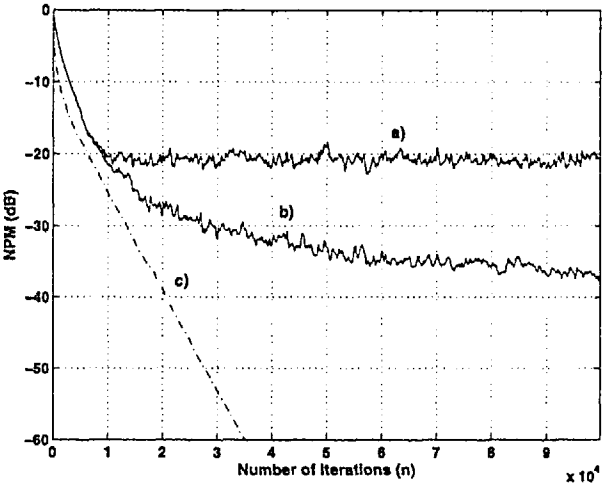

Figure 5.10: Normalized projection misalignment in the noisy case (SNR = 20 dB) for the unconstrained MCLMS with step-size a) $\mu_1(n)$, b) $\mu_2(n)$ and c) $\tilde{\mu}_{\text{opt}}(n)$.

Experiment 2: a case with additive white noise

Next, the case of additive noise was investigated for an example case with SNR=20 dB. The three implementations of the variable step-sizes were employed with the unconstrained MCLMS. The resulting NPM is plotted against time iterations n in Fig. 5.10 for a) the optimal step-size from [67] $\mu_1(n)$, b) the approximate optimal step-size $\mu_2(n)$ in (5.38) and c) the optimal step-size $\tilde{\mu}_{\text{opt}}(n)$. Two interesting observations can be made from this: (i) the implementation with the general optimal step-size $\tilde{\mu}_{\text{opt}}$ converges towards zero even under noisy conditions and thus provides an indication of the obtainable NPM and (ii) using approximations like that of (5.38) can provide a significant improvement in terms of misalignment as seen in Fig. 5.10b.

In summary, a Wiener step-size was derived for the unconstrained multichannel LMS algorithm for blind channel identification. This step-size is calculated at every iteration. In order to implement the optimal step-size in noisy environments an approximate Wiener step-size has been formulated. Simulation results have demonstrated that the proposed self-adaptive step-size gives significantly better NPM compared to a fixed step-size or the optimal step-size algorithm in [67] in the presence of observation noise.

5.5 Common Roots Detection and Estimation and the Common Zeros Problem in Adaptive Blind System Identification

It was described in Section 5.2.1 that multichannel blind system identification relies on the identifiability condition of all channel transfer functions being coprime, i.e. that they do not have any zeros in common. This prerequisite is common to many blind system identification algorithms, e.g. [46,63,64,66,72] and also for multichannel inversion [68,80]. Despite its significance this property has received little attention in the literature. One reason for not studying the common zero problem is the difficulty of reliably factoring high degree polynomials such as those arising in acoustic signal processing.

In this section, a new method is derived for adaptive and exact identification of the common roots between two polynomials without having to factor the polynomials. This is a two-step approach where, first, the distinct zero components of the two polynomials are identified blindly followed by the second step of (non-blind) estimation of the common zeros component. This approach can also be used repeatedly to detect the number of common zeros. Furthermore, the algorithm is applied to investigate how close roots have to be in order to be detected as common, which appears to not be necessarily only when they are identical. As a consequence of this, it is demonstrated that the performance of the adaptive blind channel estimation algorithms is degraded severely when the zeros between two channels get close.

The idea of detecting and identifying the common roots of polynomials without having to factor them has been of interest for many years in mathematics and control systems theory. Mainly the interest has been in binary decisions of whether or not two polynomials are coprime and methods utilizing the Sylvester matrix of the polynomial coefficients are commonly applied [124]. An approximate common factor estimation and detection method was proposed in [125] and was recently extended further to the detection and identification of common roots in more than two polynomials [126].

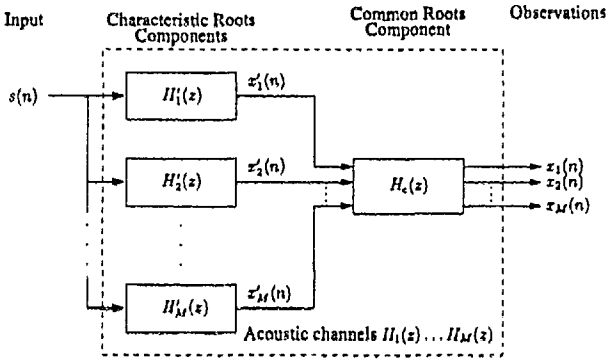

Figure 5.11: SIMO system with common zeros.

5.5.1 The Common Zeros Problem Formulation

Here the SIMO system of Fig. 5.1 is considered in a noise free environment such that $\nu_m(n) = 0$. The mth observation is then given by

$$x_m(n) = \boldsymbol{h}_m^T \boldsymbol{s}(n), \tag{5.40}$$

where $\boldsymbol{h}_m$ and $\boldsymbol{s}(n)$ were defined in Section 5.2. This can be written in the z-domain

$$X_m(z) = H_m(z)S(z), \tag{5.41}$$

where $X_m(z)$, $H_m(z)$ and $S(z)$ are the z-transforms of $x_m(n)$, $\boldsymbol{h}_m$ and $s(n)$ respectively. Given the input and the output sequences, $s(n)$ and $x_m(n)$, the objective is to find the zeros that are common to all transfer functions, $H_m(z)$, $m = 1, 2, \ldots, M$.

Let $H_c(z)$ denote the component with the roots common to all transfer function polynomials and let $H_m'(z)$ denote the characteristic zeros component of the mth channel, i.e. those zeros contained in $H_m(z)$ but not in $H_l(z)$, $l = 1, 2, \ldots, M$, $m \neq l$. The transfer functions in (5.41) can now be rewritten

$$H_m(z) = H_c(z)H_m'(z), \qquad m = 1, 2, \ldots, M, \tag{5.42}$$

where $\deg[H_c(z)] \leqslant \deg[H_m(z)]$. This decomposition is illustrated in Fig. 5.11. Thus, the problem is to detect the number of common roots, $\deg[H_c(z)]$, and to identify the common roots component, $H_c(z)$.

5.5.2 Adaptive Common Roots Estimation

A new method for identification of common roots is now derived. This process consists of two stages. First, the components of the impulse response that do not contain common roots are blindly identified using the algorithm described in Section 5.3.3. Then, using the estimates from stage one, the common roots component of the transfer function is identified.

Sage 1: Estimating the characteristic roots components

Initially, it is assumed that the number of common roots is known, which is not true in practice, and a method for detecting the number of common roots is provided in Section 5.5.3. By substituting (5.42) into (5.41), the system output equations can be written

$$\begin{aligned} X_m(z) &= S(z)H_c(z)H_m'(z) \\ &= X_c(z)H_m'(z), \qquad m = 1, 2, \ldots, M \end{aligned} \tag{5.43}$$

where $X_c(z) = S(z)H_c(z)$ is the signal component common to all M channels.

Since the transfer functions, $H_m'(z)$, do not contain any common zeros, they can be identified blindly using, for example, the MCLMS algorithm [63,66] from Section 5.3.3. The error signal based on the cross-relation between the channels defined in (5.12) can be formulated using only the portions of the transfer functions with the characteristic zeros

$$e_{ml}'(n) = \boldsymbol{x}_m^T(n)\hat{\boldsymbol{h}}_l' - \boldsymbol{x}_l^T(n)\hat{\boldsymbol{h}}_m', \quad m, l = 1, 2, \ldots, M, \; m \neq l, \tag{5.44}$$

where $\hat{\boldsymbol{h}}_m' = [\hat{h}_{m,0}'(n) \; \hat{h}_{m,1}'(n) \; \ldots \; \hat{h}_{m,L'-1}'(n)]^T$ are the estimates of $\boldsymbol{h}_m'$ at time n and L' is the length of the characteristic zeros components of the channels. The room impulse response length L is assumed to be known *a priori*, which is common practice in blind channel estimation. Consequently, all impulse responses can be estimated simultaneously with the algorithm derived in Section 5.3.3, which, assuming only real data, becomes

$$\hat{\boldsymbol{h}}'(n+1) = \frac{\hat{\boldsymbol{h}}'(n) - 2\mu[\tilde{\boldsymbol{R}}(n)\hat{\boldsymbol{h}}'(n) - J(n)\hat{\boldsymbol{h}}'(n)]}{\|\hat{\boldsymbol{h}}'(n) - 2\mu[\tilde{\boldsymbol{R}}(n)\hat{\boldsymbol{h}}'(n) - J(n)\hat{\boldsymbol{h}}'(n)]\|}, \tag{5.45}$$

where $\hat{\boldsymbol{h}}'(n) = [\hat{\boldsymbol{h}}_1'^T(n) \; \hat{\boldsymbol{h}}_2'^T(n) \; \ldots \; \hat{\boldsymbol{h}}_M'^T(n)]^T$ is a vector of the concatenated impulse responses.

Step 2: Estimating the common roots components

The channel estimates obtained from (5.45) are utilized now, to generate the intermediate signals $X'_m(z) = S(z)\hat{H}'_m(z)$, $m = 1, 2, \ldots, M$ and the common zero component, $H_c(z)$, can be found by estimating the outputs $X'_m(z)$. The estimation error for the mth channel can be written

$$e_{c,m}(n) = x_m(n) - x'^T_m(n)\hat{h}_c, \tag{5.46}$$

where $x'_m(n) = [x'_m(n)\ x'_m(n-1)\ \ldots\ x'_m(n-L_c+1)]^T$ is the mth channel input vector at time n, $\hat{h}_c(n)$ is the vector of the estimated common zero component and $L_c = \deg[H_c(z)]+1$ is the length of h_c.

Combining the error functions of all channels, a cost function is formulated

$$\begin{aligned} J_c(n) &= E\{e^2_{c,1}(n) + e^2_{c,2}(n) + \ldots + e^2_{c,M}(n)\} \\ &= \sum_{m=1}^{M} E\{e^2_{c,m}(n)\}. \end{aligned} \tag{5.47}$$

The common zeros component of the M transfer functions is thus found by minimizing the cost function in (5.47)

$$\hat{h}_{c,opt} = \arg\min_{\hat{h}_c} J_c(n). \tag{5.48}$$

Subsequently, a least mean square (LMS) adaptive algorithm [45] is employed to solve the minimization problem in (5.48) and to efficiently estimate the common zeros component. The iterative update of the estimated coefficients is written

$$\hat{h}_c(n+1) = \hat{h}_c(n) - \mu_c \nabla J_c(n) \tag{5.49}$$

where μ_c is the adaptation step size. The gradient is calculated, taking the partial derivatives of $J_c(n)$ with respect to h_c

$$\begin{aligned} \frac{\partial J_c(n)}{\partial h_c} &= \sum_{m=1}^{M} E\{\frac{\partial e^2_{c,m}(n)}{\partial h_c}\} \\ &= \sum_{m=1}^{M} E\{2\frac{\partial e_{c,m}(n)}{\partial h_c} e_{c,m}(n)\} \\ &= -2\sum_{m=1}^{M} E\{x'_m(n) e_{c,m}(n)\}. \end{aligned} \tag{5.50}$$

Inserting the estimation errors from (5.46) into (5.50) the expression for the gradient becomes

$$\nabla J_c(n) = -2\sum_{m=1}^{M} \boldsymbol{p}_m + 2\sum_{m=1}^{M} \boldsymbol{R}_m \boldsymbol{h}_c, \tag{5.51}$$

where $\boldsymbol{R}_m = E\{\boldsymbol{x}'_m(n)\boldsymbol{x}'^T_m(n)\}$ is the autocorrelation matrix of the input signal and $\boldsymbol{p}_m = E\{\boldsymbol{x}'_m(n)x_m(n)\}$ is the cross-correlation between the input and the desired output for the mth channel.

Considering the instantaneous estimates of the autocorrelation matrix, $\tilde{\boldsymbol{R}}_m = \boldsymbol{x}'_m(n)\boldsymbol{x}'^T_m(n)$, and the cross-correlation vector, $\tilde{\boldsymbol{p}}_m = \boldsymbol{x}'_m(n)x_m(n)$, the instantaneous estimate of the gradient in (5.51) can be written

$$\tilde{\nabla} J_c(n) = -2\boldsymbol{X}'(n)\boldsymbol{e}(n), \tag{5.52}$$

where $\boldsymbol{X}'(n) = [\boldsymbol{x}'_1(n)\ \boldsymbol{x}'_2(n)\ \ldots\ \boldsymbol{x}'_M(n)]$ is the input matrix and $\boldsymbol{e}(n) = [e_1(n)\ e_2(n)\ \ldots\ e_M(n)]^T$ is the error vector at time n.

Finally, substituting the result in (5.52) into (5.49), the final coefficient update equation is obtained:

$$\hat{\boldsymbol{h}}_c(n+1) = \hat{\boldsymbol{h}}_c(n) + 2\mu_c \boldsymbol{X}'(n)\boldsymbol{e}(n). \tag{5.53}$$

5.5.3 Common Roots Detection

This far it has been assumed that the order of the common zeros component is known. Since this is not the case in practice it is necessary to detect automatically the number of common zeros. One solution is to perform the identification repeatedly starting with the full channel length L then reducing this by 1 at each repetition and monitoring the mean squared error after convergence. In this way, the initial assumption is that there are no common roots and then to increment the number of common roots for each subsequent repetition. The true number of common roots is indicated when the mean square error, $J_c(n)$, is minimum as will be demonstrated by the simulation results in the following section.

5.5.4 Simulation Results: common zeros detection and identification

Simulations and results are presented to demonstrate the proposed algorithm. As a performance metric, the normalized projection misalignment (NPM) is used, which was defined in Section 5.2.2.

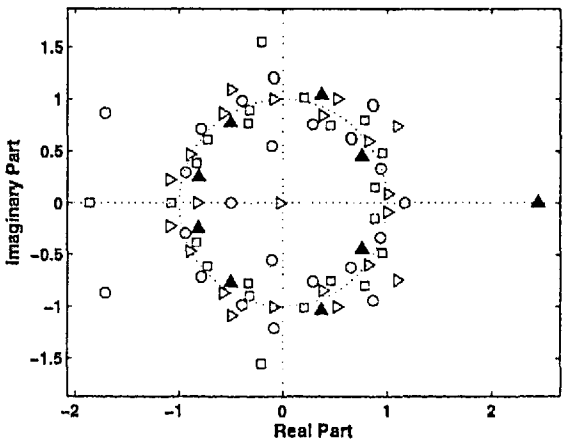

Figure 5.12: Channel zeros for channel one (circles), channel two (squares), channel three (triangles) and the common zeros (filled triangles).

Experiment 1: common roots estimation

For the first experiment, a system was used, comprising three random channels of length $L = 32$ with eight known common roots, i.e. $L_c = 9$. The channel zeros are plotted in Fig. 5.12, showing the characteristic roots for channel one (circles), two (squares), and three (triangles) and the common roots (filled triangles). The transfer functions were explicitly designed such that the minimum inter-channel separation between characteristic roots satisfies the relation $\Delta_{m,l}(u) \geq 0.1$, $\forall u$, where $\Delta_{m,l}(u) = \min\{|z_m(u) - \mathbf{z}_l|\}$, $u = 0, 1, \ldots, L'-2$, $m, l = 1, 2, \ldots, M$ is the distance between the uth zero in the mth transfer function, $z_m(u)$, and any other zero of the lth transfer function, $\boldsymbol{z}_l = [z_m(0)\ z_m(1)\ \ldots\ z_m(L'-2)]$ being a vector of the mth channel characteristic zeros.

The algorithm was run with step-size $\mu = 10^{-5}$ for the blind identification and $\mu_c = 0.2$ for the common zeros estimation. The system under consideration was excited with white Gaussian noise. The results are summarized in Table 5.1 where it is shown that the blind channel identification stage converged, in terms of NPM, to -60 dB in 39974 iterations while the second part, the common root estimation adaptive filter, converged to -60 dB in 240 iterations. It is important to note that the blind channel estimator would continue to converge, while the common roots identification has a convergence floor governed by the accuracy of Step 1.

Component Estimation	Convergence to NPM(n) = −60dB
Characteristic zeros	$n = 39974$
Common zeros	$n = 240$

Table 5.1: Number of iterations, n, required for the algorithm to converge to NPM = −60 dB for a channel of length $L = 32$ and with 8 common roots.

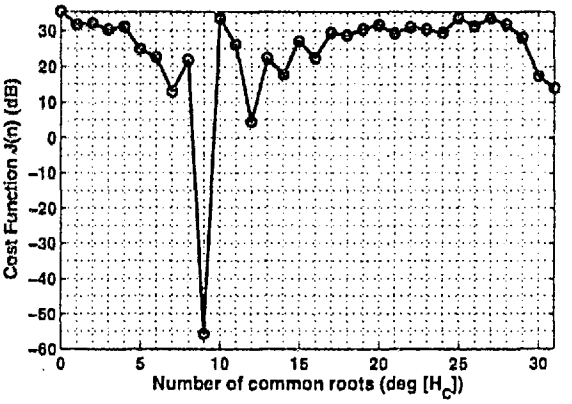

Figure 5.13: Cost function after $n = 200000$ iterations vs. estimated number of common zeros $\deg[H_c(z)]$.

Experiment 2: common roots detection

Subsequently, the case where the order of the common zeros component is unknown was considered and the number of common roots is detected by repetition as described in Section 5.5.3. The correct number of common roots gives the minimum mean squared error in the common roots identification algorithm. Using the same channels in Fig. 5.12, the algorithm was run repeatedly increasing the assumed number of common zeros each time. The algorithm was let to run for $n = 200000$ iterations at each repetition. The result is shown in Fig. 5.13 where it can be seen that the mean squared error after convergence is minimum for 8 common zeros, which is the desired outcome.

Experiment 3: common and near common zeros

Finally, an experiment was performed to examine how close two zeros have to be in order to be detected as common by the algorithm. For clarity and tractability, different channels were used here with only two common zeros as shown in Fig. 5.14 where the characteristic

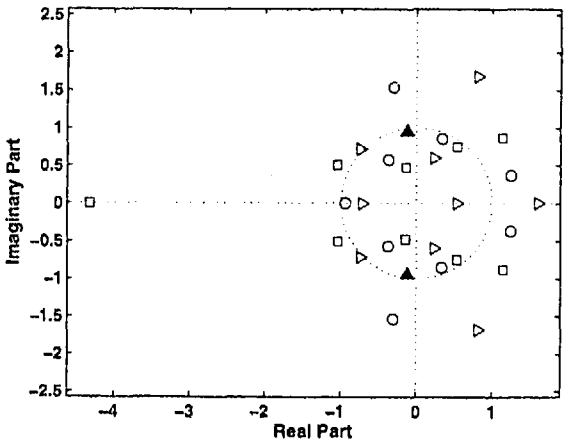

Figure 5.14: Channel zeros for channel one (circles), channel two (squares), channel three (triangles) and the common zeros (filled triangles).

roots for channel one are marked with circles, those for channel two with squares, the zeros for channel three are marked with triangles and the common roots with filled triangles. Here the common zeros are divided into two classes. A zero is considered to be 'common' if it occurs in all channels of a multichannel system at the same location in z. Zeros are considered 'near-common' if they are not identical in all channels but are nevertheless sufficiently close to be identified as common by the adaptive algorithm.

The inter-channel roots distance was set to satisfy $\Delta_{m,l}(u) \geq 0.2$, $\forall u$. Then, keeping all the remaining zeros fixed, the 'common' zeros were separated to a distance of $\Delta_c(u) = 0.2$ and moved towards each other at the steps $\Delta_c = 10^{-i}$ for $i = 1, 2, \ldots, 6$. $\Delta_c(u) = |z_{c,m}(u) - z_{c,l}(u)|$, $u = 0, 1, \ldots, L_c - 2$ is the distance between the near common zeros. The result is shown in Fig. 5.15. it can be observed that zeros which are common to within $\Delta_c(u) = 0.01$ can be correctly identified as common. This gives rise to two interesting points. First, the algorithm is resistant to small perturbations of the common zeros of two polynomials. Second, the identification of the full length channels is severely degraded at that distance. The latter also conforms with the result in [66], for an ill-conditioned channel.

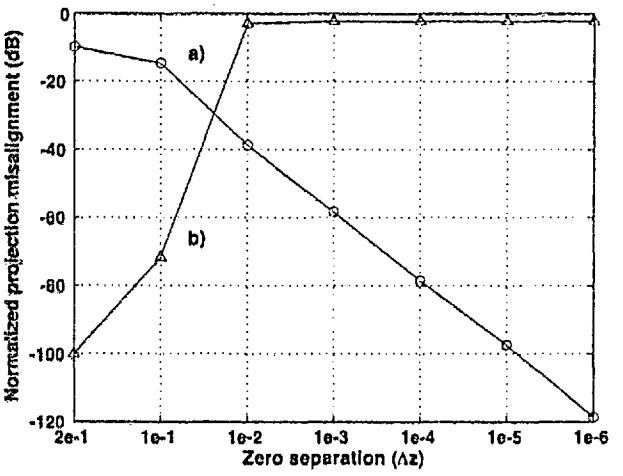

Figure 5.15: Misalignment after convergence vs. near common zeros separation for a) the common zeros estimation algorithm and b) blind estimation of the full channel.

5.5.5 Common Zeros in Acoustic Systems

From the results presented in Section 5.5.2, it is evident that near common zeros can significantly perturb the blind system identification algorithm. An obvious consequent issue is then to investigate the existence of such common or near-common zeros in acoustic systems, which forms the topic of this section.

For long FIR systems, the density of zeros in the vicinity of the unit circle in z is expected to be high [127] and therefore there will be many common or near-common zeros. This has been exactly characterized for random channels in [127], where it is shown that the zeros of high order random polynomials tend to cluster on the unit circle with uniformly distributed phases. This effect is demonstrated here with a simulation example. Figure 5.16a shows the mean distance of the zeros from the unit circle versus an increasing channel length for a random impulse response, a simulated acoustic impulse response and a measured acoustic impulse response. Figure 5.16b shows the corresponding variance of the phases of the zeros for the same channels. It is observed that for systems responses with lengths $L > 200$ the zeros indeed do cluster around the unit circle and that their normalized phases θ/π (i.e. varying in the interval $[-1, 1]$) tend to a uniform distribution for which the variance is $\sigma^2 = 1/3$ [128]. Thus, it can be expected that the blind system identification algorithm will be affected by near-common zeros for long systems.

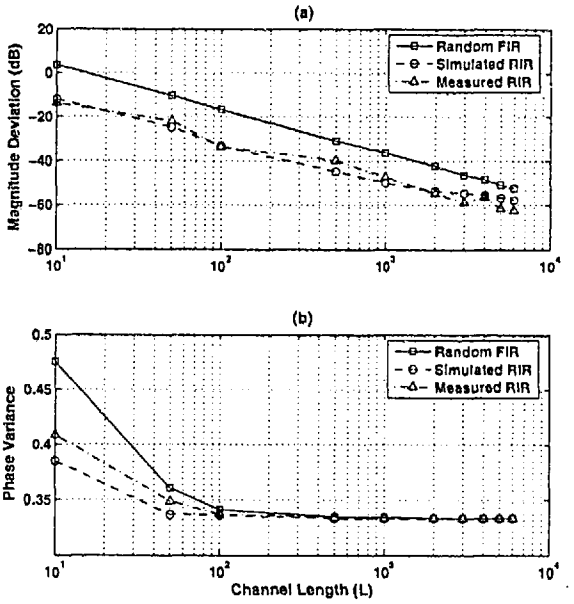

Figure 5.16: Example of the zeros of channels showing (a) that they cluster near the unit circle and (b) with uniformly distributed phases.

In summary, a new approach was developed for finding the common roots components in polynomials using adaptive filters in a signal processing framework. The most attractive feature of this method lies in its ability to identify the components due to the common roots of $M \geq 2$ polynomials without having to factor these. It was demonstrated by simulations that this approach can accurately detect the number of common roots and the impulse response of these components. From the simulation results two interesting observations were brought forward. First, the accuracy of blind channel algorithms is strongly affected by the distance of zeros and that two or more inter-channel zeros do not have to be exactly equal to be considered common in adaptive blind channel identification. Second, this same feature makes the common root detection algorithm robust to perturbations that result in small separation of the roots that should otherwise be considered common. Finally, it was demonstrated that, for large order systems, near common zeros can be expected to always exist and thus, long channels will be difficult to estimate with the MCLMS algorithm in its present state.

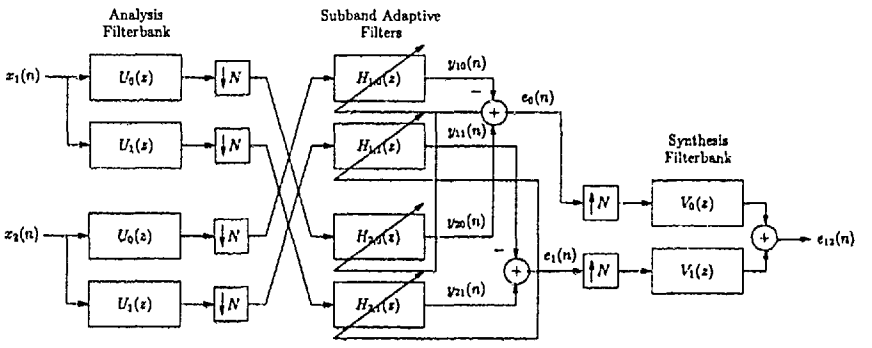

Figure 5.17: Two channel, two subband, blind system identification.

5.6 Adaptive Blind System Identification in Decimated Subbands

In blind acoustic system identification, the impulse responses involved are generally very long. From the simulation result in Section 5.5.5 it is evident that, for such systems, common or near-common zeros are likely to exist, which will limit the ability of the MCLMS algorithm to identify these systems and, consequently, its practical applicability. Therefore, it is necessary to partition the problem such that the length of the impulse response is reduced.

One obvious approach to investigate is the processing in decimated subbands [114], which has been applied successfully in, for example, the related field of acoustic echo cancellation [115,129], where it is shown to significantly improve the convergence behaviour of the adaptive filters involved, especially when the input is non-white [115]. The remainder of this section will present a method of adaptive blind system identification in subbands. First, the validity of the cross-relation error in decimated subbands is analysed and it is shown that this holds under some conditions on the filterbank, which can be satisfied approximately using oversampled filterbank structures. However, the scaling factor inherent in the adaptive algorithm is problematic since it is not uniform across the subbands. Therefore, this needs to be solved before the subband approach can find practical applications. Nevertheless, preliminary simulation results assuming knowledge of one true filter tap in each subband are presented, indicating that improvement can be achieved, which motivates further studies on the topic.

5.6.1 The Subband Cross-Relation

The cross-relation error in subbands is now investigated. The subband filter impulse responses are derived and the necessary conditions on the filterbank are formulated. Consider the system diagram shown in Fig. 5.17. This is the extension of Fig. 5.2 to the subband case, where each subband is processed separately. The output of the mth channel filter, $H_{m,i}(z)$, in the ith subband is given by

$$Y_{ml,i}(z) = \frac{1}{N}\sum_{k=0}^{N-1} U_i(z^{1/N}W_N^k)X_l(z^{1/N}W_N^k)H_{m,i}(z) \tag{5.54}$$

It then follows that the error signal arising from the cross-relation between the mth and the lth channels, given in (5.12), can be written for the ith subband as

$$E_{ml,i}(z) = \frac{1}{N}\sum_{k=0}^{N-1} U_i(z^{1/N}W_N^k)X_l(z^{1/N}W_N^k)H_{m,i}(z) - U_i(z^{1/N}W_N^k)X_m(z^{1/N}W_N^k)H_{l,i}(z), \tag{5.55}$$

where $W_N = e^{j\frac{2\pi}{N}}$. Subsequently, the total fullband output error of the system is

$$\begin{aligned} E_{ml}(z) &= \sum_{i=0}^{K-1} V_i(z)E_{ml,i}(z) \qquad (5.56)\\ &= \frac{1}{N}\sum_{i=0}^{K-1} U_i(zW_N^k)V_i(z)S(zW_N^k)\sum_{k=0}^{N-1}\left(H_l(zW_N^k)H_{m,i}(z^N) - H_m(zW_N^k)H_{l,i}(z^N)\right). \end{aligned}$$

It can now be seen from (5.56) that the aliasing components do not allow for a single set of filters, $H_{m,i}(z^N)$, that minimize the total output error to exist in the structure of Figure 5.17. However, if a restriction is imposed on the subband filters such that aliasing effects can be considered negligible, i.e.

$$U_i(zW_N^k)V_i(z) \approx 0, \quad i = 0, 1, \ldots, K-1,\ k = 1, 2, \ldots, N-1, \tag{5.57}$$

then the expression for the error in (5.56) reduces to

$$E_{ml}(z) = \frac{1}{N}S(z)\sum_{i=0}^{K-1}\left(U_i(z)H_{m,i}(z^N)H_l(z) - U_i(z)H_{l,i}(z^N)H_m(z)\right)V_i(z). \tag{5.58}$$

It is recognized that this expression is of similar form to that of the single channel non-blind case in [129]. In particular, if the following relation is defined between the fullband

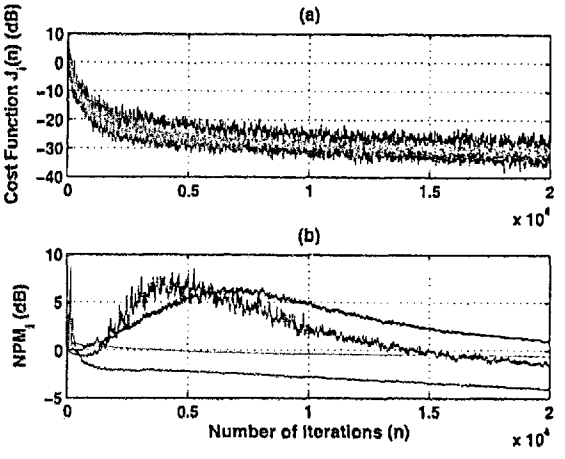

Figure 5.18: Adaptive blind system identification in subbands. Convergence behaviour for: (a) Cost function trajectory and (b) NPM.

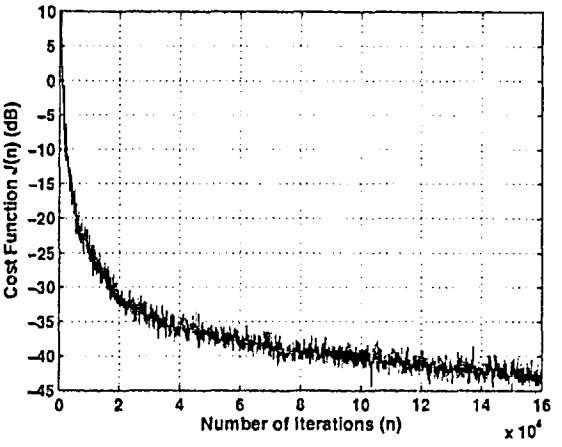

Figure 5.19: Fullband mean squared error from subband adaptive filters.

and the subband filters

$$U_i(z)H_{m,i}(z^N) = U_i(z)H_m(z), \quad m = 1, 2, \ldots, M, \; i = 0, 1, \ldots, K-1 \tag{5.59}$$

and then substituting (5.59) into (5.57), the error signal becomes

$$E_{ml}(z) = \frac{1}{N} S(z) \left(H_m(z)H_l(z) - H_l(z)H_m(z) \right) \sum_{i=0}^{K-1} U_i(z)V_i(z). \tag{5.60}$$

Thus, it can be seen that the cross-relation between channels will hold in the subband case irrespective of the reconstruction properties of the filterbank. This shows that there exist a set of subband filters, which, if accurately identified, result in a correct cross-relation error. This is important since it implies that the complex MCLMS algorithm can be applied in each subband for blind identification of the subband filters. Furthermore, it was shown in [129] that the length of these subband filters is of the order L/N, which can significantly reduce the length of the adaptive filters when N is large.

5.6.2 Simulation Results: subband blind system identification

This section will present simulation results in order to demonstrate two distinct convergence behaviours in the subband adaptive filters that have been observed in a multitude

of simulations. Moreover, the validity of the cross-relation error correspondence between the fullband output and the subband filters is confirmed. A generalized DFT (GDFT) filterbank [115] (which will be discussed in more detail in Chapter 6) with eight subbands and a decimation factor of six is used for these experiments. This results in four subband adaptive filters as only the first half of the subbands need processing in the framework of the GDFT filterbank [115] since the remaining half are the corresponding complex conjugates. Furthermore, a method to calculate the equivalent subband filters given a fullband impulse response [129] is discussed in Chapter 6, where the decomposition is applied to subband inverse filter design. The theoretical subband filters calculated with this method are used here as reference in the evaluation. An arbitrary system with $M = 5$ random channels of length $L = 128$ taps is excited with white Gaussian noise.

Figure 5.18a shows the cost function trajectories of all subband adaptive filters and the corresponding misalignment in terms of NPM is shown Fig. 5.18b. Moreover, the fullband mean squared error resulting from the subband adaptive filters is depicted in Fig. 5.19. The following observations can be made from this result: (i) the error is minimized in each subband, (ii) this also results in a minimized global error and (iii) the resulting filters do not appear to match those calculated with subband decomposition method. Before commenting on these results, a second simulation example will be given.

In this experiment, it is assumed that one tap in each subband filter is known. The simulation conditions are otherwise identical to those of the previous experiment. The known tap value is enforced by substituting it into the estimated channel vectors at every iteration as in [69,120]. One consequence of this is that the phase and magnitude ambiguity in the subband estimates are alleviated. The outcome of this experiment is shown in Figs. 5.20a and 5.20b for the cost function trajectory and the misalignment respectively. The fullband error is shown in Fig. 5.21. Also in this case the error is minimized, however, in contrast to the previous experiment, the subband filters converge towards the theoretical filters from subband decomposition. One speculative explanation to this discrepancy is that the filters in the subbands from (5.59) are not unique as discussed in [129] and many filters can be found which satisfy the relation between the subband and the fullband filters. When the the adaptation is constrained with one-tap substitution, it is thus forced to one particular solution. Therefore, the estimates in the first experiment are not necessarily wrong systems and it is possible that both cases would result in the same fullband system. However, this is not easy to evaluate due to the complex scalar problem.

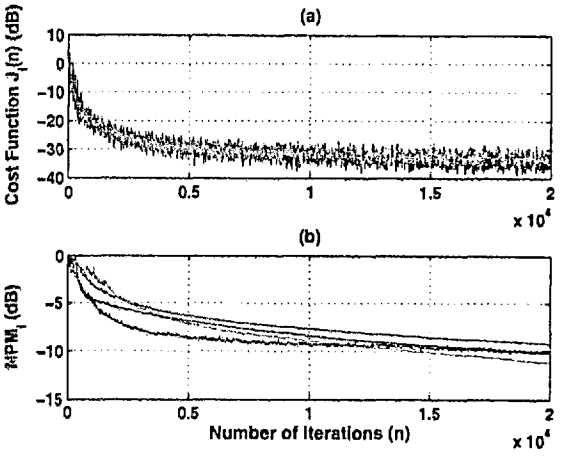

Figure 5.20: Adaptive blind system identification in subbands with one-tap substitution. Convergence behaviour in terms of: (a) Cost function trajectory and (b) NPM.

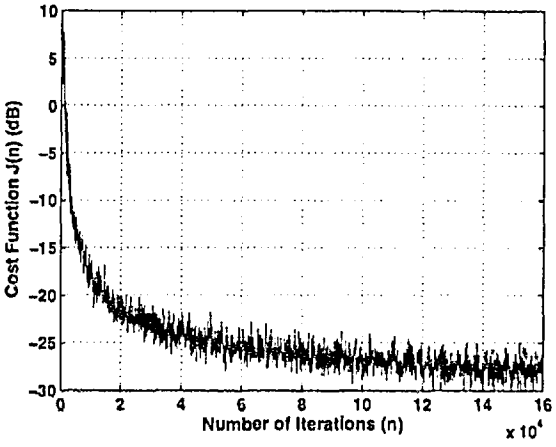

Figure 5.21: Fullband mean squared error from subband adaptive filters with one-tap substitution.

In summary, it has been shown that the cross-relation error holds for the subband case if aliasing is suppressed in the subbands. The relation between the error minimization and the estimated channels is difficult to evaluate, due to the complex scale factor. Nevertheless, the initial steps towards subband adaptive blind system identification have been made. Crucial steps for further work are study of the solutions of the subband filters, and solutions to the scale ambiguity.

5.7 Application to Acoustic Impulse Responses

The experiments in this chapter have been performed on random impulse responses and white Gaussian noise as input. In this section, two example simulation results are presented: (i) an acoustic system excited with white Gaussian noise and (ii) an acoustic system excited with real speech. The acoustic system comprises a linear microphone array with $M = 5$ microphones separated by 0.1 m and a source positioned at 1.5 m from the array. The room was of dimensions $4 \times 4 \times 3$ m and reverberation time $T_{60} = 0.3$ s. The impulse responses were truncated to different lengths in the range $L = 2^k$, $k = 7, 8, 9, 10$, i.e. ranging from 128 to 1024 taps. The sampling frequency was set to $f_s = 8$ kHz. The impulse response at one of the microphones for $L = 1024$ is shown in Fig. 5.22. The MCLMS algorithm was executed using the optimal step-size and in a noise-free case.

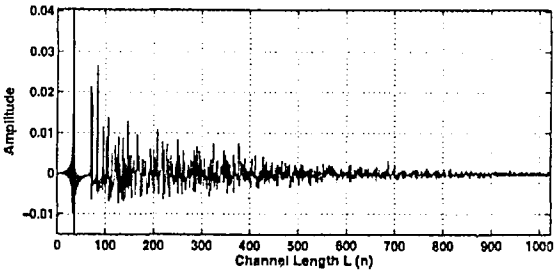

Figure 5.22: Room impulse response for the experiments in Section 5.7.

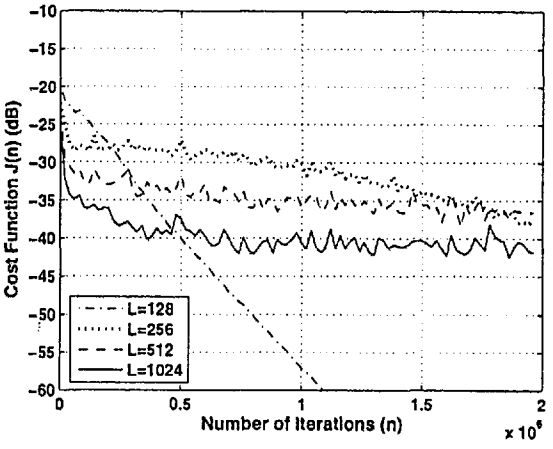

Figure 5.23: Cost function trajectory for varying channel length using white Gaussian noise input.

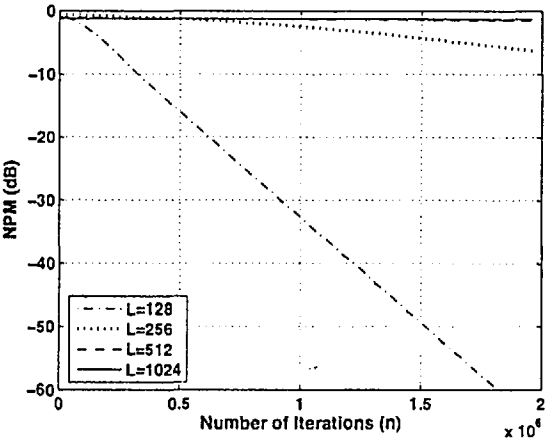

Figure 5.24: Normalized projection misalignment for for varying channel length using white Gaussian noise input.

Figure 5.23 shows the cost function trajectory for the MCLMS algorithm with white Gaussian noise input and the corresponding result in terms of NPM is shown in Fig. 5.24. It is seen that the convergence is adversely affected by the increase in channel length. Only at $L = 128$ taps does the algorithm converge. In Figs. 5.25 and 5.26, the results for the experiment performed with the speech input are shown. Similar outcome is apparent in this case, although the convergence rate is reduced considerably. These results are in agreement with the discussion in Section 5.5.5, and near common zeros is the likely cause to the convergence degradation.

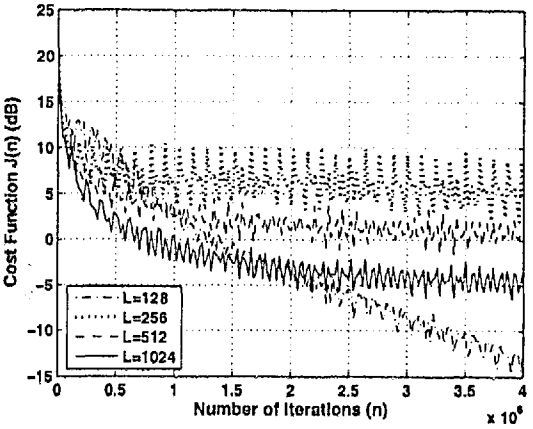

Figure 5.25: Cost function trajectory for varying channel length using speech input.

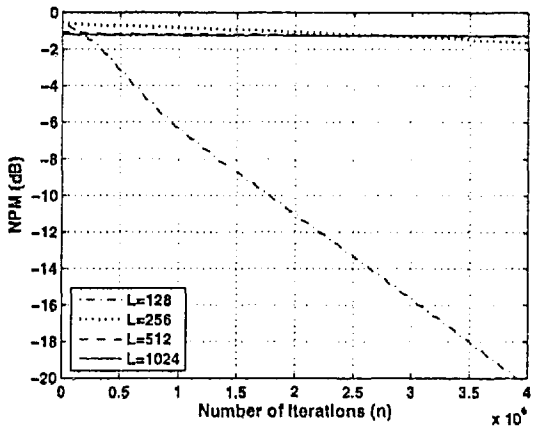

Figure 5.26: Normalized projection misalignment for for varying channel length using speech input.

5.8 Summary

Adaptive blind system identification with application to acoustic systems has been discussed. A complex multichannel least mean squares (MCLMS) algorithm was derived and it was shown that it can identify correctly an unknown multichannel system up to an arbitrary complex factor. The effects of this factor were discussed for both real and complex signals. Two problems in blind system identification were highlighted, noise and common zeros. The first of these problems was addressed with the derivation and implementation of an optimal step-size for the adaptive blind system identification algorithm. It was demonstrated that this step-size can significantly improve the performance in noisy and in noise-free conditions. A common zeros detection and identification algorithm was developed and employed to investigate of the effects of common zeros on adaptive blind system identification algorithms. It was shown that zeros between two or more channels do not have to be exactly identical in order to degrade the performance of the adaptive MCLMS algorithm. It was further demonstrated that near-common zeros are likely occur in acoustic systems. Consequently, a preliminary study on the use of subband processing was presented as a potential strategy to reduce the adaptive filter length and to improve the sensitivity to common zeros. It was then shown there exists a set of subband filters which satisfy the inter-channel cross relation and that minimize the overall error formed by this cross-relation. However, gain and phase ambiguities inherent in blind system iden-

tification are problematic in evaluation and practical applicability of this method. Finally, the application to acoustic impulse responses was demonstrated for various channel lengths and it could be concluded that with the MCLMS in its present state, is limited short impulse responses ($L < 200$). It is essential for future work on this topic to emphasize the partitioning of the channels or other methods that reduce the occurrence of common and near-common zeros.

Chapter 6

Speech Dereverberation with Approximate Room Impulse Response Estimates

IN practical scenarios, only approximate room impulse response estimates can be obtained due to measurement noise and source location fluctuations. Moreover, room impulse responses are generally non-minimum phase and involve filters with length of several thousand taps. The non-minimum phase property makes single channel inversion difficult and only approximate inversion can be achieved. This problem is alleviated when multiple observations are available and exact inversion can be achieved.

This chapter presents a study of inverse filter design using approximate impulse response estimates with application to speech dereverberation. It is demonstrated that exact filtering, although achievable, may not be desirable when the room impulse response estimates are inaccurate and some form of trade off is necessary. Consequently, a multichannel method utilizing decimated oversampled subbands is proposed for equalization of long channels, where the fullband room impulse response is decomposed into its subband equivalents before inversion. This method is shown to provide substantial computational reductions in terms of floating point operations at the cost of small magnitude distortion due to the use of non-ideal subband filters. Moreover, it is demonstrated that the proposed technique is more robust to inaccuracies in the impulse responses.

6.1 Room Impulse Response Inversion

Consider an estimate, $\hat{\boldsymbol{h}}_m$, of the room impulse responses $\boldsymbol{h}_m$ using either direct (non-blind) measurement [13] or blind system identification as discussed in Chapter 5. Dereverberation can be achieved, in principle, by an inverse system with response $G_m(z)$ satisfying the relation

$$G_m(z)H_m(z) = \kappa z^{-\tau}, \quad m = 1, 2, \ldots, M \tag{6.1}$$

where τ and κ are an arbitrary delay and and scale factor respectively. Equivalently, this can be written in the time domain

$$\boldsymbol{h}_m^T \boldsymbol{g}_m = \kappa\delta(n - \tau), \tag{6.2}$$

where $\boldsymbol{g}_m = [g_{m,0}\ g_{m,1} \cdots g_{m,L_\mathrm{i}-1}]^T$ is the L_i-tap impulse response of the inverse filter and $\delta(n)$ is an impulse. Thus, the problem of inverse filtering is to find the filter $G_m(z)$.

In the case of a minimum phase system, a stable inverse filter can be found by simply replacing the zeros of the filter $H_m(z)$ with poles [130], i.e.

$$G_m(z) = \frac{1}{H_m(z)}. \tag{6.3}$$

However, in practice, inversion of the acoustic impulse responses is not straightforward due to the following issues:

(i) Most practical room impulse responses possess non-minimum phase characteristics [27] and (6.3) is not applicable.

(ii) The average difference between maxima and minima in the room transfer function are in excess of 10 dB [26, 83] and thus, the room transfer function may contain spectral nulls that, after inversion, give strong peaks in the spectrum causing narrow band noise amplification.

(iii) Designing inverse filters with inaccurate impulse response estimates will cause distortion in the processed signal [83].

(iv) The room impulse response can be several thousand taps in length [13].

Several alternative approaches to acoustic channel inversion have been proposed in the literature and will be reviewed in the remainder of this section. The rest of this chapter is organized as follows. In the remainder of this Section, a review of the existing single and multichannel methods of acoustic inverse filter design methods is presented and their pros and cons are discussed. In Section 6.2 a multichannel approach using oversampled, decimated subbands is derived and the computational savings of this technique are analyzed. Simulation results are provided to demonstrate the performance of the method in Section 6.3. Section 6.4 demonstrates the application of inverse filtering to speech dereverberation and, finally, a concluding summary is provided in Section 6.5.

6.1.1 Single Channel Inversion: approximate equalization

Two main techniques for single channel ($M = 1$) inverse filtering can be found in the literature. These are homomorphic inversion [12, 75, 77] and least squares (LS) inversion [75, 130]. Since the room impulse responses are generally non-minimum phase, the inverse filters are of infinite length and non-causal, and only approximate equalization can be achieved with these single channel methods [80].

In homomorphic inverse filtering, the impulse response is decomposed in the cepstral domain, into a minimum phase component, $H_{\text{mp},m}(z)$, and an all-pass component, $H_{\text{ap},m}(z)$, such that [12]

$$H_m(z) = H_{\text{mp},m}(z)H_{\text{ap},m}(z). \tag{6.4}$$

An exact and stable inverse filter can be found for the minimum phase component according to (6.3). The all-pass component has unit magnitude, $|H_{\text{ap},m}(z)| = 1$, and can be equalized e.g. using matched filtering [77]. This is equivalent to a two stage approach where magnitude and phase are equalized separately. Since the phase equalization results in very long inverse filters, studies were performed on the applicability of magnitude equalization only [27, 77]. The findings show, by subjective and objective experiments, that, magnitude equalization results in strong residual echoes in the processed speech signal and is thus insufficient in the context of speech dereverberation. Homomorphic inversion suffers from the difficulty in the separation of the minimum phase and all-pass components, which was improved with an iterative algorithm proposed by Radlović and Kennedy [77].

An alternative approach is to formulate an optimization problem for which the least squares solution results in an optimal estimate of the inverse filter [75, 130], i.e.

$$\hat{G}_m(z) = \min_{G_m(z)} \|G_m(z)H_m(z) - z^{-\tau}\|^2, \tag{6.5}$$

where $\hat{G}_m(z)$ is the optimal, in the least squares sense, FIR estimate of the true inverse system, $G_m(z)$. In the time domain (6.5) is given by

$$\hat{g}_m = \min_{g_m} \|h_m^T g_m - \delta(n-\tau)\|^2, \tag{6.6}$$

where $\hat{g}_m = [\hat{g}_{m,0}\ \hat{g}_{m,1} \cdots \hat{g}_{m,L_i-1}]^T$ is the time domain equivalent of $\hat{G}_m(z)$. The delay, τ, is introduced to make the inverse system causal. An optimal delay can be found with, for example, an iterative search. However, it has been shown that the inverse filter design is not very sensitive to the exact choice of delay [75] and often it is taken as $\tau = L/2$, which is the approach adopted here.

In a comparative study between these two techniques, Mourjopoulos [75] concluded that the LS method, although sometimes less accurate than the homomorphic counterpart, is more efficient to use in practice. However, there are two main drawbacks of these single channel methods: a long processing delay and extremely long inverse filters (ideally of infinite length). The processing delay in particular can be problematic in many communications applications. On the other hand, the advantage of these types of inverse filters is that they are generally less sensitive to inexact and/or undermodelled room impulse response estimates which are common in practice. This is due to the approximate nature of the filters. An inherent feature in the LS FIR inverse filters is to only partially equalize deep nulls (depending on the length of the inverse filter), which can be advantageous to avoid the problem in point (ii) on page 134.

6.1.2 Multichannel Inversion: exact equalization

In the multichannel case, exact equalization of non-minimum phase impulse responses can be achieved by simultaneously inverting all $M \geq 2$ impulse responses as shown in Fig. 6.1. Under the condition that there are no common zeros between the channel transfer functions, the following relation between the transfer functions and their inverses can be

written [68,80]

$$\sum_{m=1}^{M} G_m(z)H_m(z) = 1. \tag{6.7}$$

This is stems from the Bezout's theorem [68], which states that for a set of polynomials with a greatest common divisor one, there exists another set of polynomials such that (6.7) is satisfied. The relation in (6.7) can be written in the time domain as

$$\sum_{m=1}^{M} \boldsymbol{H}_m \boldsymbol{g}_m = \boldsymbol{H}\boldsymbol{g} = \boldsymbol{u}, \tag{6.8}$$

where

$$\boldsymbol{H}_m = \begin{bmatrix} h_{m,0} & 0 & \dots & 0 \\ h_{m,1} & h_{m,0} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h_{m,L-1} & \dots & \vdots & 0 \\ 0 & h_{m,L-1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & h_{m,L-1} \end{bmatrix}, \tag{6.9}$$

is the mth channel convolution matrix, $\boldsymbol{H} = [\boldsymbol{H}_1\ \boldsymbol{H}_2\ \dots\ \boldsymbol{H}_M]^T$, $\hat{\boldsymbol{g}} = [\hat{\boldsymbol{g}}_1^T\ \hat{\boldsymbol{g}}_2^T\ \dots\ \hat{\boldsymbol{g}}_M^T]^T$ is a vector of the concatenated inverse filter impulse responses, and $\boldsymbol{u} = [1\ 0\ \dots\ 0]^T$ is a vector with the desired output. An optimization problem can then be formulated from (6.8) as

$$\hat{\boldsymbol{g}} = \min_{\boldsymbol{g}} \|\boldsymbol{H}\boldsymbol{g} - \boldsymbol{u}\|^2, \tag{6.10}$$

and the multichannel least squares equalization filters can be calculated [68]

$$\hat{\boldsymbol{g}} = \boldsymbol{H}^{+}\boldsymbol{u}, \tag{6.11}$$

where $\boldsymbol{H}^{+} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T$ is the matrix pseudo-inverse [96]. The dimensions of the matrix $\boldsymbol{H}_m$ are $(L + L_i - 1) \times L_i$ and those of the vector $\boldsymbol{u}$ are $(L + L_i - 1) \times 1$. These dimensions, governed by the length of the RIR, L, and the length of the inverse filter, L_i, indicate the length of the resulting inverse filters, which is bounded by [68]

$$L_i \leq \frac{L-1}{M-1}. \tag{6.12}$$

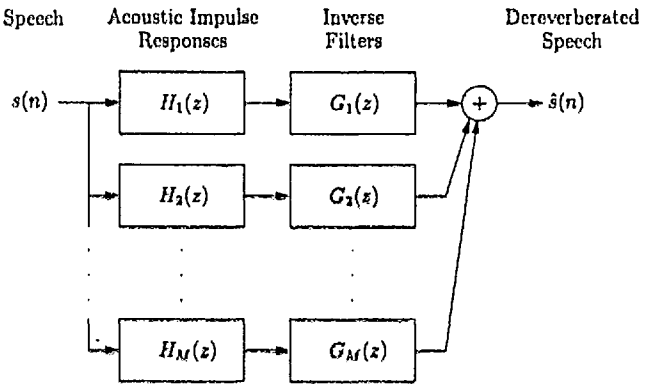

Figure 6.1: Multichannel equalization system.

In the special case when the length is chosen such that the relation in (6.12) is an equality, the matrix $\boldsymbol{H}$ becomes square and the pseudo-inverse in (6.11) reduces to a standard matrix inverse. The solution is then equivalent to that of the the MINT method proposed by Miyoshi and Kaneda [80]. However, as pointed out in [68], it may not always be possible to choose this optimal length when more than two channels are considered, since the relation in (6.12) may not give an integer result. Therefore, a larger length is often chosen in practice. For the subsequent work in this chapter L_{i} is taken as

$$L_{\mathrm{i}} = \left\lceil \frac{L-1}{M-1} \right\rceil . \tag{6.13}$$

Modified versions of the multichannel LS inverse filters include a decimated subband implementation of the MINT algorithm [81] and an implementation in an adaptive framework [82].

Under ideal conditions of exact knowledge of the impulse responses and no common zeros between the corresponding transfer functions, perfect equalization with no delay is possible. However, in many practical situations, exact inversion can be problematic [83] due to, for example, undermodelled estimates of $\boldsymbol{h}_m$ or when the channel estimates contain even moderate estimation errors. The latter point is illustrated in the following simulation example.

An arbitrary system with two random channels $\boldsymbol{h}_m$ of length $L = 16$ was used. The true impulse responses were corrupted with additive noise to model estimation errors ranging from 0 to -60 dB of normalized projection misalignment. For each case, the impulse

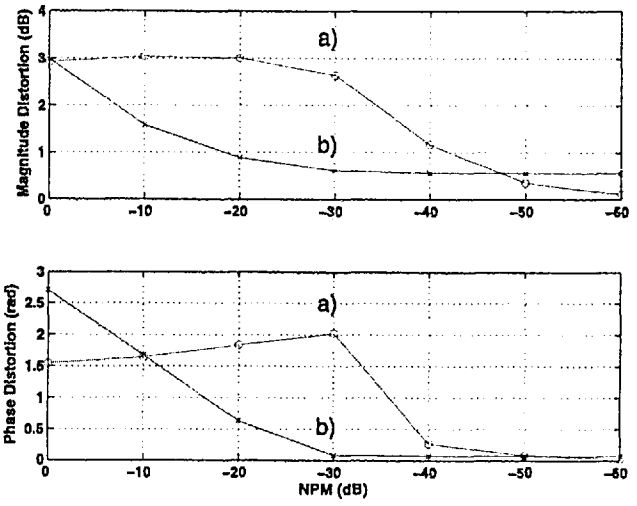

Figure 6.2: Magnitude and phase distortions vs. NPM for a) exact multichannel inverse filtering with MINT and b) approximate single channel LS inverse filtering.

response was equalized using two channel MINT and single channel LS inverse methods. The magnitude and phase distortions were evaluated using the measures (6.37) and (6.38), defined in Section 6.3. The results are displayed in Fig. 6.2 for (a) exact inverse filtering with MINT and (b) approximate LS filtering. It can be seen that equalization using MINT inversion introduces significant spectral distortion for NPM levels greater than -40 dB, which is unlikely to be achieved by current blind channel estimation techniques as demonstrated in Chapter 5. As an alternative, single channel LS inverse filter design is more robust to channel estimation errors as shown in Fig. 6.2, although equalizers with very high order are typically required as in this example case where $L_i = 10L$. This result is also in accordance with the results reported in [83] and [41] where the authors studied equalization of room impulse responses measured at a different location to that at the point of processing. A study of the effect of delay constraints in the context of acoustic channel inversion for dereverberation was presented in [131]. Its was shown that, for exact inversion, observation noise can be amplified (as in (ii) above) whereas LS solutions generally introduce significant delay and a trade-off between the two is necessary. A summary of the pros and cons of the the approximate and exact inverses can be seen in Table 6.1.

Table 6.1: Exact vs. approximate acoustic equalization filters.

	Pros	Cons
Exact inverse	+ Only require filters of orders similar to room impulse responses. + No processing delay.	- Undermodelled responses cause distortions. - Deep nulls in room responses may amplify noise when inverted. - Inexact impulse response estimates give useless inverses.
Approximate inverse	+ Less sensitive to inexact room impulse response estimates.	- Very long inverse filters. - Large processing delay to accommodate non-minimum-phase inverse.

Several approaches have been reported in the literature, addressing various of the above mentioned aspects of room impulse response equalization. Bharitkar et al [88] proposed to perform spatial averaging on the measured impulse responses and to design an inverse filter based on these spatially averaged impulse responses, such that errors due to impulse response fluctuations are reduced. In [131] the authors proposed to modify the desired signal in the LS inverse filter design, such that the late reverberation is equalized while the early reflections are preserved, which was shown to reduce the sensitivity to noise. Haneda et al [19, 132] proposed to decompose the room transfer functions into common acoustical poles and the non-common zeros FIR filters. This would reduce the sensitivity to fluctuations and also the length of the equalization filters. Mourjopoulos [76] suggested the use of an AR model of the room transfer functions rather than the all-zero model in order to reduce the filter order. Furthermore, he introduced the idea of vector quantization using a library of measured impulse responses for equalization in multiple locations in a room. The AR model of room transfer functions was further exploited by Hopgood and Rayner in a single channel subband equalization approach [20]. A subband version of the MINT method was also studied experimentally in [81]. Finally, Hikichi et al [133, 134] proposed adding a regularization term in the multichannel inversion method, which was demonstrated to add robustness to noise and room impulse response fluctuations.

6.2 Multichannel Subband Equalization

From Section 6.1, it was seen that exact inverse filters, although possible, may not be desirable. It is evident that approximate inverse filtering is preferable in realistic environments where the channel estimates are not exact and the microphone observations are corrupted by noise. Furthermore, reduction of the filter length is an important issue to consider. Long impulse responses, noise, and inaccuracies in the room impulse response estimates are interrelated as was shown in Chapter 5. Single channel LS filtering provides a means for approximate filtering, however, it requires very long filters and results in large processing delays, often inappropriate for speech applications.

In this section, a new multichannel inverse filtering scheme is presented. This method first finds the equivalent subband filters, given a fullband room impulse response, and then performs multichannel least squares equalization in each subband. It is shown that this approach significantly reduces the computational burden of the equalization process and also, simulations demonstrate that it is less sensitive to inaccuracies in the room impulse responses compared to the fullband multichannel LS equalizer.

6.2.1 Multichannel LS Inversion in Oversampled Subbands

A subband version of MINT was first investigated in [81], however, this was an experimental approach where the subband filters were first estimated using a reference signal. The relation between the fullband and the subband impulse responses was not considered and finally, the results showed unsuccessful speech dereverberation using the proposed subband method. In [20] a more rigorous approach was taken and the relation between fullband and subband filters was studied for an AR model of the room impulse response. A non-blind adaptive method for multichannel equalization in oversampled subbands was proposed in [135] and was shown to provide significant improvement over the fullband counterpart. Oversampled subbands [114] have been used successfully in reducing the complexity of acoustic signal processing problems such as, for example, acoustic echo cancellation, where significant improvements have been demonstrated in the convergence of the subband adaptive filters [115,129,135]. This motivates the study of multichannel least squares inversion in subbands.

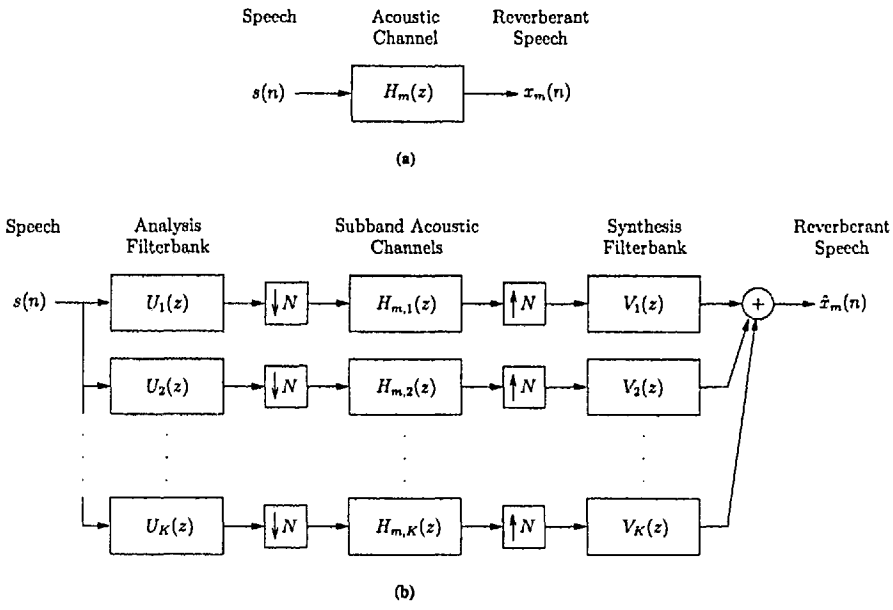

Figure 6.3: System diagrams for (a) a fullband filtering process and (b) a subband filtering process.

Subband decomposition

Consider the two systems depicted in Fig. 6.3a and 6.3b, showing a fullband filtering process of the signal $s(n)$ for the mth observation, $x_m(n)$ and a subband filtering process of the same input signal resulting in an observation $\hat{x}_m(n)$. The objective of the subband decomposition is to find a set of subband filters, $H_{m,i}(z)$, $i = 0, 1, \ldots, K$, given the fullband filter $H_m(z)$, such that, when the filtering operation is performed in the subbands, the reconstructed fullband signal, $\hat{x}_m(n)$, is equivalent to the output of the fullband filter up to an arbitrary scale factor, κ, and an arbitrary delay, τ, i.e.

$$x_m(n) = \kappa \hat{x}_m(n - \tau). \tag{6.14}$$

This type of decomposition has found applications in, for example, filtering of compressed MPEG audio signals where the subband signals can be modified without having to reconstruct them first [136]. In [136], the relations between the fullband and the subband filters were derived for a critically decimated cosine modulated filter bank [114]. It is shown that it is necessary to apply, in addition to the subband filters, cross filters

between all subbands. The authors demonstrated, however, that it is sufficient to consider only the cross-filters due to frequency bands adjacent to the subband under consideration. Reilly et al [129] studied the subband decomposition for complex oversampled filterbanks in the context of acoustic echo cancellation and showed that a good approximation can be obtained with a diagonal filtering matrix, i.e. only one filter for the subband under consideration, under the assumption that the oversampled filterbanks manage to suppress aliasing due to non-ideal subband filters. This approach will be adopted here for the multichannel decomposition.

The generalized discrete Fourier transform (GDFT) filterbank [115] is employed in the development of the subband decomposition. The advantages of this filterbank include straightforward implementation of fractional oversampling and also computationally efficient implementations [115]. Due to the oversampling, sufficient aliasing suppression can be achieved in the subbands. Within the framework of the GDFT filterbank, the analysis filters, $u_i(n)$, are calculated from a single prototype filter, $u_{\rm pr}(n)$, with bandwidth $\frac{2\pi}{K}$ according to the relation [115]

$$u_i(n) = u_{\rm pr}(n) e^{j\frac{2\pi}{K}(i+i_0)(n+n_0)}, \tag{6.15}$$

where the properties of the frequency and time offset terms, i_0 and n_0, are discussed in, for example, [115]. For the work presented here these values are set to $n_0 = 0$ and $i_0 = 1/2$ as in [129]. It can be shown [115] that a corresponding set of synthesis filters satisfying near perfect reconstruction can be obtained from the time-reversed, conjugated version of the analysis filters

$$v_i(n) = u_i^*(L_{\rm pr} - n - 1), \tag{6.16}$$

where $L_{\rm pr}$ is the length of the prototype filter and, consequently, the length of all the filterbank analysis and synthesis filters. Although, this filter design results in complex subband signals, for an even K, only $K/2$ subbands need to be processed since the remaining subbands are complex conjugates of these. The filterbank in this chapter uses $K = 32$ subbands as an example, with prototype filter length set to $L_{\rm pr} = 512$ taps and a decimation factor, $N = 24$. The prototype filter was designed using the iterative least squares method proposed by Weiss and Stewart [115], giving an estimated aliasing suppression of -82 dB. The magnitude response of the analysis filters is shown in Fig. 6.4.

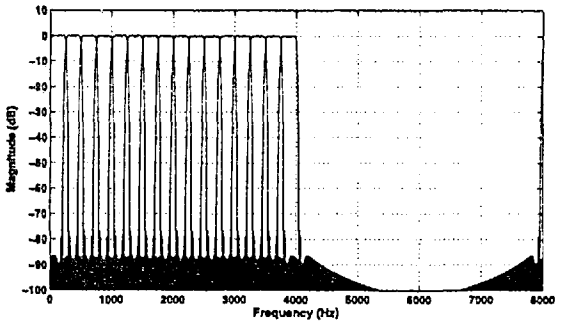

Figure 6.4: Subband filters of length $L_{\mathrm{pr}} = 512$ taps for $K = 32$ subbands, decimated by $N = 24$.

The subband decomposition proposed in [129] is now reviewed. Consider the output of the system in Fig. 6.3b, which can be written

$$\hat{X}_m(z) = \frac{1}{N} \sum_{i=0}^{K/2-1} \sum_{k=0}^{N-1} S(zW_N^k) U_i(zW_N^k) H_{m,i}(z^N) V_i(z), \tag{6.17}$$

where $W_N = e^{j\frac{2\pi}{N}}$. Under the assumption of an ideal prototype filter, the aliasing components are suppressed such that

$$U_i(zW_N^k)V_i(z) = 0, \ k > 0, \ \forall i \tag{6.18}$$

and the expression in (6.17) reduces to

$$\hat{X}_m(z) = S(z)\frac{1}{N} \sum_{i=0}^{K/2-1} U_i(z) H_{m,i}(z^N) V_i(z). \tag{6.19}$$

In practice, the condition in (6.18) holds only approximately depending on the prototype filter design.

Subsequently, if the filters in each subband $H_{m,i}(z^N)$ are chosen to satisfy the relation

$$U_i(z)H_{m,i}(z^N) = U_i(z)H_m(z), \quad i = 0, 1, \ldots, \frac{K}{2} - 1, \tag{6.20}$$

and after substitution of (6.20) into (6.19), the output of the filterbank becomes

$$\hat{X}_m(z) = S(z)H_m(z)\frac{1}{N}\sum_{i=0}^{K/2-1} U_i(z)V_i(z). \tag{6.21}$$

The analysis and synthesis filters in the GDFT filterbank employed here [115] are designed such that near perfect reconstruction is obtained at the output, i.e.

$$\Re\left\{\sum_{i=0}^{K/2-1} U_i(z)V_i(z)\right\} \approx \kappa z^{-\tau}. \tag{6.22}$$

If aliasing is suppressed in the subbands the condition in (6.22) can be satisfied by making the analysis and synthesis filters power complimentary [115]. Consequently, the expression in (6.21) reduces to

$$\Re\left\{\hat{X}_m(z)\right\} \approx S(z)H_m(z)\frac{\kappa}{N}z^{-\tau} = X_m(z)\frac{\kappa}{N}z^{-\tau}, \tag{6.23}$$

which is the desired outcome stated in (6.14), where the output of the filterbank is approximately equal to the output of the fullband filter up to an arbitrary delay and an arbitrary scale factor. The problem is now to solve for $H_{m,i}(z)$ in (6.20). The downsampled version of (6.20) is

$$\frac{1}{N}\sum_{i=0}^{N-1} U_i(z^{1/N}W_K^i)H_{m,i}(z) \approx \frac{1}{N}\sum_{i=0}^{N-1} U_i(z^{1/N}W_N^i)H_i(z^{1/N}W_N^i). \tag{6.24}$$

It can be seen that the subband equivalent filters can be obtained by feeding the fullband filter's impulse response through the analysis filterbank and the decimators followed by deconvolution of the decimated analysis filters. The accuracy of the approximation is governed by the aliasing in the subbands.

It was shown in [129], that the subband filters in Fig. 6.3b can be estimated by solving the following optimization problem

$$\hat{h}_{m,i} = \arg\min_{h_{m,i}} \|U_i h_{m,i} - p_i\|^2, \tag{6.25}$$

where

$$U_i = \begin{bmatrix} u_{i,0} & 0 & \dots & 0 \\ u_{i,N} & u_{i,0} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u_{i,L_{\text{pr}}-1} & \dots & \vdots & 0 \\ 0 & u_{i,L_{\text{pr}}-1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & u_{i,L_{\text{pr}}-1} \end{bmatrix} \tag{6.26}$$

is the convolution matrix resulting from the decimated analysis filters, $h_{m,i} = [h_{m,0}^{(i)}\ h_{m,1}^{(i)}\ \dots\ h_{m,\tilde{L}-1}^{(i)}]^T$ is the impulse response of the ith subband and the mth channel and $\boldsymbol{p}_i = [p_i(0)\ p_i(N)\ \dots\ p_i(L_{\boldsymbol{p}_i}-1)]^T$, with $p_i(n) = u_i(n) * h_m(n)$. From the dimensions of the vectors and the matrices in (6.25), the length of the component $\boldsymbol{p}_i$ is

$$L_{\boldsymbol{p}_i} = \left\lceil \frac{L + L_{\text{pr}} - 1}{N} \right\rceil, \tag{6.27}$$

where $\lceil \cdot \rceil$ denotes the ceiling operator. The length of the subband filters, $h_{m,i}$, can be defined in terms of the length of the fullband filter, L, the prototype filter length, L_{pr}, and the decimation factor, N as

$$\tilde{L} = \left\lceil \frac{L + L_{\text{pr}} - 1}{N} \right\rceil - \left\lceil \frac{L_{\text{pr}}}{N} \right\rceil + 1. \tag{6.28}$$

A system of equations can then be formulated from (6.25) and the least squares solution for the subband filters, $\hat{h}_{m,i}$, can be written

$$\hat{h}_{m,i} = U_i^{+} \boldsymbol{p}_i, \quad i = 1, 2, \dots, \frac{K}{2} - 1 \tag{6.29}$$

In summary, given a fullband filter corresponding to room impulse response, $H_m(z)$, and $K/2$ subband filters satisfying perfect reconstruction and aliasing suppression in the subbands, a set of subband filters, $H_{m,i}(z)$, of the order L/N, can be estimated such that the filtering applied in the subbands results in an equivalent outcome as that of the fullband filtering model. For a large N this can result in significant order reduction of the very long room impulse responses.

Subband Inverse Filtering

The inverse filters can now be designed for each subband using the filters $H_{m,i}(z)$ obtained from (6.29). Here, this is done utilizing the multichannel LS filter design from (6.11), which now becomes

$$\hat{\boldsymbol{g}}_i = \boldsymbol{H}_i^{+}\boldsymbol{u}, \quad i = 1, 2, \ldots, \frac{K}{2} - 1, \tag{6.30}$$

such that for each subband

$$\sum_{m=1}^{M} G_{m,i}(z) H_{m,i}(z) = 1, \quad i = 1, 2, \ldots, \frac{K}{2} - 1. \tag{6.31}$$

Thus, equalization is achieved by applying the inverse filters, $\hat{\boldsymbol{g}}_{m,i}$, to the subband signals of the reverberant observations in each subband i, $\forall i$ and then an equalized fullband signal is constructed. Assuming that an exact inverse is achieved in each subband, the accuracy of the final result will depend on the reconstruction properties of the filterbank, on the accuracy of aliasing suppression and consequently, on the prototype filters.

6.2.2 Computational Complexity

One of the objectives of the subband implementation presented in Section 6.2.1 is to reduce the computational burden involved in solving the least squares problem for the inverse filter design and thus, to allow for the equalization of the very long impulse responses involved in acoustic signal processing. In this Section, a comparative analysis of the computations involved in the fullband and the subband implementations is presented, in order to appreciate the amount of computational savings gained with the proposed method.

The following analysis considers the computations required for solving a least squares problem using the normal equations. The computational comparison is made between the solution of the fullband least squares inverse filters and the subband least squares inverse filters, where the latter also includes the cost of the subband decomposition. The comparison is made by counting the number of floating point operations (flops) involved in solving either one of the above mentioned problems. The definition of a flop is adopted from that given by Golub and van Loan [96], were one flop is either one real multiplication or one real addition. It is noted in [96], that flop counts only provide a crude estimate for the efficiency of an algorithm since they do not include any other com-

putational overhead caused by the hardware. Nevertheless, for the purpose of this study, such overhead is be assumed similar in both cases and thus, the flop count is indicative of the computational load.

Consider the general optimization problem $\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$, which has a least squares solution $\boldsymbol{x}_{\mathrm{LS}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$, where $\boldsymbol{A}$ is an arbitrary real valued $p \times q$ matrix and $\boldsymbol{b}$ is a real valued $p \times 1$ vector. The number of flops required to solve this problem using the normal equations is given by [96]

$$pq^2 + \frac{q^3}{3}. \tag{6.32}$$

The fullband inverse filter calculation in (6.11) comprises the matrix $\boldsymbol{H}$ with dimensions $(L + L_{\mathrm{i}} - 1) \times (ML_{\mathrm{i}})$ and the vector $\boldsymbol{u}$ with dimensions ML_{i}, where the length of the inverse filter is given in (6.13). Thus, $p = (L + L_{\mathrm{i}} - 1)$ and $q = (ML_{\mathrm{i}})$. Substituting these values into (6.32), the number of flops required for the fullband inverse filter computation can be written

$$(ML_{\mathrm{i}})^2(L + L_{\mathrm{i}} - 1) + \frac{(ML_{\mathrm{i}})^3}{3}. \tag{6.33}$$

The subband inverse filter takes into consideration two separate calculations: the cost of the subband inverse filter computation (6.30) and the cost of the subband decomposition (6.29). Moreover, for these calculations, the data is complex. In general, one complex multiply requires four real multiplies and one complex addition requires two real additions. Thus, for the complex case, the expression in (6.32) is multiplied by a factor of four to account for the complex processing. This approximation is excessive, but indicates the computational burden nevertheless. Finally, the total cost is considered for all $K/2$ subbands

The subband inverse filter design in (6.30) consists of the matrix $\boldsymbol{H}_i$ with dimensions $(\tilde{L} + \tilde{L}_{\mathrm{i}} - 1) \times (M\tilde{L}_{\mathrm{i}})$ and the vector $\boldsymbol{u}_i$ with dimensions $M\tilde{L}_{\mathrm{i}}$, where the length of the inverse filter is now $\tilde{L}_{\mathrm{i}} = \left\lceil \frac{\tilde{L}-1}{M-1} \right\rceil$. This leads to $p = (\tilde{L} + \tilde{L}_{\mathrm{i}} - 1)$ and $q = (M\tilde{L}_{\mathrm{i}})$. The number of flops required for the subband inverse filter computation is then given by

$$4(M\tilde{L}_{\mathrm{i}})^2(\tilde{L} + \tilde{L}_{\mathrm{i}} - 1) + \frac{4(M\tilde{L}_{\mathrm{i}})^3}{3}. \tag{6.34}$$

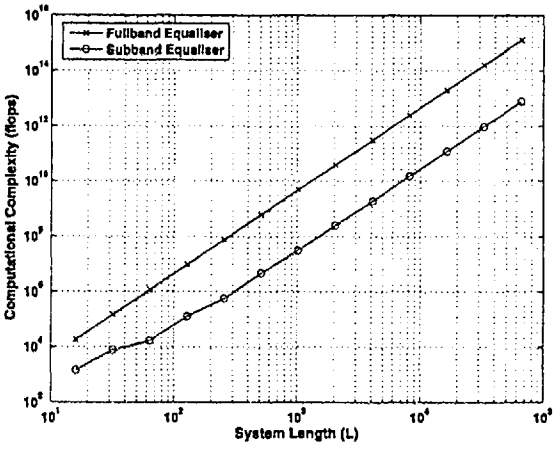

Figure 6.5: Computational complexity in terms of floating point operation count for the fullband and the subband equalizers.

Next, the complex subband decomposition problem in (6.29) contains a matrix $\boldsymbol{U}_i$ with dimensions $L_{\boldsymbol{p}_i} \times \tilde{L}$ and the vector $\boldsymbol{p}_i$ with dimensions $\tilde{L} \times 1$. Consequently, $p = L_{\boldsymbol{p}_i}$, $q = \tilde{L}$ and the number of flops required for the computation are

$$4(L_{\boldsymbol{p}_i}\tilde{L})^2 + \frac{4\tilde{L}^3}{3}, \tag{6.35}$$

where $L_{\boldsymbol{p}_i}$ and $\tilde{L}$ are given by (6.27) and (6.28) respectively.

Bringing together the results from (6.34) and (6.35), the total flops required for the subband inverse filter design can be expressed as

$$2K\left((M\tilde{L}_i)^2(\tilde{L}+\tilde{L}_i-1)+(L_{\boldsymbol{p}_i}\tilde{L})^2+\frac{(M\tilde{L}_i)^3+\tilde{L}^3}{3}\right). \tag{6.36}$$

It can be seen that the the main factor of the computational complexity is the system length and thus, the improvement achieved by the subband approach will depend on the number of subbands and on the decimation ratio. An example is given in Fig. (6.5) where the computational complexity is calculated for the fullband case and for the subband case using (6.33) and (6.36) respectively. The subband implementation for this example is that presented in Section 6.2.1 with $K = 32$ subbands decimated by $N = 24$. On aver-

age over all lengths, the subband approach reduces the required flops by a factor of 120. As an alternative interpretation, it can be seen from the graph that the computational gain for this example is of the order of 6, which is a significant improvement. For example, a fullband system of $L = 4000$ taps is processed by the subband implementation at approximately the same cost as a $L = 700$ tap system is processed by the fullband system.

6.3 Simulations and Results

Simulation results are now presented to demonstrate the performance of the proposed multichannel LS inverse filtering method in oversampled subbands derived in Section 6.2. Three experiments were performed to show the following: (i) an illustration of the performance with long impulse responses, (ii) effects of varying the channel length for various impulse responses and (iii) inverse filtering with inexact impulse responses.

Given the discrete Fourier transform of the equalized impulse response, $H_{\mathrm{EQ}}(k) = |H_{\mathrm{EQ}}(k)|e^{j\theta_{\mathrm{EQ}}(k)}$, the magnitude and phase distortions were evaluated separately. Two objective measures were employed. First, a magnitude distortion metric is used, which estimates the standard deviation, σ_{EQ}, of the equalized magnitude response from its mean level and is defined as [83]

$$\sigma_{\mathrm{EQ}} = \sqrt{\frac{1}{N_{\mathrm{FT}}} \sum_{k=0}^{N_{\mathrm{FT}}-1} \left(10\log_{10}|H_{\mathrm{EQ}}(k)| - \bar{H}_{\mathrm{EQ}}\right)^2} \ \mathrm{dB}, \tag{6.37}$$

with

$$\bar{H}_{\mathrm{EQ}} = \frac{1}{N_{\mathrm{FT}}} \sum_{k=0}^{N_{\mathrm{FT}}-1} 10\log_{10}|H_{\mathrm{EQ}}(k)|,$$

where N_{FT} is the number of frequency points considered. The second metric, Δ_{EQ}, is a measure of the deviation of the unwrapped phase from a linear fit to its values and is defined here as

$$\Delta_{\mathrm{EQ}} = \sqrt{\frac{1}{N_{\mathrm{FT}}} \sum_{k=0}^{N_{\mathrm{FT}}-1} \left(\theta_{\mathrm{EQ}}(k) - \bar{\theta}_{\mathrm{EQ}}(k)\right)^2}, \tag{6.38}$$

where $\bar{\theta}_{\mathrm{EQ}}(k)$ is the least squares fit to the phase at frequency point k. The least squares fit is calculated over all frequencies and represents linear phase.

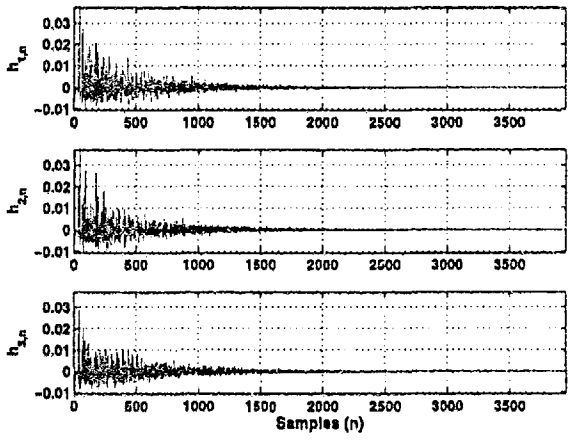

Figure 6.6: Simulated room impulse responses for $M = 3$ channels of length $L = 4000$.

Experiment 1: example with room impulse responses

This experiment shows a typical outcome of the inversion procedure using long acoustic channels. An acoustic system with $M = 3$ channels is simulated with the impulse responses shown in Fig. 6.4. The channel lengths were $L = 4000$ taps which is equivalent to $T_{60} = 0.5$ s at a sampling frequency $f_s = 8$ KHz. The impulse responses shown in Fig. 6.6 were used as input to the subband equalizer and the resulting output is shown in Fig. 6.7 where Fig. 6.7a shows the time domain result and Fig. 6.7b shows the corresponding magnitude response. It can be seen that near perfect equalization is achieved, however, there is some spectral distortion, especially in the low frequency region. This is partly due to the approximation of the subband filters and distortion resulting from the filterbank reconstruction. The accuracy of this depends on the ability of the prototype filter to suppress aliasing and on the oversampling ratio. The delay in the equalized impulse in Fig. 6.7a is due to the filterbank and is related to the order of the prototype filter L_{pr}.

Experiment 2: magnitude and phase distortion for different channel lengths

In this experiment, the investigation is concerned with the performance of the subband inversion algorithm. A system with $M = 5$ observations is assumed and two types of impulse responses were considered: (i) random channels and (ii) simulated impulse responses where the first three of the latter are shown in Fig. 6.6. The channel length was varied for

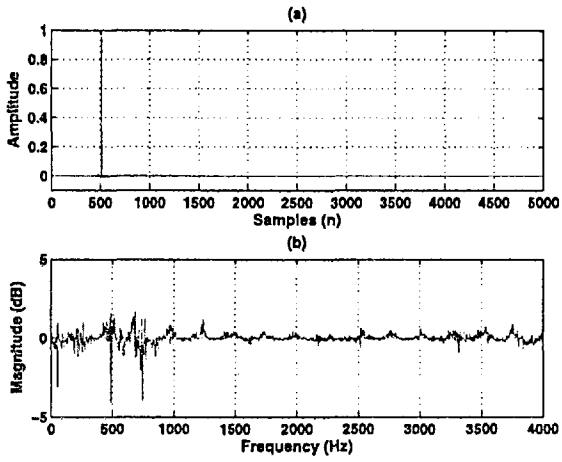

Figure 6.7: Equalized responses in (a) time domain and (b) magnitude response, using the subband multichannel least squares method.

$L = \{100, 300, 500, 1000, 2000, 4000\}$ taps. The simulated room impulse responses were truncated to the desired lengths. For each case the magnitude and the phase distortions were measured according to (6.37) and (6.38). The resulting plot is given in Fig. 6.8 for a) an average of ten different random channels realizations at each channel length and b) the simulated room impulse responses. In agreement with the example in Experiment 1, the magnitude distortion resulting from the filterbank is apparent here and seems to reside on average at 0.4 dB for the random channels and 0.2 dB for the simulated impulse responses. Nevertheless, these errors are generally small as will be seen in the next experiment. There appears to be no apparent dependence on the length of the channel.

Experiment 3: inversion with inexact impulse response estimates

Finally, experiments were performed to investigate the sensitivity of the subband approach to inexact impulse responses. The results are then compared to the corresponding fullband multichannel LS inversion. First, a system with $M = 5$ observations was assumed with random channels of length $L = 512$ taps. Random noise was added to the channels to produce an NPM varying between 0 and -80 dB. The results are shown in Fig. 6.9a for the fullband case and Fig. 6.9b for the proposed subband implementation are the averaged outcome from 10 different random channels. Interestingly, it is observed that the subband case exhibits similar properties as the single channel least squares inversion

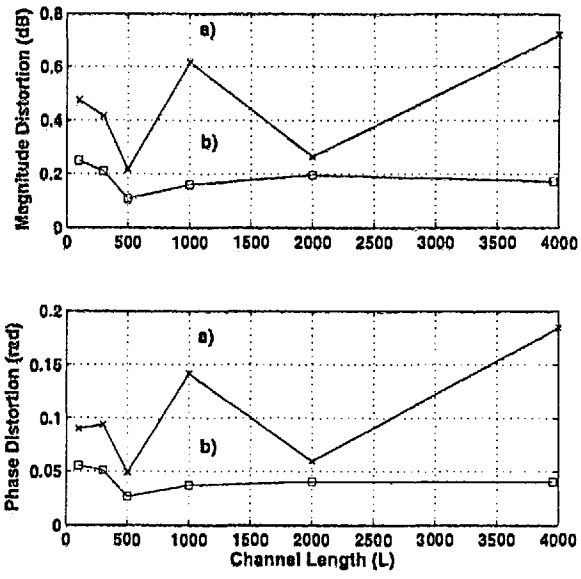

Figure 6.8: Magnitude and phase distortions vs. Channel length for a) random channel impulse responses and b) a simulated acoustic impulse response.

shown in Fig. 6.2. Thus, the subband multichannel LS method appears to be less sensitive to inaccurate channel estimates while benefiting from the shorter filters of multichannel inversion. Possible explanations to this are that the subband filters are approximations of the fullband response and that the filter lengths are reduced. The second statement is motivated by the observation that the sensitivity to errors in the channels estimates is increased with increased channel length, which can be seen by comparing the outcomes of Fig. 6.2 and 6.9. The outcome of an additional experiment with simulated acoustic impulse responses, truncated to $L = 512$ is shown in Fig. 6.10, where the results are similar to the random channels.

6.4 Application to Speech Dereverberation

The subband method is applied to speech dereverberation using similar conditions as for the experiments in Chapter 4, such that the results can be compared with those of the SMERSH method. Test data from the APLAWD database comprising the sentence "George made the girl measure a good blue vase" uttered by a a male talker was used

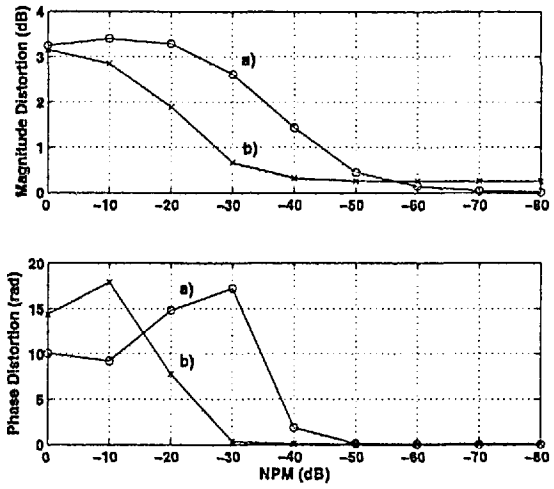

Figure 6.9: Magnitude and phase distortions vs. NPM for a) inverse filtering with the fullband multichannel LS method and b) subband multichannel LS inverse filtering with random impulse responses.

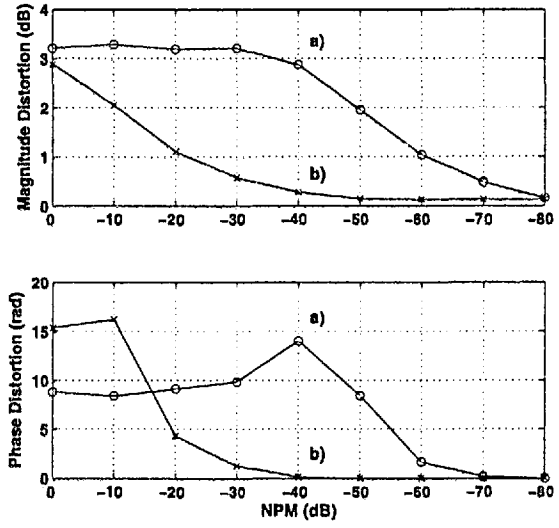

Figure 6.10: Magnitude and phase distortions vs. NPM for a) inverse filtering with the fullband multichannel LS method and b) subband multichannel LS inverse filtering with simulated room impulse responses.

as an example. The sampling frequency was $f_s = 8$ kHz. A room with dimensions $5 \times 4 \times 3$ m was simulated using the source-image method [12] described in Section 2.3. An eight element linear microphone array with 0.05 m spacing between adjacent microphones was assumed 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 NPM were simulated ranging from 0 to $-\infty$. The delay-and-sum beamformer was used as a baseline algorithm. Only one talker is used for this experiment in contrast to the ten talkers used for the dereverberation experiment in Chapter 4. This is because the inversion algorithm depends less on the talker than does the speech enhancement algorithm. The segmental SRR and the Bark spectral distortion measures described in Section 2.4 were applied to quantify the improvement of the processed speech.

The results in terms of Segmental SRR vs. reverberation time, T_{60}, are shown in Fig. 6.11 for reverberant speech, speech at the output of the DSB and speech inverse filtered with the proposed method and with various levels of NPM in the impulse responses ranging from 0 to $-\infty$. The legend in the figure indicates the different plots. Figure 6.12 shows the corresponding improvement for all cases in terms of the difference between the reverberant and the processed speech results. In all cases the inverse filtering approach provides significant improvements over the DSB, but the improvement degrades as the NPM increases, as could be expected. The results evaluated with the Bark spectral distortion are shown in Fig. 6.13 for reverberant speech, speech at the output of the DSB and speech inverse filtered with the proposed method and with various levels of NPM in the impulse responses ranging from 0 to $-\infty$. The corresponding improvement is depicted in Fig. 6.14. It can be seen that for NPM < -30 dB, the dereverberated speech signals are equivalent (in terms of BSD) with the clean speech for all reverberation times considered. This was also confirmed with listening tests. The listening tests also showed that, in the case of NPM = 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 does correspond accurately to the expected outcome based on the results presented in Figs. 6.9 and 6.10, where are large spectral distortion is observed at NPM = 0 dB.

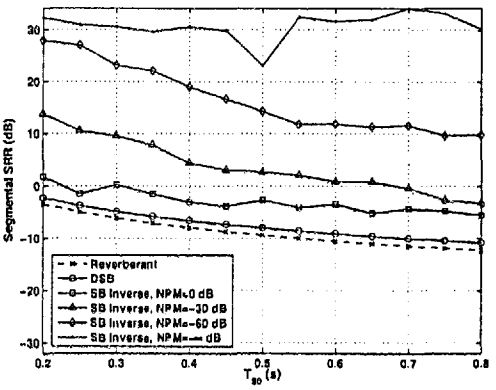

Figure 6.11: Segmental SRR for reverberant speech at the microphone closest to the talker, speech at the output of a DSB and speech processed with subband inverse filtering for NPM = $\{0, -30, -60, -\infty\}$ dB.

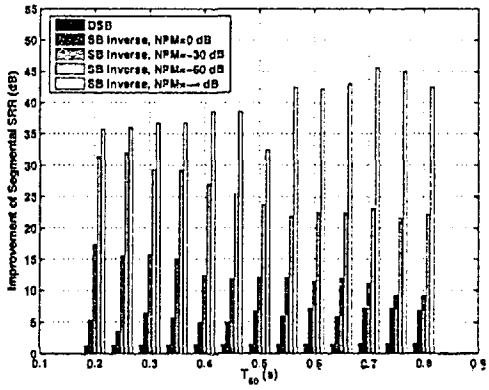

Figure 6.12: Improvement of Segmental SRR with DSB and subband inverse filtering for NPM = $\{0, -30, -60, -\infty\}$ dB.

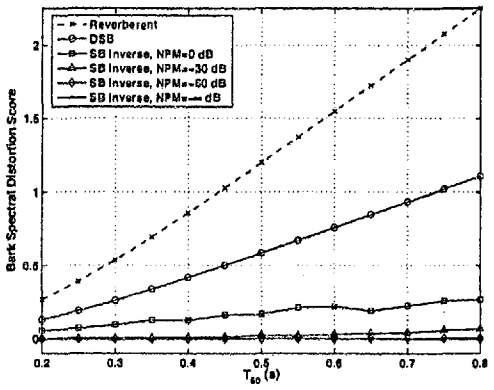

Figure 6.13: Bark spectral distortion score for reverberant speech at the microphone closest to the talker, speech at the output of a DSB and speech processed with subband inverse filtering for NPM = $\{0, -30, -60, -\infty\}$ dB.

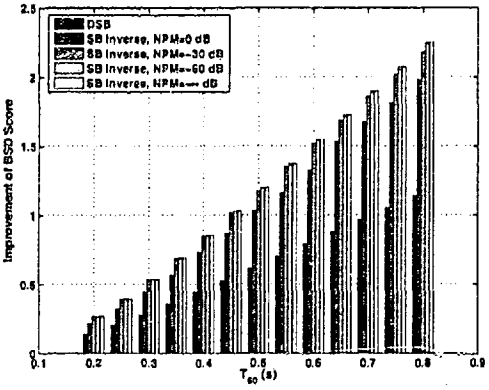

Figure 6.14: Improvement of Bark spectral distortion score with with DSB and subband inverse filtering for NPM = $\{0, -30, -60, -\infty\}$ dB.

6.5 Summary

Inverse filtering of acoustic impulse responses has been discussed both for single and for multiple microphones. Single microphone approaches can provide only approximate inversion, require very long inverse filters and result in long processing delay due to the non-minimum phase property of room impulse responses. On the other hand, exact inversion with no delay and with inverse filters of similar order to the room impulse responses is possible in the multimicrophone case. However, this is very sensitive to estimation errors in the impulse responses compared to the more robust single channel methods.

A new algorithm was derived operating on decimated, oversampled subband signals, where the fullband impulse response is decomposed into equivalent filters in the subbands and multichannel least squares inversion is applied in each subband. It was shown that this method results in substantial computational savings at the cost of small spectral distortion due to the subband filters. Simulation results were presented to evaluate the performance of this method and inversion of channels of several thousand taps was demonstrated. Experimental results indicated that the new method is more robust to impulse response estimation errors, which is due to a combination of shorter filters and approximation of the filtering in the subbands. 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 NPM < -30 dB.

Chapter 7

Conclusions

THIS final chapter summarizes and concludes the work presented in this thesis. First Section 7.1 presents a résumé of the main problems addressed. The core results are highlighted in Section 7.2 and finally, in Section 7.3 an outlook and guidelines for future research in speech dereverberation are presented.

7.1 Résumé

This thesis has dealt with the problem of blind speech dereverberation. That is, dereverberation of speech using only speech signals obtained in a reverberant room by one or more microphones positioned at a distance from the talker. The problem was formulated assuming an FIR model for the room impulse responses and dereverberation techniques were broadly categorized into: (i) beamforming, (ii) speech enhancement and (iii) blind system identification and inversion.

Delay-and-sum beamforming. The delay-and-sum beamformer (DSB) was chosen as a baseline method. It is the simplicity and yet reasonable dereverberation performance with little perceptual distortion which has made the DSB an attractive technique for comparing with newly developed methods. However, employing results from statistical room acoustics, it was shown here that the dereverberation performance of the DSB is limited in practice and in particular when the number of sensors is small ($M \leq 8$). These limitations of the DSB were further demonstrated with simulation experiments, both in terms of Bark spectral distortion and Signal-to-Reverberation Ratio.

Reverberant speech enhancement using linear prediction. A study on linear prediction of reverberant speech formed the first part of this thesis. Using a statistical

model of the room transfer functions, it was shown that, in terms of spatial expectation, the AR coefficients of anechoic voiced speech can be estimated accurately from reverberant observations. Moreover, it was demonstrated that multichannel LPC, where the AR coefficients are simultaneously estimated from multiple reverberant observations, provide the best estimates of the clean speech AR coefficients.

Consequently, it was demonstrated that the room impulse response affects mainly in the prediction residual, where it introduces erroneous peaks of similar strength to the peaks due to GCIs in the original excitation sequence. This fact, in combination with the knowledge of the general structure of the prediction residual of speech was then employed to develop a new dereverberation algorithm, which utilizes temporal averaging of neighbouring larynx cycles on spatially averaged speech observations. The ability to perform larynx cycle segmentation in speech at the output of the DSB is crucial and it was demonstrated that such segmentation can be performed with satisfactory accuracy using the DYPSA algorithm. The improvement in speech quality applying the new algorithm was demonstrated with various experiments in terms of segmental signal-to-reverberation ratio and Bark spectral distortion.

Adaptive blind SIMO system identification. The second major component of the thesis constituted a study on blind SIMO system identification and inversion. The cross-relation between channels and the resulting eigendecomposition method for blind identification of SIMO systems were introduced. An adaptive solution, the complex MCLMS, based on minimization of an error formed from the cross-relation was derived. Problems identified for this algorithm included sensitivity to noise, robustness to common zeros and channel order estimation.

First, the problem of noise sensitivity was addressed by the derivation of a new variable step-size, which is the optimal step-size for each iteration, given the latest channel estimates. Since this optimal step-size relies on the knowledge of the true system, an approximate implementation was proposed, which was demonstrated to significantly improve the identification performance both in noisy and noise-free cases.

Next, an adaptive method for the identification of common roots in polynomials was developed and was employed in a study of the sensitivity of the MCLMS algorithm to common zeros. It was shown that zeros do not have to be exactly identical to perturb the system identification process in the adaptive algorithm and thus, that not only common

but also near-common zeros are problematic. It was then demonstrated that near-common zeros are very likely to occur in large order acoustic systems.

An attempt to reduce the system length and, consequently, to improve performance of the MCLMS algorithm was made using a multirate adaptive filtering structure. It was shown that for oversampled filterbanks, there exists a set of subband filters which minimize the global cross-relation error between channels. One remaining problem of the subband blind system identification is the (complex) scalar ambiguity in each subband inherent in the MCLMS algorithm.

Dereverberation with approximate impulse response estimates. In most realistic scenarios, the inversion algorithm will operate on approximate estimates of the true acoustic paths and on noisy observations. In conformity with blind system identification, the very long impulse responses are problematic and in addition, the non-minimum phase nature of room impulse responses makes their inversion difficult. The non-minimum phase problem can be avoided by using multiple observations or approximate filtering with a single observation. Multichannel approaches enable exact, delayless inversion with inverse filters of similar order as the impulse responses but are sensitive to inaccuracies in the estimated impulse responses. In contrast, approximate single channel inversion has the advantage of lesser sensitivity to noise and erroneous impulse responses but suffers from extremely long inverse filters and modelling delay.

A new multichannel least squares (LS) inversion method operating in decimated subbands was derived to address the above mentioned issues. The fullband impulse response is first decomposed into equivalent subband filters followed by multichannel least squares inversion in each subband. It was shown that this implementation is much much more efficient in terms of floating point operations and allows inversion of very long impulse responses at the cost of a small spectral distortion. Moreover, it was demonstrated that the new method is less sensitive to inaccuracies in the system estimates. Finally, the subband multichannel LS approach was applied in the context of speech dereverberation, where simulation results indicated a large improvement in terms of segmental SEGSRR and BSD, in cases when the channel estimates are good (NPM$\geq$ 30dB).

Table 7.1: Collected results - Segmental SRR and BSD vs. different T_{60} for speech: at the closest mic. (Reverb.), at the output of a DSB, processed with the SMERSH algorithm and with subband multichannel inverse filtering with impulse responses estimates of NPM=-30dB (SBInv.)

	Segmental SRR (dB)				BSD (Score)			
T_{60}(s)	Reverb.	DSB	SMERSH	SBInv. NPM=-30dB	Reverb.	DSB	SMERSH	SBInv. NPM=-30dB
0.2	-4.7703	-3.0165	-2.6211	13.8211	0.3369	0.1584	0.1375	0.0012
0.3	-7.8983	-6.1295	-5.2296	9.5300	0.6520	0.3077	0.2225	0.0037
0.4	-10.0047	-8.1849	-6.8218	4.4280	1.0147	0.4811	0.3191	0.0105
0.5	-11.5847	-9.7183	-8.1168	2.7091	1.4005	0.6654	0.4163	0.0235
0.6	-12.8312	-10.9213	-8.9888	0.9076	1.7961	0.8531	0.5062	0.0279
0.7	-13.8645	-11.9222	-9.9036	-0.4135	2.1953	1.0418	0.6048	0.0408
0.8	-14.7304	-12.7604	-10.6671	-3.2155	2.5935	1.2297	0.7325	0.0669

7.2 Summary of Main Results

The main aim of this work was to reduce the effects of room reverberation in speech signals obtained by multiple microphones in a reverberant room. Two methods were proposed: (i) a speech enhancement method operating on the linear prediction residual of the reverberant speech and (ii) a method comprising adaptive blind system identification and multichannel inversion. Simulation results for both methods and for the DSB are collected in Table 7.1. The core results can be summarized as follows:

- The SMERSH method operating on the prediction residual reduces reverberation and gives a significant performance improvement over the DSB, especially for long reverberation times.

- There are two major advantages with SMERSH: (i) it does not require any knowledge of the room impulse responses and (ii) it contains no computationally burdensome components, and could be implemented in a real-time system. However, the level of reverberation reduction is limited.

- Despite the advances of the adaptive MCLMS algorithm for blind system identification of room impulse responses, its practical applicability is limited due to inability of tackling channel undermodelling and long impulse responses. However, it was demonstrated to accurately estimate channels of the order of 100 taps even for speech input signal.

- Multichannel equalization of, possibly noisy estimates, of the room impulse responses can be performed efficiently using the newly proposed subband implementation.
- Blind system identification/inversion methods are very computationally intensive. Nevertheless, in case of successful channel estimation, dereverberation can be performed with results superior to the speech enhancement method.

7.3 Future Directions for Speech Dereverberation

This last section of the thesis will present some directions for future development and possible extensions of the ideas presented in the foregoing chapters and finally, the work is concluded with an outlook of the future of speech dereverberation research.

Extensions to the current work

The study of the AR coefficients in Chapter 3 did not consider windowing. From the results obtained with the algorithm, it appeared that no significant degradation in speech quality is introduced which could be attributed to the estimated AR coefficients. However, the work in Chapter 3 can be extended by including the consideration of a windowing function in the analysis and to study the effects of the windowing in relation to, for example, reverberation time.

An important factor in the SMERSH algorithm is the correct identification of the GCI's in reverberant speech. In Chapter 4 DYPSA was applied at the output of the delay-and-sum beamformer. Instead, an alternative approach could be to incorporate the multichannel data into DYPSA such that GCI candidates are generated for each channel and then using the spatial information to eliminate erroneous candidates. Furthermore, a study of the inverse filter arising from the enhanced larynx cycles would be interesting. For example, partial knowledge of the room impulse responses could be incorporated.

The work presented on blind system identification could be extended in several ways. An important consideration would be the solution to the gain and phase ambiguity in the blind system identification algorithm for the application to subband processing. One alternative is to consider real subbands by using, for example, single sideband filterbanks [115] which then results only in gain ambiguity in the channel estimates. A post-processing stage could then be considered where the fullband signal is corrected at the boundaries of the subband filters. If the equalization is performed in the subbands, the post-processing

could operate on the output speech signal and could then make use of a speech production model. Alternatively, it may be possible to constrain the adaptive algorithm such that the scalar constant can be inferred. Yet another important task would involve a study of the relation between the cross-relation error minimization, the system mismatch and the zeros of the unknown systems such that improved algorithms can be devised.

The subband multichannel inversion algorithm could be studied further in terms of computational efficiency. The results presented in Section 6.2.2 can be extended to investigate the relations between number of subbands, decimation ratio, subband filter length, number of sensors and length of the filter to be equalized. In this way, an optimal (in some sense) configuration can be found for a particular application.

Subsequently, the method can be extended by applying different processing in each subband to reduce the sensitivity to erroneous channel estimates and observation noise. This could, for example, include a combination of exact inversion, approximate inversion and no processing. The selection of inversion technique for a particular subband could be made to depend on the expected accuracy of the channel estimates and on the dynamic range of the transfer function in each subband. Furthermore, the subbands could be weighted differently before the final reconstruction using some perceptual criteria, such as A-weighting curves.

Outlook

From the work presented here and also from other published results, it seems that speech enhancement algorithms are the most promising in terms of practical applications at the time of writing. In particular, the use of speech models is attractive and will be important also in future developments since it in a sense reduces the 'blindness' of the dereverberation problem.

Nevertheless, blind system identification/inversion and in particular adaptive solutions are very attractive and should be encouraged despite the difficulty of the problem. The two most important features for the success of such methods in the future are: (i) ability to system undermodelling and (ii) decomposition of the problem, such as decimated subband processing.

The reverberation models and evaluation of the processed speech will have to be addressed. It is important to reconsider the models of room reverberation used when the problem is formulated. The FIR model commonly assumed, even though convenient

for algorithmic development, results in very long filters and may be difficult to use in practice. Evaluation schemes for reverberant and dereverberated speech, which in some sense represent a mean subjective opinion, are yet to emerge, such that newly published algorithms can be assessed easier more reliably based on graphic results. However, the final optimal judge will be human hearing.

This list is not exhaustive and highlights not only some prospective directions of research, but also the amount of open research questions available in this exciting field of study. Over the last few years there has been a substantial increase in interest towards (multichannel) speech dereverberation, both from academic and industrial institutions and many exciting developments can be expected to appear in the not so distant future.

Bibliography

[1] M. Omologo, P. Svazier, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Commun.*, vol. 25, no. 1, pp. 75–95, Aug. 1998.

[2] G. Schmidt, "Applications of acoustic echo control - an overview," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 9–16.

[3] Mobile Operators Association, "History of cellular mobile communications," [Online] Available: http://www.mobilemastinfo.com/information/history.htm, 2005.

[4] UK BBC, "Iceland comes first in broadband," [Online] Available: http://news.bbc.co.uk/1/hi/technology/4903776.stm, Apr. 2006.

[5] S. L. Gay and J. Benesty, Eds., *Acoustic Signal Processing for Telecommunication*, Kluwer Academic Publishers, Mar. 2000.

[6] G. M. Davis, Ed., *Noise Reduction in Speech Applications*, CRC Press, May 2002.

[7] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer-Verlag Berlin and Heidelberg GmbH & Co. K, May 2005.

[8] Polycom, "Polycom communicator," [Online] Available: http://www.polycom.com/, 2006.

[9] Plato, *The Republic*, Penguin Books Ltd, 2003.

[10] J. W. S. Rayleigh, *The theory of sound*, Dover Publications Inc, 1976.

[11] W. C. Sabine, *Collected papers on acoustics*, Dover publishers, 1964.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[13] H. Kuttruff, *Room Acoustics*, Taylor & Francis, 4 edition, Oct. 2000.

[14] Bluetooth SIG, "The official bluetooth wireless info site," [Online] Available: http://www.bluetooth.com/, 2006.

[15] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, 1 edition, 2001.

[16] Acoustic Magic, "Voicetracker array microphone," [Online] Available: http://www.acousticmagic.com/.

[17] Microsoft Corporation, "Microphone array support in windows vista," [Online] Available: http://www.microsoft.com/whdc/device/audio/MicArrays.mspx.

[18] J. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, Apr 1991.

[19] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 320–328, Apr 1994.

[20] J. R. Hopgood and P. J. W. Rayner, "A probabilistic framework for subband autoregressive models applied to room acoustics," in *Proc. IEEE Workshop Statistical Signal Processing*, Aug 2001, pp. 492 – 495.

[21] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 476 – 488, Sep 2003.

[22] T. Paatero, "Modeling of long and complex responses using kautz filters and time-domain partitions," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 313–316.

[23] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.

[24] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, June 2003.

[25] D. B. Ward, "On the performance of acoustic crosstalk cancellation in a reverberant environment," *J. Acoust. Soc. Amer.*, vol. 110, no. 2, pp. 1195–1198, Aug. 2001.

[26] M . R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *J. Audio Eng. Soc.*, vol. 35, no. 5, pp. 299–305, May 1987.

[27] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, July 1979.

[28] D. T. Murphy and D. M. Howard, "Modelling and directionally encoding the acoustics of a room," *IEE Electronics Lett.*, vol. 34, no. 9, pp. 864–865, Apr. 1998.

[29] G. R. Campos and D. M. Howard, "On the computational efficiency of different waveguide mesh topologies for room acoustic simulation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 1063–1072, Sept. 2005.

[30] T. I. Laakso, V. Valimaki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay," *IEEE Signal Processing Mag.*, vol. 13, no. 1, pp. 30 – 60, Jan. 1996.

[31] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819–829, June 1992.

[32] A. S. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[33] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing For Telecommunication*, S. L. Gay and J. Benesty, Eds., pp. 261–279. Kluwer Academic Publishers, 2000.

[34] S. M. Griebel, *A Microphone Array System for Speech Source Localization, Denoising and Dereverberation*, Ph.D. thesis, Harvard University, Cambridge, Massachusetts, Apr. 2002.

[35] J. Tackett, "Iso 226 equal-loudness-level contour signal," http://www.mathworks.com/matlabcentral/fileexchange/ loadFile.do?objectId=7028&objectType=file, Jan. 2005.

[36] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Paris, France, Sept. 2006.

[37] J. Y. C. Wen and P. A. Naylor, "An evaluation measure for reverberant speech using tail decay modelling," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.

[38] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Signal Processing Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[39] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Commun.*, vol. 20, no. 3-4, pp. 229–240, Dec. 1996.

[40] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, 1 edition, 1975.

[41] F. Talantzis and D. B. Ward, "Robustness of multi-channel equalization in an acoustic reverberant environment," *J. Acoust. Soc. Amer.*, vol. 114, no. 2, pp. 833–841, Aug. 2003.

[42] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

[43] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985.

[44] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, no. 1-2, pp. 207–222, Oct. 1993.

[45] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, N.J., 4 edition, 2001.

[46] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[47] T. Nishiura, S. Nakanura, and K. Shikano, "Speech enhancement by multiple beamforming with reflection signal equalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2001, vol. 1, pp. 189–192.

[48] E. Jan and J. Flanagan, "Microphone arrays for speech processing," in *International Symposium on Signals, Systems, and Electronics*, Oct. 1995, pp. 373–376.

[49] E. Jan, P. Svazier, and J. L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *IEEE International Symposium on Circuits and Systems (ISCAS '95)*, May 1995, vol. 2, pp. 1460–1463.

[50] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 425–437, Sept. 1997.

[51] Y. Grenier and S. Affes, "Microphone array response to speaker movements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1997, vol. 1, pp. 247–250.

[52] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 3, pp. 437–466, Sept. 1968.

[53] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[54] J. B. Allen, "Synthesis of pure speech from a reverberant signal," U.S. Patent No. 3786188, Jan. 1974.

[55] S. M. Griebel and M. S. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Pocono Manor, Pennsylvania, Sept. 1999.

[56] S. M. Griebel and M. S. Brandstein, "Microphone array speech dereverberation using coarse channel estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 1, pp. 201–204.

[57] B. Yegnanarayana, S. R. Mahadeva Prasanna, and K. Sreenivasa Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, vol. 1, pp. 541–544.

[58] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 6, pp. 3701–3704.

[59] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Language Porcessing*, vol. 14, no. 3, pp. 774–784, May 2006.

[60] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Single-microphone blind dereverberation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer Verlag, 1 edition, Apr. 2005.

[61] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, Mar. 2005, vol. 4, pp. iv/173 – iv/176.

[62] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Processing*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[63] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[64] L. Gürelli and C. L. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 143–149, Jan. 1995.

[65] S. Gannot and M. Moonen, "Subspace methods for multi-microphone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, Oct. 2003.

[66] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Process.*, vol. 82, no. 8, pp. 1127–1138, Aug. 2002.

[67] Y. Huang, J. Benesty, and J. Chen, "Optimal step size of the adaptive multichannel LMS algorithm for blind SIMO identification," *IEEE Signal Processing Lett.*, vol. 12, no. 3, pp. 173–176, Mar. 2005.

[68] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept. 2005.

[69] Md. K. Hasan, J. Benesty, P. A. Naylor, and D. B. Ward, "Improving robustness of blind adaptive multichannel identification algorithms using constraints," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sept. 2005.

[70] N. D. Gaubitch, Md. K. Hasan, and P. A. Naylor, "Noise robust adaptive blind identification using spectral constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toulouse, France, May 2006, pp. V–93 – V–96.

[71] K. Furuya and Y. Kaneda, "Two-channel blind deconvolution for non-minimum phase impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 1315–1318.

[72] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 5, pp. 392–396, Sept. 1996.

[73] A. P. Petropulu and C. L. Nikias, "Blind deconvolution using signal reconstruction from partial higher order cepstral information," *IEEE Trans. Signal Processing*, vol. 41, no. 6, pp. 2088–2095, June 1993.

[74] M. Triki and D. T. M. Slock, "Delay-and-predict equalization for blind speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toulouse, France, May 2006, pp. V97 – V100.

[75] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1982, vol. 7, pp. 1858–1861.

[76] J. N. Mourjopoulos, "Digital equalization of room acoustics," *J. Audio Eng. Soc.*, vol. 42, no. 11, pp. 884–900, Nov. 1994.

[77] B. D. Radlović and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[78] M. Tohyama, R. H. Lyon, and T. Koike, "Source waveform recovery in a reverberant space by cepstrum dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1993, vol. 1, pp. 157–160.

[79] R. A. Kennedy and B. D. Radlović, "Iterative cepstrum-based approach for speech dereverberation," in *Proc. of the Fifth International Symposium on Signal Processing and Its Applications (ISSPA '99)*, Aug. 1999, vol. 1, pp. 55–58.

[80] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[81] K. Yamada, J. Wang, and F. Itakura, "Recovering of broad band reverberant speech signal by sub-band MINT method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 969–972.

[82] P. A. Nelson, F. Orduña-Brustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 3, pp. 185–192, Nov. 1995.

[83] B. D. Radlović, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[84] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 3, pp. 309–321, June 1975.

[85] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 6, pp. 488–494, Dec. 1976.

[86] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 3, pp. 197–210, June 1978.

[87] P. A. Nelson and S. J. Elliott, *Active Control of Sound*, Academic Press, 1993.

[88] S. Bharitkar, P. Hilmes, and C. Kyriakakis, "Robustness of spatial average equalization: A statistical reverberation model approach," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3491–3497, Dec. 2004.

[89] F. Talantzis, D. B. Ward, and P. A. Naylor, "Expected performance of a family of blind source separation algorithms in a reverberant room," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2004, vol. 4, pp. 61–64.

[90] T. Gustafsson, B. D. Rao, and M Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 791–803, Nov. 2003.

[91] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of Linear Prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Kyoto, Japan, Sept. 2003, pp. 99–102.

[92] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Tech. Rep., University College London, June 1987.

[93] J. Makhoul, "Linear Prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[94] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Macmillan, 1993.

[95] M. Kendall, A. Stuart, and J. K. Ord, *Kendall's Advanced Theory of Statistics*, vol. 1, Hodder Arnold, 6 edition, October 1994.

[96] G. H. Golub and C. F. van Loan, *Matrix computations*, John Hopkins series in the mathematical sciences. London: John Hopkins University Press, 3 edition, 1996.

[97] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 3, pp. 283 – 296, June 1976.

[98] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637–655, Aug. 1971.

[99] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Porcessing*, vol. 15, no. 1, pp. 34–43, 2007.

[100] G. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[101] B. Yegnanarayana, J. M. Naik, and D. G. Childers, "Voice simulation: factors affecting quality and naturalness," in *Proc. Conf. of the Association for Computational Linguists*, Stanford, California, USA, July 1984, pp. 530–533.

[102] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Mulit-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[103] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, Dec. 1990.

[104] M. Huckvale, "Speech filing system: Tools for speech research," [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/, July 2003.

[105] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2002, vol. 1, pp. I–349 – I–352.

[106] F. J. Harris, "On the use of windows for harmonic analysis with the Discrete Fourier Transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[107] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug. 1997.

[108] Z. Xu and M. K. Tsantsanis, "Blind channel estimation for long code multiuser CDMA systems," *IEEE Trans. Signal Processing*, vol. 48, no. 4, pp. 988–1001, Apr. 2000.

[109] K. F. Kaaresen and T. Taxt, "Multichannel blind deconvolution of seismic signals," *Geophysics*, vol. 63, no. 6, pp. 2093–2107, Nov. 1998.

[110] L. Tong and S. Perreau, "Multichannel blind identification: from subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951–1968, Oct. 1998.

[111] X. Lin, N. D. Gaubitch, and P. A. Naylor, "Two-stage blind identification of SIMO systems with common zeros," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.

[112] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Lett.*, vol. 5, no. 7, pp. 174–176, July 1998.

[113] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[114] P. P. Vaidyanathan, *Multirate systems and filter banks*, Prentice Hall, 1993.

[115] S. Weiss and R. W. Stewart, *On adaptive filtering in oversampled subbands*, Shaker Verlag, 1998.

[116] S. Roy and C. Li, "A subspace blind channel estimation method for OFDM systems without cyclic prefix," *IEEE Trans. Wireless Commun*, vol. 1, no. 4, pp. 572–579, Oct. 2002.

[117] H. Gazzah, P. A. Regalia, J. Delmas, and K. Abed-Meraim, "A blind multichannel identification algorithm robust to order overestimation," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1449–1558, June 2002.

[118] B. Widrow, J. McCool, and M. Ball, "The complex LMS algorithm," *Proc. IEEE*, vol. 63, no. 4, pp. 719–720, April 1975.

[119] C. Avendano, J. Benesty, and D. R. Morgan, "A least squares component normalization approach to blind channel identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, Arizona, Mar. 1999, vol. 4, pp. 1797 – 1800.

[120] R. Ahmad, A. W. H. Khong, Md. K. Hasan, and P. A. Naylor, "An extended normalized multichannel fims algorithm for blind channel identification," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.

[121] R. W. Harris, D. M. Chabries, and F. A. Bishop, "A variable step (vs) adaptive filter algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 2, pp. 309–316, Apr. 1986.

[122] T. Aboulnasr and K. Mayyas, "A robust variable step size LMS-type algorithm: analysis and simulations," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 631–639, Mar. 1997.

[123] N. D. Gaubitch, Md. K. Hasan, and P. A. Naylor, "Generalized optimal step-size for blind multichannel LMS system identification," *IEEE Signal Processing Lett.*, vol. to appear, 2006.

[124] T. Kailath, *Linear Systems*, Prentice Hall, Englewood Cliffs, N.J., 1980.

[125] P. Stoica and T. Söderström, "Common factor detection and estimation," *Automatica*, vol. 33, no. 5, pp. 1985–1989, May 1997.

[126] M. Agrawal, P. Stoica, and P. Ågren, "Common factor estimation and two applications in signal processing," *Signal Process.*, vol. 84, no. 2, pp. 421–429, Feb. 2004.

[127] C. P. Hughes and A. Nikeghbali, "The zeros of random polynomials cluster uniformly near the unit circle," 2004.

[128] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., 3 edition, 1991.

[129] J. P. Reilly, M. Wilbur, M. Seibert, and N. Ahmadvand, "The complex subband decomposition and its application to the decimation of large adaptive filtering problems," *IEEE Trans. Signal Processing*, vol. 50, no. 11, pp. 2730–2743, Nov. 2002.

[130] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Prentice Hall, 3 edition, 1996.

[131] M. Hofbauer and H. Loeliger, "Limitations for FIR multi-microphone speech dereverberation in the low-delay case," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sept. 2003, pp. 103–106.

[132] Y. Haneda, S. Makino, and Y. Kaneda, "Multiple-point equalization of room transfer functions by using common acoustical poles," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 4, pp. 325–333, Jul 1997.

[133] T. Hikichi, M. Delcroix, and M. Miyoshi, "On robust inverse filter design for room transfer function fluctuations," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sept. 2006.

[134] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sept. 2006, pp. 1–4.

[135] S. Weiss, G. W. Rice, and R. W. Stewart, "Multichannel equalization in subbands," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1999, pp. 203–206.

[136] C. A. Lanciani and R. W. Schafer, "Subband-domain filtering of MPEG audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1999, vol. 2, pp. 917–920.

Appendix A

Delay-and-Sum Beamformer Performance in Dereverberation

A.1 Proof of Theorem 1

Proof. The problem is to find the expected improvement in DRR at the beamformer output compared with the DRR of the best microphone, which is normally the microphone closest to the source, i.e.

$$\mathscr{E}\{\overline{\mathrm{DRR}}\} = 10\log_{10}\left(\frac{\mathscr{E}\{\mathrm{DRR}_{\mathrm{DSB}}\}}{\mathscr{E}\{\mathrm{DRR}'\}}\right), \tag{A.1}$$

where $\mathscr{E}\{\cdot\}$ is the spatial expectation which is discussed in detail in Section 3.2, 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 delay-and-sum beamformer.

Consider an array of $M > 1$ microphones with uniform response shaping weights, $w_m = 1/M$, and steering delay filters are $e^{-j2\pi f\tau_m}$, where τ_m is a delay to compensate for propagation delays of the mth direct path. The transfer function at the output of the DSB is then

$$\tilde{H}(e^{j\omega}) = \frac{1}{M}\sum_{m=1}^{M} H_m(e^{j\omega})e^{-j2\pi f\tau_m}, \tag{A.2}$$

where $H_m(e^{j\omega})$ is the room transfer function from the source to the mth microphone.

Now results from statistical room acoustics (SRA) are employed. These are reviewed in detail in Section 3.2 and are reproduced here for convenience. In SRA it is

assumed that the room transfer function of the mth microphone can be expressed in terms of a direct and a reverberant component, i.e.

$$H_m(e^{j\omega}) = H_{d,m}(e^{j\omega}) + H_{r,m}(e^{j\omega}). \tag{A.3}$$

It is further assumed that the direct and reverberant components are uncorrelated and thus, the spatial expectation of the cross terms arising in the squared magnitude of the transfer function are zero. Consequently, the expected energy density spectrum can be written

$$\mathcal{E}\{|H_m(e^{j\omega})|^2\} = |H_{d,m}(e^{j\omega})|^2 + \mathcal{E}\{|H_r(e^{j\omega})|^2\}, \tag{A.4}$$

The direct component, $H_{d,m}(e^{j\omega})$, is given by [83]

$$H_{d,m}(e^{j\omega}) = \frac{e^{jkD_m}}{4\pi D_m} \tag{A.5}$$

and the reverberant component, $\mathcal{E}\{|H_r(e^{j\omega})|^2\}$, can be written [83]

$$\mathcal{E}\{|H_r(e^{j\omega})|^2\} = \left(\frac{1-\alpha}{\pi A\alpha}\right), \tag{A.6}$$

where α is the average wall absorption coefficient and A is total wall surface area. The spatial correlation of the reverberant components between the mth and the lth microphones is [25]

$$\mathcal{E}\{H_{r,m}(e^{j\omega})H^*_{r,l}(e^{j\omega})\} = \left(\frac{1-\alpha}{\pi A\alpha}\right)\frac{\sin k\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_l\|}{k\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_l\|}, \tag{A.7}$$

where $(\cdot)^*$ denotes the complex conjugate.

Using (A.5) and (A.6) the expected DRR of the closest microphone microphone can be written [83]

$$\begin{aligned}\mathcal{E}\{\text{ DRR}'\} &= \frac{|H_{d,m}(e^{j\omega})|^2}{E\{|H_r(e^{j\omega})|^2\}} \\ &= \frac{\alpha A}{16\pi D'^2(1-\alpha)},\end{aligned} \tag{A.8}$$

where $D' = \min_m(D_m)$ is the distance between the source and the closest microphone.

Similarly, the expected power density spectrum of the beamformer can be expressed

in terms of a direct and reverberant component as in (A.4), such that $\mathscr{E}\{|\bar{H}(e^{j\omega})|^2\} = |\bar{H}_d(e^{j\omega})|^2 + \mathscr{E}\{|\bar{H}_r(e^{j\omega})|^2\}$, where the direct component becomes

$$\begin{aligned} |\bar{H}_d(e^{j\omega})|^2 &= \left| \frac{1}{M} \sum_{m=1}^{M} H_{d,m}(e^{j\omega}) e^{-j2\pi f \tau_m} \right|^2 \\ &= \frac{1}{(4\pi M)^2} \sum_{m=1}^{M} \sum_{n=1}^{M} \frac{1}{D_m D_l}, \end{aligned} \tag{A.9}$$

and the reverberant component can be shown to be

$$\mathscr{E}\{|\bar{H}_r(e^{j\omega})|^2\} = \left| \frac{1}{M} \sum_{m=1}^{M} \mathscr{E}\{H_{r,m}(e^{j\omega})\} e^{-j2\pi f \tau_m} \right|^2 \tag{A.10}$$

$$= \left(\frac{1-\alpha}{M^2 \pi A \alpha} \right) \sum_{m=1}^{M} \sum_{n=1}^{M} \frac{\sin k\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_l\|}{k\|\boldsymbol{\ell}_m - \boldsymbol{\ell}_l\|} \cos(k[D_m - D_l]),$$

where the delay has been set to $\tau_m = D_m/c$ as in [41].

The expected direct-to-reverberant ratio of the DSB is then obtained by

$$\mathscr{E}\{\mathrm{DRR}_{\mathrm{DSB}}\} = \frac{|\bar{H}_d(e^{j\omega})|^2}{\mathscr{E}\{|\bar{H}_r(e^{j\omega})|^2\}}, \tag{A.11}$$

where $|\bar{H}_d(e^{j\omega})|^2$ and $\mathscr{E}\{|\bar{H}_r(e^{j\omega})|^2\}$ are defined in (A.9) and (A.10) respectively.

Finally, substituting (A.11) and (A.8) into (A.1), gives the result stated in (2.18).

□