Single- and multi-microphone speech dereverberation using spectral enhancement

Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 25 juni 2007 om 14.00 uur

door

Emanuël Anco Peter Habets

geboren te Maastricht

Dit proefschrift is goedgekeurd door de promotor:

prof.dr.ir. J.W.M. Bergmans

Copromotoren:
dr.ir. P.C.W. Sommen
en
dr. S. Gannot

Printed by Eindhoven University Press.

Kindly supported by Technology Foundation STW, applied science division of NWO and the Technology Programme of the Ministry of Economic Affairs.

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Habets, Emanuël A.P.

Single- and multi-microphone speech dereverberation using spectral enhancement / by Emanuël Anco Peter Habets. - Eindhoven : Technische Universiteit Eindhoven, 2007.
Proefschrift. - ISBN 978-90-386-1544-8
NUR 962
Trefw.: spraakverwerking / digitale geluidstechniek / digitale signaalverwerking / elektro-akoestiek.
Subject headings: speech enhancement / interference suppression / acoustic signal processing / reverberation.

Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement

"Essentially, all models are wrong, but some are useful."

- George E.P. Box (1919-present)

Samenstelling promotiecommissie:

prof.dr.ir. A.C.P.M. Backx (voorzitter)
prof.dr.ir. J.W.M. Bergmans (promotor)
dr.ir. P.C.W. Sommen (co-promotor)
dr. S. Gannot (co-promotor)
dr. P.A. Naylor
prof.dr.-ing. R. Martin
prof.dr.ir. A. Hirschberg
ing. C.P. Janse

Summary

Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement

In speech communication systems, such as voice-controlled systems, hands-free mobile telephones, and hearing aids, the received microphone signals are degraded by room reverberation, background noise, and other interferences. This signal degradation may lead to total unintelligibility of the speech and decreases the performance of automatic speech recognition systems.

In the context of this work reverberation is the process of multi-path propagation of an acoustic sound from its source to one or more microphones. The received microphone signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*), and reflections that arrive after the early reverberation (commonly called *late reverberation*). Reverberant speech can be described as sounding distant with noticeable echo and colouration. These detrimental perceptual effects are primarily caused by late reverberation, and generally increase with increasing distance between the source and microphone. Conversely, early reverberations tend to improve the intelligibility of speech. In combination with the direct sound it is sometimes referred to as the *early speech component*.

Reduction of the detrimental effects of reflections is evidently of considerable practical importance, and is the focus of this dissertation. More specifically the dissertation deals with dereverberation techniques, i.e., signal processing techniques to reduce the detrimental effects of reflections. In the dissertation, novel single- and multi-microphone speech dereverberation algorithms are developed that aim at the suppression of late reverberation, i.e., at estimation of the early speech component. This is done via so-called spectral enhancement techniques that require a specific measure of the late reverberant signal. This measure, called spectral variance, can be estimated directly from the received (possibly noisy) reverberant signal(s) using a statistical reverberation model and a limited amount of *a priori* knowledge about the acoustic channel(s) between the source and the microphone(s).

In our work an existing single-channel statistical reverberation model serves as a starting point. The model is characterized by one parameter that depends on the acoustic characteristics of the environment. We show that the spectral variance estimator that is based on this model, can only be used when the source-microphone distance is larger

than the so-called critical distance. This is, crudely speaking, the distance where the direct sound power is equal to the total reflective power. A generalization of the statistical reverberation model in which the direct sound is incorporated is developed. This model requires one additional parameter that is related to the ratio between the direct sound energy and the sound energy of all reflections. The generalized model is used to derive a novel spectral variance estimator. When the novel estimator is used for dereverberation rather than the existing estimator, and the source-microphone distance is smaller than the critical distance, the dereverberation performance is significantly increased.

Single-microphone systems only exploit the temporal and spectral diversity of the received signal. Reverberation, of course, also induces spatial diversity. To additionally exploit this diversity, multiple microphones must be used, and their outputs must be combined by a suitable spatial processor such as the so-called delay and sum beamformer. It is not *a priori* evident whether spectral enhancement is best done before or after the spatial processor. For this reason we investigate both possibilities, as well as a merge of the spatial processor and the spectral enhancement technique. An advantage of the latter option is that the spectral variance estimator can be further improved. Our experiments show that the use of multiple microphones affords a significant improvement of the perceptual speech quality.

The applicability of the theory developed in this dissertation is demonstrated using a hands-free communication system. Since hands-free systems are often used in a noisy and reverberant environment, the received microphone signal does not only contain the desired signal but also interferences such as room reverberation that is caused by the desired source, background noise, and a far-end echo signal that results from a sound that is produced by the loudspeaker. Usually an acoustic echo canceller is used to cancel the far-end echo. Additionally a post-processor is used to suppress background noise and residual echo, i.e., echo which could not be cancelled by the echo canceller. In this work a novel structure and post-processor for an acoustic echo canceller are developed. The post-processor suppresses late reverberation caused by the desired source, residual echo, and background noise. The late reverberation and late residual echo are estimated using the generalized statistical reverberation model. Experimental results convincingly demonstrate the benefits of the proposed system for suppressing late reverberation, residual echo and background noise. The proposed structure and post-processor have a low computational complexity, a highly modular structure, can be seamlessly integrated into existing hands-free communication systems, and affords a significant increase of the listening comfort and speech intelligibility.

Samenvatting

Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement

In spraakcommunicatiesystemen, zoals spraakherkenningssystemen, hands-free telefoons en gehoorapparaten, worden de microfoon signalen gedegradeerd door nagalm, achtergrond ruis en andere stoorbronnen. Deze degradatie van het signaal kan ertoe leiden dat de spraak geheel onverstaanbaar wordt en dat automatische spraakherkenningssystemen niet meer goed functioneren.

In de context van dit werk is nagalm het proces van het meervoudig voortplanten van een akoestisch geluid van de bron naar één of meerdere microfoons. Het microfoon signaal bestaat over het algemeen uit drie onderdelen: een rechtstreeks geluid; reflecties die kort na het directe geluid ontvangen worden (ook wel *vroege nagalm* genoemd); en reflecties die na de vroege nagalm ontvangen worden (ook wel *late nagalm* genoemd). Galmende spraak kan omschreven worden als spraak die op een afstand gehoord wordt en duidelijke echo en spectrale kleuring vertoont. Deze effecten, die de perceptie verslechteren, worden voornamelijk veroorzaakt door de late nagalm en worden gewoonlijk groter naarmate de afstand tussen bron en microfoon wordt vergroot. Anderzijds zal vroege nagalm de verstaanbaarheid van spraak verbeteren. In combinatie met het directe geluid worden ze ook wel het *vroege spraakcomponent* genoemd.

Het terugdringen van de nadelige effecten van reflecties is erg belangrijk voor praktische toepassingen en is het hoofdthema van dit proefschrift. Meer specifiek zal dit proefschrift ontgalmingstechnieken behandelen, d.w.z. signaalbewerkingstechnieken die de nadelige effecten van reflecties terugdringen. In dit proefschrift zijn nieuwe spraakontgalmingsalgoritmes ontwikkeld, uitgaande van één of meerdere microfoons. Deze algoritmes zijn gericht op onderdrukking van de late nagalm of, in andere woorden, op het schatten van het vroege spraakcomponent. Dit gebeurt via zogenoemde spectrale verbeteringstechnieken die een specifieke maat van de late nagalm vereisen. Door het toepassen van een statistisch model voor nagalm en een beperkte voorkennis over het akoestische kanaal (kanalen) tussen de bron en microfoon(s) kan deze maat, spectrale variantie genoemd, direct worden afgeleid van het ontvangen galmende (mogelijk ruis bevattende) signaal (of signalen).

In ons werk dient een bestaand enkel-kanaals statistisch model voor nagalm als startpunt. Het model wordt gekarakteriseerd door een parameter die afhankelijk is van de

akoestische eigenschappen van de omgeving. We laten zien dat de spectrale variantie schatter, die gebaseerd is op dit model, alleen gebruikt kan worden indien de bron-microfoon afstand groter is dan de zogenaamde kritische afstand. Dit is grofweg de afstand waarbij het vermogen van het directe geluid gelijk is aan het vermogen van alle reflecties tezamen. Eveneens is een generalisatie ontwikkeld van het statistische model voor nagalm waarin het directe geluid is opgenomen. Dit model vereist een extra parameter die gerelateerd is aan de verhouding tussen de energie van het directe geluid en de energie van alle reflecties. Het gegeneraliseerde model wordt gebruikt om een nieuwe schatter voor de spectrale variantie af te leiden. Als de bron-microfoon afstand kleiner is dan de kritische afstand, dan wordt bij toepassing van de nieuwe schatter de ontgalming duidelijk verbeterd ten opzichten van de bestaande schatter.

Systemen die gebruik maken van slechts één microfoon benutten alleen temporele en spectrale diversiteit. Nagalm vertoont echter ook spatiële diversiteit. Om ook deze diversiteit te benutten, moeten meerdere microfoons worden gebruikt. De output van deze verschillende microfoons moet dan gecombineerd worden door een passende spatiële processor, zoals de zogehete delay-and-sum beamformer. Het is niet zonder meer duidelijk of spectrale verbetering van voor of na de spatiële processor zou moeten worden toegepast. Daarom onderzoeken we beide mogelijkheden, alsook een combinatie van de spatiële processor en de spectrale verbeteringstechniek. Een voordeel van deze laatste optie is dat de spectrale variantie schatter nog verder verbeterd kan worden. Onze experimenten tonen aan dat het gebruik van meerdere microfoons een significante verbetering van de perceptuele spraakkwaliteit tot gevolg heeft.

De toepasbaarheid van de in dit proefschrift ontwikkelde theorie wordt gedemonstreerd aan de hand van een hands-free communicatiesysteem. Daar hands-free systemen vaak gebruikt worden in een lawaaiige en galmende omgeving, bevat het ontvangen microfoon signaal niet alleen het gewenste signaal, maar ook verstoringen. We hebben het dan over verstoringen zoals nagalm die veroorzaakt wordt door de gewenste bron, achtergrond ruis en een far-end echo signaal dat het gevolg is van het geluid dat wordt geproduceerd door de luidspreker. Normaal gesproken wordt een akoestische echo onderdrukker gebruikt om de far-end echo te onderdrukken. Daar wordt dan een post-processor aan toegevoegd om achtergrond ruis en residu echo, ofwel echo die niet voldoende gereduceerd kon worden door de akoestische echo onderdrukker, te onderdrukken. In dit werk hebben we een nieuwe structuur en post-processor voor een akoestische echo onderdrukker ontwikkeld. De post-processor onderdrukt zowel late nagalm veroorzaakt door de gewenste bron, residu echo en achtergrond ruis. De late nagalm en de residu echo worden geschat door gebruik te maken van het gegeneraliseerde statistische model voor nagalm. Experimentele resultaten demonstreren duidelijk de voordelen van het voorgestelde systeem ter onderdrukking van late nagalm, residu echo en achtergond ruis. De ontwikkelde oplossing heeft een lage rekencomplexiteit, een sterk modulaire structuur, kan naadloos geïntegreerd worden in bestaande hands-free communicatiesystemen en staat een significante verbetering van het luistercomfort en de spraak verstaanbaarheid toe.

Glossary

List of Acronyms

AEC	Acoustic Echo Canceller
AIR	Acoustic Impulse Response
Alcons	articulation loss for consonants
AP	Affine Projection
AR	Auto Regressive
ASR	Automatic Speech Recognition
ATF	Acoustic Transfer Function
BEM	Boundary Element Method
BSD	Bark Spectral Distortion
CMN	Cepstral Mean Normalization
CMVN	Cepstral Mean and Variance Normalization
DOA	Direction of Arrival
DRR	Direct to Reverberation Ratio
EDC	Energy Decay Curve
EDR	Energy Decay Relief
ELR	Early to Late reverberation Ratio
ERLE	Echo Return Loss Enhancement
FDTD	Finite-Difference Time-Domain
FEM	Finite Element Method
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GSC	Generalised Sidelobe Canceller
HERB	Harmonicity based dEReverBeration
HMM	Hidden Markov Model
HMP	Hidden Markov Process

IFFT	Inverse Fast Fourier Transform
IIR	Infinite Impulse Response
IMCRA	Improved Minima Controlled Recursive Averaging
IPNLMS	Improved Proportionate NLMS
ITU-T	International Telecommunications Union
KLT	Karhunen-Loeve transform
LCMV	Linear Constrained Minimum Variance
LEM	Loudspeaker Enclosure Microphone
LIME	LInear-predictive Multi-input Equalization
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LSA	Log Spectral Amplitude
LSD	Log Spectral Distortion
LTI	Linear Time-Invariant
MAP	maximum a posteriori
MBSD	Modified Bark Spectral Distortion
MCLT	Modulated Complex Lapped Transform
MFCC	Mel-frequency cepstral coefficient
MI	Modulation Index
MIMO	Multi-Input Multi-Output
MLS	Maximum Length Sequence
MOS	Mean Opinion Score
MSC	Mean Square Coherence
MTF	Modulation Transfer Function
MVDR	Minimum Variance Distortionless Response
NLMS	Normalized Least Mean Square
OM-LSA	Optimally-Modified Log Spectral Amplitude
PDF	probability distribution function
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
RLS	Recursive Least Squares
SEA	Statistical Energy Analysis
SIMO	Single-Input Multi-Output
SIR	Signal to Interference Ratio
SISO	Single-Input Single-Output
SNR	Signal to Noise Ratio

SRA	Statistical Room Acoustics
SRR	Signal to Reverberation Ratio
STFT	short-time Fourier transform
STI	Speech Transmission Index
TF	Time-Frequency
TF-GSC	Transfer Function Generalized Sidelobe Canceller
WER	Word Error Rate
WGN	White Gaussian Noise
XOR	exclusive-or

Notations

x	scalar quantity
$\mathbf{x}$	vector quantity
$\mathbf{X}$	matrix quantity
$x(n)$	function of a discrete variable n
$x(t)$	function of a continuous variable t
x_n	function of a finite discrete variable n

Operators

$x * y$	linear convolution
x^*	complex conjugate of x
$\mathbf{x}^T$	non-conjugate vector transpose
$\|x\|$	absolute value of x
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
$\mathrm{Re}\{\cdot\}$	imaginary component of a complex number
$\mathrm{Im}\{\cdot\}$	real component of a complex number
$\mathcal{E}\{\cdot\}$	mathematical expectation
$\mathcal{E}_\theta\{\cdot\}$	spatial expectation

Symbols and Variables

A	equivalent wall absorbtion coefficient

$\bar{a}$	Sabine's absorption coefficient
$\bar{\alpha}$	average absorption coefficient
c	speed of sound in m/s
$r_{xx}(t, t+\tau)$	auto-correlation of signal x
$\bar{\delta}$	average damping constant
ρ_0	density of the propagation medium at equilibrium
f_s	sampling frequency in Hz
f_g	Schroeder frequency
$H(\omega)$	Acoustic Transfer Function
$h(t)$	Acoustic Impulse Response
M	number of microphones
$\mathbb{V}_s$	source volume
$\mathbf{r}$	receiver position (x, y, z)
$\mathbf{r}_s$	source position (x_s, y_s, z_s)
RT_{60}	reverberation time
S	total wall surface
$\gamma(l, k)$	*a posteriori* Signal to Interference Ratio
$\xi(l, k)$	*a priori* Signal to Interference Ratio
$X(l, k)$	spectral signal component

Contents

Chapter 1

Introduction

1.1 Scope and Motivation

The work presented in this dissertation is motivated by the rapidly growing market of speech communications systems. Typical speech communication systems are hands-free (mobile) telephones, voice-controlled systems, and hearing aids. The main user benefit of hands-free telephones is that they enable the user to walk around freely without wearing a headset or a microphone, and hence provide a natural way of communication. Voice-controlled systems are, for example, used in an operating room where they allow surgeons and nurses to freely move around the patient. Obviously, the main benefit of hearing aid applications is to increase the hearing capacity, enabling a hearing-aid user to interact better with other people. In all these examples, the desired acoustical source can be positioned at a considerable distance from the microphone (see Fig. 1.1). As illustrated in Fig. 1.1, the desired source produces sound waves. Some of these wave travel directly to the microphone. The resulting direct signal can be degraded by reverberation, background noise, and other interferences.

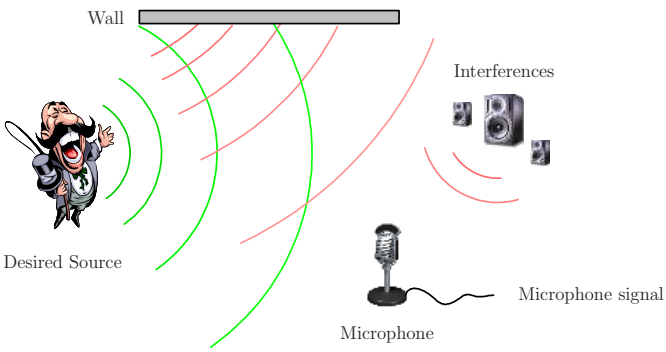


Figure 1.1 *Illustration of a desired source, a microphone, and interfering sources.*

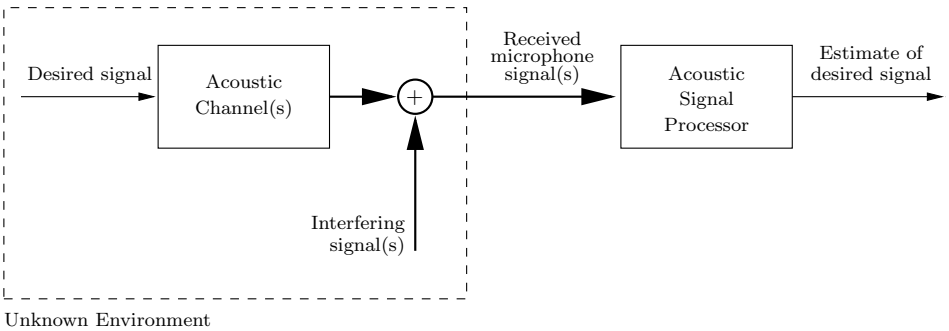


Figure 1.2 *Application of acoustic signal processing concerned with the estimation of a desired signal.*

To counteract the degradations caused by reverberation, background noise and other interferences, high-performance acoustic signal processing techniques are required. In the context of this work reverberation is the process of multi-path propagation of an acoustic sound from its source to one or more microphones. Sound is a disturbance of mechanical energy that propagates through matter, e.g., a gas, as a wave. Under the influence of a sound wave, variations of gas density and pressure occur, both of which are functions of time and position. The difference between the instantaneous pressure and and the static pressure is called the sound pressure. In this dissertation a microphone is used to transform the pressure (or pressure gradient) present in the air immediately in front of the microphone into an electrical signal. For simplicity we will assume that the microphone is ideal, i.e., that its electrical output is identical (except for a non-dimensionless scaling factor) to the local sound pressure. For this reason we will not distinguish between them in this dissertation. A block diagram which describes an application of acoustic signal processing is illustrated in Fig. 1.2. Here the sound that is produced by the desired source, designated as the *desired signal* or the *anechoic signal*, is 'transmitted' over the *acoustic channel(s)*, and in combination with the interfering signal(s) it results in the received microphone signal(s). The thick lines in Fig. 1.2 denote one or more signals, whereas the thin lines denote one signal. The interfering signals can either describe interfering sounds or electrical interferences, such as sensor noise. The received microphone signal(s) are then processed using the acoustic signal processor to estimate the desired signal.

A major challenge in acoustic signal processing originates from the degradation of the desired signal by the acoustic channel within an enclosed space, e.g., an office room or living room. Because the microphone cannot always be located near the desired source, the received microphone signals are typically degraded by (i) reverberation introduced by the multi-path propagation of the desired sound to the microphones and (ii) noise introduced by interfering sources. While state-of-the-art acoustic signal processing algorithms are available to reduce noise, the development of practical algorithms that can reduce the degradations caused by reverberation has for a long time been one of the 'holy grails'. The main difference between noise and reverberation is that the degrading component in case of reverberation is *dependent* on the desired signal, whereas in case of noise it can be assumed to be *independent* of the desired signal. It

should be noted that many, if not all, existing acoustic signal processing techniques fail completely or experience a dramatically reduced performance when reverberation is present, e.g., existing source localization and source separation techniques.

Reverberant speech can be described as sounding distant with noticeable colouration and echo. These detrimental perceptual effects generally increase with increasing distance between the source and the microphone. Furthermore, with the spread in the time of arrival of reflections at the microphone, reverberation causes blurring of speech phonemes. These detrimental effects seriously degrade the intelligibility, the performance of voice-controlled systems, and the performance of speech coding algorithms that are used in telephone systems. Reduction of these detrimental effects is evidently of considerable practical importance, and is the focus of this dissertation. The algorithms that reduce these detrimental effects are called *speech dereverberation algorithms*.

To reduce the effects of reverberation by means of acoustic signal processing the physical properties of reverberation need to be understood. Therefore, reverberation in enclosed spaces is discussed in Section 1.2. From the above discussion it is evident that reverberation degrades speech intelligibility and the performance of automatic speech recognition systems. In order to develop effective algorithms which counteract the degrading affect of reverberation it is important to know how reverberation effects the speech intelligibility and automatic speech recognition. This will be discussed in Sections 1.3, and 1.4, respectively. The problem statement will then be formulated in Section 1.5. In Section 1.6 the outline of this dissertation is given.

1.2 Reverberation in Enclosed Spaces

Reverberation, a central theme of this dissertation, is intuitively described by the concept of reflections. The desired source produces wavefronts, which propagate outward from the source. The wavefronts reflect off the walls of the room and superimpose at the microphone. In Fig. 1.3 this is illustrated with an example of a direct path and a single reflection. Due to differences in the lengths of the propagation paths to the microphone and in the amount of sound energy absorbed by the walls, each wavefront arrives at the microphone with a different amplitude and phase. The term reverberation designates the presence of delayed and attenuated copies of the source signal in the received signal.

Reverberation is the process of multi-path propagation of an acoustic signal from its source to the microphone. The received signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*), and reflections that arrive after the early reverberation (commonly called *late reverberation*). The combination of the direct sound and early reverberation is sometimes referred to as the *early sound component*. The different sound components will now be discussed in more detail.

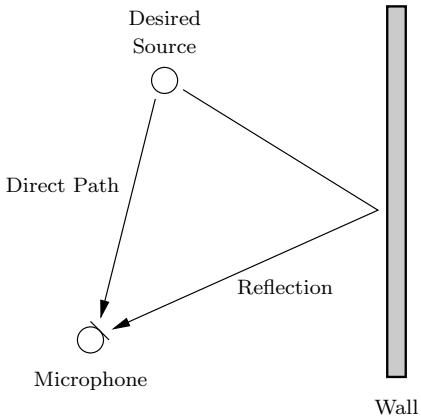


Figure 1.3 *Illustration of the direct path and a single reflection from the desired source to the microphone.*

Direct Sound The first sound that is received through free-field, i.e., without reflection, is the direct sound. In case the source is not in line of sight of the observer there is no direct sound. The delay between the initial excitation of the source and its observation is dependent on the distance and the velocity of the sound.

Early Reverberation A little time later the sounds which were reflected off one or more surfaces (walls, floor, furniture, etc.) will be received. These reflected sounds are separated in both time and direction from the direct sound. The reflected sounds form a sound component which is usually called *early reverberation*. Early reverberation will vary as the source or the microphone moves within the space, and gives us information about the size of the space and the position of the source in the space. Early reverberation is not perceived as a separate sound to the direct sound so long as the delay of the reflections does not exceed a limit of approximately 80-100 ms with respect to the arrival time of the direct sound. Early reverberation is actually perceived to reinforce the direct sound and is therefore considered useful with regard to speech intelligibility. This is often referred to as the *precedence effect*. This reinforcement is what makes it easier to hold conversations in closed rooms compared with outdoors. Early reverberation is mainly important in so-called small-room acoustics since the walls, ceiling and floor are really close. Early reverberation also causes a spectral distortion called colouration.

Late Reverberation Late reverberation results from reflections which arrive with larger delays after the arrival of the direct sound. They are perceived either as separate echoes, or as reverberation, and impair speech intelligibility.

The acoustic channel between a source and a microphone can be described by an Acoustic Impulse Response (AIR). This is the signal that is measured at the microphone in response to a source that produces a 'sound impulse'. The AIR can be divided into three segments, the *direct path*, *early reflections*, and *late reflections*, as

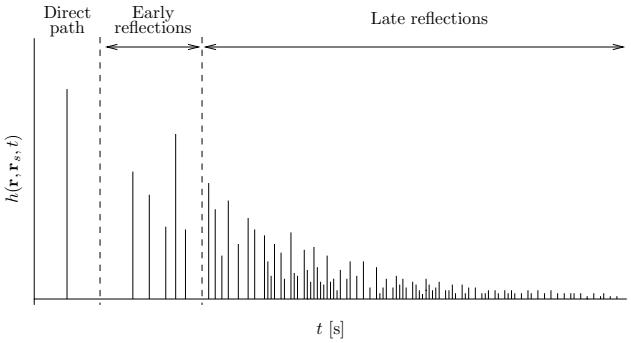


Figure 1.4 *A schematic representation of an acoustic impulse response.*

illustrated in Fig. 1.4. The convolution of these segments with the desired signal results in the direct sound, early reverberation, and late reverberation, respectively. From a signal processing perspective, early reflections appear as separate delayed impulses in the AIR, whilst late reflections appear as a continuum. Furthermore, it is important to note that the energy of the reflections decays at an exponential rate. This exponential decay is a well-known property of the AIR, which has motivated the notion of *reverberation time*. The reverberation time quantifies the severity of reverberation within a room, and is denoted by RT_{60}. It is defined as the time that is necessary for a 60 dB decay of the sound energy after switching off a sound source. A detailed discussion of the reverberation time can be found in Chapter 2, on page 44.

The time- and space-variant AIR $h(\mathbf{r}, \mathbf{r}_s, t, t')$ is defined as the response of the acoustic channel between the source at position $\mathbf{r}_s$ and the microphone at position $\mathbf{r}$ at time instant t due to a unit impulse applied at time t'. The observed signal at position $\mathbf{r}$ at time t is then given by

$$z(\mathbf{r}, t) = \int_{-\infty}^{\infty} \int_{\mathbb{V}_s} h(\mathbf{r}, \mathbf{r}_s, t, t') s(\mathbf{r}_s, t') \, \mathrm{d}\mathbf{r}_s \, \mathrm{d}t', \tag{1.1}$$

where $s(\mathbf{r}_s, t')$ denotes the source signal at position $\mathbf{r_s}$, time t', and $\mathbb{V}_s$ denotes the source volume.

The Fourier transform of the AIR at time t is called the Acoustic Transfer Function (ATF) and is denoted by $H(\mathbf{r}, \mathbf{r}_s, t; \omega)$, where ω denotes the angular frequency. The ATF defines the frequency response of the system relating the sound source to the sound pressure at the microphone, and is probably the most frequently used function to describe an acoustic channel. For reverberant environments it looks like a random function, which cannot be predicted in advance without detailed knowledge of the acoustic and geometric parameters of the room. The degree of randomness can be characterized by the *spectral deviation*, denoted by σ, as defined in Chapter 2, Section 2.7.3. Even though a reverberant sound field possesses an inherent randomness, it also possesses an underlying structure. Insight into the structure of the ATF can be obtained using the acoustic wave equation, which governs the propagation of acoustic waves through a material medium. There are various room acoustic models for the

ATF which will be discussed in this Chapter 2. Most of these models, e.g., all-zero, all-pole and zero-pole, depend on hundreds and sometimes thousands of parameters. Since the acoustic channels in real rooms are too complex to model explicitly, Statistical Room Acoustics (SRA) is often used. SRA provides a statistical description of the ATF and AIR in terms of a few key quantities, e.g., source-microphone distance, and reverberation time.

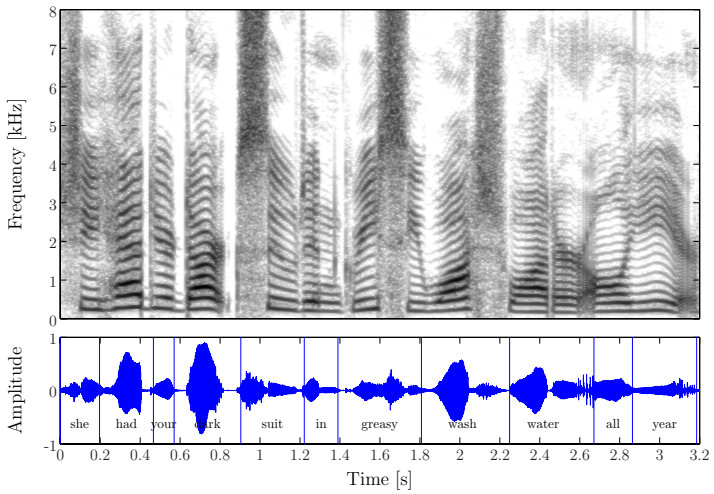
In most cases it is not feasible to measure the ATF during operation of the acoustic signal processing algorithm since the desired signal is unknown at the receiver side. In some cases the desired signal is known at the receiver side, such that the ATF can be calculated. Another problem is that the ATF changes rapidly as one moves away from the original point of measure [1, 2, 3], and that it is sensitive to source position, temperature, the positioning of room furnishings, and movements in the room. Therefore, even if it is feasible to measure the ATF in real-time, frequent re-measurement may be needed.

If the distance between the source and the microphone changes, the energy which is related to the direct path changes, while the combined energy of the early and late reflections is approximately constant. The distance at which the direct path energy is equal to the combined energy of the early and late reflections is called the *critical distance*. It should be understood that if the distance between a source and a microphone is larger than the critical distance, then the reflective energy is larger than the direct path energy.

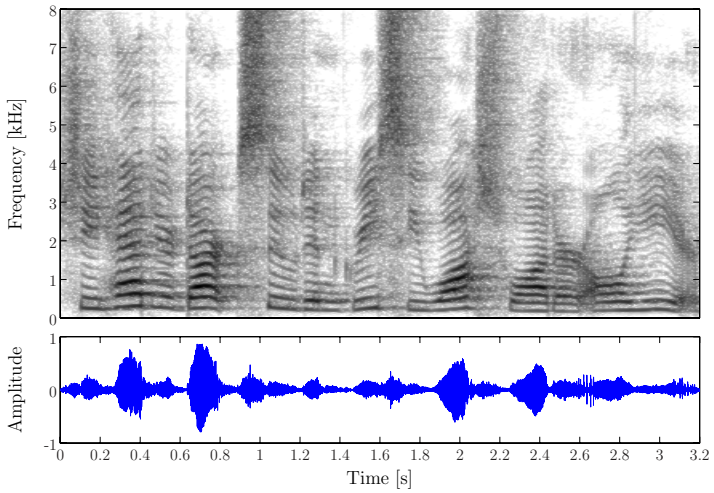
1.3 Effects of Reverberation on Speech Perception

Reverberant speech can be described as sounding distant with noticeable echo and colouration. The effects of reverberation on speech are clearly audible, and visible in the spectrogram and waveform of a speech signal. In Fig. 1.5(a) the spectrogram and waveform (including transcript) of an anechoic speech signal are depicted. The speech signal was taken from the TIMIT speech database [4]. The speech formants, which are defined as the resonance frequencies associated with the vocal tract [5], can clearly been seen in the spectrogram. It can also be seen that the phonemes are well separated in time. The anechoic signal of Fig. 1.5(a) was transmitted in an office room and its response was measured at a distance of 0.5 m from the source. The spectrogram and waveform of the received signal are shown in Fig. 1.5(b). The distortion of the speech signal that is caused by the acoustic channel is clearly visible. Blurring of the speech formants is visible in the spectrogram and the smearing of the phonemes in time is visible in both the spectrogram and the waveform. Due to this smearing the empty spaces between words and syllabi are filled by reverberation, and subsequent phonemes overlap. These distortions result in an audible difference between the anechoic and the reverberant speech, and degraded speech intelligibility and fidelity.

For the development of effective dereverberation algorithms it is of great importance to have a good understanding of the effects of reverberation on speech perception.



(a) Spectrogram (top) and waveform (bottom) of an anechoic speech signal.



(b) Spectrogram (top) and waveform (bottom) of the measured reverberant speech signal.

Figure 1.5 *Spectrograms and waveforms of (a) an anechoic speech signal (FAKS0:SA1) taken from the TIMIT speech database [4], and (b) the reverberant version of this measured at a distance of 0.5 m in an office room with a reverberation time of 0.5 s.*

Therefore, we will first investigate which physical properties of an enclosed space determine the speech intelligibly and quality (Section 1.3.1). Reverberation is well understood in terms of physical acoustics, but it is not yet well understood how the afore mentioned distortions of the speech affect the intelligibility [6]. In Section 1.3.2

we will describe two factors that contribute to the reduction of speech intelligibility, viz., self-masking and overlap-masking. For normal listeners in small reverberant rooms the distortion caused by reverberation seems to go largely unnoticed. The underlying reason is discussed in Section 1.3.3 and is used later in the development of dereverberation algorithms.

1.3.1 Speech Intelligibility and Quality

The physical properties of the enclosed space as well as the location of the source and the listener within the space have a large influence on the reverberation [7]. Reverberation and background noise cause noticeable changes in the speech quality and determine speech intelligibility in an enclosed space. It would be convenient to assume that reverberation solely reduces intelligibility, but this assumption is incorrect [8]. Acoustic engineers often consider reflections desirable since they increase the amplitude of the signal reaching a listener. This increase in amplitude can increase speech intelligibility if it raises the speech level above the ambient noise levels [9]. The work of Lochner and Burgers [10] demonstrates this effect for single reflections, and more recently, Watkins and Holt [11] demonstrate this effect for complex early reflections. Early reflections do not enhance intelligibility when the sound pressure level in the anechoic and reverberant recordings is equal.

The integration property, commonly called inertia, of the human auditory system lead to the integration of the early reverberation and the direct sound, and increases the apparent strength of the direct sound. This property was reported in 1935 by Aigner and Strutt [12]. They were the first to suggest an acoustic-energy-ratio based measure to quantify the effects of background noise and room acoustics on speech intelligibility. They called their measure the impression Q, which is given by

$$Q = \frac{E_{\mathrm{d}} + E_{\mathrm{e}}}{E_{\mathrm{l}} + E_{\mathrm{n}}}, \tag{1.2}$$

where E_{d} is the direct sound energy, E_{e} is the early part of the reflected sound energy, which in this case is defined as coming to the ear not later than 60 milliseconds after the direct sound, E_{l} is the late part of reflected sound energy coming later than 60 milliseconds, and E_{n} is the noise energy. Aigner and Strutt went further by putting a lower threshold of $Q = 1$ for a satisfactory *sound impression*. Eq. 1.2 essentially forms the basis of most speech intelligibility metrics that were later developed. It is important to note that the sound impression can be improved by reducing either E_{l} or E_{n}.

Consonants play a much more significant role in speech intelligibility than vowels. If the consonants are heard clearly, the speech can be understood more easily. In 1971 Peutz [13] proposed a measure called articulation loss for consonants (Alcons) which quantifies the reduction in perception of consonants due to reverberation. The calculation of the measure depends on the distance between the source and the microphone. The articulation loss can be decreased, i.e., the speech intelligibility can be increased,

by either decreasing the source-microphone distance or the reverberation time, and by increasing the room volume.

In 1980 Berkley investigated the perception of speech based on application of a room simulation program providing a collection of well-controlled realistic room responses with different room acoustic parameters [14]. Using the results obtained from listening tests Berkley concluded that the perception of reverberation is mainly based on a two-dimensional perceptual space. The two components are *colouration* and *echo*. Berkley showed that the spectral deviation σ, was well correlated with the subjective perception of this colouration component. The echo component is directly related to the reverberation time RT_{60}. Note that the amount of late reverberation is increased when RT_{60} is increased.

In 1982 Allen [15] reported a formula to predict the *subjective preference* of reverberant speech. Their main result is given by the equation

$$P = P_{\max} - \sigma \mathrm{RT}_{60}, \tag{1.3}$$

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

Jetzt [16] showed that the spectral deviation σ is related to the Direct to Reverberation Ratio (DRR), which is defined as the direct path energy (E_{d}) divided by the total reflective energy ($E_{\mathrm{e}} + E_{\mathrm{l}}$). It should be noted that within the same room σ is approximately constant, and reaches its maximum asymptotic value if the source-microphone distance is larger than the critical distance, which is defined as the distance at which the direct path energy is equal to the total reflective energy (see Section 2.7.1). When the source-microphone distance is smaller than the critical distance the spectral deviation σ can be used to determine the DRR. In the same room shorter source-microphone distances result in higher DRR, and less spectral deviation and thus colouration in case the source-microphone distance is smaller than the critical distance.

From the above discussion it can be concluded that late reverberation and noise are the main causes of the degradation in speech intelligibility. Furthermore, the perceptual speech quality, which is related to the subjective preference and sound impression, is related to two physical properties of reverberation, i.e., colouration and reverberation time. It should be noted that these properties are not independent since the amount colouration depends on the reverberation time, room volume, and source-microphone distance. The reverberation time RT_{60} is not only important from a perceptual point of view but it also characterizes the 'shape' of the AIR, as discussed in Section 1.2. Therefore, the reverberation time RT_{60} is an important measure that plays a crucial role in our work. These insights will be vital to the work presented in this dissertation.

1.3.2 Overlap-Masking and Self-Masking

Reverberation is well understood in terms of physical acoustics, but it is not yet well understood how reverberation affects the speech intelligibility. In order to develop effective speech dereverberation techniques it is vital to understand how reverberation affects the intelligibility.

The reduction in speech intelligibly caused by late reverberation is especially noticeable for listeners with non-native speakers [17], and hearing impairments [18]. The reason for this reduction is not entirely known [6]. Bolt and MacDonald [19] and Nábělek et al. [20] propose two contributing factors to the degradation of reverberant speech: *self-masking* and *overlap-masking*.

Self-masking refers to the time and frequency alterations of an individual phoneme [19, 20]. Reverberation slows sound onsets and decays of transient sounds. For example, the sound of an isolated /t/, which is basically a transient noise burst, becomes less abrupt in the presence of reverberation [6]. Furthermore, due to the temporal smearing caused by reverberation the formant transitions between vowels are disrupted. These disruptions reduce the phonetic information that is required for identification.

Overlap-masking occurs when a preceding phoneme and its reflections mask a subsequent phoneme [19, 20]. An example of overlap-masking is two phonemes with similar or different frequency content occurring sequentially with a brief delay between them. Because of reverberation, the initial phoneme will endure and may overlap the second phoneme and its associated reverberation. This overlap-masking impoverishes the second phoneme. This masking effect can, for example, be seen in Fig. 1.5, where the /a/ and the /sh/ sounds of the word 'wash' at $t = 2$ s are well separated in the anechoic signal but overlap in the reverberant signal.

1.3.3 Binaural Intelligibility Advantage

The distortion caused by reverberation in small rooms seems to go largely unnoticed by normal listeners. Furthermore, there is a difference in speech intelligibility between monaural (meaning 'one ear') and binaural (meaning 'two ears') listening. Most listeners benefit from binaural listening when reverberation exists. This indicates that the listeners binaural system processes the two signals to reduce reverberation. Binaural listening enables the auditory system to 'work out' the distance and the direction of sound sources, and to detect certain sounds at much lower intensity levels than if only one ear is used. Reverberation induces spatial diversity, i.e., the direct sound and the reflections arrive from different directions. Spatial diversity is apparently exploited when two ears are used. This diversity can also be exploited by acoustic signal processing algorithms via a spatial processor which combines multiple microphone signals. In this work we also pursue this option in Chapter 5.

Reverberation may be assumed to act as uncorrelated noise that masks subsequent reverberant phonemes. Libbey and Rogers [6] investigated the possibility that the binaural system suppresses uncorrelated reverberation received at each ear. They found

that the binaural intelligibility advantage using reverberation-like noise is not as large as in real reverberation. This demonstrates that only a portion of the total binaural intelligibility advantage is caused by the fact that reverberation acts as uncorrelated noise at each ear. The latter fact is used in the development of our acoustic signal processing algorithms. Since reverberation acts as uncorrelated noise we can use the coherency between two microphone signals to distinguish between the direct sound and the reverberation.

1.4 Effects of Reverberation on Automatic Speech Recognition

The performance of Automatic Speech Recognition (ASR) systems relies on the quality of the speech input. While reasonable recognition performance is commonly achieved when the source-microphone distance is small, the performance tends to decrease rapidly when this distance increases. The main problem is that the Signal to Noise Ratio (SNR) and DRR decrease when the distance increases.

Automatic speech recognition systems can be divided into two groups: *isolated-word recognition*, and *continuous speech recognition*. Isolated-word recognition systems assume that words are uttered in a discrete manner so that there are silences at the beginning and the end of each word. Continuous speech recognition systems are able to process continuous speech. The later systems are more complex because word boundaries are not known *a priori* and are often ambiguous. This ambiguity is increased when reverberation is present, and can have a negative influence on the recognition performance.

Fig. 1.6 shows a block diagram of a typical speech recognition system. First, the speech signal is pre-processed, to reduce distortion caused by, for example, lip radiation, background noise, and reverberation. Secondly, feature vectors are extracted from the pre-processed speech signal using short-time segments of the speech signal. These feature vectors are meant to characterize the essential information present in the speech signal. Based on these feature vectors the most likely text is found by the decoder using two types of knowledge sources, viz., acoustic knowledge and linguistic knowledge. The acoustic model contains the acoustic knowledge that is required to be able to decode the features into words or phonemes, and the language model contains linguistic knowledge that is required to decode these words or phonemes into text. The acoustic and linguistic knowledge is acquired in a training phase that is required prior to the decoding step.

The influence of reverberation on the performance of a state-of-the-art speech recognition system is shown in Fig. 1.7. For this recognition experiment, the speaker-independent large–vocabulary continuous speech recognition system was used that has been developed at the ESAT-PSI speech group of the K.U.Leuven[1]. An overview

[1]Special thanks to prof. H. Van hamme, dr. J. Duchateau, and dr. K. Eneman.

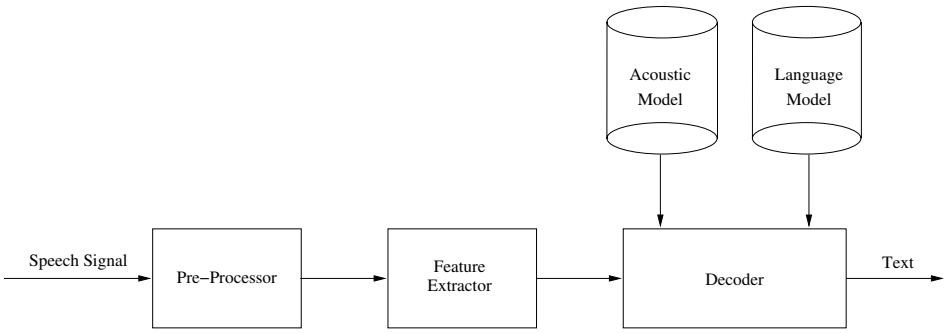
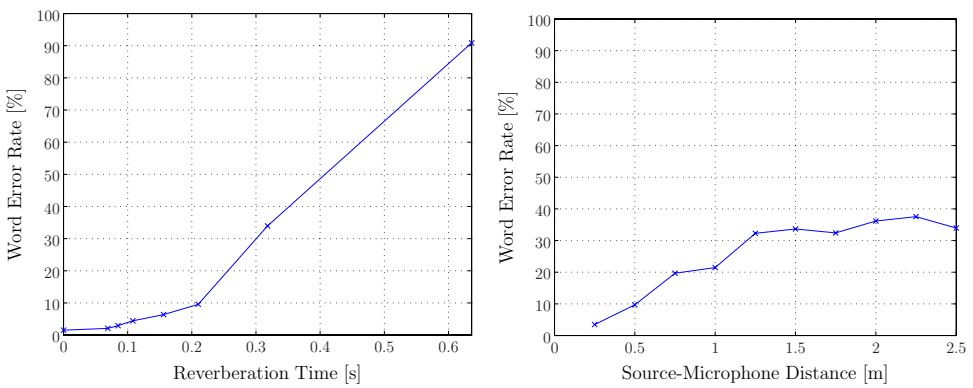


Figure 1.6 *Flow diagram of an Automatic Speech Recognition system.*



(a) Fixed source-microphone distance ($D = 3$ m). (b) Fixed reverberation time ($\mathrm{RT}_{60} = 0.32$ s).

Figure 1.7 *Speech recognition performance in a reverberant environment.*

of the acoustic model that was used can be found in [21] and the decoder is described in [22]. The reverberant signals were generated by convolving the anechoic speech signals from the test set, taken from the Wall Street Journal speech corpus, with synthetic acoustic impulse responses. These responses were generated using the room impulse response generator described in Appendix A. In Fig. 1.7(a) the Word Error Rate (WER) is shown for various reverberation times and a fixed source-microphone distance of 3 m. These results demonstrate that the error rate increases rapidly for reverberation times larger than 0.2 s. In Fig. 1.7(b) the WER is shown for various distances and a fixed reverberation time of 0.32 s. Note that the WER increases with increasing source-microphone distance.

From this simple example it is clear that the effects of reverberation on the ASR system are rather severe. Compensation for the difference between the signal that was used for training of the acoustic model and the received microphone signal can be achieved either by pre-processing of the signal or by post-precessing of the feature vectors, or both. To be able to develop an efficient and robust speech dereverberation algorithm

that can be used in conjunction with an ASR system it is important to know which speech degradations affect the speech recognition performance. Therefore, we will now briefly discuss the pre-processor, feature extractor, and decoder.

1.4.1 Pre-Processor

The received speech signal can be enhanced prior to the feature extraction step using a pre-processor to increase the recognition performance.

To compensate for the effect of the low-pass nature of voiced speech, a pre-emphasis filter is commonly used in the first pre-processing step. The low-pass nature of voiced speech is usually associated with the low-pass nature of the voiced excitation that is produced by the glottal source. The lip radiation, which is commonly modelled using a first-order high-pass filter [23], partially compensates for the low-pass nature of the voiced excitation. The pre-emphases filter compensates for the remaining low-pass nature of the voiced excitation by emphasizing the high frequency components and attenuating the low frequency components. The pre-emphasis filter is commonly implemented using a simple high-pass filter.

Other acoustic signal processing techniques can be used to compensate for the distortions caused by the acoustic channel and by noise. Noisy environments, such as cafeteria or car interiors, can severely degrade the recognition rate of speech recognition systems, sometimes rendering these systems useless. Current acoustic signal processing techniques are not able to properly cope with impulsive and non-stationary noise, whereas quite successful techniques have been developed for slow-varying or stationary noise. The so-called spectral enhancement methods are the most popular techniques for noise reduction, mostly because of their simplicity and effectiveness. Unlike noise, reverberation is correlated with the anechoic speech signal. Although many speech dereverberation algorithms are available in the literature, up to now, only few of these algorithms have been 'successfully' used as a pre-processing step for automatic speech recognition [24, 25]. The acoustic signal processing algorithm that is developed in this work can also be used as a pre-processor for an ASR system.

1.4.2 Feature Extractor

For automatic speech recognition, the acoustic signal needs to be parameterized to extract the speech information it contains. The parameters are described in the form of so-called features. The features are calculated from the received, and possibly pre-processed, speech signal. Since speech varies over time, it is more appropriate to analyse the speech signal in short time intervals where the signal is more stationary. The features are computed from short-time speech segments of 20 to 30 ms with an overlap of 50 to 75%. Frequently used features are related to cepstral coefficients, which are obtained by calculating the inverse Fourier transform of the logarithm of the spectrum of a short-time speech segment. The lower-order cepstral coefficients

represent the vocal tract impulse response. In an effort to take auditory characteristics into consideration the Mel-frequency cepstral coefficients (MFCC) were proposed [26]. These coefficients are calculated from a so-called Mel frequency scale rather than the short-term spectrum of a speech segment. The Mel frequency scale is closely related to the frequency scale of the human auditory system. The time derivatives of the MFCC are usually appended to the feature vector to capture the dynamics of speech. A speech utterance is then represented as a sequence of these feature vectors. It is important to note that most features are derived from the short-term amplitude spectrum of the short-time speech segments, and the short-term phase spectrum is disregarded.

Features can be post-processed to improve recognition performance in adverse environments using additional techniques:

1. One technique, which is often used for robust speech recognition, is applied to the cepstral coefficients and is called Cepstral Mean Normalization (CMN) [26]. CMN is used for removing short-term invariant linear channel distortion in speech signals. Convolutional distortions caused by early reflections and different microphones result in an additive offset of the cepstral coefficients. This offset can be determined by estimating the mean for each cepstral coefficient from a sequence of cepstral coefficients. Subtracting the mean from the distorted cepstral coefficients will provide an estimate of the undistorted cepstral coefficients. It has been observed that CMN produces robust features for the distortions caused by early reflections [27]. Although the CMN technique is simple and fast, its effectiveness is limited to the short-term invariant linear channel distortion caused by early reflections, while late reflections cannot be properly handled.

2. Cepstral Mean and Variance Normalization (CMVN) [28] is a simple and frequently used technique for improving the robustness in speech recognition. The mean and variance for each cepstral coefficient are estimated from a finite sequence of cepstral coefficients. First, the mean is subtracted from each cepstral coefficient, as in CMN. Secondly, each cepstral coefficient is scaled independently in order to have unity variance. Due to the subtraction and scaling the first and second moments of the feature distributions are forced to be the same for both the training and test conditions. Therefore, the mismatch between these conditions is reduced.

By post-processing the feature vectors using above techniques the distance between the feature vectors that are used in the training phase and the feature vectors that are calculated from the noisy and reverberation speech signal can be reduced. Therefore, the recognition performance of the ASR system in a noisy and reverberant environment can be improved to some extent. It should be noted that the above mentioned techniques assume that the feature vectors are independent of each other. However, for a reverberant signal the feature vectors are dependent since the arrival time of late reflections is much larger than the length of the short-time speech segments (in seconds).

1.4.3 Decoder

The decoding step is used to find the optimal sequence of phonemes or words given a sequence of observed features. For this the decoder requires additional knowledge, which is stored in the form of acoustic and language models. These models must be known prior to recognition. If the received features do not match those used in the training phase it becomes extremely difficult to decode them.

Recently, Sehr et al. [29] proposed a decoder that can be used in a reverberant environment. Unlike conventional decoders, it implicitly accounts for the dependence of successive feature vectors due to the reverberation. This is done via a combined acoustic model consisting of a conventional Hidden Markov Model (HMM), modelling the anechoic speech, and a reverberation model. Since the HMM is independent of the acoustic environment, it needs to be trained only once using anechoic speech. The training of the reverberation model is based on a set of room impulse responses for the corresponding acoustic environment. In a simulation of an isolated-digit recognition task in a highly reverberant room, the proposed method achieves a 60% reduction of the WER compared to a conventional HMM trained on reverberant speech, at the cost of an increased decoding complexity. It should be noted that the effects of late reverberation have more impact on continuous speech recognition compared to isolated-word recognition. Although these results are promising, it is unclear how large the improvement is when continuous speech recognition is performed.

1.5 Objectives

Reduction of the detrimental effects of reverberation is evidently of considerable practical importance, and is the focus of this dissertation. More specifically the dissertation deals with dereverberation techniques, i.e., acoustic signal processing techniques to reduce the detrimental effects of reverberation.

From the discussion in the previous sections it has become clear that:

1. The speech intelligibility and quality, as well as the performance of speech recognition systems, are affected primarily by late reverberation.

2. Due to changes in the source or microphone position, temperature, positioning of room furnishings, and movements in the room, the acoustic channel cannot be assumed to be time-invariant.

3. In any practical situation the desired signal is not only degraded by reverberation but also by other interferences, e.g., electronic sensor noise, thermal noise, background noise.

In the development of a dereverberation technique that is effective, efficient and robust, these issues must be taken into account. Furthermore, the spatial diversity which

is induced by reverberation, i.e., the spatial separation of the direct sound and reflections, can be exploited when multiple microphones are used. In case the developed dereverberation technique is used as a pre-processor for an ASR system, the short-term phase spectrum can often be disregarded.

Since the early days of acoustic signal processing researchers have developed numerous algorithms to counteract the detrimental effects of reverberation. Only few of these are useful in practice, and their performance is limited. In Chapter 3 an extensive literature survey is presented in which we have categorized the reverberation reduction processes depending on whether or not the AIR needs to be estimated. We then obtain two main categories, i.e., *reverberation suppression* and *reverberation cancellation*. In Chapter 3 it will become clear that so-called suppression techniques, which are successfully used for the suppression of interfering signals, are more promising than so-called cancellation techniques. We may also conclude that those techniques that require hundreds and sometimes thousands of parameters which are usually hard to estimate, result in less robust solutions compared to those techniques that require only a few parameters.

Against this background, our main objective is to develop effective, efficient, and robust speech dereverberation techniques which can be used to suppress late reverberation in a possibly noisy environment. The use of quantifiable properties of reverberation, e.g., the reverberation time, help us to minimize the number of parameters that are required.

1.6 Outline and main contributions

In this section a chapter by chapter overview is given, summarizing the main contributions of this work. Additionally, references to the publications that have been produced in the course of this work are provided.

Chapter 2 provides background information on room acoustics for later chapters (especially for Chapters 3, 5 and 6).

Although researchers have worked on speech dereverberation for three decades, at this point in time there are only a few dissertations about speech dereverberation and there is no extensive literature survey available. We have categorized the reverberation reduction processes depending on whether or not the acoustic impulse response needs to be estimated. We then obtain two main categories, i.e., *reverberation suppression* and *reverberation cancellation*. Approaches within these categories can be divided into smaller sub-categories depending on the amount of knowledge about the source and channel that is utilized. In **Chapter 3** an extensive literature survey is presented, with examples of different methods of each sub-category.

Subjective and objective measures that can be used to assess the dereverberation quality are very important in our research. Therefore, an overview of such measures is provided in **Chapter 4**. Many (often vaguely defined) objective measures were proposed in the past. Therefore, it is extremely difficult to compare the performance

of different algorithms. At this point in time there are no standardized objective measures available to evaluate the dereverberation quality. Some existing objective measures used in this dissertation are analysed in this chapter. The relevance of these measures with respect to perceptual factors such as the colouration and the reverberation time is described. Furthermore, a novel time-frequency representation of the reverberant signal is described. In this representation the spectrogram and the instantaneous DRR have been combined. The representation reveals which time-frequency components are affected most by the reverberation. Publications related to this part of the dissertation are [30, 31].

In **Chapter 5** of this dissertation, novel single- and multi-microphone speech dereverberation algorithms are developed that focus on the suppression of late reflections. Single-microphone techniques are described which estimate the early speech component via so-called spectral enhancement techniques that require a measure of the late reflections and a measure of the background noise. These measures, called late reverberant spectral variance and noise spectral variance, respectively, can be estimated directly from the noisy and reverberant microphone signal(s). Two spectral enhancement techniques that can be used to enhance the received reverberant signal are described. The first technique is based on spectral subtraction, and the second technique is based on the Optimally-Modified Log Spectral Amplitude (OM-LSA) estimator. Several modification of these spectral enhancement techniques are proposed to increase their performance. Single-microphone systems only exploit spectral diversity and temporal diversity, i.e., the separation in time of direct sound and reflections. Reverberation, of course, also induces spatial diversity. To be able to additionally exploit this diversity multiple microphones must be used, and their outputs must be combined by a suitable spatial processor such as the delay and sum beamformer. It is not *a priori* evident whether spectral enhancement is best done before or after the spatial processor. For this reason we investigate both possibilities, as well as configuration in which spectral enhancement and a novel spatial processor are merged. An advantage of the latter configuration is that the spectral variance estimator can be further improved. Our experiments show that the use of multiple microphones affords a significant improvement of the perceptual speech quality. Publications related to this part of the dissertation are [32, 33, 34, 35, 31, 36, 37, 38].

The spectral enhancement techniques that are developed in Chapter 5 require an estimate of the spectral variance of the late reverberant signal component. In **Chapter 6** we develop a novel estimator for this spectral variance. In this chapter an existing single-channel statistical reverberation model which can be used to derive such an estimator is described. The model is characterized by one parameter that depends on the characteristics of the environment. It is shown that the statistical reverberation model is closely related to the physical energy balance in an ideal diffuse environment. We find that the spectral variance estimator that is based on this model, can only be used when the source-microphone distance is larger than the critical distance. A generalization of the statistical reverberation model in which the direct sound is incorporated is developed, and a novel spectral variance estimator is derived. This model requires one additional parameter that is related to the DRR. Compared to the existing estimator the proposed estimator improves the dereverberation performance when

the source-microphone distance is smaller than the critical distance. Furthermore, the extension of the single-channel statistical reverberation model to multiple channels is developed. This extension admits a further improvement of the spectral variance estimator and the resulting dereverberation performance. A solution for estimating the amount of late reverberant spectral variance in a noisy environment is also developed. Publications related to this part of the dissertation are [32, 39, 37].

In **Chapter 7** the applicability of the theory described in the previous chapters is demonstrated using a hands-free device. This work is the result of our collaboration with dr. S. Gannot from the Bar-Ilan University and prof.dr. I. Cohen from the Technion - Israel Institute of Technology. Since hands-free devices are often used in a noisy and reverberant environment, the received microphone signal does not only contain the desired signal (commonly called near-end signal) but also interferences such as room reverberation that are caused by the near-end signal, background noise, and a far-end echo signal that results from a sound that is produced by the loudspeaker. An acoustic echo canceller is commonly used to estimate the echo path between the loudspeaker and the microphone, the estimated echo signal is then used to cancel the far-end echo. Additionally a post-processor is used to suppress background noise and residual echo, i.e., echo which could not be cancelled by the echo canceller. In this work a novel structure and post-processor for an acoustic echo canceller are developed. The proposed system is unique in the following ways. Firstly, the acoustic echo path is divided into three parts that contain (i) the direct path and a few early reflections, (ii) remaining early reflections, (iii) late reflections. The echo related to the first part is cancelled using a classical acoustic echo canceller technique. The echo related to the second part is estimated using an adaptive filter and is suppressed by the post-processor. The echo related to the third part is estimated using the statistical reverberation model described in Chapter 6, and is also suppressed by the post-processor. Note that some of the required model parameters can be estimated using the second part of the acoustic echo path which is estimated by the afore mentioned adaptive filter. Secondly, an advanced spectral enhancement technique is used to suppress late reverberation caused of the near-end source, residual echo, and background noise. Experimental results convincingly demonstrate the benefits of the proposed system for suppressing late reverberation, residual echo and background noise. Publications related to this part of the dissertation are [36, 40].

The conclusion, **Chapter 8**, summarizes the main contributions of this work and gives directions for further research.

Synthetic room impulse responses are often created using the image method developed by Allen and Berkley. In **Appendix A** this method is described and a Matlab® implementation in the form of a MEX-function is provided. Various improvements are made to incorporate the directivity of the microphone and to ensure proper inter-microphone phase relations, which are very important in the case of Single-Input Multi-Output and Multi-Input Multi-Output systems. Some extra features are added which allow the design of less complex AIRs.

The OM-LSA estimator is often used for noise reduction. In **Appendix B** an exten-

sion of this estimator is provided, which improves the suppression of multiple interferences, more specifically one non-stationary and one stationary interference. Three methods to estimate the *a priori* Signal to Interference Ratio are discussed, viz., decision-directed, causal and non-causal recursive estimation.

Chapter 2

Room-Acoustics Prerequisites

2.1 Introduction

This chapter introduces some basic theoretical properties of acoustics, which are important for understanding why particular room acoustic models are used throughout this work.

There are many different techniques for analysing the acoustics of a room and, in general, each of these techniques applies to a different frequency range of the audible spectrum, i.e., no single analytic or numerical technique can currently model the entire audible spectrum. The Acoustic Transfer Function (ATF) is defined as the frequency response of the system relating the sound source to the sound pressure at the microphone, and is probably the most frequently used function to describe an acoustic channel. Insight into the structure of the ATF can be obtained using the acoustic wave equation, which governs the propagation of acoustic waves through a material medium. Various room acoustic models for the ATF will be discussed in this chapter, e.g., pole-zero, all-zero, all-pole, common pole-zero. Since the acoustic channels in real rooms are too complex to model explicitly, Statistical Room Acoustics (SRA) is often used. SRA provide a statistical description of the ATF and the Acoustic Impulse Response (AIR) in terms of a few key quantities, e.g., source-microphone distance, room volume, and reverberation time. The reverberation time, which is a measurement of the severity of reverberation within a room, is discussed in more detail. Furthermore, the complex sound pressure in a room is studied using SRA.

When a sound is produced in a room, the reflections from the walls produce a sound energy distribution that becomes increasingly uniform with time. Eventually, the distribution of energy may be assumed to be completely uniform, and the direction of energy flow at a specific location in the room may be considered to be random. To provide useful insights in the transient behaviour of a room, the fundamental differential equation that describes the conservation of energy in a room will be discussed.

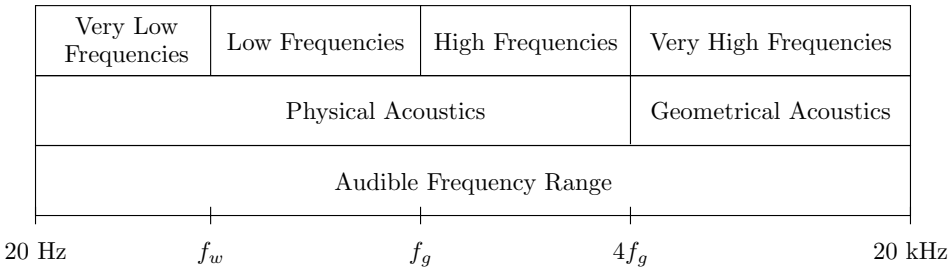


Figure 2.1 *The audible spectrum, divided into four regions.*

It should be understood that the inverse of an ATF or AIR can be used to equalize the corresponding acoustic channel. However, the ATFs and AIRs are usually non-minimum-phase, and can be decomposed into minimum-phase and excess-phase components. Since stable and causal inverse of a non-minimum-phase transfer function does not exist, often only the minimum-phase component is inverted. Therefore, we discuss the contribution of the excess-phase. Furthermore, the simulation and measurement of room acoustics are discussed.

The structure of this chapter is as follows. Various techniques which are used for analysing the acoustics of a room are discussed in Section 2.2. The acoustic wave equation is discussed in Section 2.3, and the ATF and AIR are defined in Section 2.4. Various deterministic and stochastic room acoustic models are discussed in Sections 2.5 and 2.6, respectively. A detailed description of the sound field and the energy conservation in a room is provided in Section 2.7. In Section 2.8 the reverberation time is discussed. The contribution of excess-phase is discussed in Section 2.9. How room acoustics can be simulated and measured is described in Sections 2.10 and 2.11, respectively.

2.2 Analysing Room Acoustics

There are many different techniques for analysing the acoustics of a room and, in general, each of these techniques applies to a different frequency range of the audible spectrum; no single analytic or numerical tool can currently model the entire audible frequency range between 20 Hz and 20 kHz. The audible spectrum can be divided into four regions, as depicted in Fig. 2.1, for each of which a different analytical tool is appropriate. Each of these four regions will be described in the following subsections.

Very Low Frequencies

If the frequency of a sound source is below $f_w = \frac{c}{2L}$, where c is the speed of sound in meters per second, and L is the largest dimension of the acoustic environment, there is

no resonant support for the sound in the room. The frequency band can be analysed using non-harmonic solutions to wave equations. For instance, in a small living room with dimensions $3 \times 5 \times 7$ m, and a sound velocity of 344 ms^{-1}, there is no resonant lower than 24.5 Hz.

Low Sound Frequencies

The next region corresponds to frequencies for which the wavelength of the sound under consideration is comparable to the dimensions of the room. The region spans from lowest resonant mode to the *Schroeder cut-off frequency*: [41]

$$f_{\text{g}} = \frac{G}{\sqrt{V\bar{\delta}}} \quad [\text{Hz}], \tag{2.1}$$

where $G \approx 5400$, V is the volume of the room in m^3, and $\bar{\delta}$ is the mean value of the damping constant associated with each resonant in the room. The average damping constant $\bar{\delta}$ is related to the reverberation time RT_{60}, dictating a 60 dB dynamic range (as will be discussed in Section 2.8), by the relation $\text{RT}_{60} = 3\log_{\text{e}}(10)/\bar{\delta}$, such that the cut-off frequency can be expressed in the well-known form: [42]

$$f_{\text{g}} = C\sqrt{\frac{\text{RT}_{60}}{V}} \quad [\text{Hz}], \tag{2.2}$$

where $C \approx 2000$ $(\text{ms}^{-1})^{3/2}$. In the frequency range $f_{\text{w}} \leq f < f_{\text{g}}$ wave acoustics are applicable for describing the acoustical properties of a room. Wave acoustics assume a harmonic sound source and are based on solutions of the wave equation (see Section 2.3). For instance, in a small living room with dimensions $3 \times 5 \times 7$ m and a reverberation time of 0.5 s, the lower sound frequencies range from 24.5 to 138 Hz.

High Sound Frequencies

The transition region, consists of the frequency components between f_{g} and, approximately, $4f_{\text{g}}$, where f_{g} is given by Eq. 2.2. In this region, the wavelengths are often too short for accurate modelling using wave acoustics, and too long for geometric acoustics. Thus, in general, a statistical treatment is employed. The boundary frequencies of this band for typical acoustic environments can be calculated using Eq. 2.2; for example, a small living room with dimensions $3 \times 5 \times 7$ m and $\text{RT}_{60} = 0.5$ s gives a transition region of 138 Hz to 552 Hz, whilst a car compartment of volume $V = 2.5$ m^3 and $\text{RT}_{60} = 0.05$ s gives a region of 282 Hz to 1131 Hz.

Very High Sound Frequencies

At very high sound frequencies geometrical room acoustics, also called ray acoustics, apply. As in geometrical optics, geometrical room acoustics employs the limiting case

of vanishingly small wavelengths. This assumption is valid if the dimensions of the room and its walls are large compared with the wavelength of the sound; a condition which is met for a wide-range of audio frequencies in standard rooms. Hence, in this frequency range, specular reflections and the sound ray[1] approach to acoustics prevail. Because the sound is represented by energy waves rather than complex pressure waves geometrical acoustics neglect wave related effects such as diffraction and interference.

2.3 Wave Equation

In principle, any complex sound field can be considered as a superposition of numerous simple sound waves (e.g., plane waves), and their propagation within a room can be considered linear if the properties of the medium in which the waves travel is assumed to be homogeneous, at rest, and independent of wave amplitude [41]. In physics, the acoustic wave equation governs the propagation of acoustic waves through a material medium. The form of the equation is a second order partial differential equation. The equation describes the evolution of velocity potential or sound pressure $p(\mathbf{r}, t)$ as a function of position $\mathbf{r} = (x, y, z)$ and time t.

For an homogeneous medium undergoing inviscid fluid flow, one can linearize the equations governing the dynamic behaviour of the fluid, namely the Euler's equation, i.e., Newton's 2$^{\text{nd}}$ law applied to fluids, the continuity equation, and the linearized state equation, to obtain the wave equation,

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = 0, \tag{2.3}$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.4}$$

is the *Laplacian* expressed in the Cartesian coordinates (x, y, z), and c is the speed of sound. The wave equation provides a good description of the propagation of sound waves of small amplitude in air. It accurately describes the pressure in the sound field provided $|p(\mathbf{r}, t)| \ll \rho_0 c^2$, where ρ_0 is the density of the propagation medium at equilibrium. In practice, two types of inhomogeneities occur: scalar inhomogeneities (spatial distribution of sound speed and density), for example, due to temperature variations in the medium, and vector inhomogeneities (spatial distribution of particle mean velocity), for example, due to the presence of fans or an air conditioning. However, the effects of these inhomogeneities are so small that they can be ignored in room acoustics.

Let us consider the wave equation in the frequency domain. The Fourier transform is

[1] A sound ray is meant as a small portion of a spherical wave with vanishing aperture, which originates from a certain point. It has well-defined direction of propagation and is subject to the same laws of propagation as light rays, apart from the different propagation attenuation.

defined as

$$P(\mathbf{r};\omega) \triangleq \mathcal{F}\{p(\mathbf{r},t)\}(\omega) = \int_{-\infty}^{\infty} p(\mathbf{r},t)\mathrm{e}^{-\iota\omega t}\,\mathrm{d}t, \tag{2.5}$$

where $\iota = \sqrt{-1}$. By applying the Fourier transform to Eq. 2.3 the time-independent Helmholtz equation is obtained, i.e.,

$$\nabla^2 P(\mathbf{r};\omega) + k^2 P(\mathbf{r};\omega) = 0, \tag{2.6}$$

where k denotes the wave number that is related to the angular frequency ω and the wave length λ through

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

In order to calculate the sound field emanating from a source in a specific room, we need an additional source function in Eq. 2.3 and boundary conditions, which describe sound reflection and absorption at the walls.

2.4 Acoustic Transfer Function

If there is a harmonic disturbance which is producing the waves, for which the source function is given by $s(\mathbf{r},t) = S(\mathbf{r};\omega)\mathrm{e}^{-\iota\omega t}$, then it appears at the right hand side of the wave equation Eq. 2.3, i.e.,

$$\nabla^2 p(\mathbf{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r},t)}{\partial t^2} = -s(\mathbf{r},t). \tag{2.7}$$

The Helmholtz equation is now given by

$$\nabla^2 P(\mathbf{r};\omega) + k^2 P(\mathbf{r};\omega) = -S(\mathbf{r};\omega). \tag{2.8}$$

For a unit-amplitude harmonic point source at position $\mathbf{r}_s \triangleq (x_s, y_s, z_s)$ we have $S(\mathbf{r};\omega) = \delta(\mathbf{r}-\mathbf{r}_s) = \delta(x-x_s)\delta(y-y_s)\delta(z-z_s)$, where $\delta(\cdot)$ denotes the Kronecker delta function. The partial differential equation in Eq. 2.8 can be solved by solving the following inhomogeneous equation:

$$\nabla^2 H(\mathbf{r},\mathbf{r}_s;\omega) + k^2 H(\mathbf{r},\mathbf{r}_s;\omega) = -\delta(\mathbf{r}-\mathbf{r}_\mathbf{s}), \tag{2.9}$$

where $H(\mathbf{r},\mathbf{r}_s;\omega)$ is the ATF, or *Green's function*, and $\mathbf{r}_s$ is the position of the source. For an arbitrary source function the desired source pressure can be calculated using the following relation

$$P(\mathbf{r};\omega) = \iiint_{\mathbb{V}_s} H(\mathbf{r},\mathbf{r}_s;\omega)S(\mathbf{r}_\mathbf{s};\omega)\,\mathrm{d}\mathbf{r}_s, \tag{2.10}$$

where $\mathbb{V}_s$ denotes the source volume, and $\mathrm{d}\mathbf{r}_s = \mathrm{d}x_s\,\mathrm{d}y_s\,\mathrm{d}z_s$ is the differential volume element of position $\mathbf{r}_s$. The sound pressure $p(\mathbf{r},t)$ can now be obtained using the inverse Fourier transform of Eq. 2.10.

The conventional way to solve Eq. 2.9 is to find the eigenfunctions, i.e., orthogonal solutions, $\Psi_{\mathbf{m}}(\mathbf{r};\omega)$ to the homogenous equation $\nabla^2 H(\mathbf{r},\mathbf{r}_s;\omega) + k^2 H(\mathbf{r},\mathbf{r}_s;\omega) = 0$, and express the inhomogeneous equation in terms of these eigenfunctions.

A general expression for the Green's function in an arbitrary sound field can be obtained using the eigenfunctions:

$$H(\mathbf{r},\mathbf{r}_s;\omega) = \sum_{\mathbf{m}} C_{\mathbf{m}}(\mathbf{r_s};\omega)\Psi_{\mathbf{m}}(\mathbf{r};\omega), \tag{2.11}$$

where each coefficient $C_{\mathbf{m}}$ is dependent on the position of the sound source. The eigenfunctions depend on the boundary conditions imposed by the enclosed space.

Free Space Green's Function

For an omnidirectional point source in an unbounded space, i.e., free space, the Green's function results in the well-known solution [41]

$$H(\mathbf{r},\mathbf{r}_s;\omega) = \frac{\mathrm{e}^{\iota\frac{\omega}{c}\|\mathbf{r}-\mathbf{r}_s\|}}{4\pi\|\mathbf{r}-\mathbf{r}_s\|}, \tag{2.12}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Classical Rectangular Room

In a rectangular room with physical dimensions L_x, L_y, L_z and rigid perfectly reflecting walls the eigenfunctions in Cartesian coordinates are [41]

$$\Psi_{\mathbf{m}}(\mathbf{r}) = \cos(k_x x)\cos(k_y y)\cos(k_z z), \tag{2.13}$$

where $\mathbf{m} = (m_x, m_y, m_z)$, $k_v = m_v\pi/L_v$ for $v \in \{x, y, z\}$, and m_v are non-negative integers. The eigenfunctions are often referred to as *modes* and have a simple physical interpretation as three-dimensional standing waves. The corresponding eigenvalues are $k_{\mathbf{m}}^2 = k_x^2 + k_y^2 + k_z^2$.

The solution for the inhomogeneous equation Eq. 2.9 for a classical rectangular room is [41]

$$H(\mathbf{r},\mathbf{r}_s;\omega) = \sum_{\mathbf{m}} \frac{\Psi_{\mathbf{m}}(\mathbf{r})\Psi_{\mathbf{m}}^*(\mathbf{r_s})}{\Lambda_{\mathbf{m}}(k^2 - k_{\mathbf{m}}^2)}, \tag{2.14}$$

where $\Lambda_{\mathbf{m}}$ is a normalization constant for the associated eigenvector defined by

$$\iiint_{\mathbb{V}} \Psi_{\mathbf{m}}(\mathbf{r})\Psi_{\mathbf{n}}^*(\mathbf{r})\,\mathrm{d}\mathbf{r} = \begin{cases} \Lambda_{\mathbf{m}}, & \text{for } m = n; \\ 0, & \text{for } m \neq n, \end{cases} \tag{2.15}$$

where $\mathbb{V} = \{(x, y, z) : 0 \le x \le L_x, 0 \le y \le L_y, 0 \le z \le L_z\}$ is the entire space of the room and $\mathrm{d}\mathbf{r} = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ is the differential volume element at position $\mathbf{r}$.

Equation Eq. 2.14 reveals the frequency domain structure of the ATF. The eigenfrequencies $\omega_{\mathbf{m}}$, related to the eigenvalues through $k_{\mathbf{m}} = \frac{\omega_{\mathbf{m}}}{c}$, are also known as the resonance frequencies of the room. At each eigenfrequency $\omega_{\mathbf{m}}$, the standing wave pattern of mode $\mathbf{m}$ resonates strongly. From Eq. 2.14 it can be seen that $H(\mathbf{r}, \mathbf{r}_s; \omega)$ increases without bound as $\omega \to \omega_{\mathbf{m}}$. Room mode $\Psi_{\mathbf{m}}(\mathbf{r})$ is said to be excited at eigenfrequency $\omega_{\mathbf{m}}$ (i.e., $\Psi_{\mathbf{m}}(\mathbf{r})$ makes a large contribution to sound pressure at this frequency). All rooms possess distinct resonances at low frequencies. However, in practical rooms, where walls are non-rigid and finitely absorbing, eigenvalues $k_{\mathbf{m}}$ have imaginary components that provide damping of resonance modes [41]. In that case $k_{\mathbf{m}} = \frac{\omega_{\mathbf{m}}}{c} + \iota\frac{\delta_{\mathbf{m}}}{c}$, where $\delta_{\mathbf{m}}$ denotes the damping constant (Q-factor). Assuming that $\delta_{\mathbf{m}} \ll \omega_{\mathbf{m}}$, Eq. 2.14 results in

$$H(\mathbf{r}, \mathbf{r}_s; \omega) = c^2 \sum_{\mathbf{m}} \frac{\Psi_{\mathbf{m}}(\mathbf{r})\Psi^*_{\mathbf{m}}(\mathbf{r_s})}{\Lambda_{\mathbf{m}}(\omega^2 - \omega^2_{\mathbf{m}} - 2\iota\delta_{\mathbf{m}}\omega_{\mathbf{m}})}. \tag{2.16}$$

The inverse Fourier transform of the frequency response of the room described by Eq. 2.14 leads to a AIR, $h(\mathbf{r}, \mathbf{r}_s, t)$. The variation of the AIR, or ATF, with source and microphone positions, is discussed in [43, 44], and its variation with temperature in [45]. The form of Eq. 2.14 leads to the justification of the use of some well-known modelling techniques used in signal processing, as discussed in the next section.

2.5 Modelling of Acoustic Transfer Functions

The ultimate aim of our work is to dereverberate a distorted signal recorded in an echoic acoustic environment. To achieve this, the acoustical properties of the room must be modelled. In this section we discuss some well-known modelling techniques in signal processing for the representation of room acoustics and the robustness to variations in the source and microphone positions.

2.5.1 Pole-Zero Modelling

Since the ATF can be expressed by a rational expression, it can be modelled by the conventional pole-zero model with poles $\{p^{\mathrm{PZ}}(m),\ m \in \{1, \ldots, P\}\}$ and zeros $\{q^{\mathrm{PZ}}(m),\ m \in \{1, \ldots, Q\}\}$, or autoregressive (AR) coefficients $\{a^{\mathrm{PZ}}(m),\ m \in \{1, \ldots, P\}\}$ and moving average (MA) coefficients $\{b^{\mathrm{PZ}}(m),\ m \in \{1, \ldots, Q+R\}\}$:

$$H^{\mathrm{PZ}}(\mathrm{z}) = C^{\mathrm{PZ}}\, \mathrm{z}^{-R} \frac{\prod_{m=1}^{Q}\left[1 - q^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]}{\prod_{m=1}^{P}\left[1 - p^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]} = \frac{\sum_{m=1}^{Q+R} b^{\mathrm{PZ}}(m)\mathrm{z}^{-m}}{1 + \sum_{m=1}^{P} a^{\mathrm{PZ}}(m)\mathrm{z}^{-m}}, \tag{2.17}$$

where $H^{\mathrm{PZ}}(\mathrm{z})$ represents the pole-zero modelled ATF, P is the number of poles, $Q+R$ is the total number of zeros including those at the origin, and C^{PZ} is a gain constant.

Since most acoustic transfer functions are stable and causal, the denominator of the transfer function must correspond to a stable causal sequence and therefore the poles must lie within the unit circle: $|p^{\mathrm{PZ}}(m)| < 1, m \in \{1, \ldots, P\}$. The AIRs are often non-minimum-phase, so the zeros $q^{\mathrm{PZ}}(m)$ may lie outside the unit circle, these zeros result in instabilities when the system is inverted. Alternatively, Eq. 2.17 may be expressed as

$$H^{\mathrm{PZ}}(\mathrm{z}) = C^{\mathrm{PZ}}\, \mathrm{z}^{-R}\, \frac{\prod_{m=1}^{Q_m} \left[1 - r^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right] \prod_{m=1}^{Q_m} \left[1 - s^{\mathrm{PZ}}(m)\mathrm{z}\right]}{\prod_{m=1}^{P} \left[1 - p^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]}, \tag{2.18}$$

where $|r^{\mathrm{PZ}}(m)| < 1, m \in \{1, \ldots, Q_m\}$ correspond to the minimum-phase component of the transfer function, and the zeros $|s^{\mathrm{PZ}}(m)| < 1, m \in \{1, \ldots, Q_m\}$ correspond to the maximum-phase component of the transfer function. Mourjopoulos and Paraskevas [46] discuss the pole-zero model in detail. From a physical point of view poles represent resonances (see Eq. 2.14), and zeros represent time delays and anti-resonances. The characteristics of all-zero and all-pole models when used to present room acoustics are described in Section 2.5.3 and 2.5.4.

2.5.2 Pole-Zero Model Decompositions

There are two decompositions of Eq. 2.18 which are useful for inverting acoustic transfer functions. The first is to write Eq. 2.18 as [47, 48]

$$H^{\mathrm{PZ}}(\mathrm{z}) = H^{\mathrm{PZ,min}}(\mathrm{z}) H^{\mathrm{PZ,max}}(\mathrm{z}), \tag{2.19}$$

where the minimum and maximum-phase components are given by

$$H^{\mathrm{PZ,min}}(\mathrm{z}) \triangleq C^{\mathrm{PZ}}\, \mathrm{z}^{-R}\, \frac{\prod_{m=1}^{Q_m} \left[1 - r^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]}{\prod_{m=1}^{P} \left[1 - p^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]} \tag{2.20}$$

and

$$H^{\mathrm{PZ,max}}(\mathrm{z}) \triangleq \prod_{m=1}^{Q_m} \left[1 - s^{\mathrm{PZ}}(m)\mathrm{z}\right], \tag{2.21}$$

respectively.

The second is to observe that Eq. 2.18 can also be decomposed into a minimum-phase function and a non-minimum-phase all-pass function [49]:

$$H^{\mathrm{PZ}}(\mathrm{z}) = H^{\mathrm{PZ,mp}}(\mathrm{z}) H^{\mathrm{PZ,ap}}(\mathrm{z}), \tag{2.22}$$

where the minimum-phase and all-pass components of the acoustic transfer function are given by

$$H^{\mathrm{PZ,mp}}(\mathrm{z}) = C^{\mathrm{PZ}}\,\mathrm{z}^{-R}\,\frac{\prod_{m=1}^{Q_m}\left[1 - r^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]\prod_{m=1}^{Q_m}\left[1 - s^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]}{\prod_{m=1}^{P}\left[1 - p^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]} \tag{2.23}$$

and

$$H^{\mathrm{PZ,ap}}(\mathrm{z}) = \frac{\prod_{m=1}^{Q_m}\left[1 - s^{\mathrm{PZ}}(m)\mathrm{z}\right]}{\prod_{m=1}^{Q_m}\left[1 - s^{\mathrm{PZ}}(m)\mathrm{z}^{-1}\right]}, \tag{2.24}$$

where $|H^{\mathrm{PZ,ap}}(\mathrm{z})| = 1$, for $\mathrm{z} = \mathrm{e}^{\iota\omega}$, $\forall\omega$.

2.5.3 All-Zero ATF Model

The ATF of Eq. 2.14 can be modelled by the conventional all-zero model or Finite Impulse Response (FIR) filter which can be represented with either zeros $\{q^{\mathrm{Z}}(m),\ m \in \{1,\ldots,Q\}\}$ or MA coefficients $\{b^{\mathrm{Z}}(m),\ m \in \{1,\ldots,Q+R\}\}$. This model can be considered as the numerator of Eq. 2.17, i.e.,

$$H^{\mathrm{Z}}(\mathrm{z}) = C^{\mathrm{Z}}\,\mathrm{z}^{-R}\prod_{m=1}^{Q}\left[1 - q^{\mathrm{Z}}(m)\mathrm{z}^{-1}\right] = \sum_{m=1}^{Q+R} b^{\mathrm{Z}}(m)\mathrm{z}^{-m}. \tag{2.25}$$

As discussed in the previous section this can also be expressed in the form:

$$H^{\mathrm{Z}}(\mathrm{z}) = C^{\mathrm{Z}}\,\mathrm{z}^{-R}\prod_{m=1}^{Q_m}\left[1 - r^{\mathrm{Z}}(m)\mathrm{z}^{-1}\right]\prod_{m=1}^{Q_m}\left[1 - s^{\mathrm{Z}}(m)\mathrm{z}\right], \tag{2.26}$$

where $|r^{\mathrm{Z}}(m)| < 1, m \in \{1,\ldots,Q_m\}$, and $|s^{\mathrm{Z}}(m)| < 1, m \in \{1,\ldots,Q_m\}$. The first product term corresponds to the minimum-phase component and the second product term corresponds to the maximum-phase component of the all-zero ATF. This expansion can be useful when calculating the inverse of the ATF.

There are several limitations of FIR filters imposed by the nature of room acoustics [44, 48, 50, 46]:

- Acoustic impulse responses are, in general, very long and an all-zero filter typically requires up to 10.000 coefficients. The number of coefficients is approximately determined by

$$n_s = \mathrm{RT}_{60} f_{\mathrm{s}} \quad [\text{samples}] \tag{2.27}$$

  where f_{s} is the sampling frequency in Hertz, and RT_{60} is the reverberation time for the enclosure. As an example, if $\mathrm{RT}_{60} = 0.5$ s and $f_{\mathrm{s}} = 44.1$ kHz, $n_s = 22050$ samples [48].

- The resulting FIR may be effective and appropriate only for very limited spatial combinations of source and microphone positions within a particular enclosure [51]. The large variations in ATF for small changes in source-microphone positions can cause invertibility problems, see Section 3.3.5. In some cases the distortion of the 'equalized' transfer function will be greater than the original distortions due to the ATF [43, 44, 50, 48, 2]. The sensitivity can be explained by the nature of room acoustics; transfer function zeros result from local cancellations of multipath sound components which are easily disturbed by slight changes in source-microphone positions [46]. This suggests that if the room impulse response is incorrectly estimated there may be problems with equalization.

Tohyama and Lyon [52] discuss the effect of truncating an impulse response and demonstrate that truncation can change the minimum-phase behaviour of an ATF into a non-minimum-phase characteristic.

2.5.4 All-Pole ATF Model

An alternative to Eq. 2.25 for the representation of Eq. 2.14 is the causal all-pole model, or Infinite Impulse Response (IIR) filter, which can be represented either by the poles $\{p^{\mathrm{P}}(m),\ m \in \{1,\ldots,P\}\}$ or by AR coefficients $\{a^{\mathrm{P}}(m),\ m \in \{1,\ldots,P\}\}$ and can be considered as the denominator of Eq. 2.17:

$$H^{\mathrm{P}}(\mathrm{z}) = \frac{C^{\mathrm{P}}}{\prod_{m=1}^{P}\left[1 - p^{\mathrm{P}}(m)\mathrm{z}^{-1}\right]} = \frac{C^{P}}{1 + \sum_{m=1}^{P} a^{\mathrm{P}}(m)\mathrm{z}^{-m}}. \tag{2.28}$$

The all-pole or autoregressive model for approximating rational transfer functions is widely used in many fields, especially in speech analysis. Typical all-pole model orders required for approximating acoustic transfer functions are in the range of $50 \leq P \leq 500$ [46]. Mourjopoulos and Paraskevas [46] state that the all-pole model orders are typically a factor 40 lower than the all-zero model orders, while several studies by Gudvnagen and Flockton [53, 54] state that the gain achieved using pole-zero over all-zero modelling of reverberant acoustic environments is not as high as generally thought throughout the literature, with reduction in the number of coefficients typically in the order of 1.2 to 1.5. These latter studies use modelling error functions to measure the fit of the pole-zero models to the complete AIR, rather than fitting the most important reverberant characteristics. Therefore, a significant reduction in model order should be expected for those applications where it is more important to model the main reverberant component rather than just minimizing the modelling error. Using least squares approximation theory Liavas and Regalia concluded that there was no substantial improvement by the use of IIR models compared to FIR models [55].

A significant advantage of the all-pole model over the all-zero model is its lower sensitivity to changes in source and observer positions. Mourjopoulos and Paraskevas

[46] conclude that in many signal processing applications dealing with room acoustics, it may be both sufficient and more effective to manipulate all-pole model coefficients rather than high order all-zero models. The all-pole model is also the basis of the technique discussed in [56] which allows the classification of all possible ATFs corresponding to different source-observer positions, thereby providing a 'codebook' for possible transmission paths in dereverberation applications. A shortcoming of the causal all-pole model filter is that, since it is causal and stable, it is minimum-phase and therefore cannot model the non-minimum-phase component of room acoustics. Nevertheless, a subband all-pole model can be used to avoid this problem since only a number of subbands considered individually have non-minimum-phase characteristics [57, 58].

2.5.5 Common Acoustical Pole-Zero Modelling

Acoustical poles are approximately independent of the source and observer position since they correspond to the resonant frequencies of the room. Standing waves occur at these resonances and can be observed at any point in the room, except at node points. However, the amplitude of the standing wave varies depending on the microphone positions. This variation is reflected in the zeros of the ATF [51]. The spatial independency of the poles was not assumed in Eq. 2.17 and 2.28. As such Eq. 2.18 can be written in the simpler form:

$$H^{\mathrm{CAPZ}}(\mathbf{r},\mathbf{r}_s;\mathrm{z}) = C^{\mathrm{CAPZ}}(\mathbf{r},\mathbf{r}_s)\,\mathrm{z}^{-R} \frac{\prod_{m=1}^{Q_m}\left[1-r^{\mathrm{CAPZ}}(\mathbf{r},\mathbf{r}_s,m)\mathrm{z}^{-1}\right]\prod_{m=1}^{Q_m}\left[1-s^{\mathrm{CAPZ}}(\mathbf{r},\mathbf{r}_s,m)\mathrm{z}\right]}{\prod_{m=1}^{P}\left[1-p^{\mathrm{CAPZ}}(m)\mathrm{z}^{-1}\right]} \tag{2.29}$$

where $\{p^{\mathrm{CAPZ}}(m),\ m \in \{1,\ldots,P\}\}$ are the common poles independent of $\mathbf{r}$ and $\mathbf{r}_s$. Nevertheless, it should be noted that the spatial independence assumption of the acoustical poles is simplistic, and other investigations on the fluctuation of ATFs within reverberant environments suggest that this may not be strictly true [2]. Eq. 2.29 is known as the CAPZ model of ATFs and was first introduced by Haneda et. al. [59, 51] and extended in [60].

2.5.6 Theoretical Pole Order

As shown in Section 2.4 the harmonic solutions of the wave equation Eq. 2.3 for a rectangular room can be found by solving Eq. 2.9. The number of modes in the harmonic solution can be counted, and it can be shown that the order of modes $N(f_u)$ for a room of dimensions L_x,L_y,L_z with volume $V = L_xL_yL_z$ up to an upper frequency

limit f_u is given by [41]

$$N(f_u) = \frac{4\pi V}{3}\left(\frac{f_u}{c}\right)^3 + \frac{\pi S}{4}\left(\frac{f_u}{c}\right)^2 + \frac{L}{8}\left(\frac{f_u}{c}\right), \tag{2.30}$$

where $V = L_x L_y L_z$ is the room's volume, $S = 2(L_x L_y + L_x L_z + L_y L_z)$ is the room's surface area and $L = 4(L_x + L_y + L_z)$ is the sum of all the edge lengths occurring in the rectangular room. If $f_u \gg 500$ Hz, and $\sqrt[3]{V} \gg \sqrt{S}$, then the last two terms in Eq. 2.30 can be ignored. The order of the all-pole model up to a given sampling frequency f_s is therefore given by $P \approx 2N(f_s/2)$, or [51]

$$P \approx \frac{\pi V}{3}\left(\frac{f_s}{c}\right)^3. \tag{2.31}$$

If the all-pole model order is the same as the theoretical order in Eq. 2.31, the all-pole model corresponds well with the actual room response. If the model order is lower than the theoretical order, the least squares estimated poles correspond to the major resonance frequencies which have high Q factors [51]. This is typically the case since, for example, a typical small office with volume 40 m^3 has $\pm 1.75 \times 10^5$ acoustic modes with natural frequencies below 3.5 kHz, giving a very high all-pole model order. Moreover, it is clear for most typical rooms that the model order given by Eq. 2.31 is much greater than the typical length of an FIR filter, as given in Eq. 2.27. Hence, Eq. 2.31 is a very loose upper bound.

2.6 Statistical room acoustics

In theory the validity of the modal expressions derived in Section 2.4 is not restricted to low frequencies. Therefore, it may seem surprising that a completely different approach based on statistical considerations, i.e., SRA, is actually more useful at medium and high frequencies than the deterministic approach described in the foregoing section. SRA provides a statistical description of the ATF between the source and microphone in terms of a few key quantities, e.g., source-microphone distance, room volume, and reverberation time. The crucial assumption of SRA is that the distribution of amplitudes and phases of individual plane waves, which sum up to produce sound pressure at some point in a room, is so close to random that the sound field is fairly uniformly distributed throughout the room volume. Evidently, using a model based on statistical considerations we can only make predictions with a certain probability. There are two reasons why SRA are considered. The first reason is that expressions based on sums of modes are in practice less useful at high frequencies. The problem is that when hundreds of complex terms are summed the result becomes very sensitive to small errors in each term. For example, the dimensions of the room might be slightly different from the dimensions used in the model. Even very small modelling errors will shift the natural frequencies of the modes, and the amplitude and phase of each of the terms that correspond to modes driven by their natural frequency may change somewhat. As a result the sum can be completely different. Secondly, statistical models can be

surprisingly powerful in the sense that they make it possible to predict a number of characteristics of, e.g., the sound field or the reverberation time in a room, on the basis of very little information.

Sabine's [61] major contribution was the introduction of statistical methods to calculate the reverberation time of a space without considering the details of the space geometry. Schroeder has extended Sabine's fundamental work [42, 62] and derived a set of statistical properties describing the frequency response of a random impulse response.

In a room response, the average number of modes per Hz, i.e., the modal density D_{m}, is approximately proportional to the square of frequency f [41]:

$$D_{\mathrm{m}}(f) \approx 4\pi V \frac{f^2}{c^3}. \tag{2.32}$$

The average number of reflections, i.e., the echo density D_{e}, is approximately proportional to the square of time t [41]:

$$D_{\mathrm{e}}(t) \approx 4\pi c^3 \frac{t^2}{V}. \tag{2.33}$$

These expressions can be demonstrated for rectangular rooms, and can be generalized to rooms of any geometry [41]. For higher frequencies and larger times, both densities become very high. These properties provide the foundation for statistical models of room responses, as developed by Schroeder [42] in the frequency domain, and more recently in the time domain by Polack [63].

2.6.1 Frequency-domain statistical model

Eq. 2.32 implies that, at high frequencies, the normal modes of a room overlap in the frequency domain, i.e., the average separation between natural frequencies is smaller than the bandwidth $\Delta f_{\mathbf{m}}$ of a mode. The bandwidth of a mode can be expressed as follows:

$$\Delta f_{\mathbf{m}} = \frac{\delta_{\mathbf{m}}}{\pi}, \tag{2.34}$$

where $\delta_{\mathbf{m}}$ is the damping constant of mode $\mathbf{m}$. A parameter which quantifies the probable number of modes which exist within the 3 dB bandwidth of any mode is the modal overlap $M(f)$, a parameter widely used in statistical energy analysis, defined by the expression

$$M(f) = \Delta f \; D_{\mathrm{m}}(f). \tag{2.35}$$

Thus, at high frequencies, any source signal will simultaneously excite several room modes. Assuming a sine-wave excitation and a microphone located in the reverberant field, the signal received by the microphone is the sum of the contributions of a large number of modes, where the amplitude and phase of each contribution varies with the microphone position. Consequently, the complex frequency response can be considered

as a space dependent stochastic process whose real and imaginary part are independent Gaussian processes having the same variance [62, 64, 41]. This two-dimensional Gaussian density arises from the central limit theorem, assuming independence between modes, and implies that the amplitude of the frequency response follows a Rayleigh distribution. These statistical properties also apply when the complex response is considered as a frequency dependent stochastic process, for a given microphone position. These properties are valid irrespective of the microphone position and the room, provided that the direct sound can be neglected compared to the reflected sound and that the frequency of interest is above the Schroeder frequency. The high modal overlap implies that the peaks in the frequency response of the acoustic transfer function do not correspond to the individual natural frequencies. Although the modal density increases as the square of frequency, as shown by Eq. 2.32, the average separation between adjacent peaks in the amplitude frequency response only depends on modal bandwidth. Accordingly, the density of peaks in the frequency domain is proportional to the reverberation time [65, 41]. The average number of maxima per Hz is approximately given by

$$D_{\mathrm{f}} \approx \frac{\sqrt{3}}{\bar{\delta}} \approx \frac{\mathrm{RT}_{60}}{4}. \tag{2.36}$$

Eq. 2.36 is obtained using the fact that the average damping constant $\bar{\delta}$ is related to the reverberation time RT_{60} by

$$\bar{\delta} = \frac{3\ \log_{\mathrm{e}}(10)}{\mathrm{RT}_{60}}. \tag{2.37}$$

This frequency-domain statistical model relies on the assumption of high modal overlap in the frequency domain, which is not valid at low frequencies. The Schroeder frequency, above which the theory is valid, has been verified experimentally in [62], and is given by Eq. 2.2. By combining Eq. 2.32, 2.2, and 2.36, one can verify that the average spacing between natural frequencies must be less than one third of the bandwidth of a mode for the theory to be valid.

The validity of the frequency-domain statistical model is subject to a set of conditions: [41, 2, 66]

1. The dimensions of the room are relatively large compared to the wavelength. For the frequencies of interest (in speech processing we are mainly interested in the band 300-3500 Hz), this condition is usually satisfied.

2. The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V (in m^3), and reverberation time RT_{60} (in seconds), which is defined as the time for the reverberation level to decay to 60 dB below the initial level, this condition is fulfilled for frequencies that exceed the Schroeder frequency: $f_{\mathrm{g}} = 2000\sqrt{\mathrm{RT}_{60}/V}$.

3. The source and the microphones are located in the interior of the room, at least a half-wavelength away from the walls, where $\lambda = c/f$ is the wavelength and f is

the frequency of the source signal. The sound field at a wall-mounted microphone can hence not be modelled as diffuse[2].

The frequency-domain statistical model can be used to derive different properties of the acoustic transfer function. The acoustic transfer function from the source to a microphone at position $\mathbf{r}$ can be expressed as the sum of the direct component, $H_\mathrm{d}(\mathbf{r}, \mathbf{r}_s; \omega)$, and a reverberant component, $H_\mathrm{r}(\mathbf{r}, \mathbf{r}_s; \omega)$, such that

$$H(\mathbf{r}, \mathbf{r}_s; \omega) = H_\mathrm{d}(\mathbf{r}, \mathbf{r}_s; \omega) + H_\mathrm{r}(\mathbf{r}, \mathbf{r}_s; \omega). \tag{2.38}$$

Under the conditions stated above, 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 [41, 64].

In the following it is assumed that the source and microphone position is represented by $\theta = [\mathbf{r}^T \ \mathbf{r}_s^T]^T$. Eq. 2.38 can now be written as

$$H(\theta; \omega) = H_\mathrm{d}(\theta; \omega) + H_\mathrm{r}(\theta; \omega). \tag{2.39}$$

The spatial expectation $\mathcal{E}_\theta\{\cdot\}$ is defined as the ensemble average over all allowable (in terms of condition 3) values of θ. The direct component of the ATF is the free-space Greens function as defined in Eq. 2.12. If we additionally assume that all realizations $\tilde{\theta}$ of θ have a constant source-microphone distance, i.e., only rotations and translation of the source-microphone position are allowed, then the direct component only depends on the distance $D = \|\mathbf{r} - \mathbf{r}_s\|$, such that $\mathcal{E}_\theta\{H_\mathrm{d}(\theta; \omega)\} = H_\mathrm{d}(D; \omega)$. From SRA, the expected density spectrum of the reverberant component is given by [41, 2]

$$\mathcal{E}_\theta\{|H_\mathrm{r}(\theta; \omega)|^2\} = \frac{1 - \bar{\alpha}}{\pi \bar{\alpha} S}, \tag{2.40}$$

with S being the total surface area of the room and $\bar{\alpha}$ the average absorption coefficient of the room walls. Although Eq. 2.40 is often used (c.f. [2, 67, 68, 69]) it should be noted that the absorption coefficient is frequency dependent due to the frequency dependent absorption coefficients of walls and other objects, and of air [41].

In [2] it is shown that the spatial expectation of the cross terms of the squared magnitude of Eq. 2.38 is zero. Hence, the spatially expected energy density spectrum of the ATF can be written as

$$\mathcal{E}_\theta\{|H(\theta; \omega)|^2\} = |H_\mathrm{d}(D; \omega)|^2 + \mathcal{E}_\theta\{|H_\mathrm{r}(\theta; \omega)|^2\}. \tag{2.41}$$

Note that only the reverberant component varies with the source and microphone position θ but its spatial expectation is independent of θ.

The spatial cross-correlation of the reverberant paths between the m^{th} and the n^{th} acoustic channel has been shown to be [67]

$$\mathcal{E}_\theta\{H_\mathrm{r}(\mathbf{r}_m, \mathbf{r}_s; \omega) H_\mathrm{r}^*(\mathbf{r}_n, \mathbf{r}_s; \omega)\} = \left(\frac{1 - \bar{\alpha}}{\pi \bar{\alpha} S}\right) \frac{\sin\left(k\|\mathbf{r}_m - \mathbf{r}_n\|\right)}{k\|\mathbf{r}_m - \mathbf{r}_n\|}, \tag{2.42}$$

[2] In case the sound field is diffuse the sound energy density and the direction of the intensity vector are uniformly distributed across the room.

where $k = 2\pi f/c = \omega/c$ is the wave number, $\theta = [\mathbf{r}_m^T\ \mathbf{r}_n^T\ \mathbf{r}_s^T]^T$, and $\mathbf{r}_m$ and $\mathbf{r}_n$ are the three-dimensional position vectors of the m^{th} and n^{th} microphone, respectively, with the origin at $(x, y, z) = (0, 0, 0)$. All realizations $\tilde{\theta}$ of θ have a constant distance between all positions, i.e., only rotations and translations of the source-array configuration are allowed.

2.6.2 Time-domain statistical model

Moorer noted the auditive resemblance between a concert hall impulse response and a white noise signal multiplied by an exponentially decaying envelope, and reported that such a synthetic response can produce, by convolution with anechoic signals, a natural sounding reverberation effect [70]. To obtain a frequency-dependent reverberation time, he suggested to use a filter bank and to sum the subband signals after multiplying them with different exponential envelopes.

Polack [63] developed a time-domain model complementing Schroeder's frequency-domain model. In this model, an acoustic impulse response is described as one realization of a non-stationary stochastic process:

$$h(t) = b(t)\mathrm{e}^{-\bar{\delta}t} \quad \text{for } t > 0, \tag{2.43}$$

where $b(t)$ is a zero-mean stationary Gaussian noise, and $\bar{\delta}$ is related to the reverberation time RT_{60} by Eq. 2.37. The random noise $b(t)$ is characterized by its power spectral density, denoted $B(f)$. According to Polack [63], the acoustic impulse response can then be observed on two different time scales:

1. A small time scale corresponding to the fast variations of the signal $b(t)$, i.e., to the order of the millisecond (measured by the temporal spreading of the autocorrelation function of $b(t)$). This corresponds, in the frequency domain, to the scale of slow variations of the power spectral density $B(f)$.

2. A large time scale corresponding to the slow variations of the temporal envelope, i.e., to the order of seconds (measured by the reverberation time), and corresponding to the scale of the fast variations of the frequency response of the room (measured by the bandwidth of the normal modes, given by Eq. 2.34).

Since these two scales differ several orders of magnitude, it is possible to separate the time variable t and the frequency variable f in the calculation of statistical quantities [63].

In the time domain there is an interval after which Polack's stochastic model becomes valid. The time-domain response can only be Gaussian if a sufficient number of reflections overlap at any time along the response. The peaks in the acoustic impulse response then no longer correspond to the arrivals of individual reflections. Since the reflection density increases with time according to Eq. 2.33, the situation is similar to that found in the frequency domain, except that the 'width' of a reflection in the time

domain cannot be defined solely with respect to the intrinsic properties of the room (unlike the bandwidth of a mode).

The spreading of a reflection in the time domain can only be expressed with reference to the bandwidth of the source excitation (which determines the spreading of the source pulse), or to the bandwidth of the microphone. If the criterion is that at least 10 reflections overlap within a characteristic time resolution of the auditory system, taken equal to 24 ms in [63], Eq. 2.33 leads to:

$$t_{\text{mix}} = 1000\sqrt{V} \quad [\text{s}]. \tag{2.44}$$

This value was also proposed in [71] as a reasonable approximation for the transition time between early reflections and late reverberation. Polack shows that the exponentially decaying stochastic model can be established within the framework of geometrical acoustics and billiard theory [63, 72], and defines the mixing time as the time it takes for a set of initially adjacent sound rays to spread uniformly across the room. By that time (if the origin is taken as the time of emission of a sound pulse by the source), the reverberation process has become diffuse, i.e., the sound energy density and the direction of the intensity vector are uniformly distributed across the room. The mixing character of a room depends on its geometry and the diffusing properties of the boundaries. When mixing is achieved, the echo density increases exponentially with time, rather than proportional to t^2 [72]. Consequently, the value $1000\sqrt{V}$ can be considered as an upper limit for the mixing time in typical 'mixing' rooms. The validity region of the stochastic time-frequency model of reverberation decays is limited to frequencies higher than the Schroeder frequency, given by Eq. 2.2, and to times later than the mixing time, for which an upper limit is given by Eq. 2.44. The validity region is illustrated in Fig. 2.2.

It is worth recalling that the late reverberation can be described by a stochastic model (implying no knowledge of the natural frequencies of the room) only above the Schroeder frequency, which takes a value of 100 Hz for a 400 m^3 room having a reverberation time of 1 second. Unlike typical bathrooms, concert halls are relatively absorbent with a relatively large volume, leading to Schroeder frequencies close to the lower limit of the audible range.

2.7 Sound Field

The complex sound pressure can be decomposed into two components. The first component, i.e., direct component $P_{\text{d}}(\mathbf{r};\omega)$, is the part of the sound that arrives from a sound source directly, without reflection or diffraction. The second component, i.e., the reverberant component $P_{\text{r}}(\mathbf{r};\omega)$, is what remains after subtracting the direct component. The complex sound pressure $P(\mathbf{r};\omega)$ can thus be expressed as

$$P(\mathbf{r};\omega) = P_{\text{d}}(\mathbf{r};\omega) + P_{\text{r}}(\mathbf{r};\omega). \tag{2.45}$$

Using statistical room acoustics it can be shown (c.f. [64]) that the real and imaginary

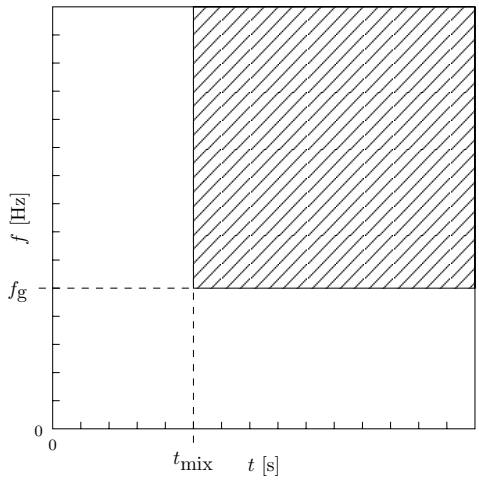


Figure 2.2 *Validity region of the statistical model in the time-frequency domain (hatched).*

parts of the complex sound pressure $P_r(\mathbf{r};\omega)$ have zero mean and are statistically independent, i.e.,

$$\mathcal{E}_\theta\{\mathrm{Re}\{P_r(\mathbf{r};\omega)\}\} = \mathcal{E}_\theta\{\mathrm{Im}\{P_r(\mathbf{r};\omega)\}\} = 0 \tag{2.46}$$

and

$$\mathcal{E}_\theta\{\mathrm{Re}\{P_r(\mathbf{r};\omega)\}\mathrm{Im}\{P_r(\mathbf{r};\omega)\}\} = 0, \tag{2.47}$$

where we have again used $\mathcal{E}_\theta\{\cdot\}$ to denote the operation of spatial averaging.

If a sound source generates a sound wave it has to deliver some energy to a fluid. The energy is carried away by the sound wave. Accordingly the amount of energy contained in one unit volume of the wave is characterized by the energy density. As with any kind of mechanical energy one has to distinguish between potential and kinetic energy density:

$$E_{pot}(t) = \frac{p^2(t)}{2\rho_0 c^2}, \tag{2.48}$$

$$E_{kin}(t) = \frac{\rho_0 |\mathbf{v}(t)|^2}{2}, \tag{2.49}$$

where $\mathbf{v}(t)$ denotes the vector of the particle velocity at time t. The instantaneous total energy density is given by

$$E(t) = E_{pot}(t) + E_{kin}(t) \quad [\mathrm{J/m^3}]. \tag{2.50}$$

In a simple travelling plane wave the sound pressure and the longitudinal component v of the particle velocity vector $\mathbf{v}$ are related by $p = \rho_0 c v$. In this case the energy density is expressed by

$$E(t) = \frac{p^2(t)}{\rho_0 c^2}. \tag{2.51}$$

Direct Sound

The time delay of the direct sound component is related to the distance D, in meters, and the sound velocity c, in meters per second, through

$$t_{\mathrm{d}} = \frac{D}{c} \quad [s]. \tag{2.52}$$

The sound energy density related to the direct sound at distance D from the source is given by [41, 2]

$$\begin{aligned} E_{\mathrm{d}} &= \frac{\mathcal{E}_\theta\{P_{\mathrm{d}}(\mathbf{r};\omega)P_{\mathrm{d}}^*(\mathbf{r};\omega)\}}{\rho_0 c^2} \\ &= \frac{QW_{\mathrm{s}}}{4\pi c D^2}, \end{aligned} \tag{2.53}$$

where Q denotes the directivity of the source compared to a sphere, and W_{s} denotes the power of the source in Watt.

Reverberant Sound

In order to deal with sound fields in a room at high frequencies, or, strictly speaking, in a room where many modes have natural frequencies within the bandwidth of any mode, a model can be adopted in which it is assumed that the sound field at a point consists of a superposition of the contributions from a number of planes waves. The plane waves are assumed to be arriving at the point considered from all possible propagation directions. One can visualize this by assuming that a point in the sound field is at the center of a sphere whose surface is divided into a very large number of segments of equal area. The line passing through the center of one of these segments and the center of the sphere defines the propagation direction associated with a particular plane wave. The field is then assumed to consist of an infinite number of plane waves, each associated with a different propagation direction. These propagation directions are defined by assuming that all segments of equal area on the surface of the sphere become infinitesimally small. In case all of these directions of propagation contribute equally to the spatially averaged energy density then the sound field is called diffuse. Although sound fields in a real room do not exactly exhibit a diffuse field behaviour, the diffuse field turns out to be a good approximation.

The spatial correlation of pressure, which is the correlation coefficient of complex sound pressure between two points, is well defined in a diffuse field [73]:

$$\frac{\mathcal{E}_\theta\{P_{\mathrm{r}}(\mathbf{r};\omega)P_{\mathrm{r}}^*(\mathbf{r}+\Delta\mathbf{r};\omega)\}}{\mathcal{E}_\theta\{|P_{\mathrm{r}}(\mathbf{r};\omega)|^2\}} = \frac{\sin(k||\Delta\mathbf{r}||)}{k||\Delta\mathbf{r}||}, \tag{2.54}$$

where $k = \omega/c$. Eq. 2.54 is equal to Eq. 2.42 normalized by the density spectrum of the reverberant component, given by Eq. 2.40. Using statistical room acoustics, the correlation in the acoustic transfer function between different frequencies can be determined. Utilizing the fact that the impulse response is approximately exponential, $\mathcal{E}_\theta\{h^2(\mathbf{r},t)\} \sim \mathrm{e}^{-t/\tau}$ with decay constant $\tau = 1/2\bar{\delta}$, the frequency correlation is [62]:

$$\frac{\mathcal{E}_\theta\{P_{\mathrm{r}}(\mathbf{r};\omega)P_{\mathrm{r}}^*(\mathbf{r};\omega+\Delta\omega)\}}{\mathcal{E}_\theta\{|P_{\mathrm{r}}(\mathbf{r};\omega)|^2\}} = \frac{1}{1+(\tau\Delta\omega)^2}. \tag{2.55}$$

The frequency correlation drops rapidly with increasing $\Delta\omega$. For example in an office, a reverberation time of 400 ms is typical. Here the frequency correlation drops rapidly beyond $\Delta\omega = 20$ Hz.

The steady-state sound energy density related to the reverberant sound component in a diffuse sound field is given by [41, 2]

$$\begin{aligned} E_\mathrm{r} &= \frac{\mathcal{E}_\theta\{P_\mathrm{r}(\mathbf{r};\omega)P_\mathrm{r}^*(\mathbf{r};\omega)\}}{\rho_0 c^2} \\ &= \frac{4W_\mathrm{s}}{cR}, \end{aligned} \tag{2.56}$$

where R denotes the room constant. In case we assume that the absorption is independent of the angle of incidence, the room constant R is given by

$$R = \frac{\bar{\alpha}S}{1-\bar{\alpha}}, \tag{2.57}$$

where $\bar{\alpha}$ denotes the average absorption coefficient and S is the total absorption area.

2.7.1 Critical Distance

The distance at which the steady-state reverberant energy equals the direct sound energy is called the critical distance or radius. Using Eq. 2.53 and 2.56 we obtain

$$\frac{Q}{4\pi D_\mathrm{c}^2} = \frac{4}{R}. \tag{2.58}$$

By solving Eq. 2.58 the critical distance D_c is obtained for a point sound source, i.e.,

$$D_\mathrm{c} = \sqrt{\frac{QR}{16\pi}} \quad [\mathrm{m}]. \tag{2.59}$$

The reverberation time can be estimated using Sabine's equation [61]

$$\mathrm{RT}_{60} = \frac{24\ \ln(10)\,V}{c\bar{a}S} \quad [\mathrm{s}] \tag{2.60}$$

where $\bar{a}$ denotes Sabine's absorption coefficient. Another well-known formula to estimate the reverberation time was derived by Eyring:

$$\mathrm{RT}_{60} = -\frac{24\ \ln(10)\,V}{c\ln(1-\bar{\alpha})\,S} \quad [\mathrm{s}]. \tag{2.61}$$

For many practical purposes it is safe to assume that the average absorption coefficient $\bar{\alpha}$ is small compared to unity. The logarithm in Eq. 2.61 can be expanded into a series

$$-\ln(1-\bar{\alpha}) = \bar{\alpha} + \frac{\bar{\alpha}^2}{2} + \frac{\bar{\alpha}^3}{3} + \ldots \tag{2.62}$$

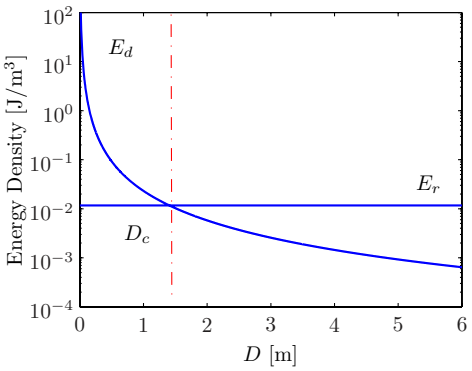


Figure 2.3 *Spatial dependence of direct (E_d) and reverberant (E_r) energy densities. The critical distance D_c is determined by Eq. 2.59.*

and all terms higher than the first order in $\bar{\alpha}$ may be neglected. In the later case the average absorption coefficient $\bar{\alpha}$ is similar, or equal, to Sabine's absorption coefficient $\bar{a}$, i.e., $\bar{\alpha} \simeq \bar{a}$. Assuming that the speed of sound in air is 344 ms^{-1}, Eq. 2.59 can be approximated by

$$D_\mathrm{c} \approx 0.1\sqrt{\frac{QV}{\pi \mathrm{RT}_{60}}} \quad [\mathrm{m}]. \tag{2.63}$$

The critical distance D_c is a function of the room parameters and the source directivity. The reverberation distance, denoted by D_h, is obtained using an omnidirectional sound source and is only a function of the room volume and the reverberation time of the room. For an omnidirectional point sound source the reverberation distance is equal to [41]

$$D_\mathrm{h} = 0.1\sqrt{\frac{V}{\pi \mathrm{RT}_{60}}} \quad [\mathrm{m}]. \tag{2.64}$$

If an observer is within the critical distance of a source, the direct energy is greater than the reverberant energy while, if the observer is outside the critical distance, the reverberant energy will be dominant. The direct and reverberant energy as a function of the source-microphone distance D are depicted in Fig. 2.3. The intelligibility of speech, for example, depends greatly on whether the observer is near the source, or far from the source. This explains why reverberation has negligible effect on intelligibility when using normal telephones or equipment where the microphone can be placed close to the sound source. Speech intelligibility is affected when the distance between the source and an omnidirectional microphone is larger than $0.3D_\mathrm{c}$, or larger than $0.5D_\mathrm{c}$ for a directional microphone. For example, when using an omnidirectional microphone in a small living room with dimensions $3 \times 5 \times 7$ m and $\mathrm{RT}_{60} = 0.5$ s, the critical distance $D_\mathrm{c} \approx 0.82$ m. Speech intelligibility would be affected when the source-microphone distance is larger than 0.25 m.

2.7.2 Energy Balance

When a sound is produced in a room the reflections from the walls and other surfaces produce a sound energy distribution in the room that becomes increasingly uniform with time. After a certain time the distribution of energy may be assumed to be completely uniform, and the direction of energy flow at a specific location in the room may be considered to be random. It should be noted that these assumptions are not true close to the absorbing surfaces or the source.

The fundamental differential equation governing the growth of sound in a room is

$$W_\mathrm{s} = V\frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t} + B\bar{\alpha}S, \tag{2.65}$$

where B denotes the irradiation strength, i.e.,

$$B = \frac{c}{4}E_\mathrm{r}(t), \tag{2.66}$$

and $\bar{\alpha}$ is the average absorption coefficient. The term $\bar{\alpha}S$ is often referred to as the equivalent absorption area of the room and is denoted by A. Eq. 2.65 can now be written as

$$W_\mathrm{s} = V\frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t} + \frac{cA}{4}E_\mathrm{r}(t). \tag{2.67}$$

This equation states that the rate at which energy is absorbed by the surfaces $\left(\frac{cA}{4}E_\mathrm{r}(t)\right)$ plus the rate $\left(V\frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t}\right)$ at which it increases throughout the room must equal the rate at which energy is being produced by the source. Using the time-constant

$$\tau = \frac{1}{2\bar{\delta}} = \frac{4V}{cA} \tag{2.68}$$

Eq. 2.67 can be rewritten as

$$\frac{4W_\mathrm{s}}{cA} = \tau\frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t} + E_\mathrm{r}(t). \tag{2.69}$$

Let us now study the energy density in case an acoustic source $W_\mathrm{s}(t)$ is switched on at $t = 0$ and switched off at time $t = t_s$. The evolution of the energy density can be divided into three different stages, i.e., the (i) sound growth, (ii) steady state, and (iii) sound decay, and is depicted in Fig. 2.4.

i) **Sound Growth:** Assuming that the source starts at time $t = 0$, the solution of the differential equation in Eq. 2.67 may be expressed as

$$E_\mathrm{r}(t) = \frac{4W_\mathrm{s}(t)}{cA}\left(1 - \mathrm{e}^{-t/\tau}\right). \tag{2.70}$$

ii) **Steady-State:** For steady-state conditions $W_\mathrm{s}(t)$ is constant and the differential quotient in Eq. 2.67 is zero. The energy density is then given by

$$E_\mathrm{r}(t) = \frac{4W_\mathrm{s}(t)}{cA}. \tag{2.71}$$

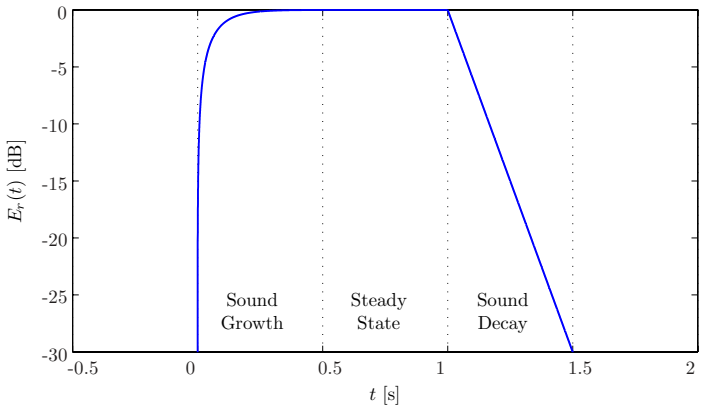


Figure 2.4 *Energy density during sound growth, steady-state, and sound decay.*

iii) **Sound Decay:** If the sound source is switched off at $t = t_s$, i.e., for $W_s(t) = 0$ for $t_s \geq 0$, the differential equation Eq. 2.67 becomes homogeneous and has the solution

$$E_r(t) = E_0 \, e^{-t'/\tau} \quad \text{for } t' \geq 0, \tag{2.72}$$

where $t' = t - t_s$, and E_0 denotes the initial energy density in J/m^3 at $t = t_s$. Eq. 2.72 may also be written in terms of the average damping constant $\bar{\delta}$ using the relation $1/\tau = 2\bar{\delta}$ [41], which results in

$$E_r(t) = E_0 \, e^{-2\bar{\delta}t'} \quad \text{for } t' \geq 0. \tag{2.73}$$

2.7.3 Spectral Deviation Measure

A pressure spectral response can be measured by placing an omnidirectional microphone at a distance D from the source. The source is driven by a pure sine wave so which is slowly varied in frequency, with the acoustic power output of the source maintained at a constant level. The microphone, which has a constant response over the frequency range, is connected to a recording device which records the magnitude of the steady-state sound pressure versus frequency. At any frequency, the complex sound pressure $P(\mathbf{r};\omega)$ at the microphone is the vector sum of a direct component $P_d(\mathbf{r};\omega)$ from the source and the reverberant components $P_r(\mathbf{r};\omega)$ which arrive at amplitudes and phases that are, in general, different from the direct component and from each other. Above the Schroeder frequency f_g it can be assumed that the reverberant components are the results of large numbers of simultaneously excited, but uncorrelated, normal modes.

In SRA it is often assumed that the real and imaginary parts $\mathrm{Re}\{P_r(\mathbf{r};\omega)\}$ and $\mathrm{Im}\{P_r(\mathbf{r};\omega)\}$ of the reverberant pressure are zero-mean with Gaussian probability densities. The expected value of the square of the magnitude of the reverberant pressure

$P_\mathrm{r}(\mathbf{r};\omega)$ may be expressed

$$\mathcal{E}\{|P_\mathrm{r}(\mathbf{r};\omega)|^2\} = \mathcal{E}\left\{(\mathrm{Re}\{P_\mathrm{r}(\mathbf{r};\omega)\})^2\right\} + \mathcal{E}\left\{(\mathrm{Im}\{P_\mathrm{r}(\mathbf{r};\omega)\})^2\right\}, \tag{2.74}$$

where $\mathcal{E}\{\cdot\}$ denotes the mathematical expectation. Without any loss in generality, it can be assumed the direct pressure, $P_\mathrm{d}(\mathbf{r};\omega)$, has only a real part. The total measured pressure $P(\mathbf{r};\omega)$ is the vector sum of $P_\mathrm{d}(\mathbf{r};\omega)$ and $P_\mathrm{r}(\mathbf{r};\omega)$. Assuming that the direct and reverberant pressures are independent we have

$$\mathcal{E}\{|P(\mathbf{r};\omega)|^2\} = |P_\mathrm{d}(\mathbf{r};\omega)|^2 + \mathcal{E}\{|P_\mathrm{r}(\mathbf{r};\omega)|^2\}. \tag{2.75}$$

The acoustic intensity level relative to the expected acoustic intensity is

$$I(\mathbf{r};\omega) = 10\log_{10}\left(\frac{|P(\mathbf{r};\omega)|^2}{\mathcal{E}\{|P(\mathbf{r};\omega)|^2\}}\right) \quad [\mathrm{dB}]. \tag{2.76}$$

The standard deviation σ of the intensity level (commonly called *spectral deviation*) is a measure for the randomness of the spectral response [16] and is given by

$$\sigma(\mathbf{r}) = \left(\frac{1}{\omega_e - \omega_s}\int_{\omega_s}^{\omega_e}\left(I(\mathbf{r};\omega) - \bar{I}(\mathbf{r})\right)^2 \mathrm{d}\,\omega\right)^{0.5} \quad [\mathrm{dB}], \tag{2.77}$$

where ω_s ($\omega_s > 2\pi f_\mathrm{g}$) and ω_e denote the start- and end-frequency of the region of interest, and $\bar{I}(\mathbf{r}) = \frac{1}{\omega_e-\omega_s}\int_{\omega_s}^{\omega_e} I(\mathbf{r};\omega)\mathrm{d}\,\omega$ denotes the average intensity level.

The spectral response between the source and the microphone is defined by the ATF $H(\mathrm{r},\mathrm{r}_s,\omega)$. Therefore, we can calculate σ directly from the ATF by

$$\sigma(\mathbf{r},\mathbf{r}_s) = \left(\frac{1}{\omega_e - \omega_s}\int_{\omega_s}^{\omega_e}\left(10\log_{10}|H(\mathbf{r},\mathbf{r}_s;\omega)|^2 - \bar{H}(\mathbf{r},\mathbf{r}_s)\right)^2 \mathrm{d}\,\omega\right)^{0.5} \quad [\mathrm{dB}], \tag{2.78}$$

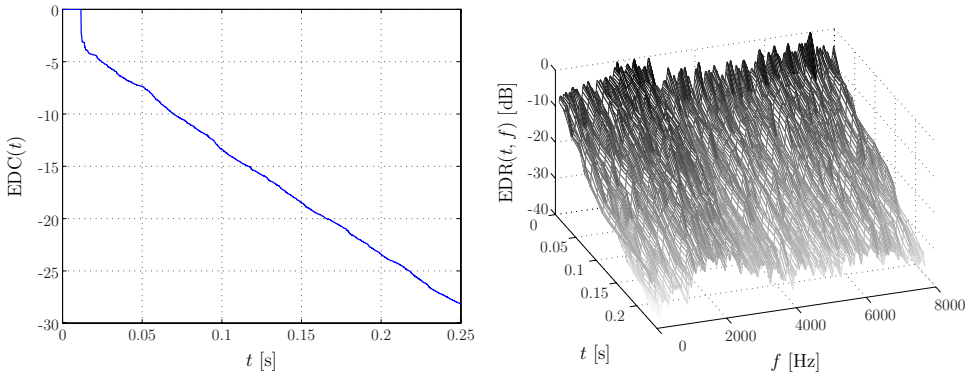
where $\bar{H}(\mathbf{r},\mathbf{r}_s) = \frac{1}{\omega_e-\omega_s}\int_{\omega_s}^{\omega_e} 10\log_{10}|H(\mathbf{r},\mathbf{r}_s;\omega)|^2\mathrm{d}\,\omega$.

Schroeder proved that σ has a theoretical upper-bound of approximately 5.57 dB [74].

2.8 Reverberation Time

Sabine's pioneering research [41] started the field of modern room acoustics and established many important concepts, most importantly the concept of reverberation time. Sabine determined that the reverberation time was proportional to the volume of the room and inversely proportional to the amount of absorption. Because the absorptive properties of materials vary as a function of frequency, the reverberation time varies as well.

The reverberation time can be measured by exciting a room to steady state with a noise signal, turning off the sound source, and plotting the resulting squared pressure



(a) (b)

Figure 2.5 *(a) Energy Decay Curve and (b) Energy Decay Relief of the AIR depicted in Fig. 2.8.*

as a function of time [75]. The time required for the resulting Energy Decay Curve (EDC) to decay 60 dB is defined as RT_{60}. The true energy decay curve can be obtained by integrating the impulse response of the room as follows: [76]

$$\mathrm{EDC}(t) = \int_t^\infty h^2(\tau)\, d\tau, \tag{2.79}$$

where $h(t)$ is the impulse response of the room which may be band-pass filtered to yield the EDC for some particular frequency band. The integral in Eq. 2.79, which is often called the Schroeder integral, computes the energy remaining in the impulse response after time t. As an example, the normalized energy decay curve[3] of the measured AIR shown in Fig. 2.8 is depicted in Fig. 2.5(a).

Jot [77] proposed a variation of the EDC that he called the Energy Decay Relief (EDR) or $\mathrm{EDR}(t, f)$. The EDR represents the reverberation decay as a function of time and frequency in a 3D plot. To compute it, the impulse response is divided into multiple frequency bands, the Schroeders integral is computed for each band, and the result are plotted as a 2D surface. As an example, the energy decay relief of the measured AIR shown in Fig. 2.8 is depicted in Fig. 2.5(b). Because the walls and air absorb the high frequencies more than the low frequencies the decay at high frequencies is faster than the decay at low frequencies. Consequently, the reverberation time at high frequencies is shorter than the reverberation time at low frequencies.

2.9 Excess-Phase

The acoustic transfer functions are usually non-minimum-phase. Since stable and causal inverses of non-minimum-phase systems do not exist, a problem investigated in

[3] Note that the EDC is often normalized with respect to the total energy of the AIR.

Section 3.3.4, it is important to known the contribution of the non-minimum-phase or excess-phase in Eq. 2.22 to the intelligibility of speech. Johansen and Rubak [78] have considered this question in detail trough a number of listening tests:

i) An anechoic speech signal is compared with the same signal filtered by the all-pass component of a real acoustic impulse response.

ii) An anechoic speech signal filtered by the minimum-phase component of a real AIR is compared with the same anechoic speech signal filtered by the complete impulse response.

The minimum-phase and all-pass component of Eq. 2.22 are extracted from a non-minimum-phase impulse response using the cepstrum method. Results from the first experiment indicate that the all-pass component affects the anechoic speech signal sufficiently for detection by the human ear. The second experiment indicates that if the excess-phase component is removed from the reverberant speech signal, there is still an audible difference, but less than the first experiment. These experiments suggest that the minimum-phase component, which contains the magnitude information, is able to partly mask the effect of excess-phase. However the ability of excess-phase to degrade speech quality is significant. The importance of phase in signals has been discussed in depth by Oppenheim and Lim [79]. Some of their comments are relevant here, particularly where they argue that phase reflects the location of events more than the magnitude. Acoustic distortion is mainly due to the arrival, or temporal locations, of early and late reflections. Much of this temporal information will be reflected in the phase response rather than the magnitude response and therefore it is important to consider excess-phase. The work in [78, 80] also reinforces the observation that the longer the reverberation time, the more excess-phase is present, and the lower the observed speech quality. Moreover, the longer the impulse response and the larger the distance between the source and the observer, the larger the degradation of the speech quality. As the distance between the source and observer increases, the measure of direct energy to reverberant energy, as discussed is Section 2.7.1, decreases rapidly, with a rate of decrease higher in rooms with small volume or long reverberation time. As a result it is crucial that dereverberant techniques do not neglect the excess-phase of the ATF. Further discussion of the importance of phase distortion due to non-minimum-phase systems may be found in a paper by Radlović and Kennedy [2].

2.10 Simulating room acoustics

Although this dissertation is primarily concerned with modelling the impulse response of a real room, it is instructive to consider room acoustic modelling methods which simulate the impulse response of a room. In this section a brief overview of various room acoustic modelling methods is presented.

Mathematically the sound propagation is described by the wave equation. An impulse response from a source to a microphone can be obtained by solving the wave

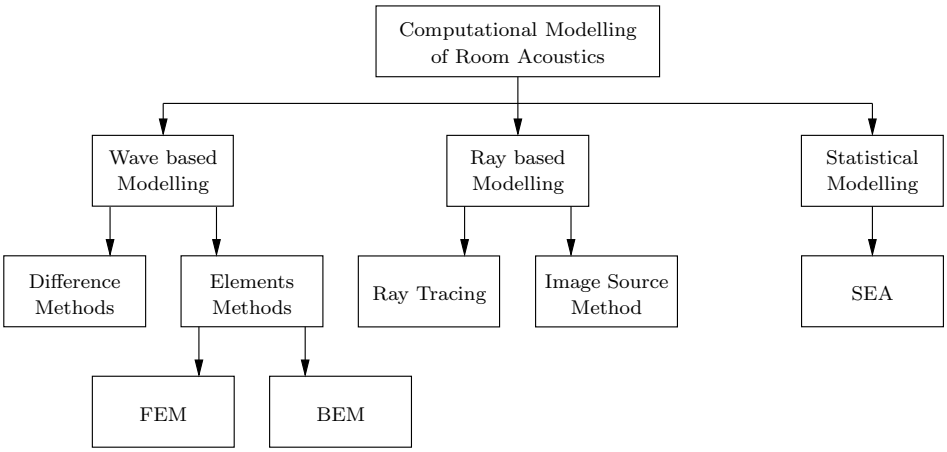


Figure 2.6 *Room acoustic models are based on sound rays (ray-based), on solving the wave equation (wave-based) or some statistical method [81].*

equation. Since it can seldom be expressed in an analytic form the solution must be approximated. There are three main modelling methods, as illustrated in Fig. 2.6, viz., wave-based, ray-based and statistical [81]. The ray-based methods, such as the ray-tracing [82] and the image-source method [83], are the most often used. The wave-based methods, such as the Finite Element Method (FEM), Boundary Element Method (BEM) [84, 85] and Finite-Difference Time-Domain (FDTD) [86] methods, are computational more demanding. In real-time auralization[4] the limited computation capacity requires simplifications. A frequently used simplification consists of modelling the direct path and early reflections individually and the late reflections by recursive digital filter structures. The statistical modelling methods, such as the Statistical Energy Analysis (SEA), have been widely used in aerospace, ship and automotive industry for high frequency noise analysis and acoustic designs. They are not suitable for auralization purposes since those methods do not model the temporal behaviour of a sound field.

Wave-based methods

The most accurate results can be achieved by the wave-based methods. An analytical solution for the wave equation can be found only in extremely simple cases such as a rectangular room with rigid walls. Therefore, numerical methods such as FEM and BEM [85, 84] are often used. The main difference between these two element methods is in the element structure. In FEM, the space is divided into volume elements, while in BEM only the boundaries of the space are divided into surface elements. The elements interact with each other according to the basics of wave propagation. The size of these elements has to be much smaller than the size of the wavelength for every

[4] Auralization is the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modelled space.

particular frequency. At high frequencies, the required number of elements becomes very high, resulting in a large computational complexity. Therefore, these methods are suitable only for low frequencies and small enclosures.

Another method for room acoustics simulation is provided by the FDTD method [86, 87]. The main principle of this method is that derivatives in the wave equation are replaced by corresponding finite differences. The FDTD method produces impulse responses that are better suited to auralization than FEM and BEM. The main benefit of the element methods over FDTD methods is that one can create a denser mesh structure where required, such as locations near corners or other acoustically challenging places.

In all the wave-based methods, the most difficult part is the definition of the boundary conditions. Typically a complex impedance is required, but it is hard to find that data in existing literature.

Ray-based methods

The ray-based methods are based on geometrical room acoustics [41], as described in Section 2.2. The most commonly used ray-based methods are the ray-tracing [82] and the image method [83]. The main difference between these methods is the way the reflection paths are calculated [81]. To model an ideal impulse response from a source to a receiver all possible sound reflection paths, commonly called rays, should be discovered. In ray-tracing methods the sound power emitted by a sound source is described by a finite number of rays. These rays propagate through space and are reflected after every collision with the room boundaries. During that time, their energy decreases as a consequence of the sound absorption of the air and of the walls involved in the propagation path. When the rays reach the receiver, an energy calculation process is performed. When all rays are processed the impulse response is obtained. Rays can be selected at random, based on a fixed interval or restricted to a given range of angles. Due to this the ray-tracing methods are by no means exhaustive, whereas the image method finds all the rays. However, while the image method is limited to geometries that are formed by planer surfaces the ray-tracing method can be applied to geometries that are formed by arbitrary surfaces.

It should be mentioned that all ray-based methods are based on energy propagations. This means that all effects involving phase differences such as refraction or interference are neglected. This is admissible if the sound signals of interest are not sinusoids or other signals with small frequency bandwidth but are composed of many spectral components covering a wide frequency range. Then it can be assumed that constructive and destructive phase effects cancel each other when two or more sound field components superimpose at a point, and the total energy in the considered point is simply obtained by adding their energies. Components with this property are often referred to as mutually incoherent [88].

The image method, which was developed by Allen and Berkley in 1979, is probably one

of the methods most commonly used in the acoustic signal processing community. In this dissertation we have used the image method to create synthetically reverberated signals. Details of this method, and an efficient implementation, can be found in Appendix A.

2.11 Acoustic Impulse Measurement

The acoustic impulse response is the main acoustical property of interest in dereverberation, and its measurement can be considered as a system identification problem.

Acoustic impulse responses can be acquired using White Gaussian Noise (WGN) sequences. Impulsive excitations are usually avoided since they are always approximated by short finite pulses and, furthermore, in order to attain a given Signal to Noise Ratio (SNR), the energy of the excitation must exceed a limit which could exceed the linear region of the devices, e.g., microphones and loudspeakers. Frequency sweeps are used when the impulse response must be measured in a short time, whereas long WGN sequences are required in order to attain a reasonable SNR.

The use of Maximum Length Sequences (MLS) forms a powerful method for the accurate determination of impulse responses in Linear Time-Invariant (LTI) systems. The method is based on the use of a deterministic pseudo-random stimulus, which is cross-correlated with the acquired response to yield the impulse response of the system under test. In literature, maximum length sequences are also known as pseudo-noise sequences, maximal-length shift register sequences or m-sequences. The MLS method is suitable for exceedingly long impulse responses, from which very finely scaled frequency responses can be calculated.

Binary maximum length sequences have a period of $2^N - 1$, where N is a positive integer. The sequences are generated recursively using a digital shift register with N binary elements (c.f. Fig. 2.7). The shift register output $x(n)$ is produced at the last register element. The output and up to three other register elements, commonly called taps, are summed together using the bitwise exclusive-or (XOR) function. The result is fed back into the first element after shifting the register contents one element towards the output. Only certain shift register tap and order combinations cycle through all of the $2^N - 1$ states, i.e., all states minus the null-state, and lead to maximum length sequences. Suitable tap combinations for common register lengths can be found in [89, 90].

In practical applications, the binary MLS is commonly mapped from values (0, 1) to symmetrical signal levels (1, -1), which is called a symmetrical MLS. Due to the sequence properties, the sum of a symmetrical MLS is always -1. This results in an important factor for impulse response measurements: the symmetrical MLS signal is practically AC coupled, and contains very little energy at DC.

The MLS has the desirable property that its normalized periodic auto-correlation

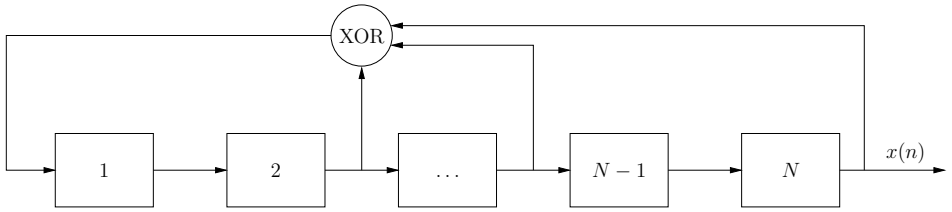


Figure 2.7 *Shift register for generating maximum length sequences.*

approximates a Dirac pulse, i.e.,

$$\begin{aligned}\tilde{r}_{xx}(m) &= \frac{1}{L}\sum_{n=0}^{L-1} x(n)x((n+m) \bmod L) \quad \text{for } |m| \leq L-1 \\ &= \hat{\delta}(m),\end{aligned} \tag{2.80}$$

where $L = 2^N - 1$, and

$$\hat{\delta}(m) = \begin{cases} 1, & \text{for } m = 0; \\ -\frac{1}{2^N-1}, & \text{for } 1 \leq |m| \leq L-1. \end{cases} \tag{2.81}$$

Eq. 2.81 can be written as

$$\hat{\delta}(m) = \frac{L+1}{L}\delta(m) - \frac{1}{L} \quad \text{for } 0 \leq |m| \leq L-1. \tag{2.82}$$

For large L the normalized periodic auto-correlation results in

$$\hat{\delta}(m) \approx \delta(m) \quad \text{for } 0 \leq |m| \leq L-1. \tag{2.83}$$

Which implies that the frequency spectrum of the MLS is approximately flat.

Due to this property the estimation of the AIR, denoted by $h(n)$, can simply be obtained by calculating the periodic cross-correlation between the transmitted MLS $x(n)$ and the received signal $y(n)$, i.e.,

$$\begin{aligned}\tilde{r}_{xy}(m) &= \frac{1}{L}\sum_{n=0}^{L-1} x(n)y((n+m) \bmod L) \quad \text{for } |m| \leq L-1 \\ &= \frac{1}{L}\sum_{n=0}^{L-1} x(n)\left(\sum_{j=0}^{L-1} h(j)x((n+m-j) \bmod L)\right) \\ &= \frac{1}{L}\sum_{j=0}^{L-1} h(j)\left(\sum_{n=0}^{L-1} x(n)x((n+m-j) \bmod L)\right) \\ &= \sum_{j=0}^{L-1} h(j)\tilde{r}_{xx}(m-j) \\ &= \hat{h}(m).\end{aligned} \tag{2.84}$$

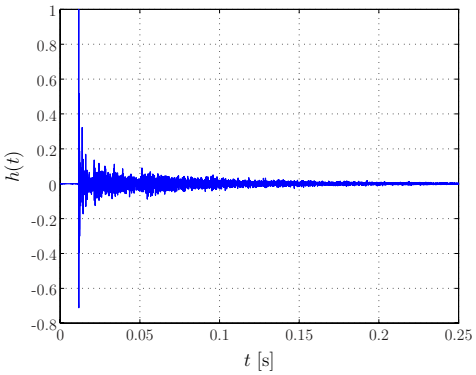


Figure 2.8 *Measured acoustic impulse response.*

Using Eq. 2.80 and 2.81 $\hat{h}(m)$ can be written as

$$\begin{aligned}\hat{h}(m) &= \sum_{j=0}^{L-1} h(j)\hat{\delta}(m-j) \\ &= h(m) + \frac{h(m) - \sum_{j=0}^{L-1} h(j)}{2^N - 1}.\end{aligned} \tag{2.85}$$

From this last equation it can be seen that the error depends on the order of the MLS and the AIR $h(m)$. Note that the periodic cross-correlation can be efficiently calculated in time domain by using the Fast Hadamard Transform [91], or in frequency domain by using the Fourier transform. The measurement noise can be reduced by averaging the impulse responses in time or frequency domain.

The AIRs that are used in this dissertation to create real reverberated signals are measured using the MLS technique. An example of a measured AIR in an office room with a reverberation time of 0.5 s is depicted in Fig. 2.8.

2.12 Summary

In this chapter we have introduced some basic theoretical acoustic properties which are important for understanding why particular models are used throughout this work. The suitability of some well-known models for the representation of room acoustics were discussed. Furthermore, the robustness of these models to variations in the source and observer position, and the effect of parameter variation on the accuracy of the model were discussed. Commonly used models include the pole-zero, all-zero, all-pole, and common-acoustical pole. However, determining which model is most robust to source and microphone movements, is an ongoing discussion.

The background of statistical room acoustics was provided. We showed when statistical room acoustics can be used, and summarized some important properties of the acoustic

transfer function that are used throughout this work.

Furthermore, the contribution of non-minimum-phase to the perception of reverberation was discussed, and it was observed that it is important that this component is not neglected since it contains most of the reverberant energy.

Finally, we showed how acoustic impulse responses can be simulated and measured in practice. In this dissertation we use the image method (see Appendix A for more details) to generate synthetic acoustic impulse responses. The real acoustic impulse responses were measured using maximum length sequences.

Chapter 3

Literature Survey

3.1 Introduction

In a patent [92] filed by Ryall in 1938, an electric signal amplifier was proposed for a two-way transmission system for voice-operated devices. In this patent the problem caused by reverberation was briefly discussed. It was noticed that reverberation was most severe at low frequencies. As such the amplification in this frequency range was limited. Since the early days of digital signal processing many dereverberation techniques have been proposed. Only recently a short review article on speech dereverberation techniques was published by Naylor and Gaubitch in [93]. However, since there is no comprehensive literature survey on speech dereverberation techniques available, we will categorize and review existing techniques in this chapter.

Reverberation reduction techniques can be divided into many categories. They may, for example, be divided into single- or multi-microphone techniques and into those primarily affecting colouration or those affecting late reverberation. We have categorized the reverberation reduction techniques depending on whether or not the acoustic impulse response needs to be estimated. We then obtain two main categories, i.e., *reverberation suppression* and *reverberation cancellation*. Techniques in the first category, reverberation suppression, do not require an estimate of the acoustic impulse response while techniques in the second category, reverberation cancellation, do require an estimate of the acoustic impulse response. Techniques within these categories can be divided into smaller sub-categories depending either on the amount of knowledge about the source or the acoustic channel that is utilized. In Fig. 3.1 the two main categories and sub-categories are depicted. The techniques used in the categories reverberation suppression and reverberation cancellation will be discussed in Sections 3.2 and 3.3, respectively.

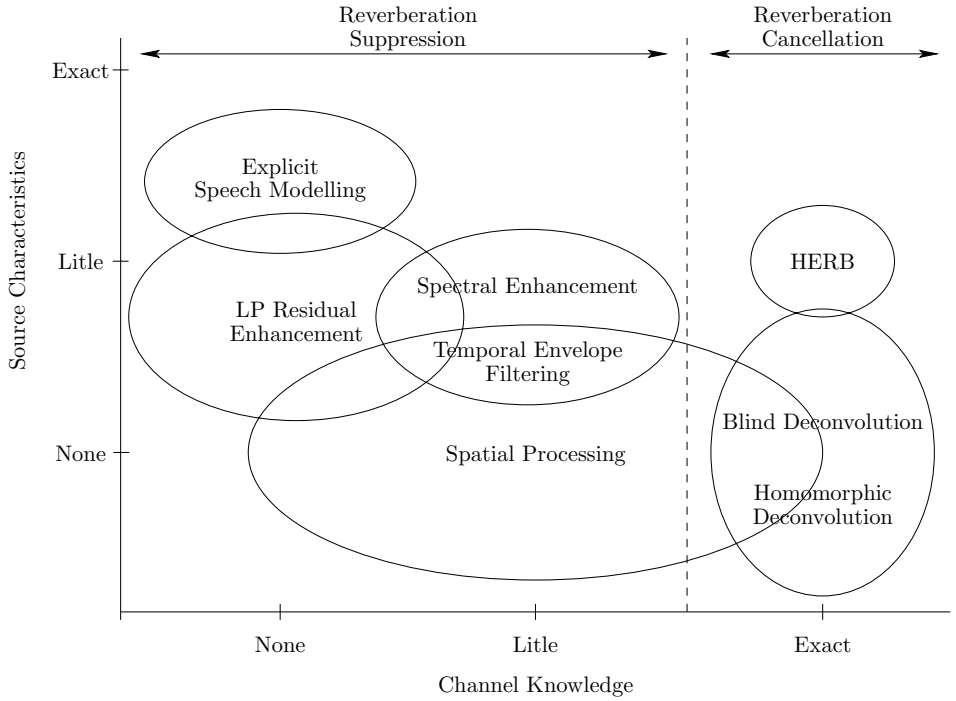


Figure 3.1 *Classification of various techniques used for speech dereverberation.*

3.2 Reverberation Suppression

In this section we will give a comprehensive overview of the reverberation suppression techniques that have been proposed until now. We have created smaller sub-categories by taking into account the amount of knowledge about either the source or the acoustic channel that is utilized, and by looking at the signal processing techniques that are involved.

3.2.1 Explicit Speech Modelling

Techniques that belong to this category exploit the underlying structure of the anechoic speech signal. Hardwick developed a Dual Excitation speech model in 1992, which was applied to the problem of speech enhancement in [94]. This model was extended by Yoo to a Generalized Dual Excitation speech model by taking pitch variations into account [95]. It should be noted that both models are mainly based upon the voiced speech segments.

Brandstein exploited the Dual Excitation model to model the speech signal in combination with spatial filtering to enhance reverberant speech in [96]. The Generalized Dual Excitation model was later used in [97].

Attias and Deng proposed a unified probabilistic framework for denoising and dereverberation of speech signals in [98]. The framework transforms the denoising and dereverberation problems into Bayes-optimal signal estimation. The key idea is to use a strong speech model that is pre-trained on a large data set of anechoic speech. The framework applies equally well to single- and multiple-microphone cases. Experiments show that the optimal estimation can outperform standard techniques such as the spectral subtraction in terms of noise suppression. Unfortunately the dereverberation performance was not evaluated separately. A disadvantage of this technique is that the result strongly depends upon the model training.

3.2.2 LP Residual Enhancement

The source-filter production model is often used for modelling speech [99]. The model describes speech production in terms of an excitation sequence exciting a time-varying all-pole filter. The excitation sequence consists of random noise for unvoiced speech and quasi-periodic pulses for voiced speech, while the filter models the human vocal tract. The all-pole filter coefficients can be estimated through Linear Prediction (LP) analysis of the recorded speech and are commonly called Linear Prediction Coefficients (LPC). The excitation sequence, or LP residual, can be now obtained by inverse filtering of the speech signal. The motivation for the proposed techniques is the observation that in reverberant environments, the LP residual of voiced speech segments contains the original impulses followed by serval other peaks due to multi-path reflections. Furthermore, an important assumption is made that the LPCs are unaffected by reverberation. Consequently, dereverberation is achieved by attenuating the peaks in the excitation sequence due to multi-path reflections, and synthesizing the enhanced speech waveform using the modified LP residual and the time-varying all-pole filter with coefficients calculated from the reverberant speech.

In Fig. 3.2 a general structure for multi-microphone LP residual enhancement is depicted. Here $\mathbf{x}(n)$ contains the samples of M microphones at discrete time n. In the LP coefficient analysis stage the poles of the time-varying all-pole filter $\hat{\mathbf{a}}(l)$ are estimated (here l denotes the time frame index), and the M LP residuals $\tilde{\mathbf{e}}(n)$ are constructed. The LP residuals are used to estimate the clean LP residual $\hat{e}(n)$. Then the estimated clean LP residual and the estimated poles are used to synthesize the speech signal $\hat{s}(n)$.

The first speech dereverberation algorithms using the LP residual enhancement technique were most likely proposed by J.B. Allen and F. Haven, from Bell Telephone

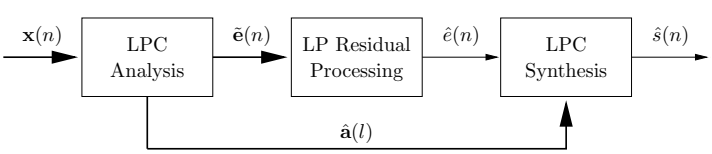


Figure 3.2 *General structure for multi-microphone LP residual enhancement.*

Laboratories Inc., in a patent that was filed in 1972 [100]. In this patent both single and multi-microphone solutions were proposed. They used a detector to distinguish between voiced and unvoiced speech frames, a pitch estimator, and a gain estimator. All signals were then used to synthesize a clean LP residual. Using the estimated vocal tract an estimate of the anechoic speech signal was constructed.

LP residuals were also used by Griebel and Brandstein in 1999. In [101] a technique for enhancing multi-channel reverberant speech using event-based processing of wavelet transform coefficients was proposed. Clustering of the wavelet extrema across multiple channels is employed to obtain a single multi-scale extrema representation from which the enhanced signal is synthesized. Processing is done in the LP residual domain, with the entire analysis being preceded by a multi-channel LPC inverse filter and followed by the corresponding forward LPC filter. The algorithm was compared to delay and sum beamforming and results were presented for reverberant and noisy conditions. By focusing on event-based data and a wavelet-domain analysis, the proposed algorithm is capable of discriminating impulses of the excitation residual generated by the desired speech signal from those brought about by multi-path echoes and uncorrelated noise. The enhanced speech derived from an all-pole synthesis with the clean excitation sequence demonstrates a robustness to environmental reverberation and additive noise.

Another multi-channel dereverberation technique is proposed by Griebel and Brandstein in [102]. They used a coarse channel modelling to modify the LP residuals of the channel data. Specifically, the incorporated channel model requires only approximate time and amplitudes of the initial multi-path reflections.

Yegnanarayana and Murthy proposed a single microphone technique to dereverberate speech [103, 104], and provided a comprehensive study on the effects of reverberation on the LP residual. The technique is based on analysis of short (2 ms) segments of data to enhance the regions in the speech signal having low Signal to Reverberation Ratio (SRR) components. The short segment analysis shows that SRR is different in different segments of speech. The processing technique involves identifying and manipulating the LP residual in three different regions of the speech signal, namely, high SRR region, low SRR region and only reverberation component region. A weighting function is derived to modify the LP residual. The weighted residual samples are used to excite the time-varying LP all-pole filter to obtain perceptually enhanced speech.

Experiments performed by Gillespie [105] showed that the kurtosis of the LP residual is a reasonable measure of reverberation. The LP residual signal becomes more Gaussian due to reverberation, consequently, the kurtosis becomes smaller. In [105] the microphone signals are processed by a sub-band adaptive filtering structure using a Modulated Complex Lapped Transform (MCLT), in which the sub-band filters are adapted to maximize the kurtosis of the LP residual of the reconstructed speech. In this way, they attain good solutions to the problem of blind speech dereverberation. Experimental results with actual data, as well as with artificially difficult reverberant situations, show very good performance, both in terms of a significant reduction of the perceived reverberation, as well as improvement in spectral deviation. However, the calculation of the kurtosis and its derivative are prone to instability [106, 107]. In

order to reduce this sensitivity, a single-channel blind dereverberation algorithm that uses a maximum likelihood approach to estimate the inverse filter was proposed by Tonelli et al. in [106], and was recently extended to multiple channels [108]. Their simulation results showed that good dereverberation is achieved even in real room, although pre-processing might be necessary, particularly for widely spaced microphones. Both the kurtosis and maximum likelihood based algorithms require sufficient spacing between the microphones.

Yegnanarayana et. al. proposed a multi-channel speech enhancement technique in [109] which is based on exploiting the features of the excitation source in speech production. The most important property is that in voiced excitation the strength of excitation is largest around the instant of glottal closure. The Hilbert envelope of the LP residual was used to derive the information of the strength of excitation. A weight function was derived by coherently combining the delay compensated Hilbert envelopes of the LP residual signals from the different microphones. The enhanced speech was again obtained by exciting the time-varying all-pole filter with the LP residual modified by the weight function. The reverberation effects are reduced significantly, however, the proposed technique introduces a significant amount of speech distortion.

The techniques proposed by Griebel, Yegnanarayana and Gillespie reduce the effects of reverberation, but do not consider the original structure of the excitation signal. The enhanced residual can differ from the original clean residual and can result in less natural sounding speech. Gaubitch and Naylor proposed to enhance the LP residual from the output of a delay and sum beamfomer [110]. They used the fact that the waveform of the LP residual between adjacent larynx-cycles[1] varies slowly, so that each such cycle can be replaced by an average of itself and its nearest neighbouring cycles. The averaging results in a suppression of spurious peaks in the LP residual caused by room reverberation. In this paper only voiced speech segments were addressed.

Above techniques rely on the observation that in reverberant conditions the LP residual contains the original excitation impulses followed by several other peaks due to reverberation. Moreover, they rely on the important assumption that the calculated LP coefficients of the all-pole filter are unaffected by the multi-path reflections of the room. Gaubitch and Naylor showed that this latter assumption holds only in a spatially averaged sense [111], and that it can not be guaranteed at a single point in space for a given room. More recently Gaubitch et al. used statistical room acoustic theory for the analysis of the Auto Regressive (AR) modelling of reverberant speech [69]. They investigated three scenarios, and showed that in terms of spatial expectation, the AR coefficients calculated from reverberant speech are approximately equivalent to those from anechoic 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 anechoic speech coefficients due to spatial correlation, which is governed by the room characteristics and the microphone array arrangement. This difference decreases as the distance between adjacent microphones is increased.

[1] The larynx-cycle starts when the glottis opens, and ends when the glottis closes. The length of a larynx-cycle is approximately 20 ms.

It was also demonstrated that AR coefficients calculated jointly from the M-channel observation provide the best approximation of the anechoic speech AR coefficients at individual source-microphone positions and in particular when the microphone separation is small < 0.3 m. Thus, in general, the M-channel joint calculation of the AR coefficients is the preferred option, specifically in the case of closely spaced microphones. It should be noted that all analyses have been performed using a single vowel, i.e., the effects of windowing, self-masking, and overlap-masking, are not considered in this paper. It is expected that proper calculation of the LP coefficients, i.e., using spatially averaged LP coefficients, improves the quality of earlier published LP residual enhancement techniques.

3.2.3 Temporal Envelope Filtering

The temporal envelope filtering techniques try to model the relation of the envelopes of the anechoic and reverberant speech waveforms, aiming at the enhancement of single microphone recorded reverberant speech. Most techniques were motivated by studies on the effect of reverberation on the Modulation Index (MI) of speech and the reduction of intelligibility in reverberant environments [112]. The speech signal is amplitude modulated, the modulation index, also called modulation depth, indicates by how much the modulated variable varies around its 'original' level. Tails produced by past acoustic events fill in low energy regions between consecutive sounds reducing the modulation depth of the original envelope and thus modifying its MI. A general structure used for temporal envelope filtering is depicted in Fig. 3.3. In the first stage the temporal envelope is extracted. In some cases the obtained envelope is used to estimate the required parameters, e.g., the reverberation time. In the next stage the envelope signal is filtered to construct an estimate of the anechoic envelope. In the final stage the anechoic speech signal is reconstructed using the fine structure of the reverberant signal. Notice that in this case the phase modifications in the fine structure are not considered.

Berkley and Mitchell (Bell Telephone Laboratories Inc.) filed a patent for automatically reducing reverberation in typical voice telecommunications systems in 1978 [113]. This system uses center clipping levels adaptive to the level of reverberation input speech. In one configuration, the speech spectrum is divided into two sub-bands, and

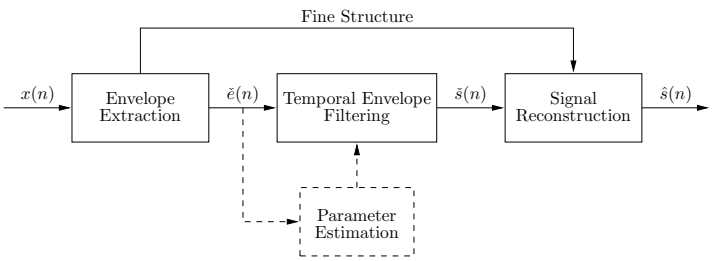


Figure 3.3 *General structure for temporal envelope filtering.*

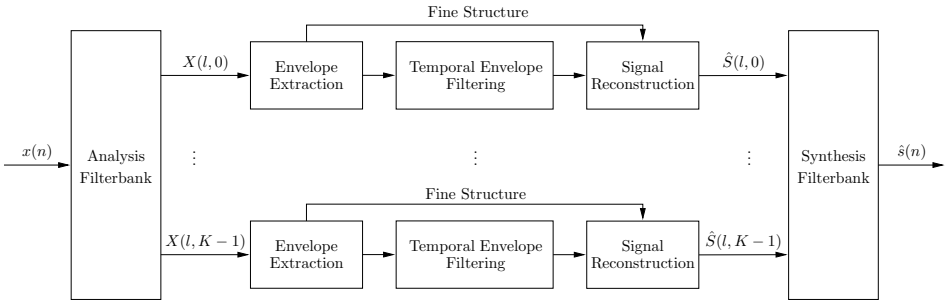


Figure 3.4 *Temporal envelope filtering in K sub-bands.*

center clipping was applied only to the lower band. They suggested to use a clipping-level holdover circuitry with exponential decay, which appears to work well for a large variety of reverberant enclosures. The complete system was implemented using analog circuits and was based on the temporal envelope of the signal.

Langhans and Strube proposed an enhancement technique for speech signals corrupted by reverberation or noise where they appropriately filtered the envelope signals in critical frequency bands based on short-time Fourier transform (STFT) [114]. They used a theoretically derived inverse Modulation Transfer Function (MTF) as high-pass filtering. The MTF of a reverberant room can be derived from a single impulse response and is defined as:

$$M(\omega) = \frac{\int_0^\infty h^2(t) \mathrm{e}^{\iota\omega t} \mathrm{d}t}{\int_0^\infty h^2(t) \mathrm{d}t}, \tag{3.1}$$

where $\iota = \sqrt{-1}$. Using Eq. 2.43 and 2.37 it can be shown that [115]:

$$|M(\omega)| = \left(1 + \left(\omega \frac{\mathrm{RT}_{60}}{13.8}\right)^2\right)^{-\frac{1}{2}}. \tag{3.2}$$

Another attempt was made by Hirsch using ad hoc high-pass filtering [116]. Avendano and Hermansky [117] used a training sequence to estimate the temporal envelope filters for each sub-band. The filters are applied to the short-term power spectrum trajectories of speech based on the STFT. Although an audible reduction of reverberation is achieved some severe artifacts are introduced.

Mourjopoulos showed that reverberation reduction can be achieved by envelope deconvolution in each frequency sub-band, which recovers the envelope of the anechoic signal from the measured speech envelope [118]. The final reconstruction of the speech waveform was performed using the original phase function. The structure of the proposed scheme is depicted in Fig. 3.4. In this work Mourjopoulos assumed that the inverse filter could be estimated in advance. Later, Hirobayashi et al. proposed the power envelope inverse filtering technique [119]. This technique differs from Mourjopoulos in the signal definition of the envelope (amplitude or power) and the carrier (sine-wave or

white noise) based on the amplitude modulation representation. Hirobayashi derived the following inverse filter using Eq. 2.43

$$H_{\text{inv}}(\mathrm{z}) = \sigma_b^2 \left(1 - \mathrm{e}^{-\frac{13.8}{\mathrm{RT}_{60} f_s}} \mathrm{z}^{-1}\right), \tag{3.3}$$

where $\mathrm{z} = \mathrm{e}^{\iota\omega t}$ and σ_b^2 denotes the variance of $b(t)$ in Eq. 2.43. Unfortunately only synthetic speech signals and impulse responses were used for testing.

The techniques proposed by Hirobayashi and Mourjopoulos in [118, 119] try to restore the temporal envelope from reverberant speech. The techniques in [114, 116, 117] attempted to restore the modulation index of reverberant speech to suppress the degradation of speech intelligibility caused by reverberation.

Recently Unoki et al. proposed an improved technique based on the MTF concept for restoring the power envelope from a reverberant speech signal [120, 121]. They used a similar structure as depicted in Fig. 3.4 and proposed a technique to blindly estimate the required parameters, i.e., σ_b^2 and the reverberation time RT_{60}, to construct the inverse filter. They concluded that the enhancement of the fine structure, i.e., phase or carrier information, should not be disregarded but needs to be addressed to improve speech intelligibility.

3.2.4 Spectral Enhancement

The spectral enhancement techniques discussed in this section achieve dereverberation by modifying the short-time spectrum of the received microphone signal(s). This technique dates back to the work of Allen et al. in 1977 [122].

Allen et al. [122] used sound recordings of two microphones, and processing was performed in the frequency domain. The process uses a sub-band technique and, within each band, the delay existing between the 'coherent part' of the two microphone signals (i.e., the actual signal and early reflections but not the late reflections) is removed by shifting the phase of one signal to align it with the other. These phase corrected signals are then summed and the entire process is referred to as co-phase and add in bands. The gain of each band is then adjusted according to the normalized cross-correlation function of the observed signals, as was devised and implemented in an analog system by Danilenko in 1968 [123]. This has the effect of attenuating bands with low-levels of 'coherence' containing mainly reverberation, while passing relatively unaltered, or slightly enhanced, frequency bands with a strong level of 'coherence' implying the presence of a strong direct component and early echoes. This technique was evaluated by normal and impaired listeners for its ability to enhance the intelligibility of isolated words recorded in a small room with relative long reverberation time of 1.3 s. In general no significant change in the average recognition score was observed [124]. Bloom improved the output by using narrower analysis bands and by deriving a gain function from a time-varying estimate of the magnitude-squared coherence function, which was smoothed in frequency domain on a critical-band basis [125]. He concluded that

these modifications do not necessarily improve the average recognition score. Similar techniques are used by Hussain in [126, 127].

Lebart et al. proposed proposed a single-microphone spectral enhancement technique for speech dereverberation [24]. An estimate of the late reverberant energy was obtained directly from the microphone signal, and dereverberation was achieved by using spectral subtraction. The proposed estimator is based on Polack's statistical reverberation model (see Section 6.3), and only requires an estimate of the reverberation time. Lebart et al. assumed that the reverberation time was frequency independent, and implicitly assumed that the energy related to the direct sound could be ignored.

Recently, Wu et al. proposed a two-stage approach for multi-microphone dereverberation [128]. In the first stage the LP residual enhancement technique proposed by Gillespie [105] was used to enhance the Direct to Reverberation Ratio (DRR). In a second stage spectral subtraction was used to reduce late reverberation. They used a heuristic function to estimate the late reverberant energy, thereby assuming that the first stage was able to reduce a significant amount of reverberation. In [129] a single-microphone solution was proposed by the same authors using a similar two-stage approach.

3.2.5 Spatial Processing

Spatial Processing techniques can be used for multi-microphone speech dereverberation. The signals can be manipulated to enhance or attenuate signals emanating from particular directions. Using these techniques the reverberant part can be spatially separated from the direct signal. Most techniques require some *a priori* knowledge of the position of the source.

Spatial processing techniques can be classified using all kinds of properties, e.g. array configuration, filter design and algorithms. An extensive overview can be found in 'Optimum Array Processing' written by Van Trees [130]. A rather educational tutorial, focusing on the enhancement of speech signals, has been written by McCowan [131]. The majority of research has been focusing upon enhancing the robustness in noisy environments. Experiments and analyses of spatial processing techniques usually do not focus on dereverberation capabilities, which makes the comparison difficult.

Fixed Beamforming

Microphone arrays achieve directionality by exploiting the fact that an incoming acoustic plane wave will generally arrive at the different microphones at slightly different times. The frequency components of these sounds could either reinforce or cancel, depending on the angle of arrival, the frequency of the component, the distance between the microphones, and the geometry of the microphone array. As a result, by summing the microphone outputs, the array develops a directional response that favours some

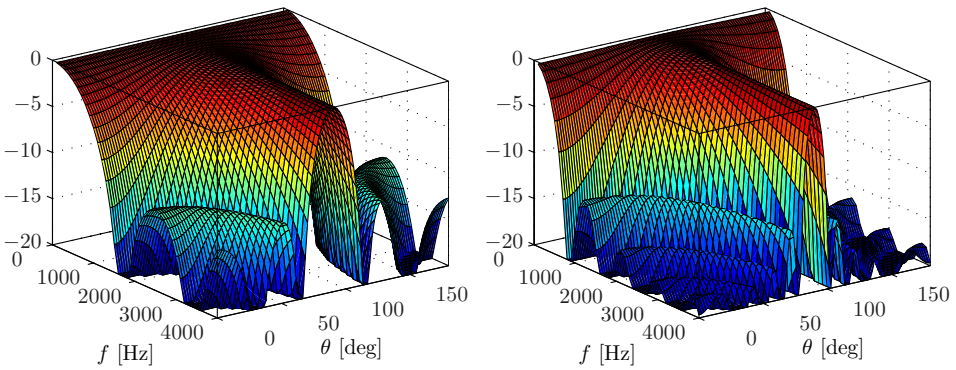


Figure 3.5 *Directivity patterns of uniform linear microphone arrays over a range of frequencies, (a) using an array of 4 elements, and (b) using an array of 8 elements.*

directions over others. Fig. 3.5 shows some sample response directivity patterns from 4-element and 8-element uniform linear arrays with microphones separated by 5 cm. Note that the largest response is always at 90 degrees, and that as the frequency increases, the width of the major response beam becomes narrower.

The arrival direction that produces the largest response does not need to be perpendicular to the axis of a linear array, as it normally would be for incident plane waves. By inserting suitable delays into each channel, an array can be designed to have its maximal response in any desired direction. The added delays ensure that, for a particular frequency, the source signal that arrives from the desired look direction at the microphones, are coherently added. This simple type of array processing is called delay and sum beamforming. Delay and sum beamforming is popular because it is easy to implement, because the directional response is stable over all environmental conditions, and because the directional response is unaffected by reverberation. Nevertheless, directional selectivity for a given number of microphones is relatively modest, as doubling the number of microphones increases the Signal to Noise Ratio (SNR) at the output of the delay and sum beamformer by only 3 dB [132]. Gaubitch and Naylor used tools from statistical room acoustics in order to predict the expected improvement in DRR at the beamformer output compared to the best microphone, which is normally the microphone closest to the source[68]. The improvement, denoted by ΔDRR, is given by

$$\begin{aligned}\Delta\text{DRR} &= 10\log_{10}\left(\frac{\mathcal{E}_\theta\{\text{DRR}_{\text{dsb}}\}}{\mathcal{E}_\theta\{\text{DRR}_{\max}\}}\right)\\ &= 10\log_{10}\left(\frac{D_{\min}\sum_{m=0}^{M-1}\sum_{n=0}^{M-1}\frac{1}{D_m D_n}}{\sum_{m=0}^{M-1}\sum_{n=0}^{M-1}\frac{\sin\left(k||\mathbf{r}_m-\mathbf{r}_n||\right)}{k||\mathbf{r}_m-\mathbf{r}_n||}\cos\left(k(D_m-D_n)\right)}\right),\end{aligned} \tag{3.4}$$

where $\mathcal{E}_\theta\{\cdot\}$ is the spatial expectation, k denotes the wave number, $\text{DRR}_{\max}$ is the

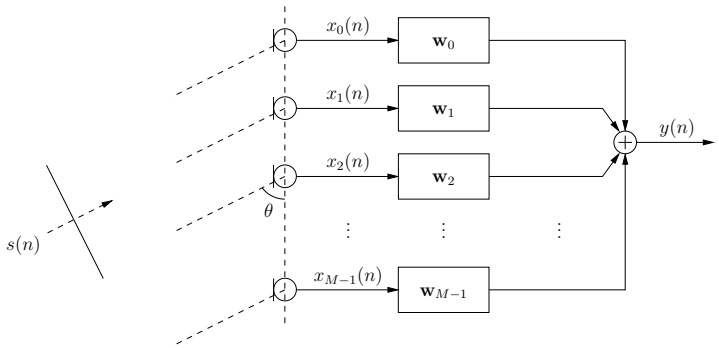


Figure 3.6 *Filter and Sum Beamformer structure, with a plane wave arriving from an angle θ.*

DRR of the microphone closest to the source (with distance $D_{\min}$) and DRR_{dsb} is the DRR at the output of the delay and sum beamfomer. The distance between the source and the m^{th} microphones is denoted by D_m. From the derived expression it could be seen that the relative improvement depends only on the microphone spacing and on the distance of the source from the array, where the effect of the latter decreases as the distance is increased. Thus, for a given geometric setup the DRR improvement is independent of the reverberation time. Simulation results were presented to confirm the validity of the derived expression.

In addition, the frequency dependence of the beam shape seen in Fig. 3.5 poses a number of problems. One is that the beam narrows as frequency increases. This can be addressed, for example, by combining nested delay and sum beamformers with different microphone spacing to produce an array with relatively constant beam width over a wide frequency range [133], a technique known as sub-array beamforming. Another consequence of the frequency dependence of the beam shape is that the frequency response depends on the arrival angle, which means that a desired speech signal arriving slightly 'off-axis' will be subjected to spectral colouration by the array. Since most current Automatic Speech Recognition (ASR) systems involve some type of spectral pattern matching, this colouration causes a mismatch between input speech and typical ASR training data, which in turn reduces ASR accuracy. A general set of solutions to these problems is available through the use of filter and sum beamforming, in which the delays of the delay and sum beamformer are replaced by linear filters. In Fig. 3.6 such a system is depicted. The m^{th} filter of length L is represented by $\mathbf{w}_m = [w_m(0)\ w_m(1)\ \ldots\ w_m(L-1)]^T$. The output of the beamformer is then constructed using

$$y(n) = \sum_{m=0}^{M-1} \mathbf{w}_m^T \mathbf{x}_m(n), \tag{3.5}$$

where $\mathbf{x}_m(n) = [x_m(n)\ x_m(n-1)\ \ldots\ x_m(n-L+1)]^T$. In principle, these filters can impose different delays or phase shifts at different frequencies, which permits a much wider and more flexible set of directivity patterns than is possible with delay and sum processing, at the cost of a much greater number of free parameters to be

determined. A comprehensive study on the design of robust filters for linear and non-linear microphone arrays can be found in [134].

Using superdirective beamforming techniques it is possible to achieve superdirectivity, i.e., spatial selectivity greater than what is obtained with conventional delay and sum beamforming. These techniques are based upon the maximization of the array gain, or directivity index, for a well-defined noise field. However, while superdirectivity can be achieved, the actual response can be extremely sensitive to practical problems such as coefficient errors, incorrect assumptions about the environment, and misalignment of sensor response or placement. For speech processing applications, superdirective techniques are useful for obtaining acceptable spatial selectivity at low frequencies for realistic array dimensions, especially when a so-called endfire array is used. Bitzer experimentally demonstrated that the ASR performance in a reverberant environment using a standard superdirective beamformer is superior to the delay and sum beamformer [135].

Adaptive Beamforming

Fixed Beamformers are easy to implement but have the obvious disadvantage that they cannot deal with a changing acoustic environment. In adaptive beamforming, the array-processing parameters are dynamically adjusted according to some optimization criterion, either on a sample-by-sample or frame-by-frame basis. Most commonly, the relevant parameters are the coefficients of FIR filters used in a filter and sum beamformer. Typically the goal of the adaptation algorithm is to maintain a fixed response to signals arriving from a desired 'look' direction, while minimizing the overall energy of the filter output. The filter can accomplish this by positioning null responses in the directions of interfering noise sources. Popular adaptive array algorithms include maximum *a posteriori* beamforming, Minimum Variance Distortionless Response (MVDR) beamforming [136], Linear Constrained Minimum Variance (LCMV) beamforming, maximum SNR beamforming [137, 138], and linear predictive beamforming. The MVDR beamforming is also known as optimum beamforming. LCMV beamforming is developed from MVDR beamforming with additional linear constraints to improve its robustness [130]. In MVDR beamforming, the directivity pattern is formed to maximize the output signal to interference plus noise ratio while maintaining a constant gain in the direction of the desired signal. The MVDR beamforming is sensitive to Direction of Arrival (DOA) estimation errors and its performance decreases significantly when an interferer is inside the mainlobe [130]. The LCMV beamforming can be implemented by placing nulls in the directions of interferers when multiple interferers are considered. One limitation of the LCMV beamforming is that the number of microphones has to exceed the number of nulls by one. An efficient implementation of a LCMV beamformer is the classical Griffiths & Jim Generalised Sidelobe Canceller (GSC) [139] which is depicted in Fig. 3.7. The GSC consists of two structures: a fixed beamformer ($\mathbf{w}_q$) which produces a non-adaptive output, and an adaptive structure for sidelobe cancelling. The adaptive structure of the GSC is preceded by a blocking matrix $\mathbf{C}_a$, which blocks signals coming from the look direction. The weights of the

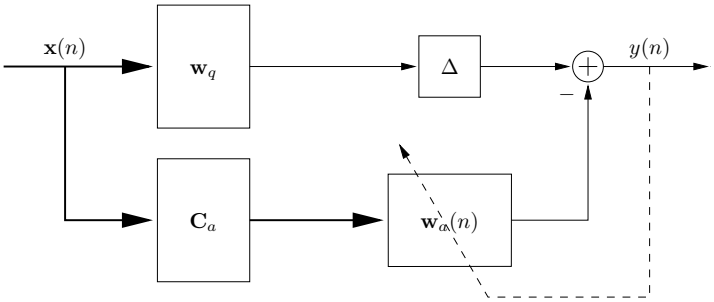


Figure 3.7 *The generalized sidelobe canceller, which is an efficient implementation of the LCMV beamformer.*

adaptive structure ($\mathbf{w}_a(n)$) are then adjusted to cancel any signal common to both structures.

Adaptive Beamforming and Reverberation

Unfortunately, many traditional adaptive algorithms become ineffective in reverberant environments. This is because the algorithms assume that the input to the system consists of a desired signal in the presence of statistically independent noise or other interference. In reverberation, the 'distortion' consists of an ensemble of attenuated and delayed replicas of the desired signal, which violates the assumption of independence between signal and noise. As an example, in case of the GSC algorithm, the leakage of the desired signal into the sidelobe cancelling path due to multi-path reflections will distort or cancel the desired signal. However, this problem can be reduced, but not solved, in case adaptation is performed in noise-only segments using a Voice Activity Detection mechanism. Hoshuyama et al. proposed to use an adaptive blocking matrix to reduce signal leakage, c.f. [140].

A more general beamformer technique, called the Transfer Function Generalized Sidelobe Canceller (TF-GSC), was proposed by Gannot et al. In [141] they derived a solution for arbitrary transfer functions, rather than relying on the assumptions that the received signals are simple delayed versions of the source signal. A sub-optimal solution was proposed using transfer function ratios that were estimated online. The blocking matrix was constructed using the same transfer function ratios, thereby significantly reducing the leakage of the desired signal. Although this solution can be used in a moderate reverberant environment it should be noted that it does not reduce the amount of reverberation.

Affes et al. [137] used a maximum SNR beamforming approach and replaced the fixed beamformer in the GSC by an adaptive beamformer which maximizes the SNR. The main advantage of the maximum SNR approach is that early reflections are coherently added to the desired speech signal, thereby increasing the DRR. Recently Warsitz and Haeb-Umbach [138] proposed a stochastic gradient ascent algorithm to maximize the SNR at the output of the filter and sum beamformer. Unfortunately the

dereverberation performance of such an approach has not been verified.

Bitzer et al. showed the theoretical amount of noise reduction obtained by a classical Griffiths & Jim GSC as a function of reverberation time [142]. They concluded that little reduction is observed for reverberation times greater than 200 ms. In a reverberant environment coherent noise fields trend to diffuse fields. Unfortunately the GSC is fundamentally unable to reduce diffuse noise sources.

Post-filtering

In practice, the basic filter and sum beamformer or adaptive beamformer, seldom exhibits the theoretical performance limits, e.g., due to incorrect assumptions about the environment, and misalignment of sensor response or placement. Furthermore, the level of improvement that can be obtained by the beamformer is often lower than the required improvement and further enhancement is desired. One technique of improving the system performance is to add a post-filter to the output of the beamformer. The post-filter enhances the beamformer output in the following ways:

- The post-filter suppresses any incoherent noise.
- The post-filter further enhances the beamformer's rejection of coherent correlated or uncorrelated noise sources not emanating from the steered direction.
- The post-filter displays robustness to minor steering errors.

Marro et al. investigated the effects on noise reduction and dereverberation of microphone arrays using post-filtering [143]. The experimental results show that for the reverberation reduction alone, the improvement yielded by the post-filter is limited. In summary, it is found that the effectiveness of such a post-filter follows that of the beamformer - if the beamformer is effective, the post-filter will further improve the system output. However, in the case where the beamformer is ineffective, the post-filter, being intrinsically linked to the beamformer performance, will be similarly ineffective. Cohen et al. proposed a multi-microphone post-processor in [144, 145] for the TF-GSC proposed by Gannot in [141]. The proposed post-processor is designed only to enhance the noise suppression, and does not reduce reverberation.

3.3 Reverberation Cancellation

The dereverberation problem can be viewed as the inverse filtering of the acoustic impulse response. In the usual formulation of the deconvolution problem, it is assumed that the system input $s(t)$ and system output $x(t)$ are both known. In the case of dereverberation and many other physical cases the system input is unknown. It is in situations of this kind that we speak of blind deconvolution (see Fig. 3.8). A good overview of existing blind deconvolution techniques can be found in [146, 147].

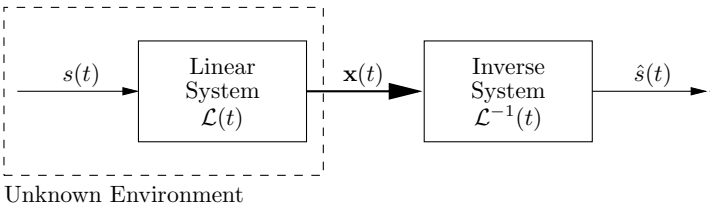


Figure 3.8 *Block diagram illustrating the background to the blind deconvolution problem.*

There are two distinct techniques to this problem of blind deconvolution: [148]

- Estimate $s(t)$ directly, or the parameters and excitation of an appropriate parametric model, as a *missing data* problem by treating the parameters of the system $\mathcal{L}(t)$ as nuisance parameters.
- Model the linear system $\mathcal{L}(t)$, estimate the parameters of the system $\mathcal{L}(t)$ by treating $s(t)$ as a nuisance parameter, and then deconvolve $x(t)$ with $\mathcal{L}^{-1}(t)$ to recover $s(t)$.

These techniques are in general based on simplistic source signal models. Many blind deconvolution techniques assume the source signal is contained within a finite support [149, 150, 151] and that its samples are independent and identically distributed (i.i.d.) [150, 152]. However, when the source signal is highly correlated, these techniques cannot be directly applied. Furthermore, many techniques assume quasi-stationarity of the system, and do not take global non-stationarity into account. Utilizing the global non-stationarity of the system allows the identification of system characteristics which may otherwise be unattainable.

In a Single-Input Single-Output (SISO) system the problem is under-constrained and can only be solved by incorporating varying degrees of prior knowledge regarding $s(t)$ and $\mathcal{L}(t)$, e.g., exploiting time diversity. A characteristic of blind deconvolution is that the source signal and impulse response of the distortion operator must be irreducible for unambiguous deconvolution [153]. An irreducible signal is one that cannot be expressed as the convolution of two or more signal components, on the understanding that the delta function is not a signal component. Suppose the channel, $h(t)$, is Linear Time-Invariant (LTI), then the observed signal may be expressed as, $x(t) = h(t) * s(t)$, where $*$ denotes convolution. If either $h(t)$ or $s(t)$ are reducible such that $h(t) = h_1(t) * h_2(t)$, and $s(t) = s_1(t) * s_2(t)$, then $x(t) = h_1(t) * h_2(t) * s_1(t) * s_2(t)$, and it is impossible to decide which component belongs to the source signal or to the distortion operator without additional knowledge. Consequently, many linear systems become reducible when they are considered stationary, and blind deconvolution is impossible. However, if, in fact, $s(t)$ and $\mathcal{L}(t)$ are both quasi-stationary and locally reducible, but possess different rates of global non-stationarity, then $s(t)$ and $\mathcal{L}(t)$ are no longer globally reducible and, therefore, in this case blind deconvolution is possible. In a Single-Input Multi-Output (SIMO) system it is also possible to exploit the spatial diversity of the received signals.

3.3.1 Blind Deconvolution

Much research has been undertaken on the topic of blind deconvolution. Multi-channel techniques appear particularly interesting because theoretically perfect inverse-filtering can be achieved if the Acoustic Impulse Responses (AIR) could be obtained *a priori*, and they do not have any common-zeros in the z-plane [154]. To achieve dereverberation without *a priori* knowledge of the room acoustics, i.e., blind dereverberation, many traditional techniques assume that the target signal is i.i.d. However, the i.i.d. hypothesis does not hold for speech-like signals. When applying such traditional deconvolution techniques to speech, the speech generating process is somehow deconvolved and the target speech signal is excessively whitened.

Hopgood used a realistic source signal model and uses Baysian parameter estimation techniques to estimate the unknown parameters [148]. The speech signal is modelled using a Block Stationary AR process while the room acoustics are modelled using an all-pole model (see Section 2.5.4). Several examples of blind deconvolution of reasonable low-order channels are investigated, and the results are encouraging.

Another technique which explores the null space of the correlation matrix, calculated from the received signals ($M \geq 2$), was developed by Gürelli and Nikias [155]. It was shown that the null space of the correlation matrix of the received signals contains information on the transfer function relating the source and the microphones. This observation constitutes the basis of their EVAM algorithm. This technique, although originally aimed at solving communication problems, has also potential in the speech processing framework and was extended by Gannot and Moonen [156, 157]. Although these techniques are supported by theory they have several drawbacks in real-life scenarios. The Generalized Eigenvalue Decomposition which is used to construct the null space of the correlation matrix is not robust enough, and quite sensitive to small estimation errors in the correlation matrix. Furthermore, the matrices involved become extremely large causing severe memory and computational requirements. Another problem arises from the wide dynamic range of the speech signal. This phenomenon may result in an erroneous estimate of the frequency response of the AIRs in the low energy bands of the input signal. Although some results are very encouraging, these drawbacks need to be investigated further.

It is well-known that a SIMO system can be equalized blindly by applying multi-channel LP to its output when the input is white. When the input is coloured, multi-channel LP will both equalize the acoustic channel and whiten the source. Triki and Slock [158, 159, 160] exploit the spatial diversity of the channel to estimate the source correlation structure, which can hence be used to determine a source whitening filter. The general structure of the proposed solution is depicted in Fig. 3.9. Multi-channel LP was then applied to the sensor signals filtered by the source whitening filter, to obtain the dereverberation filters. The estimated dereverberation filters were then applied to the received signals to obtain the dereverberated speech signal. They increase the derevreberation accuracy by exploit the time diversity of the source signal, i.e., they average dereverberation filters that are computed at different time frames. It should be noted that the order of the whitening filter is of great importance, and affects the

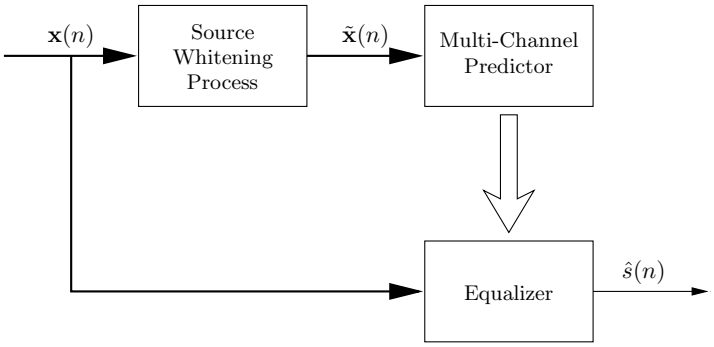


Figure 3.9 *Equalization with pre-whitening.*

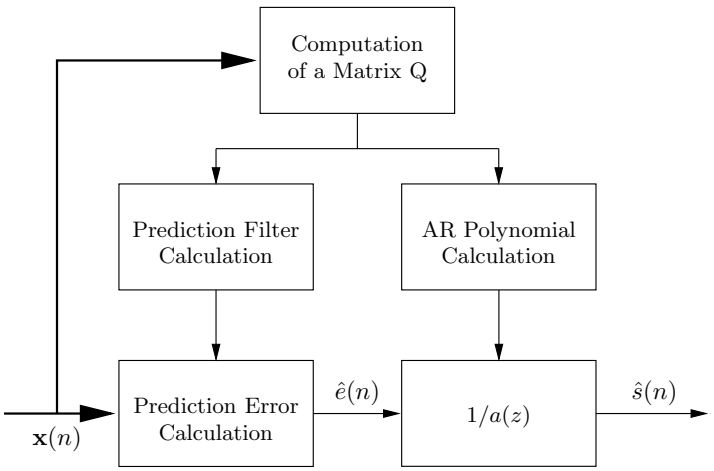


Figure 3.10 *Linear-predictive multi-input equalization.*

dereverberation performance of the complete system. Simulation results showed that the proposed solution is capable of reducing reverberation in a noise-free environment.

In [161, 162, 163] Delcroix et al. proposed a two-channel dereverberation algorithm called LInear-predictive Multi-input Equalization (LIME) with a view to solving the whitening problem of traditional blind deconvolution techniques. The structure of the proposed technique is depicted in Fig. 3.10. Delcroix et al. also used multi-channel LP, however, they allowed the source to be whitened and restored the colouration in a final stage. The prediction filter and AR polynomial $a(\mathrm{z})$ are calculated from a matrix Q which in practice can be calculated using

$$\mathbf{Q} = \left(\mathcal{E}\{\mathbf{x}(n-1)\mathbf{x}^T(n-1)\}\right)^+ \mathcal{E}\{\mathbf{x}(n-1)\mathbf{x}^T(n)\}, \tag{3.6}$$

where $\mathbf{A}^+$ is the Moore-Penrose generalized inverse of matrix $\mathbf{A}$. Simulation results showed that LIME could achieve almost perfect dereverberation for short duration impulse responses [162]. However, the presence of numerically overlapping zeros among

the channels prevent the possibility to deal with longer room impulse responses [164]. It is known [165] that the Acoustic Transfer Function (ATF) may have a large number of zeros close to the unit circle in the z-plane. Consequently, for a small number of microphones, the channels would contain numerically overlapping zeros and the dereverberation algorithm would perform poorly. In [164], Delcroix et al. showed how the use of spatial information, obtained by increasing the number of microphones, enables one to deal with long duration impulse responses. They also proposed to use cepstral mean normalization [166] to reduce the effect of the remaining distortions caused by the overlapping zeros. Experiments showed that, for a reverberation time of 0.2 seconds, dereverberation is possible in the presence of a coloured noise source of SNR of 5 dB. In [167] Delcroix et al. address the speech dereverberation problem in the presence of a coherent noise source. They showed that the LIME algorithm can achieve both dereverberation and noise reduction.

3.3.2 Homomorphic Deconvolution

A well-known homomorphic deconvolution technique is the cepstum-based technique [27, 168, 169, 170]. The underlying motivation is the fact that deconvolution in the time-domain corresponds to division in the frequency-domain and subtraction in the cepstrum-domain. These techniques generally perform poorly for reverberant speech for several reasons: framing effects, cepstral overlap of speech and late reverberation.

3.3.3 HERB

Nakatani et al. proposed a single-microphone speech dereverberation technique called Harmonicity based dEReverBeration (HERB) [171, 172]. HERB explicitly uses the fact that the source signal has a harmonic structure, in the design of a dereverberation filter. HERB estimates the dereverberation filter as an average of filters that transform observed reverberant signals into the output of an adaptive harmonic filter. The output of an adaptive harmonic filter corresponds to a rough estimation of the harmonic components of direct signals in the observed signals. The dereverberation filter, $W(k)$, is calculated as follows:

$$W(k) = A\left(\frac{\hat{X}(l,k)}{X(l,k)}\right), \tag{3.7}$$

where $X(l,k)$ and $\hat{X}(l,k)$ are discrete STFTs of an observed reverberant signal and the output of an adaptive harmonic filter at time frame l and frequency bin k, respectively. Here $A(\cdot)$ is a function that calculates the weighted average of $\hat{X}/X$ for each k over different time frames. This filter has been proven to approximate the inverse filter of the acoustic transfer function between a speaker and a microphone. In a recent paper [173] Kinoshita et al. evaluated the effect on speech intelligibility, and the potential to use HERB as a preprocessing algorithm for ASR. In both cases HERB seems to be able to decrease the Word Error Rate (WER) of the ASR system. The main disadvantage is that they required more than 5000 reverberant words, i.e., more than 60 minutes

of speech data, to acquire the dereverberation filter under the assumption that the system is time-invariant.

The conventional formulation of HERB does not provide an analytical framework within which the dereverberation performance can be optimized. In [174], Nakatani et al. reformulated HERB as a maximum a posteriori (MAP) estimation problem, in which the dereverberation filter was determined as one that maximizes the a posteriori probability given the observed signals. They derived a closed-form solution to this problem by assuming that the a posteriori probability is given by means of a Gaussian probability distribution function (PDF) under a harmonicity constraint. The proposed solution is shown to become identical to that of conventional HERB when a particular type of PDF is adopted, and it reveals the physical meaning of the weight with which conventional HERB calculated the average transfer function.

3.3.4 Inversion of mixed-phase impulse responses

In case the parameters of the system, or more specifically the filter coefficients of the acoustic impulse response have been identified, the question that remains is how to invert these long, mixed phase impulse responses, to produce inverse filters that are causal, stable and have finite length. The question was addressed in the last years by many authors, but the most efficient results are those of Mourjopoulos, and of Miyoshi and Kaneda. Mourjopoulos proposes two general techniques for inversion of finite length sequences, namely, homomorphic and least squares. Miyoshi and Kaneda proposed the well-known MINT principle which is closely related to the Bezout identity. In this section these techniques will be discussed.

Homomorphic techniques

The homomorphic techniques require decomposition of the AIR prior to inversion [175]. As shown in Section 2.5.2 the acoustical impulse response can be decomposed in two components: a minimum-phase one, containing all the zeros which fall inside the unit circle on the z-plane, and a maximum-phase component, containing all the poles which fall outside the unit circle (it is assumed that no pole falls exactly on the unit circle). The decomposition of a mixed-phase impulse response in the minimum and maximum phase components is not easy. It was approached both by homomorphic decomposition and by complex cepstral separation, but in general the results are poor. Once the components are separated, the minimum phase component can be inverted directly, because taking the Inverse Fast Fourier Transform (IFFT) of the reciprocal of its Fast Fourier Transform (FFT) yields a finite, stable and causal inverse impulse response. The same approach is unsuccessful for the maximum-phase component, as its inverse is causal and unstable or acausal and stable. However, the maximum-phase component needs to be time-reversed before getting inverted, and then time-reversed again. The acausality introduced by this process has to be eliminated by means of a time delay, causing no practical problem to not-real-time processing. Finally, the

inverse of the minimum and maximum phase components are convolved, producing the final approximate inverse filter.

Least squares techniques

The inverse of a non-minimum phase AIR is two-sided (acausal) and in general infinite in length [175]. In case a causal inverse filter is desired the inverse filter can only be approximated when the inverse filter is of infinite length. Treitel and Robinson [176] have shown that the inversion can be improved by incorporating a delay between the input and desired output sequence.

The inverse filter of length L, denoted by $\mathbf{w}_\tau = [w_\tau(0), \ldots, w_\tau(L-1)]^T$, can be found by minimizing the following cost function:

$$J(\tau) = \sum_{j=0}^{L-1} \left(\delta(j-\tau) - f_\tau(j)\right)^2, \tag{3.8}$$

where $f_\tau(j) = \sum_{i=0}^{L-1} w_\tau(i)h(j-i)$, h denotes the time-invariant acoustic impulse response, and τ denotes the additional delay.

MINT

The drawback in the conventional inverse filtering techniques seems to result from the use of only one channel. However, many systems in room acoustics can be modified to multiple output systems by adding extra microphones. In case the acoustical impulse responses do not have any common zeros an exact inverse of the system can be constructed using the MINT principle as proposed by Miyoshi and Kaneda [154]. MINT makes use of some fundamental results of multi-variable system theory. The advantage of this technique is that the inverse system consists of finite and causal filters.

The dereverberation problem can be generalized for an arbitrary M-input channel system, leading to the following set of relations

$$x_m(n) = h_m(n) * s(n) \quad \text{for } 0 \leq m \leq M, \tag{3.9}$$

and

$$\hat{s}(n) = \sum_{m=0}^{M-1} g_m(n) * x_m(n), \tag{3.10}$$

where $x_m(n)$, $h_m(n)$, $g_m(n)$ are respectively the m^{th} observation, transfer function and equalizer of the corresponding acoustic channel. For a multi-channel structure, equalization is achieved by finding a set of filters with impulse response $g_m(n)$ so that

$$\delta(n-\tau) = \sum_{m=0}^{M-1} g_m(n) * h_m(n). \tag{3.11}$$

This expression it is closely related to the Bezout identity, which states that there exists polynomials $G_m(\mathrm{z})$ such that

$$\sum_{m=0}^{M-1} G_m(\mathrm{z}) H_m(\mathrm{z}) = 1 \tag{3.12}$$

holds if the polynomials $H_m(\mathrm{z})\ \forall\ 0 \leq m \leq M-1$ have no common zeros. The algebraic decomposition that satisfies the Bezout identity is in general not unique and the algorithm reported in [154] calculates one of the possible solutions for the equalizers $g_m(n)\ \forall\ 0 \leq m \leq M-1$.

Putnam investigate the numerical precision of multiple input inverse filtering techniques [177]. He showed that in practice, the finite precision of measured impulse responses along with the inversion of poorly conditioned matrices, may pose numerical limitations. However, from his results we may conclude that the condition number decreases, and hence the numerical performance is enhanced as the number of microphones is increased.

3.3.5 Equalization Robustness

The AIR strongly depends on the source and microphone positions. Mourjopoulos [44] experimentally demonstrated that the AIR can vary drastically with changes in the source and microphone positions and orientations. Radlovic et al. [1, 2] demonstrated that even small variations, of the order of a tenth of the acoustic wavelength, can cause large degradations in the equalized acoustic impulse response. In this connection, it was found that as long as diffuse-field conditions are met, the room size, geometry and reverberation time have no significant effect on the spatial extent of the zone of equalization. Outside of such a region, equalization is ineffective and may actually yield performance worse than having no equalizer at all. This implies that sound equalization in practical environments may be an ill-posed problem [2].

Equalization using a multi-channel system has been recently investigated by Talantzis [3]. Multi-channel equalization shows an increase in equalization robustness compared to single-channel equalization. However, the equalization quality still remains highly restricted to a fraction of the wavelength. It was also found that the distance ratio, e.g. the distance between the source and microphones and the inner distance between the microphones, and the value of the wavelength are the parameters that mostly affect the equalization process whereas the room size and reverberation characteristics are of small significance.

Note that in all of the above cases exact equalization is assumed. The equalization zone may be extended if an inverse filter is used which is designed to approximate equalization for different source-microphone positions.

3.4 Summary

In theory, blind deconvolution of the acoustic channel allows perfect dereverberation. However, at this point in time there are no techniques available to blindly estimate AIRs in a realistic environment. Another problem is caused by the fact that small estimation errors can result in a highly distorted output signal.

Currently, most practical and robust solutions can be found in the techniques involving spatial processing. Fixed beamformers are capable of suppressing specular reflections. However, they only achieve a limited amount of reverberation suppression. Adaptive beamformers generally achieve superior noise suppression performance in noise-dominated environments compared to fixed beamformers, but their dereverberation capability is similar to that of the fixed beamformers. Adaptive beamfomers that are based on minimizing the output energy, are fundamentally unable to achieve good performance in reverberant environments simply because the reflections of the desired signal violate the assumption of independence between signal and noise.

The spectral enhancement techniques, and especially the single-microphone technique proposed by Lebart [24], are capable of suppressing late reverberation. The advantage of this technique is that it requires only a limited amount of *a priori* information, viz., the reverberation time of the room. Unfortunately some speech distortion is introduced by the spectral enhancement technique which decrease the quality of the dereverberated speech. In Chapter 6 it will be shown that the technique proposed in [24] can only be used when the source-microphone distance is larger than the critical distance.

Chapter 4

Dereverberation Quality Measures

4.1 Introduction

Many signal processing algorithms have been developed to enhance the quality of distorted speech. The speech quality can be quantified using subjective and objective measures. By comparing the speech quality before and after processing, one can investigate the speech quality improvement. In this chapter subjective and objective quality measures that can be used to determine the dereverberation quality will be briefly discussed.

In general, objective quality measures can be classified into intrusive (also known as end-to-end or reference) measures, and non-intrusive (also known as single-ended or no-reference) measures. The intrusive measures compare the distorted signal with the undistorted signal, which is usually called the reference signal. The non-intrusive measures do not require a reference signal, i.e., the speech quality is determined given only the distorted speech signal. In this chapter we will focus on intrusive measures, i.e., we assume that the undistorted signal is available.

Reliable quantitative measurement of the level of reverberation in a speech signal is particularly difficult and a unanimously accepted methodology has yet to emerge [93]. If an objective quality measure could be found that highly correlates with the results obtained from subjective test, its utility would be undeniable. Existing objective measures have been adopted to determine the dereverberation quality, e.g., Segmental Signal to Noise Ratio and the Bark Spectral Distortion [58] measures. Independent research has shown that the reverberation time and the spectral deviation are important perceptual factors that determine the quality of reverberant speech (see Section 1.3). Therefore, the relation of different objective measures with respect to these important factors is investigated.

The spectrogram and waveform are often used to represent a reverberant signal. How-

ever, from these representations, is not always clear how severely the signal is degraded by reverberation. In this chapter we describe a novel time-frequency representation of the reverberant signal. In this representation the spectrogram and instantaneous Direct to Reverberation Ratio are combined. This representation reveals which time-frequency components are affected most by reverberation and provides more insight than the spectrogram and the waveform do.

The structure of this chapter is as follows. In Section 4.2 a novel time-frequency representation of the reverberant signal is developed. In Section 4.3 a frequently used subjective measure is briefly described. Objective measures are then discussed in Section 4.4. The objective measures are analysed in Section 4.5. Finally, conclusions are given in Section 4.6.

4.2 Visual Representation

The waveform and the spectrogram are often used to visualize the time and time-frequency content of the signal, respectively. The spectrogram is the most frequently used time-frequency representation, and is determined from the short-time Fourier transform (STFT) of the signal. The STFT of the signal $z(n)$ is denoted by $Z(l,k)$, where l is the time frame and k the frequency bin (more details can be found in Chapter 5). Smearing in time can be observed in both the time and the time-frequency representation (see for example Figs. 1.5(a) and 1.5(b)). Spectral deviations are often clearly visible in the time-frequency representation. However, in both representations it is not possible to see how large the amount of reverberation is with respect to the direct signal, i.e., how severe the disturbance is. Therefore a novel time-frequency representation is proposed.

As a starting point the standard spectrogram is used, where the log amplitude values of $Z(l,k)$, for all l and k, are mapped to a colour value. In the proposed representation the colour is determined not only for the log amplitude value of $Z(l,k)$ but also from the instantaneous Direct to Reverberation Ratio (DRR). The instantaneous DRR is determined from two signals, i.e., the STFT of the direct signal $z_{\mathrm{d}}(n)$ and the reverberant signal $z_{\mathrm{r}}(n)$, which are denoted by $Z_{\mathrm{d}}(l,k)$ and $Z_{\mathrm{r}}(l,k)$, respectively. Note that $Z(l,k) = Z_{\mathrm{d}}(l,k) + Z_{\mathrm{r}}(l,k)$.

In order to get a proper colour representation a specific subset of colours is chosen from the hue-saturation-value colour-map. The hue can be used to select a desired colour (in this case red). The colour value (also known as the brightness) ranges from 0 to 1, and is calculated using the received signal $Z(l,k)$:

$$\mathrm{Value}(l,k) = \frac{20\log_{10}(|Z(l,k)|) - Z_{\min}}{Z_{\max} - Z_{\min}} \tag{4.1}$$

where $Z_{\min} = \min_{l,k}\{20\log_{10}(|Z(l,k)|)\}$, and $Z_{\max} = \max_{l,k}\{20\log_{10}(|Z(l,k)|)\}$. The saturation of the colour also ranges from 0 (shades of gray) to 1 (no white component),

and is calculated according to the instantaneous DRR:

$$\text{Saturation}(l,k) = \frac{\text{DRR}(l,k) - \text{DRR}_{\min}}{\text{DRR}_{\max} - \text{DRR}_{\min}} \tag{4.2}$$

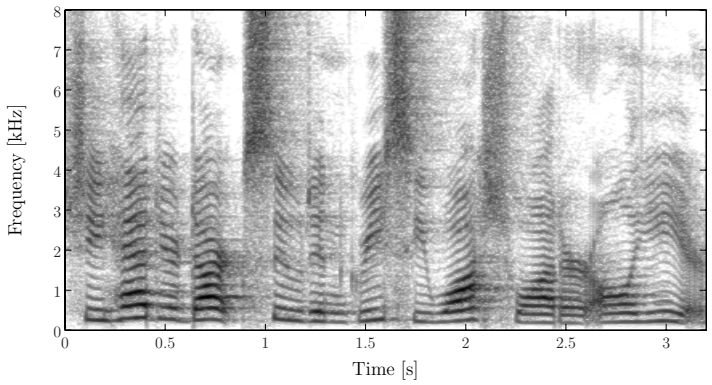
where $\text{DRR}(l,k) = 20\log_{10}(|Z_{\text{d}}(l,k)|/|Z_{\text{r}}(l,k))$, $\text{DRR}_{\min} = \min_{l,k}\{\text{DRR}(l,k)\}$, and $\text{DRR}_{\max} = \max_{l,k}\{\text{DRR}(l,k)\}$.

Two examples of a spectrogram and the proposed time-frequency representation of a reverberant signal measured at a distance of 0.25 m and 2 m, are depicted in Figs. 4.1 and 4.2, respectively. In case the signal value and the DRR value are high, the colour brightness and saturation will be high, and indicates no distortion. For decreasing DRR values the colour gradually becomes grayer. In case the signal value decreases the brightness of the colour drops, i.e., for low signal values the colour is black. In case the signal value is high and the DRR value is low the colour will be white, i.e., the direct signal is masked by reverberation. It can clearly been seen that the signal depicted in Fig. 4.1 is less affected by reverberation than the signal depicted in Fig. 4.2.

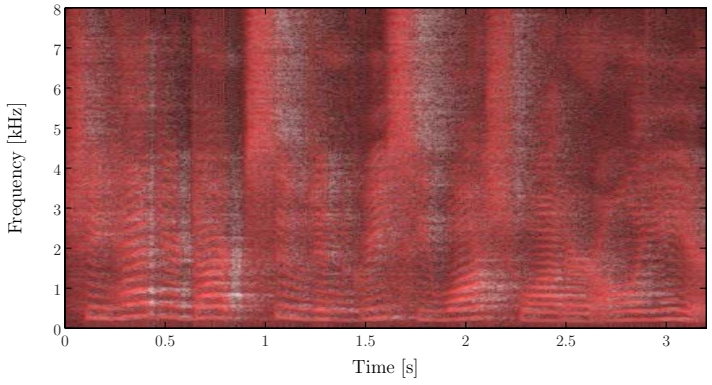
4.3 Subjective Measures

Subjective speech quality measures can be obtained using subjective listening tests in which human participants rate the performance of a system or quality of a signal in accordance with an opinion scale [178]. The International Telecommunications Union (ITU-T) has standardized the most commonly used methods for measuring the subjective quality of speech transmission over voice communication systems. For both listening-only and conversational tests the ITU-T recommends the use of a speech quality rating on a 5-point category scale, which is commonly known as the listening-quality scale [178]. An alternative speech quality scale that is used in listening-only tests is the listing-effort scale. In conversational tests a binary conversation difficulty scale is usually employed. These scales are listed in Table 4.1.

A listening test is performed by a number of subjects that listen to recordings that are degraded by an acoustic channel, and enhanced by the algorithm under test. The subjects provide their opinion on the quality of each signal, or the effort required to understand it, using the listening-quality scale or listening-effort scale, respectively. In conversational tests, subjects use a voice communication system before providing their opinion on its quality. Mean Opinion Score (MOS) is the averaged opinion score cross subjects and indicates the subjective quality of the system or algorithm under test. To obtain a realistic variability in the opinion scores, a large numbers of subjects is required. Therefore, the main drawback of subjective testing is cost [179]. Even with a large amount of subjects, the variance of MOS can still be high. Furthermore, the quality that is expected by a customer will be different depending on whether the device is an expensive conference system or a cheap mobile telephone. The constraints imposed by the need to limit the cost and the amount of subjects also limit the ability to test the system or algorithm under different environmental conditions. Hence,



(a) Spectrogram.



(b) Proposed representation.

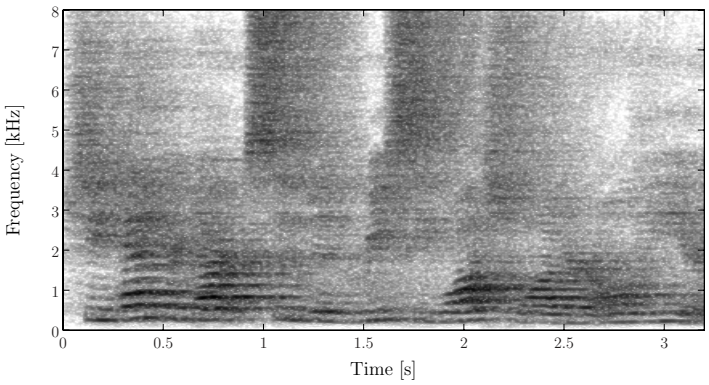
Figure 4.1 *Time-frequency representation of a reverberant signal* (RT_{60} = 500 *ms*, D = .25 *m*) *using (a) the spectrogram, and (b) the proposed method.*

it would be more practical if an automatic assessment system would exist whereby quality measures could be obtained.

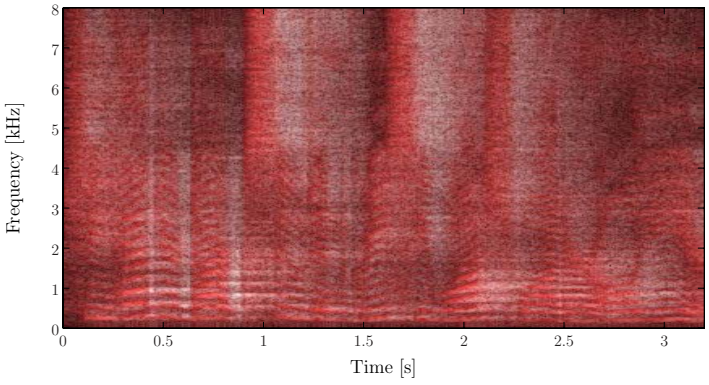
4.4 Objective Measures

With the rapidly evolving speech enhancement systems and voice communication systems, there is an increasing need for robust objective speech quality measures that correlate well with subjective speech quality. During the design and validation stages of algorithms, codecs, and communication systems, objective quality measures are valuable assessment tools. Over the last two decades, researchers have developed different measures based on various speech analysis models [180, 179].

In general, objective speech quality measures can be categorized in three domains:



(a) Spectrogram.



(b) Proposed representation.

Figure 4.2 *Time-frequency representation of a reverberant signal* ($RT_{60} = 1$ *s*, $D = 2$ *m) using (a) the spectrogram, and (b) the proposed method.*

time domain, spectral domain or perceptual domain. The time domain measures are generally applicable to analogue or waveform coding systems in which the receiver reproduces the waveform. However, they can also be used to determine the speech quality improvement. Signal to Noise Ratio (SNR) and segmental SNR are typical time domain measures [180]. Spectral domain measures are usually preferred above time-domain measures and are less influenced by possible time misalignments between the original and the received or processed signals. Most spectral domain measures are related to speech codec design. Perceptual domain measures are based on models of the human auditory system, compared to time and spectral domain measures they have and have a higher change of predicting the subjective quality of speech. Theoretically, perceptually relevant information is both sufficient and necessary for a precise assessment of perceived speech quality [179].

Most objective measures are intrusive perceptually-based measures. They are based on psychoacoustics considerations and are trained on subjective databases to represent

Listening-Quality Scale:

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

Listening-Effort Scale:

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

Conversation Difficulty Scale:

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

Table 4.1 *ITU-T recommended speech quality measurement scales [178].*

human perception. Among these perceptual models, the ITU-T has standardized the Perceptual Evaluation of Speech Quality (PESQ) in 2001 as ITU-T Recommendation P.862 [181]. PESQ predicts the listening quality of a speech signal which is degraded by codecs, background noise and packet loss.

4.4.1 Intrusive Measures

In this section various intrusive time and frequency domain measures are summarized.

Segmental Signal to Reverberation Ratio

The instantaneous segmental Signal to Reverberation Ratio (SRR) [93] of the l^{th} frame is defined similar to the segmental SNR [180], i.e.,

$$\text{SRR}_{\text{seg}}(l) = 10\log_{10}\left(\frac{\sum_{n=lR}^{lR+N-1} z_{\text{d}}^2(n)}{\sum_{n=lR}^{lR+N-1}(z_{\text{d}}(n)-\hat{z}_{\text{d}}(n))^2}\right) \quad [\text{dB}], \tag{4.3}$$

where N is the frame length in samples ($f_{\text{s}}N$ is usually equal to 32 ms), R is the frame rate in samples, $z_{\text{d}}(n)$ is the direct signal, which is a delayed version of the anechoic signal, and $\hat{z}_{\text{d}}(n)$ is the enhanced signal. The frame rate R is usually chosen such that the frames overlap $50\% - 75\%$. The mean segmental SRR is then obtained by averaging Eq. 4.3 over all frames.

Log Spectral Distortion

One of the oldest distortion measures proposed for speech and the one which most closely resembles traditional difference distortion measures is formed by the L_p norm of the difference of the log spectra of the desired signal $z_{\mathrm{d}}(n)$ and the enhanced signal $\hat{z}_{\mathrm{d}}(n)$. In most cases short-time spectra are used, which are obtained using the STFT of the signals. The STFT of the signal $z(n)$ is denoted by $Z(l,k)$, where l denotes the time frame and k the frequency bin. The frame length is usually between 32 and 64 ms long, and the overlap is $50\% - 75\%$. The L_p norm of the difference between $Z_{\mathrm{d}}(l,k)$ and $\hat{Z}_{\mathrm{d}}(l,k)$, in the l^{th} frame, is defined as

$$\mathrm{LSD}(l) = \left(\frac{2}{K} \sum_{k=0}^{\frac{K}{2}-1} \left| \mathcal{L}\{\hat{Z}_{\mathrm{d}}(l,k)\} - \mathcal{L}\{Z_{\mathrm{d}}(l,k)\} \right|^p \right)^{\frac{1}{p}} \quad [\mathrm{dB}], \tag{4.4}$$

where $\mathcal{L}\{X(l,k)\} \triangleq \max\{20\log_{10}(|X(l,k)|), \delta\}$ is the log spectrum confined to about 50 dB dynamic range ($\delta = \max_{l,k}\{20\log_{10}(|X(l,k)|)\} - 50$). The mean Log Spectral Distortion (LSD) is obtained by averaging Eq. 4.4 over all frames containing speech. The most common choices for p are 1, 2, and ∞, yielding mean absolute, root mean square, and maximum deviation, respectively.

Modulation Spectrum

Speech can be considered as a sequence of sounds with a continuously varying spectrum. These spectral variations lead to fluctuations of the envelope of the signal within individual frequency bands. Based on this concept, Steeneken and Houtgast [182] derived the Speech Transmission Index (STI), which is a powerful and widely accepted measure to predict the effect of room acoustics on speech intelligibility. This measure quantifies the speech intelligibility in terms the spectral content of the signal envelope. Plomp [183] derived a so-called modulation spectrum of speech by determining the spectrum of the signal envelope within each frequency bands. The modulation spectra show that the relevant modulation frequencies of speech are roughly in a range between 0.1 to 40 Hz and that the strongest fluctuations are between 3 to 5 Hz. The shape of the modulation spectrum is about the same for all octave bands, but the amount of modulation, denoted by the Modulation Index (MI), differs between the bands. The strongest modulations are found within the frequency band that is centred around 500 Hz, while the low-frequency bands contain less modulation.

The temporal envelope filtering techniques (as described in Section 3.2.3) are motivated by studies of the effect of reverberation on the MI of the speech signal. Tails produced by past acoustic events fill in low energy regions between consecutive sounds, and reduce the modulation depth of the original envelope and thus modifying its MI. In case the degraded signal is successfully dereverberated the modulation spectrum should be restored.

The modulation index as function of modulation frequency can be calculated using the following steps:

i) The speech signal is first analysed using an octave filter bank. The filter bank can also be bypassed resulting in a broad-band analysis.

ii) For each octave band the envelope is estimated by taking the magnitude of a standard Hilbert-transform, which results in the Hilbert envelope. The Hilbert envelope is low-pass filtered with a 50-Hz low-pass filter and then downsampled to a frequency of 200 Hz. The low-pass filtering removes any fine-structure components of the speech signal, such as the fundamental frequency of the speaker. The resulting signal will be referred to as *the envelope signal*.

iii) For each octave band the Power Spectral Density (PSD) of the envelope signal is estimated using a standard Welch procedure [184]. The parameters used for the Welch procedure are a window length of 8 seconds, a Hanning window with 40% overlap between successive windows.

iv) The intensity values of the PSD are summed over modulation frequencies for each octave band and are normalized using the DC-component of the PSD. The normalization is set to reach a value of 0 dB (m=1) for an amplitude modulated sine-wave. This calculation results in a modulation spectrum defined in the intensity domain.

4.4.2 Intrusive Perceptually-Based Measures

In this section various intrusive measures are summarized that are perceptually motivated.

Bark Spectral Distortion

The Bark Spectral Distortion (BSD) can be classified as a perceptual domain measure that transforms the speech signal into a perceptually relevant domain which incorporates human auditory models [58]. Studies have shown that the correlation coefficient of this measure with MOS scores is above 0.9 [58, 99]. The BSD metric makes use of the Bark spectra L_{z_d} and $L_{\hat{z}_\mathrm{d}}$, of the direct signal z_d and the enhanced signal $\hat{z}_\mathrm{d}(n)$, respectively. The Bark spectrum is calculated by going through three steps, i.e., critical band filtering, equal loudness pre-emphasis and phon to sone conversion. The input to this process is the magnitude squared spectrum for the current analysis frame with index l. The output is denoted by $L_x(l, k_s)$, where the index k_s denotes Bark frequency bin.

When the Bark spectra are calculated the BSD score can be obtained by

$$\mathrm{BSD} = \frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_{k_s=1}^{K_s}\left(L_{z_\mathrm{d}}(l,k_s) - L_{\hat{z}_\mathrm{d}}(l,k_s)\right)^2}{\sum_{k_s=1}^{K_s}\left(L_{z_\mathrm{d}}(l,k_s)\right)^2}, \tag{4.5}$$

where L denotes the number of analysis frames. The resulting BSD score for a speech signal is the weighted average of the BSD scores for all of the analysis frames.

The Modified Bark Spectral Distortion (MBSD) is a modification of the BSD in which the concept of a noise-masking threshold is incorporated [185]. Due to this threshold the MBSD measure differentiates between audible and inaudible distortions. The MBSD measure assumes that loudness differences below the noise masking threshold are not audible and are therefore excluded from the calculation of the perceptual distortion. The MBSD uses a simple cognition model to calculate the distortion value.

Reverberation Decay Tail

Recently, Wen and Naylor proposed a novel objective measure called the Reverberation Decay Tail (R_{DT}) [186]. The R_{DT} jointly characterizes the relative energy in the tail of the AIR and the rate of decay.

As a reference for this speech quality measure the direct speech signal $z_{\mathrm{d}}(n)$ is used. The analysis of the test signal and the reference signal are performed in the Bark spectral domain. First so-called end-points are detected in each Bark spectral bin. The end-points are defined as time instants at which the speech energy abruptly falls. Secondly, the decay and corresponding absolute decay tail energy and direct path energy are calculated for all end-points and Bark spectral bins. Note that the direct path energy is calculated using the Bark spectrum of the reference signal. Secondly, the decay, absolute decay tail energy, and direct path energy are averaged over all detected end-points, and subsequently over all Bark spectral bins, and result in λ_{avg}, D_{avg}, and A_{avg}, respectively. Finally the R_{DT} can be calculated by:

$$R_{\mathrm{DT}} = \frac{A_{\mathrm{avg}}}{\lambda_{\mathrm{avg}} D_{\mathrm{avg}}}. \tag{4.6}$$

Note that higher R_{DT} values correspond to either a higher amount of relative energy in the tail or a slower decay rate.

In [30] the R_{DT} measure was tested using three dereverberation methods. The results were compared to the subjective amount of reverberation indicated by 26 normal hearing subjects. The results showed a high correlation between the R_{DT} values and the amount of reverberation perceived by the subjects.

Perceptual Evaluation of Speech Quality

The objective measures described in ITU-T Recommendation P.862 (February 2001) is known as PESQ [181]. It is the result of several years of development and is applicable not only to speech codecs but also to intrusive measurements. Real systems may include filtering and variable delay, as well as distortions due to channel errors and low bit-rate codecs. The PSQM measure as described in ITU-T Recommendation

P.861 (February 1998), was only recommended for use in assessing speech codecs, and was not able to take proper account of filtering, variable delay, and short localized distortions. PESQ addresses these effects with transfer function equalization, time alignment, and a new algorithm for averaging distortions over time. The validation of PESQ included a number of experiments that specifically tested its performance across combinations of factors such as filtering, variable delay, coding distortions and channel errors. It is recommended that PESQ be used for speech quality assessment of 3.1 kHz (narrow-band) handset telephony and narrow-band speech codecs [181].

PESQ compares an original signal $s(t)$ with a degraded signal $z(t)$ that is the result of passing $s(t)$ through a communications system, or with the enhanced signal $\hat{s}(t)$ calculated by the enhancement system. The output of PESQ is a prediction of the perceived quality that would be given to $z(t)$, or $\hat{s}(t)$, by subjects in a subjective listening test. In the first step of PESQ a series of delays between original input and test signal are computed, one for each time interval for which the delay is significantly different from the previous time interval. For each of these intervals a corresponding start and stop point is calculated. The alignment algorithm is based on the principle of comparing the confidence of having two delays in a certain time interval with the confidence of having a single delay for that interval. The algorithm can handle delay changes both during silences and during active speech parts. Based on the set of delays that are found PESQ compares the original signal with the aligned test signal of the device under test using a perceptual model. The key to this process is transformation of both the original and test signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). This is achieved in several stages: time alignment, level alignment to a calibrated listening level, time frequency mapping, frequency warping, and compressive loudness scaling. The internal representation is processed to take account of effects such as linear filtering and local gain variations that may have little perceptual significance if they are not too severe. This is achieved by limiting the amount of compensation and making the compensation lag behind the effect. Thus minor, steady state differences between the original and degraded speech are compensated. More severe effects, or rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance. The internal representation process allows a small number of quality indicators to be used to model all subjective effects. In PESQ, two error parameters are computed in the cognitive model; these are combined to give an objective listening quality score.

4.4.3 Intrusive Channel-Based Measures

There are a number of objective measures for indicating the quality and intelligibility of filtered speech for a given impulse response. In this section different objective measures are summarized that are derived from the Acoustic Impulse Response (AIR). Additional measures can be found in [41].

In case a reverberation cancellation algorithm is used (see Section 3.3) it is possible to calculate the total impulse response that described the system between the source to the output of the algorithm. Ideally this response should be equal to a (possibly scaled and delayed) Dirac pulse. The total impulse response can also be evaluated using one of the objective measures summarized in this section. The improvement of the objective measure can be determined subsequently by calculating the objective measure of a reference impulse response, e.g., the impulse response of the system between the source and the closest microphone.

Direct to Reverberation Ratio

The most straightforward objective measure is called the DRR or SRR and is defined as:

$$\begin{aligned} \mathrm{DRR} &= 10\log_{10}\left(\frac{E_\mathrm{d}}{E_\mathrm{r}}\right) \\ &= 10\log_{10}\left(\frac{\sum_{n=0}^{n_\mathrm{d}} h^2(n)}{\sum_{n=n_\mathrm{d}+1}^{\infty} h^2(n)}\right) \quad [\mathrm{dB}], \end{aligned} \tag{4.7}$$

where the direct sound arrives at n_d. Due to finite sampling of the AIR the arrival time n_d, which is measured with respect to the source signal, will usually not fall onto an exact sample moment. When synthetic AIRs are used the direct path can be computed separately (see Appendix A). However, when dealing with measured impulse responses the direct path component, and therefore the related energy, can not be determined precisely. Therefore, $n_\mathrm{d} f_\mathrm{s}$ is often taken 8-16 ms larger than the approximate arrival time of the direct sound.

It should be noted that the DRR depends on the distance between the source and the microphone and on the reverberation time of the room. We can express the DRR using Eq. 2.53 and Eq. 2.56 as:

$$\mathrm{DRR} = 10\log_{10}\left(\frac{QR}{16\pi D^2}\right), \tag{4.8}$$

where Q is the directivity factor, R is the room constant (given by Eq. 2.57), and D is the source-microphone distance. Note that the room constant is inversely proportional to the reverberation time.

Early to Total Sound Energy Ratio

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

energy ratio:

$$D = \frac{E_{\mathrm{e}}}{E_{\mathrm{t}}} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_{\mathrm{e}}} h^2(n)}{\sum_{n=0}^{\infty} h^2(n)}\right) \quad [\mathrm{dB}], \tag{4.9}$$

where $n_{\mathrm{e}}f_{\mathrm{s}}$ is usually set to 50 or 80 ms. The time (in milliseconds) is often used as a subscript, i.e., in case $n_{\mathrm{e}}f_{\mathrm{s}} = 50$ ms the early to total sound energy ratio is denoted by D_{50}.

Early to Late reverberation Ratio

Another objective criterion is called the Early to Late reverberation Ratio (ELR) or *Clarity Index* (originally Klarheitsmaß) and is defined as:

$$C = \frac{E_{\mathrm{e}}}{E_{\mathrm{l}}} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_{\mathrm{e}}} h^2(n)}{\sum_{n=n_{\mathrm{e}}+1}^{\infty} h^2(n)}\right) \quad [\mathrm{dB}], \tag{4.10}$$

where $n_{\mathrm{e}}f_{\mathrm{s}}$ is usually set to 50 or 80 ms. The time (in milliseconds) is often used as a subscript, i.e., in case $n_{\mathrm{e}}f_{\mathrm{s}} = 50$ ms the ELR is denoted by C_{50}.

4.5 Analysis

In this section frequently used intrusive quality measures are analysed. In Section 1.3 it was shown that the reverberation time and spectral deviation provide useful information related to the reverberant speech quality and intelligibility. Rather than studying the relation between the objective measures and subjective test results, we have studied the relation between the objective measure and some basic objective measures derived from the AIR, e.g., reverberation time and spectral deviation.

The speech fragment used in the experiments consists of male and female speech ($f_{\mathrm{s}} = 16$ kHz) taken from the TIMIT database [4], and is 40 seconds long. The reverberant speech fragments are obtained by convolving the anechoic speech signal with an AIR that was generated using the image method (see Appendix A).

4.5.1 Segmental Signal to Reverberation Ratio

In Figs. 4.3(a) and 4.3(b) the segmental SRR is shown for different DRR values, for a reverberation time of 300 and 600 ms, respectively. It can clearly be seen that there is an almost linear relation between the the segmental SRR and the DRR. The offset between the two values is caused by the frame-by-frame processing of the segmental

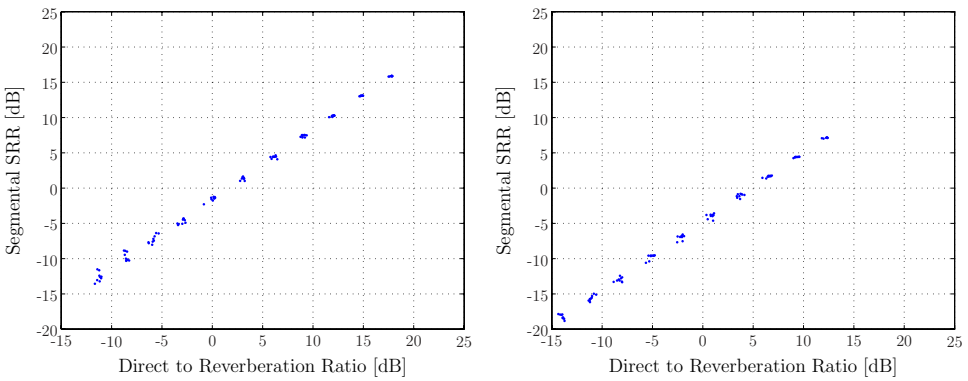


Figure 4.3 *Segmental SRR versus the DRR for (a)* $\mathrm{RT}_{60} = 300$ *ms, and (b)* $\mathrm{RT}_{60} = 600$ *ms.*

SRR, and depends on the signal and on the reverberation time. For a stationary signal the segmental SRR and DRR are equal. As with the segmental SNR and the global SNR it is expected that the correlation between the segmental SRR and the subjective speech quality is larger than the correlation between the DRR and the subjective speech quality. Note that most of the reverberation is masked when the instantaneous segmental SRR is large. In those frames where the instantaneous segmental SRR is low the reverberation will be clearly audible. Therefore, it is expected that the mean segmental SRR correlates much better with the subjective speech quality than the DRR.

In Fig. 4.4(a) the spectral deviation, which is defined as the standard deviation of the energy spectrum of the AIR [16], is shown for different DRR values (obtained using different reverberation times and source-microphone distances). The results are consistent with those obtained by Jetzt in [16], and the average maximum value of the spectral deviation is consistent with the theoretical maximum of 5.57 dB derived by Schroeder [74]. In Fig. 4.4(b) the spectral deviation is shown for different segmental SRR values. It can clearly be seen that the shape of the function is similar to the shape obtained using the DRR. However, due to the 'error' caused by the frame-by-frame processing of the segmental SRR there is a slight offset which mainly depends on the reverberation time.

In Fig. 4.5 the relation between the reverberation time and the segmental SRR is shown at for distance of 0.5 and 2 m. The segmental SRR is monotonically decreasing with the reverberation time and is almost completely independent of the source-microphone distance D.

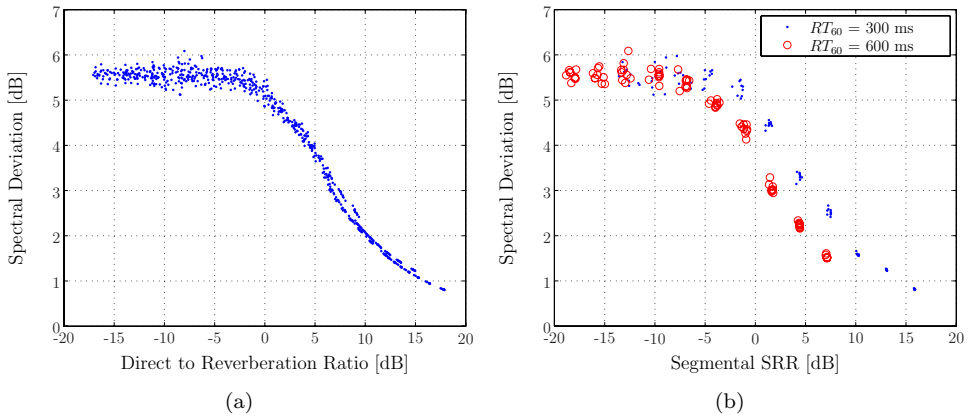


Figure 4.4 *DRR and Segmental SRR versus the spectral deviation.*

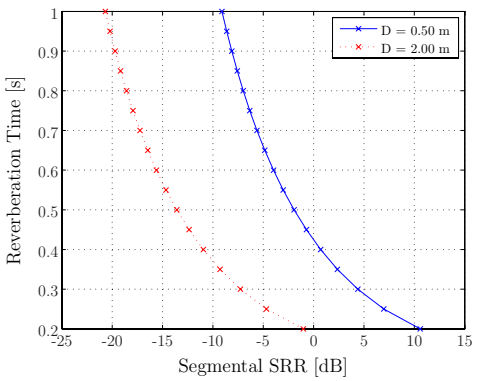


Figure 4.5 *Segmental SRR versus the reverberation time.*

4.5.2 Bark Spectral Distortion and Log Spectral Distance

The relation between the Bark Spectral Distortion and the spectral deviation, and the reverberation time, is shown in Figs. 4.6(a) and 4.6(b), respectively. For Fig. 4.6(a) the BSD values were calculated using different reverberation times and source-microphone distances. The results demonstrate that only very low BSD values correspond to a decrease in spectral deviation. Hence, a decrease in BSD does not necessarily mean that the spectral deviation is decreased. The relation between the BSD and the reverberation time depends on the distance D between the source and the microphone, as depicted in Fig. 4.6(b). For $D = 0.5$ m there is an almost linear relation between the two values. However, for $D = 2$ m the relation is non-linear. It should be noted that similar relations were found between the Log Spectral Distance (norm-1) and the spectral deviation and reverberation time.

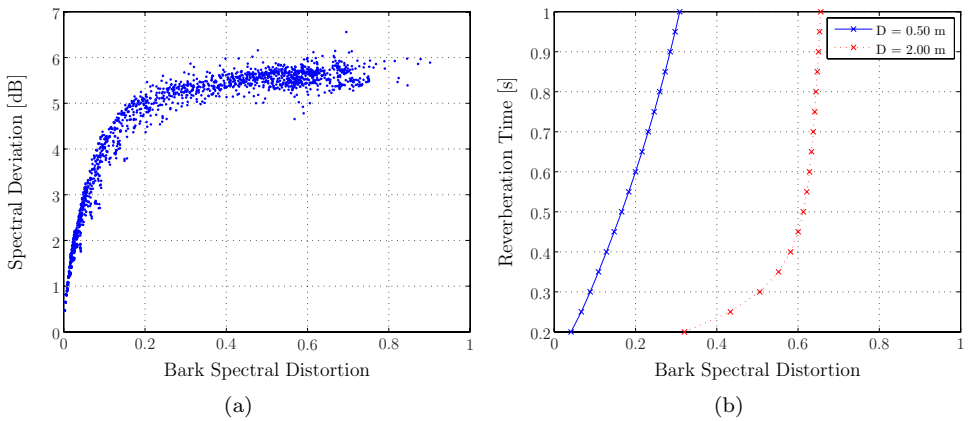


Figure 4.6 *Bark Spectral Distortion versus (a) the spectral deviation and (b) the reverberation time.*

4.5.3 Reverberation Decay Tail

The relation between the recently proposed Reverberation Decay Tail R_{DT} measure and the reverberation time, and the source-microphone distance, is shown in Figs. 4.7(a) and 4.7(b), respectively. The results shown in Fig. 4.7(a) clearly indicate an almost linear relation between the R_{DT} and the reverberation time. The R_{DT} measure is proportional to the average absolute decay tail energy A_{avg} in Eq. 4.6. Since the average absolute decay tail energy is proportional to D^2 (like the reverberation energy), the R_{DT} measure should be proportional to D^2. Fig. 4.7(b) shows that this relation is approximately true. The relation between the R_{DT} measure and the spectral deviation for $\mathrm{RT}_{60} = 300$ ms, and $\mathrm{RT}_{60} = 600$ ms is shown in Fig. 4.8. From these results it can be seen that the R_{DT} measure does depend on the amount of colouration, which is determined by the spectral deviation. It should be noted that the R_{DT} measure was developed to be independent of the coloration effect. However, Wen et al. considered a different kind of colouration in [186], in which colouration was introduced due to a strong early reflection, which causes a strong modulation in the power spectrum of the AIR.

4.5.4 PESQ

The relation between the PESQ score and the spectral deviation, and the reverberation time, is shown in Figs. 4.9(a) and 4.9(b), respectively. The PESQ score is inversely proportional to the reverberation time and increases for small source-microphone distances, i.e., for higher DRR. Due to the large variations in spectral deviation for a single score the PESQ score does not reveal much information with respect to the spectral deviation. Note that the PESQ score was not developed to determine the

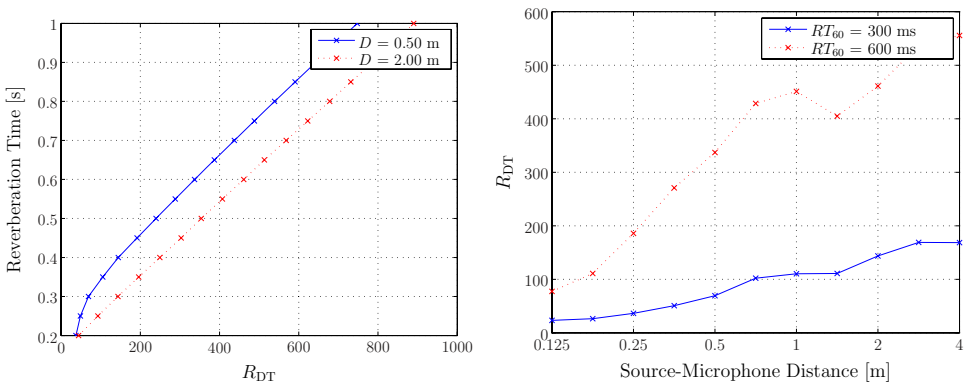


Figure 4.7 *Reverberation Decay Tail versus (a) the reverberation time and (b) the source-microphone distance.*

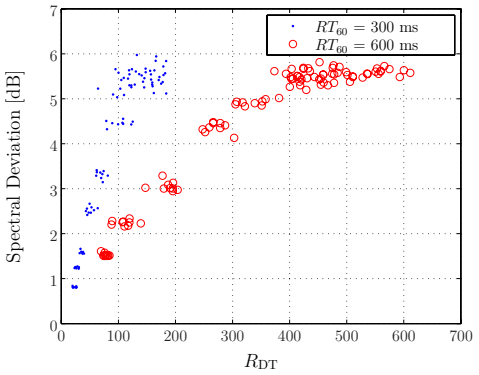
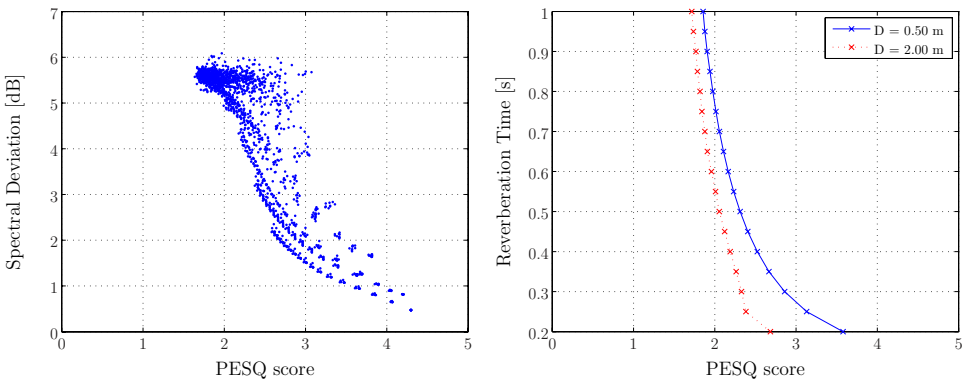


Figure 4.8 *Reverberation Decay Tail versus the spectral deviation.*

speech quality in a reverberant environment. Since it is sensitive to other distortions that are perceptually important we use it as an additional measure.

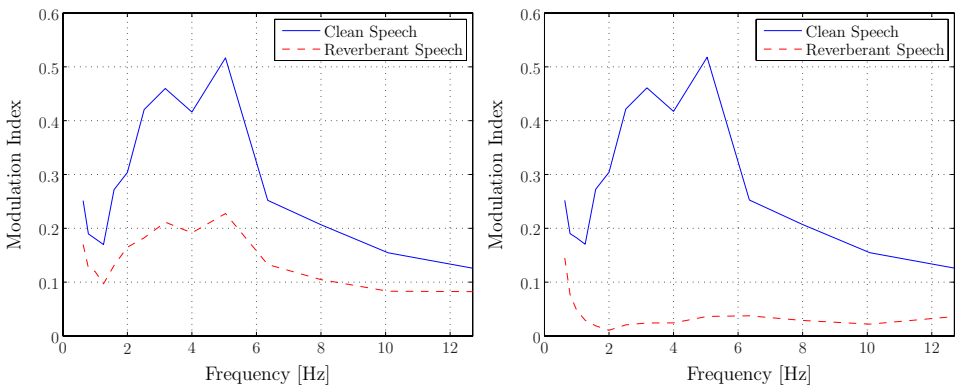
4.5.5 Modulation Spectrum

The full-band modulation spectrum of a 40 seconds anechoic and reverberant speech fragment are depicted in Figs. 4.10 and 4.11, respectively. As mentioned in Section 4.4.1 the modulation spectrum exhibits strong fluctuations, i.e., a large modulation index, between 3 and 5 Hz. This example demonstrated that, even for a reverberation time of 100 ms, the modulation index has decreased due to the reverberation. According to Steeneken and Houtgast [182] the decreased modulation index indicates a decreased speech intelligibility.



(a) (b)

Figure 4.9 *PESQ score versus (a) the spectral deviation and (b) the reverberation time.*



(a) $RT_{60} = 100$ ms, $D = 0.25$ m. (b) $RT_{60} = 300$ ms, $D = 1$ m

Figure 4.10 *Full-band modulation spectrum for the anechoic and the reverberant speech fragment.*

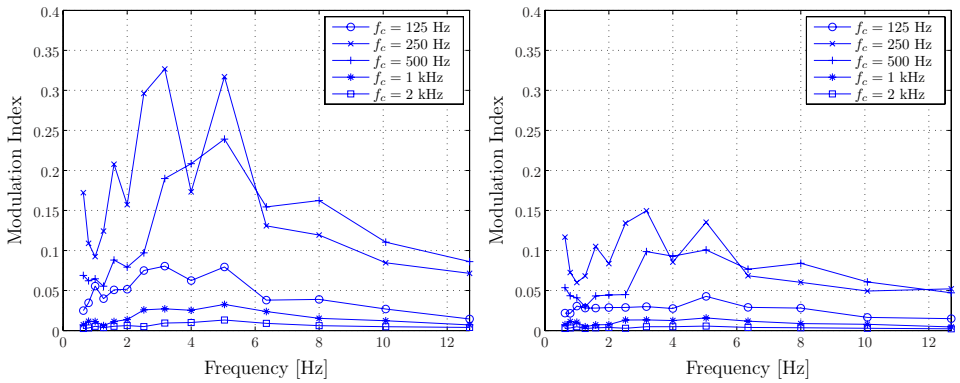


Figure 4.11 *Sub-band modulation spectrum for (a) the anechoic and (b) the reverberant* ($\mathrm{RT}_{60} = 100$ *ms,* $D = 0.25$ *m) speech fragment.*

4.5.6 Discussion

In case dereverberation is achieved by means of *reverberation cancellation*, as defined in Chapter 3, it is possible to calculate the total transfer function that describes the system between the source and processed signal. Hence, intrusive channel-based measures can be used to evaluate the dereverberation performance of the dereverberation technique. However, in case dereverberation is achieved by means of *reverberation suppression* the intrusive channel-based measures can not always be used. Many of the reverberation suppression techniques apply a time-variant and non-linear operation to the reverberant signal. This makes it impossible to calculate a single transfer function which describes the system between the source and processed signal. In the latter case other intrusive or non-intrusive objective measures can be used to evaluate the performance.

The segmental SRR is a very useful quantitative objective measure that exhibits an almost linear relation with the DRR which can be calculated directly from the AIR. It was shown that the segmental SRR depends on both the spectral deviation and reverberation time.

The BSD measure is less sensitive to changes in the reverberation time than the segmental SRR. Although the BSD measure does not reveal much information about the reverberation time and spectral deviation it can be used to determine the 'similarity' between the processed signal and the anechoic signal. The relation of the LSD (with norm-1) measure with the reverberation time and the spectral deviation is similar to that of the BSD measure. The advantage of the LSD measure is that it can be easily be evaluated for a single time frame and exhibits a lower computational complexity than the BSD measure.

The relation between the R_{DT} measure and the reverberation time is very strong.

The relation with the spectral deviation is similar to the segmental SRR. It should be noted that the R_{DT} measure gives only a global indication and does not reveal any information about the speech quality. Let us for example assume that the Reverberation Decay Tail measure indicates an improvement from 600 to 50, which would indicate a significant decrease in reverberant energy. However, this improvement does not provide any information about the overall sound quality of the signal under test, because even a completely silent signal would result in a R_{DT} value of zero. Therefore, it should always be used in combination with an other quality measure, e.g., the LSD, BSD, or PESQ measure.

The PESQ measure is not designed to designed to determine the quality of reverberant speech. It was shown that the PESQ score only decreases slightly for increasing reverberation time. However, the PESQ measure is sensitive to other distortions that are perceptually important. In this dissertation it is therefore used to determine the speech quality when additional interferences, such as noise and echo, are considered.

Finally the effect of reverberation on the modulation spectrum was studied. The fact that the modulation index decreases due to reverberation, or increases by the dereverberation technique, can also be seen from a simple waveform representation. Furthermore, the modulation index does not reveal detailed information concerning the quality of the reverberant or processed speech, and therefore is of less interest compared to other quality measures.

4.6 Conclusions

In this chapter some frequently used objective speech quality measures that are useful to determine the dereverberation quality were discussed and analysed. Additionally, a novel time-frequency representation was proposed to visualize a reverberant speech signal.

Research has shown that the reverberation time and the spectral deviation are important perceptual factors that determine the quality of reverberant speech (see Section 1.3). Therefore, it is important to understand the relation of different objective measures with respect to these perceptual factors. Furthermore, an objective measure is required which indicates the overall quality of the dereverberated signal.

In Table 4.2 we have summarized how the intrusive objective measures that were analysed and discussed in Section 4.5 are related to the colouration, reverberation time, and overall speech quality. The signs are used to indicate how strong the relation between the perceptual factors and the objective measures are. The results depicted in Table 4.2 indicate that there is no objective measure available which is mainly influenced by the colouration, i.e., spectral deviation, and thus remains a topic for further research.

Objective Measure	Colouration	Reverberation Time	Overall Speech Quality
segmental SRR	+	+	+
BSD	−	−	++
LSD	−	−	+
R_{DT}	+	++	−−
PESQ	−−	−	++

Table 4.2 *The relation between some important perceptual factors and the tested intrusive objective measures (++ = strong, . . ., −− = weak)*

In Chapter 5 and 7 the segmental SRR, LSD (norm-1), BSD and PESQ will mainly be used to evaluate the performance of the proposed dereverberation techniques. In case the suppression of more than one interference is evaluated, for example background noise and late reverberation, we prefer to use the term segmental Signal to Interference Ratio rather than the segmental SRR or segmental SNR.

Chapter 5

Single- and Multi-Microphone Dereverberation

5.1 Introduction

In speech communication systems, such as voice-controlled systems, hands-free mobile telephones, and hearing aids, the received microphone signals are degraded by room reverberation, background noise, and other interferences.

Reverberation is the process of multi-path propagation of an acoustic sound from its source to one or more microphones. The received signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*), and reflections that arrive after the early reflections (commonly called *late reverberation*). The manner in which the received signal is affected by reverberation is characterized by the Acoustic Transfer Function (ATF), which is defined as the frequency response of the system relating the sound source to the sound pressure at the receiver. Reverberant speech can be described as sounding distant with noticeable echo and colouration. The colouration can be characterized by the spectral deviation, which is defined as the standard deviation of the energy spectrum of the Acoustic Impulse Response (AIR) [16]. These detrimental perceptual effects generally increase with increasing distance between the source and receiver. For example, reverberation has a negligible effect in telephony systems with traditional handsets. However, in hands-free systems, reverberation affects the quality and intelligibility of speech and is a significant problem for both telecommunications and speech recognition applications.

Reverberation is highly dependent on the physical properties of the enclosed space as well as on the location of the source and the listener within the space [7]. It would be convenient to assume that reverberation solely reduces intelligibility [8], but this assumption is incorrect. In the development of techniques that enhance the quality and intelligibility of reverberant speech, it is important to understand which reverberation

effects are detrimental.

Allen [15] reported a formula to predict the quality of reverberant speech. The main result is given by the equation

$$P = P_{\max} - \sigma \mathrm{RT}_{60}, \tag{5.1}$$

where P is the subjective preference in some arbitrary units, $P_{\max}$ is the maximum possible preference, σ is the spectral deviation in decibels, and RT_{60} is the reverberation time in seconds. According to this formula, decreasing either the spectral deviation or the reverberation time results in an improvement in speech quality.

In continuous utterances, speech offsets, i.e., sudden transitions from continuous speech sound to silence, are relatively rare in the wideband signal. However, they are more common in the narrowband signals which arise through frequency analysis in the auditory periphery. Speech offsets and onsets can appear at the output of the auditory filters during a continuous utterance where the sound's spectrum changes, i.e., at spectral transitions [187]. The detrimental effects of reverberation on speech intelligibility have been attributed to two types of masking. Nábělek et al. [20] found evidence of *overlap-masking*, whereby a preceding phoneme can mask a subsequent segment, and of a *self-masking* of cues within consonants that have time-varying characteristics. It is understood that overlap-masking is primarily caused by late reverberation and self-masking is primarily caused by early reverberation. A more elaborate discussion of these masking types can be found in Section 1.3.2.

Reduction of the detrimental effects of reflections is evidently of considerable practical importance. In this chapter we investigate the application of acoustic signal processing techniques to the enhancement of the quality of speech that is distorted in a reverberant acoustic environment. Novel single- and multi-microphone speech dereverberation algorithms are developed that aim at the suppression of late reverberation, i.e., at estimation of the early speech component. This is done via so-called spectral enhancement techniques that require a specific measure of the late reverberant signal. This measure, called spectral variance, can be estimated directly from the received (possibly noisy) reverberant signal(s) using a statistical reverberation model and a limited amount of *a priori* knowledge about the acoustic channel(s) between the source and the receiver(s). Furthermore, in any practical situation additional interference such as sensor or computer fan noise will be present. Therefore, the joint suppression of late reverberation and other interferences will be discussed.

In [24] a single-microphone speech dereverberation technique based on spectral subtraction was introduced to reduce the effect of overlap-masking in a noise-free environment. The described technique estimates the short-term Power Spectral Density (PSD) of late reverberation based on a statistical reverberation model. The magnitude subtraction technique used by Lebart et al. in [24] can introduce noticeable distortions in the signal. Therefore, the use of a more advanced spectral enhancement technique will be investigated.

In our work an existing single-channel statistical reverberation model serves as a starting point. The model is characterized by one parameter that depends on the acoustic

characteristics of the environment. In Chapter 6 it is shown that the spectral variance estimator that is based on this model, can only be used when the source-microphone distance is smaller than the so-called critical distance. This is, crudely speaking, the distance where the direct sound power is equal to the total reflective power. A generalization of the statistical reverberation model in which the direct sound is incorporated is developed. This model requires one additional parameter that is related to the ratio between the direct sound energy and the sound energy of all reflections. The generalized model is used to derive a novel spectral variance estimator. Compared to the existing estimator the novel estimator improves the dereverberation performance when the source-microphone distance is smaller than the critical distance. In this chapter it is assumed that an estimate of the so-called late reverberant spectral variance is available.

Single-microphone systems only exploit the spectral diversity and the temporal diversity of the received signal. Reverberation, of course, also induces spatial diversity. To additionally exploit this diversity, multiple microphones must be used, and their outputs must be combined by a suitable spatial processor as described in Section 3.2.5. An example of such a spatial processor is the delay and sum beamformer. It is not *a priori* evident whether spectral enhancement is best done before or after the spatial processor. For this reason we investigate both possibilities, as well as a merge of the spatial processor and the spectral enhancement technique. An advantage of the latter option is that the spectral variance estimator can be further improved.

The structure of this chapter is as follows. In Section 5.2 of this chapter the problem is first formulated. Two spectral enhancement techniques are described in Section 5.3. In Section 5.4 three different multi-microphone techniques are proposed and discussed. The performance for different reverberation times using synthetic and measured AIRs is discussed in Section 5.5, where the delay and sum beamformer is used as a reference. Finally, conclusions are provided in Section 5.6.

5.2 Problem Formulation

Reverberation is the process of multi-path propagation of an acoustic signal $s(n)$ from its source to one or more microphones, where n denotes the discrete time index. The observed signal at the m^{th} microphone can be written as

$$x(\mathbf{r}_m, n) = z(\mathbf{r}_m, n) + v(\mathbf{r}_m, n), \tag{5.2}$$

where $\mathbf{r}_m$ denotes the position of the m^{th} microphone, $z(\mathbf{r}_m, n)$ the reverberant signal, and $v(\mathbf{r}_m, n)$ background noise. The reverberant signal can be expressed as

$$z(\mathbf{r}_m, n) = \sum_{j=0}^{L_\mathrm{h}-1} h_j(\mathbf{r}_m, \mathbf{r}_s, n) s(n-j) \tag{5.3}$$

where $h_j(\mathbf{r}_m, \mathbf{r}_s, n)$ denotes the j^{th} coefficient of the impulse response of the acoustic channel from the source to m^{th} microphone, and L_h denotes the length of the impulse

response. The position of the source is denoted by $\mathbf{r}_s$. In vector notation the impulse response is denoted by $\mathbf{h}(\mathbf{r}_m, \mathbf{r}_s, n) = [h_0(\mathbf{r}_m, \mathbf{r}_s, n), \dots h_{L_\mathrm{h}-1}(\mathbf{r}_m, \mathbf{r}_s, n)]^T$.

The aim of dereverberation is to form $\hat{s}(n)$, an estimate of $s(n)$, from $\{x(\mathbf{r}_m, n) \mid m = 0, \dots, M-1\}$. This is a blind problem since neither the signal $s(n)$ nor the acoustic impulse responses $\{\mathbf{h}(\mathbf{r}_m, \mathbf{r}_s, n) \mid m = 0, \dots, M-1\}$ are available. Furthermore, typical acoustic impulse responses are time-varying with several thousand coefficients, making the estimation extremely difficult. It should be noted that for increasing source-microphone distances the Direct to Reverberation Ratio (DRR) decreases, and the reverberation becomes dominant (see Section 2.7). In this situation dereverberation becomes very important and challenging. From the discussion in the introduction and in Section 1.3 it becomes clear that the intelligibility of the reverberant speech can be improved by reducing the amount of late reverberation. Hence, it is not necessarily to completely dereverberate the received signal.

Since the goal is to reduce late reverberation the AIR is split into two segments, $\mathbf{h}_\mathrm{e}(n)$ and $\mathbf{h}_\mathrm{l}(n)$, so that

$$h_j(n) = \begin{cases} h_{\mathrm{e},j}(n), & 0 \le j < N_\mathrm{l}; \\ h_{\mathrm{l},j}(n), & N_\mathrm{l} \le j \le L_\mathrm{h} - 1; \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

The parameter N_l can be chosen depending on the application or subjective preference. Usually N_l is chosen such that $\mathbf{h}_\mathrm{e}(n)$ consists of the direct signal and a few early refections and $\mathbf{h}_\mathrm{l}(n)$ consists of all later reflections, and hence $N_\mathrm{l}/f_\mathrm{s}$ ranges from 40 to 80 ms, where f_s denotes the sampling frequency. Eq. 5.2 can be rewritten using Eq. 5.4:

$$x(\mathbf{r}_m, n) = \underbrace{\sum_{j=0}^{N_\mathrm{l}-1} h_j(\mathbf{r}_m, \mathbf{r}_s, n) s(n-j)}_{z_\mathrm{e}(\mathbf{r}_m, n)} + \underbrace{\sum_{j=N_\mathrm{l}}^{L_\mathrm{h}-1} h_j(\mathbf{r}_m, \mathbf{r}_s, n) s(n-j)}_{z_\mathrm{l}(\mathbf{r}_m, n)} + v(\mathbf{r}_m, n). \tag{5.5}$$

The signal $z_\mathrm{e}(\mathbf{r}_m, n)$ is commonly called the *early speech component*.

In the chapter, novel single- and multi-microphone speech dereverberation algorithms are developed that aim at the suppression of the late reverberant signal $z_\mathrm{l}(n)$ and the background noise $v(n)$, i.e., at estimation of the early speech component $z_\mathrm{e}(n)$.

An overview of the developed single-microphone dereverberation algorithm using spectral enhancement, including the Time-Frequency (TF) analysis and synthesis, is depicted in Fig. 5.1. Compared to the algorithm proposed by Lebart et al. in [24] this algorithm also includes the estimation and reduction of noise.

The observed signal $x(n)$, is transformed into the time-frequency domain by applying the short-time Fourier transform (STFT). Specifically,

$$X(l,k) = \sum_{n=0}^{K-1} x(n + lR) w(n) \mathrm{e}^{-\iota \frac{2\pi k}{K} n} \tag{5.6}$$

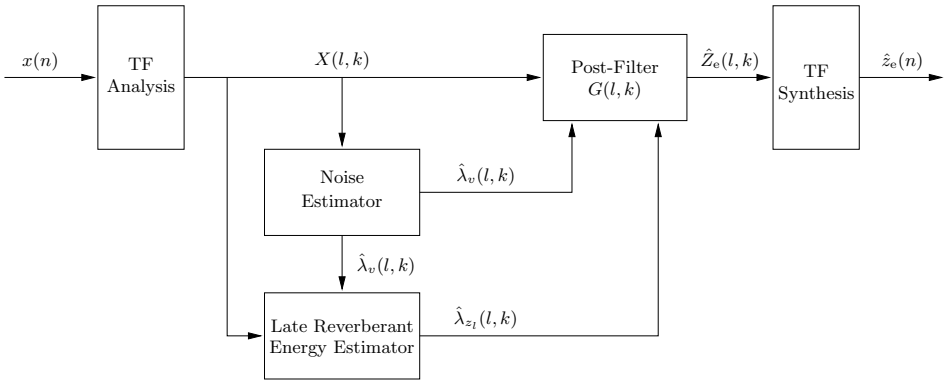


Figure 5.1 *Flow diagram of the developed single-microphone dereverberation algorithm.*

where $\iota = \sqrt{-1}$, l is the time frame index ($l = 0, 1, \dots$), k is the frequency-bin index ($k = 0, 1, \dots, K-1$), $w(n)$ is an analysis window of size K (e.g., a Hamming window), and R is the frame rate (number of samples separating two successive frames).

First the signal $X(l,k)$ is used to estimate the noise spectral variance $\lambda_v(l,k) = \mathcal{E}\{|V(l,k)|^2\}$, where $V(l,k)$ is defined as the STFT of the noise signal $v(n)$. The noise spectral variance can be obtained using power spectral density estimation during noise only periods, or by using a minimum statistics approach [188, 189, 190]. Secondly, the late reverberant spectral variance is estimated, which is defined as $\lambda_{z_1}(l,k) = \mathcal{E}\{|Z_1(l,k)|^2\}$, where $Z_1(l,k)$ denotes the STFT of the signal component $z_1(n)$. The estimation of $\lambda_{z_1}(l,k)$ will be described Chapter 6. Due to the non-stationarity of the source and due to the statistical properties of the AIR we can assume that the early and late reflections are statistically independent (see Chapter 6). Therefore, we can suppress the late reverberant signal by treating it as an additive noise term.

Numerous techniques for the enhancement of noisy speech degraded by statistically independent additive noise have been proposed in the literature. An elaborate discussion will be provided in Section 5.3 where two post-filters will be proposed to suppress the late reverberant signal and the background noise. The post-filter estimates the STFT $Z_{\mathrm{e}}(l,k)$ of the early speech signal $z_{\mathrm{e}}(n)$, and requires an estimate of the late reverberated spectral variance $\hat{\lambda}_{z_1}(l,k)$ and the noise spectral variance $\hat{\lambda}_v(l,k)$. The early speech spectrum $Z_{\mathrm{e}}(l,k)$ is constructed by applying a time and frequency dependent gain function $G(l,k)$ to $X(l,k)$, i.e.,

$$\hat{Z}_{\mathrm{e}}(l,k) = G(l,k)X(l,k). \tag{5.7}$$

Since the speech signal $z_{\mathrm{e}}(n)$ is assumed real, once we estimate $\{Z_{\mathrm{e}}(l,k) \mid k = 1, \dots, K/2\}$, the spectral coefficients for $K/2 < k \leq K-1$ are obtained by $\hat{Z}_{\mathrm{e}}(l,k) = \hat{Z}^*_{\mathrm{e}}(l, K-k)$, where $*$ denotes complex conjugation. The DC component $\hat{Z}_{\mathrm{e}}(l,0)$ is set to zero. Given an estimate $\hat{Z}_{\mathrm{e}}(l,k)$ for the STFT of the early speech signal, an

estimate for the early speech signal is obtained by applying the inverse STFT,

$$\hat{z}_{\mathrm{e}}(n) = \sum_{l} \sum_{k=0}^{K-1} \hat{Z}_{\mathrm{e}}(l,k)\tilde{w}(n-lR)\mathrm{e}^{\iota\frac{2\pi}{K}(n-lR)k}. \tag{5.8}$$

where $\tilde{w}(n)$ is a synthesis window that is bi-orthogonal to the analysis window $w(n)$ [191]. The inverse STFT is efficiently implemented by using the weighted overlap-add method [192].

5.3 Spectral Enhancement

In this section we will further elaborate on the spectral enhancement techniques are used to estimate the early speech component.

Spectral enhancement of noisy speech has been a challenging problem for many researchers for over thirty years, and is still an active research area, see for example [193, 194, 195] and references therein. Only recently these techniques have been used for speech dereverberation [24, 196]. Spectral enhancement of noisy speech is often formulated as estimation of speech spectral components from a speech signal degraded by statistically independent additive noise. In this section we consider spectral enhancement methods for single microphone setups. The situation of single microphone setups is particularly difficult under non-stationary noise and low Signal to Noise Ratio (SNR) conditions, since no reference signal is available for the estimation of the noise.

One of the earlier methods, and perhaps the most well-known method, is spectral subtraction [197, 198]. According to this method, an estimate of the short-term power spectral density of the clean signal is obtained by subtracting an estimate of the power spectral density of the background noise from the short-term power spectral density of the degraded signal. The square root of the resulting estimate is considered an estimate of the spectral magnitude of the speech signal. Subsequently, an estimate for the signal is obtained by combining the spectral magnitude estimate with the complex exponential of the phase of the noisy signal. This method generally results in random narrow-band fluctuations in the residual noise, also known as musical tones, which are annoying and disturbing to the perception of the enhanced signal. Many variations have been developed to cope with musical tones [197, 199, 200, 201, 202], including spectral subtraction techniques based on masking properties of the human auditory system [203, 204]. The spectral subtraction makes minimal assumptions about the signal and noise, and when carefully implemented, it produces enhanced signals that may be acceptable for certain applications.

Statistical methods [205, 206, 207, 208, 209] are often designed to minimize the expected value of some distortion measure between the clean and estimated signals. This method requires reliable statistical models for the speech and noise signals, a perceptually meaningful distortion measure, and an efficient signal estimator. A statistical speech model and perceptually meaningful distortion measure, which are the most ap-

propriate for spectral enhancement, have not yet been determined. Hence, the variety of statistical methods for spectral enhancement mainly differ in the statistical model [205, 207, 208], distortion measure [210, 211, 212], and the particular implementation of the spectral enhancement algorithm [195].

Spectral enhancement based on Hidden Markov Processes (HMP) tries to circumvent the assumption of specific distributions for the speech and noise processes [213, 214, 215, 216]. The probability distributions of the two processes are first estimated from long training sequences of anechoic speech and noise samples, and then used jointly with a given distortion measure to derive an estimator for the speech signal. The HMP-based speech enhancement relies on the type of training data [217, 218]. It works best with the trained types of noise, but often worse with other type of noise. Furthermore, improved performance generally entails more complex models and higher computational requirements. While hidden Markov models have been successfully applied to automatic recognition of anechoic speech signals [219], they were not found to be sufficiently refined models for speech enhancement applications [194].

Subspace methods [220, 221, 222, 223] attempt to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace. Spectral enhancement is performed by removing the noise subspace and estimating the speech signal from the remaining subspace. The signal subspace decomposition can be achieved by either using the Karhunen-Loeve transform (KLT) via eigenvalue decomposition of a Toeplitz covariance estimate of the noisy vector [220, 222], or by using the singular value decomposition of a data matrix [224, 225]. Linear estimation in the signal-plus-noise subspace is performed with the goal of minimizing signal distortion while masking the residual noise by the signal.

5.3.1 Spectral Subtraction

Lebart et al. proposed to use spectral subtraction for speech dereverberation of noise-free speech in [24]. In this section we will briefly describe this procedure. Additionally, we include the ability to suppress background noise.

The spectral subtraction technique can be related to the estimation of a short-time spectral attenuation factor. Since the early spectral, late reverberant spectral, and noise spectral components are assumed to be statistically independent, the short-time spectral attenuation factor is adjusted as a function of the *a posteriori* Signal to Interference Ratio (SIR) for each time and frequency. The *a posteriori* SIR is defined as

$$\gamma(l,k) = \frac{|X(l,k)|^2}{\lambda_{z_1}(l,k) + \lambda_v(l,k)}. \tag{5.9}$$

where $\lambda_{z_1}(l,k) = \mathcal{E}\{|Z_1(l,k)|^2\}$ denotes the late reverberant spectral variance, and $\lambda_v(l,k) = \mathcal{E}\{|V(l,k)|^2\}$ denotes the noise spectral variance which can be obtained using power spectral density estimation during noise only periods, or by using a minimum statistics approach [188, 189, 190]. We have used the Improved Minima Controlled Re-

cursive Averaging (IMCRA) algorithm proposed by Cohen [190] to obtain an estimate of the noise spectral variance $\lambda_v(l,k)$ directly from $X(l,k)$.

The short-time spectral attenuation factor can be defined as [197, 226]

$$G(l,k) = \left(1 - \left(\frac{1}{\gamma(l,k)}\right)^{\beta_1}\right)^{\beta_2}. \tag{5.10}$$

Methods like magnitude subtraction ($\beta_1 = 1/2$, $\beta_2 = 1$), power subtraction ($\beta_1 = 1$, $\beta_2 = 1/2$) and Wiener estimation ($\beta_1 = 1$, $\beta_2 = 1$) are special cases of Eq. 5.10.

Similar to the findings in [24] we concluded that magnitude subtraction gives better performance compared to power subtraction and Wiener estimation. This results in the following gain function

$$G(l,k) = 1 - \frac{1}{\sqrt{\gamma(l,k)}}. \tag{5.11}$$

However, in all frames it is possible that for some frequencies the estimated amplitude spectrum $\sqrt{\hat{\lambda}_{z_1}(l,k) + \hat{\lambda}_v(l,k)}$ of the interference is larger than the instantaneous amplitude $|X(l,k)|$ of the received spectrum. Since this could lead to negative estimates for the amplitude of the early speech spectrum $Z_e(l,k)$, the gain function $G(l,k)$ is usually put to zero (i.e., half-wave rectification). However, because of the non-stationary character of the speech signal, this non-linear rectification leads to a specific kind of residual noise which consists of short tones with randomly distributed frequencies. Different techniques have been proposed to eliminate this annoying residual noise, e.g., by averaging the (instantaneous) noisy speech spectrum over a number of frames, or by using non-linear spectral subtraction techniques [227].

Here, the residual noise problem is alleviated using two standard modifications. The first modification consists of replacing the *a posteriori* SIR in Eq. 5.11 by the *a priori* SIR $\xi(l,k)$ plus one, i.e., $\gamma(l,k) = \xi(l,k) + 1$. The *a priori* SIR is defined as

$$\xi(l,k) = \frac{\mathcal{E}\{|Z_e(l,k)|^2\}}{\lambda_{z_1}(l,k) + \lambda_v(l,k)}, \tag{5.12}$$

where $\mathcal{E}\{|Z_e(l,k)|^2\}$ denotes the early speech spectral variance which is not available in practice. Ephraim and Malah [205] proposed a very useful estimation method for the *a priori* SIR, which is known as the decision-directed estimation method. The *a priori* SIR can be substituted by the following expression:

$$\xi(l,k) = \eta \frac{|\hat{Z}_e(l-1,k)|^2}{\hat{\lambda}_{z_1}(l-1,k) + \hat{\lambda}_v(l-1,k)} + (1-\eta)\max\{\gamma(l,k) - 1, 0\}. \tag{5.13}$$

The first term, $\frac{|\hat{Z}_e(l-1,k)|^2}{\hat{\lambda}_{z_1}(l-1,k)+\hat{\lambda}_v(l-1,k)}$, represents the *a priori* SIR resulting from the processing of the previous frame. The second term, $\max\{\gamma(l,k) - 1, 0\}$, is a maximum

likelihood estimate for the *a priori* SIR, based entirely on the current frame. The parameter η ($0 \leq \eta \leq 1$) denotes a weighting factor that controls the trade-off between the interference reduction and the transient distortion brought into the signal [205]. A larger value of η results in a greater reduction of the residual noise, but at the expense of attenuated speech onsets and audible modifications of transient components. As a compromise, a value 0.98 of η was determined by simulations and informal listening tests for the purpose of noise reduction [205]. The second modification of the standard gain function consists of using a gain floor ($G_{\min}$) which constrains the minimum value of the gain function. The gain floor also gives us the possibility to control the maximum amount of interference reduction. The gain floor is usually considered to be a constant value. However, from a perceptual point of view it is desired to have a constant residual background noise level. Therefore the gain floor is made signal dependent by using

$$\tilde{G}_{\min}(l,k) = G_{\min}\frac{\lambda_v(l,k)}{\lambda_{z_1}(l,k)+\lambda_v(l,k)}. \tag{5.14}$$

The derivation, and a more elaborate discussion, of this modification can be found in Appendix B. Applying above modifications to the standard gain function in Eq. 5.11 results in the following gain function

$$G(l,k) = \max\left\{1 - \frac{1}{\sqrt{\xi(l,k)+1}}, \tilde{G}_{\min}(l,k)\right\}. \tag{5.15}$$

5.3.2 OM-LSA Estimator

In this section we propose to use a more advanced spectral enhancement technique based on statistical signal modelling. The Optimally-Modified Log Spectral Amplitude estimator is used to obtain an estimate of the desired spectral component $Z_{\mathrm{e}}(l,k)$. The minimum mean-square error Log Spectral Amplitude (LSA) estimator proposed by Ephraim and Malah [205] minimizes

$$\mathcal{E}\left\{\left(\log(A(l,k)) - \log(\hat{A}(l,k))\right)^2\right\}, \tag{5.16}$$

where $A(l,k) = |Z_{\mathrm{e}}(l,k)|$ denotes the spectral speech amplitude, and $\hat{A}(l,k)$ is its optimal estimator. Assuming spectral coefficients are conditionally independent given their variances [208], the LSA estimator is defined as

$$\hat{A}(l,k) = \exp\left(\mathcal{E}\{\log(A(l,k))|Z_{\mathrm{e}}(l,k)\}\right). \tag{5.17}$$

The LSA gain function, based on a Gaussian statistical model, is given by

$$G_{\mathrm{LSA}}(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\exp\left(\frac{1}{2}\int_{\zeta(l,k)}^{\infty}\frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t\right), \tag{5.18}$$

where

$$\xi(l,k) = \frac{\mathcal{E}\{|Z_{\mathrm{e}}(l,k)|^2\}}{\lambda_{z_1}(l,k)+\lambda_v(l,k)}, \tag{5.19}$$

$$\gamma(l,k) = \frac{|X(l,k)|^2}{\lambda_{z_l}(l,k) + \lambda_v(l,k)}, \tag{5.20}$$

and

$$\zeta(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\,\gamma(l,k). \tag{5.21}$$

The Optimally-Modified Log Spectral Amplitude (OM-LSA) spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty [228]. Given two hypotheses, $\mathrm{H}_0(l,k)$ and $\mathrm{H}_1(l,k)$, which indicate speech absence and speech presence, respectively, we have

$$\begin{aligned} \mathrm{H}_0(l,k) &: X(l,k) = Z_{\mathrm{l}}(l,k) + V(l,k), \\ \mathrm{H}_1(l,k) &: X(l,k) = Z_{\mathrm{e}}(l,k) + Z_{\mathrm{l}}(l,k) + V(l,k). \end{aligned} \tag{5.22}$$

Based on a Gaussian statistical model, the speech presence probability is given by

$$p(l,k) = \left\{1 + \frac{q(l,k)}{1-q(l,k)}\,(1+\xi(l,k))\exp\left(-\zeta(l,k)\right)\right\}^{-1}, \tag{5.23}$$

where $q(l,k)$ is the *a priori* signal absence probability [228], which will be discussed on page 105.

The OM-LSA gain function is given by,

$$G_{\mathrm{OM-LSA}}(l,k) = \{G_{\mathrm{H}_1}(l,k)\}^{p(l,k)}\,\{G_{\mathrm{H}_0}(l,k)\}^{1-p(l,k)}, \tag{5.24}$$

with $G_{\mathrm{H}_1}(l,k) = G_{\mathrm{LSA}}(l,k)$ and $G_{\mathrm{H}_0}(l,k) = G_{\min}$. The lower-bound constraint for the gain when the signal is absent is denoted by $G_{\min}$, and specifies the maximum amount of reduction in those frames.

In our case the lower-bound constraint does not result in the desired result since the late reverberant signal can still be audible. Our goal is to suppress the late reverberant signal down to the noise floor, given by $G_{\min}\,V(l,k)$, where $V(l,k)$ denotes the STFT of the noise signal $v(n)$. In Appendix B we derive $G_{\mathrm{H}_0}(l,k)$ that meets this goal, i.e.,

$$G_{\mathrm{H}_0}(l,k) = G_{\min}\frac{\lambda_v(l,k)}{\lambda_{z_l}(l,k) + \lambda_v(l,k)}. \tag{5.25}$$

A priori SIR estimators

Many researchers believe that the main advantage of the LSA estimator is related to the decision-directed estimator, proposed by Ephraim and Malah [205]. Other estimators like the causal recursive estimator, and non-causal recursive estimator, were recently proposed by Cohen [208]. Compared to the decision-directed *a priori* SIR estimator the non-causal *a priori* SIR estimator is more reliable during speech onsets, and hence reduces distortion in these periods.

Rather than using one *a priori* Signal to Interference Ratio it is possible to calculate one value for each interference. By doing this, one gains control over the interference reduction level, and the *a priori* SIR estimation approach, of interference. Note that in some cases it might be desirable to reduce one of the interferences at the cost of larger speech distortion, while other interferences are reduced less to avoid distortion. Due to the separation we can control the tradeoff between noise reduction and distortion of each of the interferences separately. A more elaborate discussion can be found in Appendix B. In this appendix we show how the decision-directed estimator, and the causal and non-causal recursive estimators, can be used to estimate the individual *a priori* SIRs. We also show how these values can be combined to obtain the total *a priori* SIR $\xi(l,k)$. It should be noted that each *a priori* SIR could be estimated using a different estimator.

A priori Signal Absence Probability

In this section we propose an efficient estimator for the *a priori* signal absence probability $q(l,k)$ that exploits spatial information ($M > 2$). This estimator uses a soft-decision approach to compute four parameters. Three parameters, i.e., $P_{\text{local}}(l,k)$, $P_{\text{global}}(l,k)$, and $P_{\text{frame}}(l)$, are proposed by Cohen in [228], and are based on the time-frequency distribution of the estimated *a priori* SIR, $\xi(l,k)$. These parameters exploit the strong correlation of speech presence in neighbouring frequency bins of consecutive frames. To further improve the *a priori* signal absence probability we propose to use a fourth parameter that exploits spatial information. Since a strong coherency between the microphone signals will indicate the presence of a direct signal, we propose to relate our fourth parameter to the Mean Square Coherence (MSC) of two microphone signals. By using the MSC we mimic the binaural overlap-masking release (see Section 1.3) that is used by the human auditory system. To use the MSC its value needs to be related to the probability $P_{\text{spatial}}(l,k)$. Since the MSC is already normalized its value lies between zero and one. To improve the estimation we first smooth the MSC in time and frequency, before the MSC value is related to $P_{\text{spatial}}(l,k)$.

The MSC is defined as

$$\Phi_{\text{MSC}}(l,k) \triangleq \frac{\mathcal{S}\{X_{21}(l,k)\}}{\mathcal{S}\{X_1(l,k)\}\mathcal{S}\{X_2(l,k)\}}, \tag{5.26}$$

where $X_{21}(l,k) = X_2(l,k)X_1^*(l,k)$, and the operator $\mathcal{S}$ denotes smoothing in time, i.e.,

$$\mathcal{S}\{Y(l,k)\} = \eta_{\text{s}}\mathcal{S}\{Y(l-1,k)\} + (1-\eta_{\text{s}})|Y(l,k)|^2, \tag{5.27}$$

where η ($0 \leq \eta_{\text{s}} \leq 1$) is the smoothing parameter. The MSC is further smoothed over different frequencies using

$$\tilde{\Phi}_{\text{MSC}}(l,k) = \sum_{i=-w_{\text{msc}}}^{w_{\text{msc}}} b_i \Phi_{\text{MSC}}(l,k+i) \tag{5.28}$$

where b is a normalized window function ($\sum_{i=-w_{\text{msc}}}^{w_{\text{msc}}} b_i = 1$) that determines the frequency smoothing.

The spatial speech presence probability $P_{\text{spatial}}(l,k)$ is related to Eq. 5.26 by

$$P_{\text{spatial}}(l,k) = \begin{cases} 0, & \tilde{\Phi}_{\text{MSC}}(l,k) \leq \Phi_{\min}; \\ 1, & \tilde{\Phi}_{\text{MSC}}(l,k) \geq \Phi_{\max}; \\ \frac{\tilde{\Phi}_{\text{MSC}}(l,k)-\Phi_{\min}}{\Phi_{\max}-\Phi_{\min}}, & \Phi_{\min} \leq \tilde{\Phi}_{\text{MSC}}(l,k) \leq \Phi_{\max}, \end{cases} \tag{5.29}$$

where $\Phi_{\min}$ and $\Phi_{\max}$ are the minimum and maximum threshold values for $\tilde{\Phi}_{\text{MSC}}(l,k)$, respectively. The proposed *a priori* speech absence probability is given by

$$\hat{q}(l,k) = 1 - P_{\text{local}}(l,k)P_{\text{global}}(l,k)P_{\text{spatial}}(l,k)P_{\text{frame}}(l). \tag{5.30}$$

In case $M > 2$ one could average the MSC over different microphones pairs to improve the estimation procedure even further. The spatially average MSC is given by

$$\bar{\Phi}_{\text{MSC}}(l,k) = \frac{2!(M-2)!}{M!} \sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \frac{\mathcal{S}\{X_{ji}(l,k)\}}{\mathcal{S}\{X_i(l,k)\}\mathcal{S}\{X_j(l,k)\}}, \tag{5.31}$$

and can then be used in Eq. 5.28. Note that for M microphones there are $\binom{M}{2} = \frac{M!}{2!(M-2)!}$ possible combinations.

5.4 Proposed Multi-Microphone Systems

Single-microphone systems only exploit the spectral diversity and the temporal diversity of the received signal. Reverberation, of course, also induces spatial diversity. To be able to additionally exploit this diversity multiple microphones must be used, and their outputs must be combined by a suitable spatial processor such as the so-called delay and sum beamformer. It is not *a priori* evident whether spectral enhancement is best done before or after the spatial processor. For this reason we investigate both possibilities, as well as a merge of the spatial processor and the spectral enhancement technique. An advantage of the latter option is that the spectral variance estimator can be further improved. In this section the multi-microphone systems are proposed and the pros and cons of each system are discussed.

5.4.1 Spatial Processor with Post-Processor

In this section we propose a multi-microphone system that consists of a spatial processor and post-processor. The post-processor consists of one of the developed single-microphone spectral enhancement techniques. The structure of the proposed system is depicted in Fig. 5.2. Here $\{x(\mathbf{r}_m, n) \mid m = 0, \ldots, M-1\}$ denote the noisy microphone signals, $y(n)$ the output of the spatial processor, and $\hat{s}(n)$ the estimated source signal. The reverberation reduction performance of such a system will depend on the performance of the spatial processor and the post-processor, and the effect

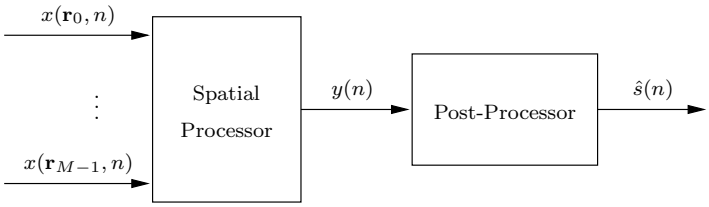


Figure 5.2 *Multi-microphone speech dereverberation system which consists of a spatial processor and a single-microphone dereverberation post-processor.*

of the spatial processor on the acoustic channel. Note that for noise reduction such a structure has been frequently used [143, 229, 230, 144, 145] to enhance the noise reduction performance of the spatial processor.

A possible spatial processor that could be used is, for example, the Transfer Function Generalized Sidelobe Canceller (TF-GSC) (see Section 3.2.5, page 65). In theory the total transfer function, which describes the system between one reference microphone and the output of the TF-GSC, is equal one. Hence, reverberation is not reduced by the spatial processor, and the total transfer function which describes the system between the source and the output of the spatial processor is equal to the reference transfer function, i.e., no reverberation is reduced by the TF-GSC. The advantage of the TF-GSC is that dereverberation can be performed in a possibly adverse noise environment, since the TF-GSC is able to suppress coherent non-stationary noise sources. In case the TF-GSC is used in conjunction with the post-processor the dereverberation performance is limited by the post-processor.

An alternative spatial processor is the delay and sum beamformer. In Section 3.2.5 we have seen that this beamformer can improve the DRR of the signal. Therefore, the DRR at the output of the spatial processor can generally be larger than 0 dB. In the proposed structure the output of the spatial processor is further enhanced using the post-processor. Therefore, the total transfer that describes the system between the source and the output of the spatial processor could be completely different from the transfer function that describes the system between the source and a single microphone. Informal listening test indicated that when the output of the delay and sum beamformer was used by the post-processor some audible distortions were introduced. Therefore, we will now analyse the total transfer function that describes the system between the source and the output of the delay and sum beamformer. Let us consider the output of the delay and sum beamformer in the frequency domain:

$$\begin{aligned} Y(\omega) &= \frac{1}{M}\sum_{m=0}^{M-1} X(\mathbf{r}_m;\omega)\mathrm{e}^{-\iota\omega\tau_m} \\ &= \frac{1}{M}\sum_{m=0}^{M-1} H(\mathbf{r}_m,\mathbf{r}_s;\omega)S(\omega)\mathrm{e}^{-\iota\omega\tau_m} \\ &= \bar{H}(\omega)S(\omega), \end{aligned} \tag{5.32}$$

where $H(\mathbf{r}_m,\mathbf{r}_s;\omega)$ denotes the acoustic transfer function that describes the system

between the source and the m^{th} microphone, $\bar{H}(\omega)$ is the total transfer function, and τ_m is propagation delay of the direct signal from the source to the m^{th} microphone. Using Statistical Room Acoustics (SRA) theory (as introduced in Section 2.6) it can be shown that the expected energy density of the total transfer $\bar{H}(\omega)$ can be expressed as:

$$\mathcal{E}_\theta\left\{|\bar{H}(\omega)|^2\right\} = \frac{1}{M^2}\left(\sum_{m=0}^{M-1}\mathcal{E}_\theta\left\{|H(\mathbf{r}_m,\mathbf{r}_s;\omega)|^2\right\} + \sum_{m=0}^{M-1}\sum_{\substack{n=0\\n\neq m}}^{M-1}\mathcal{E}_\theta\left\{H(\mathbf{r}_m,\mathbf{r}_s;\omega)H^*(\mathbf{r}_n,\mathbf{r}_s;\omega)\right\}\mathrm{e}^{-\iota\omega(\tau_m-\tau_n)}\right), \quad (5.33)$$

where $\mathcal{E}_\theta\{\cdot\}$ denotes the spatial expectation. It follows that

$$\mathcal{E}_\theta\left\{|H(\mathbf{r}_m,\mathbf{r}_s;\omega)|^2\right\} = \frac{1}{16\pi^2 D_m^2} + \frac{1-\bar{\alpha}}{\pi\bar{\alpha}S}, \quad (5.34)$$

and

$$\mathcal{E}_\theta\left\{H(\mathbf{r}_m,\mathbf{r}_s;\omega)H^*(\mathbf{r}_n,\mathbf{r}_s;\omega)\right\} = \frac{\mathrm{e}^{\iota k'(D_m-D_n)}}{16\pi^2 D_m D_n} + \left(\frac{1-\bar{\alpha}}{\pi\bar{\alpha}S}\right)\frac{\sin\left(k'\|\mathbf{r}_m-\mathbf{r}_n\|\right)}{k'\|\mathbf{r}_m-\mathbf{r}_n\|}. \quad (5.35)$$

where $k' = \omega/c$ is the wave number, and c is the sound velocity. Let us assume that the delay and sum beamformer is perfectly steered in the direction of the source, i.e., $\tau_m = D_m/c$. The expected energy density of the total transfer $\bar{H}(\omega)$ can then be expressed as:

$$\mathcal{E}_\theta\left\{|\bar{H}(\omega)|^2\right\} = \frac{1}{M^2}\sum_{m=0}^{M-1}\sum_{n=0}^{M-1}\frac{1}{16\pi^2 D_m D_n} + \frac{1}{M}\left(\frac{1-\bar{\alpha}}{\pi\bar{\alpha}S}\right) + \frac{1}{M^2}\left(\frac{1-\bar{\alpha}}{\pi\bar{\alpha}S}\right)\sum_{m=0}^{M-1}\sum_{\substack{n=0\\n\neq m}}^{M-1}\frac{\sin\left(k'\|\mathbf{r}_m-\mathbf{r}_n\|\right)}{k'\|\mathbf{r}_m-\mathbf{r}_n\|}\cos(k'(D_m-D_n)), \quad (5.36)$$

Note that the first two terms in Eq. 5.36 are frequency independent and that the last term in Eq. 5.36 results from the spatial correlation between the channels. The last term becomes large when the receivers are closely spaced, i.e., $\|\mathbf{r}_m-\mathbf{r}_n\| \to 0$, and is likely the origin of the audible distortions in the signal $\hat{s}(n)$.

Our main goal is to exploit the reverberation reduction performance of both the spatial processor and the post-processor. However, the spatial processor might have an undesired influence on the total transfer function that describes the system between the source and the output of the spatial processor. As an example we have seen that the delay and sum beamformer can introduce 'distortions' due to the spatial correlation between the acoustic channels. These distortions can violate the assumptions on which the late reverberant spectral variance estimator is based, and hence deteriorate the performance of the post-processor.

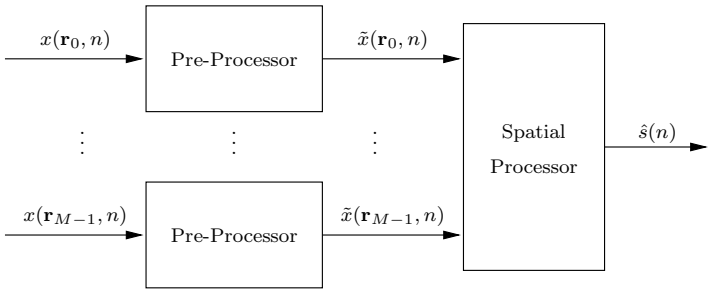


Figure 5.3 *Multi-microphone speech dereverberation system using several single-microphone dereverberation pre-processors.*

5.4.2 Pre-Processor with Spatial Processor

In the previous section we have seen that the spatial processor can have an undesired influence on the signal that is used by the post-processor. By using the same post-processor prior to the spatial processor, i.e., as a pre-processor, we eliminate the influence of the spatial processor on the post-processor. The alternative multi-microphone system thus consists of a single-microphone dereverberation pre-processor for each microphone, followed by a spatial processor, as depicted in Fig. 5.3. Here $x(\mathbf{r}_m, n)$ and $\tilde{x}(\mathbf{r}_m, n)$ denote the m^{th} noisy microphone signal and pre-processed signal, respectively. The pre-processed signals are processed by the spatial processor to obtain an estimate of the source signal $\hat{s}(n)$. It should be noted that the complexity has increased significantly, since we now need M pre-processors compared to one post-processor. The proposed structure can have many advantages. In Section 3.2.5 we have already noticed that many traditional adaptive spatial processors become ineffective in reverberant environments. Due to the reverberation reduction these problems can be partially solved. Furthermore, the spatial processor can be used to reduce early reflections which are unaffected by the pre-processor. A possible disadvantage of this systems is related to the fact that the pre-processor introduces non-linear distortions in the signal. In case an adaptive spatial processor is used such distortions can influence the adaptation of the spatial processor. In Section 6.5 we will develop a possible technique to estimate the late reverberant spectral variance in a noisy environment. This estimator can be used when an estimate of the interference is available. Unfortunately it is likely to fail in an adverse noise environment, where highly coherent and/or non-coherent non-stationary interferences are present.

5.4.3 Joint Multi-Microphone Dereverberation

The final system is the joint multi-microphone dereverberation system, as depicted in Fig. 5.4, where $\mathbf{X}(l,k) = [X(\mathbf{r}_0, l, k), \ldots, X(\mathbf{r}_{M-1}, l, k)]^T$. In this system the spatial processor and spectral enhancement techniques have been combined. Furthermore, a novel spatial processor was used that does not introduce the problems discussed in the first multi-microphone system. Additionally, all microphone signals are used to

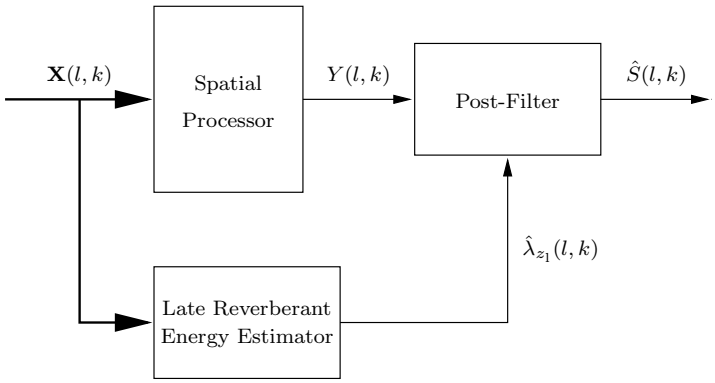


Figure 5.4 *Joint multi-microphone speech dereverberation system.*

estimate the late reverberant spectral variance.

In Section 5.4.1 we have showed that the spatial processor can have an undesired influence on the total transfer function that describes the system between the source and the input of the post-processor. As an example we have showed that delay and sum beamformer introduces a spatial correlation term (see Eq. 5.36). To avoid the introduction of this spatial correlation term we propose the following spatial processor:

$$Y(\omega) = Q(\omega)\mathrm{e}^{\iota\phi_{\mathrm{dsb}}(\omega)} \tag{5.37}$$

where

$$Q(\omega) = \left(\frac{1}{M}\sum_{m=0}^{M-1}|X(\mathbf{r}_m;\omega)\mathrm{e}^{-\iota\omega\tau_m}|^2\right)^{\frac{1}{2}} \tag{5.38}$$

denotes the amplitude spectrum, and

$$\phi_{\mathrm{dsb}}(\omega) = \arg\left\{\frac{1}{M}\sum_{m=0}^{M-1}X(\mathbf{r}_m;\omega)\mathrm{e}^{-\iota\omega\tau_m}\right\} \tag{5.39}$$

denotes the phase spectrum of $Y(\omega)$.

Eq. 5.37 can be expressed as

$$Y(\omega) = \tilde{H}(\omega)S(\omega), \tag{5.40}$$

where $\tilde{H}(\omega)$ denotes the total transfer function of the proposed spatial processor. Let us assume that the array is perfectly steered in the direction of the source, i.e., $\tau_m = D_m/c$. In case the source is in far-field the waves can be modelled as plane waves. Hence the amplitudes of the signals are identical and only the phase differs [130]. Since the phase spectrum of $Y(\omega)$ is equivalent to that of the delay and sum beamformer, i.e., $\arg\{\tilde{H}(\omega)\} = \arg\{\bar{H}(\omega)\}$, the directivity pattern of the proposed spatial processor is equivalent to that of the delay and sum beamformer.

We will now compare the expected energy density of the proposed spatial processor with that of the delay and sum beamformer in Eq. 5.36. The expected energy density of $\tilde{H}(\omega)$ can be expressed as:

$$\mathcal{E}_\theta \left\{ |\tilde{H}(\omega)|^2 \right\} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{1}{16\pi^2 D_m^2} + \frac{1}{M} \left(\frac{1-\bar{\alpha}}{\pi \bar{\alpha} S} \right). \tag{5.41}$$

By averaging the power spectra of received microphone signals rather than there complex spectra we avoid the spatial correlation terms. Furthermore, due to spatial averaging we reduce the spectral deviation, of the transfer function $\tilde{H}(\omega)$. According to Eq. 5.1 we can expect that this can further enhance the subjective speech quality.

5.4.4 Discussion

In the previous sections we proposed three possible multi-microphone speech dereverberation systems. In terms of the computational complexity the first multi-microphone system, i.e., spatial processor with post-processor, is most efficient. In terms of reverberation reduction and possible distortions this method might not be the best choice. In case the microphones are closely spaced the spatial correlation among the acoustic channels introduce undesired components at the output of the spatial processor. The second and third system are both very appealing, and the eventual choice for a specific system will in practice depend on the available processing power and the environmental conditions. It should be noted that in an adverse noise environment the noise has a much larger impact on the speech intelligibility than the late reverberant energy (see for example Eq. 1.2). This means that in such a situation dereverberation might not be the first priority.

5.5 Experiments and Results

In this section we present and discuss the results that were obtained using a single microphone, and multiple microphones. We will compare the results of the three multi-microphone dereverberation system. The first system consists of a delay and sum beamformer followed by a single-microphone speech dereverberation post-processor and will be denoted by SP-SMD (Spatial Processor - Single Microphone Dereverberation). The second system, denoted by SMD-SP (Single Microphone Dereverberation - Spatial Processor), consists of a pre-processor for each microphone signal and a delay and sum beamformer. The final system, denoted by MMD (Multi-Microphone Dereverberation), consists of the proposed spatial processor (Eq. 5.37) and a post-processor that uses the spatially averaged late reverberant spectral variance. The reverberant signals were generated using synthetic and measured acoustic impulse responses. As a reference we evaluated the performance of a standard delay and sum beamformer (DSB).

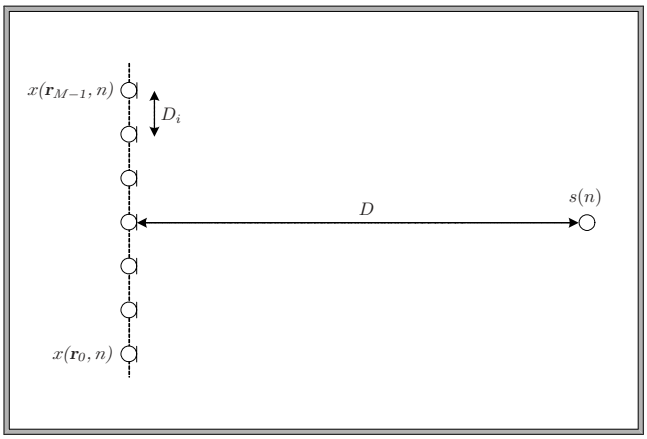


Figure 5.5 *Experimental setup with a uniform linear microphone array.*

The reverberant microphone signals are obtained by convolving 40 seconds (male and female) speech data from the TIMIT speech database [4] with different AIRs. The experimental setup is depicted in Figure 5.5. A number of microphones were uniformly spaced on a straight line, with inner microphone spacing $D_\mathrm{i} = 5$ cm. The source-array distance D is defined as the distance between the source and the center of the array, and ranges from 1 to 3 m. The dimensions of the room are 5 m x 6 m x 4 m (length x width x height). The synthetic AIRs are constructed using the image method for modelling small room acoustics [83], modified to accommodate fractional sample delays according to [231], with reverberation times from 250 to 650 milliseconds. The real AIRs were measured in an office room with dimensions 7.3 m x 6.2 m x 3.2 m (length x width x height) using the Maximum Length Sequence (MLS) technique that is described in Section 2.11. The reverberation time of the office room was 0.529 seconds. The parameters that were used for these experiments are shown in Table 5.1. All *a priori* SIRs are estimated using the decision-directed estimator (Eq. 5.19).

$f_\mathrm{s} = 8000$ Hz	$T_l = 48$ ms	$G_\mathrm{min}^\mathrm{dB} = 18$ dB	$\beta^\mathrm{dB} = 9$ dB
$\eta = 0.95$	$b =$ Hanning window	$w_\mathrm{msc} = 9$	$\Phi_\mathrm{min} = 0.2$
$\Phi_\mathrm{max} = 0.65$	$\eta_s = 0.35$		

Table 5.1 *Parameters used for these experiments.*

5.5.1 Reverberation Suppression

Using synthetic AIRs

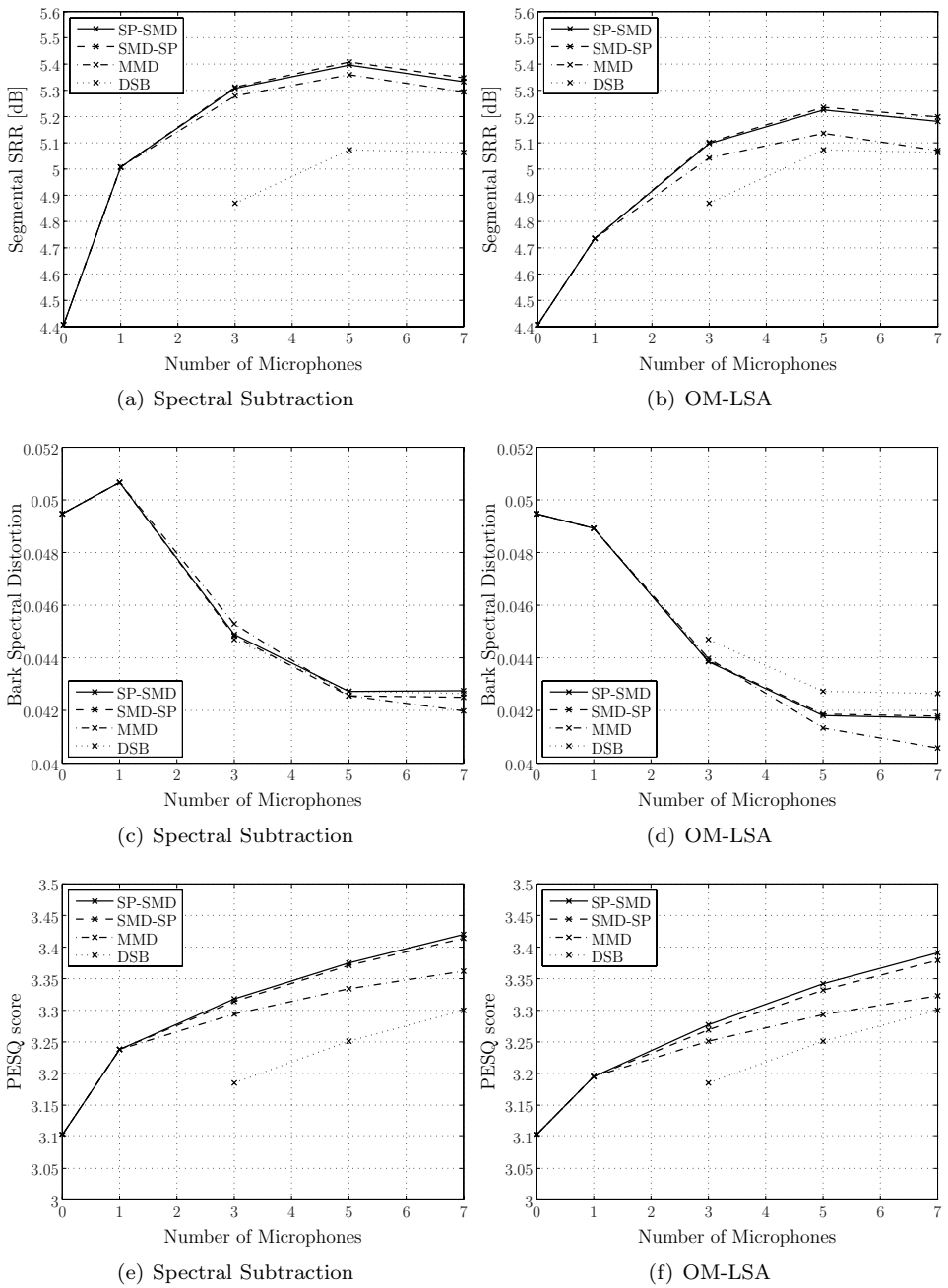
In this section we evaluated the performance of the developed multi-microphone dereverberation systems using 3, 5 and 7 microphones. The results obtained using a single microphone are shown as a comparison. In Figs. 5.6 and 5.7 the segmental Signal to Reverberation Ratio (SRR), Bark Spectral Distortion (BSD) and Perceptual Evaluation of Speech Quality (PESQ) score are depicted for $D = 1$ m and $\mathrm{RT}_{60} = 0.25$ s, and for $D = 3$ m and $\mathrm{RT}_{60} = 0.5$ s, respectively. The results depicted for $M = 0$ were obtained from the unprocessed reverberant signal (center microphone). The dotted lines represent the results that were obtained using the delay and sum beamformer.

The segmental SRR that was obtained by the proposed systems is much higher than the segmental SRR that was obtained by the delay and sum beamformer. The speech distortion in terms of the Bark spectral distortion is equal or lower than the speech distortion of the delay and sum beamformer for $D = 1$ m and $\mathrm{RT}_{60} = 0.25$ s. For $D = 3$ m and $\mathrm{RT}_{60} = 0.5$ s, there is slightly more distortion. From these results we also see that the segmental SRRs and PESQ scores increase, and the BSDs decrease, when more microphones are used. The performances of these systems in terms of segmental SRR, BSD and PESQ scores are very comparable. Informal listening tests indicated an audible distortion in the output of the SP-SMD system, which was not audible in the other systems. This distortion is most likely caused by the spatial correlation between the acoustic channels, as described in Section 5.4.1.

Both the spectral subtraction technique and the OM-LSA estimation technique, perform very well. At first one might prefer the spectral subtraction technique since it achieves higher segmental SRRs. However, since the spectral subtraction technique results in higher BSDs compared to the OM-LSA estimation technique, the results are not necessarily better. In general we can say that, compared to the spectral subtraction technique, the OM-LSA estimation technique obtains a lower segmental SRR and a lower BSD.

Using measured AIRs

In this section we have evaluated the performance of the developed multi-microphone dereverberation systems for 3, 5 and 7 microphones using real AIRs. The results obtained using a single microphone are shown as a comparison. The results for $D = 1$ m and $D = 3$ m are shown in Figs. 5.8 and 5.9, respectively. The results depicted for $M = 0$ were obtained from the unprocessed reverberant signal (center microphone). The dotted lines represent the results that were obtained using the delay and sum beamformer. The results obtained using the measured AIRs are similar to those obtained using the synthetic AIRs. In general the OM-LSA estimation technique results in a lower BSD and a lower segmental SRR. However, at a distance of 3 m the OM-LSA estimation technique introduces slightly more distortion, i.e., lower PESQ scores and BSD values, compared to the spectral subtraction technique.



(a) Spectral Subtraction (b) OM-LSA

(c) Spectral Subtraction (d) OM-LSA

(e) Spectral Subtraction (f) OM-LSA

Figure 5.6 *Objective measures obtained for the four systems, i.e., the spatial processor with post-processor (SP-SMD), the pre-processors with spatial processor (SMD-SP), the joint multi-microphone dereverberation system (MMD), and the delay and sum beamformer (DSB). The post-processor and pre-processors are based on (a-c-d) spectral subtraction, and (b-d-f) the OM-LSA estimator. The reverberant signals were created using synthetic AIRs with a reverberation time of 250 ms. The source-array distance was 1 m and T_1 was set to 48 ms.*

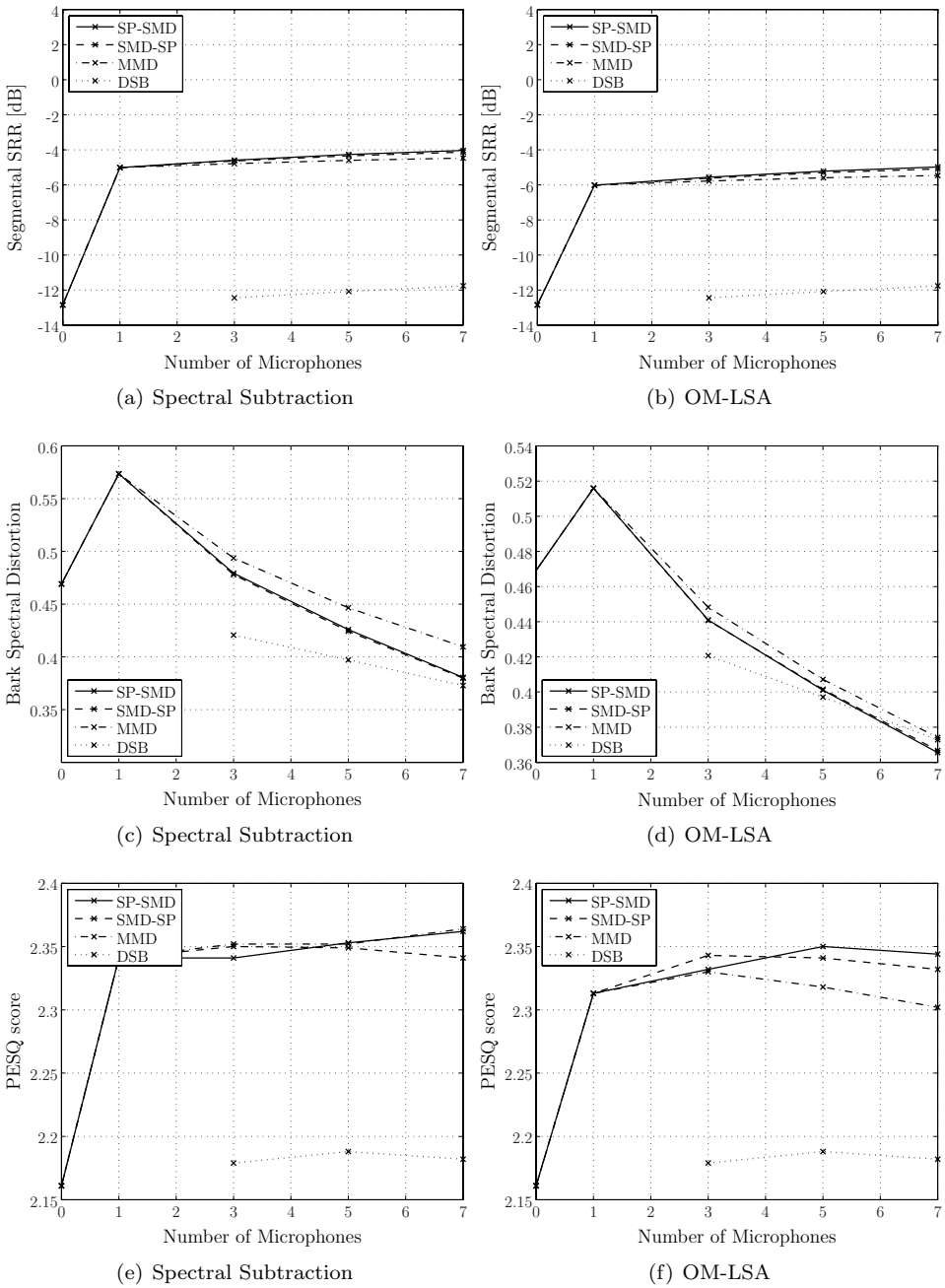
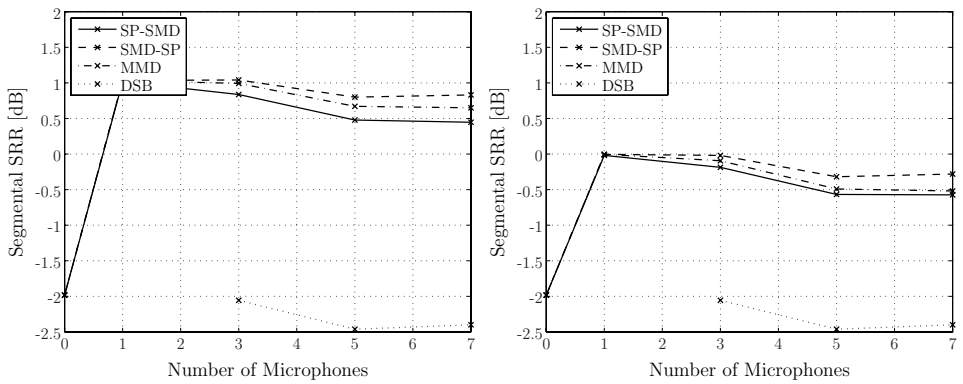
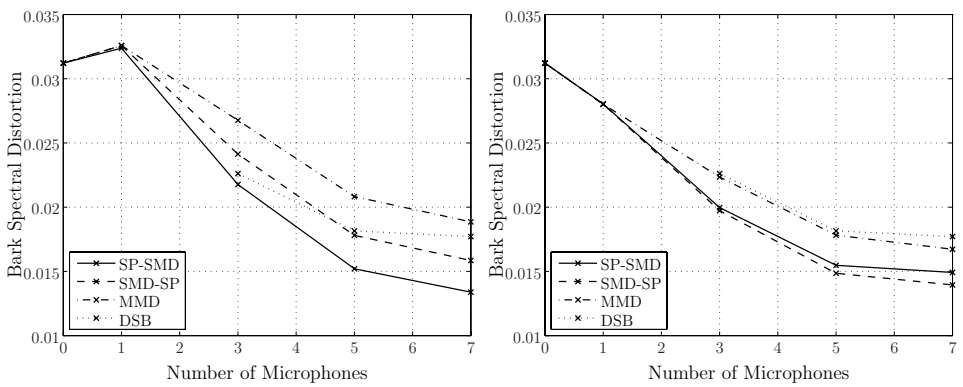


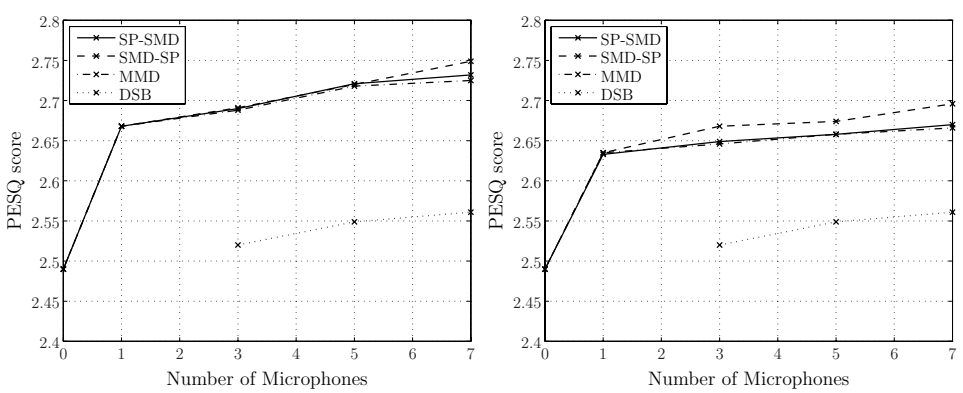
Figure 5.7 *Objective measures obtained for the four systems, i.e., the spatial processor with post-processor (SP-SMD), the pre-processors with spatial processor (SMD-SP), the joint multi-microphone dereverberation system (MMD), and the delay and sum beamformer (DSB). The post-processor and pre-processors are based on (a-c-d) spectral subtraction, and (b-d-f) the OM-LSA estimator. The reverberant signals were created using synthetic AIRs with a reverberation time of 500 ms. The source-array distance was 3 m and T_1 was set to 48 ms.*



(a) Spectral Subtraction (b) OM-LSA



(c) Spectral Subtraction (d) OM-LSA



(e) Spectral Subtraction (f) OM-LSA

Figure 5.8 *Objective measures obtained for the four systems, i.e., the spatial processor with post-processor (SP-SMD), the pre-processors with spatial processor (SMD-SP), the joint multi-microphone dereverberation system (MMD), and the delay and sum beamformer (DSB). The post-processor and pre-processors are based on (a-c-d) spectral subtraction, and (b-d-f) the OM-LSA estimator. The reverberant signals were created using measured AIRs with a reverberation time of 529 ms. The source-array distance was 1 m and T_1 was set to 48 ms.*

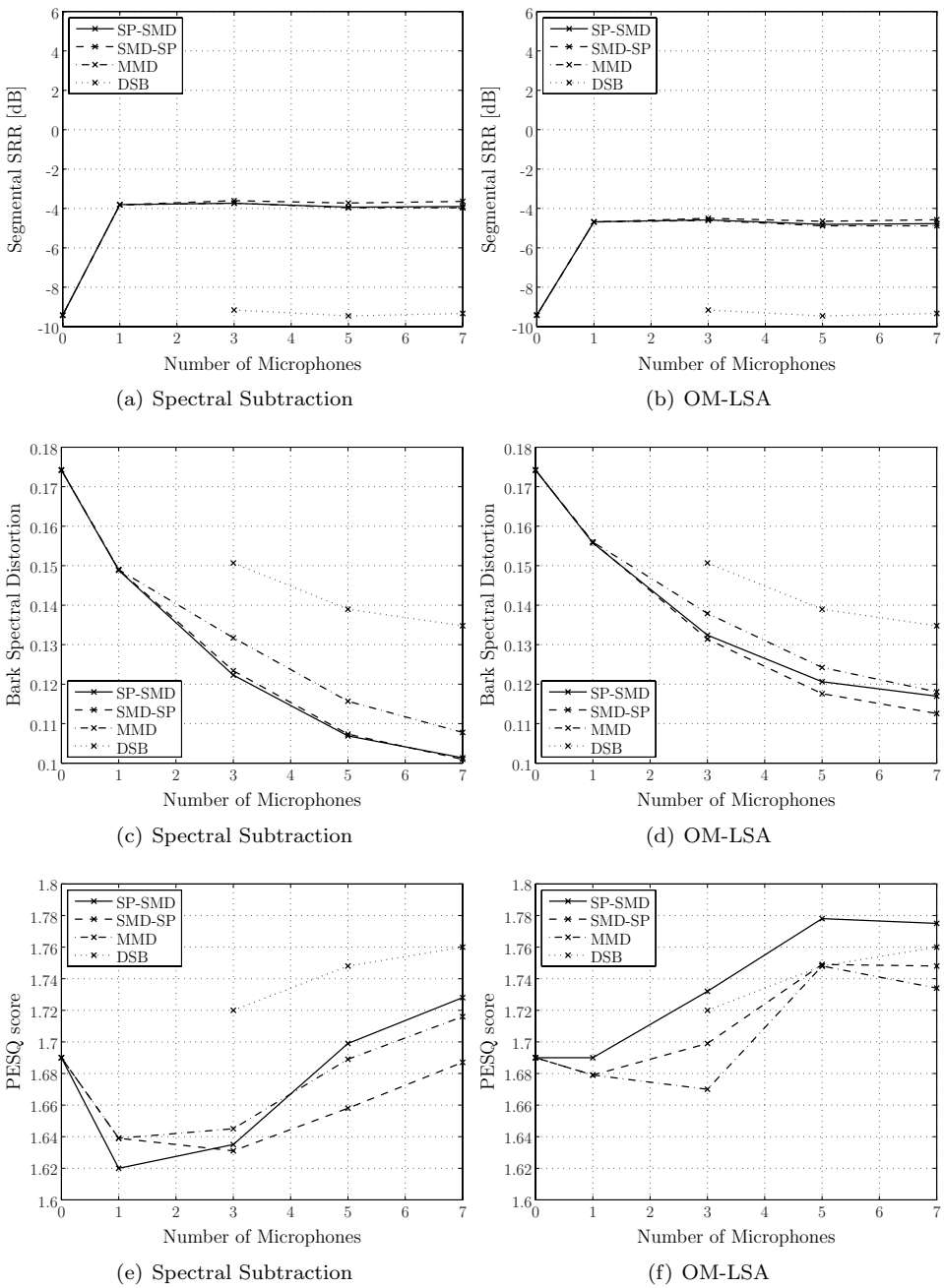


Figure 5.9 *Objective measures obtained for the four systems, i.e., the spatial processor with post-processor (SP-SMD), the pre-processors with spatial processor (SMD-SP), the joint multi-microphone dereverberation system (MMD), and the delay and sum beamformer (DSB). The post-processor and pre-processors are based on (a-c-d) spectral subtraction, and (b-d-f) the OM-LSA estimator. The reverberant signals were created using measured AIRs with a reverberation time of 529 ms. The source-array distance was 3 m and T_1 was set to 48 ms.*

5.5.2 Reverberation and Noise Suppression

In this section we have evaluated the performance of the developed single-microphone dereverberation techniques in the presence of noise.

We have first evaluated the performance using synthetic AIRs with $\mathrm{RT}_{60} = 0.5$ s. The segmental SNR of the microphone signal ranges from 10 dB till 30 dB. In Figs. 5.10 and 5.11 the segmental SIR and BSD are depicted for $D = 0.5$ m and $D = 3$ m, respectively. The segmental SIRs and BSDs were calculated using the reverberant microphone signal (unprocessed (UP)), the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). We can clearly see the increase in segmental SIR when joint reverberation and noise suppression is applied. Furthermore, the BSD obtained using the OM-LSA estimation technique is much lower than the BSD obtained using the spectral subtraction technique, while the segmental SIR is slightly lower.

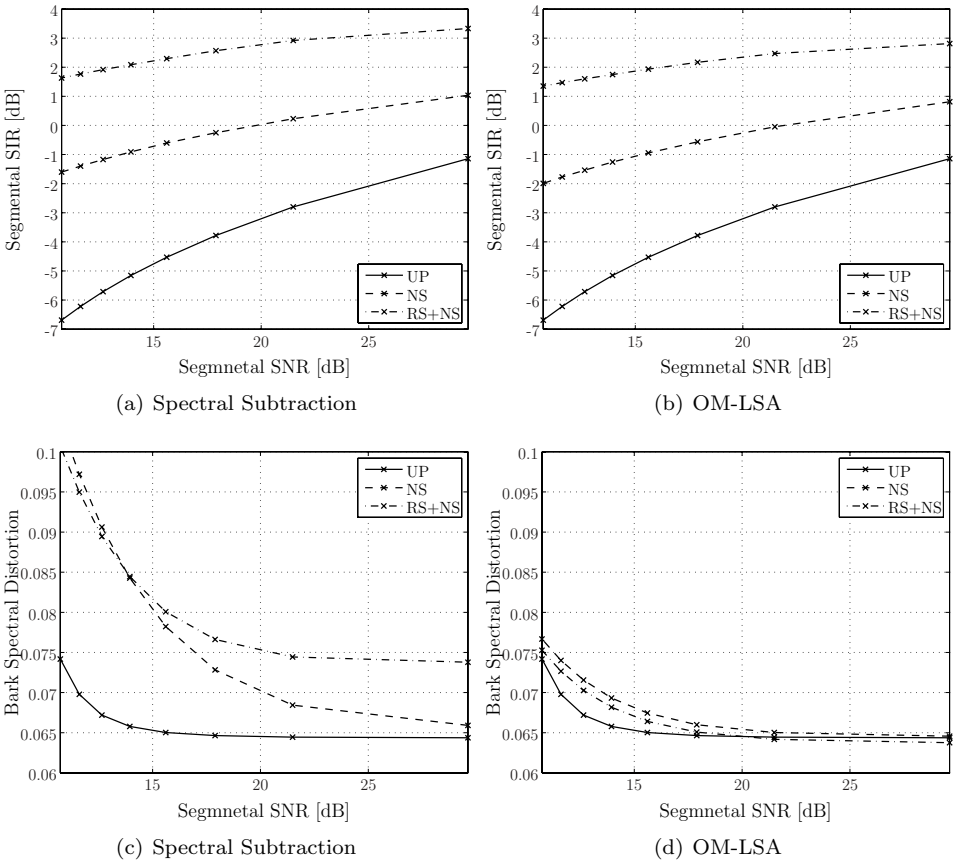
We now evaluate the performance using a real AIR that was measured in an office room ($\mathrm{RT}_{60} = 0.529$ s). In Figs. 5.12 and 5.13 the segmental SIRs and BSDs are shown for $D = 1$ m and $D = 3$ m, respectively. The BSD that was obtained by the OM-LSA estimation technique was much lower than the BSD that was obtained by the spectral subtraction technique. Again a significant increase in segmental SIR is achieved when joint reverberation and noise suppression is applied.

As an example a female speech signal corrupted by reverberation and noise (segmental SNR was 21 dB and 12.7 dB) was enhanced using the spectral subtraction technique and the OM-LSA estimation technique. The spectrogram and waveform of the direct signal, reverberant signal, early speech signal, microphone signal, and processed signals (spectral subtraction technique and OM-LSA estimation technique) are shown in Figs. 5.14 and 5.15. One can clearly see that signal that results from the spectral subtraction technique is more distorted than the signal that results from the OM-LSA estimation technique, e.g., at frequencies above 3 kHz the speech that results from the spectral subtraction technique is almost completely suppressed.

5.6 Conclusions

In this chapter we have described how late reverberation and background noise can be suppressed using single- and multi-microphones and a spectral enhancement technique. Two known spectral enhancement techniques, viz., spectral subtraction and the OM-LSA estimator have been used. Additional modifications of the spectral enhancement techniques were developed to enhance the performance when background noise and late reverberation are suppressed.

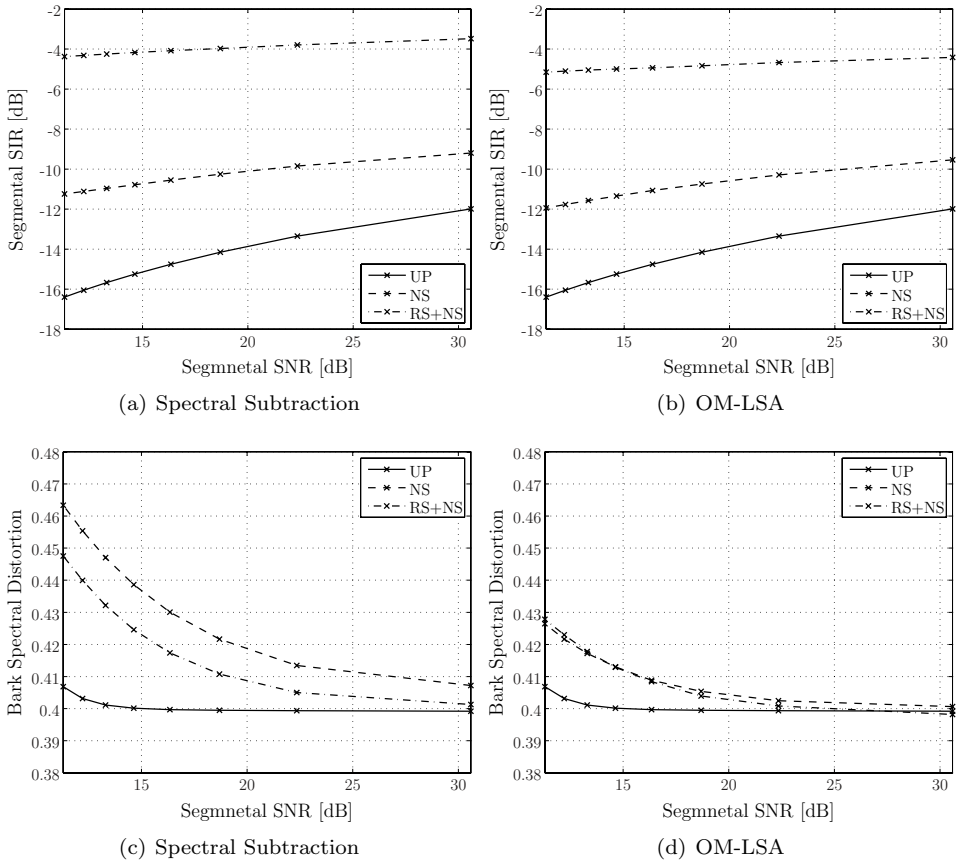
Three multi-microphone speech enhancement systems were proposed. The first system consists of a spatial processor and a post-processor, and has a very low computational complexity. The performance in terms of segmental SRR, BSD, and PESQ score is



(a) Spectral Subtraction (b) OM-LSA

(c) Spectral Subtraction (d) OM-LSA

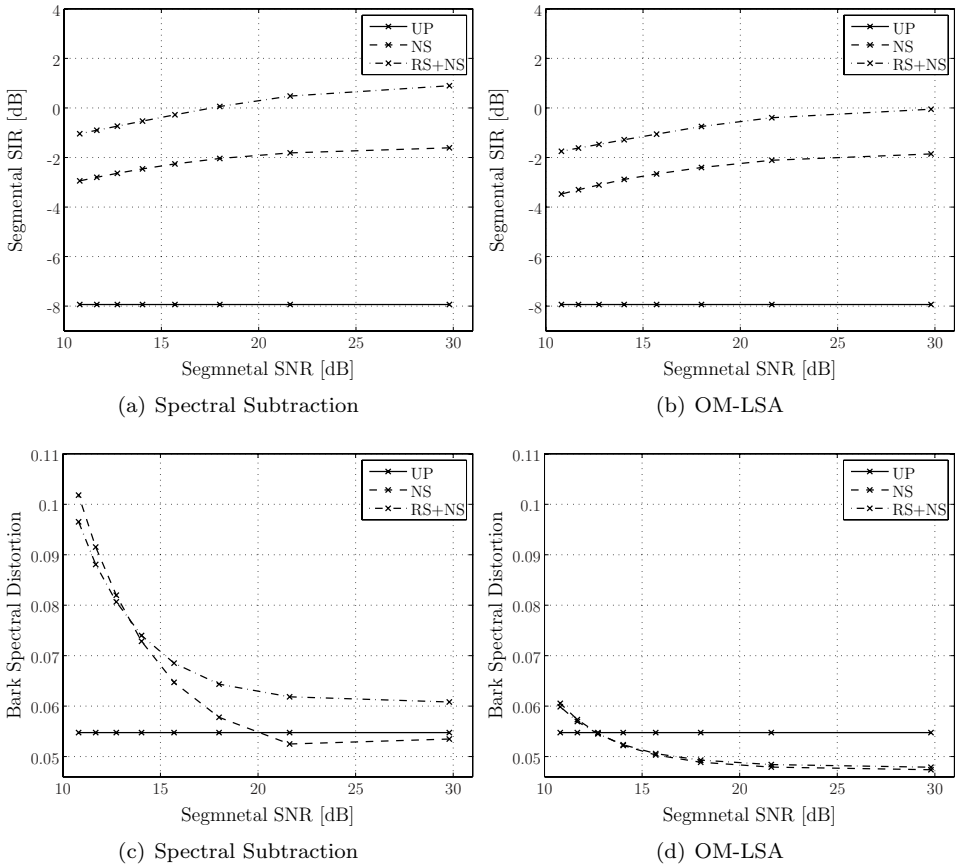
Figure 5.10 *The segmental SIRs and BSDs obtained from the unprocessed (UP) reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). The results are shown for (a-c) the spectral subtraction technique and (b-d) the OM-LSA estimation technique, using a synthetic AIR and one microphone. The source-microphone distance was 0.5 m, the reverberation time 500 ms, and T_1 was set to 48 ms.*



(a) Spectral Subtraction (b) OM-LSA

(c) Spectral Subtraction (d) OM-LSA

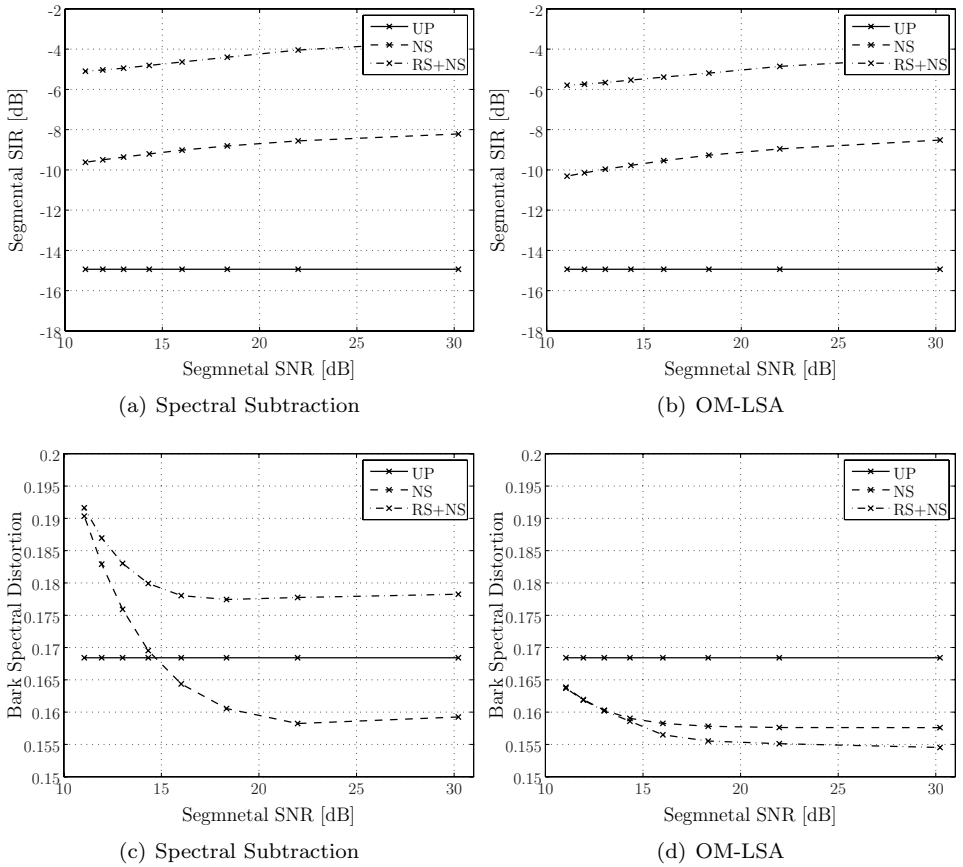
Figure 5.11 *The segmental SIRs and BSDs obtained from the unprocessed (UP) reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). The results are shown for (a-c) the spectral subtraction technique and (b-d) the OM-LSA estimation technique, using a synthetic AIR and one microphone. The source-microphone distance was 2 m, the reverberation time 500 ms, and T_1 was set to 48 ms.*



(a) Spectral Subtraction (b) OM-LSA

(c) Spectral Subtraction (d) OM-LSA

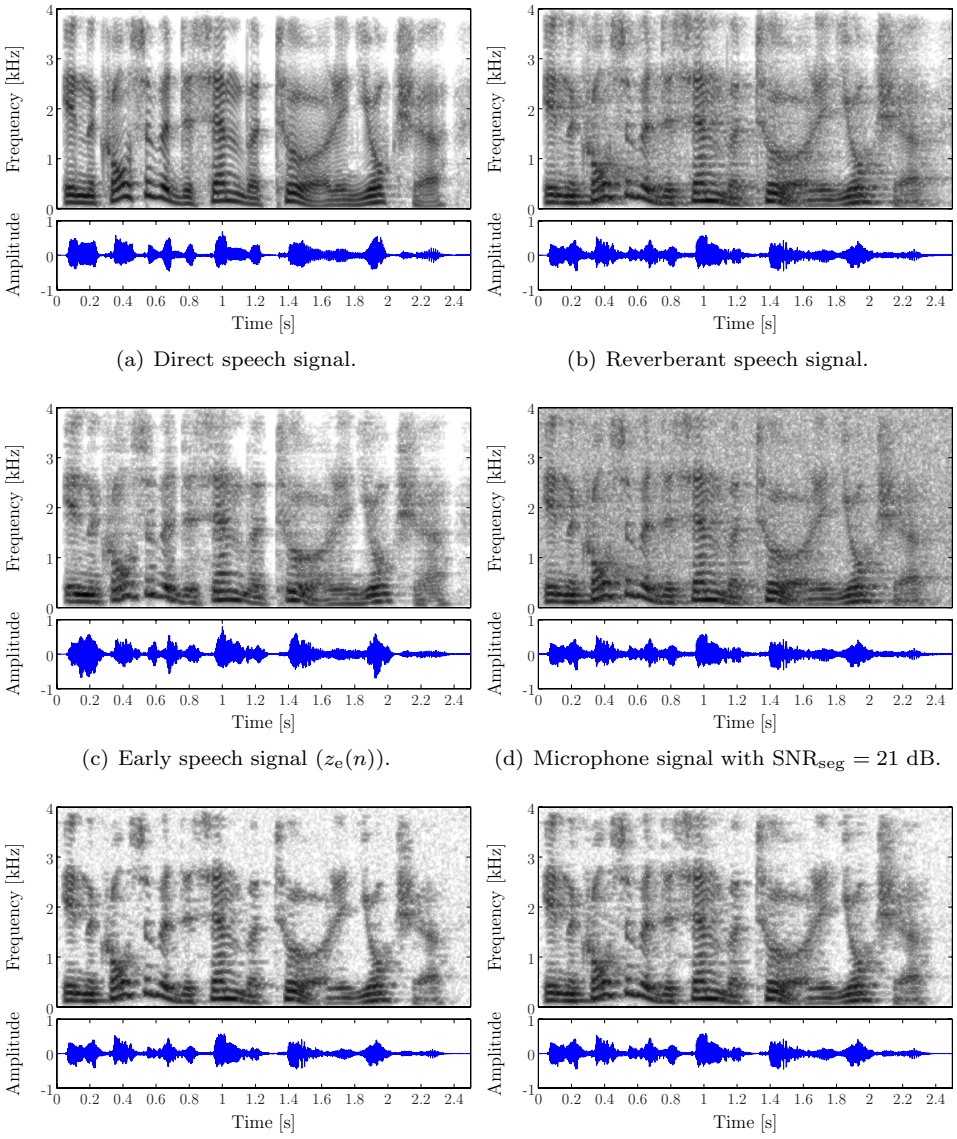
Figure 5.12 *The segmental SIRs and BSDs obtained from the unprocessed (UP) reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). The results are shown for (a-c) the spectral subtraction technique and (b-d) the OM-LSA estimation technique, using a measured AIR and one microphone. The source-microphone distance was 1 m, the reverberation time 529 ms, and T_1 was set to 48 ms.*



(a) Spectral Subtraction (b) OM-LSA

(c) Spectral Subtraction (d) OM-LSA

Figure 5.13 *The segmental SIRs and BSDs obtained from the unprocessed (UP) reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). The results are shown for (a-c) the spectral subtraction technique and (b-d) the OM-LSA estimation technique, using a measured AIR and one microphone. The source-microphone distance was 3 m, the reverberation time 529 ms, and T_1 was set to 48 ms.*



(a) Direct speech signal.

(b) Reverberant speech signal.

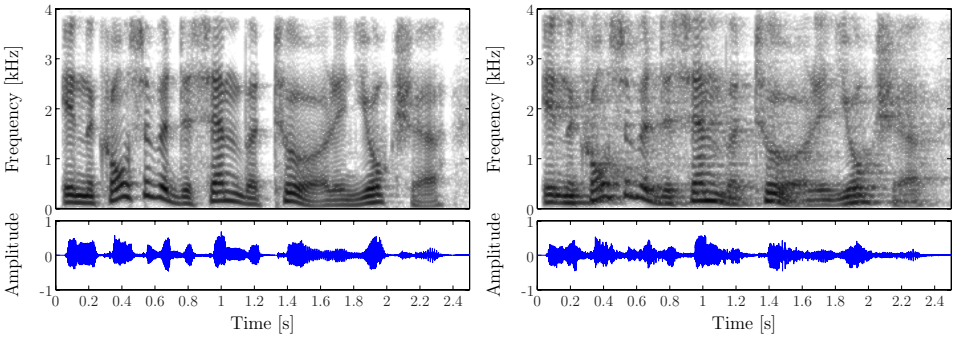
(c) Early speech signal ($z_{\mathrm{e}}(n)$).

(d) Microphone signal with $\mathrm{SNR}_{\mathrm{seg}} = 21$ dB.

(e) Spectral Subtraction

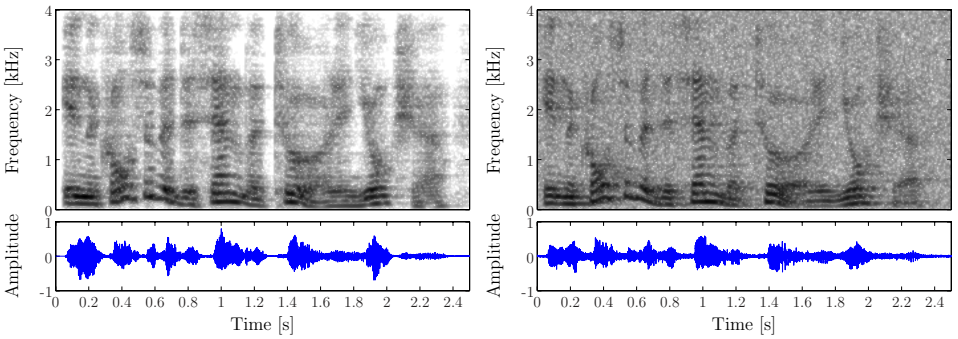
(f) OM-LSA

Figure 5.14 *Spectrograms and waveforms of the direct, reverberant, early, microphone (*$\mathrm{SNR}_{\mathrm{seg}} = 21$ *dB), and enhanced speech signal. The enhanced speech was obtained using joint reverberation and noise suppression using (c) the spectral subtraction technique, and (d) the OM-LSA estimation technique. A measured AIR with a reverberation time of 529 ms was used to generate the reverberant signal. The source-microphone distance was 1 m and* T_{l} *was set to 48 ms.*



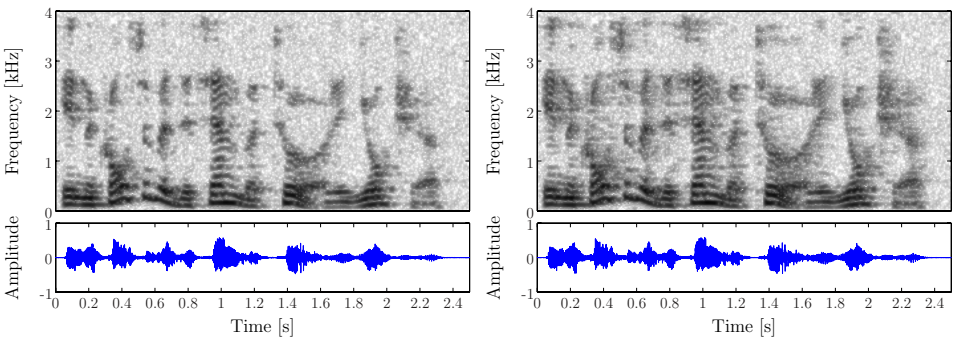
(a) Direct speech signal.

(b) Reverberant speech signal.



(c) Early speech signal.

(d) Microphone signal with $\mathrm{SNR}_{\mathrm{seg}} = 12.7$ dB.



(e) Spectral Subtraction

(f) OM-LSA

Figure 5.15 *Spectrograms and waveforms of the direct, reverberant, early, microphone (*$\mathrm{SNR}_{\mathrm{seg}} = 12.7$ *dB), and enhanced speech signal. The enhanced speech was obtained using joint reverberation and noise suppression using (c) the spectral subtraction technique, and (d) the OM-LSA estimation technique. A measured AIR with a reverberation time of 529 ms was used to generate the reverberant signal. The source-microphone distance was 1 m and* T_1 *was set to 48 ms.*

very good. However, in conjunction with the delay and sum beamformer this system introduces some audible distortions due to the spatial correlation between the acoustic channels. The second system consists of multiple pre-processors (one for each microphone) and a spatial processor, and has a much higher computational complexity than the first system. The overall performance of this system is very good. The third and final system uses a novel spatial processor. The output of this spatial processor is enhanced using a spatially averaged estimate of the late reverberant spectral variance. This system has a complexity that is larger than the first system but lower than the second system. The segmental SRRs and BSDs that were obtained using the third system are slightly worse compared to the second and first system. However, there are no audible distortions like in the first system. Furthermore, in [30] we evaluated three multi-microphone speech dereverberation techniques using subjective listening tests. The first dereverberation technique is a standard delay and sum beamfomer, the second technique is a Linear Prediction based enhancement technique proposed by Gaubitch et al. in [110], and the third is an early version of the spectral subtraction based dereverberation technique described in Chapter 5. The subjective listening test was performed according to the guidelines of International Telecommunications Union (ITU-T) Recommendation Series-P for subjective testing [178, 232]. Using the listening tests, we have estimated the subjective perception of colouration, reverberation decay tail effect, and the overall speech quality. A total of 26 normal hearing subjects was subjected to 64 speech files, with a male and a female talker for eight acoustic setups (different distances and reverberation times), and speech processed with the three dereverberation algorithms. Calibration speech examples were given to assist listeners in identifying colouration and reverberation decay tail effects. Compared to the received reverberant microphone signal the results for the spectral enhancement based technique indicated that the colouration was approximately equal, the reverberation decay tail effect was reduced, and the overall speech quality was improved. Based on the intelligibility tests performed by Allen [15] we expect that the intelligibility of the processed signal has been improved.

Results demonstrate that high reverberation suppression can be achieved with low speech distortion over a wide range of source-microphone distances and reverberation times.

Chapter 6

Late Reverberant Spectral Variance Estimation

6.1 Introduction

In speech communication systems, such as voice-controlled systems, hands-free mobile telephones, and hearing aids, the received microphone signals are degraded by room reverberation, background noise, and other interferences.

Reverberation is the process of multi-path propagation of an acoustic sound from its source to one or more microphones. The received signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*), and reflections that arrive after the early reflections (commonly called *late reverberation*). Reverberant speech can be described as sounding distant with noticeable echo and colouration. These detrimental perceptual effects generally increase with increasing distance between the source and microphone. From the discussion in Section 1.3 and 1.4 it has become clear that the late reverberation is the main cause of these detrimental effects. The combination of the direct sound and the early reverberation are sometimes referred to as the *early speech component*.

Reduction of the detrimental effects of reflections is evidently of considerable practical importance, and is the focus of this dissertation. In Chapter 5 dereverberation algorithms, i.e., signal processing algorithms that reduce the detrimental effects of reflections have been developed. The developed novel single- and multi-microphone speech dereverberation algorithms aim at the suppression of late reverberation, i.e., at estimation of the early speech component. This is done via so-called spectral enhancement techniques that require a specific measure of the late reverberant signal. This measure, called spectral variance, needs to be estimated directly from the received (possibly noisy) reverberant signal(s). In this chapter we derive such an estimator using a statistical reverberation model that requires a limited amount of *a priori*

knowledge about the acoustic channel(s) between the source and the microphone(s).

In this chapter an existing single-channel statistical reverberation model that was proposed by Polack [63], serves as a starting point. The model is characterized by a parameter that depends on the reverberation time RT_{60} of the environment. Detailed information about the reverberation time can be found in Section 2.8. In this chapter it is shown that the spectral variance estimator that is based on Polack's model, can only be used when the source-microphone distance is larger than the critical distance. This is the distance where the direct path energy is equal to the total reflective energy, i.e., early plus late reflective energy. A generalization of the statistical reverberation model in which the direct sound is incorporated is developed. This model requires one additional parameter, denoted by κ, which is related to the ratio between the direct path energy and the sound energy of all reflections. The generalized model is used to derive a novel spectral variance estimator, for the so-called late reverberant spectral variance. When the novel estimator rather than the existing estimator is used for dereverberation and the source-microphone distance is smaller than the critical distance, the dereverberation performance is significantly increased. It will be shown that an estimate of the parameter κ can be obtained blindly. Under the assumption that the enclosed space is ergodic, which means that the resulting sound field is diffuse, different realizations of this stochastic reverberation process can be obtained by varying either the position of the microphone or the position of the source. Therefore, we can replace the ensemble average that is used in the derivation of the spectral variance estimator by spatial averaging. Since spatial averaging can be performed by using multiple microphones, we can use them to improve the late reverberant spectral variance estimation. Experimental results demonstrate the performance of the spectral variance estimator. Furthermore, we analyse how sensitive the estimators are with respect to errors in the two parameters RT_{60} and κ.

The structure of this chapter is as follows. In Section 6.2 the problem is formulated. In Section 6.3 we will first discuss Polack's statistical model, and introduce the generalized statistical model. Both models are used to derive a late reverberant spectral variance estimator in Section 6.4. In Section 6.5 it will be shown how the late reverberant spectral variance can be estimated from a noisy observation using additional processing. A more elaborate discussion on the (blind) estimation of the required parameters RT_{60} and κ can be found in Section 6.6. Simulation results will be shown in Section 6.7, and demonstrate the feasibility of the two estimators and the sensitivity with respect to estimation errors of the required parameters.

6.2 Problem Formulation

The reverberant signal results from the convolution of the anechoic speech signal $s(t)$ and the causal time-invariant Acoustic Impulse Response (AIR). The AIR function is

denoted by $h(t)$, such that

$$z(t) = \int_{-\infty}^{t} s(\theta)h(t-\theta)\,\mathrm{d}\theta. \tag{6.1}$$

Since our goal is to suppress late reverberation the AIR is divided into two segments, $h_\mathrm{e}(t)$ and $h_\mathrm{l}(t)$ so that

$$h(t) = \begin{cases} h_\mathrm{e}(t), & \text{for } 0 \leq t < T_\mathrm{l}; \\ h_\mathrm{l}(t), & \text{for } t \geq T_\mathrm{l}; \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

The parameter T_l can be controlled depending on the application or subjective preference. Usually T_l is chosen such that $h_\mathrm{e}(t)$ consists of the direct path and a few early reflections and $h_\mathrm{l}(t)$ consists of all later reflections, and hence ranges from 40 to 80 ms.

Eq. 6.1 can be rewritten using Eq. 6.2:

$$z(t) = \underbrace{\int_{-\infty}^{t} s(\theta)h_\mathrm{e}(t-\theta)\,\mathrm{d}\theta}_{z_\mathrm{e}(t)} + \underbrace{\int_{-\infty}^{t} s(\theta)h_\mathrm{l}(t-\theta)\,\mathrm{d}\theta}_{z_\mathrm{l}(t)}. \tag{6.3}$$

Our main purpose is to obtain an estimate of the late reverberant spectral variance $\lambda_{z_\mathrm{l}}(l,k)$, which is defined as

$$\lambda_{z_\mathrm{l}}(l,k) = \mathcal{E}\{|Z_\mathrm{l}(l,k)|^2\}, \tag{6.4}$$

where $\mathcal{E}\{\cdot\}$ denotes mathematical expectation, and $Z_\mathrm{l}(l,k)$ denotes the discrete short-time Fourier transform (STFT) of $z_\mathrm{l}(t)$.

6.3 Statistical Reverberation Models

In this section Polack's statistical reverberation model is discussed, and a generalized statistical reverberation model is proposed. It will be show that Polack's statistical model is closely related to the energy balance equation which was described in Section 2.7.2.

Since the acoustic behaviour in real rooms is too complex to model explicitly, Statistical Room Acoustics (SRA) is often used. SRA provides a statistical description of the transfer function that describes the system between the source and microphone in terms of a few key quantities, e.g., source-microphone distance, room volume, and reverberation time. The crucial assumption of SRA is that the distribution of amplitudes and phases of individual plane waves, which sum up to produce sound pressure at some point in a room, is so close to random that the sound field is fairly uniformly

distributed throughout the room volume. A more elaborate discussion about SRA can be found in Section 2.6.

The validity of this model is subject to the following conditions that are assumed to hold: [41, 2, 66]

i) The dimensions of the room are relatively large compared to the wavelength.

ii) The average spacing of the resonance frequencies (commonly called modes) of the room must be smaller than one third of their bandwidth. In a room with volume V (in m^3), and reverberation time RT_{60} (in seconds), which is defined as the time for the reverberation level to decay to 60 dB below the initial level, this condition is fulfilled for frequencies that exceed the Schroeder frequency: $f_{\mathrm{g}} = 2000\sqrt{\mathrm{RT}_{60}/V}$.

iii) The source and the microphones are located in the interior of the room, at least a half-wavelength away from the walls.

6.3.1 Existing Statistical Reverberation Models

Moorer [70] noted the auditive resemblance between a concert hall impulse response and a white noise signal multiplied by an exponentially decaying envelope, and reported that such a synthetic response can produce a natural sounding reverberation effect (by convolution with anechoic signals).

Polack [63] developed a time-domain model complementing Schroeder's frequency domain model. Polack modelled the AIR as a realization of a non-stationary stochastic process, according to

$$h(t) = \begin{cases} b(t)\mathrm{e}^{-\bar{\delta}t}, & \text{for } t \geq 0; \\ 0, & \text{otherwise}, \end{cases} \tag{6.5}$$

where $b(t)$ is a white zero-mean Gaussian stationary noise signal, and the average damping constant $\bar{\delta}$ is linked to the reverberation time RT_{60} through

$$\bar{\delta} = \frac{3\log_{\mathrm{e}}(10)}{\mathrm{RT}_{60}}. \tag{2.37}$$

In contrast to Polack's statistical model in Eq. 6.5 the average damping constant $\bar{\delta}$ is frequency dependent due to the frequency dependent reflection coefficients of walls and other objects, and the frequency dependent absorption coefficient of air [41]. We will take this dependency into account by using a different model in each sub-band k, i.e.,

$$h(t,k) = \begin{cases} b(t,k)\mathrm{e}^{-\bar{\delta}(k)t}, & \text{for } t \geq 0; \\ 0, & \text{otherwise}, \end{cases} \tag{6.6}$$

where $b(t,k)$ is a white zero–mean Gaussian stationary noise signal, and the average damping constant $\bar{\delta}(k)$ in the k^{th} sub-band is linked to the frequency dependent reverberation time $\text{RT}_{60}(k)$ through

$$\bar{\delta}(k) = \frac{3\log_{\text{e}}(10)}{\text{RT}_{60}(k)}. \tag{6.7}$$

For the derivation of the late reverberant spectral variance estimator the full-band model in Eq. 6.5 will be used. Since the sub-bands are assumed to be statistically independent, and the estimation will be performed for each sub-band k, the incorporation of $\bar{\delta}(k)$ will be straightforward.

In the early nineties, Polack [72] concluded that the reverberation process is statistical when the number of simultaneous echo arrivals reaches a limit of about 10. In this case the echo density is high enough, such that the space can be considered to be in a fully diffused or mixed state. The essential requirement is ergodicity, which requires that any given echo trajectory in the space will eventually reach all points. The ergodicity is determined by the shape of the enclosure and the surface reflection properties. It should be noted that non-ergodic shapes will exhibit much longer mixing times and may not even have an exponential decay. Nevertheless, while it may not be true that all acoustic environments can be modelled using this stochastic model, it is sufficiently accurate for most spaces. Kuttruff [41] concluded that the model is valid in case the distance between the source and the measurement point is greater than the critical distance. The critical distance indicates the distance at which the direct path energy equals the energy of the early and late reflections, i.e., the Direct to Reverberation Ratio (DRR) equals 0 dB. In Section 2.7.1 we have shown that for an omnidirectional source the critical distance is defined as

$$D_{\text{h}} = 0.1\sqrt{\frac{V}{\pi \text{RT}_{60}}} \quad [\text{m}]. \tag{2.64}$$

This implies that Polack's statistical model is only valid when the DRR is less than 0 dB, i.e., the source-microphone distance needs to be larger than the critical distance.

The energy envelope of the AIR can be expressed as

$$\mathcal{E}_h\{h^2(t)\} = \sigma^2 \text{e}^{-2\bar{\delta}t}, \tag{6.8}$$

where σ^2 denotes the variance of $b(t)$ and $\mathcal{E}_h\{\cdot\}$ denotes ensemble averaging over h, i.e., over different realizations of the stochastic process in Eq. 6.5. Under the assumption that our space is ergodic we may evaluate the ensemble average in Eq. 6.8 by spatial averaging so that different realizations of this stochastic process are obtained by varying either the position of the receiver or the source [233]. Note that the same stochastic process will be observed, for all allowable positions (in terms of the third SRA condition), provided that the time origin be defined with respect to the signal emitted by the source, and not with respect to the arrival time of the direct sound at the receiver.

6.3.2 Generalized Statistical Reverberation Model

Polack's statistical model has been found extremely helpful to derive an estimator for the late reverberant spectral variance in case the DRR is smaller than 0 dB [24, 33], i.e., in case the source-microphone distance is larger than the critical distance. In case the DRR is larger than 0 dB, which indicates that the source-microphone distance is smaller than the critical distance, the late reverberant spectral variance is overestimated, which will result in severe spectral distortion of the dereverberated signal. In this section a generalized statistical reverberation model is proposed which is used to derive a novel spectral variance estimator. Furthermore, like Polack's statistical model the proposed statistical model could also be used to create artificial reverberation.

The AIR $h(t)$, can be split into two segments, $h_{\mathrm{d}}(t)$ and $h_{\mathrm{r}}(t)$:

$$h(t) = \begin{cases} h_{\mathrm{d}}(t), & \text{for } 0 \leq t < T_{\mathrm{r}}; \\ h_{\mathrm{r}}(t), & \text{for } t \geq T_{\mathrm{r}}; \\ 0, & \text{otherwise.} \end{cases} \tag{6.9}$$

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

$$h_{\mathrm{d}}(t) = \begin{cases} b_{\mathrm{d}}(t)\mathrm{e}^{-\bar{\delta}t}, & \text{for } 0 \leq t < T_{\mathrm{r}}; \\ 0, & \text{otherwise,} \end{cases} \tag{6.10}$$

where $b_{\mathrm{d}}(t)$ is a white zero–mean Gaussian stationary noise signal and $\bar{\delta}$ is linked to the reverberation time RT_{60} through Eq. 2.37. The reverberant component $h_{\mathrm{r}}(t)$ is described using the following model:

$$h_{\mathrm{r}}(t) = \begin{cases} b_{\mathrm{r}}(t)\mathrm{e}^{-\bar{\delta}t}, & \text{for } t \geq T_{\mathrm{r}}; \\ 0, & \text{otherwise,} \end{cases} \tag{6.11}$$

where $b_{\mathrm{r}}(t)$ is a white zero–mean Gaussian stationary noise signal. Under the SRA conditions the direct and reverberant component of the AIR are uncorrelated [2]. Therefore, it is reasonable to assume that $b_{\mathrm{d}}(t)$ and $b_{\mathrm{r}}(t)$ are uncorrelated, i.e., $\mathcal{E}\{b_{\mathrm{d}}(t)b_{\mathrm{r}}(t+\tau)\} = 0$.

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

$$\mathcal{E}_h\{h^2(t)\} = \begin{cases} \sigma_{\mathrm{d}}^2\mathrm{e}^{-2\bar{\delta}t}, & \text{for } 0 \leq t < T_{\mathrm{r}}; \\ \sigma_{\mathrm{r}}^2\mathrm{e}^{-2\bar{\delta}t}, & \text{for } t \geq T_{\mathrm{r}}; \\ 0, & \text{otherwise,} \end{cases} \tag{6.12}$$

where σ_d^2 and σ_r^2 denote the variance of $b_\mathrm{d}(t)$ and $b_\mathrm{r}(t)$, respectively. When $\sigma_\mathrm{d}^2 < \sigma_\mathrm{r}^2$, the contribution of the direct path can be ignored. Therefore, it is assumed that $\sigma_\mathrm{d}^2 \geq \sigma_\mathrm{r}^2$.

Note that the proposed model is equivalent to Polack's statistical model in case $\sigma_\mathrm{d}^2 = \sigma_\mathrm{r}^2$.

6.3.3 Relation with Energy Balance Equation

As discussed in Section 6.3.1 Polack's statistical model is only valid when the source-microphone distance is larger than the critical distance. We will now prove that the reverberant energy density that can be predicted by Polack's model is related to the fundamental differential equation governing the growth of the reverberant energy density in a room, which was discussed in Section 2.7.2 and is also known as the energy balance equation. This relation also proves that the influence of the direct path energy has not been taken into account by Polack.

The energy balance equation is given by

$$\frac{4W_\mathrm{s}(t)}{cA} = \tau \frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t} + E_\mathrm{r}(t), \qquad [2.69]$$

where W_s denotes the power of the source in Watt, $E_\mathrm{r}(t)$ the reverberant energy density in J/m^3, τ the decay constant, c the sound velocity in m/s, and A the equivalent absorption area of the room. Under the assumption that $W_\mathrm{s}(t) = 0$ for $t \leq 0$, the general solution for this differential equation is

$$E_\mathrm{r}(t) = \frac{1}{\tau}\mathrm{e}^{-t/\tau} \int_0^t \mathrm{e}^{\theta/\tau} W(\theta)\, \mathrm{d}\theta, \qquad (6.13)$$

with $W(t) = \frac{4W_\mathrm{s}(t)}{cA}$.

Proof of Eq. 6.13. First $\frac{4W_\mathrm{s}(t)}{cA}$ in Eq. 2.69 is replaced by $W(t)$ and both sides are divide by τ:

$$\frac{1}{\tau}W(t) = \frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t} + \frac{1}{\tau}E_\mathrm{r}(t). \qquad (6.14)$$

Secondly, using the integrating factor $\mathrm{e}^{t/\tau}$ and the fact that

$$\frac{\mathrm{d}\left(\mathrm{e}^{t/\tau}E_\mathrm{r}(t)\right)}{\mathrm{d}t} = \mathrm{e}^{t/\tau}\left(\frac{\mathrm{d}E_\mathrm{r}(t)}{\mathrm{d}t} + \frac{1}{\tau}E_\mathrm{r}(t)\right) \qquad (6.15)$$

Eq. 6.14 can be written as

$$\frac{\mathrm{d}\left(\mathrm{e}^{t/\tau}E_\mathrm{r}(t)\right)}{\mathrm{d}t} = \frac{1}{\tau}\mathrm{e}^{t/\tau}W(t). \qquad (6.16)$$

Now the left and right hand side can be integrated, such that

$$\begin{aligned} \mathrm{e}^{t/\tau} E_{\mathrm{r}}(t) &= \frac{1}{\tau} \int_0^t \mathrm{e}^{\theta/\tau} W(\theta)\, \mathrm{d}\theta + C \\ E_{\mathrm{r}}(t) &= \mathrm{e}^{-t/\tau} \left(\frac{1}{\tau} \int_0^t \mathrm{e}^{\theta/\tau} W(\theta)\, \mathrm{d}\theta + C \right). \end{aligned} \tag{6.17}$$

□

Let us consider a zero-mean source signal $s(t)$, with variance $\sigma_s^2(t) = \mathcal{E}\left\{s^2(t)\right\}$. The variance of the reverberant signal is defined as $\sigma_z^2(t) = \mathcal{E}\left\{z^2(t)\right\}$, using Eq. 6.1 and 6.5 it follows that

$$\begin{aligned} \sigma_z^2(t) &= \mathcal{E}\left\{ \int_0^t s(\theta) h(t-\theta)\, \mathrm{d}\theta \int_0^t s(\theta') h(t-\theta')\, \mathrm{d}\theta' \right\} \\ &= \int_0^t \mathcal{E}\left\{s^2(\theta)\right\} \mathcal{E}\left\{h^2(t-\theta)\right\}\, \mathrm{d}\theta \\ &= \sigma^2 \mathrm{e}^{-2\bar{\delta}t} \int_0^t \mathrm{e}^{2\bar{\delta}\theta} \sigma_s^2(\theta)\, \mathrm{d}\theta. \end{aligned} \tag{6.18}$$

Since the time-constant τ is related to the average damping constant by $1/\tau = 2\bar{\delta}$ (c.f. [41]) Eq. 6.18 can be written as

$$\sigma_z^2(t) = \sigma^2 \mathrm{e}^{-t/\tau} \int_0^t \mathrm{e}^{\theta/\tau} \sigma_s^2(\theta)\, \mathrm{d}\theta. \tag{6.19}$$

For $\sigma_s^2(\theta) = W(\theta)$ it is clear that $\sigma_z^2(t)$ is proportional to $E_{\mathrm{r}}(t)$ given by Eq. 6.13.

6.4 Late Reverberant Spectral Variance Estimator

In this section two estimators for the late reverberant spectral variance are derived using Polack's statistical model and the proposed generalized statistical model.

6.4.1 Estimator based on Polack's Statistical Model

The auto-correlation of the reverberant signal z at time t and lag τ for a fixed source-microphone configuration is defined as

$$r_{zz}(t, t+\tau) = \mathcal{E}_z\{z(t)z(t+\tau)\}, \tag{6.20}$$

where $\mathcal{E}_z\{\cdot\}$ denotes ensemble averaging with respect to z. For one realization of h we have

$$r_{zz}(t, t+\tau; h) = \int_{-\infty}^{t} \int_{-\infty}^{t+\tau} \mathcal{E}_s\{s(\theta)s(\theta')\} h(t-\theta) h(t+\tau-\theta')\, \mathrm{d}\theta \mathrm{d}\theta', \tag{6.21}$$

where $\mathcal{E}_s\{\cdot\}$ denotes ensemble averaging with respect to s. Since there is no physical relation between the stochastic processes h and s, these processes can be assumed to be statistical independent. The spatially averaged auto-correlation is

$$\begin{aligned} r_{zz}(t, t+\tau) &= \mathcal{E}_h\{r_{zz}(t, t+\tau; h)\} \\ &= \int_{-\infty}^{t} \int_{-\infty}^{t+\tau} \mathcal{E}_s\{s(\theta)s(\theta')\}\mathcal{E}_h\{h(t-\theta)h(t+\tau-\theta')\}\, \mathrm{d}\theta \mathrm{d}\theta'. \end{aligned} \tag{6.22}$$

Using Eq. 6.5 and the fact that $b(t)$ consists of a zero-mean white Gaussian noise signal, it follows that

$$\mathcal{E}_h\{h(t-\theta)h(t+\tau-\theta')\} = \sigma^2 \mathrm{e}^{-2\bar{\delta}t} \mathrm{e}^{\bar{\delta}(\theta+\theta'-\tau)} \delta(\theta - \theta' + \tau), \tag{6.23}$$

where $\delta(\cdot)$ denotes the Dirac function. Accordingly,

$$\begin{aligned} r_{zz}(t, t+\tau) &= \mathrm{e}^{-2\bar{\delta}t} \int_{-\infty}^{t} \mathcal{E}_s\{s(\theta)s(\theta+\tau)\}\, \sigma^2 \mathrm{e}^{2\bar{\delta}\theta}\, \mathrm{d}\theta \\ &= \mathrm{e}^{-2\bar{\delta}t} \int_{t-T_\mathrm{l}}^{t} \mathcal{E}_s\{s(\theta)s(\theta+\tau)\}\, \sigma^2 \mathrm{e}^{2\bar{\delta}\theta}\, \mathrm{d}\theta \\ &\quad + \mathrm{e}^{-2\bar{\delta}t} \int_{-\infty}^{t-T_\mathrm{l}} \mathcal{E}_s\{s(\theta)s(\theta+\tau)\}\, \sigma^2 \mathrm{e}^{2\bar{\delta}\theta}\, \mathrm{d}\theta. \end{aligned} \tag{6.24}$$

The auto-correlation at time t can be divided into two terms, as shown in Eq. 6.24. The first term depends on the direct signal between time $t - T_\mathrm{l}$ and t, whereas the second depends on the late reverberant signal and is responsible for overlap-masking. Let us consider the spatially averaged auto-correlation at time $t - T_\mathrm{l}$

$$r_{zz}(t - T_\mathrm{l}, t - T_\mathrm{l} + \tau) = \mathrm{e}^{-2\bar{\delta}(t-T_\mathrm{l})} \int_{-\infty}^{t-T_\mathrm{l}} \mathcal{E}_s\{s(\theta)s(\theta+\tau)\}\, \sigma^2 \mathrm{e}^{2\bar{\delta}\theta}\, \mathrm{d}\theta. \tag{6.25}$$

Now the auto-correlation at time t can be expressed as

$$r_{zz}(t, t+\tau) = r_{z_\mathrm{e} z_\mathrm{e}}(t, t+\tau) + r_{z_\mathrm{l} z_\mathrm{l}}(t, t+\tau), \tag{6.26}$$

with

$$r_{z_\mathrm{e} z_\mathrm{e}}(t, t+\tau) = \mathrm{e}^{-2\bar{\delta}t} \int_{t-T_\mathrm{l}}^{t} \mathcal{E}_s\{s(\theta)s(\theta+\tau)\}\, \sigma^2 \mathrm{e}^{2\bar{\delta}\theta}\, \mathrm{d}\theta, \tag{6.27}$$

$$r_{z_\mathrm{l} z_\mathrm{l}}(t, t+\tau) = \mathrm{e}^{-2\bar{\delta}T_\mathrm{l}} r_{zz}(t - T_\mathrm{l}, t - T_\mathrm{l} + \tau). \tag{6.28}$$

In practice the signals can be considered as stationary over periods of time that are short compared to the reverberation time RT_{60}. This is justified by the fact that the exponential decay is very slow, and that speech is quasi-stationary. Let T_s be the time span over which the speech signal can be considered stationary, which is usually

around 20-40 ms [99]. Under the assumption that $T_\mathrm{s} \leq T_\mathrm{l} \ll \mathrm{RT}_{60}$, the counterparts of Eq. 6.26 and Eq. 6.28 in terms of the short-term power spectral densities are:

$$P_{zz}(t,f) = P_{z_\mathrm{e}z_\mathrm{e}}(t,f) + P_{z_\mathrm{l}z_\mathrm{l}}(t,f), \tag{6.29}$$

$$P_{z_\mathrm{l}z_\mathrm{l}}(t,f) = \mathrm{e}^{-2\bar{\delta}T_\mathrm{l}} P_{zz}(t-T_\mathrm{l},f). \tag{6.30}$$

In the STFT domain we then have:

$$\lambda_z(l,k) = \lambda_{z_\mathrm{e}}(l,k) + \lambda_{z_\mathrm{l}}(l,k), \tag{6.31}$$

$$\lambda_{z_\mathrm{l}}(l,k) = \mathrm{e}^{-2\bar{\delta}T_\mathrm{l}} \lambda_z(l-N_\mathrm{l},k), \tag{6.32}$$

where $N_\mathrm{l} = \frac{T_\mathrm{l} f_\mathrm{s}}{R}$, R denotes the frame rate in samples of the STFT, and f_s the sampling frequency.

The frequency dependency of the average damping constant $\bar{\delta}$ can now be introduced by replacing $\bar{\delta}$ in Eq. 6.32 by $\bar{\delta}(k)$, i.e.,

$$\lambda_{z_\mathrm{l}}(l,k) = \mathrm{e}^{-2\bar{\delta}(k)T_\mathrm{l}} \lambda_z(l-N_\mathrm{l},k). \tag{6.33}$$

The spectral variance of the received microphone signal $\lambda_z(l,k)$ is estimated by using an exponentially weighted moving-average filter, also known as a first-order low-pass filter, with filter-constant $\eta_z^\mathrm{d}(k)$ $(0 \leq \eta_z^\mathrm{d}(k) < 1)$, i.e.,

$$\hat{\lambda}_z(l,k) = \eta_z^\mathrm{d}(k)\hat{\lambda}_z(l-1,k) + (1-\eta_z^\mathrm{d}(k))|Z(l,k)|^2. \tag{6.34}$$

The low-pass filtering is required to get a smoother estimate of the spectral variance, which can be interpreted as an estimate of energy in the room measured at the corresponding microphone position. If the filter-constant is too large, such that too much smoothing is applied, the spectral variance is over-estimated during the free-decay, which occurs when the source has become silent. It can be shown that the filter-constant $\eta_z^\mathrm{d}(k)$ is related to the time-constant $\tau_z(k)$ of the filter by:

$$\eta_z^\mathrm{d}(k) = \frac{\tau_z(k)}{\tau_z(k) + \frac{R}{f_\mathrm{s}}}. \tag{6.35}$$

Let us assume that the reverberation time $\mathrm{RT}_{60}(k)$ of the room has been estimated, then the average damping constant $\bar{\delta}(k)$ can be calculated using Eq. 6.7. The corresponding decay constant $\tau(k)$ of the room is then given by $1/2\bar{\delta}(k)$ (see Chapter 2). To ensure that no over-estimation can occur we require that $\tau_z(k) \leq \tau(k)$. The upper-bound for the filter-constant $\eta_z^\mathrm{d}(k)$ is thus related to the average damping constant $\bar{\delta}(k)$ given by:

$$\eta_z^\mathrm{d}(k) = \frac{\frac{1}{2\bar{\delta}(k)}}{\frac{1}{2\bar{\delta}(k)} + \frac{R}{f_\mathrm{s}}}. \tag{6.36}$$

Simulation results and informal listening tests have suggested that in practice $\eta_z^\mathrm{d}(k)$ should be chosen slightly higher than the upper-bound. Apparently a slight over-estimation, and hence suppression, yields better results than a small amount of residual

late reverberation. Unfortunately the proposed filter-constant is often too high to accurately track onsets in the reverberant speech signal. Since the filter-constant value cannot be lowered we propose to use two filter-constant values, such that the tracking abilities are improved. In case the received power spectrum $|Z(l,k)|^2$ is smaller than, or equal to, $\hat{\lambda}_z(l-1,k)$ the value $\eta_z^{\mathrm{d}}(k)$ is used. In any other case the filter-constant $\eta_z^{\mathrm{a}}(k)$ $(0 \leq \eta_z^{\mathrm{a}}(k) < \eta_z^{\mathrm{d}}(k))$ is chosen. The filter-constant is now signal dependent, i.e.,

$$\eta_z(l,k) = \begin{cases} \eta_z^{\mathrm{d}}(k), & \text{for } |Z(l,k)|^2 \leq \hat{\lambda}_z(l-1,k); \\ \eta_z^{\mathrm{a}}(k), & \text{otherwise.} \end{cases} \tag{6.37}$$

The spectral variance $\hat{\lambda}_z(l,k)$ is then calculated using $\eta_z(l,k)$:

$$\hat{\lambda}_z(l,k) = \eta_z(l,k)\hat{\lambda}_z(l-1,k) + (1-\eta_z(l,k))|Z(l,k)|^2. \tag{6.38}$$

Finally the late reverberant spectral variance $\lambda_{z_{\mathrm{l}}}(l,k)$ can be estimated by substituting $\hat{\lambda}_z(l,k)$, which is given by Eq. 6.38, in Eq. 6.33.

6.4.2 Estimator based on the Generalized Statistical Model

Using Eq. 6.9 the received signal $z(t)$ is expressed as

$$\begin{aligned} z(t) &= \int_{-\infty}^{t} s(\theta)h(t-\theta)\,d\theta \\ &= \int_{t-T_{\mathrm{r}}}^{t} s(\theta)h_{\mathrm{d}}(t-\theta)\,d\theta + \int_{-\infty}^{t-T_{\mathrm{r}}} s(\theta)h_{\mathrm{r}}(t-\theta)\,d\theta. \end{aligned} \tag{6.39}$$

The auto-correlation $r_{zz}(t,t+\tau) = \mathcal{E}_z\{z(t)z(t+\tau)\}$ of the reverberant signal z at time t and lag τ for a fixed source-microphone configuration, i.e., one realization of h, is

$$\begin{aligned} r_{zz}(t,t+\tau;h) = &\int_{t-T_{\mathrm{r}}}^{t}\int_{t-T_{\mathrm{r}}+\tau}^{t+\tau} \mathcal{E}_s\{s(\theta)s(\theta')\}h_{\mathrm{d}}(t-\theta)h_{\mathrm{d}}(t+\tau-\theta')\,d\theta d\theta' \\ &+ \int_{-\infty}^{t-T_{\mathrm{r}}}\int_{-\infty}^{t-T_{\mathrm{r}}+\tau} \mathcal{E}_s\{s(\theta)s(\theta')\}h_{\mathrm{r}}(t-\theta)h_{\mathrm{r}}(t+\tau-\theta')\,d\theta d\theta'. \end{aligned} \tag{6.40}$$

Using Eq. 6.11 it follows that

$$\mathcal{E}_h\{h_{\mathrm{d}}(t-\theta)h_{\mathrm{d}}(t+\tau-\theta')\} = \sigma_{\mathrm{d}}^2\,\mathrm{e}^{-2\bar{\delta}t}\mathrm{e}^{\bar{\delta}(\theta+\theta'-\tau)}\delta(\theta-\theta'+\tau), \tag{6.41}$$

and

$$\mathcal{E}_h\{h_{\mathrm{r}}(t-\theta)h_{\mathrm{r}}(t+\tau-\theta')\} = \sigma_{\mathrm{r}}^2\,\mathrm{e}^{-2\bar{\delta}t}\mathrm{e}^{\bar{\delta}(\theta+\theta'-\tau)}\delta(\theta-\theta'+\tau). \tag{6.42}$$

Note that $\mathcal{E}\{b_{\mathrm{d}}(t)b_{\mathrm{r}}(t+\tau)\} = 0$ implies that $\mathcal{E}\{h_{\mathrm{d}}(t)h_{\mathrm{r}}(t+\tau)\} = 0$.

The spatially averaged auto-correlation of Eq. 6.40 is

$$\begin{aligned} r_{zz}(t,t+\tau) &= \mathcal{E}_h\{r_{zz}(t,t+\tau;h)\} \\ &= r_{z_\mathrm{d}z_\mathrm{d}}(t,t+\tau) + r_{z_\mathrm{r}z_\mathrm{r}}(t,t+\tau), \end{aligned} \tag{6.43}$$

with

$$r_{z_\mathrm{d}z_\mathrm{d}}(t,t+\tau) = \sigma_\mathrm{d}^2\, \mathrm{e}^{-2\bar{\delta}t} \int_{t-T_\mathrm{r}}^{t} \mathcal{E}\{s(\theta)s(\theta+\tau)\}\mathrm{e}^{2\bar{\delta}\theta}\, d\theta, \tag{6.44}$$

and

$$\begin{aligned} r_{z_\mathrm{r}z_\mathrm{r}}(t,t+\tau) &= \sigma_\mathrm{r}^2\, \mathrm{e}^{-2\bar{\delta}t} \int_{-\infty}^{t-T_\mathrm{r}} \mathcal{E}\{s(\theta)s(\theta+\tau)\}\mathrm{e}^{2\bar{\delta}\theta}\, d\theta \\ &= \sigma_\mathrm{r}^2\, \mathrm{e}^{-2\bar{\delta}t} \int_{t-2T_\mathrm{r}}^{t-T_\mathrm{r}} \mathcal{E}\{s(\theta)s(\theta+\tau)\}\mathrm{e}^{2\bar{\delta}\theta}\, d\theta \\ &\quad + \sigma_\mathrm{r}^2\, \mathrm{e}^{-2\bar{\delta}t} \int_{-\infty}^{t-2T_\mathrm{r}} \mathcal{E}\{s(\theta)s(\theta+\tau)\}\mathrm{e}^{2\bar{\delta}\theta}\, d\theta. \end{aligned} \tag{6.45}$$

The first term in Eq. 6.43 depends on the direct signal between time $t-T_\mathrm{r}$ and t, and the second depends on the reverberant signal.

Let us consider the spatially averaged auto-correlation at time $t-T_\mathrm{r}$:

$$r_{zz}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau) = r_{z_\mathrm{d}z_\mathrm{d}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau) + r_{z_\mathrm{r}z_\mathrm{r}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau), \tag{6.46}$$

with

$$r_{z_\mathrm{d}z_\mathrm{d}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau) = \sigma_\mathrm{d}^2\, \mathrm{e}^{-2\bar{\delta}(t-T_\mathrm{r})} \int_{t-2T_\mathrm{r}}^{t-T_\mathrm{r}} \mathcal{E}\{s(\theta)s(\theta+\tau)\}\mathrm{e}^{2\bar{\delta}\theta}\, d\theta, \tag{6.47}$$

and

$$r_{z_\mathrm{r}z_\mathrm{r}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau) = \sigma_\mathrm{r}^2\, \mathrm{e}^{-2\bar{\delta}(t-T_\mathrm{r})} \int_{-\infty}^{t-2T_\mathrm{r}} \mathcal{E}\{s(\theta)s(\theta+\tau)\}\mathrm{e}^{2\bar{\delta}\theta}\, d\theta. \tag{6.48}$$

The term $r_{z_\mathrm{r}z_\mathrm{r}}(t,t+\tau)$ in Eq. 6.45 can be expressed as

$$\begin{aligned} r_{z_\mathrm{r}z_\mathrm{r}}(t,t+\tau) &= \kappa\, \mathrm{e}^{-2\bar{\delta}T_\mathrm{r}} r_{z_\mathrm{d}z_\mathrm{d}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau) \\ &\quad + \mathrm{e}^{-2\bar{\delta}T_\mathrm{r}}\, r_{z_\mathrm{r}z_\mathrm{r}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau), \end{aligned} \tag{6.49}$$

with $\kappa = \sigma_\mathrm{r}^2/\sigma_\mathrm{d}^2$. Here $\kappa \leq 1$, since it is assumed that $\sigma_\mathrm{d}^2 \geq \sigma_\mathrm{r}^2$. Eq. 6.49 can be rewritten using Eq. 6.46:

$$\begin{aligned} r_{z_\mathrm{r}z_\mathrm{r}}(t,t+\tau) &= \mathrm{e}^{-2\bar{\delta}T_\mathrm{r}}\,(1-\kappa)\, r_{z_\mathrm{r}z_\mathrm{r}}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau) \\ &\quad + \kappa\, \mathrm{e}^{-2\bar{\delta}T_\mathrm{r}}\, r_{zz}(t-T_\mathrm{r},t-T_\mathrm{r}+\tau). \end{aligned} \tag{6.50}$$

The late reverberant component can now be obtained using

$$r_{z_\mathrm{l}z_\mathrm{l}}(t,t+\tau) = \mathrm{e}^{-2\bar{\delta}(T_\mathrm{l}-T_\mathrm{r})} r_{z_\mathrm{r}z_\mathrm{r}}(t-T_\mathrm{l}+T_\mathrm{r},t-T_\mathrm{l}+T_\mathrm{r}+\tau). \tag{6.51}$$

Note that for $\kappa = 1$, i.e., $\sigma_{\mathrm{d}}^2 = \sigma_{\mathrm{r}}^2$, Eq. 6.50 and 6.51 result in Eq. 6.28.

Under the same assumptions as in Section 6.4.1, the counterpart of Eq. 6.51 in terms of the short-term power spectral densities is:

$$P_{z_{\mathrm{l}}z_{\mathrm{l}}}(t,f) = \mathrm{e}^{-2\bar{\delta}(T_{\mathrm{l}}-T_{\mathrm{r}})} P_{z_{\mathrm{r}}z_{\mathrm{r}}}(t - T_{\mathrm{l}} + T_{\mathrm{r}}, f), \tag{6.52}$$

where

$$P_{z_{\mathrm{r}}z_{\mathrm{r}}}(t,f) = \mathrm{e}^{-2\bar{\delta}T_{\mathrm{r}}}\,(1-\kappa)\,P_{z_{\mathrm{r}}z_{\mathrm{r}}}(t - T_{\mathrm{r}}, f) + \kappa\,\mathrm{e}^{-2\bar{\delta}T_{\mathrm{r}}} P_{zz}(t - T_{\mathrm{r}}, f). \tag{6.53}$$

We now define $T_{\mathrm{r}} f_{\mathrm{s}}$ equal to the frame rate R of the STFT. In the STFT domain Eq. 6.52 is then given by

$$\lambda_{z_{\mathrm{l}}}(l,k) = \mathrm{e}^{-2\bar{\delta}\left(T_{\mathrm{l}} - \frac{R}{f_{\mathrm{s}}}\right)} \lambda_{z_{\mathrm{r}}}(l - N_{\mathrm{l}} + 1, k), \tag{6.54}$$

and Eq. 6.53 by

$$\lambda_{z_{\mathrm{r}}}(l,k) = \mathrm{e}^{-2\bar{\delta}\frac{R}{f_{\mathrm{s}}}}\,(1-\kappa)\,\lambda_{z_{\mathrm{r}}}(l-1,k) + \kappa\,\mathrm{e}^{-2\bar{\delta}\frac{R}{f_{\mathrm{s}}}}\lambda_z(l-1,k). \tag{6.55}$$

Remember that T_{l} is chosen such that $N_{\mathrm{l}} = \frac{T_{\mathrm{l}} f_{\mathrm{s}}}{R}$ is an integer value.

When taking the frequency dependency of κ and $\bar{\delta}$ into account we obtain

$$\lambda_{z_{\mathrm{l}}}(l,k) = \mathrm{e}^{-2\bar{\delta}(k)\left(T_{\mathrm{l}} - \frac{R}{f_{\mathrm{s}}}\right)} \lambda_{z_{\mathrm{r}}}(l - N_{\mathrm{l}} + 1, k), \tag{6.56}$$

and

$$\lambda_{z_{\mathrm{r}}}(l,k) = \mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}}\,(1-\kappa(k))\,\lambda_{z_{\mathrm{r}}}(l-1,k) + \kappa(k)\,\mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}}\lambda_z(l-1,k). \tag{6.57}$$

Finally, the late reverberant spectral variance $\lambda_{z_{\mathrm{l}}}(l,k)$ can be estimated given $\hat{\lambda}_z(l,k)$, $\hat{\kappa}(k)$, and $\hat{\bar{\delta}}(k)$, by calculating Eq. 6.56 and 6.57.

The parameter $\hat{\bar{\delta}}$ can be calculated given an estimate of the reverberation time RT_{60}, which can be estimated directly from the AIR using Schroeder's method (see Section 2.8). However, until now we did not discuss how the parameter κ can be obtained, or to which physical quantity it is related. It will now be shown that the ratio $\sigma_{\mathrm{r}}^2/\sigma_{\mathrm{d}}^2$ is related to the DRR, which is defined as

$$\frac{E_{\mathrm{d}}}{E_{\mathrm{r}}} = \frac{\int_0^{T_{\mathrm{r}}} h^2(t)\,\mathrm{d}t}{\int_{T_{\mathrm{r}}}^{\infty} h^2(t)\,\mathrm{d}t}. \tag{6.58}$$

It should be noted that the DRR can be estimated directly from the AIR using Eq. 4.7. However, the AIR is not known *a priori* in many practical situations. The blind estimation of κ and RT_{60} will be discussed in Section 6.6. In general the DRR is frequency dependent, c.f. [41], hence κ is frequency dependent. Therefore, κ can be calculated in different sub-bands to improve the accuracy of the model. Using the model in Eq. 6.6 the direct and reverberant energy can be expressed as

$$E_{\mathrm{d}}(k) = \int_0^{T_{\mathrm{r}}} \sigma_{\mathrm{d}}^2(k)\mathrm{e}^{-2\bar{\delta}(k)t}\,\mathrm{d}t = \frac{\sigma_{\mathrm{d}}^2(k)}{2\bar{\delta}(k)}\left(1 - \mathrm{e}^{-2\bar{\delta}(k)T_{\mathrm{r}}}\right) \tag{6.59}$$

and

$$E_{\mathrm{r}}(k) = \int_{T_{\mathrm{r}}}^{\infty} \sigma_{\mathrm{r}}^2(k) \mathrm{e}^{-2\bar{\delta}(k)t} \, \mathrm{d}t = \frac{\sigma_{\mathrm{r}}^2(k)}{2\bar{\delta}(k)} \mathrm{e}^{-2\bar{\delta}(k)T_{\mathrm{r}}}, \tag{6.60}$$

respectively. Where $\sigma_{\mathrm{d}}^2(k)$ and $\sigma_{\mathrm{r}}^2(k)$ denote the variances of $b_{\mathrm{d}}(t,k)$ and $b_{\mathrm{r}}(t,k)$, respectively. Now the parameter $\kappa(k)$ can be expressed in terms of $E_{\mathrm{d}}(k)$ and $E_{\mathrm{r}}(k)$:

$$\kappa(k) = \frac{\sigma_{\mathrm{r}}^2(k)}{\sigma_{\mathrm{d}}^2(k)} = \frac{1 - \mathrm{e}^{-2\bar{\delta}(k)T_{\mathrm{r}}}}{\mathrm{e}^{-2\bar{\delta}(k)T_{\mathrm{r}}}} \frac{E_{\mathrm{r}}(k)}{E_{\mathrm{d}}(k)}. \tag{6.61}$$

Furthermore, we should keep in mind that the DRR, and thus κ, depends on the distance between the source and microphone. Therefore, spatial averaging can only be performed over those microphone signals that have the same source-microphone distance.

6.5 Estimation in a Noisy Environment

In practice the microphone signal generally consists of the reverberant signal and an ambient noise term $v(t)$, i.e.,

$$\begin{aligned} x(t) &= z(t) + v(t) \\ &= \int_{-\infty}^{t} s(\theta) h(t-\theta) \, \mathrm{d}\theta + v(t). \end{aligned} \tag{6.62}$$

In Section 5.3 we developed a method which allows joint reverberation and noise reduction under the assumption that an unbiased estimate of the late reverberant spectral variance is available. However, in case the noisy signal $x(t)$ is used directly a fraction of the noise will be part of our estimated late reverberant spectral variance, i.e., the estimate will be biased.

In Section 6.5.1 it is shown how an unbiased estimate can be obtained using an additional pre-processing step. Additionally, it will be shown in Section 6.5.2 that this bias can be predicted in case the noise characteristics are time-invariant.

6.5.1 Unbiased Estimation

In terms of the discrete STFT Eq. 6.62 can be expressed as

$$X(l,k) = Z(l,k) + V(l,k). \tag{6.63}$$

For the estimation of the late reverberant spectral variance $\lambda_{z_l}(l,k)$, an estimate of the power spectrum $|Z(l,k)|^2$ is required. The power spectrum of the reverberant spectral component $Z(l,k)$ can be estimated by minimizing

$$\mathcal{E}\left\{ \left(A(l,k) - |\hat{Z}(l,k)|^2 \right)^2 \right\}, \tag{6.64}$$

where $A(l,k) = |Z(l,k)|^2$, and $\hat{Z}(l,k) = G_{\text{SP}}(l,k)X(l,k)$. As shown in [234] this leads to the following spectral gain function

$$G_{\text{SP}}(l,k) = \sqrt{\frac{\xi_{\text{SP}}(l,k)}{1+\xi_{\text{SP}}(l,k)}\left(\frac{1}{\gamma_{\text{SP}}(l,k)} + \frac{\xi_{\text{SP}}(l,k)}{1+\xi_{\text{SP}}(l,k)}\right)}, \tag{6.65}$$

where

$$\xi_{\text{SP}}(l,k) = \frac{\lambda_z(l,k)}{\lambda_v(l,k)} \tag{6.66}$$

and

$$\gamma_{\text{SP}}(l,k) = \frac{|X(l,k)|^2}{\lambda_v(l,k)}, \tag{6.67}$$

denote the *a priori* and *a posteriori* Signal to Interference Ratios (SIR), respectively. It should be noted that $\lambda_z(l,k)$ is not known *a priori*. However, the *a priori* SIR can be estimated using the decision-directed estimator proposed by Ephraim and Malah [210], or by the causal or non-causal recursive estimators proposed by Cohen [208] (also see Appendix B for more details on the causal and non-causal *a priori* SIR estimators). An estimate of the noise spectral variance $\lambda_v(l,k)$ is obtained using the Improved Minima Controlled Recursive Averaging (IMCRA) approach [190]. The estimate $\hat{A}(l,k)$ of the power spectrum $|Z(l,k)|^2$ is then given by:

$$\hat{A}(l,k) = \left(G_{\text{SP}}(l,k)\right)^2 |X(l,k)|^2. \tag{6.68}$$

An unbiased estimate of the reverberant speech spectral variance $\lambda_z(l,k)$ can now be obtained by substituting $|Z(l,k)|^2$ in Eq. 6.38 by $\hat{A}(l,k)$. The estimate $\hat{\lambda}_z(l,k)$ can then be used in Eq. 6.57 to estimate the reverberant spectral variance $\lambda_{z_{\text{r}}}(l,k)$. Subsequently, $\hat{\lambda}_{z_{\text{r}}}(l,k)$ can be used in Eq. 6.56 to estimate the late reverberant spectral variance $\lambda_{z_{\text{l}}}(l,k)$.

6.5.2 Bias Estimation and Correction

It will now be shown that the bias can be predicted in case the noise is time-invariant. When the bias has been predicted it can be used to apply a correction to the estimated late reverberant spectral variance $\lambda_{z_{\text{l}}}(l,k)$[1], as shown in Fig. 6.1. This correction will circumvent any over-subtraction of the noise by the post-filter $G(l,k)$. Although the proposed modification reduces the overall complexity of our system it should be noted that the unbiased estimator in Section 6.5.1 is more general, and can also be used when dealing with non-stationary interferences (see for example Chapter 7).

In case the noise is assumed to be time-invariant the complexity of our system can be reduced by omitting the estimation step that was proposed in Section 6.5.1. Let us assume that the noise spectral variance $\lambda_v(l,k)$ is not suppressed prior to the

[1]Note that a correction could also be applied to the noise spectral variance. Thereby, reducing the noise spectral variance $\lambda_v(l,k)$ with the noise that is already included in $\lambda_{x_{\text{l}}}(l,k)$.

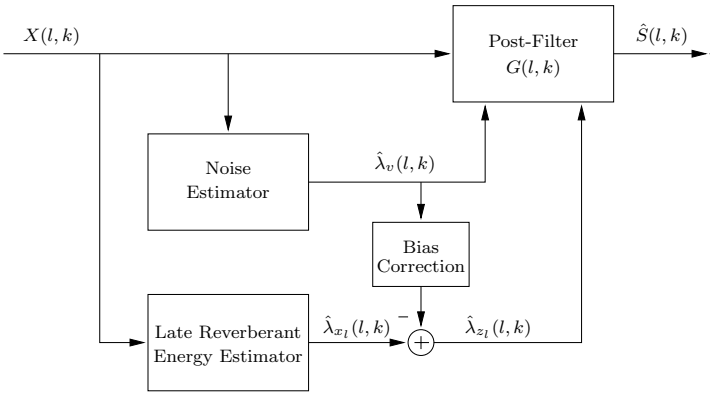


Figure 6.1 *Joint reverberation and noise reduction with bias correction.*

estimation of the late reverberant spectral variance. Assuming that the reverberant signal and noise are uncorrelated, and $\lambda_v(l,k) = \varepsilon(k)$, we have

$$\lambda_x(l,k) = \lambda_z(l,k) + \varepsilon(k), \tag{6.69}$$

where $\varepsilon(k)$ denotes the frequency dependent bias. To study the effect of this bias on our estimator we first recast Eq. 6.53 in a non-recursive manner:[2]

$$\begin{aligned}\lambda_{x_\mathrm{r}}(l,k) &= \mathrm{e}^{-2\bar{\delta}\frac{R}{f_\mathrm{s}}}\,(1-\kappa)\,\lambda_{x_\mathrm{r}}(l-1,k) + \kappa\,\mathrm{e}^{-2\bar{\delta}\frac{R}{f_\mathrm{s}}}\lambda_x(l-1,k)\\ &= \sum_{j=1}^{\infty}\lambda_x(l-j-1,k)\left(\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_\mathrm{s}}}\right)^j + \kappa\mathrm{e}^{-2\bar{\delta}\frac{R}{f_\mathrm{s}}}\lambda_x(l-1,k),\end{aligned} \tag{6.70}$$

where $\lambda_{x_\mathrm{r}}(l,k)$ is the late reverberant spectral variance that is obtained from $\lambda_x(l,k)$. Now Eq. 6.69 is substituted in Eq. 6.70. After rearranging the terms this results in

$$\begin{aligned}\lambda_{x_\mathrm{r}}(l,k) = \sum_{j=1}^{\infty}\lambda_x(l-j-1,k)\left(\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_\mathrm{s}}}\right)^j + \kappa\mathrm{e}^{-2\bar{\delta}\frac{R}{f_\mathrm{s}}}\lambda_z(l-1,k)\\ + \sum_{j=1}^{\infty}\varepsilon(k)\left(\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_\mathrm{s}}}\right)^j + \kappa\mathrm{e}^{-2\bar{\delta}\frac{R}{f_\mathrm{s}}}\varepsilon(k).\end{aligned} \tag{6.71}$$

The spectral variance $\lambda_{x_\mathrm{r}}(l,k)$ can be expressed as

$$\lambda_{x_\mathrm{r}}(l,k) = \lambda_{z_\mathrm{r}}(l,k) + \varepsilon_{z_\mathrm{r}}(k), \tag{6.72}$$

where $\varepsilon_{z_\mathrm{r}}(k)$ denotes the bias in $\lambda_{x_\mathrm{r}}(l,k)$. The infinite geometrical series in the last term of Eq. 6.71 can be expanded, i.e,

$$\sum_{j=1}^{\infty}\varepsilon(k)\left(\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_\mathrm{s}}}\right)^j = \varepsilon(k)\frac{\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_\mathrm{s}}}}{1-\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_\mathrm{s}}}}. \tag{6.73}$$

[2]The frequency dependency of $\bar{\delta}$ and κ has been omitted to simply the notation.

We can see that the noise will bias the estimated spectral variance of the reverberant speech $\lambda_{z_r}(l,k)$, i.e.,

$$\varepsilon_{z_r}(k) = \varepsilon(k)\left(\frac{\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_s}}}{1-\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_s}}} + \kappa\mathrm{e}^{-2\bar{\delta}\frac{R}{f_s}}\right). \tag{6.74}$$

The spectral variance of the late reverberant signal component $\lambda_{x_l}(l,k)$ is calculated using Eq. 6.56, i.e.,

$$\lambda_{x_l}(l,k) = \mathrm{e}^{-2\bar{\delta}(T_l-\frac{R}{f_s})}\lambda_{x_r}(l-N_l+1,k). \tag{6.75}$$

It can easily be verified that $\lambda_{x_l}(l,k) = \lambda_{z_l}(l,k)+\varepsilon_{z_l}(k)$, where $\varepsilon_{z_l}(k)$ denoted the bias in $\lambda_{x_l}(l,k)$ which is given by

$$\varepsilon_{z_l}(k) = \varepsilon(k)\mathrm{e}^{-2\bar{\delta}(T_l-\frac{R}{f_s})}\left(\frac{\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_s}}}{1-\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_s}}} + \kappa\mathrm{e}^{-2\bar{\delta}\frac{R}{f_s}}\right). \tag{6.76}$$

It should be noted that in case $\kappa = 1$ Eq. 6.76 reduces to

$$\varepsilon_{z_l}(k) = \varepsilon(k)\mathrm{e}^{-2\bar{\delta}T_l}. \tag{6.77}$$

Finally, the correction can be applied to the estimated late reverberant spectral variance $\hat{\lambda}_{x_l}(l,k)$, i.e.,

$$\hat{\lambda}_{z_l}(l,k) = \hat{\lambda}_{x_l}(l,k) - \hat{\lambda}_v(l,k)\mathrm{e}^{-2\bar{\delta}(T_l-\frac{R}{f_s})}\left(\frac{\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_s}}}{1-\kappa(1-\kappa)\mathrm{e}^{-4\bar{\delta}\frac{R}{f_s}}} + \kappa\mathrm{e}^{-2\bar{\delta}\frac{R}{f_s}}\right), \tag{6.78}$$

where $\hat{\lambda}_{z_l}(l,k)$ denotes the corrected late reverberant spectral variance.

6.6 Reverberation Time and DRR Estimator

In order to estimate the late reverberant spectral variance an estimate the reverberation time RT_{60} of the room, and the parameter κ is required. In Section 6.4.2 it was already shown that the parameter κ is related to the DRR.

Partially blind methods to estimate the reverberation time have been developed in which the characteristics of the room are 'learnt' using neural network approaches [235]. Another method uses a segmentation procedure for detecting gaps in the signals, and tracks the sound decay curve [24, 196]. Recently, a blind method has been proposed by Ratnam et al. based on a maximum-likelihood estimation procedure [236]. Most of these methods can also be applied to sub-band signals in order to estimate the frequency dependent reverberation time. In some applications, e.g., echo cancellation, one could estimate the reverberation time using the estimate echo path (see Chapter 7). It is reasonable to assume that the reverberation time is approximately constant in the room. For some applications, e.g, audio or video-conferencing where a fix setup is used the reverberation could also be obtained using a calibration process.

To the best of our knowledge there are no blind estimation procedures available to acquire an estimate the DRR. In many practical situations the distance between the source and the microphone will vary. Since the DRR depends on the distance between the source and the microphone it is important that the parameter κ can be estimated online. Remember that the parameter κ was introduced to prevent over-estimation of the reverberant spectral variance $\hat{\lambda}_{z_\mathrm{r}}(l,k)$. In case κ is too large the spectral variance $\hat{\lambda}_{z_\mathrm{r}}(l,k)$ could become larger than $|Z(l,k)|^2$, which indicates that over-estimation has occurred. Therefore, the value of κ should be lowered. Furthermore, during the free-decay, which occurs after an offset of the source signal, $\hat{\lambda}_{z_\mathrm{r}}(l,k)$ should be approximately equal to $|Z(l,k)|^2$. Estimation of κ could therefore be performed after a speech offset. Unfortunately, the detection of speech offsets is rather difficult. From the above discussion is has become clear that κ should at least fulfill the following condition: $|Z(l,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(l,k) \geq 0$.

We propose to estimate the parameter κ adaptively. When speech is detected and $|Z(l,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(l,k) < 0$ the value of κ is lowered, when $|Z(l,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(l,k) > 0$ the value of κ is raised slowly. When $|Z(l,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(l,k) = 0$ the value of κ is assumed to be correct. This update scheme can be implemented as follows:

$$\hat{\kappa}(l+1) = \begin{cases} \hat{\kappa}(l) + \mu_\kappa \left(1 - \dfrac{\sum_{k=0}^{\frac{K}{2}-1} \hat{\lambda}_{z_\mathrm{r}}(l,k)}{\sum_{k=0}^{\frac{K}{2}-1} |Z(l,k)|^2}\right), & \text{speech present;} \\ \hat{\kappa}(l), & \text{otherwise,} \end{cases} \tag{6.79}$$

where μ_κ $(0 < \mu < 1)$ denotes the step-size. After each update step $\hat{\kappa}(l+1,k)$ is constrained, such that $0 < \hat{\kappa}(l+1) \leq 1$. Experimental results that demonstrate the feasibility of the proposed estimator can be found in Section 6.7.

6.7 Simulation Results

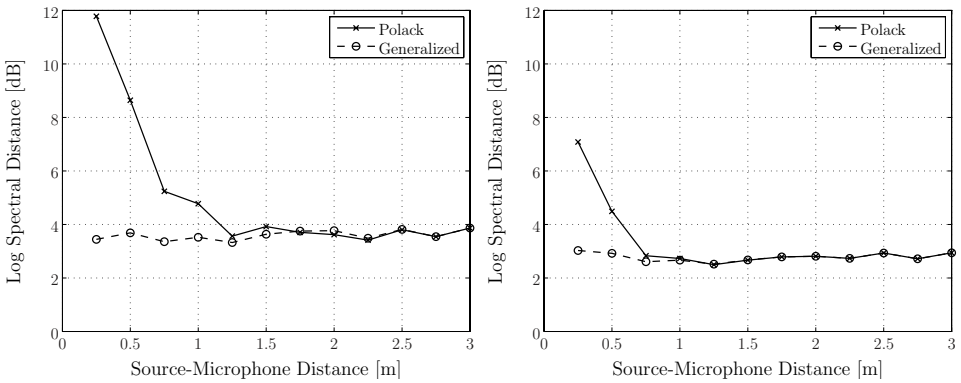
In this section we evaluate the accuracy of the two late reverberant spectral variance estimators that are developed in this chapter.

The speech fragment used in our experiments consists of male and female speech ($f_\mathrm{s} = 16$ kHz) taken from the TIMIT database [4], and is 40 seconds long. The reverberant speech fragments are obtained by convolving the anechoic speech signal with an AIR that was generated using the image method (see Appendix A).

The accuracy of the estimators is determined by the Log Spectral Distance between λ_{z_l} and $\hat{\lambda}_{z_\mathrm{l}}$, and is calculated using

$$\mathrm{LSD} = \frac{2}{KL} \sum_{l} \sum_{k=0}^{\frac{K}{2}-1} \left| \mathcal{L}\{\lambda_{z_\mathrm{l}}(l,k)\} - \mathcal{L}\{\hat{\lambda}_{z_\mathrm{l}}(l,k)\} \right| \quad [\mathrm{dB}], \tag{6.80}$$

where L denotes the number of frame, and $\mathcal{L}\{\lambda(l,k)\} \triangleq \max\{10\log_{10}(\lambda), \delta\}$ is the log spectrum confined to about 50 dB dynamic range ($\delta = \max_{l,k}\{10\log_{10}(\lambda)\} - 50$).



(a) $\mathrm{RT}_{60} = 0.3$ s, $T_\mathrm{l} = 48$ ms (b) $\mathrm{RT}_{60} = 0.6$ s, $T_\mathrm{l} = 48$ ms

Figure 6.2 *The LSD for a range of source-microphone distances.*

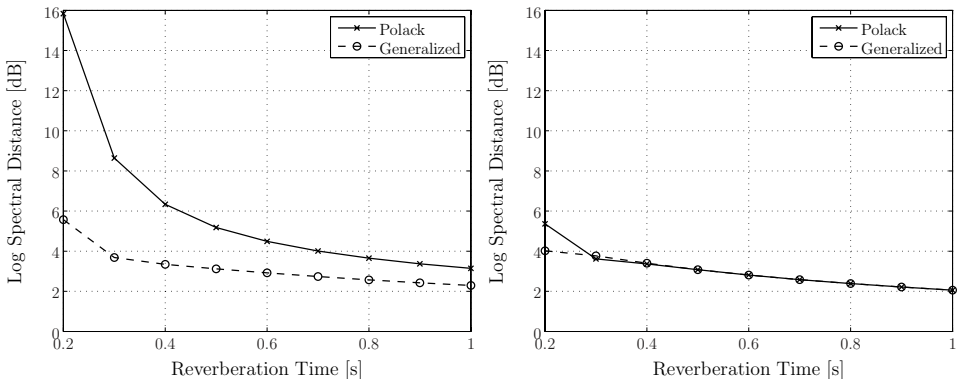
6.7.1 Estimation in a Noise-Free Environment

The first experiment is related to the estimation of the late reverberant spectral variance from a noise-free observation using a single microphone. For evaluation purposes we used the (full-band) reverberation time which was measured directly from the AIRs using Schroeder's method [76]. The damping constant $\bar{\delta}$ can then be calculated using Eq. 2.37. The DRR was calculated from the same AIR using Eq. 6.58. The parameter κ was calculated using Eq. 6.61. It should be noted that this procedure only gives an approximate value of RT_{60} and κ.

In Fig. 6.2 the Log Spectral Distortion (LSD) at various distances ($D = 0.25, 0.5, \ldots,$ 3 m) is shown for both estimators, and a reverberation time of 0.3 and 0.6 s. The critical distance (as defined in Eq. 2.59) was 1.17 and 0.82 m for a reverberation time of 0.3 and 0.6 s, respectively. It can clearly been seen that for a source-microphone distance smaller than the critical distance the proposed estimator, which is based on the generalized model, results in a smaller LSD compared to the estimator, which is based on Polack's model.

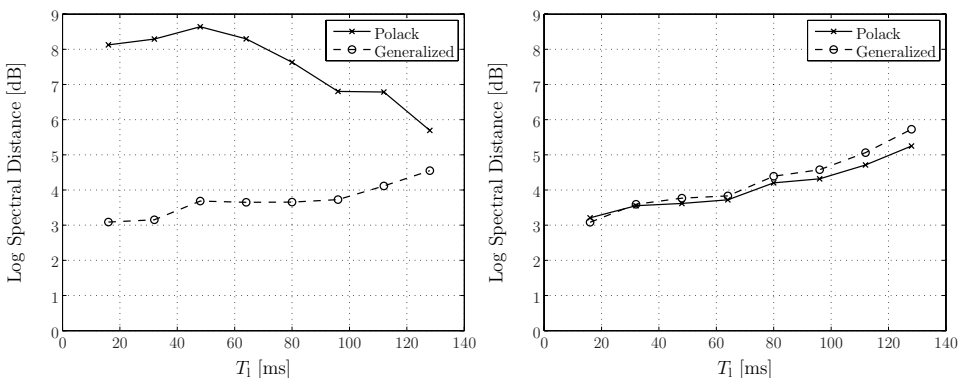
In Fig. 6.3 the LSD is shown for different reverberation times ($\mathrm{RT}_{60} = 0.2, 0.3, \ldots, 1$ s), and a source-microphone distance of 0.5 and 2 m. For both source-microphone distances the LSD decreases when the reverberation time increases. The increase in reverberation time increases the echo density (see Eq. 2.33), and hence the randomness of the AIR. For a source-microphone distance of 0.5 m the proposed late reverberant spectral variance estimator performs much better than the estimator that is based on Polack's statistical reverberation model.

In Fig. 6.4 the LSD is shown for $T_\mathrm{l} = 16, 32, \ldots, 128$ ms, $\mathrm{RT}_{60} = 0.3$ s, and a source-microphone distance of 0.5 and 2 m. When using the estimator that is based on Polack's statistical model the LSD slightly decreases for increasing values of T_l, and $D = 0.5$ m. When the generalized estimator is used the LSD increases for increasing



(a) $D = 0.5$ m, $T_\mathrm{l} = 48$ ms (b) $D = 2$ m, $T_\mathrm{l} = 48$ ms

Figure 6.3 *The LSD for reverberation times between 0.2 s and 1 s.*

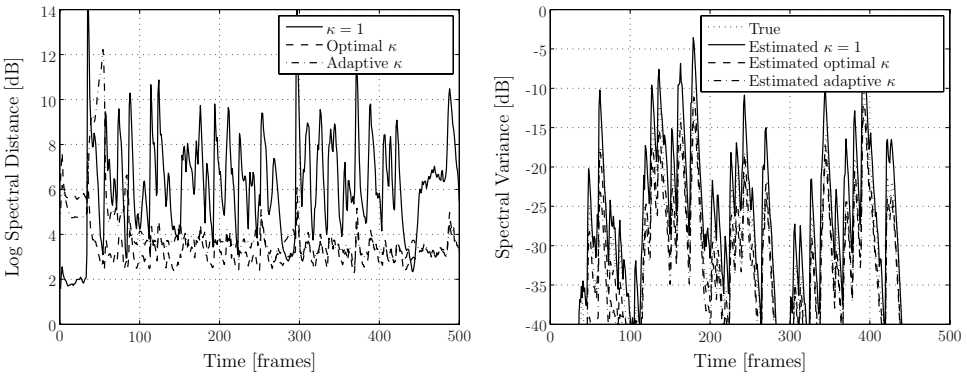


(a) $\mathrm{RT}_{60} = 0.3$ s, $D = 0.5$ m (b) $\mathrm{RT}_{60} = 0.3$ s, $D = 2$ m

Figure 6.4 *The LSD for different values of* T_l.

values of T_l for $D = 0.5$ and 2 m. The results demonstrate that the accuracy of the late reverberant spectral variance estimator reduces for larger values of T_l. It is understood that small model mismatches, e.g., errors in RT_{60} and κ, have a larger influence on the estimation when T_l increases.

In the foregoing experiments the reverberation time RT_{60} and κ were measured using the AIR, resulting in an 'optimal' value of κ. Since the AIR is not known *a priori* in practice RT_{60} and κ are unknown. In the literature various solutions have been proposed to estimate the reverberation time blindly, as discussed in Section 6.6. In the following experiment we assume that the reverberation time ($\mathrm{RT}_{60} = 0.4$ s) is known, whereas κ is estimated adaptively using the method developed in Section 6.6. The source-receiver distance was set to 0.5 m and $T_\mathrm{l} = 48$ ms. The instantaneous LSD



(a) $\mathrm{RT}_{60} = 0.4$ s, $D = 0.5$ m, $T_\mathrm{I} = 48$ ms (b) $\mathrm{RT}_{60} = 0.4$ s, $D = 0.5$ m, $T_\mathrm{I} = 48$ ms

Figure 6.5 *a) The instantaneous LSD using $\kappa = 1$, optimal κ and adaptive κ, b) the spectral variance for $k = 24$ of λ_{z_l}, $\hat{\lambda}_{z_\mathrm{l}}$ using $\kappa = 1$, $\hat{\lambda}_{z_\mathrm{l}}$ using the optimal κ and $\hat{\lambda}_{z_\mathrm{l}}$ using the adaptive κ.*

between λ_{z_l} and $\hat{\lambda}_{z_\mathrm{l}}$ is shown in Fig. 6.5(a). The late reverberant spectral variance $\hat{\lambda}_{z_\mathrm{l}}$ was obtained using $\kappa = 1$, the optimal κ obtained from the AIR and the adaptive κ that was obtained using Eq. 6.79 with $\mu_\kappa = 0.01$. The initial value of κ was set to 0.01. It can be seen that the instantaneous LSD that was obtained using $\kappa = 1$, i.e., using Polack's model, is much larger than the instantaneous LSD that was obtained using the generalized model. Furthermore, the instantaneous LSD obtained using the adaptive κ slowly converges to the instantaneous LSD that was obtained using the optimal κ. In Fig. 6.5(b) we have shown the 'true' and three estimated late reverberant spectral variances for a single frequency bin ($k = 24$). The estimated spectral variances were obtained using $\kappa = 1$, the 'optimal' κ and the adaptive κ. It can clearly be seen that the late reverberant spectral variance is over-estimated when $\kappa = 1$. However, the late reverberant spectral variance that was obtained using the adaptive κ slowly converges to the late reverberant spectral variance that was obtained using the optimal κ. This experiment demonstrates the possibility of estimating κ blindly in case the reverberation time is know.

6.7.2 Estimation in a Noisy Environment

In this section we have studied the estimation of the late reverberant spectral variance in case background noise is present. The additive noise $v(n)$ was speech-like noise, taken from the NOISEX-92 database [237]. The spectral variance of the noise was estimated from the noisy microphone signal $x(n)$ using the IMCRA approach [190].

In Fig. 6.6 the LSD is shown that is obtained for different segmental Signal to Noise Ratio (SNR) values using the standard estimation and the noisy observation $x(n)$, the unbiased estimation as described in Section 6.5, and the standard estimation using the

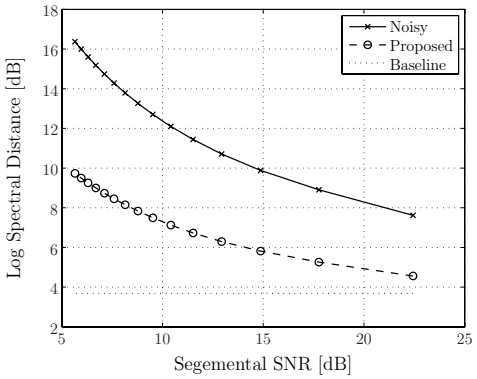


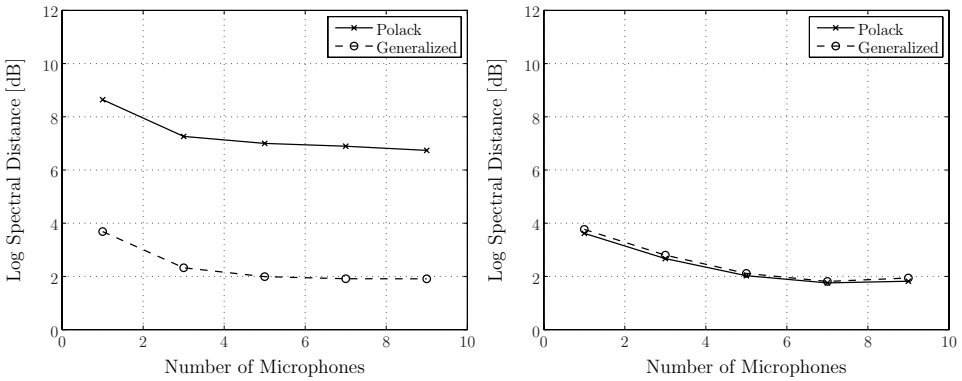
Figure 6.6 *The LSD between the optimal spectral variance and the estimated spectral variance that was obtained using the biased and unbiased estimator* ($\mathrm{RT}_{60} = 0.3$ *s*, $D = 0.5$ *m*).

noise-free signal $z(n)$ (baseline). The results demonstrate that the unbiased estimator reduces the deviation caused by the background noise.

6.7.3 Estimation using Multiple Microphones

We will now analyse the LSD when multiple microphones are used. For this evaluation we used the spatially averaged (full-band) reverberation time which was measured directly from the AIRs using Schroeder's method [76]. The damping constant $\overline{\delta}$ can then be calculated using Eq. 2.37. The DRR was calculated from the same AIRs using Eq. 6.58. First an estimate of κ was calculated for each source-microphone pair using Eq. 6.61. Finally, the results were averaged over all source-microphone pairs. It should be noted that this procedure only gives an approximate value for κ, and that the spacial averaging may only be performed in case the source-microphone distances are approximately equal. In this case the microphones were positioned around the source at equal distance.

The LSD obtained for $M = 1, 3, \ldots, 9$ microphones at a distance of 0.5 and 2 m is shown in Fig. 6.7. The results demonstrate that the LSD obtained using both late reverberant spectral variance estimators decreases for increasing number of microphones. It should be noted that the largest improvement in LSD is achieved when three or five microphones are used instead of one. Furthermore, we can see that at a distance of 0.5 m the LSDs obtained using Polack's model are larger than the LSDs obtained using the generalized model. At a distance of 2 m the LSDs obtained using Polack's model and the generalized model are similar.



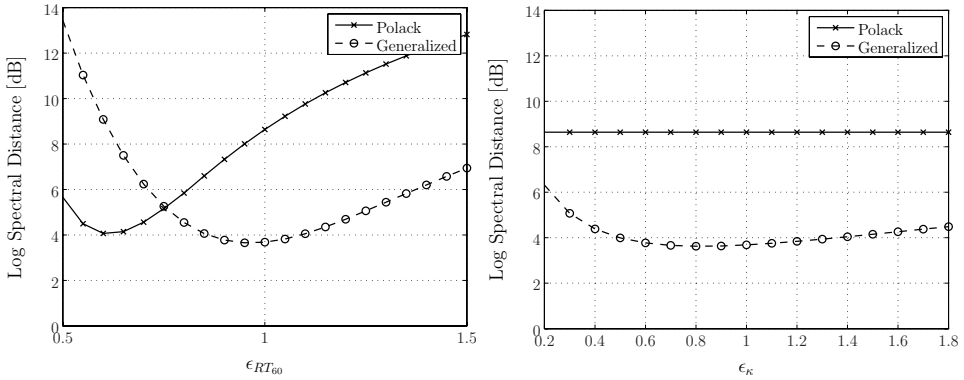
(a) $\mathrm{RT}_{60} = 0.3$ s, $D = 0.5$ m (b) $\mathrm{RT}_{60} = 0.3$ s, $D = 2$ m

Figure 6.7 *LSD obtained using multiple microphones.*

6.7.4 Parameter Estimation Errors

The proposed late reverberant spectral variance estimator requires an estimate of RT_{60} and κ. In general these parameters can not be acquired perfectly, for example due to background noise or changes in the room, or in the position of the source or the microphone. The sensitivity to RT_{60} and κ were studied by introducing errors ranging from -50% till $+50\%$ ($\epsilon_{\mathrm{RT}_{60}} = 0.5, 0.6, \ldots, 1.5$), and from -80% till $+80\%$ ($\epsilon_{\kappa} = 0.2, 0.1, \ldots, 1.8$), respectively. In this experiment only one microphone signal was used.

The LSD that was obtained using $\widehat{RT}_{60} = \epsilon_{\mathrm{RT}_{60}}\mathrm{RT}_{60}$ is shown in Fig. 6.8(a). The results demonstrate that minimum distortion is achieved for the 'correct' RT_{60} value when the generalized estimator is used. For the estimator that is based on Polack's model the minimum LSD that was obtained for $\epsilon_{\mathrm{RT}_{60}} \approx 0.625$, and suggest that for distances smaller than the critical distance an improvement is expected when a correction is applied to the reverberation time. It can be shown that the correction is related to the DRR. Although this correction can decrease the average LSD it should be noted that correcting the reverberation time results in an under-estimation during the free-decay, which occurs after an speech offset. Especially in these periods the improvement of the proposed generalized estimator is much larger, since it uses the correct reverberation time. The LSD that was obtained using $\hat{\kappa} = \epsilon_{\kappa}\kappa$ is shown in Fig. 6.8(b). Since the estimator that is based on Polack's model is not a function of κ the LSD is not affected by the changes in κ. The LSD results that were acquired using the proposed late reverberant spectral variance estimator are also shown in Fig. 6.9 as a function of $\epsilon_{\mathrm{RT}_{60}}$ and ϵ_{κ}.



(a) $\mathrm{RT}_{60} = 0.3$ s, $D = 0.5$ m (b) $\mathrm{RT}_{60} = 0.3$ s, $D = 0.5$ m

Figure 6.8 *LSD with errors in (a)* RT_{60} *and (b)* κ.

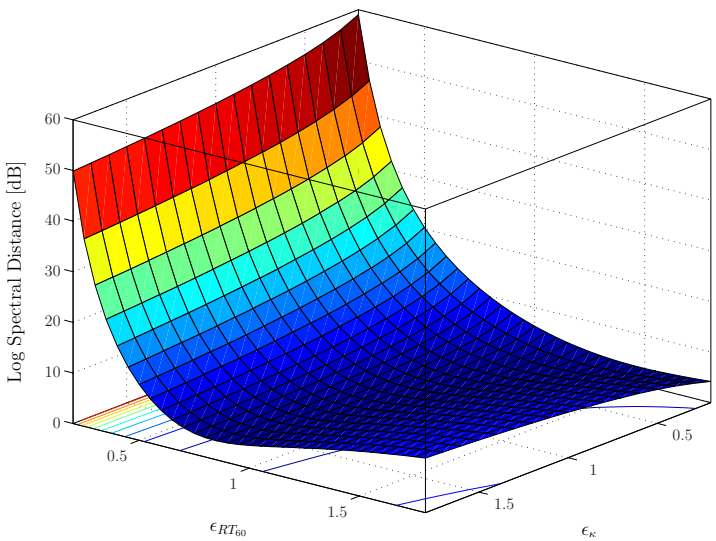


Figure 6.9 *3D plot of the LSD values obtained using the generalized spectral variance estimator with errors in* RT_{60} *and* κ *(*$\mathrm{RT}_{60} = 0.3$ *s,* $D = 0.5$ *m,* $M = 1$*).*

6.8 Conclusions

In this chapter we have described Polack's statistical reverberation model, and proved that it arcuately models the reverberant energy density which is described by the energy balance equation. It was shown that the energy envelope of the AIR can be obtained by spatial averaging. Additionally, a generalized statistical reverberation model was developed, which can be used to take the energy of the direct path into account. The proposed generalization is important in case the source-microphone distance is smaller than the critical distance. Two estimators were derived for the late

reverberant spectral variance. The first is based on Polack's statistical model and the second is based on the generalized statistical reverberation model.

The estimation of the late reverberant spectral variance is based on a noise-free observation of the signal. The influence of noise on the estimation procedure was discussed, and a method to acquire an unbiased estimate of the late reverberant spectral variance was proposed. Furthermore, it was shown that, for time-invariant noise sources, the bias is related to the noise spectral variance and can be calculated directly from the estimation model parameters. A correction can then be applied to the estimated late reverberant spectral variance to avoid over-estimation.

Simulation results demonstrated the feasibility of the developed late reverberant spectral variance estimators.

Chapter 7

Joint Dereverberation and Residual Echo Suppression

7.1 Introduction

Hands-free devices, such as a mobile telephone, are one of the most important and well-known applications for acoustic echo cancellation systems [238]. The acoustic echo cancellation system allows the possibility to conveniently use a hands-free device while maintaining a high user satisfaction in terms of speech distortion and acoustic echo attenuation. Furthermore, in driving cars the hands-free functionality if often required by law.

Hands-free devices are often used in a noisy enclosed environment. Therefore, the microphone will not only receive the desired sound (commonly called *near-end sound*) but also reverberation and background noise. These distortions degrade the fidelity and intelligibility of speech and the recognition performance of automatic speech recognition systems.

7.1.1 Problem Statement

A conventional acoustic echo canceller and Loudspeaker Enclosure Microphone (LEM) system are depicted in Fig. 7.1. The microphone signal is denoted by $y(n)$, where n denotes the discrete time index. The microphone signal $y(n)$ consists of a reverberant speech component $z(n)$, an acoustic echo $d(n)$, and a noise component $v(n)$. The reverberant near-end speech component is given by

$$z(n) = \sum_{j=0}^{L_{\mathrm{a}}-1} a_j(n)s(n-j), \tag{7.1}$$

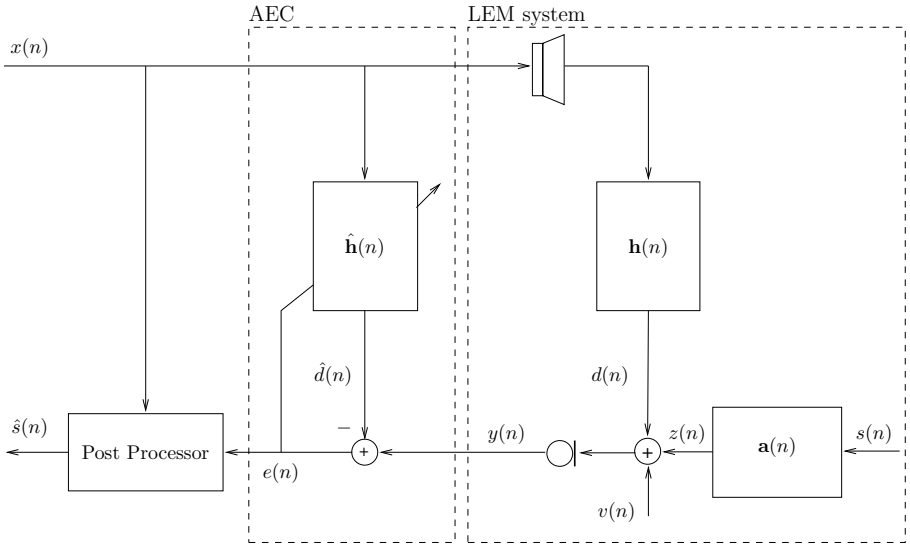


Figure 7.1 *Conventional acoustic echo canceller with post-processor.*

where $a_j(n)$ denotes the j^{th} coefficient of the Acoustic Impulse Response (AIR) that describes the system between the near-end source and the microphone at time n, L_{a} is the length of the AIR, and $s(n)$ denotes the anechoic speech signal. The acoustic echo is

$$d(n) = \sum_{j=0}^{L_{\text{h}}-1} h_j(n)x(n-j), \tag{7.2}$$

where $h_j(n)$ denotes the j^{th} coefficient of the acoustic echo path at time n, L_h is the length of the acoustic echo path, and $x(n)$ denotes the far-end speech signal.

The ultimate goal is to obtain an estimate $\hat{z}_{\text{d}}(n)$ of the direct speech signal $z_{\text{d}}(n)$, which is a delayed and attenuated version of $s(n)$. However, most systems try to estimate the reverberant speech signal $z(n)$ or the noisy reverberant speech signal $z(n) + v(n)$. In this chapter we show how an estimate of the dereverberated near-end speech signal can be obtained.

The error signal of the acoustic echo canceller is given by

$$\begin{aligned} e(n) &= y(n) - \hat{d}(n) \\ &= z(n) + d(n) + v(n) - \hat{d}(n) \\ &= z(n) + e_r(n) + v(n), \end{aligned} \tag{7.3}$$

where $\hat{d}(n)$ denotes the estimated echo signal, and $e_r(n) = d(n) - \hat{d}(n)$ denotes the *residual echo signal*.

In general the residual echo signal $e_r(n)$ will not be zero. Three main reasons are stated here: [239, 240]

1) The length of the acoustic echo path $\mathbf{h}(n) = [h_0(n), \ldots, h_{L_\mathrm{h}-1}(n)]^T$ is approximately defined by $\mathrm{RT}_{60} f_\mathrm{s}$, where f_s denotes the sampling frequency, and RT_{60} denotes the reverberation time in seconds. Due to practical reasons, e.g., complexity, slow convergence of long adaptive filters, and robustness, the length of the estimate $\hat{\mathbf{h}}(n)$ has to be limited to a certain length, L_AEP. Thus, the amount of echo resulting from the unmodelled part of the echo path cannot be removed.

2) The adaptation is not ideal, i.e., the system mismatch vector

$$\boldsymbol{\Delta}(n) = \mathbf{h}'(n) - \hat{\mathbf{h}}(n), \tag{7.4}$$

where $\mathbf{h}'(n) = [h_0(n), \ldots, h_{L_\mathrm{AEP}-1}(n)]^T$ is a truncated version of $\mathbf{h}(n)$, is not zero.

3) The echo may contain some components that are non-linear with respect to $x(n)$. Since the model is linear, these components cannot be handled correctly. This non-linearity has a twofold influence. First, it contributes to the residual echo. Secondly, it influences the convergence of the adaptive filter, which may lead to even larger residual echo components.

A measure to express the effect of the echo cancellation is the Echo Return Loss Enhancement (ERLE): [239]

$$\mathrm{ERLE} = 10 \log_{10} \left(\frac{\mathcal{E}\left\{d^2(n)\right\}}{\mathcal{E}\left\{\left(d(n) - \hat{d}(n)\right)^2\right\}} \right), \tag{7.5}$$

where $\mathcal{E}\{\cdot\}$ denotes mathematical expectation. The maximum achievable ERLE is thus determined by the amount of residual echo.

Under the assumption that the acoustic echo path $\mathbf{h}(n)$ is exponentially decaying, Breining et al. [241] showed that the maximum ERLE is given by

$$\mathrm{ERLE}_\mathrm{max} = 60 \frac{L_\mathrm{AEP}}{f_\mathrm{s} \mathrm{RT}_{60}} \ \mathrm{dB}, \tag{7.6}$$

where L_AEP denotes the length of the estimated acoustic echo path, which is usually smaller than the length of the true echo path L_h. The reverberation time is defined as the time necessary for a 60 dB decay of the sound energy after switching off the sound source. It should be noted that this result is only valid in case the distance between the loudspeaker and the microphone is larger than the critical distance as defined in Eq. 2.59.

7.1.2 Review of previous work

In this section we give a short review of related previous work. An extensive review can be found in [238, 242, 239].

In the 1960s voice-controlled switching systems have been developed to suppress the acoustic echo of the far-end speaker. These systems strongly suppress the hands-free microphone signal whenever the far-end signal is detected. The main drawback of this technique is that it leads to an unacceptable half-duplex connection between the two ends of the communication system [243]. Furthermore, due to this technique the background noise is also strongly suppressed, and the remaining background noise sounds very unnatural. Therefore, voice-controlled switches are nowadays implemented in conjunction with comfort noise injection.

The acoustic echo cancellation problem is usually solved by using an adaptive filter in parallel to the acoustic echo path [238, 241, 242, 239]. The adaptive filter is used to generate a signal that is a replica of the acoustic echo signal. An estimate of the near-end speech signal is then obtained by subtracting the estimated acoustic echo signal, i.e., the output of the adaptive filter, from the microphone signal. Highly sophisticated control mechanisms have been proposed for fast and robust adaptation of the adaptive filter coefficients in realistic acoustic environments [239, 240]. As mentioned in the previous section there is always a residual echo after the echo canceller, and it is widely accepted that echo cancellers alone will not be able to deliver a sufficient echo attenuation [242, 239, 240, 244]. Turbin et al. compared three post-filtering techniques to reduce the residual echo and concluded that the spectral subtraction technique, which is commonly used for noise reduction, was the most efficient [245]. In a reverberant environment there can be a large amount of late residual echo due the deficient length of the adaptive filter. In [244] Enzner proposed a recursive estimator for the short-term Power Spectral Density (PSD) of the late residual echo signal using an estimate of the reverberation time of the room. The reverberation time can be estimated directly from the estimated echo path. The estimated short-term PSD of the late residual echo signal is then suppressed using a spectral enhancement technique.

Hands-free devices are often used in a noisy environment. In some applications like hands-free terminal devices, noise reduction becomes necessary due to the relatively large distance between the microphone and the mouth of the speaker. The first attempts to develop a combined echo and noise reduction system can be attributed to Grenier et al. [246, 247] and to Yasukawa [248]. Both employ more than one microphone. A survey of these systems can be found in [239, 249]. In [250] Gustafsson et al. proposed two post-filters for residual echo and noise reduction. The first post-filter was based on the Log Spectral Amplitude estimator [205] and was extended to attenuate multiple interferences. The second post-filter was psychoacoustically motivated.

Martin and Vary proposed a system for joint acoustic echo cancellation, dereverberation, and noise reduction using two microphones [251]. A similar system was developed by Dörbecker and Ernst in [252]. In both papers dereverberation was performed by exploiting the coherence between the two microphones. This approach was proposed

first by Allen et al. in [253]. Bloom [125] found that this dereverberation approach had no statistically significant effect on intelligibility, even though the measured average reverberation time and the perceived reverberation time were considerably reduced by the processing.

Until now there were no single microphone solutions for the suppression of reverberation, acoustic echo, and background noise.

7.1.3 Scope and organization

In this chapter we propose a complete single microphone system that is capable of suppressing late reverberation of the near-end speech signal, acoustic echo and background noise, with a small amount of speech distortion. We focus on the residual echo caused by under-modelling of the acoustic echo path. Detailed information can be found in Section 7.2.

In our approach the acoustic echo path is divided into three non-overlapping parts. The first part, which contains the direct path and a few early reflections, is cancelled using a regular adaptive filter. The length of this adaptive filter is relatively short, resulting in fast adaptation and tracking of the filter coefficients. The second and third parts of the acoustic echo path result in the residual echo signal, which is suppressed by the developed post-filter. The second part of the acoustic echo path contains a few early reflections and some late reflections. The residual echo signal related to the second part of the acoustic echo path is estimated using a second adaptive filter and is suppressed by the post-filter. To increase the robustness of the residual echo suppression, we will use the short-term PSD of the output of the second adaptive filter rather than the complete signal. An important room characteristic, namely the reverberation time, is obtained from the second adaptive filter. The tail part, i.e., third part, of the acoustic echo path has a much simpler and continuous structure than the previous parts. The short-term PSD of the residual echo related to the tail of the acoustic echo path is estimated using a statistical reverberation model and the estimated reverberation time. A detailed description of the two residual echo estimators and the estimation of the room characteristics can be found in Section 7.3.

In Section 7.4 we show how an estimate of the late reverberant spectral variance of the near-end speech signal is obtained using a statistical reverberation model. In this section we provide an alternative derivation of the late reverberant spectral variance estimator with so-called direct path compensation. The obtained estimator is equal to the estimator derived from the generalized statistical reverberation model developed in Section 6.3.2. The estimated late reverberant spectral variance is used by the post-filter to dereverberate the near-end speech signal.

A single post-filter is then applied to the error signal $e(n)$ to obtain an estimate of the dereverberated near-end speech signal. Detailed information about the post-filter is provided in Section 7.5.

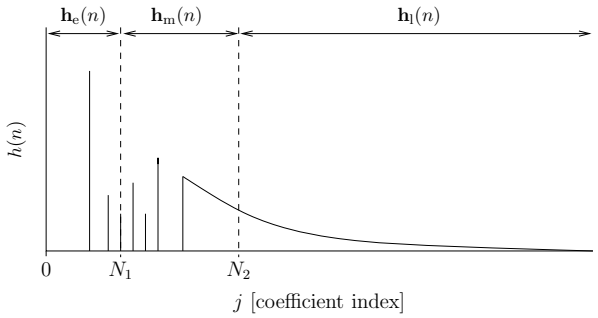


Figure 7.2 *Schematic representation of the acoustic echo path* $\mathbf{h}(n)$.

Experimental results are presented in Section 7.6. Finally, discussions and conclusions are presented in Sections 7.7 and 7.8, respectively.

7.2 Proposed Solution

We propose to divide the impulse response of the acoustic echo path $\mathbf{h}(n)$ into three parts, $\mathbf{h}_\mathrm{e}(n)$, $\mathbf{h}_\mathrm{m}(n)$ and $\mathbf{h}_\mathrm{l}(n)$ (see Fig. 7.2) such that

$$h_j(n) = \begin{cases} h_{\mathrm{e},j}(n), & 0 \leq j < N_1; \\ h_{\mathrm{m},j-N_1}(n), & N_1 \leq j < N_2; \\ h_{\mathrm{l},j-N_2}(n), & N_2 \leq j \leq L_\mathrm{h} - 1; \\ 0, & \text{otherwise}, \end{cases} \tag{7.7}$$

where j denotes the coefficient index, and L_h denotes the length of the acoustic echo path. The values N_1 and N_2, in samples, are chosen such that $\mathbf{h}_\mathrm{e}(n)$ consists of the direct path and a few early reflections, $\mathbf{h}_\mathrm{m}(n)$ consists of a few early and late reflections (i.e., mixed reflections), and $\mathbf{h}_\mathrm{l}(n)$ consists of all later reflections (i.e., late reverberation). N_1/f_s usually ranges from 32 to 64 ms, depending on the distance between the loudspeaker and the microphone, and N_2/f_s is usually larger than 128 ms.

The AIR from the near-end source to the microphone is divided into two parts (see Fig. 7.3) such that

$$a_j(n) = \begin{cases} a_{\mathrm{e},j}(n), & 0 \leq j < N_\mathrm{l}; \\ a_{\mathrm{l},j}(n), & N_\mathrm{l} \leq j \leq L_\mathrm{a} - 1; \\ 0, & \text{otherwise}, \end{cases} \tag{7.8}$$

where L_a denotes the length of the AIR, N_l is chosen such that $\mathbf{a}_\mathrm{e}(n)$ consists of the direct path and a few early reflections, and $\mathbf{a}_\mathrm{l}(n)$ consists of all later reflections, i.e., late reverberation. $N_\mathrm{l}/f_\mathrm{s}$ usually ranges from 40 to 80 ms.

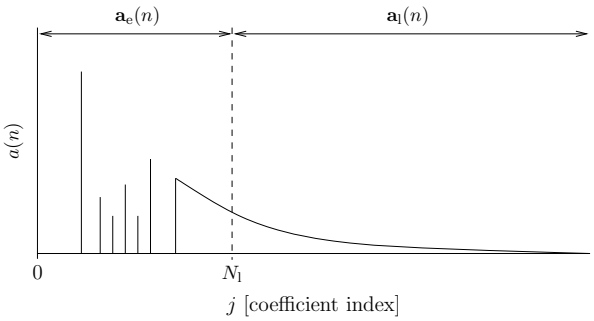


Figure 7.3 *Schematic representation of the acoustic impulse response* $\mathbf{a}(n)$.

The observed microphone signal is given by

$$\begin{aligned} y(n) &= z(n) + d(n) + v(n) \\ &= \sum_{j=0}^{N_{\mathrm{l}}-1} a_j(n)s(n-j) + \sum_{j=N_r}^{L_{\mathrm{a}}-1} a_j(n)s(n-j) + d(n) + v(n) \\ &= z_{\mathrm{e}}(n) + z_{\mathrm{l}}(n) + d(n) + v(n), \end{aligned} \tag{7.9}$$

where $z_{\mathrm{e}}(n)$ is the desired near-end speech component, and $z_{\mathrm{l}}(n)$ denotes the late reverberant near-end speech component. Note that $z_{\mathrm{e}}(n)$ is affected only by the early reflections. In Section 7.4 we will explain why it is sufficient to estimate $z_{\mathrm{e}}(n)$ rather than $s(n)$.

The proposed system with post-processor is depicted in Fig. 7.4. The Acoustic Echo Canceller (AEC) and post-processor are discussed in the following sub-sections. The developed algorithm is summarized in Alg. 1.

7.2.1 Acoustic Echo Cancellation (AEC)

An adaptive filter is used to cancel the echo signal related to the first part of the acoustic echo path, i.e., $\mathbf{h}_{\mathrm{e}}(n)$. It should be noted that one could introduce an artificial delay in the system to compensate for the delay introduced by the loudspeaker-microphone distance. This artificial delay can for example be determined in case the loudspeaker-microphone distance is known, which is often the case for hands-free telephones. By doing this the length of the adaptive filter can be reduced. As an example we use a standard Normalized Least Mean Square (NLMS) algorithm to estimate $\mathbf{h}_{\mathrm{e}}(n)$, where $\hat{\mathbf{h}}_{\mathrm{e}}(n) = [\hat{h}_{\mathrm{e},0}(n), \hat{h}_{\mathrm{e},1}(n), \ldots, \hat{h}_{\mathrm{e},L_{\mathrm{e}}-1}(n)]^T$. The length of the adaptive filter is $L_{\mathrm{e}} = N_{\mathrm{l}}$. The update equation for the NLMS algorithm is given by

$$\hat{\mathbf{h}}_{\mathrm{e}}(n+1) = \hat{\mathbf{h}}_{\mathrm{e}}(n) + \mu(n)\frac{\mathbf{x}(n)e(n)}{\mathbf{x}^T(n)\mathbf{x}(n) + \delta_{NLMS}}, \tag{7.10}$$

where $\mu(n)$ $(0 < \mu < 2)$ denotes the step-size, δ_{NLMS} $(\delta_{NLMS} > 0)$ the regularization factor, and $\mathbf{x}(n) = [x(n), \ldots, x(n - L_{\mathrm{e}} + 1)]^T$ the far-end speech signal. Note that

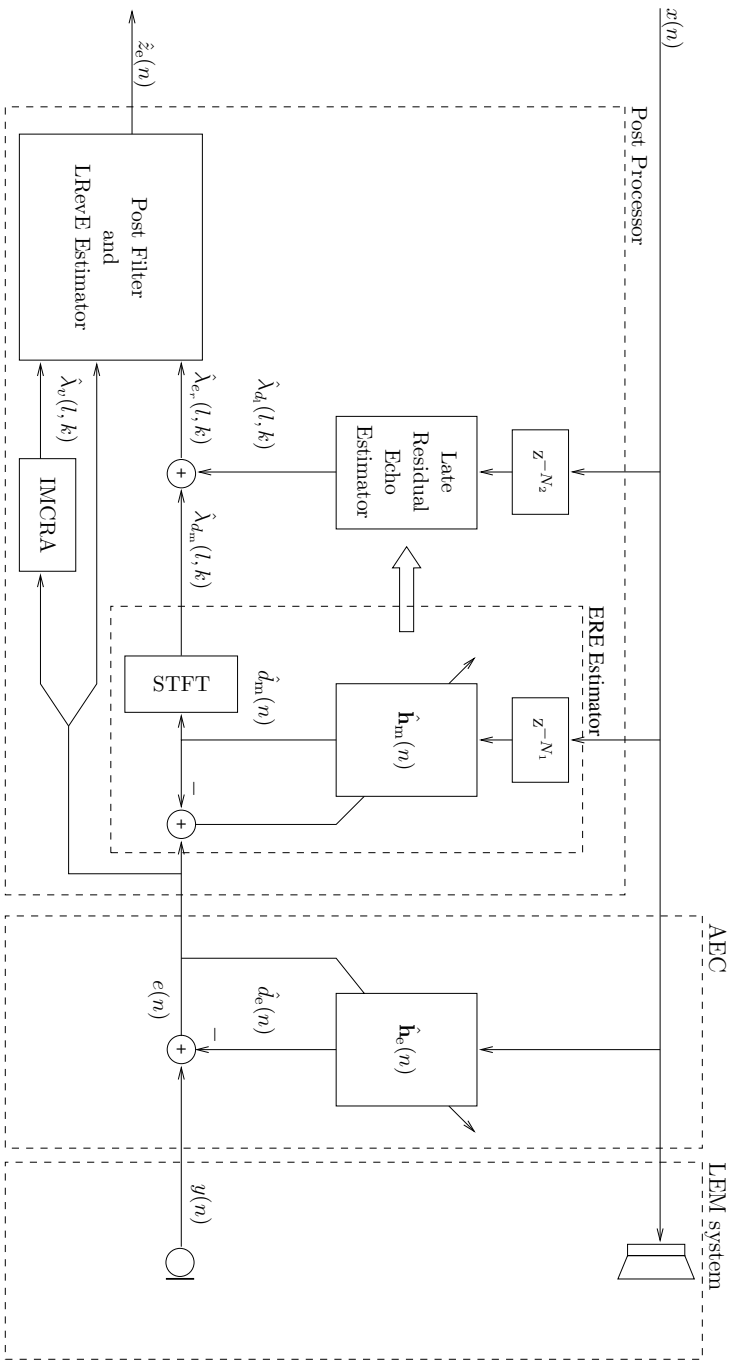


Figure 7.4 *Proposed acoustic echo canceller with post-processor. The structure of the post-filter and Late Reverberant Energy (LRevE) estimator can be found in Fig. 7.6.*

Algorithm 1 Summary of the developed algorithm.

1) **Acoustic Echo Cancellation**: Update the adaptive filter $\hat{h}_{\mathrm{e}}(n)$ and calculate $\hat{d}_{\mathrm{e}}(n)$.

2) **Estimate Early Residual Echo**: Update the adaptive filter $\hat{h}_{\mathrm{m}}(n)$ and calculate $\hat{d}_{\mathrm{m}}(n)$.

3) **Estimate Reverberation Time**: Estimate $\mathrm{RT}_{60}(n)$ using Eq. 7.22.

4) **Calculate STFT transform**: Calculate the STFT of $e(n) = y(n) - \hat{d}_{\mathrm{e}}(n)$, $\hat{d}_{\mathrm{m}}(n)$, and $x(n)$.

5) **Estimate Background Noise**: Estimate $\lambda_v(l,k)$ using [190].

6) **Estimate Late Residual Echo**: Calculate $\tilde{c}(l,k)$ using (7.33) and $\hat{\lambda}_{d_{\mathrm{l}}}(l,k)$ using (7.35).

7) **Estimate Late Reverberant Energy**: Calculate $G_{\mathrm{SP}}(l,k)$ using Eq. 7.38-7.41. Estimate $\lambda_z(l,k)$ using Eq. 7.42, and calculate $\hat{\lambda}_{z_{\mathrm{l}}}(l,k)$ using Eq. 7.51 and 7.53.

8) **Perform Post-Filtering**:

   (a) Calculate the *a posteriori* SIR using Eq. 7.58 and the individual *a priori* SIR using Eq. 7.72-7.73 with $\vartheta \in \{z_{\mathrm{l}}, e_r, v\}$, the total *a priori* Signal to Interference Ratio (SIR) can then be calculated using Eq. 7.74-7.75.

   (b) Calculate the speech presence probability using Eq. 7.61.

   (c) Calculate the gain function $G_{\mathrm{OM-LSA}}(l,k)$ using Eq. 7.56, Eq. 7.67, Eq. 7.61, and Eq. 7.63.

   (d) Calculate $\hat{Z}_{\mathrm{e}}(l,k)$ using Eq. 7.15.

9) **Calculate inverse STFT transform**: Calculate the output $\hat{z}_{\mathrm{e}}(n)$ by applying the inverse STFT to $\hat{Z}_{\mathrm{e}}(l,k)$.

other, more advanced, algorithms can be used, i.g., Recursive Least Squares (RLS) or Affine Projection (AP), see for example [239] and the references therein. Since the first part of the acoustic echo path is sparse, one might use the Improved Proportionate NLMS (IPNLMS) algorithm proposed by Benesty and Gay [254].

Double-talk occurs during periods when the far-end speaker and the near-end speaker are simultaneously talking and can seriously affect the convergence and tracking ability of the adaptive filter. Double-talk detectors and optimal step-size control methods have been proposed to alleviate this problem [255, 239, 256]. These methods are

beyond the scope of this chapter. Here we choose the step-size $\mu(n)$ as follows,

$$\mu(n) = \begin{cases} 0, & \text{during double-talk;} \\ \mu, & \text{otherwise,} \end{cases} \tag{7.11}$$

where the double-talk periods are detected manually.

The estimated echo signal can be calculated using

$$\hat{d}_{\mathrm{e}}(n) = \sum_{j=0}^{L_{\mathrm{e}}-1} \hat{h}_{\mathrm{e},j}(n) x(n-j). \tag{7.12}$$

7.2.2 Post-Processor

The post-processor contains four estimators and a post-filter. Two estimators are related to the residual echo, i.e., early residual echo and late residual echo (see Section 7.3). The third estimator is used to estimate the noise spectral variance. We use the Improved Minima Controlled Recursive Averaging (IMCRA) algorithm proposed by Cohen [190] to obtain an estimate of the noise spectral variance in each time frame and frequency bin directly from $e(n)$. The fourth estimator (see Section 7.4) is related to the late reverberant energy of the near-end speech signal, and requires the estimated short-term PSD of the residual echo and background noise. This estimate is used to dereverberate the near-end speech signal $z(n)$.

The post-processor also contains a single post-filter, which is applied to the error signal $e(n)$, to obtain an estimate of the near-end speech component $z_{\mathrm{e}}(n)$.

Using Eq. 7.9 we can rewrite the error signal, defined in Eq. 7.3, as

$$e(n) = z_{\mathrm{e}}(n) + z_{\mathrm{l}}(n) + e_r(n) + v(n). \tag{7.13}$$

Using the short-time Fourier transform (STFT), we have in the time-frequency domain

$$E(l,k) = Z_{\mathrm{e}}(l,k) + Z_{\mathrm{l}}(l,k) + E_r(l,k) + V(l,k), \tag{7.14}$$

where k represents the frequency bin, and l the time frame.

The spectral speech component $\hat{Z}_{\mathrm{e}}(l,k)$ is estimated by applying a spectral gain function $G_{\mathrm{OM-LSA}}$, see Section 7.5, to each spectral component $E(l,k)$, i.e.,

$$\hat{Z}_{\mathrm{e}}(l,k) = G_{\mathrm{OM-LSA}}(l,k)\, E(l,k). \tag{7.15}$$

The dereverberated near-end speech signal $\hat{z}_{\mathrm{e}}(n)$ can be obtained using the inverse STFT and the weighted overlap-add method.

7.3 Residual Echo Estimation

The residual echo signal is divided into two parts, which are related to $\mathbf{h}_{\mathrm{m}}(n)$ and $\mathbf{h}_{\mathrm{l}}(n)$, respectively. In Fig. 7.5 a typical Acoustic Impulse Response and its Energy Decay Curve (EDC) are depicted. The EDC is obtained by backward integration of the squared AIR, as proposed by Schroeder in [76], and is normalized with respect to the total energy of the AIR. The second part of the acoustic echo path, i.e., $\mathbf{h}_{\mathrm{m}}(n)$, consists of early reflections and a few late reflections. This part is estimated using a second adaptive filter. It should be noted that early reflections appear as separate delayed impulses in the acoustic echo path, whilst late reflections appear as a continuum. The short-term PSD of the output of the adaptive filter is used to suppress the corresponding residual echo signal. The tail, i.e., third part, of the acoustic echo path is related to $\mathbf{h}_{\mathrm{l}}(n)$ and has a much simpler structure than the second part.

Small position changes and movements in the room can have a large influence on the acoustic echo path. These momentary mismatches between the exact and estimated acoustic echo path can result in speech distortions during double-talk and a momentary increase of residual echo. However, the power spectrum of the acoustic echo path is less sensitive to small position changes and movements in the room than the complete acoustic echo path. Therefore, we propose to use the short-term PSD of the output of the second adaptive filter to suppress the related residual echo.

In the following sub-sections we describe how the short-term PSDs of the residual echo, related to $\mathbf{h}_{\mathrm{m}}(n)$ and $\mathbf{h}_{\mathrm{l}}(n)$, are estimated.

7.3.1 Early Residual Echo (ERE)

A second adaptive filter of length $L_{\mathrm{m}} = N_2 - N_1$ is denoted by $\hat{\mathbf{h}}_{\mathrm{m}}(n)$, and is used to estimate $\mathbf{h}_{\mathrm{m}}(n)$. As an example we used a standard NLMS algorithm. Since this adaptive filter is used to estimate a distinct part of the acoustic echo path, which has different characteristics than $\mathbf{h}_{\mathrm{e}}$, the filter coefficient updates might be chosen differently. It should be noted that some *a priori* knowledge of the AIR can be used to increase the robustness and the convergence, see for example [257]. Also note that the second adaptive filter is placed in series with the first adaptive filter. Compared to a parallel solution or equivalent partitioned adaptive filters we have two separate error signals such that the step-size can be controlled separately, e.g., depending on the Signal to Noise Ratio (SNR). The estimated echo signal $\hat{d}_{\mathrm{m}}(n)$ is determined by

$$\hat{d}_{\mathrm{m}}(n) = \sum_{j=0}^{L_{\mathrm{m}}-1} \hat{h}_{\mathrm{m},j}(n) x_{\mathrm{m}}(n-j), \tag{7.16}$$

where $x_{\mathrm{m}}(n) = x(n - N_1)$.

The STFT of $\hat{d}_{\mathrm{m}}(n)$ is defined as

$$\hat{D}_{\mathrm{m}}(l,k) = \sum_{n=0}^{L_w-1} w(n)\hat{d}_{\mathrm{m}}(mR+n)\mathrm{e}^{-\iota\frac{2\pi k}{K}n} \quad \text{for } k = \{0,\ldots,K-1\}, \tag{7.17}$$

where $\iota = \sqrt{-1}$, $w(n)$ is the analysis window, e.g., a Hamming window, with length L_w, K represents the length of the DFT, and R denotes the frame rate, which is defined as the length of the analysis window minus the length of the window overlap. The estimated short-term PSD related to the early residual echo is given by

$$\hat{\lambda}_{d_{\mathrm{m}}}(l,k) = \left|\hat{D}_{\mathrm{m}}(l,k)\right|^2. \tag{7.18}$$

This estimate will be used in the post-filter to suppress the early residual echo.

7.3.2 Late Residual Echo (LRE)

In [244] Enzner proposed a recursive estimator for the short-term PSD of the late residual echo. In Fig. 7.5 it can clearly be seen that the tail of the AIR exhibits an exponential decay, and the tail of the EDC exhibits a linear decay. The recursive estimator exploits the fact that the exponential decay rate is directly related to the reverberation time of the room, which can be estimated using the estimated echo path. Additionally, the estimator requires a second parameter that specifies the initial power of the late residual echo.

In this section we derive an essentially equivalent recursive estimator starting in the time-domain rather than directly in the frequency domain as in [244]. Enzner used the entire estimated echo path to estimate the required parameters, viz., the reverberation time and the initial power, which are both assumed to be frequency independent. It should, however, be noted that these parameters are usually frequency dependent [41]. Furthermore, in many applications the distance between the loudspeaker and the microphone is small, which results in a strong direct echo. Since the presence of either a delay or a strong direct echo results in an erroneous estimate of both the reverberation time and the initial power (c.f. [258]) we propose to use the filter coefficients of the second adaptive filter to estimate the reverberation time. To improve the regularity of the decay ramp, and thus the estimation, we apply a linear curve fit to the smoothed decay envelope, i.e., the EDC, rather than a direct fit to the log-envelope of the estimated echo path as proposed in [244].

Note that we can assume that the reverberation time in the room is independent of the position in the room [41]. Therefore, we can use the estimated reverberation time for the dereverberation of the near-end speech signal.

Using a statistical reverberation model and the estimated reverberation time we can estimate the spectral variance of the residual echo related to $\mathbf{h}_{\mathrm{l}}(n)$. First, we determine the reverberation time using the filter coefficients of $\hat{\mathbf{h}}_{\mathrm{m}}(n)$. Second, we estimate

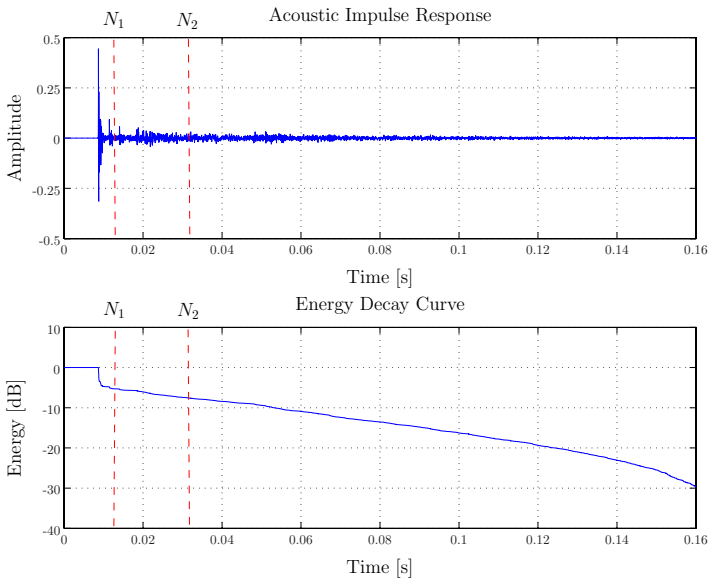


Figure 7.5 *Typical Acoustic Impulse Response and related Energy Decay Curve.*

the spectral variance of the residual echo related to $\hat{\mathbf{h}}_{\mathrm{l}}(n)$ based on a statistical reverberation model and the estimated reverberation time.

Parameter Estimation

We propose to estimate the reverberation time directly from $\hat{\mathbf{h}}_{\mathrm{m}}(n)$. Similar to Schroeder's method (see Section 2.8 for more details) we first calculate the EDC of $\hat{\mathbf{h}}_{\mathrm{m}}(n)$. Secondly, a straight line is fitted to part of the EDC values to obtain the slope of the EDC. It should be noted that the last EDC values are not useful due to the finite length of $\hat{\mathbf{h}}_{\mathrm{m}}(n)$ and due to the final misalignment of the adaptive filter coefficients. Therefore, we use only a dynamic rage of 20 dB[1] to determine the slope of the EDC. Finally, the reverberation time is updated using an adaptive scheme. A detailed description can be found in Alg. 2.

In general, the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air [41]. Instead of applying the above procedure to $\hat{\mathbf{h}}_{\mathrm{m}}(n)$, we can apply the above procedure to a band-pass filtered version of $\hat{\mathbf{h}}_{\mathrm{m}}(n)$. We used six 1-octave band filters to acquire a higher frequency resolution. The six reverberation time values are interpolated and extrapolated to obtain an estimate of $\widehat{\mathrm{RT}}_{60}(u,k)$ for each frequency bin, where $u \triangleq nR_{\mathrm{EDC}}$ and R_{EDC} denotes the estimation rate of the EDC. In the sequel the time index u is omitted for simplification.

[1] It might be necessary to decrease the dynamic range when L_{m} is small or the reverberation time is long.

Algorithm 2 Estimation of the reverberation time using $\hat{\mathbf{h}}_{\mathrm{m}}(n)$.

1) Calculate the Energy Decay Curve of $\hat{\mathbf{h}}_{\mathrm{m}}(u)$, where u equals nR_{EDC} and R_{EDC} denotes the estimation rate, using

$$\mathrm{EDC}(u,n) = 20\log_{10}\left(\sum_{j=n}^{L_{\mathrm{m}}-1}\left(\hat{h}_{\mathrm{m},j}(u)\right)^2\right) \quad \text{for } 0 \le n \le L_{\mathrm{m}}-1.$$

2) A straight line is fitted to part of the EDC data points, using a least squares approach. The regression coefficient at time u, denoted by $q(u)$, of the line is obtained by minimizing the following cost function:

$$J(p(u),q(u)) = \sum_{n=n_s}^{n_e}\left(\mathrm{EDC}(u,n) - (p(u)+q(u)\,n)\right)^2, \tag{7.19}$$

where n_s $(0 \le n_s < L_{\mathrm{m}}-1)$ and n_e $(n_s < n_e \le L_{\mathrm{m}}-1)$ denote the start-time and end-time of EDC values that are used, respectively. A good choice for n_s and n_e is given by

$$n_s = \arg\min_n \left|\frac{\mathrm{EDC}(u,n)}{\mathrm{EDC}(u,0)} + 5\right| \tag{7.20}$$

and

$$n_e = \arg\min_n \left|\frac{\mathrm{EDC}(u,n)}{\mathrm{EDC}(u,0)} + 25\right|, \tag{7.21}$$

respectively.

3) The reverberation time $\widehat{\mathrm{RT}}_{60}(u)$ can now be calculated using

$$\widehat{\mathrm{RT}}_{60}(u) = \widehat{\mathrm{RT}}_{60}(u-1) + \mu_{\mathrm{RT}_{60}}\left(\frac{60}{q(u)f_{\mathrm{s}}} - \widehat{\mathrm{RT}}_{60}(u-1)\right), \tag{7.22}$$

where $\mu_{\mathrm{RT}_{60}}$ denotes the adaptation step-size.

Late Residual Echo Estimation

The remaining late residual error due to under-modelling of $\mathbf{h}(n)$ is related to the tail of the acoustic echo path. In the sequel we assume that $L_{\mathrm{h}} = \infty$. The late residual echo $d_{\mathrm{l}}(n)$ is given by

$$d_{\mathrm{l}}(n) = \sum_{j=0}^{\infty} h_{\mathrm{l},j}(n)x_{\mathrm{l}}(n-j), \tag{7.23}$$

where $x_{\mathrm{l}}(n) = x(n-N_2)$.

The short-term PSD of $d_{\mathrm{l}}(n)$ is defined as

$$\lambda_{d_{\mathrm{l}}}(l,k) \triangleq \mathcal{E}\left\{|D_{\mathrm{l}}(l,k)|^2\right\}. \tag{7.24}$$

In the STFT domain we can approximate $D_{\mathrm{l}}(l,k)$ by

$$D_{\mathrm{l}}(l,k) \approx \sum_{i=0}^{\infty} H_{\mathrm{l},i}(l,k)\; X\left(l-i-\frac{N_2}{R},k\right), \tag{7.25}$$

where $H_{\mathrm{l},i}(l,k)$ is related to the STFT of $\mathbf{h}_{\mathrm{l}}(mR)$, and i denotes the frame index of $\mathbf{H}_{\mathrm{l}}(l,k)$. Note that N_2 should be chosen such that N_2/R is an integer value.

Using Eq. 7.24, Eq. 7.25, and the assumption that

$$\mathcal{E}\{H_{\mathrm{l},i}(l,k)H_{\mathrm{l},i+\tau}(l,k)\} = 0 \;\; \forall \tau \neq 0 \; \forall l, \tag{7.26}$$

we can express $\lambda_{d_{\mathrm{l}}}(l,k)$ as

$$\begin{aligned}\lambda_{d_{\mathrm{l}}}(l,k) &\approx \mathcal{E}\left\{\sum_{i=0}^{\infty}|H_{\mathrm{l},i}(l,k)|^2 \;\left|X\left(l-i-\frac{N_2}{R},k\right)\right|^2\right\}\\ &\approx \sum_{i=0}^{\infty}\mathcal{E}\left\{|H_{\mathrm{l},i}(l,k)|^2\right\}\;\mathcal{E}\left\{\left|X\left(l-i-\frac{N_2}{R},k\right)\right|^2\right\}.\end{aligned} \tag{7.27}$$

The energy envelope of $H_{\mathrm{l},i}(l,k)$ is given by

$$\mathcal{E}\{|H_{\mathrm{l},i}(l,k)|^2\} = c(l-i,k)\;\mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}i}, \tag{7.28}$$

where $c(l-i,k)$ denotes the initial power in the k^{th} sub-band at time $(l-i)R$. The exponential decay rate $\bar{\delta}(k)$ is related to the frequency dependent reverberation time $\mathrm{RT}_{60}(k)$ through

$$\bar{\delta}(k) \triangleq \frac{3\ln(10)}{\mathrm{RT}_{60}(k)}. \tag{7.29}$$

Using Eq. 7.28, and the fact that $\lambda_x(l,k) = \mathcal{E}\{|X(l,k)|^2\}$, we can rewrite Eq. 7.27 as

$$\begin{aligned}\lambda_{d_{\mathrm{l}}}(l,k) &\approx \sum_{i=0}^{\infty} c(l-i,k)\;\mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}i}\;\lambda_x\left(l-i-\frac{N_2}{R},k\right)\\ &\approx \sum_{i'=-\infty}^{l} c(i',k)\;\mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}(l-i')}\;\lambda_x\left(i'-\frac{N_2}{R},k\right)\\ &\approx \mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}}\lambda_{d_{\mathrm{l}}}(l-1,k) + c(l,k)\;\lambda_x\left(l-\frac{N_2}{R},k\right).\end{aligned} \tag{7.30}$$

The initial power $c(l,k)$ can be calculated using the following expression

$$c(l,k) = \left|\sum_{j=0}^{L_w-1}\hat{h}_{\mathrm{l},j}(lR)\mathrm{e}^{-\iota\frac{2\pi k}{K}j}\right|^2 \quad \text{for } k = \{0,\ldots,K-1\}. \tag{7.31}$$

Since $\hat{\mathbf{h}}_{\mathrm{l}}(n)$ is not available, we use the last L_w coefficients of $\hat{\mathbf{h}}_{\mathrm{m}}(n)$ and extrapolate the energy using the estimated decay. We then obtain an estimate of $c(l,k)$ by

$$\hat{c}(l,k) = \mathrm{e}^{-2\bar{\delta}(k)\frac{L_w}{f_s}} \left| \sum_{j=0}^{L_w-1} \hat{h}_{\mathrm{m},L_{\mathrm{m}}-L_w+j}(mR)\mathrm{e}^{-\iota\frac{2\pi k}{K}j} \right|^2 \quad \text{for } k = \{0,\ldots,K-1\}. \tag{7.32}$$

The estimated initial power $\hat{c}(l,k)$ might contain some spectral zeros, which can easily be removed by smoothing $\hat{c}(l,k)$ along the frequency axis using

$$\tilde{c}(l,k) = \sum_{i=-w}^{w} b_i \hat{c}(l,k+i), \tag{7.33}$$

where b is a normalized window function ($\sum_{i=-w}^{w} b_i = 1$) that determines the frequency smoothing. The initial power $\tilde{c}(l,k)$ can then be used in (7.35) to estimate $\lambda_{d_{\mathrm{l}}}(l,k)$. Note that $\tilde{c}(l,k)$ can be updated at a lower rate to reduce the complexity of the late residual echo estimator.

An alternative method to estimate the initial power $c(l,k)$ is to minimize

$$\mathcal{E}\left\{\left(|E(l,k) - \hat{D}_m(l,k)|^2 - \hat{\lambda}_{d_{\mathrm{l}}}(l,k;c(l,k))\right)^2\right\}, \tag{7.34}$$

with respect to $c(l,k)$.

Using the estimated reverberation time $\widehat{\mathrm{RT}}_{60}(k)$ and Eq. 7.29 we obtain an estimate of the exponential decay rate $\bar{\delta}(k)$. Using the initial power $\tilde{c}(l,k)$ we can now estimate the short-term PSD $\lambda_{d_{\mathrm{l}}}(l,k)$ using

$$\hat{\lambda}_{d_{\mathrm{l}}}(l,k) = \mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_s}} \hat{\lambda}_{d_{\mathrm{l}}}(l-1,k) + \tilde{c}(l,k)\ \hat{\lambda}_x\left(l - \frac{N_2}{R}, k\right), \tag{7.35}$$

where $\hat{\lambda}_x(l,k)$ can be calculated using

$$\hat{\lambda}_x(l,k) = \eta_x \hat{\lambda}_x(l-1,k) + (1-\eta_x)|X(l,k)|^2, \tag{7.36}$$

where η_x ($0 \leq \eta_x < 1$) denotes the smoothing parameter.

7.4 Late Reverberant Energy Estimation

In this section we explain how the Late Reverberant Energy (LRevE) of the near-end speech signal $z(n)$ is estimated. Suppression of the late reverberant energy will partially dereverberate the near-end speech signal. Since the first part of the acoustic impulse response $\mathbf{a}(n)$, i.e., $\mathbf{a}_{\mathrm{e}}(n)$, remains unaltered we do not equalize the spectral colouration caused by the early reflections. As discussed in Section 1.3 we can increase the speech quality and intelligibility by reducing the energy related to the late

reflections. It should be noted that early reflections may even contribute to the speech quality and intelligibility.

There are two main issues that have to be dealt with. First, we require an estimate of the spectral variance of the reverberant signal $Z(l,k)$ for the estimation of the late reverberant energy (Section 7.4.1). Second, we need to compensate for the energy contribution of the direct path, as explained in Section 7.4.2.

7.4.1 Reverberant Energy Estimation

The spectral variance of the reverberant spectral component $Z(l,k)$, i.e., $\lambda_z(l,k)$, is estimated by minimizing

$$\mathcal{E}\left\{\left(A(l,k)-|\hat{Z}(l,k)|^2\right)^2\right\}, \tag{7.37}$$

where $A(l,k)=|Z(l,k)|^2$ and $\hat{Z}(l,k)=G_{\mathrm{SP}}(l,k)E(l,k)$.

As shown in [234] this leads to the following spectral gain function

$$G_{\mathrm{SP}}(l,k)=\sqrt{\frac{\xi_{\mathrm{SP}}(l,k)}{1+\xi_{\mathrm{SP}}(l,k)}\left(\frac{1}{\gamma_{\mathrm{SP}}(l,k)}+\frac{\xi_{\mathrm{SP}}(l,k)}{1+\xi_{\mathrm{SP}}(l,k)}\right)}, \tag{7.38}$$

where

$$\xi_{\mathrm{SP}}(l,k)=\frac{\lambda_z(l,k)}{\lambda_{e_r}(l,k)+\lambda_v(l,k)} \tag{7.39}$$

and

$$\gamma_{\mathrm{SP}}(l,k)=\frac{|E(l,k)|^2}{\lambda_{e_r}(l,k)+\lambda_v(l,k)}, \tag{7.40}$$

denote the *a priori* and *a posteriori* Signal to Interference Ratios (SIR), respectively. The *a priori* SIR is estimated using the decision-directed estimator proposed by Ephraim and Malah [210]. Estimates of the spectral variance of the noise in the error signal $e(n)$, i.e., $\lambda_v(l,k)$, are obtained using the IMCRA approach [190]. An estimate of the residual echo spectral variance is given by

$$\hat{\lambda}_{e_r}(l,k)=\hat{\lambda}_{d_{\mathrm{m}}}(l,k)+\hat{\lambda}_{d_{\mathrm{l}}}(l,k). \tag{7.41}$$

An estimate of the power spectrum of the reverberant signal $z(n)$ is then obtained by:

$$\hat{A}(l,k)=\left(G_{\mathrm{SP}}(l,k)\right)^2|E(l,k)|^2. \tag{7.42}$$

The power spectrum spectral $\hat{A}(l,k)$ is then smoothed over time using a first-order low-pass filter, as described in Chapter 6, Section 6.4, to obtain an estimate of the late reverberant spectral variance $\lambda_{z_{\mathrm{l}}}(l,k)$.

7.4.2 Direct Path Compensation

In Section 6.4 we have shown that, using Polack's statistical room impulse response model [63], the late reverberant spectral variance can be estimated directly from the spectral variance of the reverberant signal using

$$\hat{\lambda}_{z_{\mathrm{l}}}(l,k) = \alpha^{\frac{N_{\mathrm{l}}}{R}}(k)\hat{\lambda}_z\left(l - \frac{N_{\mathrm{l}}}{R}, k\right), \tag{7.43}$$

where $\alpha(k) = \mathrm{e}^{-2\bar{\delta}(k)\frac{R}{f_{\mathrm{s}}}}$ $(0 \leq \alpha(k) < 1)$ and $\bar{\delta}(k)$ is given by Eq. 7.29. The value N_{l} should be chosen such that $\frac{N_{\mathrm{l}}}{R}$ is an integer value, where R denotes the frame rate of the STFT.

In Section 6.3 we have shown that the late reverberant spectral variance estimator that was derived using Polack's statistical reverberation model can only be used when the energy of the direct path is small compared to the reverberant energy. However, in many practical situations, the source is close to the microphone, and the contribution of the energy related to the direct path cannot be neglected. To alleviate this problem we have developed a novel estimator based on a generalized statistical model (see Section 6.4). We will now present an alternative derivation of this estimator to compensate for the energy related to the direct path.

The energy envelope of the acoustic impulse response in the k^{th} sub-band can be modelled as

$$\tilde{A}_k(\mathrm{z}) = E_{\mathrm{d}}(k) + E_{\mathrm{r}}(k)\tilde{R}_k(\mathrm{z}), \tag{7.44}$$

where $E_{\mathrm{d}}(k)$ and $E_{\mathrm{r}}(k)$ are related to the amount of direct and reverberant energy in the k^{th} sub-band, respectively, and $\tilde{R}_k(\mathrm{z})$ denotes the normalized energy envelope of the reverberant part of the acoustic impulse response, which starts at $l = 1$, i.e.,

$$\tilde{R}_k(\mathrm{z}) = \frac{1-\alpha(k)}{\alpha(k)} \sum_{m=1}^{\infty} (\alpha(k))^m \, \mathrm{z}^{-m}, \tag{7.45}$$

such that

$$\frac{1-\alpha(k)}{\alpha(k)} \sum_{m=1}^{\infty} (\alpha(k))^m = 1. \tag{7.46}$$

By expanding the series in Eq. 7.45 we obtain

$$\tilde{R}_k(\mathrm{z}) = \frac{1-\alpha(k)}{\alpha(k)} \frac{\alpha(k)\mathrm{z}^{-1}}{1-\alpha(k)\mathrm{z}^{-1}}. \tag{7.47}$$

To eliminate the contribution of the energy of the direct path in $\hat{\lambda}_z(l,k)$, we propose to apply the following filter to $\hat{\lambda}_z(l,k)$,

$$F_k(\mathrm{z}) = \frac{E_{\mathrm{r}}(k)\tilde{R}_k(\mathrm{z})}{E_{\mathrm{d}}(k) + E_{\mathrm{r}}(k)\tilde{R}_k(\mathrm{z})}. \tag{7.48}$$

We now define $\kappa(k)$, which is inversely proportional to the direct to reverberant energy ratio in the k^{th} sub-band, as

$$\kappa(k) \triangleq \frac{1-\alpha(k)}{\alpha(k)} \frac{E_{\text{r}}(k)}{E_{\text{d}}(k)}. \tag{7.49}$$

Using the normalized energy envelope $\tilde{R}_k(\text{z})$, as defined in Eq. 7.47, Eq. 7.48 and Eq. 7.49 we obtain

$$F_k(\text{z}) = \frac{\alpha(k)\kappa(k)\text{z}^{-1}}{1-\alpha(k)\left(1-\kappa(k)\right)\text{z}^{-1}}. \tag{7.50}$$

Using the difference equation related to the filter in Eq. 7.50 we obtain an estimate of the reverberant energy with compensation of the direct path energy, i.e.,

$$\hat{\lambda}'_z(l,k) = \alpha(k)\left(1-\kappa(k)\right)\hat{\lambda}'_z(l-1,k) + \alpha(k)\kappa(k)\hat{\lambda}_z(l-1,k). \tag{7.51}$$

We require that $0 < \kappa(k) \leq 1$ to ensure the stability of the filter since $\alpha(k)\left(1-\kappa(k)\right)$ should always be larger than, or equal to, zero. This requirement is also reasonable from a physical point of view since only the source can increase reverberation energy in the room, i.e., the contribution of $\hat{\lambda}'_z(l-1,k)$ to $\hat{\lambda}'_z(l,k)$ should always by smaller than, or equal to, $\alpha(k)$. In case $E_{\text{d}}(k) \gg E_{\text{r}}(k)$, i.e., $\kappa(k)$ is small, $\hat{\lambda}'_z(l,k)$ mainly depends on $\alpha(k)\hat{\lambda}'_z(l-1,k)$. In case $E_{\text{d}}(k) \ll E_{\text{r}}(k)$ we reach the upper bound of $\kappa(k)$, i.e., $\kappa(k)=1$, and $\hat{\lambda}'_z(l,k)$ is equal to

$$\hat{\lambda}'_z(l,k) = \alpha(k)\hat{\lambda}_z(l-1,k). \tag{7.52}$$

We now replace $\hat{\lambda}_z(l,k)$ in Eq. 7.43 by the spectral variance with direct path compensation, i.e., $\hat{\lambda}'_z(l,k)$, to obtain the late reverberant energy $\hat{\lambda}_{z_{\text{l}}}(l,k)$, i.e.,

$$\hat{\lambda}_{z_{\text{l}}}(l,k) = \alpha^{\frac{N_{\text{l}}}{R}-1}(k)\hat{\lambda}'_z\left(l-\frac{N_{\text{l}}}{R}+1,k\right). \tag{7.53}$$

In this chapter we assume that $\kappa(k)$ is known *a priori*. In practice $\kappa(k)$ could be estimated adaptively using the method proposed in Section 6.6.

7.5 Post-Filter

At this point we have obtained estimates of the spectral variance of all interferences, i.e., late reverberation of the near-end speech signal, residual echo, and background noise. In this section we develop a single post-filter, i.e., spectral gain function, based on the Optimally-Modified Log Spectral Amplitude (OM-LSA) estimator proposed by Cohen [228]. The spectral gain function is applied to the spectrum of the error signal $e(n)$, in order to suppress all interferences. The spectral gain function, which depends on both time and frequency, is a function of the *a posteriori* and *a priori* SIRs. The *a posteriori* SIRs can be estimated directly given the 'noisy' observation and an estimate of the spectral variance of each interference. The estimation of the *a priori* SIR is slightly more complicated and is discussed in Section 7.5.2.

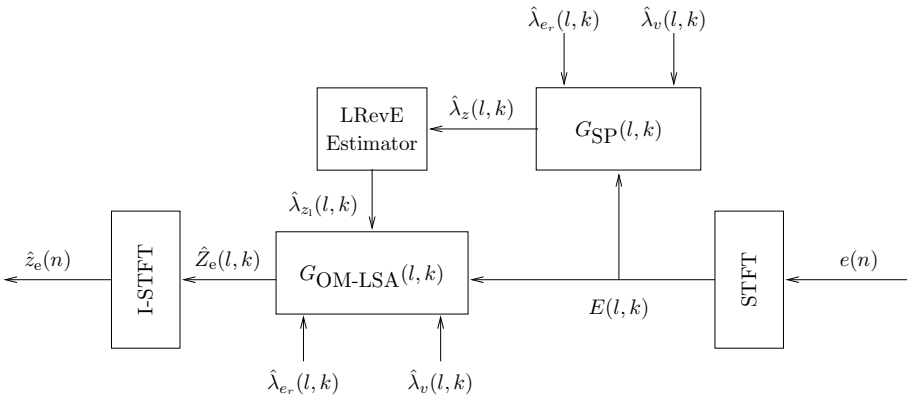


Figure 7.6 *Post-Filter and Late Reverberant Energy (LRevE) Estimator.*

7.5.1 Modified OM-LSA Estimator

We used a modified version of the Optimally-Modified Log Spectral Amplitude estimator to obtain an estimate of the desired spectral component $Z_{\text{e}}(l,k)$. In Appendix B we have developed a similar modification for one non-stationary and one stationary interference. In this chapter we extend this idea to three interferences and apply it to a specific application. The Log Spectral Amplitude (LSA) estimator proposed by Ephraim and Malah [205] minimizes

$$\mathcal{E}\left\{\left(\log(A(l,k)) - \log(\hat{A}(l,k))\right)^2\right\}, \tag{7.54}$$

where $A(l,k) = |Z_{\text{e}}(l,k)|$ denotes the spectral speech amplitude, and $\hat{A}(l,k)$ is its optimal estimator. Assuming spectral coefficients are conditionally independent given their variances [208], the LSA estimator is defined as

$$\hat{A}(l,k) = \exp\left(\mathcal{E}\{\log(A(l,k))|E(l,k)\}\right). \tag{7.55}$$

The LSA gain function is given by

$$G_{\text{LSA}}(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)} \exp\left(\frac{1}{2}\int_{\zeta(l,k)}^{\infty} \frac{\text{e}^{-t}}{t}\,\text{d}t\right), \tag{7.56}$$

where

$$\xi(l,k) = \frac{\lambda_{z_{\text{e}}}(l,k)}{\lambda_{z_1}(l,k) + \lambda_{e_r}(l,k) + \lambda_v(l,k)}, \tag{7.57}$$

$$\gamma(l,k) = \frac{|E(l,k)|^2}{\lambda_{z_1}(l,k) + \lambda_{e_r}(l,k) + \lambda_v(l,k)}, \tag{7.58}$$

and

$$\zeta(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\,\gamma(l,k). \tag{7.59}$$

The OM-LSA spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty [228]. Given two hypotheses, $\mathrm{H}_0(l,k)$ and $\mathrm{H}_1(l,k)$, which indicate speech absence and speech presence, respectively, we have

$$\begin{aligned}\mathrm{H}_0(l,k) &: E(l,k) = Z_{\mathrm{l}}(l,k) + E_r(l,k) + V(l,k),\\ \mathrm{H}_1(l,k) &: E(l,k) = Z_{\mathrm{e}}(l,k) + Z_{\mathrm{l}}(l,k) + E_r(l,k) + V(l,k),\end{aligned} \tag{7.60}$$

where $E_r(l,k) = D_{\mathrm{m}}(l,k) + D_{\mathrm{l}}(l,k)$. Based on a Gaussian statistical model, the speech presence probability is given by

$$p(l,k) = \left\{1 + \frac{q(l,k)}{1-q(l,k)}\left(1+\xi(l,k)\right)\exp\left(-\zeta(l,k)\right)\right\}^{-1}, \tag{7.61}$$

where $q(l,k)$ is the *a priori* signal absence probability. In [228] an efficient estimator for $q(l,k)$ is proposed. This estimator uses a soft-decision approach to compute three parameters, i.e., $P_{\mathrm{local}}(l,k)$, $P_{\mathrm{global}}(l,k)$, and $P_{\mathrm{frame}}(l)$, which and are based on the time-frequency distribution of the estimated *a priori* SIR, $\xi(l,k)$. These parameters exploit the strong correlation of speech presence in neighbouring frequency bins of consecutive frames. The estimated *a priori* signal absence probability $\hat{q}(l,k)$ is then given by

$$\hat{q}(l,k) = 1 - P_{\mathrm{local}}(l,k)P_{\mathrm{global}}(l,k)P_{\mathrm{frame}}(l). \tag{7.62}$$

The OM-LSA gain function is given by,

$$G_{\mathrm{OM-LSA}}(l,k) = \left\{G_{\mathrm{H}_1}(l,k)\right\}^{p(l,k)}\left\{G_{\mathrm{H}_0}(l,k)\right\}^{1-p(l,k)}, \tag{7.63}$$

with $G_{\mathrm{H}_1}(l,k) = G_{\mathrm{LSA}}(l,k)$ and $G_{\mathrm{H}_0}(l,k) = G_{\min}$. The lower-bound constraint for the gain when the signal is absent is denoted by $G_{\min}$, and specifies the maximum amount of reduction in those frames.

In our case the lower-bound constraint does not result in the desired result since the late reverberant signal and residual echo can still be audible. Our goal is to suppress the late reverberant signal and the residual echo down to the noise floor, given by $G_{\min}\ V(l,k)$. We apply $G_{\mathrm{H}_0}(l,k)$ to those time-frequency frames where the desired signal is assumed to be absent, i.e., the hypothesis $\mathrm{H}_0(l,k)$ is assumed to be true, such that

$$\hat{Z}_{\mathrm{e}}(l,k) = G_{\mathrm{H}_0}(l,k)\left(Z_{\mathrm{l}}(l,k) + E_r(l,k) + V(l,k)\right). \tag{7.64}$$

The desired solution for $\hat{Z}_{\mathrm{e}}(l,k)$ is

$$\hat{Z}_{\mathrm{e}}(l,k) = G_{\min}(l,k)\ V(l,k). \tag{7.65}$$

Assuming that all interferences are uncorrelated, minimizing

$$\mathcal{E}\left\{\left|G_{\mathrm{H}_0}(l,k)\left(Z_{\mathrm{l}}(l,k) + E_r(l,k) + V(l,k)\right) - G_{\min}(l,k)\ V(l,k)\right|^2\right\} \tag{7.66}$$

results in

$$G_{\mathrm{H}_0}(l,k) = G_{\min}\frac{\hat{\lambda}_v(l,k)}{\hat{\lambda}_{z_{\mathrm{l}}}(l,k) + \hat{\lambda}_{e_r}(l,k) + \hat{\lambda}_v(l,k)}. \tag{7.67}$$

7.5.2 *A priori* SIR estimator

Many researchers believe that the main advantage of the LSA estimator is related to the decision-directed estimator, proposed by Ephraim and Malah [205]. Rather than using one *a priori* Signal to Interference Ratio it is possible to calculate one value for each interference. By doing this, one gains control over the interference reduction level, and the *a priori* SIR estimation approach, of each interference. Note that in some cases it might be desirable to reduce one of the interferences at the cost of larger speech distortion, while other interferences are reduced less to avoid distortion. Due to the separation we can control the tradeoff between noise reduction and distortion of each of the interferences separately. Gustafsson et al. also used separate *a priori* SIRs in [250, 259] for two interferences, i.e., background noise and residual echo. In this section we show how the decision-directed estimator can be used to estimate the individual *a priori* SIRs, and we propose a slightly different way of combining them. It should be noted that each *a priori* SIR could be estimated using a different approach, e.g., the non-causal *a priori* SIR estimator proposed by Cohen in [208]. Note that the non-causal *a priori* SIR estimator can reduce distortion in speech onsets compared to the decision-directed estimator.

The *a priori* SIR in Eq. 7.57 can be written as

$$\frac{1}{\xi(l,k)} = \frac{1}{\xi_{z_1}(l,k)} + \frac{1}{\xi_{e_r}(l,k)} + \frac{1}{\xi_v(l,k)}, \tag{7.68}$$

with

$$\xi_\vartheta(l,k) = \frac{\lambda_{z_e}(l,k)}{\lambda_\vartheta(l,k)}, \tag{7.69}$$

where $\vartheta \in \{z_1, e_r, v\}$.

The decision-directed estimator is given by

$$\hat{\xi}(l,k) = \max\left\{\eta\frac{\hat{A}^2(l-1,k)}{\lambda(l-1,k)} + (1-\eta)\psi(l,k), \xi_{\min}\right\}, \tag{7.70}$$

where $\psi(l,k) = \gamma(l,k) - 1$ is the *instantaneous* SIR, $\gamma(l,k)$ is the *a posteriori* SIR as defined in Eq. 7.58,

$$\lambda(l',k) \triangleq \lambda_{z_1}(l',k) + \lambda_{e_r}(l',k) + \lambda_v(l',k), \tag{7.71}$$

and $\xi_{\min}$ is a lower-bound constraint on the *a priori* SIR. The weighting factor η ($0 \leq \eta < 1$) controls the tradeoff between the amount of noise reduction and distortion. The *a priori* SIR $\xi_\vartheta(l,k)$, as defined in Eq. 7.69, can be obtained using the following expression

$$\hat{\xi}_\vartheta(l,k) = \max\left\{\eta_\vartheta\frac{\hat{A}^2(l-1,k)}{\lambda_\vartheta(l-1,k)} + (1-\eta_\vartheta)\psi_\vartheta(l,k), \xi_{\min,\vartheta}\right\}, \tag{7.72}$$

where

$$\begin{aligned}\psi_\vartheta(l,k) &= \frac{\lambda(l,k)}{\lambda_\vartheta(l,k)}\psi(l,k) \\ &= \frac{|E(l,k)|^2 - \lambda(l,k)}{\lambda_\vartheta(l,k)},\end{aligned} \tag{7.73}$$

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

In case the near-end speech signal is very small, and the late reverberant and/or residual echo signals are very small, the *a priori* SIRs $\xi_{z_1}(l,k)$ and/or $\xi_{e_r}(l,k)$ may be unreliable since $\lambda_{z_e}(l,k)$ and $\lambda_{z_1}(l,k)$ and/or $\lambda_{e_r}(l,k)$ are close to zero. In the sequel we assume that there is always a certain amount of background noise. We propose to calculate $\xi(l,k)$ using only the most important and reliable *a priori* SIRs as follows[2]

$$\xi(l,k) = \begin{cases} \xi_v, & 10\log_{10}\left(\frac{\lambda_v}{\lambda_{z_1}+\lambda_{e_r}}\right) > \beta^{\mathrm{dB}}; \\ \xi', & \text{otherwise}, \end{cases} \tag{7.74}$$

and

$$\xi'(l,k) = \begin{cases} \frac{\xi_{e_r}\xi_v}{\xi_{e_r}+\xi_v}, & 10\log_{10}\left(\frac{\lambda_{e_r}}{\lambda_{z_1}}\right) > \beta^{\mathrm{dB}}; \\ \frac{\xi_{z_1}\xi_v}{\xi_{z_1}+\xi_v}, & 10\log_{10}\left(\frac{\lambda_{z_1}}{\lambda_{e_r}}\right) > \beta^{\mathrm{dB}}; \\ \frac{\xi_{z_1}\xi_v\xi_{e_r}}{\xi_v\xi_{e_r}+\xi_{z_1}\xi_{e_r}+\xi_{z_1}\xi_v}, & \text{otherwise}, \end{cases} \tag{7.75}$$

where the threshold β^{dB} specifies the level difference in dB. In case the noise level is β^{dB} higher than the level of residual echo and late reverberation (in dB), the total *a priori* SIR, $\xi(l,k)$, will be equal to $\xi_v(l,k)$. Otherwise $\xi(l,k)$ will be calculated depending on the level difference between $\lambda_{z_1}(l,k)$ and $\lambda_{e_r}(l,k)$ using Eq. 7.75. In case the level of residual echo is β^{dB} larger than the level of late reverberation, $\xi(l,k)$ will depend on both $\xi_v(l,k)$ and $\xi_{e_r}(l,k)$. In case the opposite is true, $\xi(l,k)$ will depend on both $\xi_v(l,k)$ and $\xi_{z_1}(l,k)$. In any other case $\xi_v(l,k)$ will be calculated using all *a priori* SIRs.

7.6 Experimental Results

In this section we present experimental results of our system. In the subsequent subsections we evaluate the residual echo suppression, the robustness with respect to echo path changes, and dereverberation performance. In Section 7.6.4 we evaluate the performance of the complete system during double-talk.

The experimental setup is depicted in Fig. 7.7. The room dimensions were 5 m x 4 m x 3 m (length x width x height). The distance between the near-end speaker and the microphone (r_s) was 1 m, the distance between the loudspeaker and microphone (r_l) was 0.5 m. The AIRs $\mathbf{a}(n)$ and $\mathbf{h}(n)$ were generated using the room impulse response

[2] The time and frequency indices at the right-hand side have been omitted.

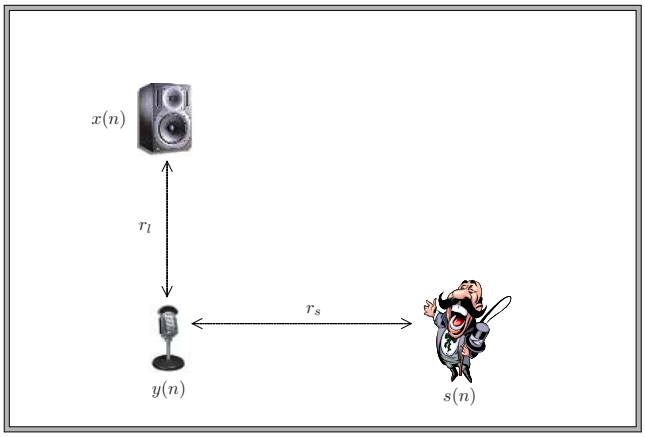


Figure 7.7 *Experimental setup.*

generator described in Appendix A. The first 250 ms of the AIRs are depicted in Fig. 7.8. The wall absorption coefficients were chosen such that the reverberation time is approximately 500 milliseconds. The microphone signal $y(n)$ was generated using Eq. 7.9. The analysis window $w(n)$ of the Short-Time Fourier Transform was a 256 point Hamming window, i.e., $L_w = 256$, and the overlap between two successive frames was set to 75%, i.e., $R = 0.25\ L_w$. The remaining parameter settings are shown in Table 7.1. The additive noise $v(n)$ was speech-like noise, taken from the NOISEX-92 database [237].

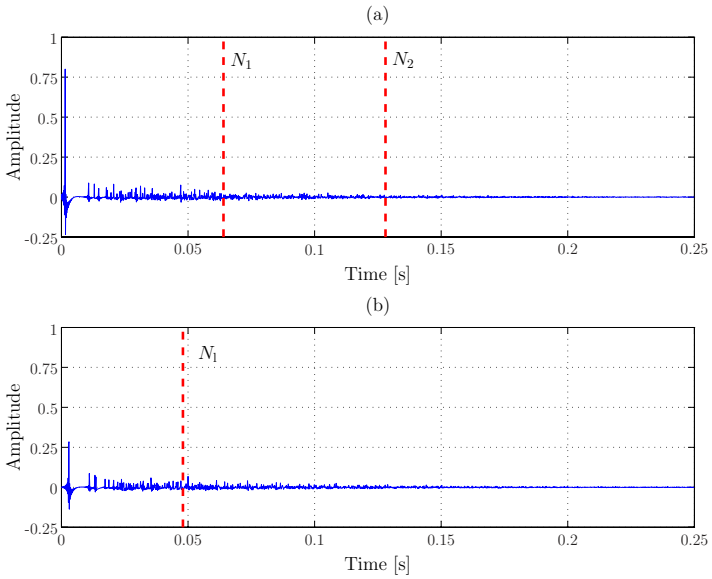


Figure 7.8 *Acoustic impulse responses (a)* $\mathbf{h}(n)$ *and (b)* $\mathbf{a}(n)$.

$f_\mathrm{s} = 8000$ Hz	$N_1 = 0.064\ f_\mathrm{s}$	$N_2 = 0.128\ f_\mathrm{s}$	$N_\mathrm{l} = 0.048\ f_\mathrm{s}$
$G_\mathrm{min}^\mathrm{dB} = 18$ dB	$\beta^\mathrm{dB} = 9$ dB	w=3	

Table 7.1 *Parameters used for these experiments.*

7.6.1 Residual Echo Suppression

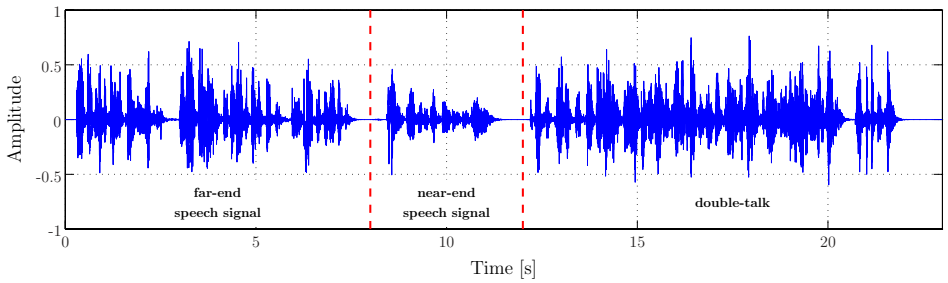
The echo cancellation performance, and more specifically the improvement due to the residual echo suppression of the post-processor, was evaluated using the Echo Return Loss Enhancement. The ERLE achieved by the first adaptive filter was calculated using

$$\mathrm{ERLE}(l) = \frac{\sum_{n=lR'}^{lR'+L'-1} d^2(n)}{\sum_{n=lR'}^{lR'+L'-1} \left(d(n) - \hat{d}_\mathrm{e}(n)\right)^2}, \tag{7.76}$$

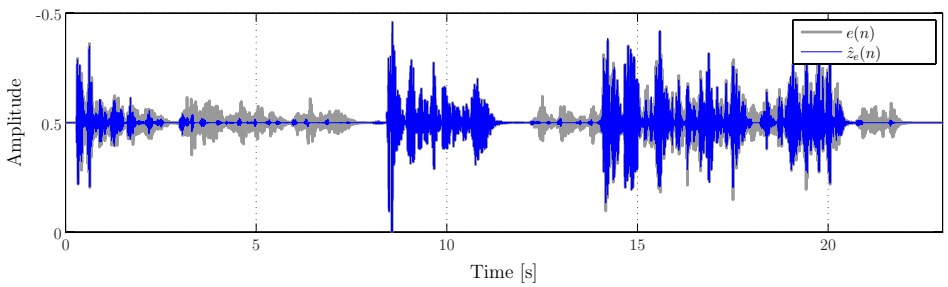
where $L' = 0.032\ f_\mathrm{s}$ is the frame length and $R' = \frac{L'}{4}$ is the frame rate. To evaluate the total echo suppression, i.e., with post-processor, we calculated the ERLE using Eq. 7.76 and replaced $d(n) - \hat{d}_\mathrm{e}(n)$ by $z(n) - \hat{z}_\mathrm{e}(n)$, where the latter term denotes the residual echo at the output of the post-filter. This experiment was conducted without noise, and the post-filter was configured such that no reverberation was reduced, i.e., $\lambda_{z_\mathrm{l}}(l,k) = 0\ \forall l\ \forall k$. The microphone signal $y(n)$, the error signal $e(n)$, and the ERLE with and without post-processor are shown in Fig. 7.9. We can see that the ERLE has significantly increased when the post-processor is used.

7.6.2 Robustness

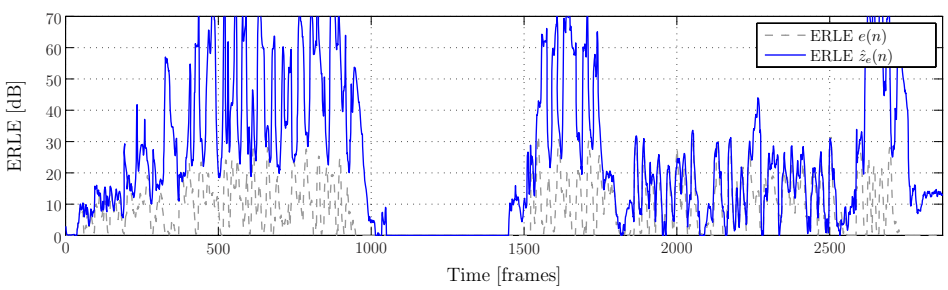
We evaluate the robustness of the system with respect to changes in the echo path when the far-end speech signal was active. Two types of changes were applied to the acoustic echo path. First, we rotated the loudspeaker over 15 degrees in the x-y plane with respect to the microphone (at 3.5 seconds). Second, we decreased the microphone-loudspeaker distance by 5 cm (at 6.1 seconds). We compared the performance of our system with a standard AEC, i.e., without post-processor, and an adaptive filter of length N_2. The time at which the position changes is marked with a dash-dotted line. The microphone signal $y(n)$, the error signal $e(n)$ of the standard AEC, and the ERLEs are shown in Fig. 7.10. Despite the fact that the ERLE drops in all cases, we can clearly see that the drop at 3.5 seconds is smaller for the proposed system. At 6.1 seconds we can see that the output signal of the developed system is less than the error signal of the standard AEC. These results indicate that the proposed system is more robust than the standard AEC. It should also be noted that the late residual echo estimator, which is very robust to changes in the room, does not require any re-convergence time, i.e., when the reverberation time has been estimated.



(a) Microphone signal $y(n)$.



(b) Error signal $e(n)$ and the processed signal $\hat{z}_e(n)$.

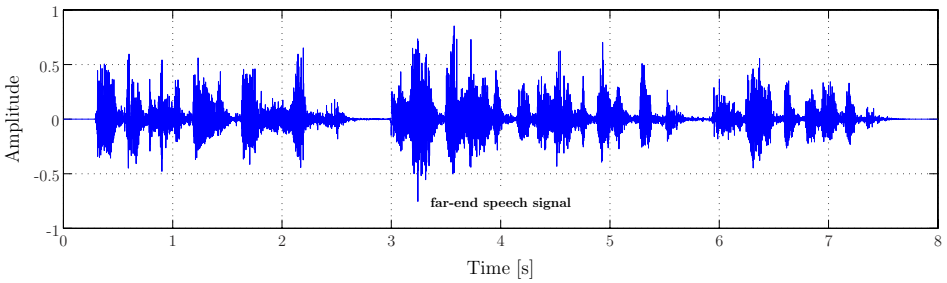


(c) Echo Return Loss Enhancement w.r.t. $e(n)$ and $\hat{z}_e(n)$.

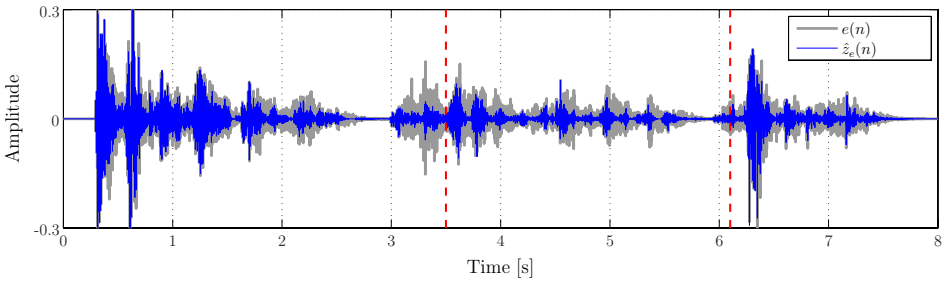
Figure 7.9 *Echo suppression performance.*

7.6.3 Dereverberation

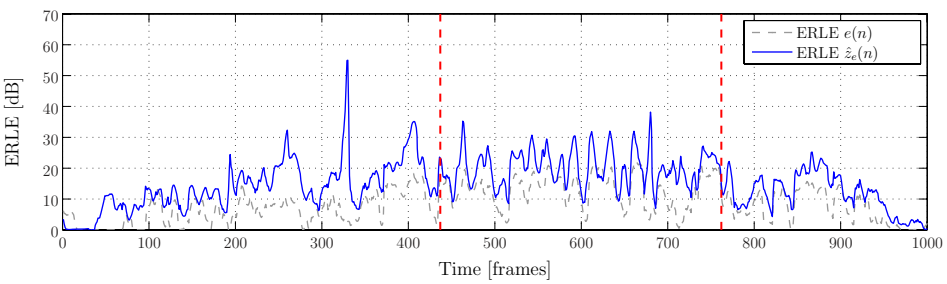
The dereverberation performance has been evaluated using the segmental SIR, Log Spectral Distortion (LSD), Bark Spectral Distortion (BSD), and PESQ. The Direct to Reverberation Ratio $\kappa(k)$ was obtained from the Energy Decay Curve of the acoustic impulse response $\mathbf{a}(n)$. An estimate of the reverberation time ($\widehat{\mathrm{RT}}_{60}(k)$) was obtained using the procedure described in Section 7.3.2. After convergence of the adaptive filters RT_{60} was 493 ms. The parameter N_1 was set to 0.048 f_s.



(a) Microphone signal $y(n)$.



(b) Error signal $e(n)$ and the processed signal $\hat{z}_\mathrm{e}(n)$.



(c) Echo Return Loss Enhancement w.r.t. $e(n)$ and $\hat{z}_\mathrm{e}(n)$.

Figure 7.10 *Echo suppression performance w.r.t. echo path changes*

The instantaneous SIR of the l^{th} frame is defined as

$$\mathrm{SIR}(l) = 10\log_{10}\left(\frac{\sum_{n=lR'}^{lR'+L'-1} z_\mathrm{e}^2(n)}{\sum_{n=lR'}^{lR'+L-1}\left(z_\mathrm{e}(n)-\upsilon(n)\right)^2}\right) \quad [\mathrm{dB}], \tag{7.77}$$

where $\upsilon \in \{y, \hat{z}_\mathrm{e}\}$. The segmental SIR, denoted by $\mathrm{SIR}_{\mathrm{seg}}$, is defined as the average instantaneous SIR over the set of frames where the near-end speech is active.

The LSD between $z_\mathrm{e}(n)$, i.e., the anechoic signal convolved with the acoustic impulse response $\mathbf{a}_\mathrm{e}(n)$, and the dereverberated signal is used as a measure of distortion. The

	SNR_{seg}	Unprocessed	Processed (NR)	Processed (RR+NR)
SIR_{seg}	5 dB	-0.53 dB	3.26 dB	4.09 dB
	10 dB	1.73 dB	4.34 dB	5.66 dB
	25 dB	4.82 dB	5.26 dB	7.57 dB
LSD	5 dB	7.62 dB	3.51 dB	3.38 dB
	10 dB	5.47 dB	3.05 dB	2.75 dB
	25 dB	2.96 dB	2.78 dB	2.05 dB
BSD	5 dB	0.0594	0.0538	0.0382
	10 dB	0.0478	0.0467	0.0286
	25 dB	0.0442	0.0441	0.0245
PESQ	5 dB	1.87	2.37	2.42
	10 dB	2.18	2.15	2.67
	25 dB	2.54	2.57	2.89

Table 7.2 *Segmental SIR, LSD, BSD, and PESQ for different segmental Signal to Noise Ratios* ($\text{RT}_{60} \approx 0.5$ *s*).

distance in the l^{th} frame is calculated using

$$\text{LSD}(l) = \frac{1}{K} \sum_{k=0}^{K-1} \left| 10 \log_{10} \left(\frac{Z'_{\text{e}}(l,k)}{\Upsilon'(l,k)} \right) \right| \quad [\text{dB}], \tag{7.78}$$

where $\Upsilon \in \{Y, \hat{Z}_{\text{e}}\}$, K denotes the number of frequency bins, and

$$P'(l,k) \triangleq \max\{|P(l,k)|^2, \delta\} \tag{7.79}$$

is the spectral power where $P \in \{Z_{\text{e}}, \Upsilon\}$, clipped such that the log-spectrum dynamic range is confined to about 50 dB, i.e., $\delta = 10^{-50/10} \max_{l,k}\{|P(l,k)|^2\}$. Finally, the LSD is defined as the average distance over all frames.

The BSD and Perceptual Evaluation of Speech Quality (PESQ) scores were calculated by comparing $z_{\text{e}}(n)$ with $y(n)$ and $\hat{z}_{\text{e}}(n)$, respectively. Both perceptually motivated objective measures are described in Chapter 4.

We tested the dereverberation performance under different segmental Signal to Noise Ratios. The segmental SNR value is determined by averaging the instantaneous SNR of those frames where the near-end speech is active. To show the improvement related to the dereverberation process we evaluated the objective measures with and without reverberation reduction. The results under different segmental SNRs, with Noise Reduction (NR), and with Reverberation Reduction and Noise Reduction (RR+NR), are shown in Table 7.2. The results show a consistent improvement in segmental SIR, LSD, BSD and PESQ values.

The instantaneous SIR and LSD results obtained with a segmental SNR of 25 dB together with the anechoic, reverberant and processed signals are presented in Fig. 7.11. Since the Signal to Noise Ratio is relatively high, the instantaneous Signal to Interference Ratio mainly relates to the amount of reverberation, such that the SIR improvement relates to the reverberation reduction. Especially in those areas where

	$\hat{\kappa}(k)$		
	$0.9 \cdot \kappa(k)$	$\kappa(k)$	$1.1 \cdot \kappa(k)$
SIR_{seg}	7.52 dB	7.57 dB	7.62 dB
LSD	2.05 dB	2.05 dB	2.05 dB
BSD	0.0250	0.0245	0.0241
PESQ	2.88	2.89	2.90

Table 7.3 *Segmental SIR, LSD, BSD and PESQ, segmental SNR = 5 dB, and* $\hat{\kappa}(k) = \{0.9 \cdot \kappa(k), \kappa(k), 1.1 \cdot \kappa(k)\}$.

	SNR_{seg}	Unprocessed	Processed (RR+NR)
SIR_{seg}	5 dB	-3.70 dB	3.51 dB
	10 dB	-3.08 dB	4.60 dB
	25 dB	-2.66 dB	5.79 dB
LSD	5 dB	10.95 dB	2.69 dB
	10 dB	9.88 dB	2.16 dB
	25 dB	9.04 dB	1.82 dB
BSD	5 dB	0.7840	0.0978
	10 dB	0.7908	0.0660
	25 dB	0.7942	0.0485
PESQ	5 dB	0.85	1.95
	10 dB	0.94	1.82
	25 dB	1.61	2.34

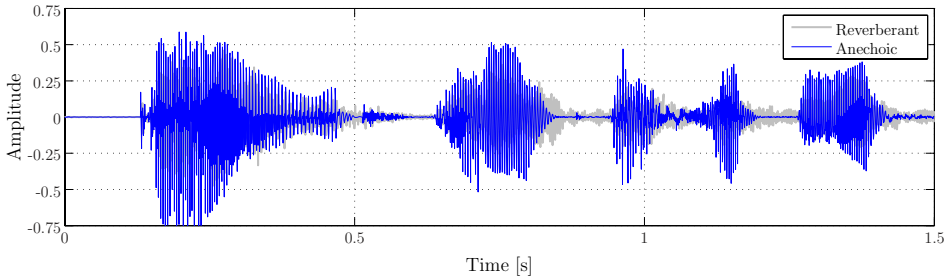
Table 7.4 *Segmental SIR, LSD, BSD, and PESQ for different segmental Signal to Noise Ratios during double-talk* ($\text{RT}_{60} \approx 0.5$ *s*).

the instantaneous SIR of the unprocessed signal is low the instantaneous SIR and LSD are increased and decreased, respectively. During speech onsets some speech distortion (negative improvement) may occur due to the decision-directed *a priori* SIR estimation (see Section 7.5.2). The spectrograms and waveforms of the near-end speech signal $z(n)$, $z_{\text{e}}(n)$, and the processed signal $\hat{z}_{\text{e}}(n)$ are shown in Fig. 7.12. From these plots it can be clearly seen that the smearing in time due to the reverberation has been reduced significantly.

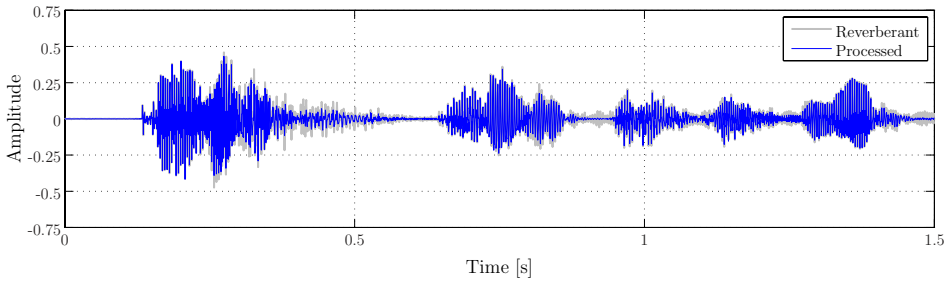
In practice the direct to reverberant energy ratio $\kappa(k)$ needs to be estimated online. To evaluate the robustness with respect to errors in $\kappa(k)$ we introduced an error of $\pm 10\%$. The SIR_{seg} and LSD using the perturbated values of $\kappa(k)$ are shown in Table 7.3. From this experiment we can see that the performance of the developed algorithm is not very sensitive to errors in the parameter $\kappa(k)$.

7.6.4 Joint Suppression Performance

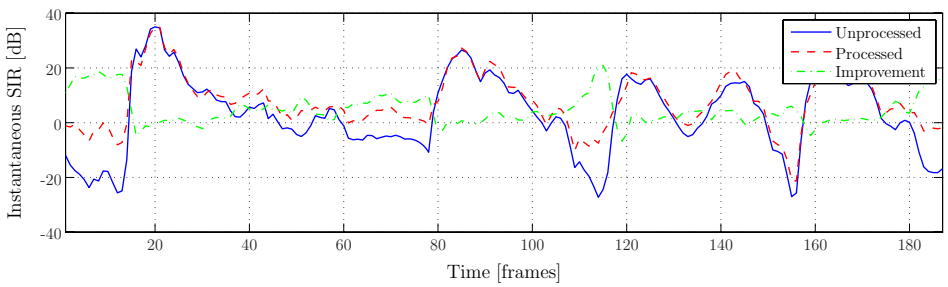
We evaluated the performance during double-talk using the segmental SIR, LSD, BSD, and PESQ at three different segmental SNR values. The results are presented in Table 7.4. The results show a significant improvement in terms of all objective measures.



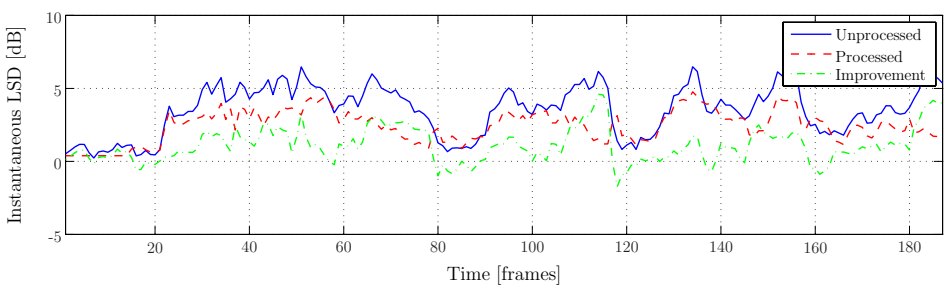
(a) Reverberant and anechoic near-end speech signal.



(b) Reverberant and estimated near-end speech signal.



(c) SIR of the unprocessed and processed near-end speech signal, and improvement.



(d) LSD of the unprocessed and processed near-end speech signal, and improvement.

Figure 7.11 *Dereverberation performance of the system during near-end speech period* ($\mathrm{RT}_{60} \approx 0.5$ *s).*

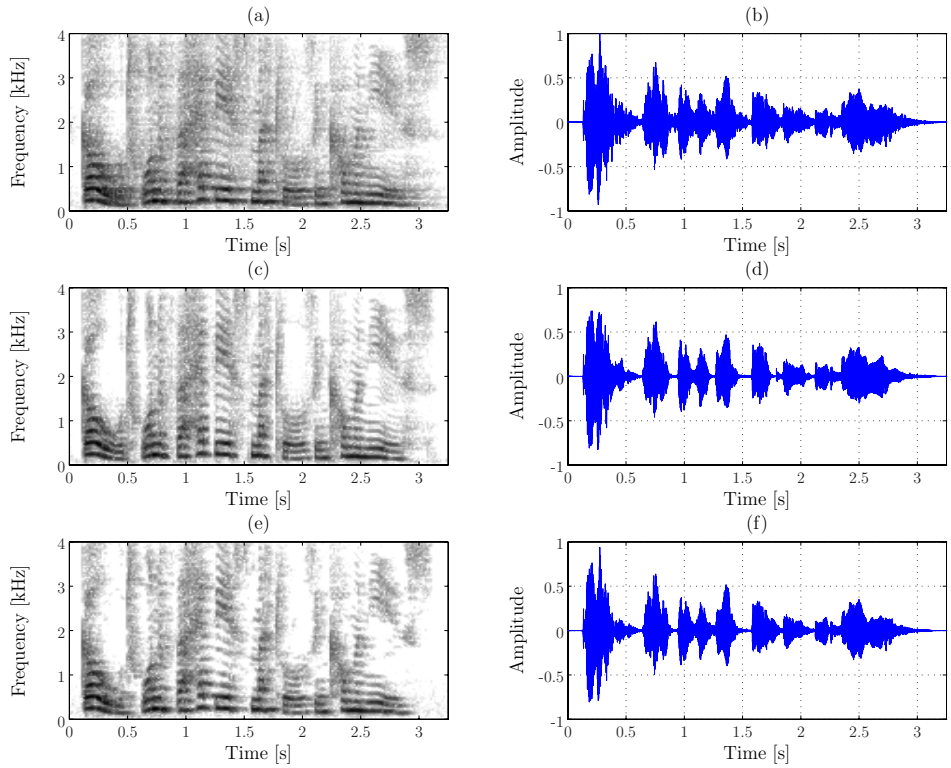


Figure 7.12 *Spectrogram and waveform of (a-b) the reverberant near-end speech signal $z(n)$, (c-d) the near-end speech signal $z_e(n)$, and (e-f) the dereverberated near-end speech signal $\hat{z}_e(n)$ (segmental SNR = 25 dB,* $\mathrm{RT}_{60} \approx 0.5$ *s).*

The spectrograms of the microphone signal, near-end speech signal and the estimated signal $\hat{z}_e(n)$ for a segmental SNR of 25 dB, and 5 dB are, respectively, shown in Figs. 7.13 and 7.14. The spectrograms clearly show how well the interferences are suppressed during double-talk.

7.7 Discussion

The lengths of the two adaptive filters can be controlled using the parameters N_1 and N_2. The choice will in general be related to the application, acoustic environment and the desired complexity and robustness. In some cases it even might be desired to make N_1 zero, which results in a complete spectral acoustic echo canceller. In other cases, where N_1 covers the direct path and early reflections, e.g., in a small enclosed space, N_2 can be chosen equal to N_1, such that the second adaptive filter is omitted, and only the late residual echo estimator is used.

We have used a standard Normalized Least Mean Squares algorithm to update the

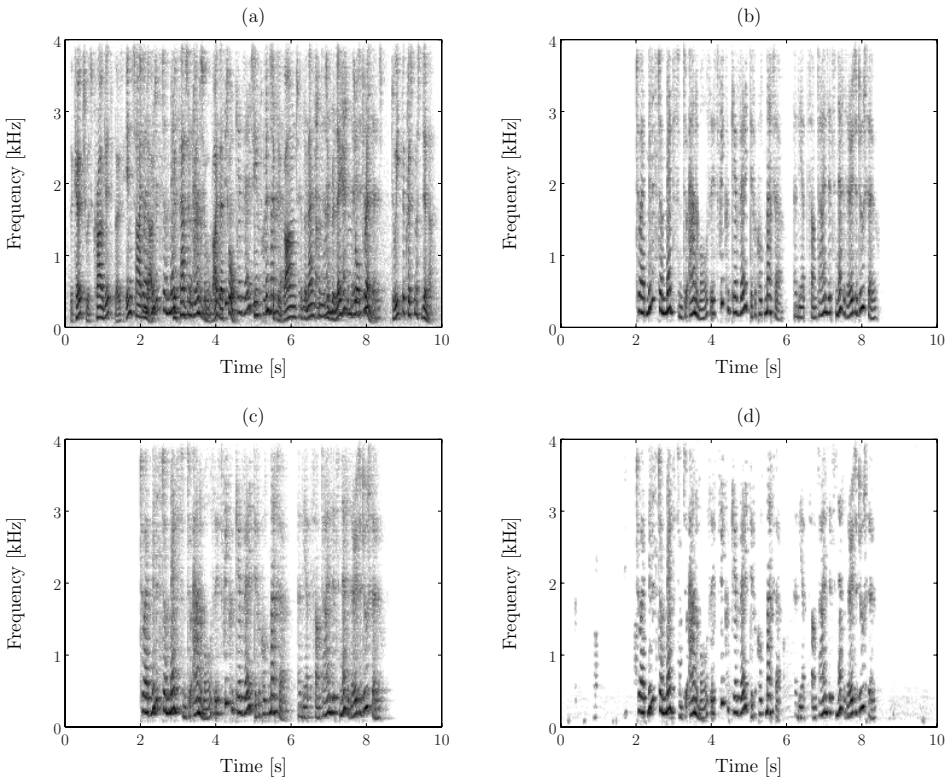


Figure 7.13 *Spectrograms of (a) the microphone signal $y(n)$, (b) the near-end speech signal $z_{\mathrm{e}}(n)$, (c) the reverberant near-end speech signal $z(n)$, and (d) the estimated signal $\hat{z}_{\mathrm{e}}(n)$, during double-talk (segmental SNR = 25 dB,* $\mathrm{RT}_{60} \approx 0.5$ *s).*

adaptive filters. Due to the choices of N_1 and N_2, the lengths of the adaptive filters are deficient. In case the far-end signal $x(n)$ is not spectrally white, the filter coefficients are biased [260, 261]. However, the filter coefficients that are mostly affected, are in the tail region of $\hat{\mathbf{h}}_{\mathrm{e}}(n)$ and $\hat{\mathbf{h}}_{\mathrm{m}}(n)$. Accordingly, this problem can be partially solved by increasing the values of N_1 and N_2, and calculating the output using the original N_1 and $N_2 - N_1$ coefficients of the filters. Alternatively, one could use a, possibly adaptive, pre-whitening filter [241], or other adaptive algorithm like AP or RLS.

The amount of reverberation reduction can be controlled using $\xi_{\min,z_{\mathrm{l}}}$ and the parameter N_{l}. It should be noted that in case the source-microphone distance is smaller than the critical distance, such that $\kappa(k) < 1$, the time $N_{\mathrm{l}}/f_{\mathrm{s}}$ can be reduced to around 8-16 ms, i.e., one or two frames, while keeping the amount of speech distortion low.

An estimate of the reverberation time is required for the late residual echo estimation and late reverberant energy estimation. In some applications, e.g., conference systems, this parameter may be determined using a calibration step. We developed a method to estimate the reverberation time online using the estimated filter $\hat{\mathbf{h}}_{\mathrm{m}}$, assuming that the convergence of the filter $\hat{\mathbf{h}}_{\mathrm{m}}$ is sufficient. Instantaneous divergence of the filter

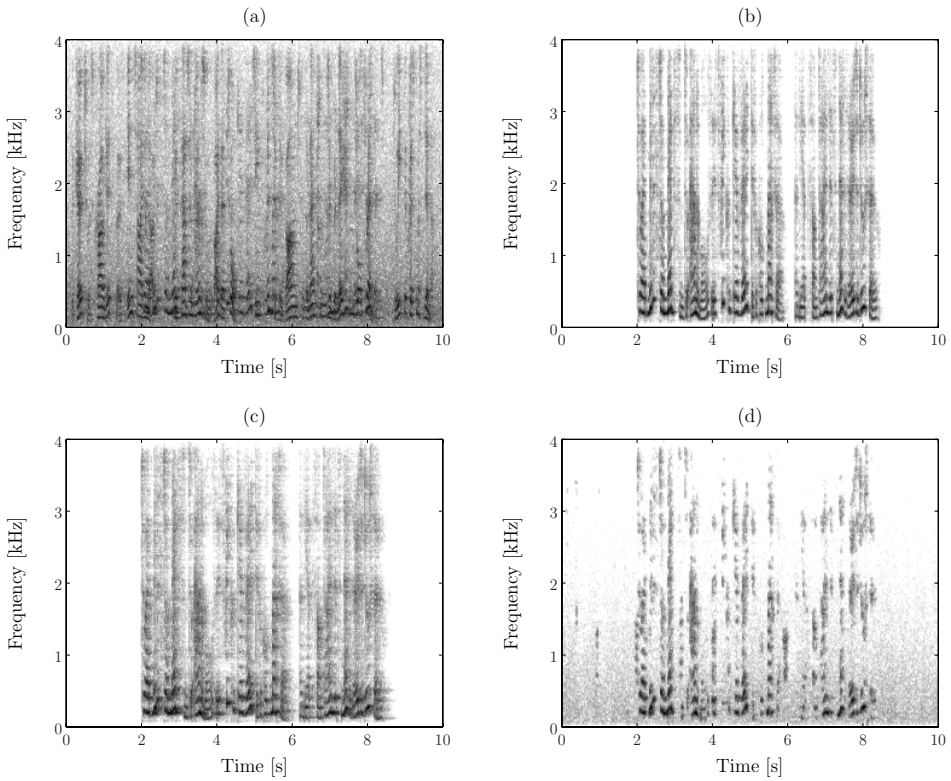


Figure 7.14 *Spectrograms of (a) the microphone signal $y(n)$, (b) the near-end speech signal $z_{\mathrm{e}}(n)$, (c) the reverberant near-end speech signal $z(n)$, and (d) the estimated signal $\hat{z}_{\mathrm{e}}(n)$, during double-talk (segmental SNR = 5 dB,* $\mathrm{RT}_{60} \approx 0.5$ *s).*

coefficients, e.g., due to false double-talk detection or echo path changes, does not significantly influence the estimation of the reverberation time due to its relatively slow update mechanism. In case the filter coefficients cannot convergence, for example due to background noise, the estimated reverberation time will be inaccurate. It should be noted that under-estimation of the reverberation time will reduce the performance of the system in terms of late residual echo and reverberation suppression. However, under-estimation will not introduce any additional distortion.

Acoustic echo cancellation solutions that are capable of handling both the residual echo and background noise are often implemented in the STFT domain and usually require two STFTs and one inverse STFT. The increase in complexity of the proposed system compared to former solutions is small since the estimation of the reverberation time and the late reverberant energy only requires a few extra operations. It should be noted that the current implementation requires three STFTs and one inverse STFT. However, the number of STFTs can be reduced to two by implementing the adaptive filters $\hat{\mathbf{h}}_{\mathrm{e}}(n)$ and $\hat{\mathbf{h}}_{\mathrm{m}}(n)$ in the STFT domain.

7.8 Conclusions

In this chapter we have developed a novel post-processor which is designed to efficiently reduce reverberation of the near-end speech signal, residual echo and background noise. We proposed to divide the acoustic echo path into three distinct parts, each of which is handled in a different way. The late residual echo is reduced in a unique and efficient way by exploiting the exponential decay of the acoustic echo path. The exponential decay, which is related to the reverberation time of the room, is identified from the estimated acoustic echo path. The estimated reverberation time is also used for dereverberation of the near-end speech signal. A single post-filter based on the Optimally-Modified Log Spectral Estimator is applied to the error signal of the acoustic echo canceller to obtain an estimate of the dereverberated near-end speech signal. Experimental results demonstrate the high performance of the proposed system, and its robustness to small position changes of the loudspeaker in the room.

Chapter 8

Conclusions and Further Research

The objective of the work presented in this dissertation is to investigate the application of spectral enhancement techniques to suppress late reverberation, residual echo and background noise. In the previous chapters, novel single- and multi-microphone techniques have been introduced to achieve these objectives. This chapter summarizes the obtained results and highlights the contribution of the present work.

8.1 Conclusions

In typical speech communication systems, such as hands-free mobile telephones, video-conferencing, voice-controlled systems, hearing aids and cochlear implants, the received microphone signals are corrupted by room reverberation, background noise, and far-end echo signals. This signal degradation may lead to total unintelligibility of speech and decreases the performance of automatic speech recognition systems. Hence, high performance acoustic signal processing techniques are required.

In this dissertation several single- and multi-microphone speech dereverberation techniques based on spectral enhancement were developed. It was shown that quantifiable properties, e.g., the reverberation time and direct to reverberation ratio, can be used to dereverberate the received microphone signals. We will now summarize the obtained results and highlight the main contributions of the individual chapters.

In **Chapter 2** we have provided some fundamental prerequisites on room acoustics. The theory described in this chapter was used throughout this dissertation.

An extensive literature survey covering a variety of speech dereverberation techniques that were developed in the last three decades has been presented in **Chapter 3**. We categorized the reverberation reduction techniques depending on whether or not the acoustic impulse response needs to be estimated. We then obtain two main categories, i.e., Reverberation Suppression and Reverberation Cancellation. Techniques within

these categories can be divided into smaller sub-categories depending on either the amount of prior knowledge about either the source or the acoustic channel that is utilized.

In **Chapter 4** we have discussed and evaluated objective measures to determine the quality of the dereverberated signal. It was shown that the segmental Signal to Reverberation Ratio and Reverberation Decay Tail are very useful quantitative objective measures, which have an almost linear relation with the reverberation time. Unfortunately there are currently no objective measures available to evaluate the change in colouration independently from the reverberation time. Furthermore, a novel time-frequency representation of a reverberant speech signal was proposed. Compared to the standard spectrogram the proposed representation clearly indicates which time-frequency components have a low Direct to Reverberation Ratio, i.e., which time-frequency components are affected most by reverberation.

Three multi-microphone speech dereverberation systems have been developed in **Chapter 5**. Two spectral enhancement techniques were used to enhance the observed, possibly noisy, reverberant speech signal. The first technique is based on spectral subtraction, and the second technique is based on the Optimally-Modified Log Spectral Amplitude (OM-LSA) estimator. Several modifications of the OM-LSA estimator were developed to increase its performance. The developed multi-microphone speech dereverberation systems exhibit a low computational complexity and are robust to small changes in the room characteristics. Results obtained using synthetic and measured acoustic impulse responses showed a significant reverberation reduction with little or no speech distortion.

In **Chapter 6** we have discussed Polack's statistical reverberation model, which can be used to estimate the late reverberant spectral variance directly from the received microphone signal. We showed that Polack's statistical reverberation model is closely related to the physical energy balance in an ideal diffuse environment. Polack's reverberation model is based on the implicit assumption that the reverberant component is dominant, and hence, that the source-microphone distance is larger than the critical distance. We proposed a generalized statistical reverberation model which was used to derive a novel late reverberant spectral variance estimator. The derived estimator can be used over a wide range of source-microphone distances. We showed that the proposed estimator is biased when background noise is present. An unbiased estimator was developed which can be used in case an estimate of the noise spectral variance is available. In case the noise is time-invariant, or slowly time-varying, the bias can be calculated explicitly from the estimate noise spectral variance. The bias can then be subtracted from the estimated late reverberant spectral variance. Compared to the unbiased estimator this simple correction reduces the computational complexity of the system.

A novel post-processor for an acoustic echo cancellation system was developed in **Chapter 7**. The post-processor can be used to suppress reverberation, residual echo, and background noise. The system is unique in the following ways. Firstly, it uses an advanced spectral enhancement technique, viz., the OM-LSA estimation technique.

This technique is used to suppress late reverberation of the near-end speech signal, residual echo, and background noise. We have developed several modifications of the OM-LSA estimation technique. The first modification provides a stationary residual noise level while suppressing multiple interferences. The second modification increases the ability to control the estimation procedure of each interference. Secondly, we divided the echo path into three parts (direct path, early and late). While the echo signal that results from the direct path is cancelled using a classical acoustic echo canceller technique, the echo signal which results from the second and third part are suppressed using the developed spectral enhancement technique. The third part, i.e., late reverberant part, is estimated using the model described in Chapter 6. The dereverberation of the near-end speech signal is based on the techniques developed in Chapters 5 and 6. The late reverberant spectral variance estimator requires an estimate of the reverberation time. However, blind estimation of the reverberation time can be very problematic. In this case we used the estimated echo path to estimate the reverberation time of the room. Experimental results convincingly demonstrate the benefits of the proposed system for suppressing late reverberation, residual echo and background noise. The system has a low computational complexity, a highly modular structure, can be seamlessly integrated into existing hands-free communication systems, and affords a significant increase of the listening comfort and speech quality.

Synthetic room impulse responses are often created using the image method developed by Allen and Berkley. In **Appendix A** this method is explained, and an efficient Matlab® implementation in the form of a MEX-function is provided. This function can be used to simultaneously calculate multiple impulse responses, i.e., from one source to multiple microphones. Various improvements are made to incorporate the directivity of the microphone and to ensure proper inter-microphone phase relations, which are very important in the case of Single-Input Multi-Output (SIMO) and Multi-Input Multi-Output (MIMO) systems. Some extra features were added which allow the design of less complex room impulse responses.

The optimal modified Log Spectral Amplitude estimator is often used for noise reduction. In **Appendix B** we provide an extension to this estimator to be able to deal with multiple interferences, more specifically to deal with one non-stationary and one stationary interference. Three methods to estimate the *a priori* Signal to Interference Ratio are discussed, viz., decision-directed, causal and non-causal recursive estimation.

Currently state-of-the-art noise reduction techniques are widely used in telecommunication and consumer equipment such as video-conferencing, hands-free mobile telephony and voice-controlled systems, hearing aids and cochlear implants. However, until now dereverberation techniques are rarely included because there are no practical and robust techniques available. Therefore, dereverberation techniques are of great interest to the high-tech industry and their costumers. The results obtained using the single- and multi-microphone speech dereverberation techniques developed in this dissertation show that a substantial amount of reverberation reduction can be achieved with little *a priori* information about the acoustic channel. Due to the low complexity and robustness of these techniques we believe that they can and will be used in future applications, like hands-free devices and hearing-aids.

8.2 Suggestions for further research

In this section we will provide several suggestions for further research.

The developed generalized late reverberant spectral variance estimator requires a limited amount of *a priori* information about the acoustic channel. As discussed in Chapter 6 and 7 this information can be obtained in some applications, e.g., hands-free mobile devices. In some applications this might be difficult, and the use of calibration procedures that need to be performed by the costumer should be avoided. Therefore, it is of great importance to *further improve the possibilities to blindly estimate the required parameters, such as the reverberation time and the direct to reverberation ratio.*

Throughout this work we mainly used objective measures to evaluate the performance of the developed dereverberation algorithms. In [30] we evaluated three multi-microphone speech dereverberation techniques using subjective listening tests. The first dereverberation technique is a standard delay and sum beamfomer, the second technique is a Linear Prediction based enhancement technique proposed by Gaubitch et al. in [110], and the third is an early version of the spectral subtraction based dereverberation technique described in Chapter 5. The subjective listening test was performed according to the guidelines of International Telecommunications Union (ITU-T) Recommendation Series-P for subjective testing [178, 232]. Using the listening tests, we have estimated the subjective perception of colouration, reverberation decay tail effect, and the overall speech quality. A total of 26 normal hearing subjects was subjected to 64 speech files, with a male and a female talker for eight acoustic setups (different distances and reverberation times), and speech processed with the three dereverberation algorithms. Calibration speech examples were given to assist listeners in identifying colouration and reverberation decay tail effects. Compared to the received reverberant microphone signal the results for the spectral enhancement based technique indicated that the colouration was approximately equal, the reverberation decay tail effect was reduced, and the overall speech quality was improved. Based on the intelligibility tests performed by Allen [15] we expect that the intelligibility of the processed signal has been improved. However, further *investigation and verification of the speech intelligibility* is of great importance to further assess, improve, and increase the utilization, of these and other dereverberation methods.

In **Chapter 5** we have discussed three multi-microphone dereverberation systems. We showed that the combination of spatial processing and a dereverberation post-processor can be very problematic due to the spatial correlation between the acoustic channels. However, spatial processing algorithms can be useful to suppress early reflections. Further research might be conducted to investigate other possibilities to suppress late reverberation in such a case. A first attempt was made in [38] using a dual microphone system. Here a reference signal was constructed by blocking the direct speech signal. This reference signal was then used to estimate the late reverberant energy adaptively. A post-filter was used to enhance the output of a delay and sum beamformer. Future research could focus on the extension to multiple microphones,

which allows better estimation of the residual reverberant energy and suppression of the early reflections. Furthermore, more realistic situations where additional interferences are present could be investigated.

Currently, the reverberation models are described in the time domain. Certain approximations are made when the late reverberant spectral variance estimator is transformed to the short-time Fourier transform (STFT) domain. To avoid this, it would be of interest to *develop a multi-channel statistical model in the STFT domain*, and to derive a late reverberant spectral variance estimator according to this model.

In this work we have processed all signals in the STFT domain. Many spectral enhancement techniques showed a subjective improvement when processing was performed in a perceptual domain, e.g., using non-uniform frequency bands. As an example, processing could be performed in the Bark spectral domain, which is often used in hearing-aid devices. Hence, it would be of interest to *investigate the performance and applicability of other time-frequency transforms.*

The late reverberant spectral variance estimators derived in **Chapter 6** are based on the fact that the envelope of the acoustic impulse response has an exponential decay. Although this assumption is true for many enclosed spaces a generalization might be of interest. In some situations we are dealing with so-called coupled rooms (for example when there is an opening between two enclosed spaces). In case the coupled rooms exhibit a different decay rate in each room, the total decay consists of a sum of exponential decays [262]. Hence, it would be of interest to *derive a late reverberant spectral variance estimator for non-exponentially decaying envelopes.*

In **Chapter 7** we have developed a post-processor for a Single-Input Single-Output (SISO) acoustic echo cancellation system. In Chapter 5 we already saw that multiple microphone can be used to increase the reverberation reduction and the decrease speech distortion compared to a single microphone. Furthermore, using multi-microphones the noise reduction can be increased. Therefore, it is of interest to *extend the post-processor to SIMO and MIMO acoustic echo cancellation systems.*

The late reverberant spectral variance estimator is based on forward estimation, i.e., only the observed microphone signals are used to estimate the late reverberant spectral variance. It should be noted that the first speech signals that are received by the microphone are free of reverberation (assuming that the room is 'in rest'. Therefore this signal could be used to 'clean' future samples. Hence *backward or forward-backward estimators might be used to estimated the late reverberant spectral variance.*

Bibliography

[1] B.D. Radlović, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, 1999, vol. 2, pp. 881–884.

[2] B.D. Radlović, R. Williamson, and R. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 311–319, 2000.

[3] F. Talantzis and D. Ward, "Investigation of performance of acoustic arrays for equalization in a reverberant environment," in *Proc. of the fourth International Conference on Digital Signal Processing*, 2002, vol. 1, pp. 247–250.

[4] J.S. Garofolo, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," Tech. rep., National Institute of Standards and Technology (NIST), Gaithersburg,Maryland, Dec. 1988.

[5] X. Huang, A. Acero, and H. Hon, *Spoken language processing. a guide to theory, algorithm and system development*, Prentice-Hall, 2001.

[6] B. Libbey and P.H. Rogers, "The effect of overlap-masking on binaural reverberant word intelligibility," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3141–3151, 2004.

[7] G.A. Soulodre, N. Popplewell, and J.S. Bradley, "Combined effects of early reflections and background noise on speech intelligibility," *Journal of Sound Vibration*, vol. 135, pp. 123–133, 1989.

[8] J.C. Steinberg, "Effects of distortion upon the recognition of speech sounds," *Journal of the Acoustical Society of America*, vol. 1, pp. 121–137, 1995.

[9] M. Hodgson and E.M. Nosal, "Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms," *Journal of the Acoustical Society of America*, vol. 111, pp. 931–939, 2002.

[10] J.P.A. Lochner and J.F. Burger, "The subjective masking of short time delayed echoes by their primary sounds and their contribution to the intelligibility of speech," *Acustica*, vol. 8, pp. 1–10, 1958.

[11] A.J. Watkins and N.J. Holt, "Effects of a complex reflection on vowel indentification," *Acustica*, vol. 86, pp. 532–542, 2000.

[12] F. Aigner and M.J.O. Strutt, "On a physiological effect of several sources of sound on the ear and its consequences in architectural acoustics," *Journal of the Acoustical Society of America*, vol. 6, no. 3, pp. 155–159, 1935.

[13] V.M.A. Peutz, "Articulation loss of consonants as a criterion for speech transmission in a room," *Journal of the Audio Engineering Society*, vol. 19, no. 11, pp. 915–919, Dec. 1971.

[14] D.A. Berkley, *Acoustical factors affecting hearing aid performance*, chapter Normal listeners in typical Rooms - Reverberation Perception, Simulation, and Reduction, pp. 3–24, University Park Press, Baltimore, 1980.

[15] J.B. Allen, "Effects of small room reverberation on subjective preference," *Journal of the Acoustical Society of America*, vol. 71, no. S1, pp. S5, 1982.

[16] J.J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *Journal of the Acoustical Society of America*, vol. 65, no. 5, pp. 1204–1211, 1979.

[17] Y. Takata and A.K. Nábělek, "English consonant recognition in noise and in reverberation by japanese and american listeners," *Journal of the Acoustical Society of America*, vol. 88, pp. 663–666, 1990.

[18] A.K. Nábělek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *Journal of Speech and Hearing Research*, vol. 24, pp. 375–383, 1981.

[19] R.H. Bolt and A.D. MacDonald, "Theory of speech masking by reverberation," *Journal of the Acoustical Society of America*, vol. 21, pp. 577–580, 1949.

[20] A.K. Nábělek, T.R. Letowski, and F.M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, 1989.

[21] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Communication*, vol. 24, no. 1, pp. 5–17, Apr. 1998.

[22] K. Demuynck, J. Duchateau, D. van Compernolle, and P. Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Communication*, vol. 30, no. 1, pp. 37–53, Jan. 2000.

[23] L.R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[24] K. Lebart, J.M. Boucher, and P.N Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, 2001.

[25] F.S. Pacheco and R. Seara, "Spectral subtraction for reverberation reduction applied to automatic speech recognition," in *Proc. of the fifth international telecommunications symposium (ITS2006)*, Fortaleza-CE, Brazil, Sept. 2006, vol. 4, pp. 581–584.

[26] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[27] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'91)*, 1991, vol. 2, pp. 977–980.

[28] O. Viikki and L. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[29] A. Sehr, M. Zeller, and W. Kellermann, "Hands-free speech recognition using a reverberation model in the feature domain," in *Proc. of the European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy, Sept. 2006.

[30] 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. of the 10th International Workshop of Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006, pp. 1–4.

[31] E.A.P. Habets, "Room Impulse Response Generator," *Internal Report*, pp. 1–17, 2006.

[32] E.A.P. Habets, "Single-Channel Speech Dereverberation based on Spectral Subtraction," in *Proc. of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC'04)*, Veldhoven, Netherland, Nov. 2004, pp. 250–254.

[33] E.A.P. Habets, "Experimental Results of a Multi-Channel Speech Dereverberation Algorithm based on a Statistical Model of Late Reverberation," in *Proc. of the first annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS'05)*, Antwerpen, Belgium, Apr. 2005, pp. 11–14.

[34] E.A.P. Habets, "Multi-Channel Speech Dereverberation based on a Statistical Model of Late Reverberation," in *Proc. of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, Mar. 2005, pp. 173–176.

[35] E.A.P. Habets, "Speech Dereverberation based on a Statistical Model of Late Reverberation using a Linear Microphone Array," in *Proc. of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA'05)*, Piscataway NJ, USA, Mar. 2005, pp. d7–d8.

[36] E.A.P. Habets, I. Cohen, and S. Gannot, "MMSE Log Spectral Amplitude Estimator for Multiple Interferences," in *Proc. of the 10th International Workshop of Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006, pp. 1–4.

[37] E.A.P. Habets and P.C.W. Sommen, "Speech Dereverberation using Spectral Subtraction and a Generalized Statistical Reverberation Model," *Submitted to Elsevier's Speech Communications journal*, June 2006.

[38] E.A.P. Habets and S. Gannot, "Dual-Microphone Speech Dereverberation using a Reference Signal," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Honolulu, USA, Apr. 2007, vol. IV, pp. 901–904.

[39] E.A.P. Habets, S. Gannot, and I. Cohen, "Dual-Microphone Speech Dereverberation in a Noisy Environment," in *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT'06)*, Vancouver, Canada, Aug. 2006.

[40] E.A.P. Habets, I. Cohen, S. Gannot, and P.C.W. Sommen, "Joint Dereverberation and Residual Echo Suppression of Speech Signals in a Noisy Environment," *Submitted to IEEE Transactions of Audio, Speech, and Language Processing*, June 2006.

[41] H. Kuttruff, *Room Acoustics*, Spon Press, London, fourth edition, 2000.

[42] M.R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, pp. 299–306, 1954.

[43] D. Cole, M. Moody, and S. Sridharan, "Position-Independent Enhancement of Reverberant Speech," *Journal of the Audio Engineering Society*, vol. 45, no. 3, pp. 142–147, 1997.

[44] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, no. 2, pp. 217–228, 1985.

[45] M. Omura, M. Yada, H. Sruwatari, S. Kajita, K. Takeda, and F. Itakura, "Compensating of Room Acoustic Transfer Functions Affected by Change of Room Temperature," in *Proc. of the IEEE International Conference*, Mar. 1999, vol. 2, pp. 941–944.

[46] J. Mourjopoulos and M. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, 1991.

[47] J. Mourjopoulos and P. Clarkson, "Dereverberation of speech using optimum control," in *Proc. of the International Conference Digital Signal Processing '84*, 1984, pp. 415–419.

[48] J. Mourjopoulos, "Digital equalization of room acoustics," *Journal of the Audio Engineering Society*, vol. 42, no. 11, pp. 884–900, 1994.

[49] A. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice Hall Inc., Englewood Cliff, NJ, 1975.

[50] J. Mourjopoulos, "Digital equalization methods for audio systems," *Journal of the Audio Engineering Society*, 1988.

[51] 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, 1994.

[52] M. Tohyama, "Transfer function phase and trucated impulse responses," *Journal of Acoustical Society of America*, vol. 5, no. 86, pp. 2025–2029, 1989.

[53] S. Gudvangen and S.J. Flockton, "Modelling of acoustic transfer functions for echo cancellers," in *IEE Proceedings of Vision, Image and Signal Processing*, 1995, vol. 142, pp. 47–51.

[54] S. Gudvangen and D. Florencio, "Comparison of pole-zero and all-zero modelling of acoustics tranfer functions," *Electronics Letters*, vol. 28, no. 21, pp. 1976–1978, 1992.

[55] A.P. Liavas and P.A. Regalia, "Acoustic Echo cancellation: Do IIR models offer better modeling capabilities than their FIR counterparts?," *IEEE Trans. Signal Processing*, vol. 46, no. 9, pp. 2499–2504, 1998.

[56] J. Mourjopoulos, A. Tsopanoglou, and N. Fakotakis, "A vector quantization approach for room transfer function classification," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'91)*, 1991, vol. 5, pp. 3593–3596.

[57] H. Wang and F. Itakura, "Dereverberation of speech signals based on sub-band envelope estimation," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E74-A, no. 11, pp. 3576–3583, 1991.

[58] H. Wang and F. Itakura, "An implementation of multi-microphone dereverberation approach as a preprocessor to the word recognition system," *Journal of the Acoustical Society of Japan*, vol. 13, no. 5, pp. 285–293, 1992.

[59] Y. Haneda, S. Makino, and Y. Kaneda, "Modeling of a room transfer function using common acoustical poles," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'92)*, 1992, vol. 2, pp. 213–216.

[60] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer functions," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 188–196, 1999.

[61] W.C. Sabine, *Collected Papers on Acoustics (Originally 1921)*, Los Altos, CA: Peninsula Publishing, 1993.

[62] M.R. Schroeder, "Frequency correlation functions of frequency responses in rooms," *Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[63] J.D. Polack, *La transmission de l'énergie sonore dans les salles*, Thèse de doctorat d'etat, Université du Maine, La mans, 1988.

[64] P.A. Nelson and S.J. Elliott, *Active Control of Sound*, Academic, London, 1993.

[65] M.R. Schroeder, "On frequency response curves in rooms: Comparison of experimental, theoretical and monte-carlo results for the average frequency spacing between maxima," *Journal of the Acoustical Society of America*, vol. 34, no. 1, pp. 76–80, 1962.

[66] F. Talantzis and D.B. Ward, "Robustness of multichannel equalization in an acoustic reverberant environment," *Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 833–841, 2003.

[67] D.B. Ward, "On the performance of acoustic crosstalk cancellation in a reverberrant environment," *Journal of Acoustical Society of America*, vol. 110, pp. 1195–1198, 2001.

[68] N.D. Gaubitch and P.A. Naylor, "Analysis of the dereverberation performance of microphone arrays," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'05)*, 2005, pp. 121–125.

[69] N.D. Gaubitch, D.B. Ward, and P.A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *Journal of the Acoustical Society of America*, vol. 120, pp. 4031–4039, Dec. 2006.

[70] J.A. Moorer, "About this reverberation business [computer music]," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979.

[71] W. Reichardt and U. Lehmann, "Raumeidruck als oberbegriff von raumlichkeit und halligkeit, erlauterungen des raumeindrucksmasses," *Acustica*, vol. 40, pp. 174–183, 1978.

[72] J.D. Polack, "Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2, pp. 235–244, 1993.

[73] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S.Edelman, and M.C. Thompson, "Measurement of correlation coecients in reverberant sound fields," *Journal of the Acoustical Society of America*, vol. 27, no. 6, pp. 1072–1077, 1955.

[74] M.R. Schroeder, "Effect of frequency and space averaging on the transmission responses of multimode media," *Journal of the Audio Engineering Society*, vol. 46, no. 2A, pp. 277–283, Aug. 1969.

[75] M.R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, no. 6, pp. 1187–1188, June 1965.

[76] M.R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

[77] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, Mar. 1992, vol. 2, pp. 221–224.

[78] L.G. Johansen and P. Rubak, "The Excess Phase in Loudspeaker/Room Transfer Functions: Can It Be Ignored in Equalization Tasks?," *Journal of the Audio Engineering Society*, 1996.

[79] A. Oppenheim and J.S. Lim, "The importance of phase in signals," in *Proc. of the IEEE Proceedings Special Issue on Digital Image Processing*, 1981, vol. 69, pp. 529–541.

[80] L.G. Johansen and P. Rubak, "Investigating speech quality by homomorphic deconvolution," in *Proc. of the first European Conference on Signal prediction and Analysis*, Prague, 1997, pp. 327–330.

[81] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[82] A. Kulowski, "Algorithmic representation of the ray tracing technique," *Applied Acoustics*, vol. 18, no. 6, pp. 449–469, 1985.

[83] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small Room Acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[84] M. Kleiner, B.I. Dalenbäck, and P. Svensson, "Auralization - an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, Nov. 1993.

[85] A. Pietrzyk, "Computer modeling of the sound field in small rooms," in *Proc. of the 15th AES Int. Conf. on Audio, Acoustics & Small Spaces*, Copenhagen, Denmark, Oct. 1998, vol. 2, pp. 24–31.

[86] D. Botteldoore, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *Journal of the Acoustical Society of America*, vol. 98, no. 6, pp. 3302–3308, 1995.

[87] L. Savioja, J. Backman, A. Järvinen, and T. Takala, "Waveguide mesh method for low-frequency simulation of room acoustics," in *Proc. of the 15th Int. Congr. Acoust. (ICA'95)*, Trondheim, Norway, June 1995, vol. 2, pp. 1–4.

[88] M.J. Crocker, *Handbook of Acoustics*, Wiley-Interscience, 1998.

[89] F.J. MacWilliams and N.J.A. Sloane, "Pseudo-random sequences and arrays," in *Proc. of the IEEE*, Dec. 1976, vol. 64, pp. 1715–1729.

[90] J. Vanderkooy, "Aspects of mls measuring systems," *Journal of the Audio Enigeering Society*, vol. 42, pp. 219–231, 1994.

[91] H. Alrutz and M.R. Schroeder, "A fast hadamard transform method for the evaluation of measurements using pseudorandoom test signals," in *Proc. of 11th International Congress on Acoustics*, Paris, France, 1983, vol. 6, pp. 235–238.

[92] L.E. Ryall, "Improvements in electric signal amplifiers incorporating voice-operated devices," G.B. Patent No. 509613, 1939.

[93] P.A. Naylor and N.D. Gaubitch, "Speech dereverberation," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'05)*, 2005.

[94] J. Hardwick, C.D. Yoo, and J.S. Lim, "Speech enhancement using the dual excitation speech model," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'93)*, 1993, pp. 367–370.

[95] C.D. Yoo, "Speech enhancement based on the generalized dual excitation model with adaptive analysis window," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, 1995, vol. 1, pp. 832–835.

[96] M.S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, 1998, vol. 6, pp. 3613–3616.

[97] M. Brandstein and S. Griebel, "Explicit Speech Modeling for Microphone Array Applications," in *Microphone arrays: signal processing techniques and applications*, M. Brandstein and D. Ward, Eds., book chapter 7, pp. 133–153. Springer, 2001.

[98] H. Attias and L. Deng, "Speech Denoising and Dereverberation Using Probabilistic Models," *Advances in Neural Information Processing Systems (NIPS'00)*, vol. 13, pp. 758–764, 2001.

[99] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, New York: MacMillan, 1993.

[100] J.B. Allen, "Synthesis of pure speech from a reverberant signal," U.S. Patent No. 3786188, 1974.

[101] S. Griebel and M. Brandstein, "Wavfelet tranform extrama clustering for multi-channel speech deverberation," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99)*, 1999.

[102] S. Griebel and M. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Nov. 2001, vol. 1, pp. 201–204.

[103] B. Yegnanarayana, "Enhancement of reverberant speech using LP residual," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, 1998, vol. 1, pp. 405–408.

[104] B. Yegnanarayana and P.S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 267–281, 2000.

[105] B.W. Gillespie, H. Malvar, and D. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, 2001, vol. 6, pp. 3701–3074.

[106] N. Mitianoudis M. Tonelli and M.E. Davies, "Maximum Likelihood approach to blind audio de-reverberation," in *Proc. of the 7th International Conference on Digital Audio Effects (DAFX'04)*, Naples, Italy, Oct. 2004, pp. 1–6.

[107] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, 2001.

[108] M. Tonelli, M.G. Jafari, and M.E. Davies, "A multi-channel Maximum Likelihood approach to de-reverberation," in *Proc. of the European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy, Sept. 2006.

[109] B. Yegnanarayana, "Speech enhancement using excitation source information," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, 2002, vol. 1, pp. 541–544.

[110] N.D. Gaubitch, P.A. Naylor, and D.B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. of the European Signal Processing Conference (EUSIPCO'04)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[111] N.D. Gaubitch, P.A. Naylor, and D. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'03)*, Kyoto, Japan, 2003, pp. 99–102.

[112] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[113] D.A. Berkley and O.M.M. Mitchell, "Removing reverberative echo components in speech signals," U.S. Patent No. 4166924, 1979.

[114] T. Langhans and H. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'82)*, 1982, vol. 7, pp. 156–159.

[115] M.R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acustica*, vol. 49, pp. 179–182, 1981.

[116] H.G. Hirsch, *Signal Processing IV: Theories and Applications*, chapter Automatic Speech Recognition in Rooms, pp. 1177–1180, Elsevier Saence Publishers B.V. (North Holland), EURASIP, 1988.

[117] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. of the fourth International Conference on Spoken Language Processing (ICSLP'96)*, 1996, vol. 2, pp. 889–892.

[118] J. Mourjopoulos and J.K. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'83)*, 1983, pp. 1144–1147.

[119] S. Hirobayashi, H. Nomura, T. Koike, and M. Tohyama, "Speech waveform recovery from a reverberant speech signal using inverse filtering of power envelope transfer function," *IEEE Trans. of the Institute of Electronics Information and Communication Engineers*, vol. J81–A, no. 10, pp. 1323–1330, 1998.

[120] K. Sakata M. Unoki, M. Furukawa and M. Akagi, "An improved method based on the mtf concept for restoring the power envelope from a reverberant signal," *Acoustical Science and Technology*, vol. 25, no. 4, pp. 232–242, 2004.

[121] K. Sakata M. Unoki, M. Furukawa and M. Akagi, "A speech dereverberation method based on mtf concept in power envelope restoration," *Acoustical Science and Technology*, vol. 25, no. 4, pp. 243–254, 2004.

[122] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[123] L. Danilenko, *Binaurales Hören in nichtstationaren diffusen Schallfeld*, Ph.D. Thesis, University of Aachen, FDR, 1968.

[124] P. Bloom, "Evaluation of a dereverberation process by normal and impaired listeners," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'80)*, 1980, pp. 500–503.

[125] P. Bloom and G. Cain, "Evaluation of Two Input Speech Dereverberation Techniques," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'82)*, 1982, vol. 1, pp. 164–167.

[126] A. Hussain, "Intelligibility assessment of a multi-band speech enhancement scheme," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, 2000, vol. 2, pp. 1045–1048.

[127] A. Hussain and D. Campbell, "Intelligibility improvements using binaural diverse subband processing applied to speech corrupted with automobile noise," in *IEE Proceedings of Vision, Image and Signal Processing*, 2001, vol. 148, pp. 127–132.

[128] M. Wu and D. Wang, "A two-stage algorithm for enhancement of reverberant speech," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, 2005, vol. 1, pp. 1085–1088.

[129] M. Wu and D. Wang, "A two-stage algorithm for enhancement of reverberant speech," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[130] H.L. Van Trees, *Optimum Array Processing*, Detection, Estimation and Modulation Theory. Wiley, 2002.

[131] I.A. McCowan, "Microphone Arrays: A Tutorial," Report, 2001.

[132] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, 1993.

[133] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms," *Journal of Acoustical Society of America*, vol. 5, no. 78, pp. 1508–1518, 1985.

[134] S. Doclo, *Multi-microphone noise reduction and dereverberation techniques for speech applications*, Ph.D. Thesis, Katholieke Universiteit Leuven, Belgium, May 2003.

[135] J. Bitzer, K. Simmer, and K.D. Kammeyer, "Multi-Microphone Noise Reduction by Post-Filter and Superdirective Beamformer," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'99)*, Sept. 1999, pp. 100–103.

[136] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[137] S. Affès 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, 1997.

[138] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principal component analysis," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Mar. 2005, vol. 4, pp. 797–800.

[139] L.J. Griffiths and C.W. Jim, "An Alternate Approach to Linearly Constrained Adaptive Beamforming," *IEEE Transaction on Antennas and Propagation*, vol. 1, no. 30, pp. 27–34, 1982.

[140] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[141] 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, 2001.

[142] J. Bitzer, K. Simmer, and K.D. Kammeyer, "Theoretical Noise Reduction Limits of the Generalized Sidelobe Canceller (GSC) for Speech Enhancement," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, Mar. 1999, vol. 5, pp. 2965–2968.

[143] C. Marro, Y. Mahieux, and K. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[144] I. Cohen, S. Gannot, and B. Berdugo, "An Integrated Real-Time Beamforming and Postfiltering System for Non-Stationary Noise Environments," *EURASIP Journal on Applied Signal Processing, vol. special issue on Signal Processing for Acoustic Communication Systems*, vol. 2003, no. 11, pp. 10641073, Oct. 2003.

[145] S. Gannot and I. Cohen, "Speech Enhancement Based on the General Transfer Function GSC and PostFiltering," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[146] S. Haykin, *Blind Deconvolution*, Prentice Hall information and system sciences series. Prentice Hall, fourth edition, 1994.

[147] S. Haykin, *Unsupervised Adaptive Filtering*, John–Wiley&Sons, second edition, 2000.

[148] J. Hopgood, *Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments*, PhD Thesis, Cambridge University, 2001.

[149] T.H. Li, "Estimation and blind deconvolution of autoregressive systems with non-stationary binary inputs," *Journal time series analysis*, vol. 14, no. 6, pp. 575–588, 1993.

[150] R. Chen and T.H. Li, "Blind restoration of linearly degraded discrete signals by Gibbs sampling," *IEEE Trans. Signal Processing*, vol. 43, pp. 2410–2413, 1995.

[151] T.H. Li and K. Mbarek, "A blind equalizer for nonstationary discrete-valued signals," *IEEE Trans. Signal Processing*, vol. 45, pp. 247–254, 1997.

[152] O. Cappé, A. Doucet, M. Lavielle, and E. Moulines, "Simulation-based methods for blind maximum-likelihood filter deconvolution," *IEEE Trans. Signal Processing*, vol. 73, no. 1, pp. 3–25, 1999.

[153] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE Trans. Signal Processing*, vol. 13, pp. 43–64, 1996.

[154] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'88)*, vol. 36, no. 2, pp. 145–152, 1988.

[155] M. Gürelli and C. Nikias, "EVAM: An Eigenvector-Based Algorithm for Multichannel Blind Deconvolution of Input Colored Signals," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 134–149, 1995.

[156] S. Gannot and M. Moonen, "Subspace Methods for Multi-Microphone Speech Dereverberation," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'01)*, Darmstadt, Germany, 2001.

[157] S. Gannot and M. Moonen, "Subspace Methods for Multimicrophone Speech Dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[158] M. Triki and D.T.M. Slock, "Blind dereverberation of quasi-periodic sources based on multichannel linear prediction," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'05)*, Eindhoven, Netherlands, Sept. 2005.

[159] M. Triki and D.T.M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06), Toulouse, France*, May 2006, vol. 5, pp. 97–100.

[160] M. Triki and D.T.M. Slock, "Iterated delay and predict equalization for blind speech dereverberation," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006.

[161] M. Miyoshi, "Estimating AR parameter-sets for linear-recurrent signals in convolutive mixtures," in *Proc. of the ICA'03*, 2003, pp. 585–589.

[162] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation of speech signals based on linear prediction," in *Proc. of the 8th International Conference on Spoken Language Processing ICSLP'04, Jeju Island, Korea*, Oct. 2004, vol. 2, pp. 877–881.

[163] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," *Acoustical Science and Technology*, vol. 26, no. 5, pp. 432–439, 2005.

[164] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 2, pp. 430–440, 2006.

[165] H. Wang and F. Itakura, "An Approach to Dereverberation Using Multi-microphone Sub-band Envelope Estimation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'91)*, 1991, vol. 2, pp. 953–956.

[166] T.F. Quatieri, *Discrete-Time Speech Signal Processing*, NJ: Prentice Hall, 2002.

[167] M. Delcroix, T. Hikichi, and M. Miyoshi, "On the use of lime dereverberation algorithm in an acoustic environment with a noise source," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, 2006, vol. 1, pp. 825–828.

[168] R. Kennedy and B. Radlović, "Iterative cepstrum-based approach for speech dereverberation," in *Proc. of the fifth International Symposium on Signal Processing and its Applications (ISSPA'99)*, 1999, vol. 1, pp. 55–58.

[169] A. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, 1994, vol. 1, pp. I–12.

[170] S. Subramaniam, A. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 392–396, 1996.

[171] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Implementation and Effects of Single Channel Dereverberation based on the Harmonic Structure of Speech," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'03)*, 2003, pp. 91–94.

[172] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, 2003, vol. 1, pp. 92–95.

[173] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Harmonicity Based Dereverberation for Improving Automatic Speech Recognition Performance and Speech Intelligibility," *IEICE Trans. on Fundamentals of Electronics Communications and Computer Sciences*, vol. E88-A, no. 7, pp. 1724–1731, 2005.

[174] T. Nakatani, B. Juang, K. Kinoshita, and M. Miyoshi, "Harmonicity based dereverberation with maximum a posteriori estimation," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*, 2005, pp. 94–97.

[175] J. Mourjopoulos, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'82)*, May 1982, pp. 858–1861.

[176] S. Treitel and E. Robinson, "The design of High-resolution digital filters," *IEEE Transaction on Geoscience Electronics*, vol. GE-4, no. 1, pp. 25–38, 1966.

[177] W. Putnam, "A numerical investigation of the invertibility of room transfer functions," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'95)*, 1995, pp. 249–252.

[178] "Methods for subjective determination of transmission quality," Recommendation P.800, International Telecommunications Union (ITU-T), Feb. 1996.

[179] D. Picovici and A.E. Mahdi, "Towards non-intrusive speech quality assessment for modern telecommunications," in *First Joint IEI/IEE Symposium of Telecom Systems Research*, Nov. 2001.

[180] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, *Objective Measures of Speech Quality*, O. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1988.

[181] "Perceptiial evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Recommendation P.862, International Telecommunications Union (ITU-T), Feb. 2001.

[182] H.J.M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[183] R. Plomp, *Hearing: physiological bases and psychophysics (R. Klinke, R. Hartmann, eds.)*, chapter The role of modulations in hearing, pp. 270–275, New York: Springer, 1983.

[184] M.B. Priestly, *Non-linear and Non-stationary Time Series Analysis*, Academic Press, New York, NY, USA, 1988.

[185] W. Yang, M. Dixon, and R. Yantorno, "A modified bark spectral distortion measure which uses noise masking threshold," in *IEEE Speech Coding Workshop*, Pocono Manor, 1997, pp. 55–56.

[186] J.Y.C. Wen and P. Naylor, "An evaluation measure for reverberant speech using tail decay modelling," in *Proc. of the European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy, 2006, pp. 1–4.

[187] S. Furui, "On the role of spectral transition for speech perception," *Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[188] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[189] I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[190] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[191] J. Wexler and S. Raz, "Discrete Gabor expansions," *Speech Processing*, vol. 21, no. 3, pp. 207–220, Nov. 1990.

[192] R.E. Crochiere and L.R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[193] J.W. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

[194] Y. Ephraim, H. Lev-Ari, and W.J.J. Roberts, *The Electronic Handbook*, chapter A brief survey of speech enhancement, p. 1512 1526, CRC Press, second edition, 2005.

[195] Y. Ephraim and I. Cohen, *The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing*, chapter Recent advancements in speech enhancement, pp. 15–12 – 15–26, R.C. Dorf, Eds. CRC Press, third edition, 2006.

[196] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," in *Proc. of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA'05)*, Piscataway NJ, USA, Mar. 2005.

[197] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 112–120, Apr. 1979.

[198] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[199] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'79)*, Washington, DC, Apr. 1979, pp. 208–211.

[200] Z. Goh, K.-C. Tan, and T.G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 287–292, May 1998.

[201] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 328–337, July 1998.

[202] H. Gustafsson, S.E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[203] D.E. Tsoukalas J.N. Mourjopoulos and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 497–514, Nov. 1997.

[204] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[205] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[206] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 341351, Sept. 2002.

[207] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[208] I. Cohen, "Relaxed Statistical Model for Speech Enhancement and A Priori SNR Estimation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 870–881, Sept. 2005.

[209] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Speech Processing*, vol. 86, no. 4, pp. 698–709, Apr. 2006.

[210] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[211] P.J. Wolfe and S.J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process., Special Issue on Digital Audio for Multimedia Communications*, vol. 2003, no. 10, pp. 1043–1051, Sept. 2003.

[212] P.C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

[213] B.H. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 6, pp. 1404–1413, Dec. 1985.

[214] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[215] H. Sheikhzadeh and L. Deng, "Waveformbased speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 5, pp. 80–91, Jan. 1994.

[216] Y. Ephraim and N. Merhav, "Hidden markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1568, June 2002.

[217] H. Sameti, H. Sheikhzadeh, L. Deng, and R.L. Brennan, "Mm-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 445–455, Sept. 1998.

[218] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Inc., Upper Saddle River, New Jersey, 1993.

[219] F. Jelinek, *Statistical methods for speech recognition*, MIT Press, Cambridge, MA, 1998.

[220] Y. Ephraim and H.L.V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[221] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 497–507, Sept. 2000.

[222] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[223] Y. Hu and P.C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 4, pp. 334–341, July 2003.

[224] S.H. Jensen, P.C. Hansen, S.D. Hansen, and J.A. Sorensen, "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[225] S. Doclo and M. Moonen, "Gsvd-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[226] J. Benesty and Y. Huang, Eds., *Adaptive Signal Processing : Applications to Real-World Problems*, Signals and Communication Technology. Springer, 2003.

[227] J. Poruba, "Speech enhancement based on nonlinear spectral subtraction," in *Proc. of the fourth IEEE International Caracas Conference on Devices, Circuits and Systems*, 2002, pp. T031–1–T031–4.

[228] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[229] I.A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, 2000, vol. 3, pp. 1723–1726.

[230] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, 2002, vol. 1, pp. 905–908.

[231] P.M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.

[232] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Recommendation P.835, International Telecommunications Union (ITU-T), Feb. 2003.

[233] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engineering Society, 103th Convenstion*, Aug. 1997.

[234] A.J. Accardi and R.V. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, 1999, vol. 1, pp. 201–204.

[235] T.J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *Journal of the Audio Engineering Society*, vol. 49, no. 4, pp. 219–230, 2001.

[236] R. Ratnam, D.L. Jones, B.C. Wheeler, W.D. O'Brien Jr., C.R. Lansing, and A.S. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[237] A. Varga and H.J.M. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, vol. 12, pp. 247–251, July 1993.

[238] G. Schmidt, "Applications of acoustic echo control an overview," in *Proc. of the European Signal Processing Conference (EUSIPCO'04)*, Vienna, Austria, 2004.

[239] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley-IEEE Press, June 2004.

[240] V. Myllyla, "Residual echo filter for enhanced acoustic echo control," *Signal Processing*, vol. 86, no. 6, pp. 1193–1205, June 2006.

[241] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic Echo Control - An Application of Very-High-Order Adaptive Filters," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, 1999.

[242] E. Hänsler, "The hands-free telephone probleman annotated bibliography," *Signal Processing*, vol. 27, no. 3, pp. 259–271, 1992.

[243] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[244] G. Enzner, *A Model-Based Optimum Filtering Approach to Acoustic Echo Control: Theory and Practice*, Ph.D. Thesis, RWTH Aachen University, Wissenschaftsverlag Mainz, Aachen, Germany, Apr. 2006, ISBN 3-86130-648-4.

[245] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, 1997, vol. 1, pp. 307–310.

[246] Y. Grenier, M. Xu, J. Prado, and D. Liebenguth, "Real-time implementation of an acoustic antenna for audio-conference," in *Proc. of the International Workshop on Acoustic Echo Control (IWAENC'89)*, Berlin, Sept. 1989.

[247] M. Xu and Y.Grenier, "Acoustic echo cancellation by adaptive antenna," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'89)*, Berlin, Sept. 1989.

[248] H. Yasukawa, "An Acoustic Echo Canceller with Sub-Band Noise Cancelling," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E75-A, no. 11, pp. 1516–1523, 1992.

[249] R. Le Bouquin Jeannès, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 808–820, 2001.

[250] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[251] R. Martin and P. Vary, "Combined acoustic echo cancellation, dereverberation and noise reduction: a two microphone approach," *Annales des Telecommunications*, vol. 49, no. 7-8, pp. 429–438, 1994.

[252] M. Dörbecker and S. Ernst, "Combination of Two-Channel Spectral Subtraction and Adaptive Wiener Post-Filtering for Noise Reduction and Dereverberation," in *Proc. of the European Signal Processing Conference (EUSIPCO'96)*, Triest, 1996.

[253] J. Allen and L. Radiner, "A unified approach to short-time fourier analysis and synthesis," *Proc. of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[254] J. Benesty and S.L. Gay, "An improved PNLMS algorithm," in *Proc. of the 27th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, 2002, pp. 1881–1884.

[255] E. Hänsler and G. Schmidt, "Hands-free telephones - joint control of echo cancellation and postfiltering," *Signal Processing*, vol. 80, pp. 2295–2305, 2000.

[256] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Processing*, vol. 86, pp. 1124–1139, June 2006.

[257] T. van Waterschoot, G. Rombouts, and M. Moonen, "MSE optimal regularization of APA and NLMS algorithms in room acoustic applications," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006, pp. 1–4.

[258] M. Karjalainen, P. Antsalo, A. Mkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *Journal of the Audio Engineering Society*, vol. 11, pp. 867–878, 2002.

[259] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, pp. 21–32, 1998.

[260] D.W.E. Schobben and P.C.W. Sommen, "On the performance of too short adaptive FIR filters," in *Proc. of the 8th Annual ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing CSSP-97*, J.P. Veen, Ed., Utrecht, Netherlands, 1997, STW, Technology Foundation, pp. 545–549, ISBN 90-73461-12-X.

[261] K. Mayyas, "Performance analysis of the deficient length LMS adaptive algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 53, no. 8, pp. 2727–2734, 2005.

[262] J.E. Summers, R.R. Torres, and Y. Shimizu, "Statistical-acoustics models of energy decay in systems of coupled rooms and their relation to geometrical acoustics," *Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 958–969, Aug. 2004.

[263] A.D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, Acoustical Society of America, 1991.

[264] E. Hänsler and G. Schmidt, "Hands-free telephones - joint control of echo cancellation and postfiltering," *Signal Processing*, vol. 80, pp. 2295–2305, 2000.

Appendix A

Room Impulse Response Generator

Many people who are working in the field of acoustic signal processing reach a point where they want to simulate room acoustics. The image method, which was developed by Allen and Berkley in 1979, is probably one of the methods most commonly used in the acoustic signal processing community. Therefore it will be discussed in more detail. A mex-function, which can be used in Matlab®, has been created to generate multi-channel Room Impulse Responses using the image method. This function enables the users to control the reflection order, room dimension, and microphone directivity.

A.1 Allen and Berkley's Image Method

The image model (Section A.1.1) can be used to simulate the reverberation in a room for a given source and microphone location. The system is treated as an Linear Time-Invariant (LTI) system whose impulse response consists of a set of delayed impulses of gradually decreasing amplitudes. Allen and Berkley developed an efficient method [83], using the image model, to compute a Finite Impulse Response (FIR) for rectangular rooms. This method and some additional refinements will we explained in Section A.1.2.

A.1.1 Image model

Fig. A.1 shows a sound source S located near a rigid reflecting wall. At destination D two signals arrive, one from the direct path and a second one from the reflection. The path length of the direct path can be directly calculated from the known locations of the source and the destination. Also shown is an image of the source, S', located behind the wall at a distance equal to the distance of the source from the wall. Because of symmetry, the triangle SRS' is isosceles and therefore the path length $SR + RD$ is

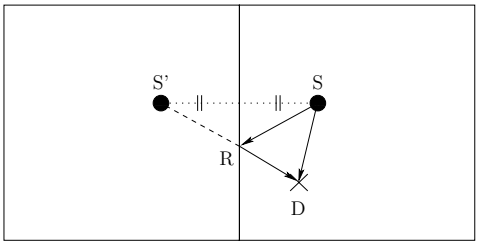


Figure A.1 *Direct path and one path involving one reflection obtained using a first level image.*

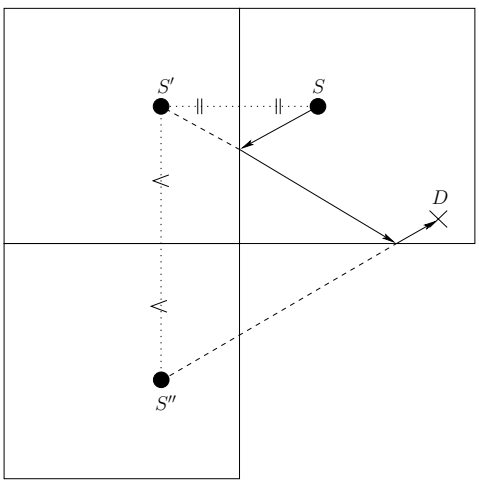


Figure A.2 *Path involving two reflections obtained using two levels of images.*

the same as $S'D$. To compute the path length of the reflected path, we can construct an image source and compute the distance between destination and image source. Also, the fact that we are computing the distance using an image means that there was one reflection in the path. Fig. A.2 shows a path involving two reflections. The length of this path can be obtained from the length of $S''D$. In Fig. A.3 the length of a path involving three reflections is obtained from the length of $S'''D$. These figures can also be extended to three dimensions to take into account reflections from the ceiling and the floor. In general the path lengths (and thus the delays) of reflections can be obtained by computing the distance between the source images and the destination. The strength of the reflection can be obtained from the path length and the number of reflections involved in the path. The number of reflections involved in the path is equal to the level of images that was used to compute the path.

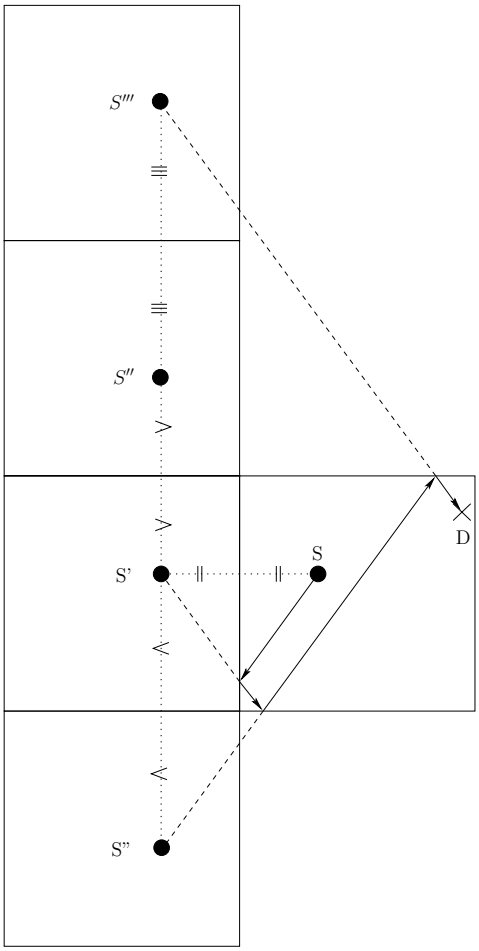


Figure A.3 *Path involving three reflections obtained using three levels of images.*

A.1.2 Image method

Consider a rectangular room with length, width and height given by L_x, L_y and L_z. Let the sound source be at a location represented by the vector $\mathbf{r}_s = [x_s\ y_s\ z_s]$ and let the microphone be at a location represented by the vector $\mathbf{r} = [x\ y\ z]$. Both vectors are with respect to the origin, which is located at one of the corners of the room. The vector joining the microphone to any of the first level images can be written as

$$\mathbf{R}_p = [x_s - x + 2qx \quad y_s - y + 2jy \quad z_s - z + 2kz]. \tag{A.1}$$

Each of the elements in the triplet $p = (q, j, k)$ can take on values 0 or 1, which means that there are 8 different combinations, $(0, 0, 0)$ to $(1, 1, 1)$. The eight possible combinations are part of a set $\mathcal{P}$. When the value of p is 1 in any dimension, then an image of the source in that direction is considered. To consider images of any level,

we add the vector $\mathbf{R}_u$ to $\mathbf{R}_p$ where

$$\mathbf{R}_u = 2[nL_x \quad lL_y \quad mL_z]. \tag{A.2}$$

Each of the elements of the triplet $u = (n, l, m)$ takes on values between -N and +N, depending on the maximum level of images that we would like to consider. We define a set $\mathcal{U}$ which contains all $(2N+1)^3$ combinations. For a given N, this method computes $8(2N+1)^3$ different paths.

The distance between any source image and the microphone can be written as

$$D = \|\mathbf{R}_u + \mathbf{R}_p\|, \tag{A.3}$$

where $\|\cdot\|$ denotes the Euclidian norm. The time delay of arrival of the reflected sound ray corresponding to any image source can be expressed as

$$t_{\text{TDOA}} = \frac{\|\mathbf{R}_u + \mathbf{R}_p\|}{c}, \tag{A.4}$$

where c denotes the sound velocity in meters per second.

Using the Free Space Green's function Eq. 2.12 we can express the transfer function from any image source to the microphone as

$$H(\mathbf{r}, \mathbf{r}_s; \omega) = \frac{\mathrm{e}^{-\iota \frac{\omega}{c} \|\mathbf{R}_u + \mathbf{R}_p\|}}{4\pi \|\mathbf{R}_u + \mathbf{R}_p\|}, \tag{A.5}$$

where $\iota = \sqrt{-1}$. In case of rigid walls, the acoustic transfer function from the source to the receiver can be expressed as [83]

$$H(\mathbf{r}, \mathbf{r}_s; \omega) = \sum_{p \in \mathcal{P}} \sum_{u \in \mathcal{U}} \frac{\mathrm{e}^{-\iota \frac{\omega}{c} \|\mathbf{R}_u + \mathbf{R}_p\|}}{4\pi \|\mathbf{R}_u + \mathbf{R}_p\|}. \tag{A.6}$$

By taking the inverse Fourier transform of Eq. A.6 we obtain the acoustic impulse response $h(\mathbf{r}, \mathbf{r}_s, t)$, i.e.,

$$h(\mathbf{r}, \mathbf{r}_s, t) = \sum_{p \in \mathcal{P}} \sum_{u \in \mathcal{U}} \frac{\delta\left(t - \frac{\|\mathbf{R}_u + \mathbf{R}_p\|}{c}\right)}{4\pi \|\mathbf{R}_u + \mathbf{R}_p\|}. \tag{A.7}$$

Note that Eq. A.7 is the exact solution to the wave equation in a rectangular, rigid-wall, room, and can be derived by calculating the inverse Fourier transform from the solution given by Eq. 2.14 [83].

Under certain assumptions (see [83] for more details) the impulse response in case of non-rigid walls can be derived. The impulse response for this source and microphone location can then be written as [83]

$$h(\mathbf{r}, \mathbf{r}_s, t) = \sum_{p \in \mathcal{P}} \sum_{u \in \mathcal{U}} \beta_{x_1}^{|n-q|} \beta_{x_2}^{|n|} \beta_{y_1}^{|l-j|} \beta_{y_2}^{|l|} \beta_{z_1}^{|m-k|} \beta_{z_2}^{|m|} \frac{\delta\left(t - \frac{\|\mathbf{R}_u + \mathbf{R}_p\|}{c}\right)}{4\pi \|\mathbf{R}_u + \mathbf{R}_p\|}, \tag{A.8}$$

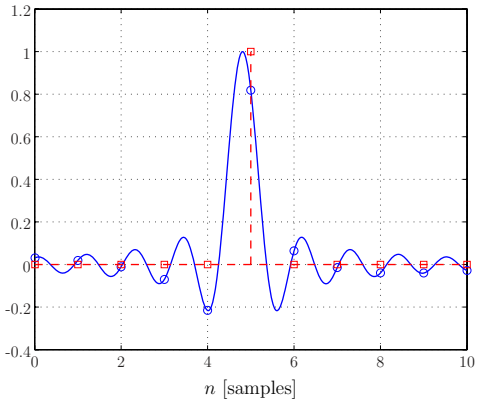


Figure A.4 *Comparison of the shifted (dashed line) and low-pass impulse (solid line) method using a fractional delay of 4.8 samples.*

where the quantities β_{x_1}, β_{x_2}, β_{y_1}, β_{y_2}, β_{z_1}, and β_{z_2} are the reflection coefficients of the six walls.

Once the impulse response has been computed this way, the source signal can be convolved with the impulse response to simulate the signal picked up by the microphone.

An important consideration while simulating the discrete version of this impulse response using a computer is that the delays given by Eq. A.4 do not always fall at sampling instants. This distortion can be ignored in many applications. However, for multiple microphone systems that depend on inter-microphone phase relations, correct simulation of arrival time relationships is critical. One way to reduce this problem is to compute the discrete impulse response at a much higher sampling frequency, decimate the impulse response to the original sampling frequency, and convolve the source signal with it. Peterson suggested another modification to the image method [231]. In this approach, each impulse in Eq. A.8 is replaced by the impulse response of a Hanning-windowed ideal low pass filter of the form

$$h_{lp}(t) = \begin{cases} \frac{1}{2}\left(1+\cos\left(\frac{2\pi t}{T_w}\right)\right)\operatorname{sinc}\left(2\pi F_c t\right), & -\frac{T_w}{2} < t < \frac{T_w}{2}; \\ 0, & \text{otherwise}, \end{cases}, \tag{A.9}$$

where T_w is the width (in time) of the impulse response and F_c is the cut-off frequency of the low-pass filter. All Acoustic Impulse Responses (AIR) used in this dissertation are generated using $T_w = 0.008 f_s$ and $F_c = 1$. Each impulse in Eq. A.8 is replaced by $h_{lp}(t)$ centred at the true delay. By doing this, true delays of arrival of the reflected signals are simulated accurately even at the original low sampling frequency. A comparison of both methods, using a delay of 4.8 samples, is depicted in Fig. A.4. Squares indicate sample values produced by Allen and Berkley's shifted impulse method and circles indicate sample values produced by Peterson's low-pass impulse method. The solid line shows the central portion of the continuous-time low-pass impulse function.

The other consideration while simulating reverberation for a room is the duration of

reverberation or the reverberation time. Formally, the reverberation time is defined as the time required for the intensities of reflected sound rays to be down 60 dB from the direct path sound ray. An empirical formula, known as Sabin-Franklin's formula [263] can be used to relate the reverberation time RT_{60} by,

$$\mathrm{RT}_{60} = \frac{24\ \ln(10)\ V}{c \sum_{i=1}^{6} S_i(1-\beta_i^2)}, \tag{A.10}$$

where β_i and S_i denote the reflection coefficient and the surface of the i^{th} wall. The volume of the room is denoted by V.

A.2 Implementation

The image method as discussed in the previous section has been implemented as a Matlab® mex-function and was written in C++. The resulting Dynamic-Link-Library (DLL) can easily be used within Matlab® as a standard Matlab® function. The C++ implementation is much faster than the equivalent Matlab® implementation. The source-code can be found in Section A.4.

The function *rir_generator* is defined as follows:

```
function [h, beta] = rir_generator(c, fs, r, s, L, beta, nsample,
                                   mtype, order, dim, orientation,
                                   hp_filter);
```

Input parameters:

Parameter	Description
c	sound velocity in m/s.
fs	sampling frequency in Hz.
r	M x 3 matrix specifying the (x,y,z) coordinates of the receiver(s) in m.
s	1 x 3 vector specifying the (x,y,z) coordinates of the source in m.
L	1 x 3 vector specifying the room dimensions (x,y,z) in m.
beta	1 x 6 vector specifying the reflection coefficients $[\beta_{x_1}\ \beta_{x_2}\ \beta_{y_1}\ \beta_{y_2}\ \beta_{z_1}\ \beta_{z_2}]$ or beta = Reverberation Time (RT_{60}) in seconds.

Optional input parameters:

Parameter	Description	Default value
nsample	number of samples to calculate.	$\mathrm{RT}_{60} f_\mathrm{s}$
mtype	type of microphone that is used [‘omnidirectional’, ‘subcardioid’, ‘cardioid’, ‘hypercardioid’, ‘bidirectional’].	‘omnidirectional’
order	reflection order.	-1
dim	room dimension (2 or 3).	3
orientation	specifies the angle (in rad) in which the microphone is pointed.	0
hp_filter	use ‘false’ to disable high-pass filter.	‘true’

Output parameters:

Parameter	Description
h	M x nsample matrix containing the calculated room impulse response(s).
beta	In case a reverberation time is specified as an input parameter the corresponding reflection coefficient is returned.

Multi-Channel Support In case more than one receiver position is specified the function *rir_generator* will calculate all AIRs at once.

Reverberation Time versus Reflection Coefficients The reflection coefficients in Eq. A.8 can be specified using the parameter ’*beta*’. In case ’*beta*’ consists of one element the program assumes a reverberation time (in seconds) is specified. The corresponding average reflection coefficient is calculated using Eq. A.10 and will be returned using the output parameter ‘*beta*’.

Reflection Order and Room Dimension In order to control the complexity of the generated AIR one can control the maximum reflection order using the parameter ‘*order*’. In case the order is chosen ‘-1’ (default value) the maximum amount of reflections, given the desired length of the AIR, is calculated. The dimension of the room can be set using the parameter ‘*dim*’. This value can either be 2 or 3 (default value).

Microphone Directivity The microphone’s directionality, or polar pattern, can also be taken into account. Different kinds of polar patterns are implemented and can be chosen using the parameter ‘*mtype*’. The signal attenuation $A(\theta)$, where θ denotes the direction of arrival, is calculated using the following standard formula:

$$A(\theta) = P + PG\cos(\theta)\,. \tag{A.11}$$

The polar pattern is controlled by P and PG, see Table A.1. The resulting polar patterns for the Omnidirectional, Cardioid, Hypercardioid and Bidirectional microphone are depicted in Fig. A.5.

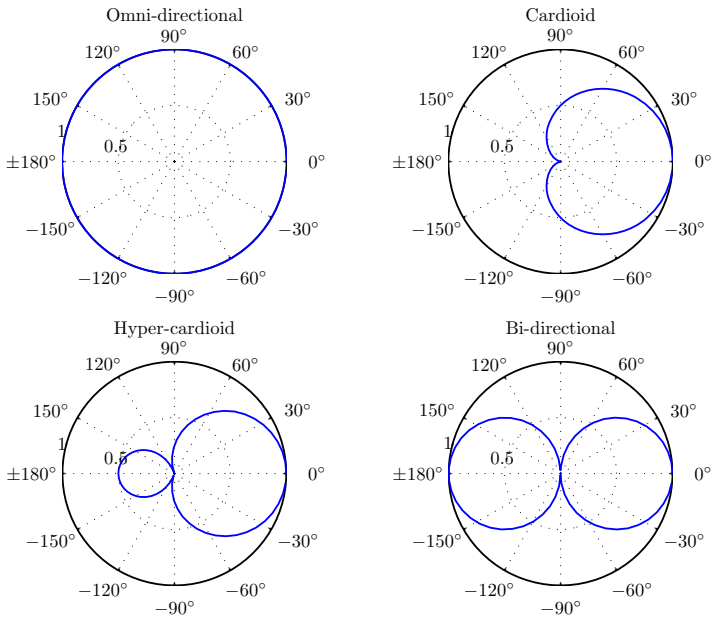


Figure A.5 *Polar plots of four different microphone polar patterns.*

Directivity Pattern	P	PG
Omnidirectional	1	0
Subcardioid	0.75	0.25
Cardioid	0.5	0.5
Hypercardioid	0.25	0.75
Bidirectional	0.25	0.75

Table A.1 *Supported polar patterns and corresponding values for P and PG.*

The angle in which the microphone is pointing can be adjusted with the parameter '*orientation*'. By default the microphone points towards the positive x-axis. Note that the microphone's directionality only takes the azimuth of the received reflection into account. The elevation of the received reflection does not influence the attenuation.

A.3 Examples

In this section some basic and more complex examples are presented in the form of a Matlab® script.

```
c = 340;                        % Sound velocity (m/s)
fs = 16000;                     % Sample frequency (samples/s)
r = [2 1.5 2];                  % Receiver position [x y z] (m)
s = [2 3.5 2];                  % Source position [x y z] (m)
L = [5 4 6];                    % Room dimensions [x y z] (m)
beta = 0.4;                     % Reverberation time (s)
n = 4096;                       % Number of samples

h = rir_generator(c, fs, r, s, L, beta, n);
```

Example A.1 *Simple example to generate one AIR.*

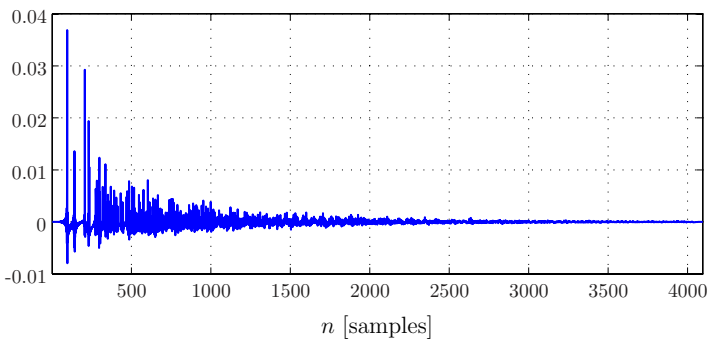


Figure A.6 *Output of Example A.1.*

```
c = 340;                        % Sound velocity (m/s)
fs = 16000;                     % Sample frequency (samples/s)
r = [2 1.5 2];                  % Receiver position [x y z] (m)
s = [2 3.5 2];                  % Source position [x y z] (m)
L = [5 4 6];                    % Room dimensions [x y z] (m)
beta = 0.4;                     % Reverberation time (s)
n = 1024;                       % Number of samples
mtype = 'omnidirectional';      % Type of microphone
order = 2;                      % Reflection order
dim = 3;                        % Room dimension
orientation=0;                  % Microphone orientation (rad)
hp_filter=1;                    % Enable high-pass filter

h = rir_generator(c, fs, r, s, L, beta, n, mtype,order,dim,orientation,hp_filter);
```

Example A.2 *Generate one AIR.*

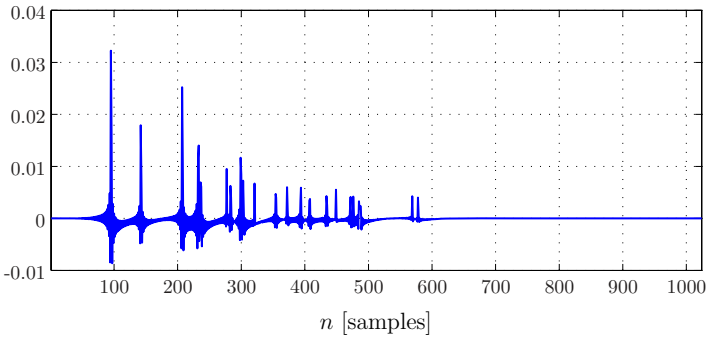


Figure A.7 *Output of Example A.2.*

```
c = 340;                        % Sound velocity (m/s)
fs = 16000;                     % Sample frequency (samples/s)
r = [2 1.5 2 ; 1 1.5 2];        % Receiver positions [x_1 y_1 z_1 ; x_2 y_2 z_2] (m)
s = [2 3.5 2];                  % Source position [x y z] (m)
L = [5 4 6];                    % Room dimensions [x y z] (m)
beta = 0.4;                     % Reverberation time (s)
n = 4096;                       % Number of samples
mtype = 'omnidirectional';      % Type of microphone
order = -1;                     % -1 equals maximum reflection order!
dim = 3;                        % Room dimension
orientation = 0;                % Microphone orientation (rad)
hp_filter = 1;                  % Enable high-pass filter

h = rir_generator(c, fs, r, s, L, beta, n, mtype, order, dim, orientation, hp_filter);
```

Example A.3 *Generate multiple AIRs.*

```
c = 340;                        % Sound velocity (m/s)
fs = 16000;                     % Sample frequency (samples/s)
r = [2 1.5 2];                  % Receiver position [x y z] (m)
s = [2 3.5 2];                  % Source position [x y z] (m)
L = [5 4 6];                    % Room dimensions [x y z] (m)
n = 4096;                       % Number of samples
beta = 0.4;                     % Reverberation time (s)
mtype = 'hypercardioid';        % Type of microphone
order = -1;                     % -1 equals maximum reflection order!
dim = 3;                        % Room dimension
orientation = pi/2;             % Microphone orientation (rad)
hp_filter = 1;                  % Enable high-pass filter

h = rir_generator(c, fs, r, s, L, beta, n, mtype, order, dim, orientation, hp_filter);
```

Example A.4 *Generate one AIR using a hypercardioid microphone.*

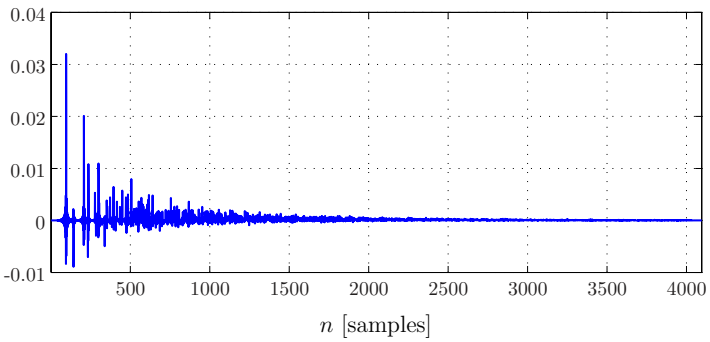


Figure A.8 *Output of Example A.4.*

A reverberant signal can now be created by filtering the anechoic signal with the generated AIR.

```
reverberant_signal = filter(h,1,clean_signal);
```

Example A.5 *Generate reverberant signal.*

A.4 Source Code

```
/*
Program      : Room Impulse Response Generator

Description  : Function for simulating the impulse response of a specified
               room using the image method [1,2].

               [1] J.B. Allen and D.A. Berkley,
               Image method for efficiently simulating small-room Acoustics,
               Journal Acoustic Society of America, 65(4), April 1979, p 943.

               [2] P.M. Peterson,
               Simulating the response of multiple microphones to a single
               acoustic source in a reverberant room, Journal Acoustic
               Society of America, 80(5), November 1986.

Author       : ir. E.A.P. Habets (e.a.p.habets@tue.nl)

Version      : 1.7.20060531

History      : 1.0.20030606 Initial version
               1.1.20040803 + Microphone directivity
                            + Improved phase accuracy [2]
               1.2.20040312 + Reflection order
               1.3.20050930 + Reverberation Time
               1.4.20051114 + Supports multi-channels
               1.5.20051116 + High-pass filter [1]
                            + Microphone directivity control
               1.6.20060327 + Minor improvements
               1.7.20060531 + Minor improvements

Copyright (C) 2003-2006 Technische University Eindhoven, The Netherlands.

This program is free software; you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation; either version 2 of the License, or
(at your option) any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public License
along with this program; if not, write to the Free Software
Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA  02110-1301  USA
*/

#define _USE_MATH_DEFINES

#include "matrix.h"
#include "mex.h"
#include "math.h"
#include <memory.h>

#define ROUND(x) ((x)>=0?(long)((x)+0.5):(long)((x)-0.5))

double sinc(double x)
{
  if (x == 0)
    return(1.);
  else
    return(sin(x)/x);
}

double sim_microphone(double x, double y, double angle, char* mtype)
{
  double a, refl_theta, P, PG;

  refl_theta = atan2(y,x) - angle;

  // Polar Pattern           P        PG
  // ----------------------------------
```

```
  // Omnidirectional        1       0
  // Subcardioid            0.75    0.25
  // Cardioid               0.5     0.5
  // Hypercardioid          0.25    0.75
  // Bidirectional          0       1

  switch(mtype[0])
  {
  case 'o':
    P = 1;
    PG = 0;
    break;
  case 's':
    P = 0.75;
    PG = 0.25;
    break;
  case 'c':
    P = 0.5;
    PG = 0.5;
    break;
  case 'h':
    P = 0.25;
    PG = 0.75;
    break;
  case 'b':
    P = 0;
    PG = 1;
    break;
  default:
    P = 1;
    PG = 0;
    break;
  };

  a = P + PG * cos(refl_theta);

  return a;
}

void mexFunction(int nlhs, mxArray *plhs[], int nrhs, const mxArray *prhs[])
{
  if (nrhs == 0)
  {
    mexPrintf("------------------------------------------------------------------------------\n"
              "| Room Impulse Response Generator ------------------------------------|\n"
              "|--------------------------------------------------------------------|\n"
              "| Function for simulating the impulse response of a specified ------|\n"
              "| room using the image method [1,2]. ---------------------------------|\n"
              "|--------------------------------------------------------------------|\n"
              "| Author ----: Emanuel A.P. Habets (e.a.p.habets@tue.nl) -------------|\n"
              "|--------------------------------------------------------------------|\n"
              "| Version ---: 1.7.20060531 ------------------------------------------|\n"
              "|--------------------------------------------------------------------|\n"
              "| Copyright (C) 2003-2006 Technische University Eindhoven, ----------|\n"
              "| The Netherlands. ---------------------------------------------------|\n"
              "|--------------------------------------------------------------------|\n"
              "| [1] J.B. Allen and D.A. Berkley, -----------------------------------|\n"
              "| -----Image method for efficiently simulating small-room Acoustics, |\n"
              "| -----Journal Acoustic Society of America, --------------------------|\n"
              "| -----65(4), April 1979, p 943. ------------------------------------|\n"
              "|--------------------------------------------------------------------|\n"
              "| [2] P.M. Peterson, -------------------------------------------------|\n"
              "| -----Simulating the response of multiple microphones to a single --|\n"
              "| -----acoustic source in a reverberant room, Journal Acoustic ------|\n"
              "| -----Society of America, 80(5), November 1986. ---------------------|\n"
              "|--------------------------------------------------------------------|\n"
              "------------------------------------------------------------------------------\n\n"
              "function [h, beta] = rir_generator(c, fs, r, s, L, beta, nsample, mtype,"
              " order, dim, orientation, hp_filter);\n\n"
              "Input parameters:\n"
              "c   = sound velocity in m/s.\n"
              "fs  = sampling frequency in Hz.\n"
              "r   = M x 3 array specifying the (x,y,z) coordinates of the receiver(s) in m.\n"
              "s   = 1 x 3 vector specifying the (x,y,z) coordinates of the source in m.\n"
              "L   = 1 x 3 vector specifying the room dimensions (x,y,z) in m.\n"
              "beta = 1 x 6 vector specifying the reflection coefficients"
              " [beta_x1 beta_x2 beta_y1 beta_y2\n"
              "      beta_z1 beta_z2] or c = Reverberation Time (T_60) in seconds.\n"
              "nsample = number of samples to calculate, default is T_60*fs.\n"
              "mtype = [omnidirectional, subcardioid, cardioid, hypercardioid, bidirectional],"
              " default is omnidirectional.\n"
              "order = reflection order, default is -1, i.e. maximum order).\n"
              "dim = room dimension (2 or 3), default is 3.\n"
              "orientation = specifies the angle (in rad) in which the microphone is pointed,"
              " default is 0.\n"
              "hp_filter = use 'false' to disable high-pass filter, the high-pass filter is"
              " enabled by default.\n\n"
              "Output parameters:\n"
              "h = M x nsample matrix containing the calculated room impulse response(s).\n"
              "beta = In case a reverberation time is specified as an input parameter, the"
              " corresponding reflection coefficient is returned.\n\n");
    return;
```

```
}
else
{
  mexPrintf("Room Impulse Response Generator (Version 1.7.20060531) by Emanuel Habets\n"
    "Copyright (C) 2003-2006 Technische University Eindhoven, The Netherlands.\n");
}

// Check for proper number of arguments
if (nrhs < 6)
  mexErrMsgTxt("Error: There are at least six input parameters required.");
if (nrhs > 12)
  mexErrMsgTxt("Error: Too many input arguments.");
if (nlhs > 2)
  mexErrMsgTxt("Error: Too many output arguments.");

// Check for proper arguments
if (!(mxGetN(prhs[0])==1) || !mxIsDouble(prhs[0]) || mxIsComplex(prhs[0]))
  mexErrMsgTxt("Invalid input aruments!");
if (!(mxGetN(prhs[1])==1) || !mxIsDouble(prhs[1]) || mxIsComplex(prhs[1]))
  mexErrMsgTxt("Invalid input aruments!");
if (!(mxGetN(prhs[2])==3) || !mxIsDouble(prhs[2]) || mxIsComplex(prhs[2]))
  mexErrMsgTxt("Invalid input aruments!");
if (!(mxGetN(prhs[3])==3) || !mxIsDouble(prhs[3]) || mxIsComplex(prhs[3]))
  mexErrMsgTxt("Invalid input aruments!");
if (!(mxGetN(prhs[4])==3) || !mxIsDouble(prhs[4]) || mxIsComplex(prhs[4]))
  mexErrMsgTxt("Invalid input aruments!");
if (!(mxGetN(prhs[5])==6 || mxGetN(prhs[5])==1) || !mxIsDouble(prhs[5]) || mxIsComplex(prhs[5]))
  mexErrMsgTxt("Invalid input aruments!");

// Load parameters
double           c = mxGetScalar(prhs[0]);
double           fs = mxGetScalar(prhs[1]);
const double*    rr = mxGetPr(prhs[2]);
int              nr_of_mics = (int) mxGetM(prhs[2]);
const double*    ss = mxGetPr(prhs[3]);
const double*    LL = mxGetPr(prhs[4]);
double*          cc;
int              nsamples;
char*            mtype;
int              order;
int              dim;
double           angle;
int              hp_filter;
double           TR;

plhs[1] = mxCreateDoubleMatrix(1, 1, mxREAL);
double* beta = mxGetPr(plhs[1]);
beta[0] = 0;

// Reflection coefficients or Reverberation Time?
if (mxGetN(prhs[5])==1)
{
  double V = LL[0]*LL[1]*LL[2];
  double S = 2*(LL[0]*LL[2]+LL[1]*LL[2]+LL[0]*LL[1]);
  TR = mxGetScalar(prhs[5]);
  double alfa = 24*V*log(10.0)/(c*S*TR);
  if (alfa >= 1)
    mexErrMsgTxt("Error: The reflection coefficients can not be calculated using the current "
                 "room parameters, i.e. room size and reverberation time.\n           Please "
                 "specify the reflection coefficients or change the room parameters.");
  beta[0] = sqrt(1-alfa);
  cc = new double[6];
  for (int i=0;i<6;i++)
    cc[i] = beta[0];
}
else
{
  cc = mxGetPr(prhs[5]);
}

// High-pass filter (optional)
if (nrhs > 11 && mxIsEmpty(prhs[11]) == false)
{
  hp_filter = (int) mxGetScalar(prhs[11]);
}
else
{
  hp_filter = 1;
}

// Microphone orientation (optional)
if (nrhs > 10 && mxIsEmpty(prhs[10]) == false)
{
  angle = (double) mxGetScalar(prhs[10]);
}
else
{
  angle = 0;
}

// Room Dimension (optional)
if (nrhs > 9 && mxIsEmpty(prhs[9]) == false)
```

```
{
  dim = (int) mxGetScalar(prhs[9]);
  if (dim != 2 && dim != 3)
    mexErrMsgTxt("Invalid input aruments!");
}
else
{
  dim = 3;
}

// Reflection order (optional)
if (nrhs > 8 &&  mxIsEmpty(prhs[8]) == false)
{
  order = (int) mxGetScalar(prhs[8]);
  if (order < -1)
    mexErrMsgTxt("Invalid input aruments!");
}
else
{
  order = -1;
}

// Type of microphone (optional)
if (nrhs > 7 &&  mxIsEmpty(prhs[7]) == false)
{
  mtype = new char[mxGetN(prhs[7])+1];
  mxGetString(prhs[7], mtype, mxGetN(prhs[7])+1);
}
else
{
  mtype = new char[1];
  mtype[0] = 'o';
}

// Number of samples (optional)
if (nrhs > 6 &&  mxIsEmpty(prhs[6]) == false)
{
  nsamples = (int) mxGetScalar(prhs[6]);
}
else
{
  if (mxGetN(prhs[5])>1)
  {
    double V = LL[0]*LL[1]*LL[2];
    double S = 2*(LL[0]*LL[2]+LL[1]*LL[2]+LL[0]*LL[1]);
    double alpha = ((1-pow(cc[0],2))+(1-pow(cc[1],2)))*LL[0]*LL[2] +
                   ((1-pow(cc[2],2))+(1-pow(cc[3],2)))*LL[1]*LL[2] +
                   ((1-pow(cc[4],2))+(1-pow(cc[5],2)))*LL[0]*LL[1];
    TR = 24*log(10.0)*V/(c*alpha);
    if (TR < 0.1)
      TR = 0.128;
  }
  nsamples = (int) (TR * fs);
}

// Create output vector
plhs[0] = mxCreateDoubleMatrix(nr_of_mics, nsamples, mxREAL);
double* imp = mxGetPr(plhs[0]);

// Temporary variables and constants (high-pass filter)
const double W = 2*M_PI*100/fs;
const double R1 = exp(-W);
const double R2 = R1;
const double B1 = 2*R1*cos(W);
const double B2 = -R1 * R1;
const double A1 = -(1+R2);
const double A2 = R2;
double       X0, Y0, Y1, Y2;

// Temporary variables and constants (image-method)
const double Fc = 1;
const int    Tw = 2 * ROUND(0.004*fs);
const double cTs = c/fs;
double*      hanning_window = new double[Tw+1];
double*      LPI = new double[Tw+1];
double*      r = new double[3];
double*      s = new double[3];
double*      L = new double[3];
double*      beta = new double[6];
double       kwnsample = pow((double) nsamples,2);
double       hu[6];
double       refl[3];
double       dist;
double       ll;
double       strength;
int          pos, fdist;
int          n1,n2,n3;
int          q, j, k;
int          m, l, n;
int          t;

for (int i = 0 ; i < 6 ; i++)
```

```
{
  if (dim == 2 && i > 3)
    beta[i] = 0;
  else
    beta[i] = cc[i];
}

s[0] = ss[0]/cTs; s[1] = ss[1]/cTs; s[2] = ss[2]/cTs;
L[0] = LL[0]/cTs; L[1] = LL[1]/cTs; L[2] = LL[2]/cTs;

// Hanning window
for (t = -Tw/2 ; t <= Tw/2 ; t++)
{
  hanning_window[t + Tw/2] = 0.5 * (1 + cos(2*M_PI*t/Tw));
}

for (int mic_nr = 0; mic_nr < nr_of_mics ; mic_nr++)
{
  // [x_1 x_2 ... x_N y_1 y_2 ... y_N z_1 z_2 ... z_N]
  r[0] = rr[mic_nr + 0*nr_of_mics] / cTs;
  r[1] = rr[mic_nr + 1*nr_of_mics] / cTs;
  r[2] = rr[mic_nr + 2*nr_of_mics] / cTs;

  n1 = ROUND((nsamples/(2*L[0])) + 2);
  n2 = 0;
  n3 = 0;

  // Generate room impulse response
  for (q = 0 ; q < 2 ; q++)
  {
    hu[0] = s[0] - r[0] + 2*q*r[0];

    for (j = 0 ; j < 2 ; j++)
    {
      hu[1] = s[1] - r[1] + 2*j*r[1];

      for (k = 0 ; k < 2 ; k++)
      {
        hu[2] = s[2] - r[2] + 2*k*r[2];

        for (n = -n1 ; n <= n1 ; n++)
        {
          ll = kwnsample - pow(2*n*L[0], 2);
          if (ll <= 0)
            n2 = 2;
          else
            n2 = ROUND(sqrt(ll)/(2*L[1]) + 2);

          hu[3] = hu[0] + 2*n*L[0];

          refl[0] = pow(beta[0], abs(n)) * pow(beta[1], abs(n+q));

          for (l = -n2 ; l <= n2 ; l++)
          {
            ll = kwnsample - pow(2*n*L[0], 2) - pow(2*l*L[1], 2);
            if (ll <= 0)
              n3 = 2;
            else
              n3 = ROUND(sqrt(ll)/(2*L[2]) + 2);

            hu[4] = hu[1] + 2*l*L[1];

            refl[1] = pow(beta[2], abs(l)) * pow(beta[3], abs(l+j));

            for (m = -n3 ; m <= n3 ; m++)
            {
              hu[5] = hu[2] + 2*m*L[2];

              refl[2] = pow(beta[4],abs(m)) * pow(beta[5], abs(m+k));

              dist = sqrt(pow(hu[3], 2) + pow(hu[4], 2) + pow(hu[5], 2));

              fdist = (int) floor(dist);
              if (abs(2*n+q)+abs(2*l+j)+abs(2*m+k) <= order || order == -1)
              {
                if (fdist+(Tw/2) <= nsamples)
                {
                  strength = sim_microphone(hu[3], hu[4], angle, mtype)
                             * refl[0]*refl[1]*refl[2]/(4*M_PI*dist*cTs);

                  for (t = 0 ; t < Tw+1 ; t++)
                    LPI[t] = hanning_window[t] * Fc * sinc(M_PI*Fc*(t-(dist-fdist)-(Tw/2)));

                  pos = fdist-(Tw/2);
                  if (pos > 0)
                    for (t = 0 ; t < Tw+1 ; t++)
                      imp[mic_nr + nr_of_mics*(pos+t)] = imp[mic_nr + nr_of_mics*(pos+t)]
                                                         + strength * LPI[t];
                  else
                    for (t = -pos ; t < Tw+1 ; t++)
                      imp[mic_nr + nr_of_mics*(pos+t)] = imp[mic_nr + nr_of_mics*(pos+t)]
                                                         + strength * LPI[t];
```

```
                    }
                  }
                }
              }
            }
          }
        }
      }

      // 'Original' high-pass filter as proposed by Allen and Berkley.
      if (hp_filter == 1)
      {
        Y0 = 0;
        Y1 = 0;
        Y2 = 0;
        for (int idx = 0 ; idx < nsamples ; idx++)
        {
          X0 = imp[mic_nr+nr_of_mics*idx];
          Y2 = Y1;
          Y1 = Y0;
          Y0 = B1*Y1 + B2*Y2 + X0;
          imp[mic_nr+nr_of_mics*idx] = Y0 + A1*Y1 + A2*Y2;
        }
      }
    }
  }
}
```

Appendix B

OM-LSA Estimator for Multiple Interferences

In this appendix we present an algorithm for robust speech enhancement based on an Optimally-Modified Log Spectral Amplitude (OM-LSA) estimator for multiple interferences. In the original OM-LSA one interference was taken into account. However, there are many situations where multiple interferences are present. Since the human ear is more sensitive to a small amount of residual non-stationary interference than to a stationary interference we would like to reduce the non-stationary interference signal down to the residual noise level of the stationary interference. Possible applications for the developed algorithm are joint speech dereverberation and noise reduction, and joint residual echo suppression and noise reduction. Additionally, we present three possible methods to estimate the a priori Signal to Interference Ratio of each of the interferences.

B.1 Introduction

Spectral enhancement has received a lot of attention in the last three decades, especially for single channel noise reduction. Recently, researchers have started to use these techniques for residual echo suppression [264, 250] and speech dereverberation [34]. In practical systems one may encounter more than one interference simultaneously.

In [250] Gustafsson et al. proposed two post-filters for residual echo and noise reduction. The first post-filter is based on the Log Spectral Amplitude estimator [205] and was extended to attenuate multiple interferences, the second post-filter was psychoacoustically motivated.

In this appendix we present an Optimally-Modified Log Spectral Amplitude (OM-LSA) estimator for multiple interferences. The OM-LSA spectral gain function, which

minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty. In the original OM-LSA, proposed by Cohen [228], one interference was taken into account. There are many applications in which we are dealing with one non-stationary and one stationary interference. Since the human ear is more sensitive to a small amount of residual non-stationary interference than to a stationary interference we would like to reduce the non-stationary interference signal down to the residual noise level of the stationary interference, such that the final residual non-stationary interference will be masked by the residual stationary interference. Possible applications for the proposed algorithm are joint speech dereverberation and noise reduction, and joint residual echo suppression and noise reduction. The OM-LSA spectral gain function is a function of the *a priori* and *a posteriori* Signal to Interference Ratios (SIR). In this appendix we additionally present three possible methods to estimate the *a priori* SIR of each of the interferences.

The outline of this appendix is as follows. The problem statement can be found in Section B.2. A brief review of the OM-LSA estimator and a modification of the spectral gain function is provided in Section B.3. In Section B.4 we will discuss three methods to estimate the *a priori* SIR for each of the interferences.

B.2 Problem Statement

Let $x(n)$, $r(n)$ and $d(n)$ denote the clean speech signal and two uncorrelated additive interference signals, respectively,

$$y(n) = x(n) + r(n) + d(n).$$

It should be noted that in case $r(n)$ and $d(n)$ are statistically independent Gaussian random variables they can be considered as one interference. The variance of the total interference is then equal to the sum of the separate variances. However, in case $r(n)$ and $d(n)$ are, for example, a non-stationary and a stationary interference, and the (maximum) amount of desirable reduction is different, their separation is preferred. The OM-LSA spectral gain function, which depends on both time and frequency, is a function of the *a priori* and *a posteriori* Signal to Interference Ratios, which are denoted by $\xi(l,k)$ and $\gamma(l,k)$, respectively. In this appendix time frames are denoted by the index l, and frequency bins are denoted by the index k. We show that one can gain control of the noise reduction level for each interference by associating a separate *a priori* SIR with each interference.

The estimated short-time Fourier transform (STFT) of the clean speech, $\hat{X}(l,k)$, is obtained by applying the spectral gain function, $G_{\text{OM-LSA}}$, to each noisy spectral component:

$$\hat{X}(l,k) = G_{\text{OM-LSA}}(l,k)Y(l,k).$$

The estimated clean speech signal can be obtained using the inverse STFT and a weighted overlap-add method.

In the sequel we assume that an estimate of the spectral variance of each interference is available at all times. In many applications, such as speech dereverberation or residual echo suppression, it is reasonable to assume that the spectral variance of the non-stationary interference can be estimated (c.f. [264, 250, 34]). The spectral variance of the stationary interference can be estimated, for example, using the Improved Minima Controlled Recursive Averaging (IMCRA) method proposed by Cohen.

B.3 OM-LSA Estimator

The Log Spectral Amplitude (LSA) estimator proposed by Ephraim and Malah [205] minimizes

$$\mathcal{E}\left\{\left(\log(A(l,k)) - \log(\hat{A}(l,k))\right)^2\right\}, \tag{B.1}$$

where $A(l,k) = |X(l,k)|$ denotes the spectral speech amplitude, and $\hat{A}(l,k)$ is its optimal estimator. Assuming statistical independent spectral components, the LSA estimator is defined as

$$\hat{A}(l,k) = \exp\left(\mathcal{E}\{\log(A(l,k))|Y(l,k)\}\right). \tag{B.2}$$

The LSA gain function is given by

$$G_{\text{LSA}}(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)} \exp\left(\frac{1}{2}\int_{\zeta(l,k)}^{\infty} \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t\right), \tag{B.3}$$

where

$$\frac{1}{\xi(l,k)} = \frac{1}{\xi_r(l,k)} + \frac{1}{\xi_d(l,k)}, \tag{B.4}$$

$$\xi_r(l,k) = \frac{\lambda_x(l,k)}{\lambda_r(l,k)}, \quad \xi_d(l,k) = \frac{\lambda_x(l,k)}{\lambda_d(l,k)}, \tag{B.5}$$

$$\gamma(l,k) = \frac{|Y(l,k)|^2}{\lambda_r(l,k)+\lambda_d(l,k)}, \tag{B.6}$$

$$\zeta(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\,\gamma(l,k), \tag{B.7}$$

$$\lambda_x(l,k) = \mathcal{E}\{|X(l,k)|^2\}, \tag{B.8}$$

and

$$\lambda_d(l,k) = \mathcal{E}\{|D(l,k)|^2\}, \text{ and } \lambda_r(l,k) = \mathcal{E}\{|R(l,k)|^2\}. \tag{B.9}$$

The OM-LSA spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty [228]. Given two hypotheses, $H_0(l,k)$ and $H_1(l,k)$, which indicate speech absence and presence, respectively, we have

$$\begin{aligned} H_0(l,k) &: Y(l,k) = R(l,k) + D(l,k), \\ H_1(l,k) &: Y(l,k) = X(l,k) + R(l,k) + D(l,k). \end{aligned} \tag{B.10}$$

Based on a Gaussian statistical model, the speech presence probability is given by

$$p(l,k) = \left\{1 + \frac{q(l,k)}{1-q(l,k)}\,(1+\xi(l,k))\exp\left(-\zeta(l,k)\right)\right\}^{-1}, \tag{B.11}$$

where $q(l,k)$ is the *a priori* speech absence probability [228].

The OM-LSA gain function is given by,

$$G_{\text{OM-LSA}}(l,k) = \{G_{\text{H}_1}(l,k)\}^{p(l,k)}\,\{G_{\text{H}_0}(l,k)\}^{1-p(l,k)}, \tag{B.12}$$

with $G_{\text{H}_1}(l,k) = G_{\text{LSA}}(l,k)$ and $G_{\text{H}_0}(l,k) = G_{\min}$. The lower-bound constraint for the gain when the signal is absent is denoted by $G_{\min}$, and specifies the maximum amount of noise reduction in noise only frames.

In our case the lower-bound constraint does not result in the desired result because $r(n)$ can still be clearly audible. To alleviate this problem we propose the following modification of G_{H_0}. Our goal is to suppress the non-stationary interference down to the noise floor, given by $G_{\min}\ D(l,k)$. We apply $G_{\text{H}_0}(l,k)$ to those time-frequency frames where the desired signal is assumed to be absent, i.e. hypothesis $\text{H}_0(l,k)$ is assumed to be true, such that

$$\hat{X}(l,k) = G_{\text{H}_0}(l,k)\,(R(l,k) + D(l,k))\,. \tag{B.13}$$

The desired solution for $\hat{X}(l,k)$ is

$$\hat{X}(l,k) = G_{\min}(l,k)\ D(l,k). \tag{B.14}$$

Assuming that the interferences are uncorrelated, minimizing

$$\mathcal{E}\left\{\left|G_{\text{H}_0}(l,k)\,(R(l,k) + D(l,k)) - G_{\min}(l,k)\ D(l,k)\right|^2\right\} \tag{B.15}$$

results in the desired solution for $G_{\text{H}_0}(l,k)$,

$$G_{\text{H}_0}(l,k) = G_{\min}\frac{\lambda_d(l,k)}{\lambda_d(l,k) + \lambda_r(l,k)}. \tag{B.16}$$

The *a posteriori* SIRs can be directly estimated given the noisy observation and an estimate of the spectral variance of each interference. The estimation of the *a priori* SIR is slightly more complicated and will be discussed in the next section.

B.4 *A priori* SIR estimator for Multiple Interferences

Many researchers believe that the main advantage of the LSA estimator is related to the decision-directed estimator, proposed by Ephraim and Malah [205]. In this section

we show how the decision-directed estimator can be used for estimating $\xi_r(l,k)$ and $\xi_d(l,k)$. We also present a causal and non-causal recursive estimation procedure for the *a priori* SIRs using the same reasoning as in [208].

The total *a priori* SIR can be calculated using Eq. B.4. However, in case the desired speech signal $x(n)$ and the non-stationary interference signal $r(n)$ are very small, the *a priori* SIR $\xi_r(l,k)$ may be unreliable since $\lambda_x(l,k)$ and $\lambda_r(l,k)$ are close to zero. Hence, the total *a priori* SIR may be unreliable as well. In the following we assume that there is always a certain amount of background noise, i.e., $\lambda_d(l,k) \neq 0$. To alleviate the foregoing problem we propose to calculate $\xi(l,k)$ using only the most important and reliable *a priori* SIRs as follows

$$\xi(l,k) = \begin{cases} \xi_d, & 10\log_{10}\left(\frac{\lambda_d(l,k)}{\lambda_r(l,k)}\right) > \beta^{\mathrm{dB}}; \\ \dfrac{\xi_d(l,k)\xi_r(l,k)}{\xi_d(l,k)+\xi_r(l,k)}, & \text{otherwise}, \end{cases} \tag{B.17}$$

where the threshold β^{dB} specifies the level difference between $\lambda_d(l,k)$ and $\lambda_r(l,k)$ in dB.

B.4.1 Decision Directed

The decision-directed estimator is given by

$$\hat{\xi}^{\mathrm{DD}}(l,k) = \max\left\{\eta\frac{\hat{A}^2(l-1,k)}{\lambda(l-1,k)} + (1-\eta)\psi(l,k),\ \xi_{\min}\right\},$$

where $\psi(l,k) = \gamma(l,k) - 1$ is the *instantaneous* SIR, $\lambda(l,k) = \lambda_r(l,k) + \lambda_d(l,k)$, and $\xi_{\min}$ is a lower-bound constraint on the *a priori* SIR. The weighting factor η ($0 \leq \eta \leq 1$) controls the tradeoff between the amount of noise reduction and distortion (e.g., musical tones). To estimate $\xi_v(l,k)$, where $v \in \{r,d\}$, we propose to use the following expression

$$\hat{\xi}_v^{\mathrm{DD}}(l,k) = \max\left\{\eta_v\frac{\hat{A}^2(l-1,k)}{\lambda_v(l-1,k)} + (1-\eta_v)\psi_v(l,k),\ \xi_{\min,v}\right\},$$

where

$$\begin{aligned} \psi_v(l,k) &= \frac{\lambda(l,k)}{\lambda_v(l,k)}\psi(l,k) \\ &= \frac{\lambda_r(l,k)+\lambda_d(l,k)}{\lambda_v(l,k)}(\gamma(l,k)-1) \\ &= \frac{|Y(l,k)|^2 - \lambda_r(l,k) - \lambda_d(l,k)}{\lambda_v(l,k)} \\ &= \frac{|Y(l,k)|^2 - \lambda(l,k)}{\lambda_v(l,k)}. \end{aligned} \tag{B.18}$$

B.4.2 Causal Recursive Estimator

In this section we propose a causal conditional estimator

$$\xi_v(l|l,k) \triangleq \frac{\lambda_x(l|l,k)}{\lambda_v(l,k)}, \tag{B.19}$$

where $v \in \{r,d\}$ and $\lambda_x(l|l,k) \triangleq \mathcal{E}\{A^2(l,k)|Y([0,\ldots,l],k)\}$, for the *a priori* SIRs given the noisy measurements up to frame l. The estimators are derived following the footsteps in [208]. Each estimator combines two steps, a 'propagation' step and an 'update' step, following the rational of the Kalman smoother, to recursively predict and update the estimate for $\lambda_x(l,k)$ as new data arrives.

Suppose we are given an estimate $\hat{\lambda}_x(l|l-1,k)$, which is conditioned on the noisy measurements up to frame $l-1$, and a new noisy spectral component $Y(l,k)$ is observed. Then, the estimate for $\lambda_x(l,k)$, can be updated by computing the conditional variance of $X(l,k)$ given $Y(l,k)$ and $\hat{\lambda}_x(l|l-1,k)$

$$\hat{\lambda}_x(l|l,k) = \mathcal{E}\left\{A^2(l,k)|\hat{\lambda}_x(l|l-1,k), Y(l,k)\right\}. \tag{B.20}$$

The result is [208]

$$\hat{\lambda}_x(l|l,k) = \frac{\hat{\xi}(l|l-1,k)}{1+\hat{\xi}(l|l-1,k)}\left(\frac{1}{\gamma(l,k)} + \frac{\hat{\xi}(l|l-1,k)}{1+\hat{\xi}(l|l-1,k)}\right)|Y(l,k)|^2. \tag{B.21}$$

The 'update' step for $\hat{\xi}_v$ is now obtained by dividing both sides by $\lambda_v(l,k)$, i.e.,

$$\begin{aligned}\hat{\xi}_v(l|l,k) &= \max\left\{\frac{\hat{\xi}(l|l-1,k)}{1+\hat{\xi}(l|l-1,k)}\left(\frac{1}{\gamma(l,k)} + \frac{\hat{\xi}(l|l-1,k)}{1+\hat{\xi}(l|l-1,k)}\right)\frac{|Y(l,k)|^2}{\lambda_v(l,k)}, \xi_{\min,v}\right\} \\ &= \max\left\{\frac{\hat{\xi}(l|l-1,k)}{1+\hat{\xi}(l|l-1,k)}\left(\frac{\lambda(l,k)}{\lambda_v(l,k)} + \gamma_v(l,k)\frac{\hat{\xi}(l|l-1,k)}{1+\hat{\xi}(l|l-1,k)}\right), \xi_{\min,v}\right\}.\end{aligned} \tag{B.22}$$

Computation of the 'update' step for the *a priori* SIR given the past noisy speech components up to frame $l-1$ requires the estimate

$$\hat{\xi}(l|l-1,k) \triangleq \frac{\hat{\lambda}_x(l|l-1,k)}{\lambda(l-1,k)}, \tag{B.23}$$

where

$$\hat{\lambda}_x(l|l-1,k) = (1-\beta)\hat{\lambda}_x(l-1|l-1,k) + \beta\hat{A}^2(l-1,k), \tag{B.24}$$

and β $(0 \le \beta \le 1)$ is related to the degree of non-stationarity of the random process $\{\lambda_x(l,k)|l=0,1,\ldots\}$. Dividing both sides of Eq. B.24 by $\lambda(l-1,k)$, we obtain the 'propagation' step

$$\hat{\xi}(l|l-1,k) = (1-\beta)\hat{\xi}(l-1|l-1,k) + \beta\frac{\hat{A}^2(l-1,k)}{\lambda(l-1,k)}, \tag{B.25}$$

where $\hat{\xi}(l-1|l-1,k)$ is calculated similar to $\xi(l,k)$ in Eq. B.17 using $\hat{\xi}_d(l-1|l-1,k)$ and $\hat{\xi}_r(l-1|l-1,k)$.

B.4.3 Non-Casual Recursive Estimator

In this section we propose a non-causal conditional estimator

$$\xi_v(l|l+L,k) \triangleq \frac{\lambda_x(l|l+L,k)}{\lambda_v(l,k)}, \tag{B.26}$$

where $v \in \{r,d\}$ and $\lambda_x(l|l+L,k) \triangleq \mathcal{E}\{A^2(l,k)|Y([0,\ldots,l+L],k)\}$, for the *a priori* SIRs given the noisy measurements up to frame $l+L$. The non-causal estimator combines two steps, a 'propagation' step and an 'update' step, following the rationale of Kalman filtering, to recursively predict and update the estimate for $\lambda_x(l,k)$ as new data arrives. The non-causal estimator also employs future spectral measurements in the process to better predict the spectral variance of the clean speech.

Let $\lambda'_x(l|l+L,k) \triangleq \mathcal{E}\{A^2(l,k)|Y([0,\ldots,l-1,l+1,\ldots,l+L],k)\}$ denote the conditional spectral variance of $X(l,k)$ given the noisy measurements up to frame $l+L$ excluding the noisy measurement at frame l. Let $\lambda_x(l|[l+1,\ldots,l+L],k) \triangleq \mathcal{E}\{A^2(l,k)|Y([l+1,\ldots,l+L],k)\}$ denote the conditional spectral variance of $X(l,k)$ given the subsequent noisy measurements $Y([l+1,\ldots,l+L],k)$.

The estimate for $\lambda_x(l,k)$ given $\lambda'_x(l|l+L,k)$ and $Y(l,k)$ can be updated by

$$\begin{aligned}\hat{\lambda}_x(l|l+L,k) &= \mathcal{E}\left\{A^2(l,k)|\lambda'_x(l|l+L,k),Y(l,k)\right\}\\ &= \frac{\hat{\xi}'(l|l+L,k)}{1+\hat{\xi}'(l|l+L,k)}\left(\frac{1}{\gamma(l,k)}+\frac{\hat{\xi}'(l|l+L,k)}{1+\hat{\xi}'(l|l+L,k)}\right)|Y(l,k)|^2,\end{aligned} \tag{B.27}$$

where

$$\hat{\xi}'(l|l+L,k) \triangleq \frac{\hat{\lambda}'_x(l|l+L,k)}{\lambda(l-1,k)} \tag{B.28}$$

is the *a priori* SIR given the noisy speech components up to frame $l+L$, excluding frame l [208].

The 'backward estimation' and 'backward-forward propagation' are exactly the same as in [208] and are presented here for completeness. The 'backward estimation' is given by

$$\hat{\xi}(l|[l+1,\ldots,l+L],k) = \begin{cases}\frac{1}{L}\sum_{n=1}^{L}\gamma(l+n,k)-\varepsilon, & \text{if non-negative;}\\ 0, & \text{otherwise,}\end{cases} \tag{B.29}$$

where ε ($\varepsilon \geq 1$) is the over-subtraction factor. The 'backward-forward propagation' is

calculated using

$$\hat{\xi}'(l|l+L,k) = \beta\frac{\hat{A}^2(l-1,k)}{\lambda(l-1,k)} + (1-\beta)\left[\beta'\hat{\xi}(l-1|l+L-1,k) + (1-\beta')\hat{\xi}(l|[l+1,\ldots,l+L],k)\right], \quad \text{(B.30)}$$

where β $(0 \leq \beta \leq 1)$ is related to the stationarity of the random process λ_x, β' $(0 \leq \beta' \leq 1)$ is associated with the reliability of the estimate $\xi(l|[l+1, l+L], k)$, and $\hat{\xi}(l-1|l+L-1,k)$ is calculated similar to $\xi(l,k)$ in Eq. B.17 using $\hat{\xi}_d(l-1|l+L-1,k)$ and $\hat{\xi}_r(l-1|l+L-1,k)$.

Dividing both sides in Eq. B.27 by $\lambda_v(l,k)$, and applying a lower-bound constraint $\xi_{\min,v}$, results in the 'update' step, i.e.,

$$\hat{\xi}_v(l|l+L,k) = \max\left\{\frac{\hat{\xi}'(l|l+L,k)}{1+\hat{\xi}'(l|l+L,k)}\left(\frac{1}{\gamma(l,k)} + \frac{\hat{\xi}'(l|l+L,k)}{1+\hat{\xi}'(l|l+L,k)}\right)\frac{|Y(l,k)|^2}{\lambda_v(l,k)},\ \xi_{\min,v}\right\}. \quad \text{(B.31)}$$

Index

List of publications by the author

[P-1] J. v.d. Laar, E.A.P. Habets, J.D.P.A. Peters, and P.A.M. Lokkart, "Adaptive blind audio signal separation on a DSP," *Proc. of the 12th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC'01)*, pp. 475–479, Nov. 2001.

[P-2] E.A.P. Habets and P.C.W. Sommen, "Optimal Microphone Placement for Source Localization using Time Delay Estimation," in *Proc. of the 13th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC'02)*, Veldhoven, Netherlands, Nov. 2002, pp. 284–287.

[P-3] E.A.P. Habets, "A Literature Study on Dereverberation in Acoustic Environments," *Internal Report*, pp. 1–50, 2003.

[P-4] E.A.P. Habets, "Single-Channel Speech Dereverberation based on Spectral Subtraction," in *Proc. of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC'04)*, Veldhoven, Netherland, Nov. 2004, pp. 250–254.

[P-5] E.A.P. Habets, "Multi-Channel Speech Dereverberation based on a Statistical Model of Late Reverberation," in *Proc. of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, Mar. 2005, pp. 173–176.

[P-6] E.A.P. Habets, "Speech Dereverberation based on a Statistical Model of Late Reverberation using a Linear Microphone Array," in *Proc. of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA'05)*, Piscataway NJ, USA, Mar. 2005, pp. d7–d8.

[P-7] E.A.P. Habets, "Experimental Results of a Multi-Channel Speech Dereverberation Algorithm based on a Statistical Model of Late Reverberation," in *Proc. of the first annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS'05)*, Antwerpen, Belgium, Apr. 2005, pp. 11–14.

[P-8] E.A.P. Habets, "Room Impulse Response Generator," *Internal Report*, pp. 1–17, 2006.

[P-9] E.A.P. Habets, I. Cohen, S. Gannot, and P.C.W. Sommen, "Joint Dereverberation and Residual Echo Suppression of Speech Signals in a Noisy Environment," *Submitted to IEEE Transactions of Audio, Speech, and Language Processing*, June 2006.

[P-10] E.A.P. Habets and P.C.W. Sommen, "Speech Dereverberation using Spectral Subtraction and a Generalized Statistical Reverberation Model," *Submitted to Elsevier's Speech Communications journal*, June 2006.

[P-11] E.A.P. Habets, S. Gannot, and I. Cohen, "Dual-Microphone Speech Dereverberation in a Noisy Environment," in *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT'06)*, Vancouver, Canada, Aug. 2006.

[P-12] E.A.P. Habets, I. Cohen, and S. Gannot, "MMSE Log Spectral Amplitude Estimator for Multiple Interferences," in *Proc. of the 10th International Workshop of Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006, pp. 1–4.

[P-13] 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. of the 10th International Workshop of Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006, pp. 1–4.

[P-14] E.A.P. Habets and S. Gannot, "Dual-Microphone Speech Dereverberation using a Reference Signal," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Honolulu, USA, Apr. 2007, vol. IV, pp. 901–904.

Acknowledgments

The work presented in this dissertation was conducted in the Signal Processing Systems (SPS) group at the Technische Universiteit Eindhoven (TU/e) in the Netherlands. First of all I would like to thank my co-promoter dr.ir. Piet Sommen for giving me the opportunity to join the 'Transparent Audio Communication' project. His support and true understanding of my personal situation during my Ph.D. has meant a lot to me. I also thank my promoter prof.dr.ir. Jan Bergmans for his support, especially during the last phase of my Ph.D. I would like to express my gratitude to my colleagues from the SPS group for their support and for creating a nice working environment. Furthermore, I would like to thank the people from the personal department, in particular Pleun Bazen and Femke Verheggen. I would also like to thank my (former) roommates, Jakob, Jeroen, Alfonso, Li, Hongming, Nicolae, and Yulia. Special thanks to Jakob van de Laar, who became one of my best friends and always made time for a short walk around the campus, proofreading my articles, and discussions.

The 'Transparent Audio Communication' project was kindly supported by Technology Foundation STW, applied science division of NWO and the Technology Programme of the Ministry of Economic Affairs of the Netherlands. I would like to thank the program officer dr.ir. Frank van den Berg and Yvonne van Scharenburg from STW for their support during this project. I also thank the members of the user committee, and in particular Kees Janse (Philips Research), Jos Leenen (Ge-ReSound) and Roland de Wild (ASTRON), for their interest in, and discussion during, this project.

During my Ph.D. I spent one month at the Bar-Ilan University in Ramat-Gan in Israel where I had the opportunity to work with dr. Sharon Gannot, who later became my second co-promoter, and prof.dr. Israel Cohen. I would like to thank them for this opportunity. This period was not only very fruitful from a research point of view, but was also a wonderful experience. Special thanks to Sharon and Dana for their hospitality and for showing me some beautiful places in Israel.

I also thank the other members of the promotion committee, viz., dr. Patrick Naylor, Prof.Dr.-Ing. Rainer Martin and prof.dr.ir. Mico Hirschberg. Their comments and suggestions have improved the presentation and quality of this work.

During the last four years time was precious. I'm thankful to my family and all my friends who understood my priorities. Special thanks to my cousins Harold and Mark,

who own an exclusive clothing shop in Maastricht (Van Overeem Van Wissen), and kept me well dressed on special occasions. I would also like to thank Aukje for her love, encouragement and patience, and her family for their support during my Ph.D. Last, but certainly not least, I would like to thank my parents for their never-ending support, love and encouragement.

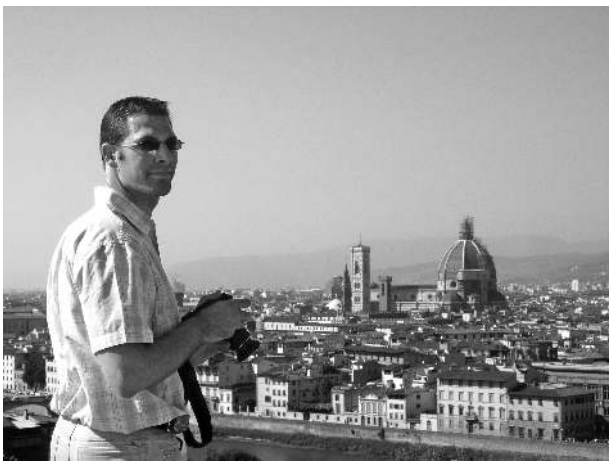
Curriculum Vitae

Emanuël A.P. Habets was born in Maastricht, The Netherlands, on April 12, 1976.

He received the B.Sc degree in electrical engineering from the Hogeschool Limburg, The Netherlands, and M.Sc degree in electrical engineering from the Technische Universiteit Eindhoven (TU/e), The Netherlands, in 1999 and 2002, respectively. From 2002 to 2006, he worked as a researcher in the Signal Processing Systems (SPS) group of the TU/e, where he carried out the work presented in this dissertation. In February 2006 he was a guest researcher at Bar-Ilan University in Ramat-Gan, Israel. Since March 2007, he is a postdoctoral researcher at the Technion - Israel Institute of Technology and at the Bar-Ilan University in Ramat-Gan, Israel.

His research interest include statistical signal processing and speech enhancement using either single or multiple microphones with applications in acoustic communication systems. His main research interest is speech dereverberation.

Mr. Habets was a member of the organization committee of the 9^{th} International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, 2005.



At the 14^{th} European Signal Processing Conference (EUSIPCO) in Florence, Italy, 2006.