Joint Dereverberation and Noise Reduction for Binaural Hearing Aids and Mobile Phones

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines Doktors der
Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Master of Science

Marco Jeub

aus Bad Neuenahr-Ahrweiler

Berichter: Universitätsprofessor Dr.-Ing. Peter Vary
Universitätsprofessor Dr. ir. Simon Doclo

Tag der mündlichen Prüfung: 01.08.2012

AACHENER BEITRÄGE ZU DIGITALEN NACHRICHTENSYSTEMEN

Herausgeber:

Prof. Dr.-Ing. Peter Vary
Institut für Nachrichtengeräte und Datenverarbeitung
Rheinisch-Westfälische Technische Hochschule Aachen
Muffeter Weg 3a
52074 Aachen
Tel.: 0241-80 26 956
Fax.: 0241-80 22 186

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar

1. Auflage Aachen:
Wissenschaftsverlag Mainz in Aachen
(Aachener Beiträge zu digitalen Nachrichtensystemen, Band 34)
ISSN 1437-6768
ISBN 978-3-86130-345-9

Wissenschaftsverlag Mainz
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 2 39 48 oder 02 41 / 87 34 34
Fax: 02 41 / 87 55 77
www.Verlag-Mainz.de

Herstellung: Druckerei Mainz GmbH,
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34; Fax: 02 41 / 87 55 77
www.Druckservice-Aachen.de

Gedruckt auf chlorfrei gebleichtem Papier

"D 82 (Diss. RWTH Aachen University, 2012)"

Acknowledgments

This thesis was written during my time as research assistant at the *Institute of Communication Systems and Data Processing (IND)* at the *Rheinisch-Westfälische Technische Hochschule Aachen (RWTH Aachen University)*. I would like to take the opportunity to thank the people who contributed to the success of this work.

At first, I am deeply grateful to my supervisor Prof. Dr.-Ing. Peter Vary for his continuous support of my work by numerous ideas and suggestions and for encouraging me in my scientific interests. I am also indebted to Prof. Dr. ir. Simon Doclo for being the co-supervisor of this thesis and showing much interest in my work.

Furthermore, I want to thank all my colleagues and permanent staff of the institute for providing a pleasant and enjoyable working environment. In particular, I am grateful to Dr.-Ing. Thomas Esch, Dr.-Ing. Heiner Löllmanni, Dipl.-Ing. Magnus Schäfer, Dipl.-Ing. Christoph Nelke, Dr.-Ing. Hauke Krüger, Dr.-Ing. Christiane Antweiler, Dipl.-Ing. Moritz Beermann and Dipl.-Ing. Matthias Pawig for their proof-readings of parts of the manuscript.

Special thanks go to the students who contributed to this work. In particular, I am indebted to Dipl.-Ing. Christian Herglotz, B.Sc. Juliana Hsu, Dipl.-Ing. Christoph Norrenbrock, B.Sc. Andreas Witte and Dipl.-Ing. Robert Bücs for their valuable contributions.

This work was accompanied by projects with Siemens and Intel and I want to thank our former project partners for the good collaboration. Special thanks to Dr. Christophe Beaugeant and Intel Mobile Communications in Sophia Antipolis, France for the good and friendly collaboration over the last three years.

I also wish to express my appreciation to Dr. Patrick Naylor and his team at Imperial College London for many fruitful discussions during my stay in their lab.

Munich, August 2012 Marco Jeub

Contents

Symbols & Abbreviations

List of Principal Symbols

Estimated values are usually indicated by a hat

$\{\cdot\}^*$	Complex conjugate
a_p	Linear prediction (LP) coefficients ($p = 1, 2, .., P$)
$b(k)$	Normal distribution
c	Sound velocity [m/s]
$C_{x_1 x_2}(\Omega)$	Magnitude squared coherence (MSC)
$\mathrm{CDR}^{(\mathrm{min\|max})}$	Limits of CDR estimate
d_{mic}	Inter-microphone distance [m]
d_{c}	Critical distance [m]
d_{LM}	Source-microphone or loudspeaker-microphone distance [m]
$\mathrm{DRR}^{(\mathrm{min\|max})}$	Limits of DRR estimate
$\mathbb{E}$	Expectation value
$E_{l\|r}$	Energy of right and left signals $x_{l\|r}(k)$
$\mathrm{EDC}(t)$	Energy decay curve (EDC)
f_{c}	Subband center frequency [Hz]
f_{s}	Sampling frequency [Hz]
$f_{\mathrm{smooth}}^{(\mathrm{min\|max})}$	Thresholds for musical noise reduction postfilter [Hz]
$f_{\mathrm{Schroeder}}$	Schroeder frequency [Hz]
f_0, μ_0	First zero-crossing of a function (time domain and corresponding frequency bin
$G(\lambda, \mu)$	Spectral weighting gains
G_{min}	Lower spectral gain threshold
$h(k)\|_{(\mathrm{Gen})}$	Generalized statistical RIR model
$h(k)\|_{(\mathrm{Polack})}$	Statistical RIR model
$h_{\mathrm{early}}(k)$	Direct and early part of RIR h(k)
$h_{\mathrm{late}}(k)$	Late reverberant part of RIR h(k)
$H_{12}(e^{j\Omega})$	Acoustic transfer function of source between the two microphones
$H_m(e^{j\Omega})$	Acoustic transfer function between source and microphone m

$h_m(k)$	Room impulse response between source and microphone m
$J_0(\cdot)$	zero-order Bessel function of first kind
J	Prediction order in [frames] (EK)
k	Discrete time index
k_d	Discrete time index of RIR where the direct sound ends
k_l	Time span after which the late reverberation begins
L	Block or frame length
L_r	Length of the discrete impulse response
M	FFT length
m	Microphone index
N	Non-redundant FFT bins for real-valued input (here: $N = M/2 + 1$)
N_l	Time span after which the late reverberation begins [No. frames]
$n_m(k)$	Noise signal of microphone m, m=1,2
p	Sound source index
P	Prediction interval in [frames] (EK)
r	Radius of head
R	Number of sound sources
$s(k)$	Clean speech signal
$s_m(k)$	Clean speech signal of microphone m
$\widehat{s}_m(k)$	Enhanced signal of microphone m
$\widehat{S}_m(\lambda, \mu)$	Enhanced signal of microphone m in the short-term DFT domain
$\widetilde{S}(\lambda, \mu)$	Pre-enhanced signal in the short-term DFT domain
T	Time constant
T_l	Time span after which the late reverberation begins [ms]
$\mathrm{T}_{60}^{(\mathrm{Sabine})}$	Reverberation time formula by Sabine
T_{60}	Reverberation time (RT) [s]
V	Room volume [m^3]
$X_{\mathrm{early}}(\lambda, \mu)$	Early reverberant signal in the short-term DFT domain
$X_{\mathrm{late}}(\lambda, \mu)$	Late reverberant signal in the short-term DFT domain
$x_{l\|r}(k)$	Input signals of left and right side of binaural hearing aid
$X_m(\lambda, \mu)$	Noisy and reverberant speech signal of microphone m in the short-term DFT domain
$x_m(k)$	Noisy and reverberant speech signal of microphone m
$\alpha^{(\mathrm{active})}$	Smoothing factor for DRR estimation in speech active frames
$\alpha^{(\mathrm{DRR})}$	Smoothing factor for DRR estimation
$\alpha^{(\mathrm{FK})}$	Smoothing factor (FK)
$\alpha^{(\mathrm{ideal})}$	Smoothing factor used for ideal PSD calculation
$\alpha^{(\mathrm{ILD})}$	Smoothing factor for ILD estimation
$\alpha^{(\mathrm{inactive})}$	Smoothing factor for DRR estimation in speech inactive frames
$\alpha^{(\mathrm{PLDNE})}$	Smoothing factor (PLDNE)
$\alpha^{(\mathrm{PSD})}$	Smoothing factor for short-term PSD calculation
$\alpha^{(\mathrm{X1\|X2})}$	Smoothing factors (PLDNE)
$\alpha^{(\mathrm{xx})}$	Smoothing factor used for reverberant speech PSD

$\alpha^{(\mathrm{coh})}$	Smoothing factor for coherence calculation
α_{DD}	Smoothing factor (DDA)
α_{nn}	Smoothing factor (Coherence noise PSD estimator)
α_{Sabine}	Absorption coefficient for Sabine RT equation
$\Delta E(\lambda, \mu)$	Short-term and frequency dependent ILD
ΔE	Interaural Level Difference (ILD)
Δt	Interaural Time Difference (ITD)
$\Delta\Phi_{\mathrm{PLD}}(\lambda, \mu)$	Power level difference of the noisy input signal (PLD)
$\Delta\Phi_{\mathrm{PLDNE}}(\lambda, \mu)$	Normalized PLD (PLDNE)
Δ	No. of frames back that late rev. is assumed to start (EK)
γ	Overestimation factor (PLD)
$\gamma(\lambda, \mu)$	Short-term interaural coherence (IC)
$\gamma^{(\mathrm{WW})}$	Attenuation factor (WW)
$\gamma_{\mathrm{thr}}(\mu)$	Interaural coherence threshold for ILD estimation
Γ_{max}	Threshold for coherence model
$\Gamma_{x_1 x_2}(e^{j\Omega})$	Coherence function
$\Gamma^{(\mathrm{mix})}_{x_1 x_2}(e^{j\Omega})$	Coherence function of a mixed coherent and diffuse noise field
$\Gamma^{(\mathrm{coh})}_{x_1 x_2}(e^{j\Omega})$	Coherence function of a coherent noise source
$\Gamma^{(\mathrm{cyl})}_{x_1 x_2}(e^{j\Omega})$	Coherence function for cylindrically isotropic or 2D noise field
$\Gamma^{(\mathrm{diff})}_{x_1 x_2}(e^{j\Omega})$	Coherence function for spherically isotropic or 3D noise field
$\kappa(\mu)$	Proportionality constant (HB)
λ	Frame index
Λ	Scaling factor for musical noise reduction postfilter
μ	Frequency index
μ'	Subband index
ν	Spreading parameter of Rayleigh distribution (WW)
Ω	Angular frequency
$\overline{X}(\lambda, \mu)$	Masking thresholds in the short-term DFT domain
$\phi_{\mathrm{min\|max}}$	Thresholds (PLDNE)
$\Phi_{x_1 x_2}(e^{j\Omega})$	Cross-PSD of x_1 and x_2
$\Phi_{x_m x_m}(e^{j\Omega})$	Auto-PSD of x_m
$\widehat{\Phi}_{\mathrm{int}}(\lambda, \mu)$	Short-term PSD of late reverberant speech and background noise
$\widehat{\Phi}_{\mathrm{nn}}(\lambda, \mu)$	Short-term PSD of background noise
$\widehat{\Phi}^{(\mathrm{late})}_{\mathrm{rev}}(\lambda, \mu)$	Short-term PSD of late reverberant speech
$\widehat{\Phi}^{(\mathrm{I})}_{nn}(\lambda, \mu)$	Coherence based noise PSD estimate (arithm. mean)
$\widehat{\Phi}^{(\mathrm{II})}_{nn}(\lambda, \mu)$	Coherence based noise PSD estimate (geom. mean)
ψ	Threshold for onset time detection
$\Psi(e^{j\Omega})$	Coherent-to-diffuse energy ratio (CDR)
ρ	Decay rate
σ^2	Variance
$\tau^{(\mathrm{VAD})}$	VAD threshold
θ	Azimuth angle of source signal s(k) [°]

ζ_{att}	Interference attenuation factor
ζ_{thr}	Threshold for musical noise postfilter

List of Abbreviations

ACELP	Algebraic Code Excited Linear Prediction
AIR	Aachen Impulse Response
AMR-NB	Adaptive Multi-Rate Narrowband
AMR-WB	Adaptive Multi-Rate Wideband
ATF	Acoustic Transfer Function
BB	Bottom-Bottom
BSD	Bark Spectral Distortion
BSS	Blind Source Separation
BT	Bottom-Top
CELP	Code Excited Linear Prediction
CI	Cochlear Implants
CDR	Coherent-to-Diffuse Energy Ratio
DCT	Discrete Cosine Transform
DDA	Decision-Directed Approach
DMSS	Dual-Microphone Spectral Subtraction
DOA	Direction-of-Arrival
DRR	Direct-to-Reverberation Energy Ratio
DSB	Delay-and-Sum Beamformer
EDC	Energy Decay Curve
ERP	Ear Reference Point
FFT	Fast Fourier Transform
GCC-PHAT	Generalized Cross-Correlation with Phase Transform
HAs	Hearing Aids
HFRP	Hands-Free Reference Point
HHP	Hand-Held Position
i.i.d.	independent and identically distributed
IC	Interaural Coherence
IFFT	Inverse Fast Fourier Transform
ILD	Interaural Level Difference
ITD	Interaural Time Difference
ITU	International Telecommunication Union
LP	Linear Prediction
LRSV	Late Reverberant Spectral Variance
LTI	Linear Time Invariant
MARDY	Multichannel Acoustic Reverberation Database at York
ML	Maximum Likelihood
MLS	Maximum Length Sequence
MRP	Mouth Reference Point
MS	Minimum Statistics

MSC	Magnitude Squared Coherence
MMSE	Minimum Mean Square Error
MWF	Multichannel Wiener Filter
NA	Noise Attenuation
NB	Narrowband
PLDNE	Power Level Difference Noise Estimator
PLD	Power Level Difference
PSD	Power Spectral Density
PESQ	Perceptual Evaluation of Speech Quality
PSEQ	Perfect Sequence
QMF	Quadrature Mirror Filter
RIR	Room Impulse Response
RT	Reverberation Time
SA	Speech Attenuation
SDW-MWF	Speech Distortion Weighted Multichannel Wiener Filter
SB-ADPCM	Sub-Band Adaptive Differential Pulse Code Modulation
SMERSH	Spatiotemporal averaging Method for Enhancement of Reverberant Speech
SIR	Signal-to-Interference Ratio
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SPP	Speech Presence Probability
SRR	Signal-to-Reverberant Energy Ratio
SRMR	Speech to Reverberation Modulation energy Ratio
SRT	Speech Reception Threshold
STI	Speech Transmission Index
SWB	Super-Wideband
VAD	Voice Activity Detector
WB	Wideband
WGN	White Gaussian Noise

1

Introduction

Speech signals captured by the microphones of a speech communication device are often distorted by interfering noise sources as well as room reverberation. Such degradations may reduce the listening comfort and the speech intelligibility. Commonly, a reduction of these two types of interferences is regarded independently. For the reduction of background noise, a large variety of algorithms exists which are already integrated in applications such as hearing aids and mobile phones. In contrast to that, the reduction of room reverberation is not tackled by most speech enhancement systems.

Thus, the target of a future speech enhancement algorithm should be a reduction of unwanted background noise *and* room reverberation while ensuring at the same time that the speech distortions are as low as possible. In the context of portable devices, the computational complexity and the algorithmic delay is of significant importance. Moreover, the algorithm should be able to adapt fast to changing acoustic conditions.

The main focus of this thesis is to develop suitable dual-channel speech enhancement algorithms for the joint reduction of reverberation *and* background noise designed and adopted to different applications:

- binaural hearings aids and
- dual-microphone mobile phones,

each with its own specific acoustic conditions.

Novel algorithms are developed and presented which exploit both the noise field coherence used by hearing aid applications as well as the *Power Level Difference* (PLD) between the two microphone signals in dual-microphone mobile phones. Furthermore, new and improved algorithms for estimating required acoustic parameters such as the *Reverberation Time* (RT) and *Direct-to-Reverberation Energy Ratio* (DRR) are introduced which are capable of a blind estimation directly from the noisy and reverberant speech signals.

In a further step it will be shown that the availability of a wireless data-link between the hearing aids of left and right ear allows for the integration of so-called binaural

speech enhancement algorithms. By means of an appropriate data exchange, the algorithms can exploit spatial information and, most importantly, preserve the binaural cues (*Interaural Time Difference* (ITD), *Interaural Level Difference* (ILD)) which are important for sound localization.

This thesis consists of three major parts: *acoustics*, *algorithms*, and *applications*.

In Chapter 2, an elaborate analysis of the acoustic environment which is relevant for hearing aids and mobile phones is given. All investigations are based on recordings and measured *Room Impulse Responses* (RIR) in realistic scenarios. Furthermore, signal processing models and estimation techniques for various acoustic properties such as the RT and DRR are derived and discussed. Due to the importance for the development of hearing aid algorithms in particular, an improved estimation technique for the binaural cues is presented.

An overview of state-of-the-art algorithms for *independent* dereverberation and noise reduction, considering single- and dual-channel algorithms, is given in Chapter 3. For both application areas, novel and improved algorithms are presented which are explicitly developed for specific acoustic conditions. In the context of hearing aid applications, the noise field coherence plays an important role and should be taken into account. For dual-microphone mobile phones, it is shown how efficient speech enhancement is possible if the PLD of speech and the interfering sources between the two microphones is taken into account. In the last part of this chapter, the joint reduction of background noise *and* reverberation is tackled by means of an interlaced combination of different algorithms and estimation techniques.

Chapter 4 shows the effectiveness of the new approaches in the context of three application scenarios:

- A novel two-stage speech enhancement algorithm for *binaural hearing aids* is proposed which explicitly preserves the binaural cues and is capable to reduce early *and* late reverberation as well as background noise. The basic idea of such a combination is that Stage I of the algorithm mainly reduces the late reverberant and background noise components, while a subsequent Wiener filter in Stage II attenuates all non-coherent signal components including early and residual late reverberation. Stage II bases on a coherence model of the reverberant sound field and takes the shadowing effects of the head into account.

  This section also comprises strategies how to extend existing single-channel algorithms to two output channels and discusses the influence of bilateral signal processing, i.e., unsynchronized processing without data exchange on the binaural cues.

- The second application example considers dual-microphone *mobile phones* and presents a novel algorithm which explicitly exploits the special acoustic conditions where a secondary microphone is placed on the top side of the device. By taking into account the power level differences of speech, noise and reverberation, an effective method is proposed which is advantageous compared to state-of-the-art methods.

- Along with these two main areas, a case study is presented to show how dereverberation can be applied to enhance *single-channel* speech recordings taken in the German parliament. For this purpose a new *psychoacoustically-motivated* dereverberation concept is introduced which is capable to increase the subjective listening impression significantly.

All algorithms presented in this work are running in real-time on a standard PC laptop and have been evaluated under realistic acoustic conditions.

Parts of the results of this thesis have been pre-published in the following references: [JV09b, JSV09, JV09a, JKAO09, JV10, JSEV10, JLV10, JSK+10, SJSV10, LYJV10, JDV11, JNBV11, JNK+11, JV11, HJN+11, JHN+12, GLJ+12]. These references are marked by an underlined label, i.e., [____], throughout the thesis.

2

Analysis and Models of the Acoustic Environment

In this chapter, an elaborate analysis of the acoustic environment which is relevant for (binaural) hearing aids and dual-microphone mobile phones is given. Besides this investigation, signal processing models and estimation techniques for various acoustic properties are derived and discussed.

An introduction into relevant acoustic principles and required acoustic parameters such as the *Reverberation Time* (RT) and the *Direct-to-Reverberation Energy Ratio* (DRR) is given in Sec. 2.1. The analysis in Sec. 2.2 deals with the characterization of the acoustic environments and is divided into discussions on the acoustic device as well as investigations on additive background noise and room reverberation. Due to the particular importance for binaural signal processing, a short review on the binaural noise field coherence is given. A discussion on efficient algorithms to estimate important acoustic parameters such as short-term noise field coherence, DRR and RT is given in Sec. 2.3. This section also presents an efficient algorithm to estimate the binaural cues as well as a novel noise field classification algorithm.

In this thesis, *binaural processing* is referred to a dual-channel input and dual-channel output processing where the left and right microphone signals are available on both sides. In contrast to that, *bilateral processing* means that each side (left and right) is performing an independent processing without data exchange.

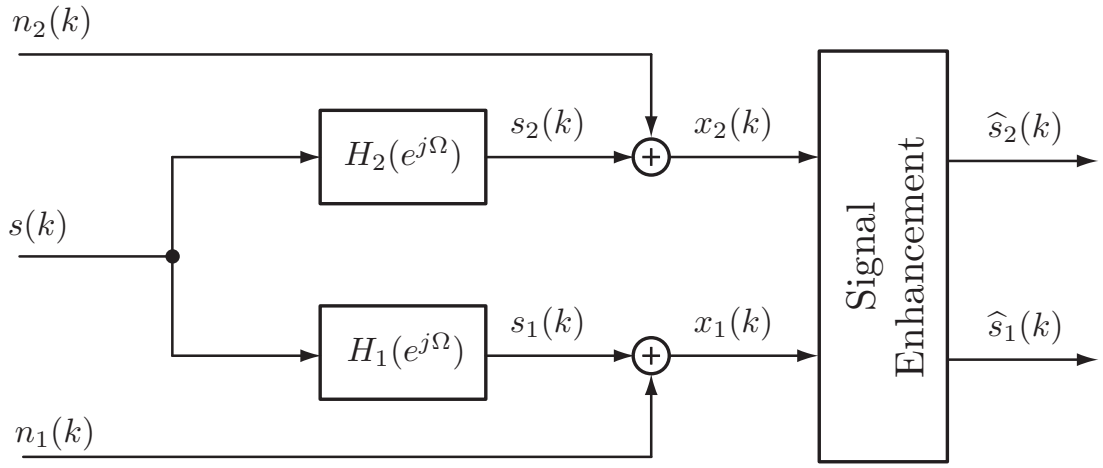
2.1 Fundamentals of Acoustics

2.1.1 System Model

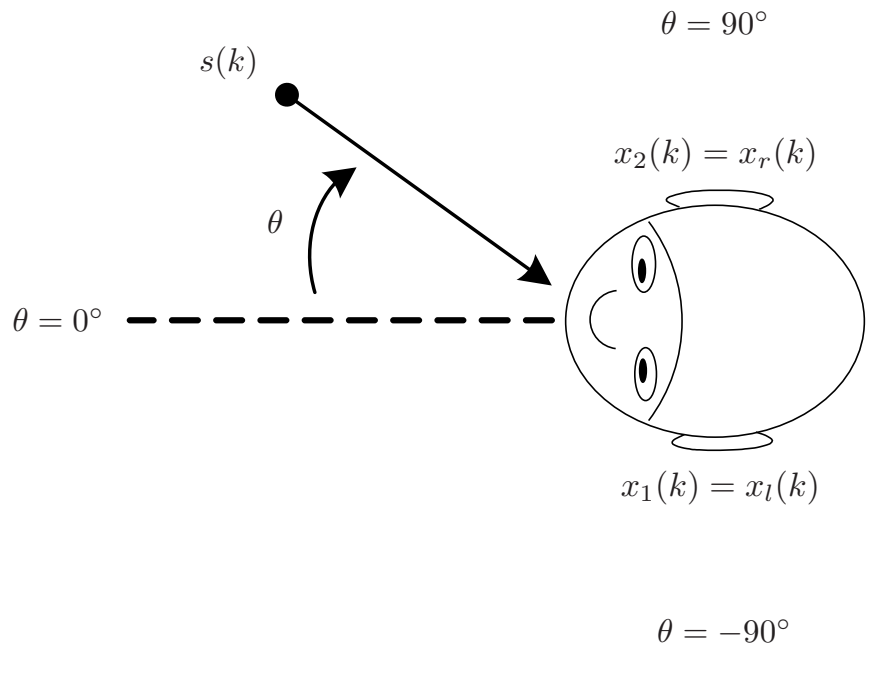
In the remainder of this thesis, the following generalized dual-channel signal model is used. The two microphone signals $x_1(k)$ and $x_2(k)$ are the inputs of the dual-channel speech enhancement system and are related to a clean speech signal $s(k)$ and additive background noise signals $n_m(k)$ as shown in Fig. 2.1 (a) with $m = 1, 2$ and discrete time index k. The noisy signals are termed $x_m(k)$ and the acoustic transfer functions

between source and microphones are denoted by $H_m(e^{j\Omega})$ in the frequency domain or $h_m(k)$ in the time-domain. The enhanced signals are named $\widehat{s}_m(k)$.

It is assumed that the noise sources are uncorrelated with the speech signal and different properties of the correlation of background noise such as incoherent, coherent or diffuse noise fields are considered.



(a)



(b)

Figure 2.1: (a) Generalized dual-channel signal model, (b) coordinate system and notation for binaural hearing aids.

Table 2.1: Main simulation parameters.

Parameter	Setting
Sampling frequency	$f_s = 16\,\text{kHz}$
Frame length	$L = 320$ (20 ms)
FFT length	$M = 512$ (incl. zero-pad.)
Frame overlap	50% (Hann window)

In the special case of binaural hearing aids, the input signals, source direction θ in the azimuth plane and ear signals $x_r(k)$, $x_l(k)$ are denoted according to Fig. 2.1 (b).

Throughout this thesis, the processing is carried out with a sampling frequency of $f_\text{s} =$ 16 kHz. For every enhancement or estimation algorithm which operates in the frequency domain, the input signals are first segmented into frames of length $L = 320$ (20 ms) with an overlap of 50%. After windowing (e.g., applying a Hann window) and zero-padding, these frames are transformed via *Fast Fourier Transform* (FFT) of length $M = 512$ into the short-term spectral domain. The corresponding spectra are denoted by $X_m(\lambda, \mu)$, where λ marks the frame index and μ the discrete frequency bin. Since the output of the FFT is half redundant for real-valued input signals, only the first $N = M/2 + 1 = 257$ frequency bin are processed. The main simulation parameters are summarized in Table 2.1. The databases for speech, background noise and room impulse responses are listed in App. A.1.

For the objective quality evaluations in this thesis, intrusive and non-intrusive instrumental measurements are used. Besides the non-intrusive measures, it is required that direct speech $s_d(k)$ and interfering sources $n_\text{int}(k)$ are available separately, given by

$$s_d(k) = s(k) * h_\text{d}(k), \tag{2.1}$$

$$n_\text{int}(k) = s(k) * h_\text{rev}(k) + n(k), \tag{2.2}$$

where $h_\text{d}(k)$ and $h_\text{rev}(k)$ denote the direct and the reverberation parts (including early and late reverberant speech) of the *Room Impulse Response* (RIR), respectively. The use of the direct speech compared to clean or anechoic speech has the great advantage that the sound wave propagation time which is inherently included in the RIR does not need to be compensated before the evaluation. The enhanced signal $\widehat{s}(k)$, filtered speech $\widetilde{s}_d(k)$ and filtered interfering signals $\widetilde{n}_\text{int}(k)$ are obtained by applying the same spectral weighting gains to each of the three individual signals in the frequency domain as illustrated in Fig. 2.2 [Gus99]. In terms of speech attenuation, *Speech to Reverberation Modulation energy Ratio* (SRMR) and *Perceptual Evaluation of Speech Quality* (PESQ) score, all signal levels are normalized to $-26\,\text{dBov}$ using the ITU-T Rec. P.56 speech voltmeter [ITU93]. Silence periods have been removed using the *Voice Activity Detector* (VAD) of the AMR-WB speech codec [3GP04c]. For most experiments, the first 30 s of all signals have been removed in order to compare the steady state performance only.

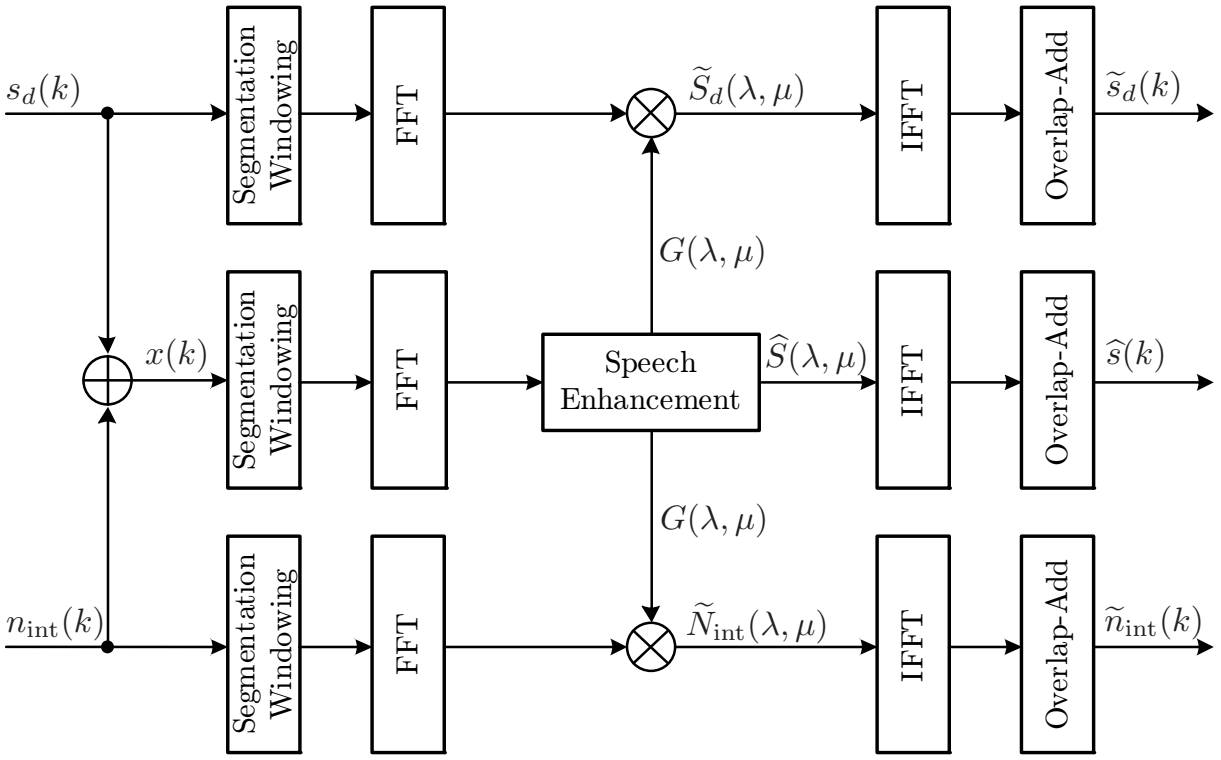


Figure 2.2: Principle of objective evaluation to determine the filtered speech signal and interfering sources.

2.1.2 Noise Field Coherence

A well-established measure for describing a key feature of any noise environment is the complex noise field coherence, which is of special interest for the development and evaluation of multi-channel speech enhancement algorithms. The complex coherence between two signals[1] $x_m(k)$ $(m = 1, 2)$ is defined in the frequency domain as [Kut09]

$$\Gamma_{x_1x_2}(e^{j\Omega}) = \frac{\Phi_{x_1x_2}(e^{j\Omega})}{\sqrt{\Phi_{x_1x_1}(e^{j\Omega}) \cdot \Phi_{x_2x_2}(e^{j\Omega})}}, \tag{2.3}$$

where $\Phi_{x_1x_1}(e^{j\Omega})$ and $\Phi_{x_2x_2}(e^{j\Omega})$ represent the auto-*Power Spectral Density* (PSD) of $x_1(k)$ and $x_2(k)$ in the Fourier domain. $\Phi_{x_1x_2}(e^{j\Omega})$ represents the cross-PSD between $x_1(k)$ and $x_2(k)$. The normalized radian frequency is given by $\Omega = 2\pi f/f_s$ with frequency variable f and sampling frequency f_s.

The frequently used term *Magnitude Squared Coherence* (MSC) is referred to the squared magnitude of Eq.(2.3) and is given by

$$C_{x_1x_2}(e^{j\Omega}) = \left|\Gamma_{x_1x_2}(e^{j\Omega})\right|^2 = \frac{\left|\Phi_{x_1x_2}(e^{j\Omega})\right|^2}{\Phi_{x_1x_1}(e^{j\Omega}) \cdot \Phi_{x_2x_2}(e^{j\Omega})} \tag{2.4}$$

[1] For the sake of clarity, the introduction of the coherence function is given with the signals $x_m(k)$ compared to Fig. 2.1 (a). When it comes to the coherence of noise only, the coherence between the noise sources $n_m(k)$ is considered.

with the property

$$0 \leq C_{x_1 x_2}(e^{j\Omega}) \leq 1. \tag{2.5}$$

A value of zero for the MSC appears for an incoherent noise field, e.g., self-noise of the microphones.

2.1.2.1 Coherence of Important Noise Fields

In a *diffuse noise field*, an infinite number of point sources are emitting in all directions simultaneously with equal energy and low spatial correlation. Assuming two microphones in the far-field and a homogeneous noise field, the spherically isotropic, diffuse or 3D coherence can be calculated by the integration over all possible directions of incident of a directional sound source. The corresponding formula reads [Kut09]

$$\Gamma^{(\mathrm{diff})}_{x_1 x_2}(e^{j\Omega}) = \mathrm{sinc}\left(\Omega f_s d_{\mathrm{mic}}/c\right) \tag{2.6}$$

with distance d_{mic} between two omnidirectional microphones and sound velocity c, where a value $c = 340\,\mathrm{m/s}$ is used throughout this thesis.

The first zero-crossing of the sinc-function reads

$$f_0 = c/(2d_{\mathrm{mic}}). \tag{2.7}$$

For a given spacing of 0.15 m, which is relevant for hearing aids, this frequency is $f_0 = 1.1\,\mathrm{kHz}$. The microphone signals are highly correlated for frequencies below f_0 while the correlation is low for frequencies above f_0. The corresponding frequency bin is termed μ_0 in the DFT domain. This boundary is important to understand the performance of coherence-based speech enhancement algorithms, e.g., which exploit the high coherence of a desired signal and the low coherence of an interference. For such algorithms, the highest achievable frequency-dependent noise attenuation is given by [VHH98]

$$-10\log_{10}\left(\mathrm{C}_{\mathrm{x_1 x_2}}(\mathrm{e}^{\mathrm{j}\Omega})\right), \tag{2.8}$$

e.g., if the background noise MSC has a value of 0.1, a maximum of 10 dB can be reduced, whereas for an MSC values of 0.8 only approx. 1 dB is the maximum achievable noise attenuation. Thus, a significant noise reduction can only be obtained for frequencies with relatively low MSC values of the interfering signals.

In the case of a non-uniform distribution of the incident sound waves, the corresponding coherence function is termed cylindrically isotropic or 2D and yields to [Kut09]

$$\Gamma^{(\mathrm{cyl})}_{x_1 x_2}(e^{j\Omega}) = J_0\left(\Omega f_s d_{\mathrm{mic}}/c\right), \tag{2.9}$$

with the zero-order Bessel function of first kind denoted by $J_0(\cdot)$.

When it comes to non-omnidirectional microphone characteristics (e.g., cardioid) or differential arrays, which is not considered in this thesis, the reader is referred to

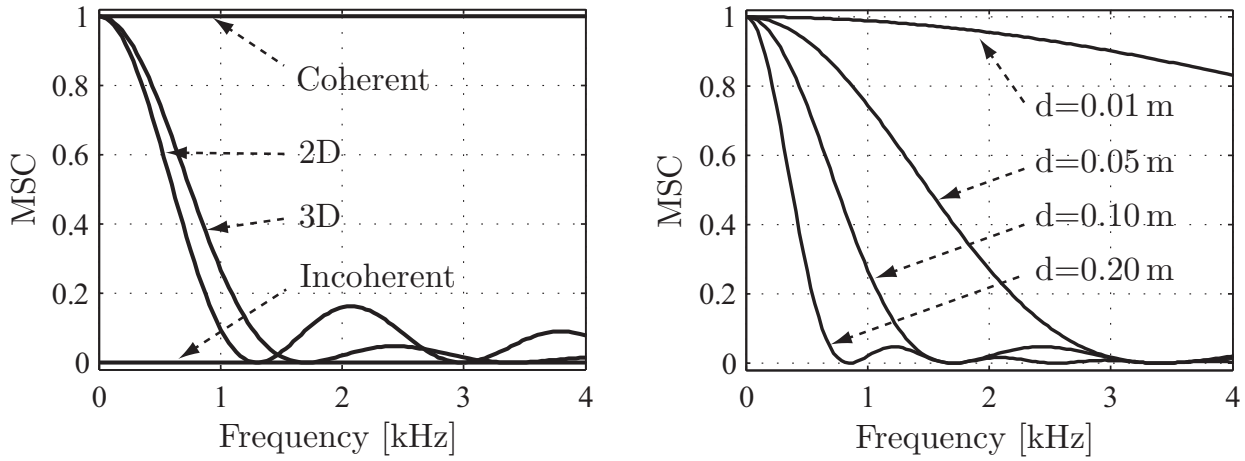


Figure 2.3: MSC of (left) incoherent, coherent, spherically isotropic (3D, diffuse) and cylindrically isotropic (2D) noise field for a fixed inter-microphone distance of $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$, (right) diffuse noise field at different $d_{\mathrm{mic}} = \{0.01, 0.05, 0.1, 0.2\}\,\mathrm{m}$.

the discussions in [ACV86, Mar95, Mar01b, Elk01]. In reverberant sound fields, the coherence can also be approximated by a diffuse noise field as discussed later, see also [JR00, Kut09, JSV09].

The coherence function of a noise source which is coherent between the microphones is given by [Kut09]

$$\Gamma_{x_1x_2}^{(\mathrm{coh})}(e^{j\Omega}) = \mathrm{e}^{-\mathrm{j}\Omega \mathrm{f_s} \mathrm{d_{mic}} \cos(\theta)/\mathrm{c}} = \cos\left(\Omega \mathrm{f_s} \mathrm{d_{mic}} \cos\left(\theta\right)/\mathrm{c}\right) - \mathrm{j}\sin\left(\Omega \mathrm{f_s} \mathrm{d_{mic}} \cos\left(\theta\right)/\mathrm{c}\right), \tag{2.10}$$

where the microphones are placed in broadside orientation in the far-field and the noise source arrives at angle θ. This corresponds obviously to a value of one for the MSC.

Figure 2.3 (left) shows exemplarily the MSC for an incoherent, coherent, spherically isotropic and cylindrically isotropic noise field at a fixed inter-microphone distance of $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$. In Fig. 2.3 (right), the characteristics of a diffuse sound field for different inter-microphone distances are plotted. It can be seen that the MSC decays rapidly for higher frequencies and that the sensor spacing has a severe impact on the MSC.

In the context of binaural hearing aids, the shadowing effect of the head has a significant impact on the coherence function. To take this into account, a binaural semi-analytical signal processing model will be discussed in Sec. 2.2.1.1 and is derived in App. B. A discussion on the short-term coherence estimation from the input signals itself is given in Sec. 2.3.1.

2.1.2.2 Mixed Noise Fields

When it comes to mixed noise fields, i.e., a superposition of diffuse and coherent noise the cross-PSD in Eq.(2.3) is given by the sum of R individual cross-PSDs [Pie78]. This summation applies also for the total auto-PSD. Hence, the generalized complex coherence function reads

$$\Gamma_{x_1x_2}^{(\mathrm{mix})}(e^{j\Omega}) = \frac{\sum_{r=1}^{R} \Phi_{x_1x_2}^{(r)}(e^{j\Omega})}{\sqrt{\sum_{r=1}^{R} \Phi_{x_1x_1}^{(r)}(e^{j\Omega}) \cdot \sum_{r=1}^{R} \Phi_{x_2x_2}^{(r)}(e^{j\Omega})}} \tag{2.11}$$

with R being the number of superposed noise fields or sources. It is assumed that all noise sources are uncorrelated with each other.

For an ideal diffuse noise field, the noise signals arrive with approximately equal power spectral density $\Phi_{\mathrm{d}}(e^{j\Omega})$ at the two microphones such that

$$\Phi_{x_1x_1}^{(\mathrm{diff})}(e^{j\Omega}) = \Phi_{x_2x_2}^{(\mathrm{diff})}(e^{j\Omega}) = \Phi_{\mathrm{d}}(e^{j\Omega}). \tag{2.12}$$

The cross-PSD in this case is given by [Pie78]

$$\Phi_{x_1x_2}^{(\mathrm{diff})}(e^{j\Omega}) = \Phi_{\mathrm{d}}(e^{j\Omega}) \cdot \mathrm{sinc}\left(\Omega f_s d_{\mathrm{mic}}/c\right). \tag{2.13}$$

For a coherent noise source which arrives at angle θ with approximately the same PSD $\Phi_{\mathrm{c}}(e^{j\Omega})$, the auto-PSD reads

$$\Phi_{x_1x_1}^{(\mathrm{coh})}(e^{j\Omega}) \approx \Phi_{x_2x_2}^{(\mathrm{coh})}(e^{j\Omega}) = \Phi_{\mathrm{c}}(e^{j\Omega}). \tag{2.14}$$

The corresponding cross-PSD can be expressed by

$$\Phi_{x_1x_2}^{(\mathrm{coh})}(e^{j\Omega}) = \Phi_{\mathrm{c}}(e^{j\Omega})\mathrm{e}^{-\mathrm{j}\Omega\mathrm{f_s}\mathrm{d_{mic}}\cos(\theta)/\mathrm{c}}. \tag{2.15}$$

If we assume that the coherent noise signal arrives without a time delay at the two sensors, i.e., $\theta = \pi/2$ (see Fig. 2.1 (b)), or that the time delay has been compensated, the cross-PSD reduces to

$$\Phi_{x_1x_2}^{(\mathrm{coh})}(e^{j\Omega}) = \Phi_{\mathrm{c}}(e^{j\Omega}). \tag{2.16}$$

To take into account different ratios of coherent and diffuse noise, the *Coherent-to-Diffuse Energy Ratio* (CDR) is introduced as

$$\Psi(e^{j\Omega}) = \frac{\Phi_{\mathrm{c}}(e^{j\Omega})}{\Phi_{\mathrm{d}}(e^{j\Omega})}. \tag{2.17}$$

Having described the auto- and cross-PSD equations, a model for the complex coherence of a mixed noise field with diffuse and coherent components can be expressed

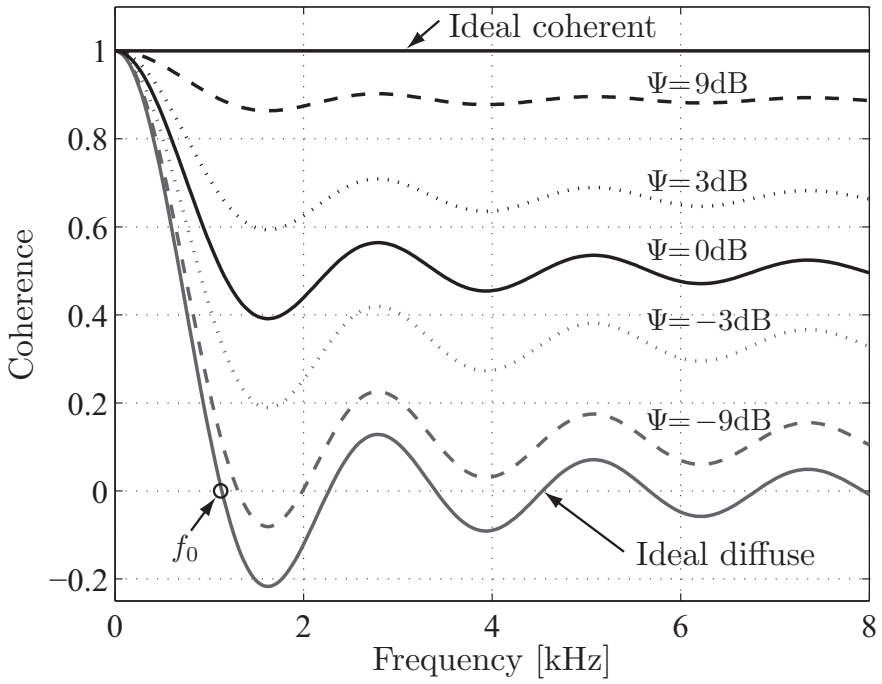


Figure 2.4: Coherence of mixed coherent and diffuse noise with varying CDR. For the simulations an inter-microphone distance of $d_{\text{mic}} = 0.15\,\text{m}$ is used. Gray curves indicate a negative CDR value in dB and black curves CDR values $\geq 0\,\text{dB}$. Additionally, the curves for ideal diffuse and ideal coherent noise fields are plotted. The first zero-crossing of the sinc-function is marked by f_0. Please note that the coherent source is assumed to arrive from the front.

with Eqs.(2.11), (2.17), $R = 2$ and by assuming $\theta = \pi/2$ for the coherent noise source by

$$\begin{aligned}\Gamma_{x_1x_2}^{(\text{mix})}(e^{j\Omega}) &= \frac{\Phi_{\text{c}}(e^{j\Omega}) + \Phi_{\text{d}}(e^{j\Omega})\text{sinc}\left(\Omega f_s d_{\text{mic}}/c\right)}{\Phi_{\text{c}}(e^{j\Omega}) + \Phi_{\text{d}}(e^{j\Omega})} \\ &= \frac{\Psi(e^{j\Omega}) + \text{sinc}\left(\Omega f_s d_{\text{mic}}/c\right)}{1 + \Psi(e^{j\Omega})}.\end{aligned} \tag{2.18}$$

For the special case where the CDR is equal for all frequencies, the three following special cases are included in Eq.(2.18):

- $\Psi \to \infty \Rightarrow \Gamma_{x_1x_2}^{(\text{mix})}(e^{j\Omega}) \to 1$: coherent noise field,
- $\Psi = 1$: same energy of diffuse and coherent noise, $\Gamma_{x_1x_2}^{(\text{mix})}(e^{j\Omega}) = \frac{1+\text{sinc}\left(\Omega f_s d_{\text{mic}}/c\right)}{2}$,
- $\Psi = 0 \Rightarrow \Gamma_{x_1x_2}^{(\text{mix})}(e^{j\Omega}) = \text{sinc}\left(\Omega f_s d_{\text{mic}}/c\right)$: diffuse noise field.

In the following, the CDR value will be expressed in dB. Figure 2.4 shows exemplarily the coherence of a coherent and diffuse noise field for different CDR values. It can be seen that for $\Psi \leq -9\,\text{dB}$ and for $\Psi \geq 9\,\text{dB}$, the coherence can be approximated by a diffuse and coherent noise field, respectively.

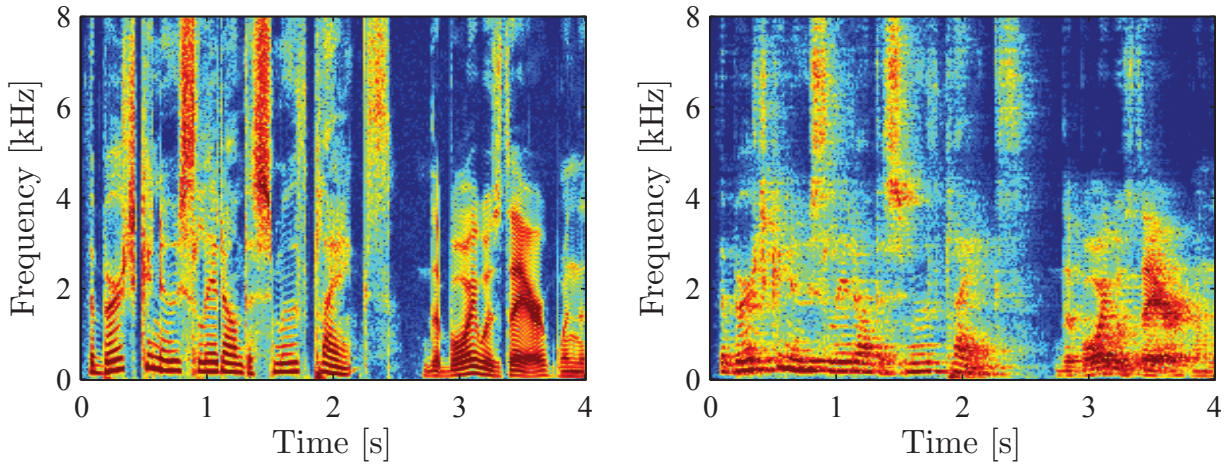


Figure 2.5: Illustration of room reverberation effects by a spectrogram for a speech signal at $f_s = 16\,\text{kHz}$: (left) anechoic speech signal, (right) reverberant speech signal (Lecture room: $d_{\text{LM}} = 4\,\text{m}$, $\text{T}_{60} = 0.69\,\text{s}$, $\text{DRR} = 1.75\,\text{dB}$).

An algorithm for noise field classification which allows to estimate the CDR in the operating range of $-9\,\text{dB} \leq \Psi \leq 9\,\text{dB}$ from a dual-channel noisy observation is presented in Sec. 2.3.6. This approach can also be used to estimate the *Direct-to-Reverberation Energy Ratio* (DRR) blindly from reverberant speech signals as shown later in Sec. 2.3.4.

2.1.3 Room Acoustics and Reverberation

Room acoustics is a special area of acoustics which describes the behavior of sound waves in an enclosed space, cf. [Kut09]. Assuming an *Linear Time Invariant* (LTI) system[2], the acoustic properties of a room in terms of sound propagation and reflections for a specific source-microphone configuration can be completely described by the *Acoustic Transfer Function* (ATF) or RIR. Given a room impulse response $h(k)$ and a discrete time anechoic speech or audio signal $s(k)$, both with sufficient limited frequency bandwidth of, e.g., $f_c = 8\,\text{kHz}$, the reverberant microphone signal can be obtained by

$$x(k) = s(k) * h(k), \tag{2.19}$$

where $*$ indicates discrete convolution. Figure 2.5 illustrates in the left subfigure the spectrogram of an anechoic speech signal. The same signal after convolving with a room impulse response is shown in the right subfigure. The effect of reverberation appears in smearing of the anechoic signal over time in this illustration.

A representation of a room impulse response of length k_r (or T_r in sec.) can be divided

[2]For the description of the acoustical behavior of the enclosure as well as for the derivation of the speech enhancement algorithms, a time-invariant RIR of finite length is assumed.

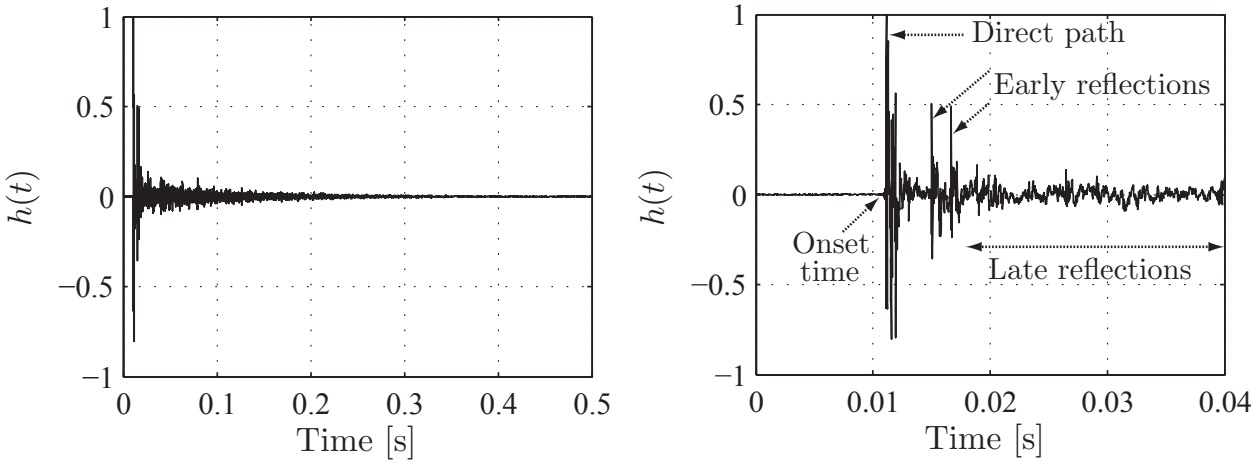


Figure 2.6: Illustration of a room impulse response: (left) RIR with full decay phase, (right) detail of direct and early parts (Lecture room: $d_{\mathrm{LM}} = 4\,\mathrm{m}$, $\mathrm{T}_{60} = 0.69\,\mathrm{s}$, $\mathrm{DRR} = 1.75\,\mathrm{dB}$).

into its direct path and early reflections as well as its late reverberant components by

$$h(k) = \begin{cases} 0 & \text{for } k < 0 \\ h_{\mathrm{early}}(k) & \text{for } 0 \leq k < k_l \\ h_{\mathrm{late}}(k) & \text{for } k_l \leq k \leq k_r, \end{cases} \tag{2.20}$$

where $h_{\mathrm{early}}(k)$ refers to the direct and early path, $h_{\mathrm{late}}(k)$ to the late path and k_l marks the time span after which the late reverberation begins (also referred to as T_l given in sec.). T_l usually lies in the range of $50 - 100\,\mathrm{ms}$, cf. [Kut09].

For the remainder of this work, three different reverberation components are defined:

- *Very early reverberation* is defined as the first 2 ms of early reflections after the direct path. In terms of dual-channel signals, this part contains all reverberation components which are coherent between the microphones,
- *Early reverberation* is given by the full early reflection part of the RIR without direct path,
- *Late reverberation* is referred to the full reverberation tail without direct path and early reflections.

For a further discussion on how to define the boundary between early reflections and late reverberation see, e.g., [HYN07, SS07]. A room impulse response of a lecture room and the indication of the different areas is depicted in Fig. 2.6.

2.1.3.1 Reverberation Time

One fundamental parameter of room acoustics is the *Reverberation Time* (RT). It is defined as the time period a sound needs to decrease by 60 dB from its initial *Sound*

Pressure Level (SPL) after being switched-off and is therefore also referred to as T_{60}. It is linked with the decay rate ρ by

$$\rho = \frac{3\ln(10)}{\mathrm{T}_{60}}. \tag{2.21}$$

Early studies by Sabine have shown that the reverberation time is proportional to the volume of the room and inversely proportional to the amount of absorption in the room [Sab21]:

$$\mathrm{T}_{60}^{(\mathrm{Sabine})} = \frac{24\ln(10)}{\mathrm{c}} \frac{\mathrm{V}}{\alpha_{\mathrm{Sabine}}\mathrm{A}}\,[\mathrm{s}], \tag{2.22}$$

with room volume V, total absorption surface area A and absorption coefficient α_{Sabine}. It can be seen that, in theory, the reverberation time is independent of the distance between source and receiver and that the effect of sound propagation delay is neglected. A related equation is given by Eyring, cf. [Sab21, Kut09].

If a continuous-time room impulse response $h(t)$ is available, the reverberation time can be measured with the so-called Schroeder method [Sch65]. Based on the *Energy Decay Curve* (EDC), which can be obtained from the the so-called *Schroeder integral* by

$$\mathrm{EDC}(t) = \int_t^\infty h^2(\tau)\mathrm{d}\tau, \tag{2.23}$$

the T_{60} can be determined by the time required for the EDC to decay by 60 dB from its initial energy level.

Due to acoustic measurement noise, e.g., self-noise of the microphones and background noise during the measurement, the EDC is usually limited by a noise floor above 60 dB. Therefore, several approaches have been proposed in the past to determine the reverberation time from a non-ideal RIR with a noise floor. As proposed by Schroeder [Sch65], a least squares curve fitting is performed in the EDC region from -5 to -35 dB. The intersection of the obtained fitting line with the 60 dB boundary gives an approximation of the 'true' reverberation time. This procedure is illustrated in Fig. 2.7 for a RIR measured in a lecture room at a loudspeaker microphone distance $d_{\mathrm{LM}} \approx 4\,\mathrm{m}$. It can be seen that the EDC is limited by a noise floor of approx. -50 dB. With the curve fitting procedure, the ground truth of the reverberation time can be determined as $\mathrm{T}_{60} \approx 0.7\,\mathrm{s}$.

The discussed Schroeder method can be applied only if the RIR is known. However, since the RIR is usually time-invariant and not known exactly in a real-system, it has to be estimated blindly or semi-blindly from the reverberant, or possibly reverberant and noisy input signal[3]. Methods for a blind estimation of the reverberation time are discussed in Sec. 2.3.3 along with a method to calculate and blindly estimate the frequency-dependent RT using a *Discrete Cosine Transform* (DCT) filterbank.

[3] System identification approaches where the RIR is estimated are not considered.

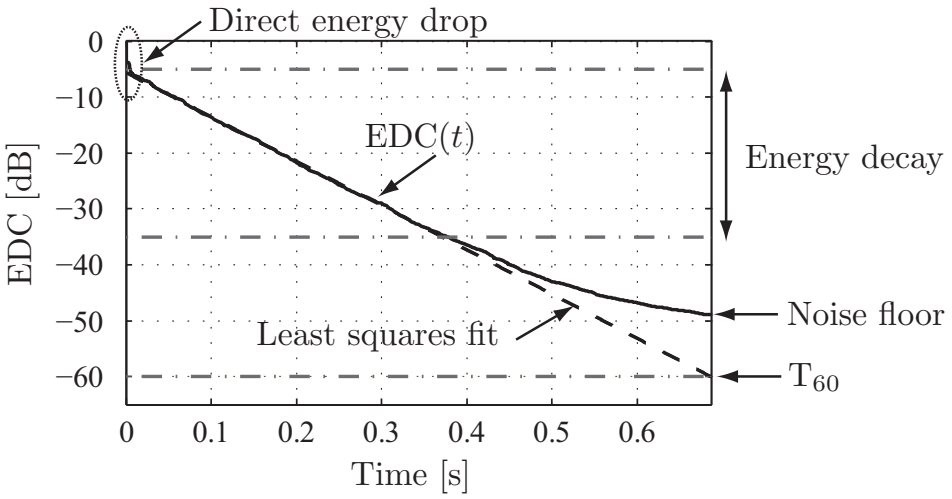


Figure 2.7: Illustration of Schroeder method to determine the reverberation time from a given room impulse response (Lecture room: $d_{\mathrm{LM}} = 4\,\mathrm{m}$, $\mathrm{T}_{60} = 0.69\,\mathrm{s}$, $\mathrm{DRR} = 1.75\,\mathrm{dB}$).

2.1.3.2 Direct-to-Reverberation Energy Ratio

A further very important method for the characterization of a RIR is to measure the energies contained in different parts of the impulse response [Kut09]:

- DRR: Direct-to-reverberation energy ratio,
- ETR: Early-to-total sound energy ratio,
- ELR: Early-to-late reverberation ratio (Clarity index).

In this thesis, only the DRR is considered of importance. It is commonly defined as

$$\frac{\mathrm{DRR'}}{[\mathrm{dB}]} = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{k_d} h^2(k)}{\sum_{k=k_d+1}^{k_r} h^2(k)} \right), \tag{2.24}$$

where k_d denotes the discrete time index where the direct sound ends. This value is usually chosen such that a few *very early reflections* are included in the direct path. Here, it is calculated by $k_d = k_0 + 2\,\mathrm{ms} \cdot \mathrm{f_s}$ with the onset time k_0. The transition times in seconds are denoted by T_d and T_0, respectively, i.e., $T_d = T_0 + 2\,\mathrm{ms}$.

When considering noisy impulse responses, the DRR is usually biased towards lower values. Here, it is suggested to take only the direct part starting from the onset time and the reverberant part up to the reverberation time into account:

$$\frac{\mathrm{DRR}}{[\mathrm{dB}]} = 10 \cdot \log_{10} \left(\frac{\sum_{k=k_0}^{k_d} h^2(k)}{\sum_{k=k_d+1}^{k_{\mathrm{RT}}} h^2(k)} \right), \tag{2.25}$$

where k_{RT} corresponds to the reverberation time in samples.

Novel estimators for the onset time as well as the DRR will be presented in Secs. 2.3.2 and 2.3.4, respectively. The frequency-dependent DRR is calculated by means a DCT filterbank which is described later.

Critical Distance

The DRR also determines the *critical distance* d_c of a sound event. The critical distance is defined as the distance from the source at which the sound energy due to the direct-path component is equal to the sound energy due to reverberation. Hence, the following two cases have to be distinguished:

- DRR $< 0\,\mathrm{dB} \rightarrow$ Source outside the critical distance d_c,
- DRR $> 0\,\mathrm{dB} \rightarrow$ Source within the critical distance d_c.

In the early work by Sabine [Sab21], the critical distance is defined as a function of RT, room volume V and directivity parameter of the sound source[4] Q as

$$d_c^{(\mathrm{Sabine})} \approx 0.1 \sqrt{\frac{Q\,V}{\pi\,\mathrm{T}_{60}}}. \tag{2.26}$$

However, since the reflecting materials and hence the absorption coefficient have different properties for different frequency ranges of the emitting sound source, the DRR as well as the critical distance should be expressed in dependency of the frequency, which is neglected in Eq.(2.26).

2.1.3.3 Statistical RIR Models

Statistical models for the room impulse response and hence, for the reverberation effects are important for the remainder of this work. Such models are used to derive estimators for the short-term PSD of the late reverberant speech. We review four time-domain statistical models which are based on the fundamental discussions by Schroeder [Sch62] and Polack [Pol88]. It has to be mentioned that all discussed models represent only a coarse approximation of reality and assume that several conditions hold, cf. [Hab10, Kut09]. One important assumption is that all models are only valid for frequencies above the Schroeder frequency which is defined as

$$f_{\mathrm{Schroeder}} \approx 2000 \sqrt{\frac{\mathrm{T}_{60}}{V}}, \tag{2.27}$$

i.e., for a room with the dimensions $V = 8 \cdot 5 \cdot 3\,\mathrm{m}^3$ and $\mathrm{T}_{60} = 0.6\,\mathrm{s}$, the statistical models are valid above $f_{\mathrm{Schroeder}} = 141\,\mathrm{Hz}$. In this contribution, this impact can be neglected because of the characteristics of speech signals where most energy is contained above the Schroeder frequency.

[4] For an omnidirectional source: $Q = 1$.

In [Pol88], the room impulse response is described as a sequence of *independent and identically distributed* (i.i.d.) random variables with zero mean and normal distribution $b(k)$, multiplied by an exponentially decaying function as

$$h(k)|_{(\mathrm{PL})} = \begin{cases} b(k)\, e^{-\rho k/f_s} & \text{for } k \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.28}$$

where ρ is the decay rate as in Eq.(2.21). The corresponding energy envelope can be expressed as

$$\mathbb{E}\{h(k)^2|_{(\mathrm{PL})}\} = \sigma^2\, e^{-2\rho k/f_s}, \tag{2.29}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation value and σ^2 the variance of $b(k)$. This model is only valid when the direct path energy is smaller than the energy of all reflections (low DRR, outside the critical distance), cf. [Hab07].

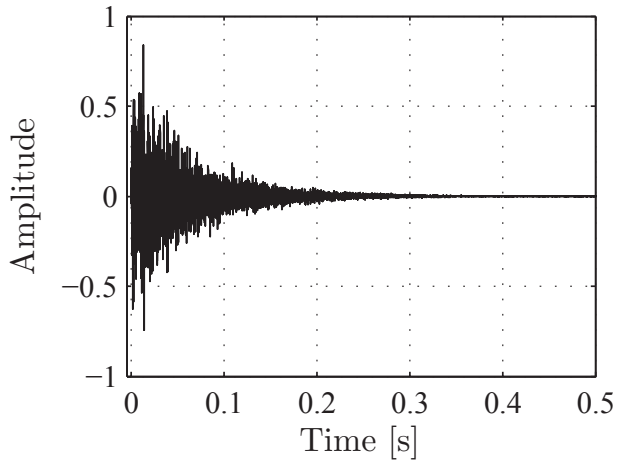
A generalized model was proposed in [Hab07, Hab10]. The RIR is now divided into two segments: one segment which corresponds to the direct path and early reflections and the second segment which describes late reverberation. Hence, this model denoted by HB can distinguish between early and late reverberation and is given by

$$h(k)|_{(\mathrm{HB})} = \begin{cases} b_{\mathrm{d}}(k)\, e^{-\rho_d k/f_s} & \text{for } 0 \leq k < k_d \\ b_{\mathrm{r}}(k)\, e^{-\rho_r k/f_s} & \text{for } k \geq k_d \\ 0 & \text{otherwise,} \end{cases} \tag{2.30}$$

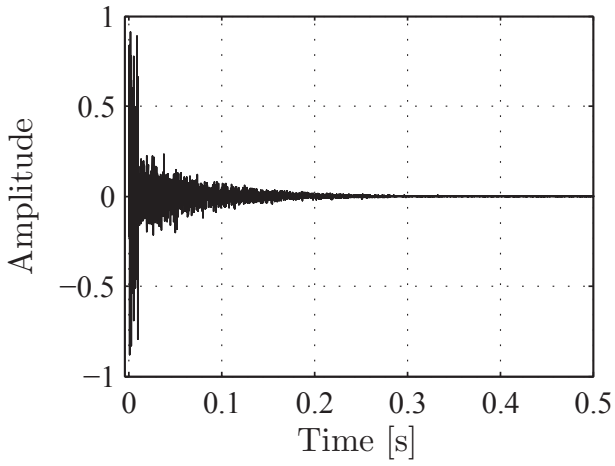
where k_d is chosen as in Eq.(2.25). $b_{\mathrm{d}}(k)$ and $b_{\mathrm{r}}(k)$ represent two uncorrelated noise sequences of early and late reverberation, respectively, which are both i.i.d. random variables with zero mean and normal distribution as in the PL model (see Eq.(2.28)). The variances of $b_{\mathrm{d}}(k)$ and $b_{\mathrm{r}}(k)$ are denoted by σ_d^2 and σ_r^2 and ρ_d and ρ_r represent two decay factors for the direct and late reverberant parts. It is further assumed that $\sigma_d^2 \geq \sigma_r^2$ (high DRR, within the critical distance). The energy envelope of Eq.(2.30) reads

$$\mathbb{E}\{h(k)^2|_{(\mathrm{HB})}\} \begin{cases} \sigma_d^2\, e^{-2\rho_d k/f_s} & \text{for } 0 \leq k < k_d \\ \sigma_r^2\, e^{-2\rho_r k/f_s} & \text{for } k \geq k_d \\ 0 & \text{otherwise.} \end{cases} \tag{2.31}$$

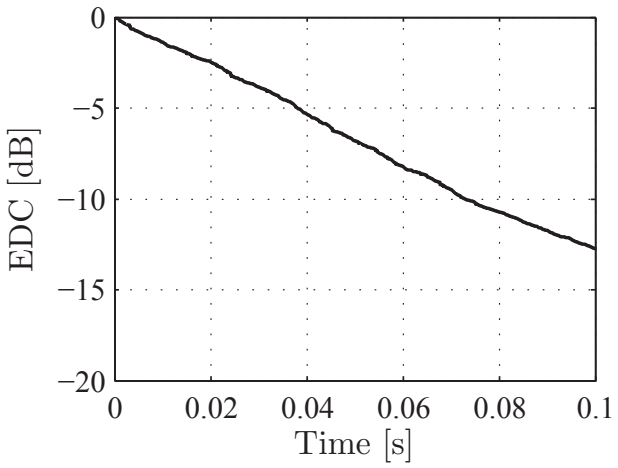
A RIR realization for the two models according to Eqs.(2.28) and (2.30) is illustrated in Fig. 2.8. In the figure, (a) shows exemplarily the RIR using the Polack model with $\mathrm{T}_{60} = 0.5\,\mathrm{s}$ and a variance of $b(k)$ equal to $\sigma^2 = 1$. A RIR realization using the generalized approach is depicted in Subfigure (b). Here, the same RT is used, but with different variances for early and late reverberation, i.e., $\sigma_d^2 = 1$ and $\sigma_r^2 = 1/4$ which corresponds to a DRR of 6 dB. The late reverberant part of the RIR is assumed to begin after 50 ms which corresponds to $k_d = 50 \cdot 10^{-3} \cdot f_s = 800$ samples at a sampling frequency of $f_s = 16\,\mathrm{kHz}$. In (c) and (d), the corresponding energy decay curves are calculated by means of the Schroeder integral applying Eq.(2.23) to the impulse



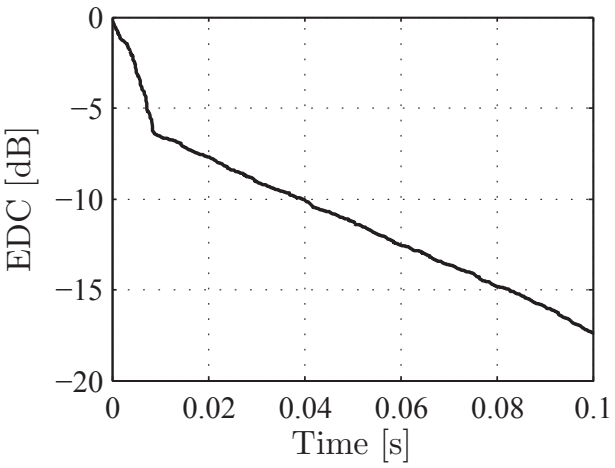
(a) Polack model using Eq.(2.28).



(b) Generalized model using Eq.(2.30) with a DRR of 6 dB.



(c) Energy decay curve of (a).



(d) Energy decay curve of (b).

Figure 2.8: Illustration of two considered statistical RIR models: (top) impulse responses, (bottom) corresponding energy decay curves.

responses. The rapid energy drop of 6 dB between direct part and the beginning of late reverberation after 50 ms can clearly be seen in Subfigure (d). It can be concluded that for high DRRs, the generalized approach is more accurate.

A further statistical RIR model, which is closely related to Eq.(2.31), was proposed by Erkelens in [EH10]. Here, the direct path is modeled by a discrete delta pulse and the late reverberant part as a zero-mean i.i.d. Gaussian process as in Eqs.(2.28),(2.30). The model reads

$$h(k)|_{(\mathrm{EK})} = \begin{cases} 1 & \text{for } k = 1 \\ b_{\mathrm{r}}(k)\, e^{-\rho k / f_s} & \text{for } k > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2.32}$$

and is later used to derive a correlation-based late reverberant speech PSD estimator

in Sec. 3.1.1.2.

An alternative RIR model has recently been published in [EH11]. The model consists of a sum of J decaying cosine functions with random phase according to

$$h(k)|_{(\cos)} = \begin{cases} 1 & \text{for } k = 1 \\ \sum_{j=1}^{J} A_j e^{-\delta_j k} \cos(\omega_\mathrm{j} \mathrm{k} + \phi_\mathrm{j}) & \text{for } k > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2.33}$$

with amplitude A_j and decay constant δ_j. The phases ϕ_j are assumed to be independent and uniformly distributed for each cosine function. The frequencies ω_j correspond physically to the modal frequencies of the room [EH11].

2.1.3.4 Source-Image RIR Model

The source-image model by Allen is associated to the geometrical room impulse response models [AB79]. The main idea behind geometrical RIR models is to replace every reflection on a wall or obstacle by a so-called virtual source or image of the source. These are derived based on the real sound source and symmetries of the reflecting walls. Hence, room reverberation is modeled by a large number of images of the emitting sound source. Details are given in [AB79]. The main advantage of this method are the different degrees of freedom. Nearly any shoebox-like geometry, source-microphone configuration, microphone type and reflection coefficients can be simulated.

The major drawback of this method is the limitation to shoebox-like rooms which does not represent a realistic scenario for most acoustic situations. Moreover, it is inherently difficult to simulate various types of materials with each having different reflection coefficients. Since the number of generated reflections is also limited, e.g., by the processing time, the energy decay can differ from the well-known exponential decay model by Polack [Pol88]. Moreover, shadowing effects of the head cannot be taken into account with this method and hence, for this thesis, measured RIRs are used instead of simulated ones.

2.1.3.5 Measured Room Impulse Responses

The limitation of the source-image method can be overcome by the measurement of room impulse responses in real environments. They can be obtained very effectively with pseudorandom sequences (e.g., *Maximum Length Sequence* (MLS) [Van94] or *Perfect Sequence* (PSEQ) [Ant08]) or sinusoid sweeps [MM01, TAV10]. The main advantage is the accurate reproduction of the acoustic properties of the measurement room. Nevertheless, once the RIR is measured, no changes of the configuration can be made. However, the use of a reasonably large database can help to lower the impact of this restraint.

The so-called *Multichannel Acoustic Reverberation Database at York* (MARDY) has been presented in [WGH+06]. It consists of real measured room impulse responses

of a room with interchangeable panels. By doing so, the acoustic properties can be changed quite easily. The authors measured the RIRs at different source-microphone distances with an eight element linear array at adjacent distances of 0.05 m. The recordings are well-suited for the evaluation of multi-channel dereverberation algorithms but not appropriate for binaural applications where the head influence plays an important role.

The *Aachen Impulse Response* (AIR) database (see App. A and [JSV09]) is a set of impulse responses that were measured in a wide variety of rooms. The initial aim of the AIR database was to allow for realistic studies of signal processing algorithms in reverberant environments with a special focus on hearing aid applications. The database is available free of charge[5]. In the first version [JSV09], it offers *Binaural Room Impulse Responses* (BRIR) measured with a dummy head in different locations with different acoustic properties, such as RT, DRR and room volume. Besides for the evaluation of dereverberation algorithms and for perceptual investigations of reverberant speech, this part of the database allows for the investigation of the head shadowing influence as well as the influence of signal processing algorithms on binaural cues since all recordings where made with and without a dummy head. In a first update [JSEV10], the database was extended to BRIRs with various azimuth angles between head and desired source. This further allows to investigate (binaural) *Direction-of-Arrival* (DOA) algorithms. Since dereverberation can also be applied to telephone speech, the extension published in [JSK+10] includes (dual-channel) impulse responses between the artificial mouth of a dummy head and a dual-microphone mock-up phone. The measurements were carried out in compliance with the *International Telecommunication Union* (ITU) standards for both the hand-held and the hands-free position. For the latest extension, new measurements were carried out in the Aula Carolina (Aachen, Germany) which is a former church with a large ground area of 570 m^2 and a high ceiling that shows very strong reverberation effects ($T_{60} = 4 - 6.6$ s).

For the sake of completeness, the use of real recordings in reverberant rooms instead of convolving an anechoic signal with the RIR has to be mentioned as an alternative. The main advantage of this procedure is the possibility to completely capture all aspects of the acoustic system omitting the LTI assumption that is necessary for the convolution-based concept described above. However, the assumption of linearity is valid as a first approximation for most real-life scenarios where at least short-term invariance can be expected as well. The small remaining advantage does definitely not outweigh the immense loss in flexibility during algorithm development and evaluation. Please refer to Table A.3 in the appendix for an overview of important RIR databases.

[5]Download link: `http://www.ind.rwth-aachen.de/air`

2.2 Acoustic Environment Analysis

2.2.1 Binaural Hearing Aids

Hearing Aids (HAs) aim to improve daily life for hearing impaired people. Even though the main purpose is to compensate for the hearing loss in specific frequency ranges, modern digital devices offer the feed-in of music or telephone signals via wireless connections. In state-of-the-art hearing aids, the devices on both sides of the head are already able to exchange control information, e.g., about the acoustic environment. In the near future, even a data-link with full audio bandwidth can be expected. A detailed introduction into modern hearing aid technologies is out of the scope of this thesis. The reader is referred to, e.g., [Kat08, HCE+05, Dil01] and the references therein for elaborate discussions and future trends.

In this section, the most relevant acoustic parameters, which are required for the development of suitable signal enhancement algorithms, will be discussed. A major difference to other speech enhancement algorithms, e.g., for mobile phones or hands-free car devices is the consideration of binaural hearing.

2.2.1.1 Head-Related Noise Field Coherence

As discussed in Sec. 2.1.2, well-known analytical models exist for homogeneous isotropic noise fields, such as cylindrically and spherically isotropic noise fields. Furthermore, certain features of mixed noise fields can be expressed by a superposition of different auto- and cross-power spectral densities. In contrast to that, the influence of head shadowing on the coherence is usually modeled heuristically. This applies for binaural hearing aids as well as binaural speech transmission systems. Early studies in [LB86] showed that the influence of the head has a severe impact on the noise field coherence and proposed a modified coherence model which is basically a curve fitting procedure derived from measurements with an artificial head. The authors in [BD98] present a model for binaural sound synthesis which is based on the binaural cues. However, this model allows to reproduce binaural sound data with correct interaural time and level difference cues, but gives no information about the binaural coherence. In [KPB09], the distance parameter d_{mic} of the free-field coherence model given by Eq.(2.6) is simply scaled in order to take the modified coherence into account.

A semi-analytical model for the binaural noise field coherence which is used throughout this work can be found in [JDV11, Dör98] and is described in more detail in App. B. The derivation employs Kirchhoff's diffraction theory and Babinet's principle, cf. [BW99]. The main advantage compared to the previous models is that arbitrary dimensions for head and microphone distances can be employed and no acoustic measurements as in [LB86, KPB09] are required.

For the verification, room impulse responses in a reverberant environment ($d_{\mathrm{LM}} = 4\,\mathrm{m}$, $\mathrm{T}_{60} = 0.69\,\mathrm{s}$, $\mathrm{DRR} = 1.75\,\mathrm{dB}$) are measured with an artificial head ($d_{\mathrm{head}} = 0.15\,\mathrm{m}$, $r_{\mathrm{head}} = 0.075\,\mathrm{m}$). The two microphones are positioned close to the

pinna at 1 cm from the ear canal such that $d_{\mathrm{mic}} = 0.17\,\mathrm{m}$. The measurements were repeated in an otherwise unchanged experimental setup after the head was removed to examine the influence of the head. In order to evaluate the noise field coherence of the late reverberant part only, the coherent direct and early parts from the impulse response were removed.

The measured impulse responses are convolved with a speech signal signal having a duration of 1 min. The coherence curves are calculated by means of the recursive periodogram approach, which is introduced in more detail in Sec. 2.3.1, from the reverberant speech signals. Figure 2.9 shows the corresponding curves for two microphones at a distance of $d_{\mathrm{mic}} = 0.17\,\mathrm{m}$ (gray). The theoretical curves represent the ideal 3D and 2D noise field without head (free-field) by Eqs.(2.6) and (2.9), respectively (red).

The corresponding 3D and 2D curves with head shadowing are determined without proof by solving (see App. B for details)

$$\Gamma_{x_1x_2}(\Omega) = \frac{2\int\limits_0^{\pi/2} \mathrm{Re}\left\{\mathrm{H}_1(\Omega,\theta)\,\mathrm{H}_2^*(\Omega,\theta)\right\}\,\sin\theta\,\mathrm{d}\theta}{\int\limits_0^{\pi/2}\left(|H_1(\Omega,\theta)|^2 + |H_2(\Omega,\theta)|^2\right)\,\sin\theta\,\mathrm{d}\theta} \tag{2.34}$$

with and without the $\sin\theta$-terms and are plotted in black. All functions are given as the squared magnitudes of the coherence function, i.e., MSC. Regarding the figure, we can conclude that the proposed coherence model accurately approximates the measured coherence. Especially for hearing-aid algorithms, assuming a cocktail-party environment, the proposed 2D model is the most appropriate one.

This coherence model can be used to investigate the influence of head-shadowing on coherence-based speech enhancement algorithms, in binaural noise reduction or dereverberation algorithms where the binaural coherence is exploited explicitly and to generate realistic binaural noise fields for simulations. A MATLAB reference implementation is available online[6].

2.2.1.2 Analysis of Background Noise

Since hearing aid users wear their devices permanently in everyday life, the acoustic situation varies often. The occurring background noise can, for example, vary from street noise to babble noise. Here, we restrict the analysis to a few specific noise types which are very challenging for hearing aid users. The so-called *cocktail-party environment* is the most prominent and most difficult situation.

For the signal processing algorithm, it is of significant importance how to characterize the existing noise field. For the evaluation, binaural background noise which was recorded with an artificial head from the ETSI background noise database [ETS09] is used. In contrast to the measurements of the AIR database where the microphones

[6]Download link: `http://www.ind.rwth-aachen.de/~bib/jeub11`

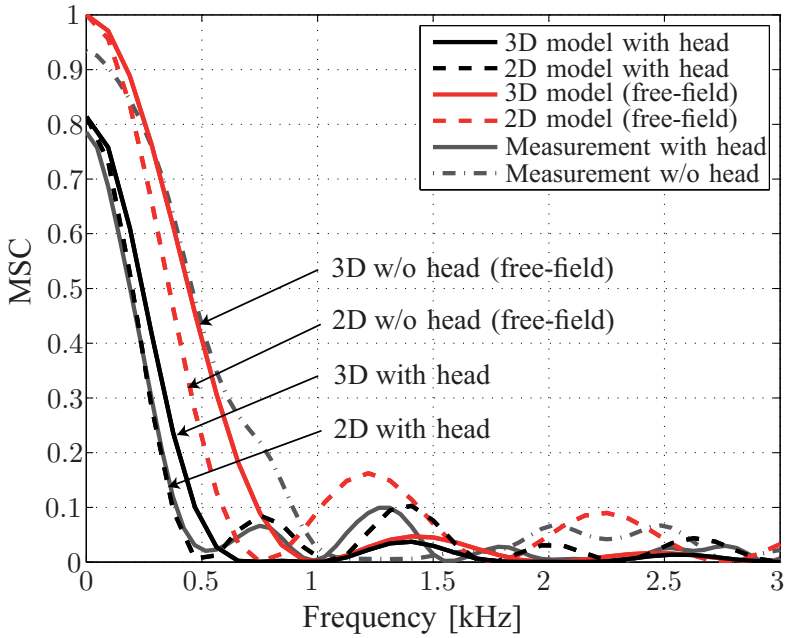


Figure 2.9: MSC of ideal diffuse noise field and shadowing influence. Plotted are the theoretical curves (free field models and the novel binaural coherence models) and results from measurements in a reverberant environment (Lecture room: $d_{\mathrm{LM}} = 4\,\mathrm{m}$, $\mathrm{T}_{60} = 0.69\,\mathrm{s}$, $\mathrm{DRR} = 1.75\,\mathrm{dB}$).

where placed next to the pinna, the ETSI database uses recordings from the ear microphone ($d_{\mathrm{mic}} = 0.15\,\mathrm{m}$) where the influence of the ear channel has been compensated by an equalization filter. In the following discussion, it will be shown exemplarily that the noise field can mainly be described as diffuse and homogeneous, which is a very important assumption for the considered speech enhancement algorithms. Figure 2.10 exemplarily shows the long-term PSD of the cafeteria and kindergarten noise for the left and right ear (using the ETSI recordings which have a maximum duration of 29 s). It can be seen that the left and right ear signals have roughly the same PSD. Hence, the common assumption of a homogeneous noise field holds for these recordings.

In order to investigate the noise field in terms of correlation among the two ear signals, the long-term MSC is calculated from the noise signals. From Fig. 2.11, we can conclude that the noise field can be characterized as diffuse. Plotted is the measured MSC of the signal containing background noise only and the theoretical MSC using the proposed binaural coherence model for an inter-microphone distance of $d_{\mathrm{mic}} = 0.15\,\mathrm{m}$.

2.2.1.3 Analysis of Reverberation

While the auditory system of normal-hearing people is usually capable of reducing room reverberation by means of binaural processing, this ability is mostly degraded for many types of hearing loss [Bla96, Kat08]. Especially for hearing impaired people, room reverberation has a distinct influence on intelligibility and listening comfort if the DRR lies below 0 dB or 10 dB, depending on the type of hearing loss [See11]. Thus,

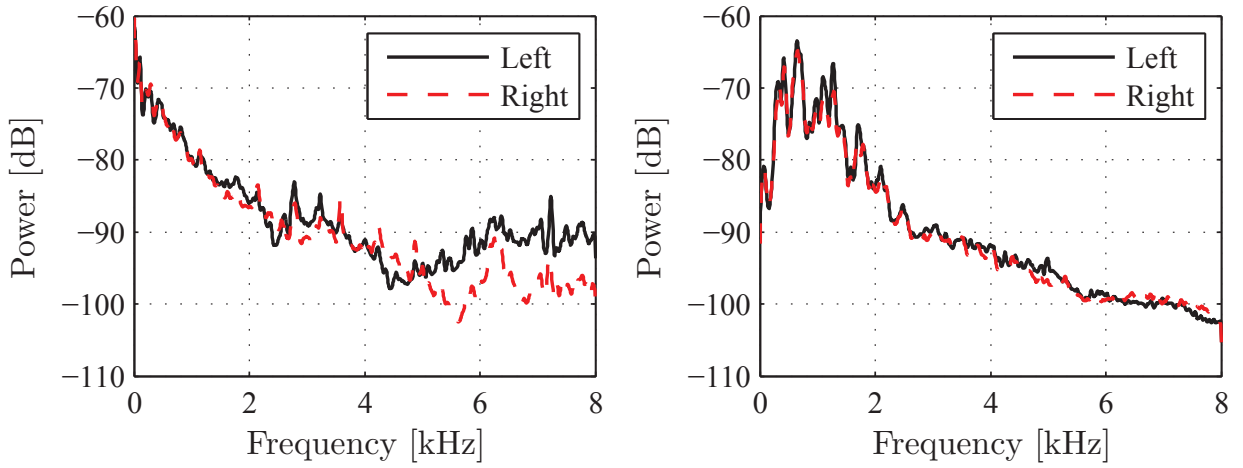


Figure 2.10: Long-term PSD of background noise recorded at left and right ear of an artificial head in two real environments: (left) cafeteria noise, (right) kindergarten noise.

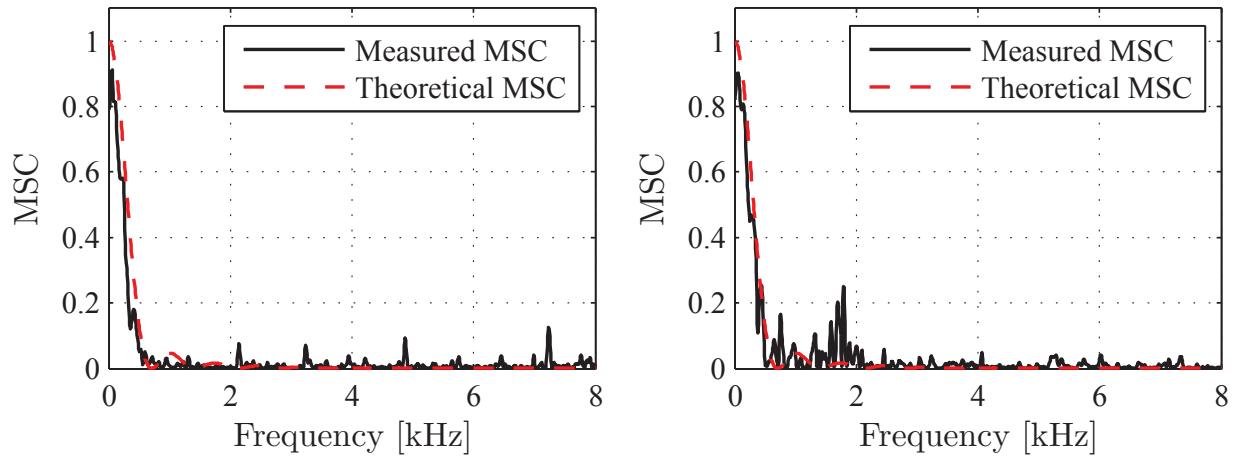


Figure 2.11: MSC of background noise recorded at left and right ear of an artificial head in two real environments: (left) cafeteria noise, (right) kindergarten noise. Measured curves are marked by the solid black line, the theoretical MSC curves using the proposed binaural coherence model are indicated by the dashed red lines.

suitable algorithms for dereverberation should be incorporated in modern hearing aids.

In the sequel of this section, the analyses in terms of RT and DRR are based on measured room impulse responses taken from the AIR database. In order to show the frequency dependency, all plots are generated by decomposing the RIR into subbands by means of a DCT filterbank and by applying the Schroeder method or the DRR calculation procedure individually to each of the subbands. This procedure is described in more detail in Sec. 2.3.3.1.

In the first analysis, the frequency-dependent reverberation time of three different rooms is shown in Fig. 2.12. For larger rooms like the considered lecture room and

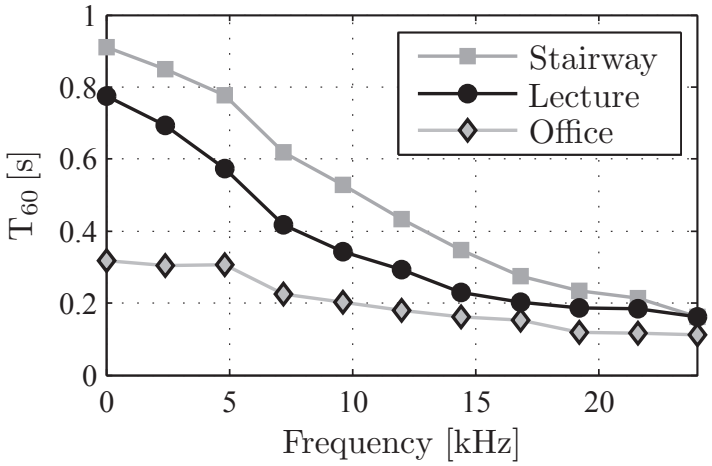


Figure 2.12: Reverberation time over frequency for three different room impulse responses from the AIR database.

stairway hall, a distinctive difference in T_{60} for smaller and higher frequencies can be observed, which is usually the case for larger rooms with highly reflecting materials such as glass or concrete walls [Kut09]. In contrast to that, the smaller office room has a nearly flat RT over frequency. From this figure, we conclude that the frequency-dependency of the RT has to be taken into account for a dereverberation algorithm.

Figure 2.13 shows in the left subfigure the DRR over the source-microphone distance d_{LM} for the stairway, lecture and office room, which are calculated from the measured RIR by applying Eq.(2.25). In the right plot, the DRR over frequency for a fixed distance of $d_{\mathrm{LM}} = 1\,\mathrm{m}$ (office, stairway) and $d_{\mathrm{LM}} = 4\,\mathrm{m}$ (lecture) is plotted. The critical distance is marked by the dashed 0 dB-line. For smaller distances of $d_{\mathrm{LM}} < 3\,\mathrm{m}$, the source is mainly within the critical distance and very high DRR values occur when the source is located at a distance of $d_{\mathrm{LM}} \approx 1\,\mathrm{m}$.

In terms of DRR over frequency, the DRR shows, as expected from the RT plots, also a high frequency-dependency. In order to take into account the special boundary conditions for binaural hearing aids, the DRR is also investigated over the azimuth angle of the source θ. The strong variation of the DRR for different azimuth angles is depicted in Fig. 2.14 and can be explained with the shadowing effects of the head. Similar dependencies could be observed for the T_{60} .

This subsection has shown that both DRR and RT exhibit a high dependency on frequency and show strong variations for different azimuth angles of the source. For binaural hearing aids, the desired source is mainly within the critical distance. This motivates the use of models which take the frequency-dependency into account as well as the use of the generalized statistical RIR model given by Eq.(2.30).

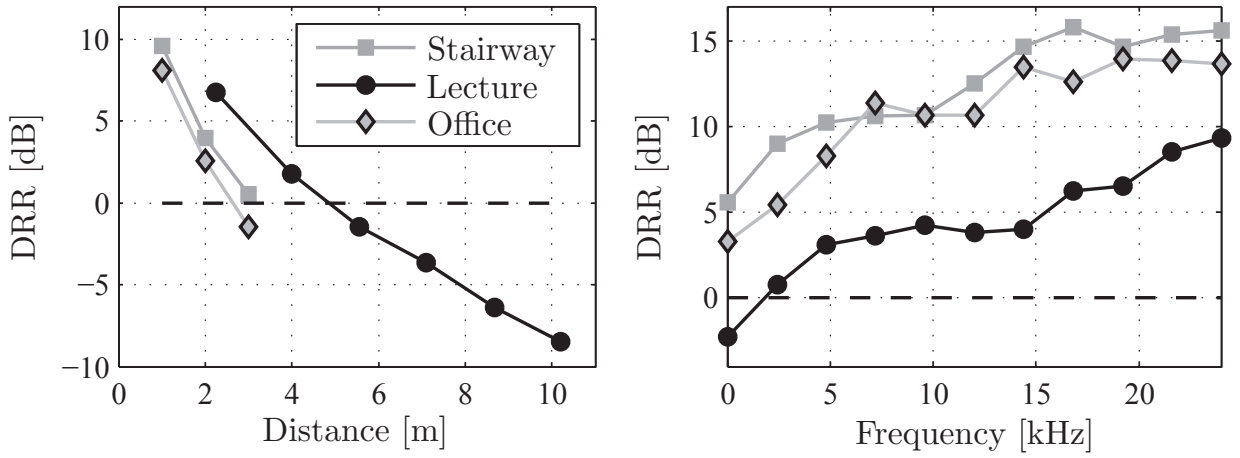


Figure 2.13: DRR: (left) frequency-independent (broadband) over the source-microphone distance d_{LM}, (right) frequency-dependent over frequency for a fixed distance of $d_{\mathrm{LM}} = 1\,\mathrm{m}$ (office, stairway) and $d_{\mathrm{LM}} = 4\,\mathrm{m}$ (lecture). The horizontal dashed lines represent the critical distance.

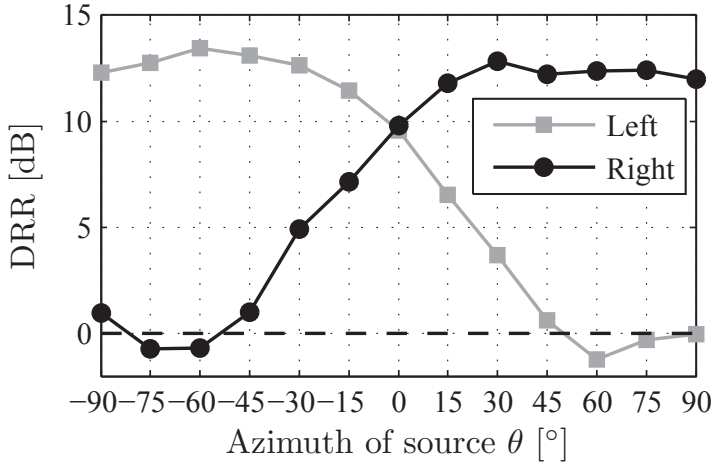


Figure 2.14: DRR measured in a stairway hall ($d_{\mathrm{LM}} = 1\,\mathrm{m}$, $\mathrm{T}_{60} = 0.72\,\mathrm{s}$) at different azimuth angles of the source signal in the presence of a dummy head.

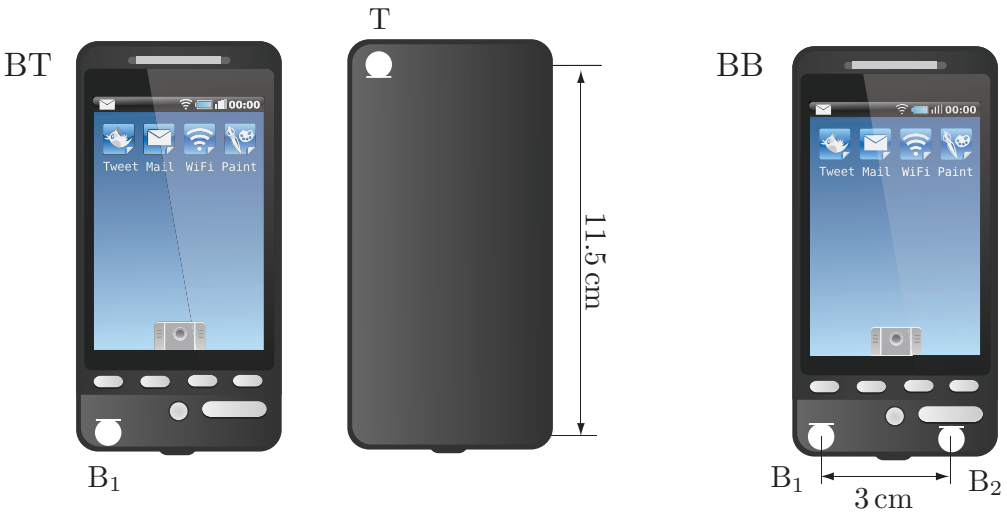


Figure 2.15: Illustration of mobile-phone with the two considered microphone positions. (left) bottom-top (BT), (right) bottom-bottom (BB).

2.2.2 Mobile Phones

Common mobile phones use a single microphone for capturing the desired speech signal. This so-called primary microphone is usually mounted at the bottom of the device in order to allow for a short acoustic path between mouth and microphone, which ensures a high direct path energy and less reverberation. Depending on the phone design, a secondary microphone can be placed either at the bottom next to the primary microphone, or on the backside on top of the device in order to capture the speech signal with a lower sound pressure level.

In the remainder of this thesis, two different dual-microphone configurations are discussed:

- *Bottom-Top* (BT): Primary microphone at the bottom on the front side, secondary microphone at the backside on top of the device,
- *Bottom-Bottom* (BB): Primary and secondary microphone both placed on the front side at the bottom of the device.

The corresponding drawings are shown in Fig. 2.15. Due to the geometrical limitations, only small inter-microphone distances are possible. Here, it is assumed that the maximum distance for BT is approximately 12 cm and 3 cm for BB.

A further important aspect regarding the acoustic situation is the handling of the phone. In order to describe the geometry of head and phone, a definition of *Ear Reference Point* (ERP) and *Mouth Reference Point* (MRP) is introduced first by means of Fig. 2.16. The MRP is located at a distance of 0.025 m in front of the lips on the horizontal axis through the center of the opening of the mouth. It is defined in the absence of any obstruction [ITU07]. The ERP, a so-called virtual point for geometric reference, is located at the entrance to the listener's ear, traditionally used for calculating telephonometric loudness ratings [ITU07].

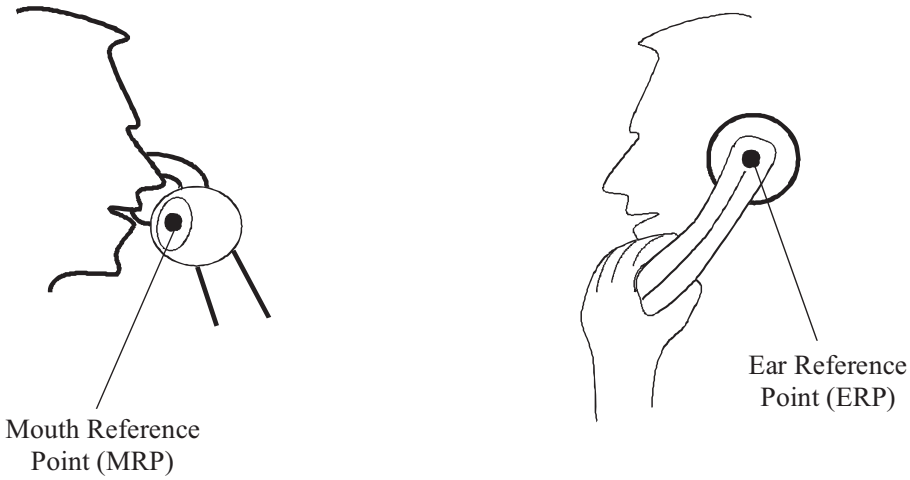


Figure 2.16: Definition of mouth and ear reference points according to ITU-T P.64 [ITU07].

Mainly two different standardized phone positions are to be distinguished:

- *Hand-Held Position* (HHP) [ITU07] and
- *Hands-Free Reference Point* (HFRP) [ITU00].

In the "classical" HHP, the phone is held directly at the ear and, hence, the smallest possible distance between mouth and bottom-microphone exists. When the phone is used in hands-free mode, a distance between mouth and primary microphone of up to 0.5 m can occur [ITU00]. This however, results in a lower direct path energy of the desired speech signal and more reverberation (low DRR). In addition, more background noise is captured. For the hand-held position, this distance is depending on the phone geometry and handling and can vary between 0.05 and 0.10 m.

2.2.2.1 Analysis of Background Noise

The following background noise analysis is based on measurements inside an acoustic chamber using the standardized multi-loudspeaker procedure described in [ETS09] to generate realistic noise fields. Here, the analysis is restricted to two important noise types: car and babble noise from [ETS09].

The recording systems consists of a HEAD acoustics HMS II.3 artificial head according to ITU-T Rec. P.58 [ITU96] including a mouth simulator. Two mock-up phones (BB and BT microphone configuration) each containing two omnidirectional Beyerdynamic MM1 measurement microphones integrated in a 6x12x3 cm^3 plastic housing are used. The microphones are placed according to Fig. 2.15. For recording the speech signals in the HHP, the phone was mounted on the artificial head by means of the HEAD acoustics HHP 3 hand-held positioner in the so-called flat handset position in accordance with ITU-T P.64 Annex D.3 [ITU07]. Further details on the measurement systems are described in App. A.

Important acoustic quantities are the PSD recorded at the positions of the three microphones for both speech and noise. Figure 2.17 shows exemplarily the long-term

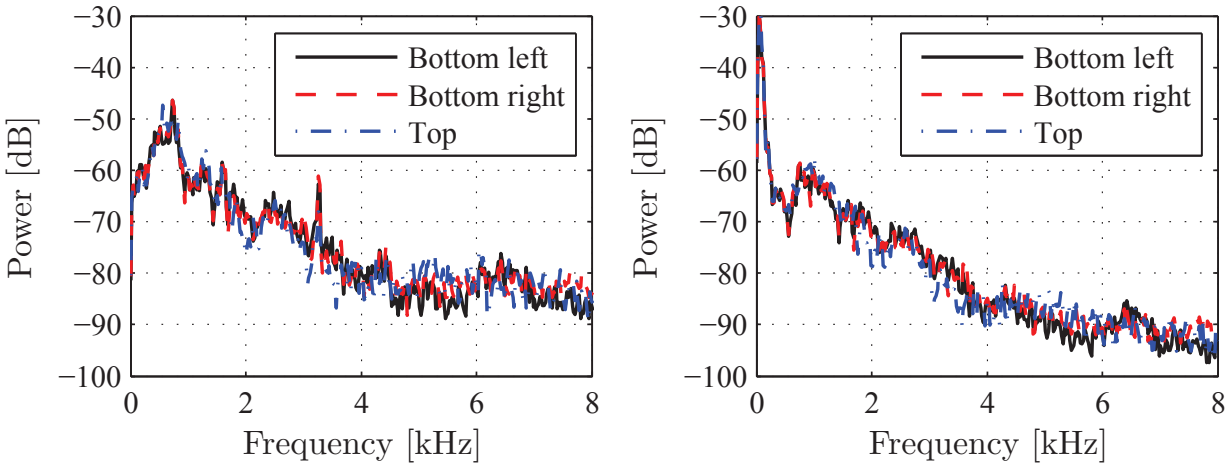


Figure 2.17: Long-term PSD of background noise captured by the three microphones: (left) pub noise, (right) traffic noise. Plotted are the PSDs for the two bottom microphones as well as the top microphone.

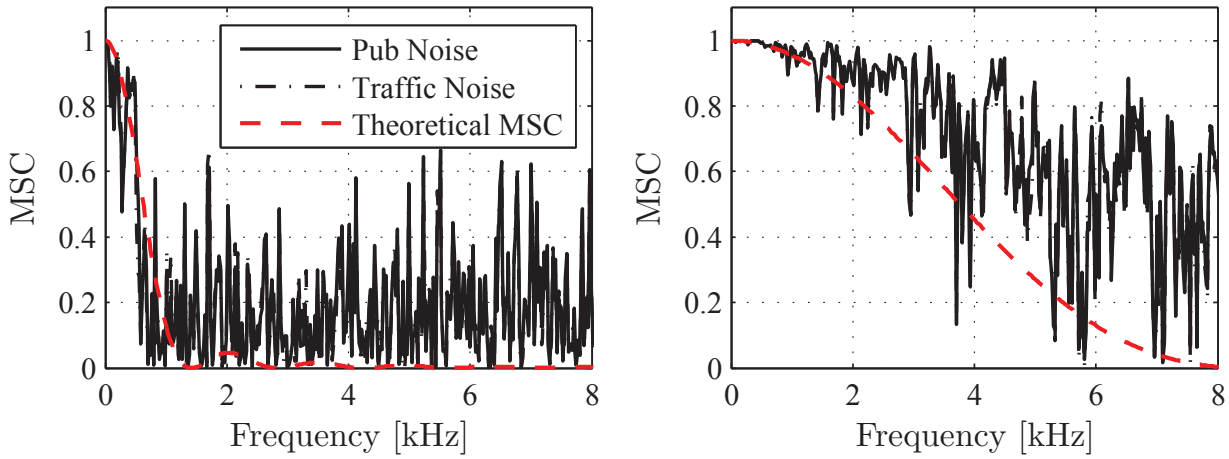


Figure 2.18: MSC of background noise: (left) between top and bottom microphones (BT), (right) between bottom left and bottom right microphones (BB).

PSDs of pub noise (left) and traffic noise (right) for the two bottom microphones (B_1, B_2) and the top microphone. It can be seen that all signals have roughly the same PSD among the microphones and hence, a homogeneous noise field exists as confirmed by analysis of further noise types. To investigate the noise field in terms of correlation among the two microphones, the MSC between bottom (B_1) and top (T) microphone and between the bottom microphones (B_1, B_2) is calculated from the noise-only signals. The MSC is compared to the theoretical MSC using the free-field diffuse model with the corresponding inter-microphone distances as depicted in Fig. 2.18. From the evaluation of our recordings we can conclude that the considered noise fields can be characterized as diffuse.

All experiments with noise-only conditions have also been verified with the same mock-up phone, which was placed outside in crowded places (here: Aachen Christmas market) and a busy street.

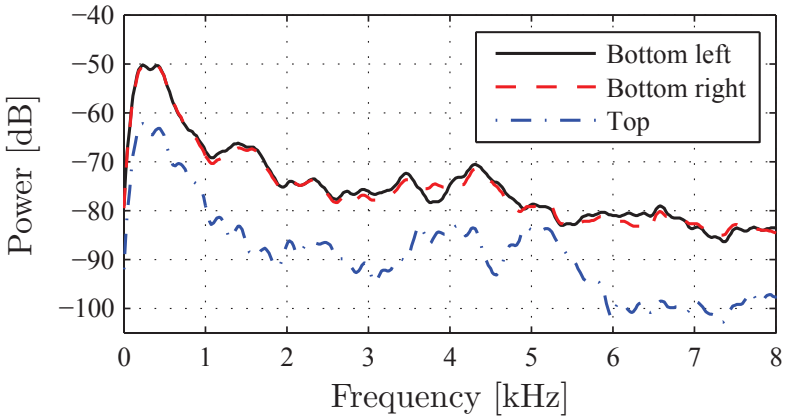


Figure 2.19: Attenuation of speech signal from artificial mouth of dummy head to the three different microphones.

2.2.2.2 Analysis of Desired Speech

Since the microphones can be placed at two different positions, i.e., BT and BB on the phone, it is of significant importance to investigate the attenuation of the desired speech signals from the mouth to the different microphones. Figure 2.19 shows the long-term PSD of a speech signal picked up by the three microphones (noise-free case) where a *Power Level Difference* (PLD) of approx. 10 dB is measured between the bottom and top microphone for all frequencies. In contrast to that, it can be seen that the power level difference between the microphone signals B_1 and B_2 can be neglected.

The MSC of the captured speech among the microphones is investigated as before and shown in Fig. 2.20. A high coherence for the entire frequency range is observed for the BB alignment. Significant notches can be seen in the MSC plot for the BT alignment and, hence, algorithms which rely on a perfect coherent speech signal at both microphones are expected to exhibit a loss in performance for this configuration. It has to be mentioned that the notches are mainly caused by reflections and scattering of the soundwaves in the acoustic chamber and the hand-held positioner.

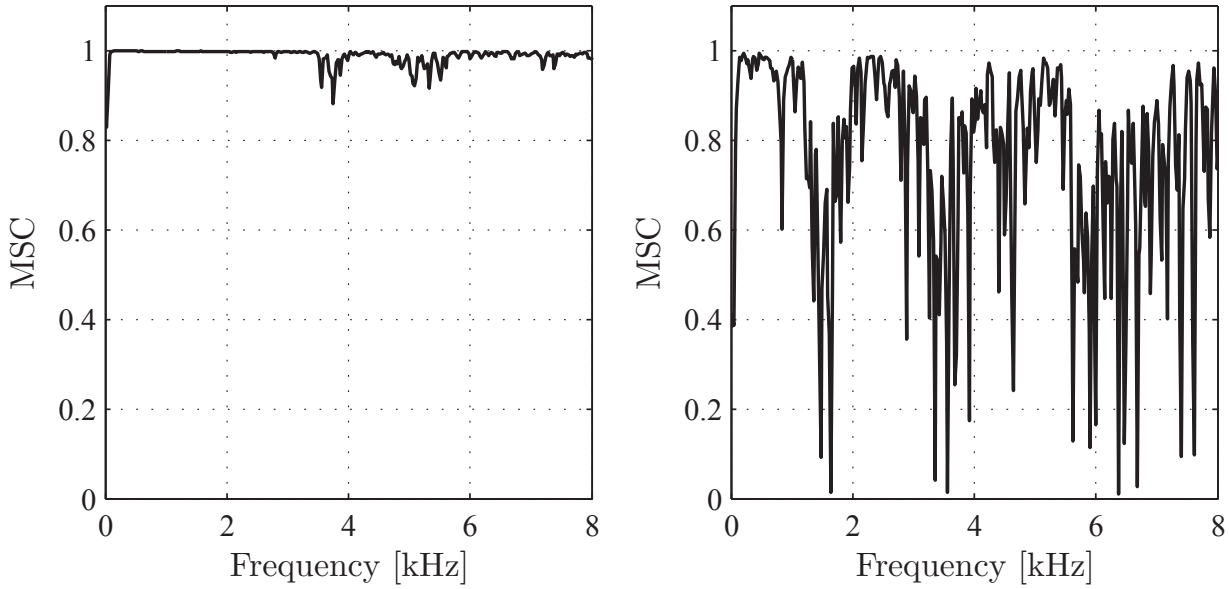


Figure 2.20: MSC of speech signal from artificial mouth of dummy head between two microphones: (left) bottom-bottom (BB), (right) bottom-top (BT).

2.2.2.3 Analysis of Reverberation

The influence of reverberation on the intelligibility in hand-held telephony is commonly assumed to be negligible. In the following, it will be shown based on measurements and a listening experiment that this statement is not always true [JSK+10]. A typical RIR from a corridor location captured with dummy head and mock-up phone is depicted in Fig. 2.21. The difference between the HHP (left) and HFRP (right) can clearly be seen. In the hands-free case, the direct signal arrives later than in the hand-held position due to the longer sound propagation delay. Moreover, the direct path energy is much lower.

For the experiments, several speech codecs are employed to the reverberant speech signal in order to emulate realistic transmitted speech signals. These signals are used for a subsequent informal listening test. In the context of *Code Excited Linear Prediction* (CELP) speech codecs, it is already known that the effects of room reverberation are reduced by means of the adaptive postfilter [CG95] employed in the speech decoder [JV09b]. However, a sufficient dereverberation cannot be obtained by such processing, especially since the postfilter is employed after decoding at the receiver side. The single-channel test signals for the experiments are generated as follows. First, speech files $s(k)$ from the TSP speech database [Kab02] are convolved with the impulse responses $h(k)$ between artificial mouth and microphones of the (BB) mock-up phone at $f_s = 48\,\text{kHz}$, providing the reverberant signals $x(k)$. Second, the reverberant speech signals $x(k)$ are downsampled to f_s', encoded and decoded independently using three different speech codecs with sampling frequency, bandwidth and bit rates as follows:

- *Adaptive Multi-Rate Narrowband* (AMR-NB) codec [3GP04a]
  $f_s' = 8\,\text{kHz}$, $3.4\,\text{kHz}$, 12.2 kbit/s

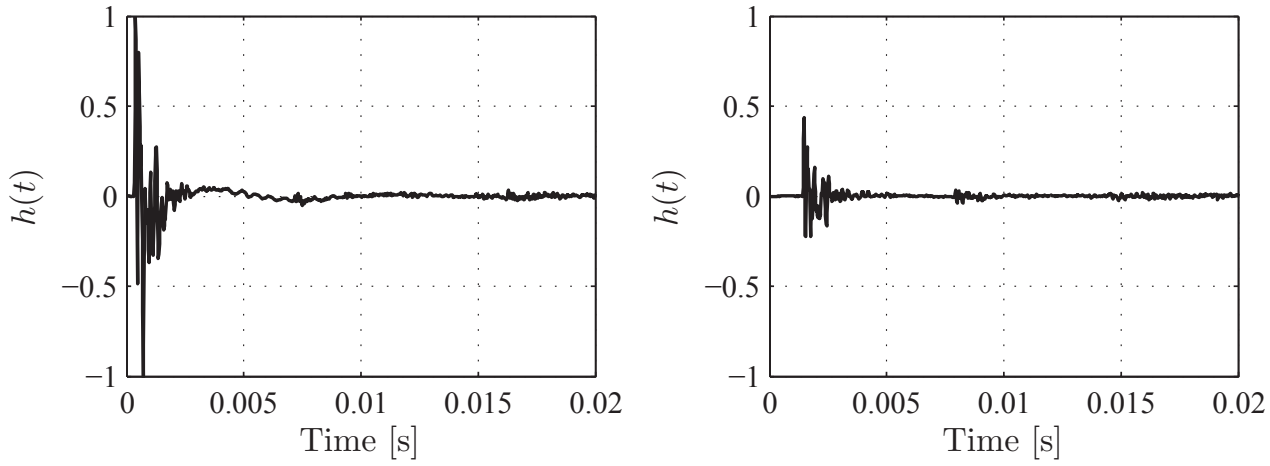


Figure 2.21: Room impulse response measured with BB mock-up phone and dummy head in a corridor: (left) *Hand-Held Position* (HHP) ($T_{60} = 0.98\,s$, DRR = 12.78 dB), (right) *Hands-Free Reference Point* (HFRP) ($T_{60} = 1.34\,s$, DRR = 6.51 dB).

Table 2.2: Channel-based measures calculated directly from the impulse responses. The results are averaged over both channels (BB).

Room	RT [s]		DRR [dB]	
	HHP	*HFRP*	*HHP*	*HFRP*
Office	0.4	0.52	12.27	5.28
Kitchen	0.42	0.52	11.18	4.62
Corridor	0.98	1.34	12.78	6.51
Stairway	1.31	1.51	10.87	4.7

- *Adaptive Multi-Rate Wideband* (AMR-WB) codec [3GP04b] $f_s' = 16\,kHz$, 7 kHz, 23.05 kbit/s
- *Super-Wideband* (SWB) speech and audio codec [GKL+09] $f_s' = 32\,kHz$, 14 kHz, 64 kbit/s

The reverberant and transcoded signals are denoted by $\tilde{x}(k)$ in the following. For simplicity, no bit errors were added.

2.2.2.4 Objective Evaluation

Table 2.2 shows the results in terms of DRR and RT where further details are provided by App. A.3. It can be seen from Table 2.2 that the room acoustic measures differ greatly between HHP and HFRP. This can be explained with the direct path between loudspeaker and microphone at the HFRP and the indirect sound propagation for the HHP. For both office and kitchen, a moderate reverberation time of approx. ≤ 0.5s was measured and no significant difference in the RT between HHP and HFRP was

Table 2.3: ITU-R BS.1284-1 five-grade impairment scale [ITU03].

5.0	Imperceptible
4.0	Perceptible but not annoying
3.0	Slightly annoying
2.0	Annoying
1.0	Very annoying

examined. However, the DRR differs by more than 6 dB between HHP and HFRP for all measured rooms. Compared to Fig. 2.13 for the binaural DRR ranges, in the HHP case used for mobile telephony, positive DRR values always occur. Regarding the corridor and stairway scenario, a high RT was measured even for the hand-held position. Since the DRR values are always positive, i.e., the source is within the critical distance, the use of the generalized statistical RIR model should be considered here as well.

A further analysis of the acoustic environment in terms of the frequency-dependent reverberation time, a coherence analysis of the reverberant sound field as well as results from the PEMO-Q objective measure [HK06] has been carried out and is given in [JSK+10]. These investigations show that the reverberant sound field is diffuse and that the reverberation time highly depends on the frequency. Furthermore, the results of the PEMO-Q measure show a high correlation to the conducted subjective listening test.

2.2.2.5 Subjective Evaluation

The listening test took place in a low-reverberant studio booth having a high sound isolation of 42 dB against exterior noise. A calibrated HEAD Acoustics PEQ V digital equalizer in combination with a Sennheiser HD600 headphone was used. During the test with 30 experienced listeners (normal hearing, age: $24 - 33$ years), 24 different signals $\tilde{s}(k)$ were presented to the participants. An anechoic speech signal of 18 s duration was processed according to the considered transmission system described before. Each of the 8 signals (4x HHP, 4x HFRP) were transcoded with the *Narrow-band* (NB), *Wideband* (WB) and *Super-Wideband* (SWB) codec after the convolution with a room impulse response. For each of the sentences, the listeners were asked to rate the impairment according to the ITU-R BS.1284-1 five-grade impairment scale (see Table 2.3 and [ITU03]). The signals could be played ad libitum before the probands had to make their judgments. Since the listeners were not asked to rate the overall speech quality but only the impairment due to room reverberation, the results of the different codecs do not represent a quality rating.

The results averaged over the scores of the 30 participants and over the three codecs are depicted in Fig. 2.22. It can be seen from the left figure, that reverberation is perceptible for all tested scenarios. The listening test shows that most listeners rated the effect for office and kitchen as perceptible but not annoying in the hand-held

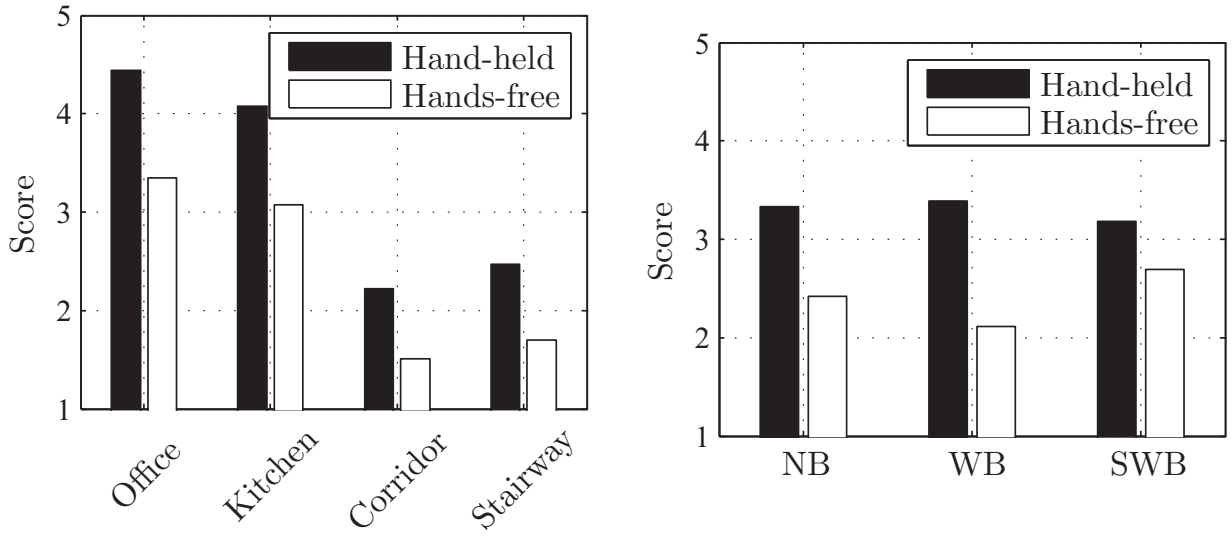


Figure 2.22: Results of listening test according to the ITU-R five-grade impairment scale: (left) averaged over different codecs and (right) averaged over different rooms for each codec.

case. In terms of the corridor and stairway sentences, the effects of reverberation are clearly perceptible and rated as annoying. As expected, the impairment scores for the hands-free positions are always lower.

In a second experiment, the subjective scores for the different codecs are evaluated and are shown in Fig. 2.22 (right). It can be observed that no significant difference exists among the tested codecs. This corresponds to the investigations in [RWS09] where different wideband codecs where investigated under reverberant conditions.

Based on the objective evaluation as well as a listening test with 30 participants, it has been shown that an impairment due to reverberation can always be observed. For small enclosures like the tested office and kitchen, the effects are mostly not rated as annoying. For other enclosures (with higher reverberation) like stairway halls or corridors, room reverberation has a strong influence on the intelligibility and we conclude that dereverberation algorithms should be applied for both hands-free and hand-held telephones.

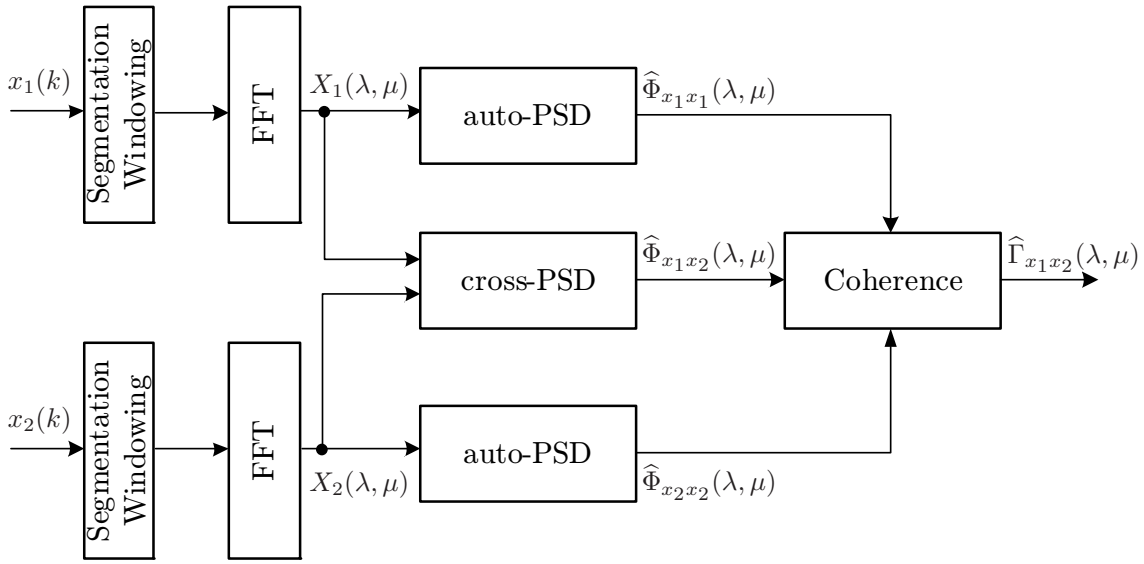


Figure 2.23: Principle of short-term coherence estimation in the frequency domain.

2.3 Estimation of Important Acoustic Parameters

2.3.1 Short-Term Coherence Estimation

To determine a short-term coherence estimate from segments of the two input signals $x_1(k)$ and $x_2(k)$ in a practical implementation, the short-term auto- and cross power spectral densities in Eq.(2.3) are calculated from periodograms $|X_m(\lambda,\mu)|^2$ according to [Wel67, CKN73] as sketched in Fig. 2.23. The estimates of the short-term auto- and cross-PSD read

$$\widehat{\Phi}_{x_1x_1}(\lambda,\mu) = \alpha^{(\mathrm{PSD})} \cdot \widehat{\Phi}_{x_1x_1}(\lambda-1,\mu) + \left(1-\alpha^{(\mathrm{PSD})}\right) \cdot |X_1(\lambda,\mu)|^2, \tag{2.35}$$

$$\widehat{\Phi}_{x_2x_2}(\lambda,\mu) = \alpha^{(\mathrm{PSD})} \cdot \widehat{\Phi}_{x_2x_2}(\lambda-1,\mu) + \left(1-\alpha^{(\mathrm{PSD})}\right) \cdot |X_2(\lambda,\mu)|^2, \tag{2.36}$$

$$\widehat{\Phi}_{x_1x_2}(\lambda,\mu) = \alpha^{(\mathrm{PSD})} \cdot \widehat{\Phi}_{x_1x_2}(\lambda-1,\mu) + \left(1-\alpha^{(\mathrm{PSD})}\right) \cdot X_1(\lambda,\mu) \cdot X_2^*(\lambda,\mu). \tag{2.37}$$

The smoothing factor $0 \leq \alpha^{(\mathrm{PSD})} \leq 1$ controls the balance between smoothing for variance reduction and tracking of non-stationary signal characteristics. Here, $\{\cdot\}^*$ denotes the complex conjugate.

In [Wel67, CKN73] it is suggested to use overlapping frames to reduce the variance of the estimate. As a compromise in terms of variance reduction and computational complexity, the common calculation using half-overlapping frames (50%) is used throughout this work.

In theory, if the signal $x_1(k)$ is a linear filtered version of the signal $x_2(k)$ or vice versa, the coherence tends always to one. However, when it comes to the short-term coherence estimation, this is valid if the block size for the coherence estimation is larger than the length of the filter impulse response, cf. [Mar95]. This effect is illustrated in Fig. 2.24 where *White Gaussian Noise* (WGN) is convolved with a set of

binaural room impulse responses (Lecture room, length of RIR: $k_r = 21845$, $\mathrm{T}_r = 2.73\,\mathrm{s}$). For the top subfigures, the late reverberation part of the RIR (see Eq.(2.20) with $\mathrm{T}_l = 100\,\mathrm{ms}$) was used for the convolution and for the bottom subfigures, only the direct part of the RIR. Plotted are the MSC curves[7] of the reverberant signals for two different block sizes: $L = 320\,(20\,\mathrm{ms})$ and $L = 2 \cdot 21845\,(2.73\,\mathrm{s})$. From the upper plot, it can be seen that the expected theoretical MSC curve (in gray) (using Eq.(2.6) with $d_{\mathrm{mic}} = 0.17\,\mathrm{m}$) is only valid for small block sizes. Once the block size reaches a multiple of the impulse response length, a MSC of one occurs for all frequencies, as expected. Regarding the signal with direct components only, as shown in the bottom subfigures, the coherence is, at least for $L > 320\,(20\,\mathrm{ms})$, independent of the block size. This can also be explained since the time span of the direct path was set to the onset time plus 2 ms (see also Eq.(2.25)) which is smaller than the block size.

Due to the practical requirement of small block sizes for the considered speech enhancement algorithms and the calculation of the coherence using Eqs.(2.35),(2.36),(2.37), it is feasible to use input signals for the simulations which are generated by convolving a speech signal with a finite length RIR. Thus, the term coherence-based dereverberation always refers to an approach which is based on estimates of the coherence in the short-term DFT domain with $L \ll \mathrm{T}_r$ in the following. It should be noted that the performance results using such input signals for the considered speech dereverberation algorithms are verified by means of real recordings.

[7] The curves are obtained using the MATLAB command `mscohere` with 50% overlap and Hann window (`hanning`).

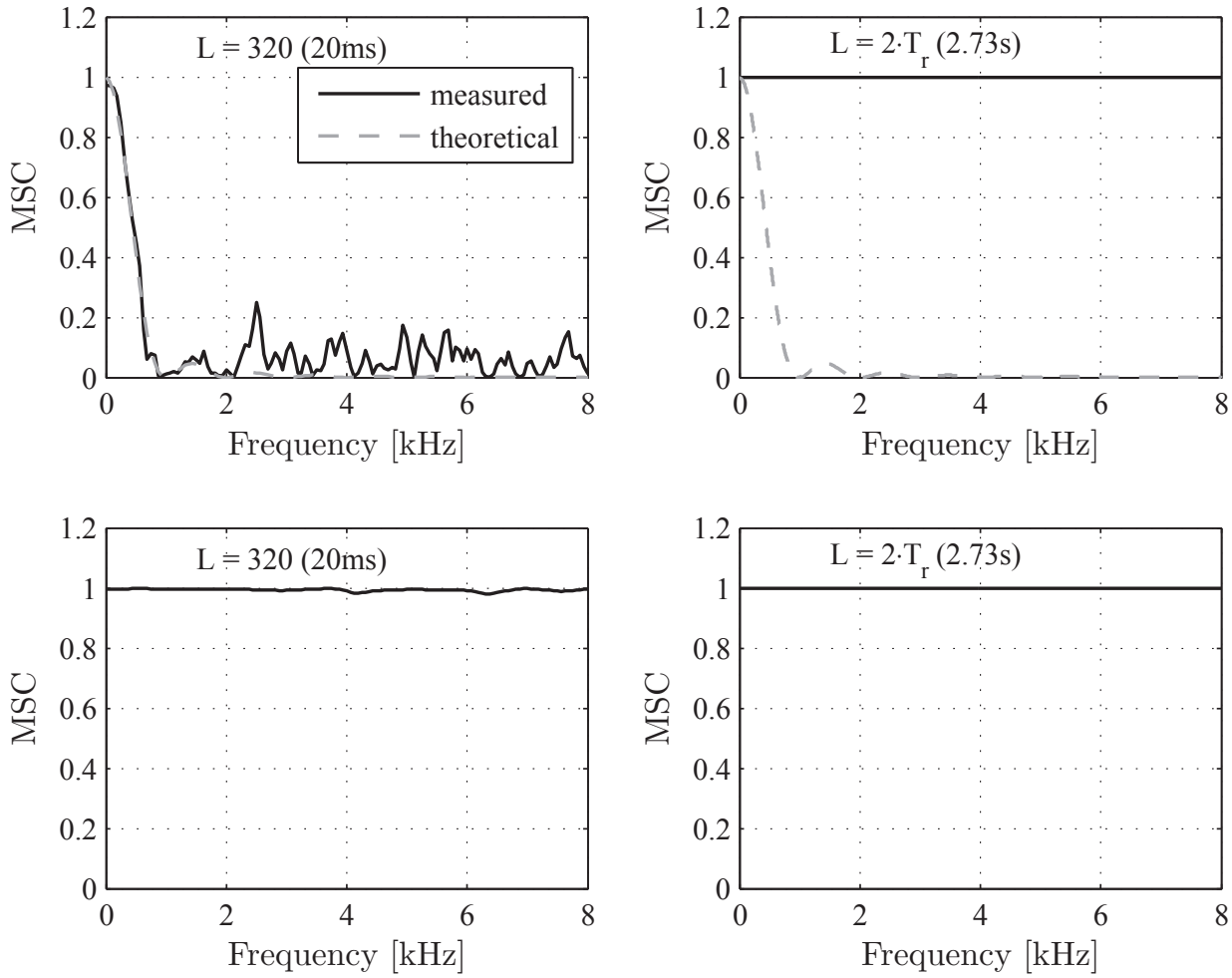


Figure 2.24: Influence of block size on short-term coherence estimate. A WGN signal is convolved with (top) the late reverberant part of a RIR and (bottom) the direct path of an RIR. The dashed gray line in the top plot represents the ideal diffuse noise model using Eq.(2.6) with $d_{\mathrm{mic}} = 0.15\,\mathrm{m}$. (Lecture room, length of RIR: $k_r = 21845$, $\mathrm{T}_r = 2.73\,\mathrm{s}$).

Table 2.4: Influence of different onset detection methods on DRR. The values are determined directly from two different room impulse responses.

	Lecture		Corridor	
Onset detection method	Conventional	Proposed	Conventional	Proposed
DRR [dB]	−1.3 dB	−3.1 dB	13.1 dB	12.9 dB

2.3.2 Estimation of the RIR Onset Time

An accurate detection of the onset time of a room impulse response has a significant influence on the estimated DRR value and hence, a very reliable estimation is essential. A general discussion on state-of-the-art algorithms to determine the onset time of a RIR is given in [DDP08, Ush10]. The authors claim that in case of a high DRR, energy based methods such as taking the maximum of the RIR give reliable estimates of the onset time. This approach is referred to the conventional approach. However, in situations when no line-of-sight between source and microphone exists, the estimation accuracy drops dramatically since the direct path peak in the RIR is not necessarily the peak with the highest energy anymore. This situation can occur, e.g., in binaural hearing when the source is located at one side of the head and, due to head shadowing, no direct path between the source and the opposed side exists. Hence, a novel method which allows for a reliable determination of the onset time is proposed:

First, the absolute square of the RIR $|h(k)|^2$ is computed and normalized. The onset time k_0 is obtained by the discrete time index when $|h(k)|^2$ is above a threshold ψ, e.g., $\psi = 0.1$. In case of very noisy impulse responses, the use of an adaptive threshold is beneficial which is, however, not required for the considered RIRs.

The advantage of the new procedure is illustrated in Fig. 2.25. The top plot shows the absolute square of the RIR. In the bottom subfigure, the rhombus marks the accurately detected onset of the RIR with the improved method while the conventional approach, indicated by the square, is only capable of detecting the maximum peak.

An example of the DRR values obtained directly from the RIR of two different rooms is given in Table 2.4. It can be seen that a significant difference by using the two onset detection methods exists and it is proposed to use the novel approach to determine the ground truth of the DRR.

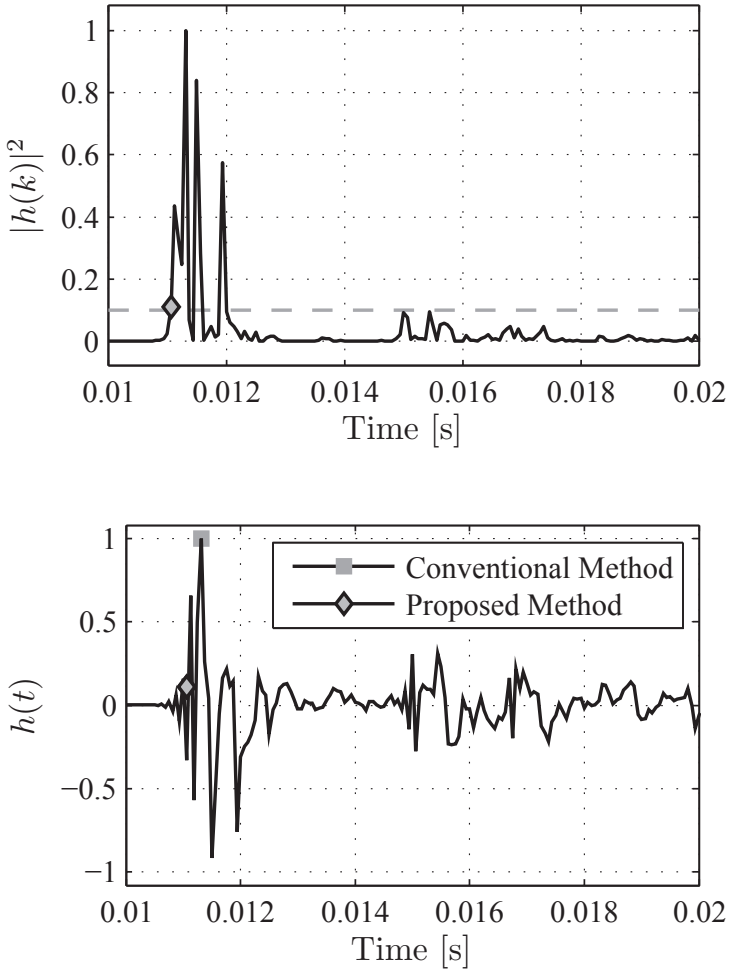


Figure 2.25: Illustration of proposed method for onset detection: (top) $|h(k)|^2$ of RIR over discrete time index given in seconds. The marker indicates the global maximum at the onset time k_0 and the horizontal dashed line marks the threshold $\psi = 0.1$, (bottom) RIR with the marked onset times using the proposed and the conventional approach.

2.3.3 Estimation of the Reverberation Time

The estimation of the RT or decay rate is important for the classification of an acoustic environment as well as an essential parameter for many dereverberation algorithms. In this thesis, the approach by [LV08a], which was later improved in [LYJV10], is used. In contrast to previous approaches for a blind RT estimation based on a *Maximum Likelihood* (ML) estimation [RJW+03, LV08a], this improved algorithm exhibits a significantly reduced computational complexity and is more suitable to track time-varying RTs. Such properties are of special importance for an application within hearing aids or mobile phones where only a very limited computational power is available. The main steps of the algorithm are outlined briefly in the following where a more detailed description is given in [LYJV10] and a comparison to alternative approaches in [GLJ+12].

The blind RT estimation is performed frame-wise on a single-microphone reverberant speech signal. In a first step, this signal is downsampled[8], e.g., by a factor of five to reduce the computational burden for estimating the RT. Afterwards, a pre-selection is performed to detect segments within the signal which possibly contain only sound decay. If such a possible sound decay is detected, the decay rate[9] ρ is estimated by ML estimation which is based a statistical model of the RIR (see Eq.(2.28)). The obtained value is used to update a histogram determined by the most recent ML estimates. The value associated with the maximum of this histogram is taken as an estimate for the RT. A recursive smoothing is finally applied to this RT value to reduce the variance of the estimate.

In order to detect changes of the RT more rapidly, a fast adaptation mechanism can be employed [LYJV10]. This comprises a second histogram with a lower number of ML estimates for the RT. If the maximum of this second histogram differs from that of the first histogram by more than 0.2 s for a certain period, the first histogram is replaced by the second histogram. For the sake of computational complexity, this fast adaptation is not employed in the considered speech enhancement systems.

An alternative promising approach which operates in the frequency domain is presented in [WHN08, Wen09]. The proposed method exploits the distribution of the energy envelope in each frequency bin. The estimated negative side variance of the decay rate distribution of the energy envelopes is mapped to the room decay rate. This mapping function has to be determined in advance in an off-line procedure. Due to the high computational complexity and drawback of the required off-line training, the abovementioned blind method [LYJV10] is beneficial for real-time applications and used throughout this thesis.

2.3.3.1 Frequency-Dependent RT Estimation

Since the considered time-domain approach [LYJV10] is only capable of estimating a frequency-independent RT, an extension to a frequency-dependent estimate is dis-

[8] Usually the downsampling operation involves a low-pass filter for alias compensation which is, however, not necessary in this special case.

[9] The relation between decay rate ρ and reverberation time is given by Eq.(2.21).

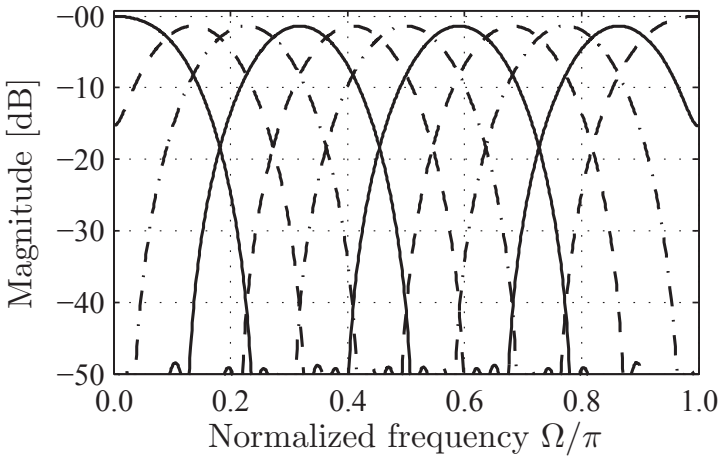


Figure 2.26: DCT analysis filterbank over normalized frequency with $M = 11$ subbands.

cussed in the following. This is of great interest since the RT exhibits a high dependency of the frequency, especially for larger rooms and rooms with highly reflective material. In the following paragraph, two models for the RT are proposed which allow to calculate the RT only in a few subbands. The basic principle is first validated for an off-line estimation where the RIR is given.

Assuming that the impulse response is available, it can be decomposed into several subbands and for each subband the RT can be calculated, e.g., by using the Schroeder method. In what follows, a uniform DCT filterbank with 11 subbands as described in [LV11] where the magnitude responses of the analysis filters are plotted in Fig. 2.26 is used. If a non-uniform decomposition is desired, it has been shown in [LV11] that a warped DCT filterbank is advantageous compared to the frequently used 1/3 Octave filterbank since the usage provides more reliable RT values at low frequencies.

An analysis of multiple impulse responses from various databases [JSV09, WGH+06, Kit10, KEA+09] has shown that a sufficient approximation of the frequency-dependent RT can be obtained if the RT is known in two or three subbands. For the first case, the RT model named *RT Model* 1 is given by the linear equation

$$\widehat{\mathrm{T}}_{60}(f) = \mathrm{T}_{60}(f_1) + \frac{\mathrm{T}_{60}(f_2) - \mathrm{T}_{60}(f_1)}{f_2 - f_1}\,(f - f_1) \tag{2.38}$$

where $\mathrm{T}_{60}(f_1)$ and $\mathrm{T}_{60}(f_2)$ are the reverberation times at subbands with center frequencies f_1 and f_2, respectively. A further improvement of the model in terms of accuracy for larger rooms can be obtained if the model is extended by a third subband RT. This model is referred to as *RT Model* 2 and given by

$$\widehat{\mathrm{T}}_{60}(f) = \begin{cases} \mathrm{T}_{60}(f_1) + \frac{\mathrm{T}_{60}(f_2)-\mathrm{T}_{60}(f_1)}{f_2-f_1}\,(f - f_2) & \text{for } \; 0 < f \leq f_2 \\ \\ \mathrm{T}_{60}(f_2) + \frac{\mathrm{T}_{60}(f_3)-\mathrm{T}_{60}(f_2)}{f_3-f_2}\,(f - f_3) & \text{for } \; f_2 < f \leq f_\mathrm{s}/2, \end{cases} \tag{2.39}$$

where $\mathrm{T}_{60}(f_3)$ denotes the RT in a third subband. The center frequencies are listed in Table 2.5 and an illustration of the two models is given in Fig. 2.27. In Subfigures (a)

Table 2.5: Center frequencies proposed for the frequency approximation models at $f_s = 16\,\text{kHz}$.

	f_1	f_2	f_3
RT Model 1	0.8 kHz	7.2 kHz	-
RT Model 2	0.8 kHz	3.2 kHz	7.2 kHz

and (c), the RT over frequency is plotted for the lecture and corridor RIR. The red solid line marks the frequency-independent RT estimate obtained by applying the Schroeder method to the full-band RIR (time-domain RT estimate). The solid line with the circle markers gives the RT when the full (11 subband) DCT filterbank is applied and the dashed-dotted and dashed lines give the approximations with the proposed RT Models 1 and 2. It can be seen that for the corridor location, the RT Model 1 gives a sufficient accuracy while in case of the lecture room, Model 2 greatly increases the accuracy. On the right side in (b) and (d), the estimation errors ΔRT between the full-band RT and the approximations with frequency-independent RT as well as RT Model 1 and 2 are shown. From this we can conclude that the proposed RT models are appropriate for a sufficient approximation of the (true) full-band RT and hence, high saving in terms of complexity can be achieved.

2.3.3.2 Performance Evaluation

The performance of the aforementioned blind RT estimation procedure [LYJV10] is evaluated in this paragraph using single-channel RIRs from the AIR database. First, frequency-independent and second frequency-dependent in combination with the two RT models. For the experiments, a speech signal with a duration of 3 min. was used to show the convergence behavior as well as the tracking performance over time. Here, only the estimation accuracy is considered. A possible influence on the performance of a speech enhancement algorithm which requires a reliable RT estimate is discussed in Sec. 3.1.4. The influence of additional background noise on the estimation accuracy is discussed in Sec. 3.3.1.

Figure 2.28 (top) shows the 'true' RT determined by applying the Schroeder method to the RIR and the estimated RT is calculated frame-wise (20 ms) from a reverberant speech signal with a duration of 3 min. The curves are plotted over the source-microphone distance d_{LM} and the results are averaged over all frames. The first 30 s are not taken into account in order to compare the steady state performance. It can be seen that especially for larger distances, the estimated RT matches greatly the true RT obtained directly from the given RIR. A deeper analysis of the estimation error is given in the lower plots. The left plot in Fig. 2.28 (bottom) represent the estimation error over the source-microphone distance and the right plot over the DRR. It is obvious that the performance of the RT estimator drops for sources within the critical distance. This can be explained by the fact that the statistical RIR model by Polack, which is the basis of the RT estimator, is valid only for sources outside the critical distance. Hence, if the model is violated, a high underestimation of the RT occurs. Therefore, a possible modification on the RT estimator should consider the

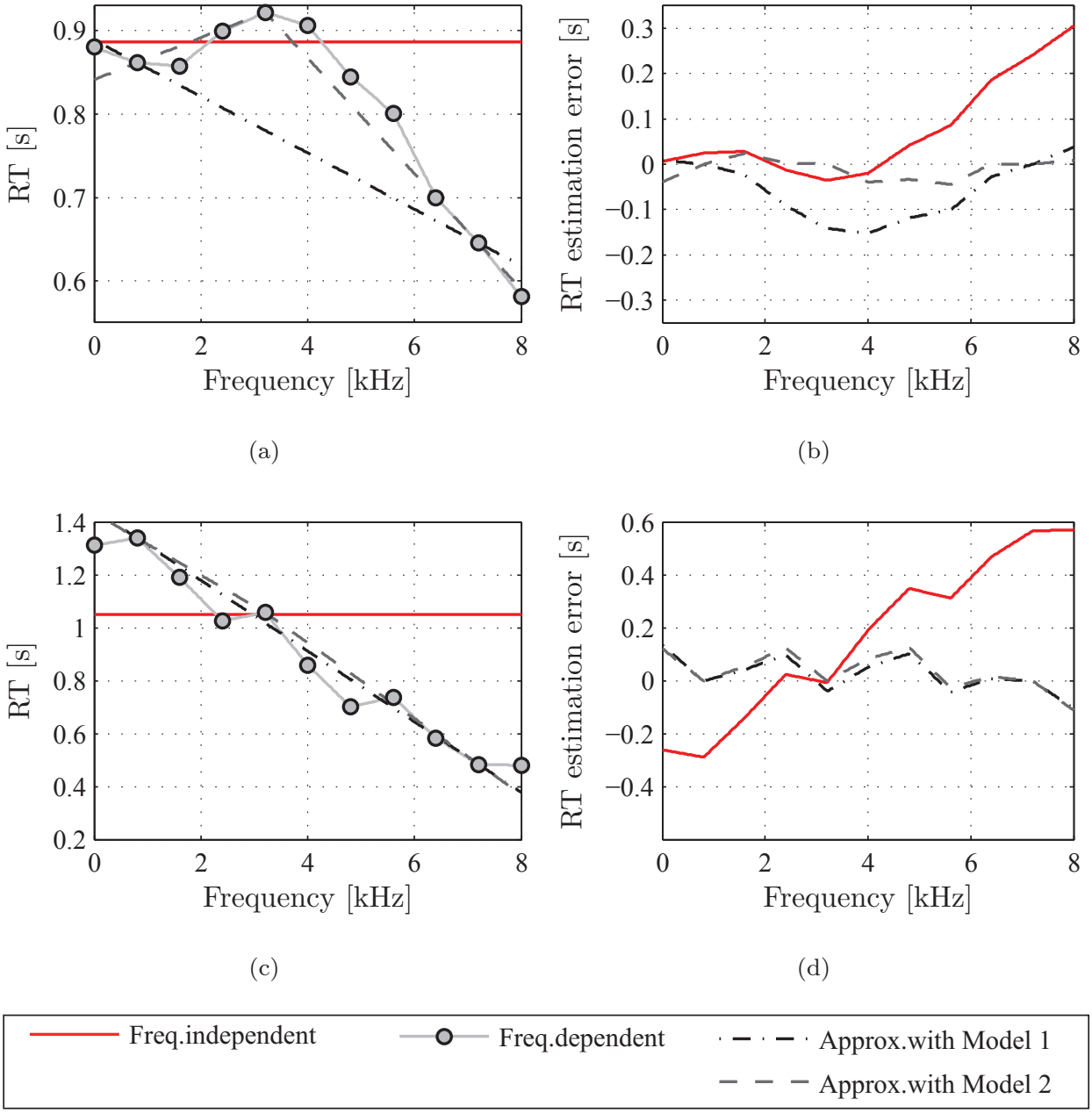


Figure 2.27: Frequency-dependent RT and estimation error for the two proposed RT models:
(a),(b) lecture room ($d_{\mathrm{LM}} = 10.2\,\mathrm{m}$, $\mathrm{T}_{60} = 0.87\,\mathrm{s}$, $\mathrm{DRR} = -3.83\,\mathrm{dB}$);
(c),(d) corridor (HHP, $d_{\mathrm{LM}} = 0.1\,\mathrm{m}$, $\mathrm{T}_{60} = 1.21\,\mathrm{s}$, $\mathrm{DRR} = 12.71\,\mathrm{dB}$).

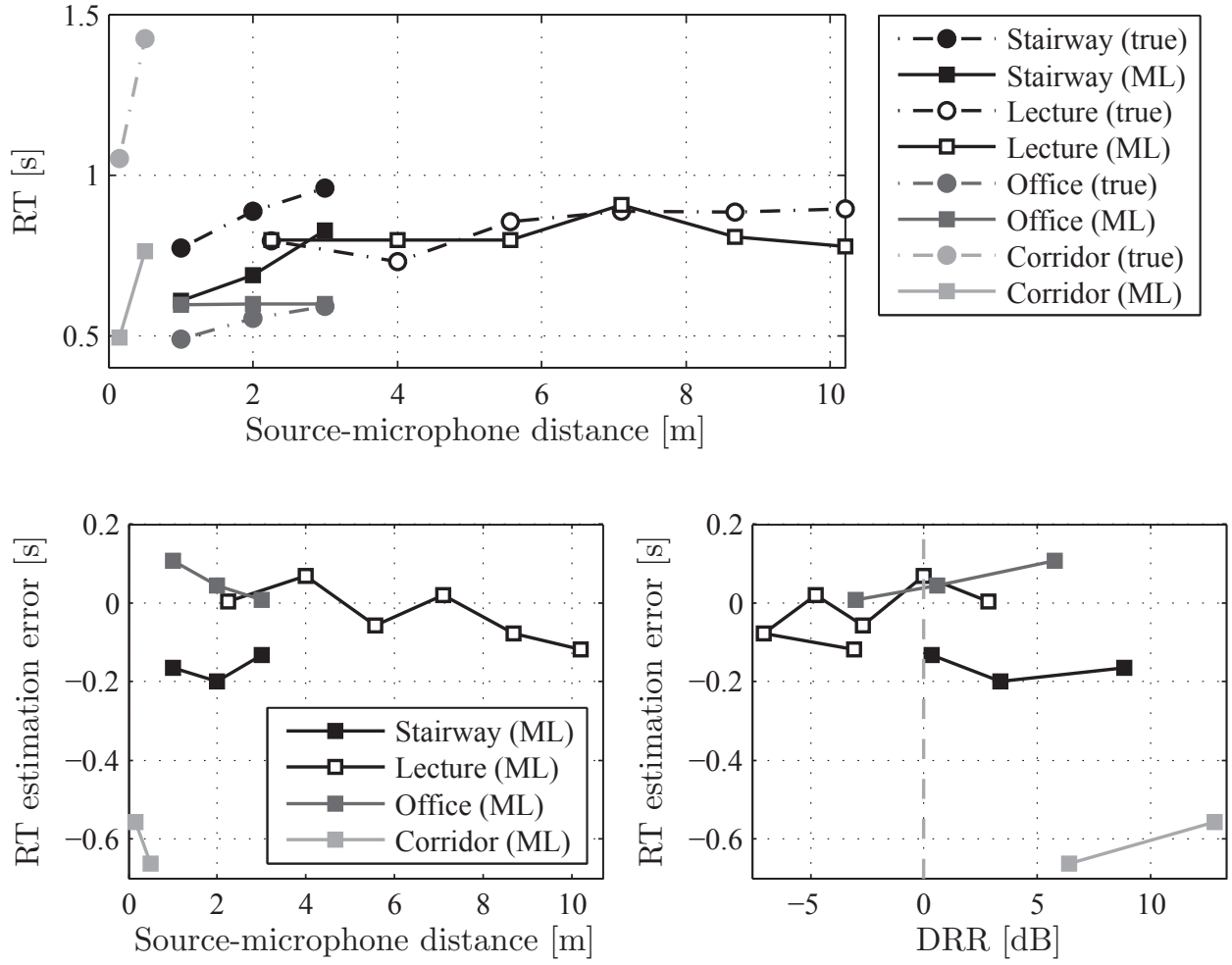


Figure 2.28: Blind RT estimation performance: (top) plotted over the source-microphone distance d_{LM}. (bottom) estimation error (left) over the source-microphone distance, (right) over the DRR where the vertical dashed line marks the critical distance.

utilization of a generalized statistical RIR models and/or to take a priori knowledge of the DRR into account. Nevertheless, a high underestimation is more favorable since a dereverberation algorithm which relies on this estimate would reduce less reverberation. A large overestimation causes a high amount of speech distortion as shown later.

In a further experiment, the tracking speed of the estimator in a changing acoustic environment is considered. Such situation can occur, e.g., when a person wearing a hearing aid is walking from one room into another. For the following experiment, the first 60 s of the anechoic speech signal is convolved with the RIR of a lecture room ($d_{\mathrm{LM}} = 4\,\mathrm{m}$, $\mathrm{T}_{60} = 0.79\,\mathrm{s}$, $\mathrm{DRR} = 1.75\,\mathrm{dB}$) and the second 60 s with the RIR of an office ($d_{\mathrm{LM}} = 2\,\mathrm{m}$, $\mathrm{T}_{60} = 0.53\,\mathrm{s}$, $\mathrm{DRR} = 2.54\,\mathrm{dB}$). For the averaging over frames, the full transition period is included. The outcome is depicted in Fig. 2.29. The algorithm requires approx. $18 - 20\,\mathrm{s}$ to adapt to the acoustic situation from the initial state. After the abrupt RT change after 60 s, the tracking takes at least $35 - 40\,\mathrm{s}$ to adopt to the new RT.

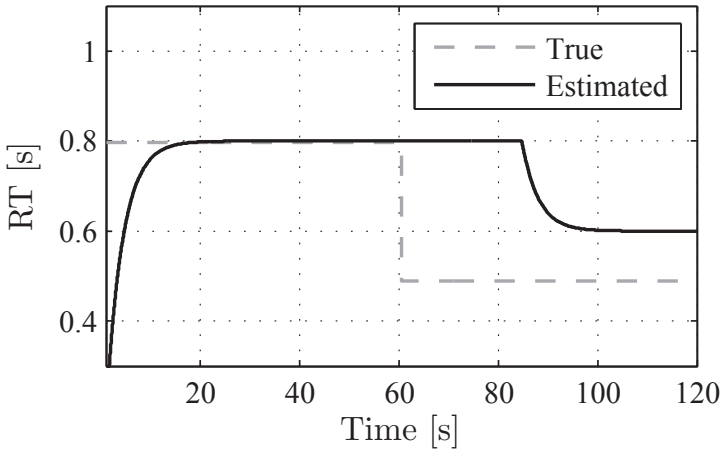


Figure 2.29: Tracking speed of blind RT estimation algorithm using two different RIRs: lecture room with $T_{60} = 0.79\,s$ and office with $T_{60} = 0.53\,s$.

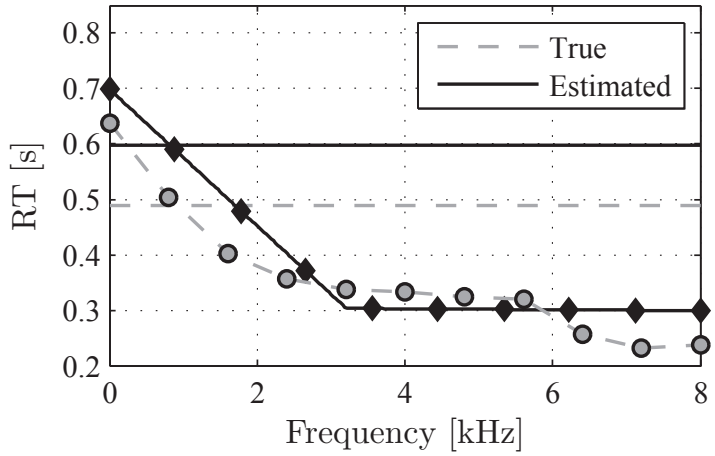


Figure 2.30: Frequency-dependent RT estimation (with markers) where the true and estimated frequency-independent RT is also plotted by the horizontal lines (w/o markers).

Finally, the last experiment evaluates the frequency-dependent blind RT estimation using the discussed frequency approximation models given by Eqs.(2.38),(2.39). Figure 2.30 shows by the dashed lines the true RT and the solid line represent the estimated RT using the RT Model 2 using Eq.(2.39) and the RIR of an office room. Additionally, the frequency-independent RTs are given by the horizontal lines. From this experiment, which was also repeated with other RIRs, we can conclude that the proposed simplification with an estimation only in three DCT subbands gives a sufficient approximation of the true frequency-dependent RT even when applied to a blind estimation procedure.

2.3.4 Estimation of the Direct-to-Reverberation Energy Ratio

When analyzing the acoustic environment, the DRR measures always refer to the acoustic channel and are calculated directly from a given impulse response. Since the DRR is also a very important parameter for some speech dereverberation algorithms, a blind estimation directly from the observed speech signal is proposed [JNBV11].

For the short-term DRR estimation, the generalized coherence model of Eq.(2.18) is considered. The included CDR can also be seen as the ratio between coherent and non-coherent noise or as the ratio between direct speech and reverberant speech. Thus, having an estimate of the coherence for a given input signal available leads easily to the desired DRR.

In order to estimate the DRR from a given coherence function, Eq.(2.18) can be rearranged to

$$\Psi(e^{j\Omega}) = \frac{\operatorname{sinc}(\Omega f_s d_{\mathrm{mic}}/c) - \Gamma_{x_1x_2}^{(\mathrm{mix})}(e^{j\Omega})}{\Gamma_{x_1x_2}^{(\mathrm{mix})}(e^{j\Omega}) - 1}, \tag{2.40}$$

where $\Gamma_{x_1x_2}^{(\mathrm{mix})} - 1 > 0$ has to be ensured for the denominator, e.g., by means of an upper threshold of the coherence with, e.g., $\Gamma_{\max} = 0.99$. Please note that Eq.(2.18) and thus, Eq.(2.40) assume that a possible time-delay of the speech signal among the microphones has been compensated.

To obtain a short-term estimate of the CDR $\widehat{\Psi}(\lambda, \mu)$, the generalized coherence function $\Gamma_{x_1x_2}^{(\mathrm{mix})}(e^{j\Omega})$ in Eq.(2.40) is replaced by the corresponding short-term estimate $\widehat{\Gamma}_{x_1x_2}(\lambda, \mu)$ using the described recursive smoothing procedure to determine the auto- and cross PSDs. Additionally, the obtained coherence estimate is smoothed over time using the constant $\alpha^{(\mathrm{coh})}$.

In order to prevent the algorithm from underestimations during speech pauses and to avoid high fluctuations of the estimate, an adaptive smoothing procedure using a novel dual-channel VAD is proposed. The VAD has a very low computational complexity compared to frequently used single-channel algorithms (see, e.g., [VM06, JKAO09]) and can be used for other applications as well. First, the mean short-term MSC, averaged over all frequency bins μ, is calculated independently for each frame without recursive smoothing of neither the MSC nor the required auto- and cross-PSD terms as

$$\overline{C}_{x_1x_2}(\lambda) = \frac{1}{M} \sum_{\mu=1}^{M} \widehat{C}_{x_1x_2}(\lambda, \mu). \tag{2.41}$$

Based on $\overline{C}_{x_1x_2}(\lambda)$, each frame is classified as either speech active or speech inactive. Depending on this classification, a smoothing constant is determined as

$$\alpha^{(\mathrm{DRR})} = \begin{cases} \alpha^{(\mathrm{inactive})} & \text{for } \overline{C}_{x_1x_2}(\lambda) < \tau^{(\mathrm{VAD})} \\ \alpha^{(\mathrm{active})} & \text{otherwise,} \end{cases} \tag{2.42}$$

with VAD constant $\tau^{(\mathrm{VAD})}$. Second, the preliminary estimate $\widehat{\Psi}(\lambda,\mu)$ is used to calculate the frequency-dependent short-term DRR by recursive smoothing as

$$\widehat{\mathrm{DRR}}(\lambda,\mu) = \alpha^{(\mathrm{DRR})} \cdot \widehat{\mathrm{DRR}}(\lambda-1,\mu) + (1-\alpha^{(\mathrm{DRR})})\,\widehat{\Psi}(\lambda,\mu) \tag{2.43}$$

and is given in the log-domain by

$$\widehat{\mathrm{DRR}}(\lambda,\mu) = 10 \cdot \log_{10}(\widehat{\mathrm{DRR}}(\lambda,\mu)). \tag{2.44}$$

The estimates are limited by a lower and upper threshold by

$$\widehat{\mathrm{DRR}}(\lambda,\mu) = \min\left\{\max\left\{\widehat{\mathrm{DRR}}(\lambda,\mu), \mathrm{DRR}^{(\mathrm{min})}\right\}, \mathrm{DRR}^{(\mathrm{max})}\right\}. \tag{2.45}$$

A frequency-independent DRR estimate can be obtained by an additional averaging over all $M-\mu_0$ frequency bins by

$$\overline{\mathrm{DRR}}(\lambda) = 10 \cdot \log_{10}\left(\frac{1}{\mathrm{M}-\mu_0}\sum_{\mu=\mu_0}^{\mathrm{M}} \widehat{\mathrm{DRR}}(\lambda,\mu)\right), \tag{2.46}$$

where μ_0 corresponds to the first root of the sinc-function (see Sec. 2.1.2.2 and Fig. 2.4). In the case of closely-spaced microphones, an averaging over the entire frequency range $\mu = 1, 2, .., M$ is performed. A MATLAB reference implementation is available online.[10]

Alternative multi-channel approaches are presented, e.g., in [HNS+10a, HNS+10b, HNS+11], where the power spectra of both the direct sound and reverberation are estimated from the spatial correlation matrix of the observed signal. In [LC08, LC10], a dual-channel algorithm for the binaural distance perception is developed which contains a DRR estimation unit. This comprises a 32-channel Gammatone filterbank and an equalization-cancellation operation which aims to decompose the signal into its direct and reverberant components. The authors in [BW08] present a single-channel dereverberation algorithm where the required DRR is estimated during speech pauses.

The following section gives a performance evaluation of these alternatives and proofs the superiority of the proposed solution for estimating the DRR value.

2.3.4.1 Performance Evaluation

In the following experiments, the performance of the proposed algorithm (Proposed) in comparison with [LC10] (Lu) and [HNS+11] (Hioka) is given. The input signals are generated by convolving a speech signal of 3 min. duration with three different binaural room impulse responses. The dual-channel input is then used to estimate the DRR with the proposed and reference methods (Lu/Hioka). Relevant simulation parameters are listed in Table 2.6. The case of additional background noise is treated later in Sec. 3.3.1.

[10] Download link: `http://www.ind.rwth-aachen.de/~bib/jeub11d.`

Table 2.6: Main simulation parameters of the novel blind DRR estimation method.

Parameter	Setting
Thresholds	$\mathrm{DRR}^{(\mathrm{min})} = -20\,\mathrm{dB}$, $\mathrm{DRR}^{(\mathrm{max})} = 20\,\mathrm{dB}$
Smoothing factors	$\alpha^{(\mathrm{PSD})} = 0.85$, $\alpha^{(\mathrm{coh})} = 0.5$, $\alpha^{(\mathrm{active})} = 0.98$, $\alpha^{(\mathrm{inactive})} = 1$
VAD threshold	$\tau^{(\mathrm{VAD})} = 0.95$

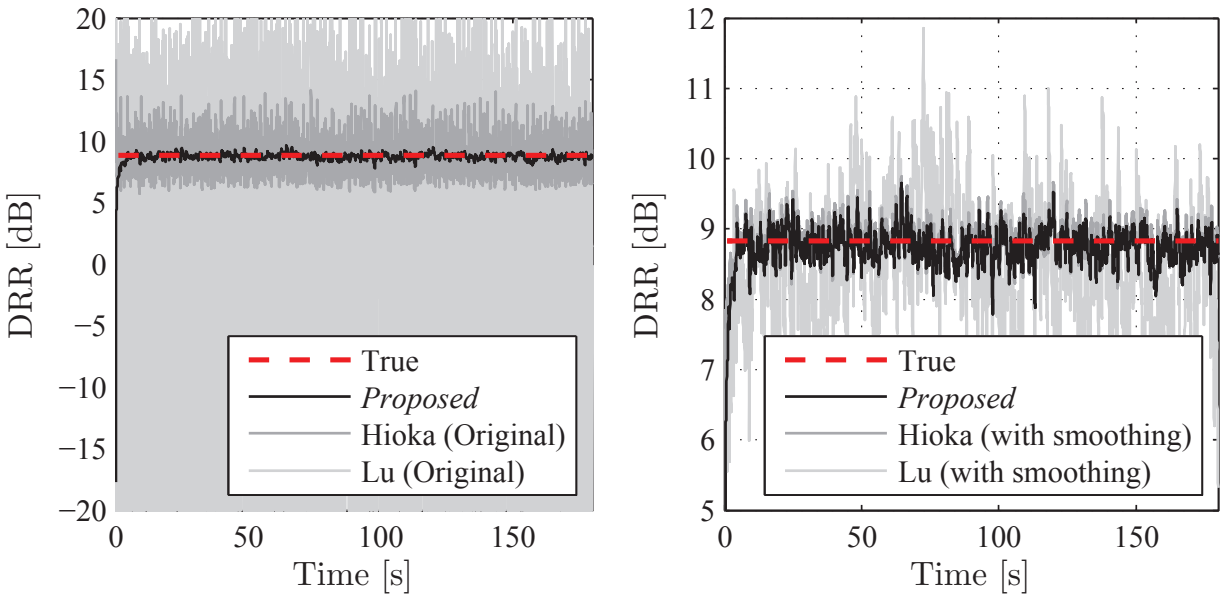


Figure 2.31: DRR estimation performance over time: (left) using the original Lu and Hioka implementations, (right) applying additionally the proposed smoothing procedure to all algorithms.

The DRR estimation results for one reverberant signal over time are depicted in Fig. 2.31 (left) where the true DRR is marked by the red dashed line. The proposed algorithm is given in black and the two reference methods in gray colors. It can be seen that using the implementation as given in the corresponding publications, high fluctuations occur for the two reference methods. In this example, the proposed algorithm already utilizes the presented adaptive smoothing procedure and thus, gives very stable results. In a next step, the proposed smoothing and averaging procedure is integrated into the two reference methods and the results are shown in Fig. 2.31 (right). The fluctuation could be reduced significantly and much better results are obtained by employing the smoothing procedure. A more detailed analysis of the right subfigure has shown that the Hioka algorithm results in more fluctuations of the DRR estimate compared to the proposed method. It can further be seen that even with the smoothing procedure, a convergence time of less than 5 s occurs for all algorithms.

The overall estimation accuracy, averaged over time, is evaluated and shown in Fig. 2.32 for lecture room and corridor as well as the Aula Carolina. Each marker shape represents an algorithm and the colors correspond to the different rooms. The

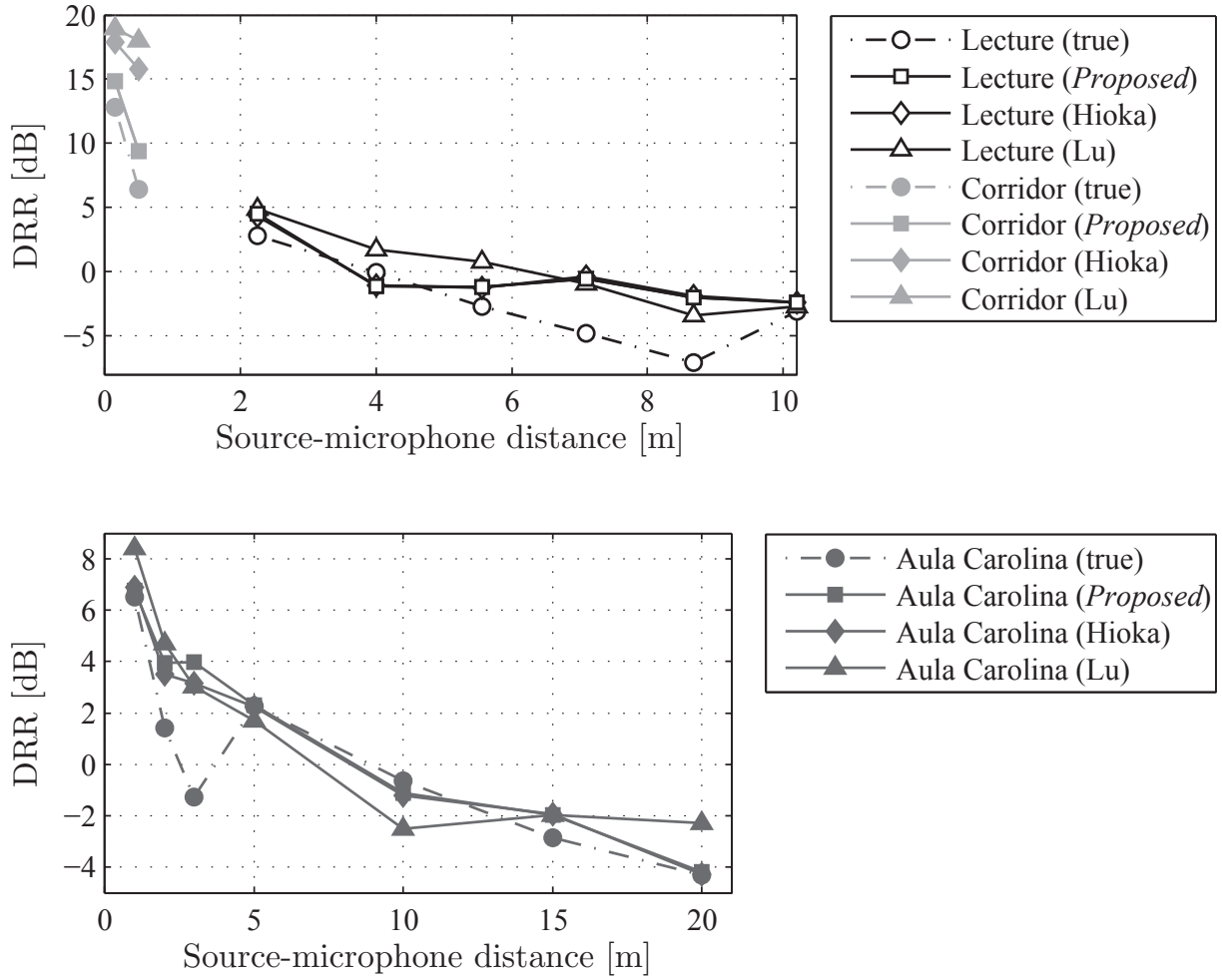


Figure 2.32: DRR estimation performance over the source-microphone distance: (top) lecture and office, (bottom) Aula Carolina.

results of the proposed and Hioka algorithm are quite similar which can be explained with the similar underlying principle of exploiting the coherence of speech and noise. The Lu algorithm also shows a good tendency but the overall estimation error is larger compared to the other two methods. It has to be mentioned that especially the estimation accuracy for sources which are located within the critical distance have to be very accurate.

In terms of computational complexity, the proposed method requires the lowest computational effort and possesses a low memory consumption. This has been verified using the normalized Matlab processing time as a rough indicator as given in Table 2.7. A matrix inversion in the Hioka approach causes a high computational load as well as a large amount of memory. The Lu method decomposes the signal into 32 subbands using a Gammatone filterbank which can be realized more efficiently than with the employed Matlab functions.

As for the RT estimator, the tracking in varying acoustic conditions is also of great interest. The estimation performance with a changing RIR from a high (Stairway with DRR = 9.5 dB) to a low DRR (Lecture with DRR = −3.8 dB) is presented in

Table 2.7: Normalized processing time of three DRR estimation algorithms using three minutes of speech material.

Algorithm	*Proposed*	Hioka	Lu
Processing time	1.0	5.9	3.5

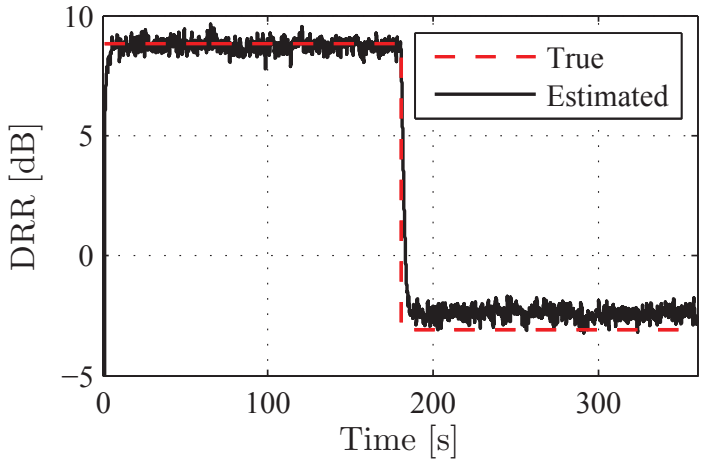


Figure 2.33: DRR estimation tracking performance of the proposed method under changing acoustic conditions using two different RIRs: stairway with DRR = 9.5 dB and lecture room with DRR = −3.8 dB.

Fig. 2.33. We can conclude that the proposed method is capable of converging very fast in changing DRR conditions at a very low convergence time.

2.3.5 Estimation of Binaural Cues

The human auditory system has a very sophisticated mechanism to analyze the spatial impression of an acoustic environment by exploiting the *binaural cues* [Bla96]. This comprises the ability for distance and direction estimation. Numerous experiments have shown that the localization in the azimuth plane is mostly based on the interaural time and level differences of the sound event. The localization in terms of elevation is carried out with the help of the spectral coloring of the input signals due to the shape of the outer ear. The distance perception is based on the direct-to-reverberation energy ratio, cf. [BH99]. Since they are the most important binaural cues for source localization in the azimuth plane, the main focus of this thesis will be on the *Interaural Time Difference* (ITD) and *Interaural Level Difference* (ILD). The estimation of these cues is important for both the development and the evaluation of binaural algorithms. In reverberant sound fields, the reliability of these estimates is not always guaranteed. The binaural cues are degraded not only by taking values different from those of the non-reverberant signals, but also by having a larger variance, which makes the localization of the source ambiguous. Hence, an improved system with an adaptive threshold will be derived which significantly decreases the variance of the estimation. The ground truth of the binaural cues is determined as the ones that

can be estimated from the reverberant signals by focusing only on the time-frequency portions that are interaurally coherent.

The interaural time difference is defined as the time delay of arrival between the left and right ear. Assuming a simple model of the head as a spherical torso and a source in the far field, the ITD can be expressed according to Fig. 2.1 (b) by [Har99]

$$\Delta t = \frac{3r}{c} \sin \theta, \tag{2.47}$$

where r is the radius of the head and c the speed of sound. For an approximate radius of $r = 8.5\,\text{cm}$, the ITD lies in the range $-750\,\mu\text{s} \leq \Delta t \leq +750\,\mu\text{s}$. These maximum and minimum values are also of interest for the DOA estimator which is incorporated in the proposed speech enhancement system (see Sec. 4.1.3). The interaural level difference in dB is given by the level differences of the signals arriving at the left and right ear:

$$\Delta E = 10 \cdot \log_{10} \left(\frac{E_l}{E_r} \right), \tag{2.48}$$

with $E_{l|r}$ being the energy of the right and left signal $x_{l|r}(k)$, respectively. As a simple rule-of-thumb, the ITD is relevant for frequencies below and the ILD for frequencies above 1.5 kHz, cf. [Bla96, Har99].

For situations with only one dominant source, a straightforward method to estimate the ITD is to calculate the cross-correlation and to measure the time lag of the maximum [Jef48]; an overview on such time delay estimation techniques can be found in [CBH06]. The ILD can simply be calculated by the energy ratio as in Eq.(2.48). However, for multiple sources or reverberant environments, both measures become unreliable, cf. [FM04].

A promising procedure to improve the estimation robustness for both ITD and ILD has been published in [FM04], where only cues are selected where the *Interaural Coherence* (IC) is above a certain threshold. By this procedure, the algorithm tries to estimate only the cues of the direct path (which correspond to the clean speech or free-field cues). This can be seen as a replication of the precedence effect in the human auditory system which mainly relies on the binaural cues of the first wave front for azimuth localization, cf. [Bla96]. An improved frame-wise estimation procedure for ILD cues will be described shortly in the following [JSEV10]. The extension to ITD estimation is straightforward, the only change is the replacement of the frame-wise ILD by a frame-wise ITD.

The input signals of both channels are first divided into frames of 20 ms (320 samples at a sampling rate of 16 kHz) with an overlap of 319 samples to allow for a detailed and precise analysis. These frames are then decomposed into 24 critical bands using a Gammatone cochlear filterbank [Sla98, PAG95]. The center frequencies are chosen according to the Glasberg and Moore model [GM90].

For each band with subband index μ' ($\mu' = 1, 2, ..., 24$) and corresponding center frequency f_c, the estimation is performed by means of recursive averaging. The

signals for each frame λ and subband μ' are denoted by $x_l(\lambda, \mu', k)$ and $x_r(\lambda, \mu', k)$. The subband ILD is calculated as

$$\Delta E(\lambda, \mu') = 10 \cdot \log_{10} \left(\frac{E_l(\lambda, \mu')}{E_r(\lambda, \mu')} \right). \tag{2.49}$$

The energies of the left and right channel are calculated by recursive averages independent for each subband

$$E_l(\lambda, \mu') = \alpha^{(\mathrm{ILD})} \cdot \sum_{k=1}^{K} x_l^2(\lambda, \mu', k) + (1 - \alpha^{(\mathrm{ILD})}) \cdot E_l(\lambda - 1, \mu'), \tag{2.50a}$$

$$E_r(\lambda, \mu') = \alpha^{(\mathrm{ILD})} \cdot \sum_{k=1}^{K} x_r^2(\lambda, \mu', k) + (1 - \alpha^{(\mathrm{ILD})}) \cdot E_r(\lambda - 1, \mu'), \tag{2.50b}$$

with K being the number of samples in each frame of 20 ms duration. The smoothing factor $\alpha^{(\mathrm{ILD})}$ is determined from the time constant $T = 10\,\mathrm{ms}$ and sampling frequency f_s in Hz as in [FM04]

$$\alpha^{(\mathrm{ILD})} = \frac{1}{T \cdot f_s}. \tag{2.51}$$

Since the per-frame ILD estimate $\Delta E(\lambda, \mu')$ gives unreliable results, especially in reverberant environments, the variance of the estimate should be decreased. One very attractive possibility is to select only cues with an interaural coherence above a certain threshold for further evaluation. The IC is estimated by the normalized cross-correlation given by

$$\gamma(\lambda, \mu') = \frac{E_{lr}(\lambda, \mu')}{\sqrt{E_l(\lambda, \mu') \cdot E_r(\lambda, \mu')}}, \tag{2.52}$$

where $E_{lr}(\lambda, \mu')$ is calculated by

$$E_{lr}(\lambda, \mu') = \alpha^{(\mathrm{ILD})} \cdot \sum_{k=1}^{K} (x_l(\lambda, \mu', k) \cdot x_r(\lambda, \mu', k)) + (1 - \alpha^{(\mathrm{ILD})}) \cdot E_{lr}(\lambda - 1, \mu'). \tag{2.53}$$

In the following, only cues with an IC above the threshold $\gamma_{\mathrm{thr}}(\mu')$ are used:

$$\Delta E_{\mathrm{sel}}(\lambda, \mu') = \{\Delta E(\lambda, \mu') | \gamma(\lambda, \mu') > \gamma_{\mathrm{thr}}(\mu')\}. \tag{2.54}$$

The choice of $\gamma_{\mathrm{thr}}(\mu')$ has a strong influence on the estimation: If $\gamma_{\mathrm{thr}}(\mu')$ is chosen too low, the cue selection process will be rendered inefficient as no significant reduction in variance can be achieved. On the other hand, if $\gamma_{\mathrm{thr}}(\mu')$ is chosen too high, the reliability of the selection will be decreased as just very few signal frames will be considered for the determination of $\Delta E_{\mathrm{sel}}(\lambda, \mu')$. In terms of reverberant signals, the necessity for different thresholds per frequency band is motivated by the high frequency-dependency of the reverberation tail, cf. [JSV09]. For the sake of brevity, the index μ' for the threshold is omitted in the following. In [FM04], a fixed threshold was given depending on the center frequency f_c of the frequency band:

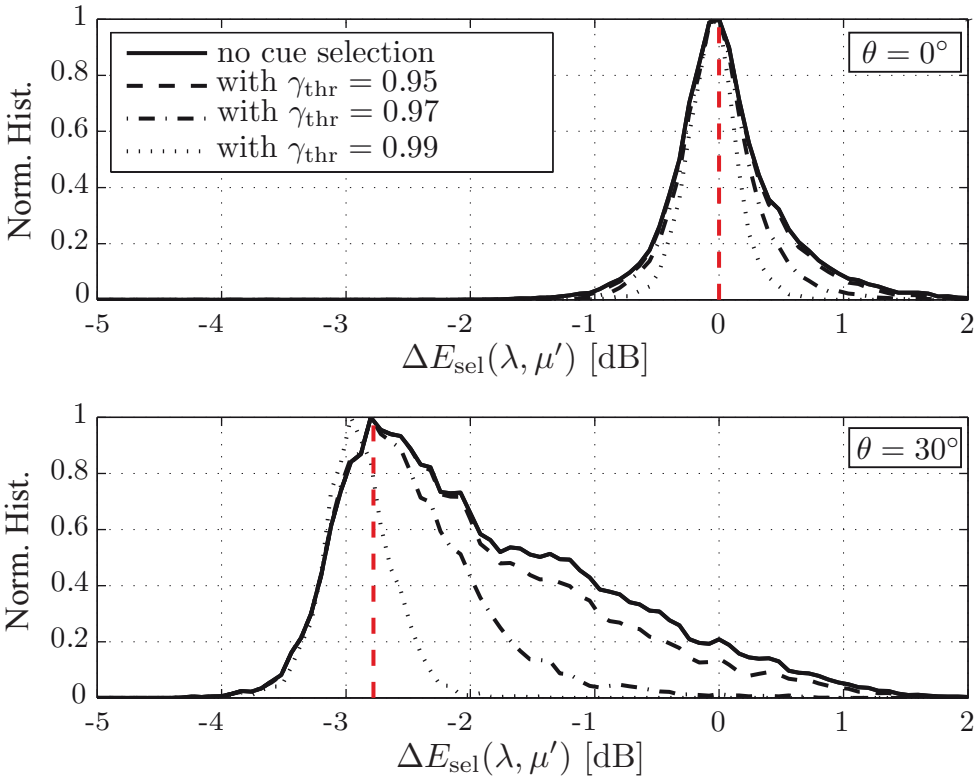


Figure 2.34: ILD estimation: normalized histograms of different cue selection procedures for a speech source from two different azimuth angles (without and with cue selection and three different thresholds γ_{thr} at center frequency $f_c = 2584\,\text{Hz}$). The signals are generated using BRIRs of a stairway hall having an average reverberation time of $T_{60} = 0.82\,\text{s}$. The vertical red dashed lines mark the correct values for this frequency band ($\mu' = 17$) in terms of the anechoic cues.

- $f_c = 500\,\text{Hz} \Rightarrow \gamma_{\text{thr}} = 0.95$,
- $f_c = 2000\,\text{Hz} \Rightarrow \gamma_{\text{thr}} = 0.99$.

However, the optimum threshold is not only depending on the center frequency but also on the azimuth angle of the source signal. In Fig. 2.34, the normalized histograms of the cue selection according to Eq.(2.54) are depicted for a center frequency of $f_c = 2584\,\text{Hz}$ (corresponds to subband $\mu' = 17$) and for the two azimuth angles $\theta = 0°$ and $\theta = 30°$. The reverberant speech is generated using eight speech files from the NTT database [NC94] convolved with binaural room impulse responses measured with a dummy head in a stairway hall. The vertical red lines mark the correct values for this frequency band in terms of the anechoic cues. These have been derived by convolving the same source signal with the (manually segmented) direct path of the corresponding BRIR.

For an azimuth angle of $\theta = 0°$, the optimum result, calculated from the direct speech signals, for the ILD estimation would be 0 dB while it would be $-2.78\,\text{dB}$ for $\theta = 30°$. The estimated ILD is shown for four different cue selection conditions:

- without any selection of cues ($\gamma_{\text{thr}} = 0$),

- with $\gamma_{\text{thr}} = 0.95$,
- with $\gamma_{\text{thr}} = 0.97$,
- with $\gamma_{\text{thr}} = 0.99$.

It can be seen that for an angle of $\theta = 0°$, the results do not differ significantly between the cue selection strategies. Since $\gamma_{\text{thr}} = 0.99$ leads to the smallest variance in the estimation, it would be the threshold of choice for this case. However, the situation changes for an angle of $\theta = 30°$ where the variance without any selection procedure is quite large and even a threshold of $\gamma_{\text{thr}} = 0.95$ does not lead to a substantial decrease in variance. A threshold of $\gamma_{\text{thr}} = 0.99$ on the other hand leads to another issue: the mean of the histogram deviates from the correct value. The reason for this direction-dependent behavior lies in the strong variation of the DRR for different azimuth angles as shown later. A decrease in DRR leads to a stronger impact of diffuse components on the estimation of ILD and ITD.

To allow for a threshold that is better suited for all frequency bands and azimuth angles, a novel adaptive procedure that makes use of the signal statistics for the determination of $\gamma_{\text{thr}}(\mu')$ from $\gamma(\lambda, \mu')$ is proposed [JSEV10]. The threshold is calculated as the 90th percentile of all individual values $\gamma(\lambda, \mu')$. This procedure guarantees that the ILD is always estimated from the most reliable values (i.e., the 10% of all individual values with the highest IC) for $\Delta E_{\text{sel}}(\lambda)$ while ensuring that single outliers never get too much weight in the calculation. For the cases that are depicted in Fig. 2.34, this procedure leads to a threshold $\gamma_{\text{thr}} = 0.99$ for $\theta = 0°$ and $\gamma_{\text{thr}} = 0.98$ for $\theta = 30°$. Similar properties can be observed for the other subbands μ'.

The improved estimator ensures a significant reduction in variance for the ILD estimate leading to more reliable results in accordance with the precedence effect in the human auditory system. Hence, it will later be used in Sec. 4.1.2 to investigate the influence of a bilateral dereverberation on the binaural cues. The concept has also been found beneficial in [CMW11b, Cor11] and is used there to determine the true ILD for experiments on the binaural cue changes due to noise reduction.

2.3.6 Noise Field Classification

Acoustic environment classification is an important component of modern hearing aids [HCE+05, Wit01] and various types of speech communication systems such as mobile phones. Depending on the acoustic environment, the device should automatically adjust the operating mode and control all integrated algorithms. When it comes to multi-microphone speech enhancement algorithms, e.g., for noise reduction or dereverberation, it is of significant interest whether the noise field can be characterized as either coherent or diffuse (non-coherent) or as a mixture. This is important since many algorithms rely, e.g., on a diffuse noise field assumption and hence, do not show sufficient enhancement performance under other noise conditions and can lastly degrade the desired signal.

In the following, a new approach for noise field classification is presented which allows to estimate the short-term ratio between coherent and diffuse noise (CDR, $\Psi(\lambda, \mu)$) from a dual-channel noisy observation [JNBV11]. The algorithm is based on an estimate of the noise field coherence from a noisy speech signal and a subsequent minima tracking. The algorithm does not require a voice activity detector and gives reliable estimates even in the presence of speech. Depending on the application, this ratio can be computed either frequency-dependent or frequency-independent. The CDR was introduced in case of mixed noise fields in Sec. 2.1.2.2 and used for a blind DRR estimator in Sec. 2.1.3.2.

The basic principle of the CDR estimator is related to the DRR estimator and utilizes Eq.(2.40) as well. In contrast to the DRR estimation, here, the direct computation for the classification of background noise is only possible in segments of speech absence or in segments with very low speech energy. This applies especially when the diffuse noise components are dominant, i.e., $\Psi \leq 0\,\text{dB}$ (see Fig. 2.4) and would restrict the application of the CDR estimator to the very limited periods of speech absence. In the following, it is discussed how to perform a reliable CDR estimation independent of speech activity. The speech signal is assumed to be the target signal for any subsequent noise reduction or classification system, i.e., a person standing in front of an hearing impaired person, or the speech signal emitted from the mouth to a mobile phone in hand-held or hands-free position. Hence, this speech signal alone would have a coherence close to one if we assume that the person is located within the critical distance (high DRR). Besides, we assume that the CDR changes slowly over time and sudden changes are not taken into account.

For the upcoming simulations, the noise signals are generated using the approach of [HCG08], where predefined spatial coherence constraints can be employed. In order to demonstrate the principle of the algorithm, WGN having the same PSD is used for the coherent and diffuse noise which are mixed at a CDR of $\Psi = -9\,\text{dB}$. All signals are uncorrelated with each other.

For the noisy signal, speech samples which are coherent among the microphones from the TSP speech database [Kab02] are summed with the mixed noise signals at a *Signal-to-Noise Ratio* (SNR)[11] of 10 dB. An inter-microphone distance of $d_{\text{mic}} = 0.15\,\text{m}$ is assumed which is a typical distance between two binaural hearing aids. For simplicity, the head shadowing is neglected for the experiments and the mixed noise field coherence model given by Eq.(2.18) is used which was derived by using the free-field diffuse model of Eq.(2.6). It is further assumed that a possible time-delay of the speech signal among the microphones has been compensated. Such simulation ensures reproducible results and are required in order to evaluate the CDR estimation accuracy which would be inherently difficult with measured data. Further simulation parameters are listed in Table 2.8.

The influence of an additional speech signal to the noise input signal on the coherence function is illustrated in Fig. 2.35. The estimated coherence for a noise-only and noisy

[11] Here, the SNR is defined as the ratio between the desired speech signal to the mixture of diffuse and coherent noise.

Table 2.8: Main simulation parameters of the novel blind CDR estimation method.

Parameter description	Setting
Frame overlap	75% (Hann window)
Thresholds	$\mathrm{CDR}^{(\mathrm{min})} = -20\,\mathrm{dB}$, $\mathrm{CDR}^{(\mathrm{max})} = 20\,\mathrm{dB}$
Smoothing factors	$\alpha^{(\mathrm{PSD})} = 0.8$, $\alpha^{(\mathrm{coh})} = 0.5$
Minima tracking window length	1.5 s

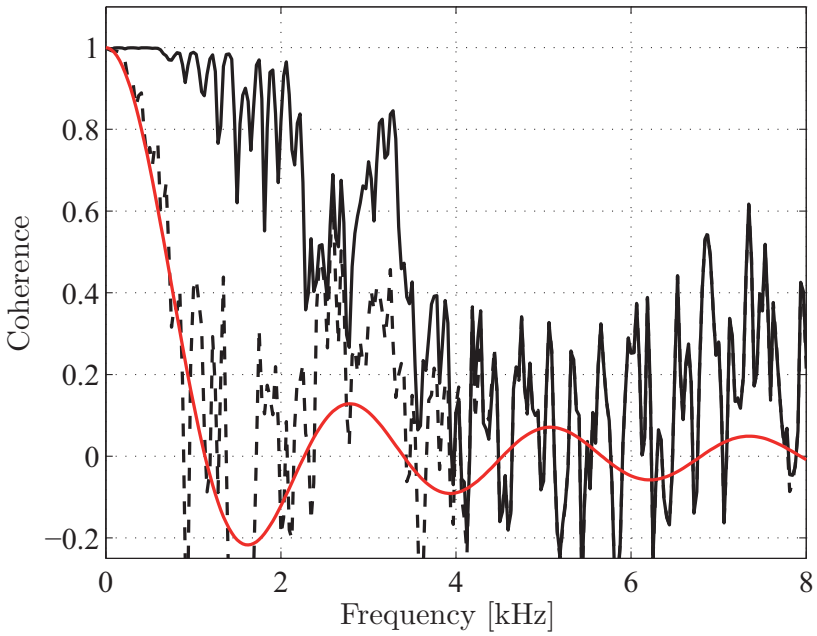


Figure 2.35: Coherence of noise (dashed) and noisy speech (solid) with frequency-independent CDR of $\Psi = -9\,\mathrm{dB}$ and additive WGN with 10 dB SNR. The ideal noise field coherence curve of the mixed noise field is computed using Eq.(2.18) with $d_{\mathrm{mic}} = 0.15\,\mathrm{m}$, $\Psi = -9\,\mathrm{dB}$ and is marked by the solid red line.

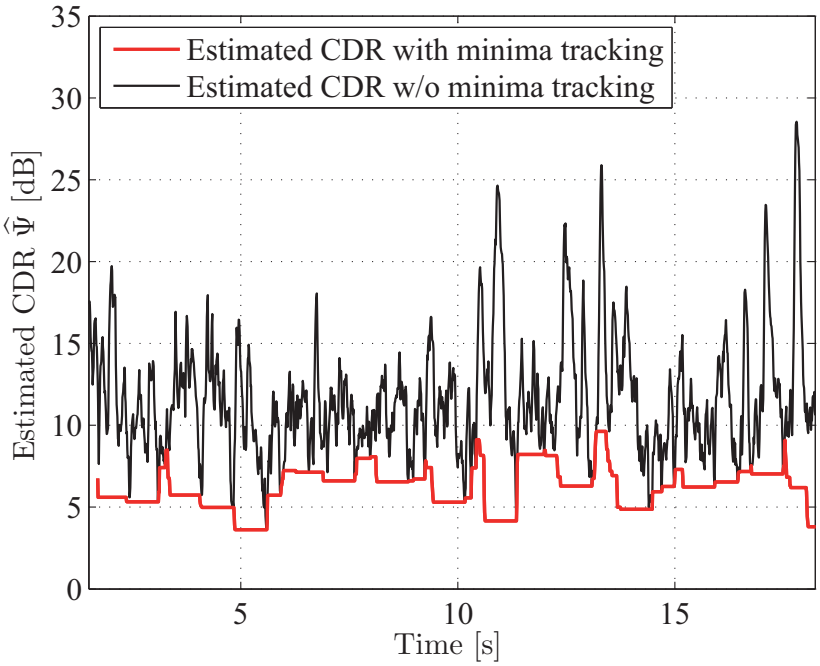


Figure 2.36: Estimated CDR without minima tracking $\widehat{\Psi}(\mu)$ and with minima tracking $\widehat{\Psi}_{(\min)}(\mu)$ for one frequency bin $\mu = 65$ ($\widehat{=}2\,\text{kHz}$). The signals are mixed with a fixed CDR $\Psi = 6\,\text{dB}$ and additive WGN with 10 dB SNR.

speech frame is given. In the case of speech absence (dashed black line), the estimated coherence follows the theoretical red solid line for the given CDR (Eq.(2.18)). When it comes to a noisy speech signal (solid black line), the estimated noise field coherence is biased towards higher values. This can be explained since the coherence of the speech signal alone would be one, as mentioned previously. Furthermore, the speech signal has the highest energy components for a frequency range up to 3 kHz. Hence, the estimated CDR would always return values for the mixing ratio which are higher than the true CDR in segments of speech activity. For very low SNR conditions ($< 0\,\text{dB}$), where the noise signal dominates the overall signal energy, this effect is negligible.

In order to counteract the biased CDR estimate, a minima tracking of the estimated CDR per frequency bin is performed using a large tracking window of 1.5 s. This concept is well-known from noise PSD estimation [Mar01a]. Figure 2.36 shows the estimated CDR over time for one specific frequency bin. In this case the CDR was set fixed to $\Psi = 6\,\text{dB}$ and the SNR to 10 dB. Depicted is the estimated CDR without minima tracking $\widehat{\Psi}$ and with minima tracking $\widehat{\Psi}_{(\min)}$. From this figure, it can be concluded that the speech signal causes a severe bias to the estimated CDR and hence, the performance of any subsequent algorithm which relies on this estimate would be degraded.

The complete CDR estimation algorithm is summarized in Fig. 2.37. The processing is performed for each frame λ and, depending on the application, a frequency-independent or frequency-dependent CDR estimate can be obtained. For complexity reasons, fixed smoothing factors for the recursive estimation of the auto- and cross-PSD and no bias correction as in [Mar01a] are employed. The usage can, however, reduce the remaining bias between the true and estimated CDR.

- transform current frame λ into the frequency domain with frequency bin $\mu = 1, ..., N$
- compute smoothed auto- and cross-PSD $\widehat{\Phi}_{x_1x_1}(\lambda,\mu)$, $\widehat{\Phi}_{x_2x_2}(\lambda,\mu)$, $\widehat{\Phi}_{x_1x_2}(\lambda,\mu)$ by recursive smoothing ($\alpha^{(\mathrm{PSD})}$); Eqs.(2.35), (2.36) and (2.37).
- compute coherence $\Gamma_{x_1x_2}(\lambda,\mu)$; Eq.(2.3) by recursive smoothing ($\alpha^{(\mathrm{coh})}$)
- compute CDR estimate $\widehat{\Psi}(\lambda,\mu)$; Eq.(2.40)
- perform minima tracking to obtain $\widehat{\Psi}_{(\min)}(\lambda,\mu)$
- average over frequency

$$\overline{\Psi}(\lambda) = \frac{1}{N-\mu_0} \sum_{\mu=\mu_0}^{N} \widehat{\Psi}_{(\min)}(\lambda,\mu) \tag{2.55}$$

- process next frame $\lambda + 1$

Figure 2.37: CDR estimation algorithm summary. The averaging over frequency is required only for a frequency-independent CDR estimate.

The overall performance of the proposed CDR estimator with and without the minima tracking is depicted in Fig. 2.38. The accuracy of the CDR estimate has been increased significantly by applying a minima tracking to the direct estimate obtained from Eq.(2.40). For the desired operating range of $-9 \leq \Psi \leq 9\,\mathrm{dB}$ (see above and Fig. 2.4), the estimator shows sufficient accuracy. Since the ideal diffuse noise field shows a high coherence for low frequencies (see Fig. 2.4 and Eq.(2.6)), no clear separation between coherence and diffuse noise is possible. The estimator is only capable for a frequency-dependent CDR estimation above $f_0/4$, which was confirmed heuristically by experiments. In further simulations we have shown that the inter-microphone distance d_{mic} should be equal or greater than 0.02 m.

2.3.6.1 Influence of Coherent Noise Direction

It has to be mentioned that the assumption for the angle of the interfering source of $\theta = \pi/2$ in the derivation of the CDR estimator (see Eq.(2.18)) does not have an influence on the estimation performance as shown in Fig. 2.39. In the simulation, a coherent noise signal from alternative angles was considered. It can be seen that the simplification in the derivation of the CDR estimator does not affect the estimation accuracy. Furthermore, the main target applications are (binaural) hearing aids and dual-microphone mobile phones where an inter-microphone spacing of 15 cm, 10 cm or 3 cm is assumed and a frame-wise processing with a typical frame length of 20 ms.

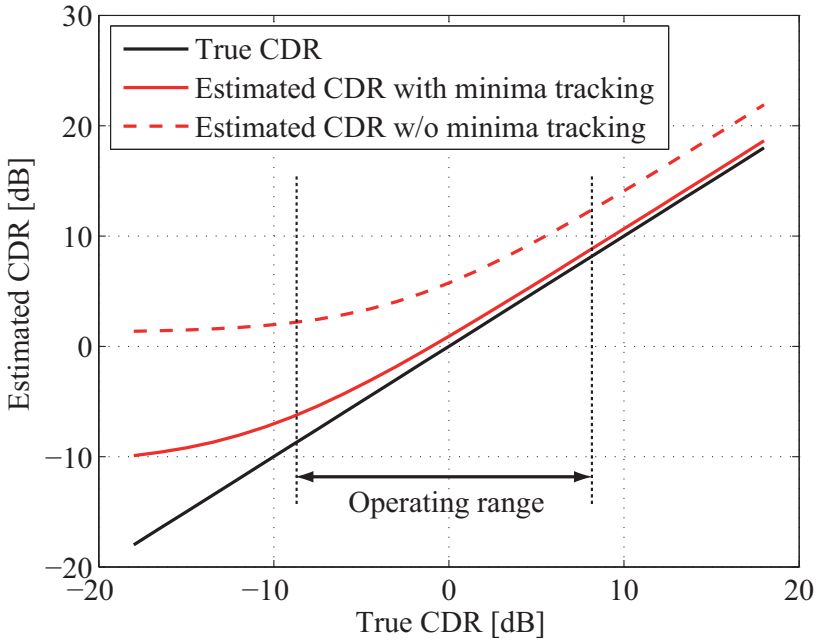


Figure 2.38: True and estimated CDR for varying CDR values. The results are averaged over all frequency bins for each CDR step and averaged over time.

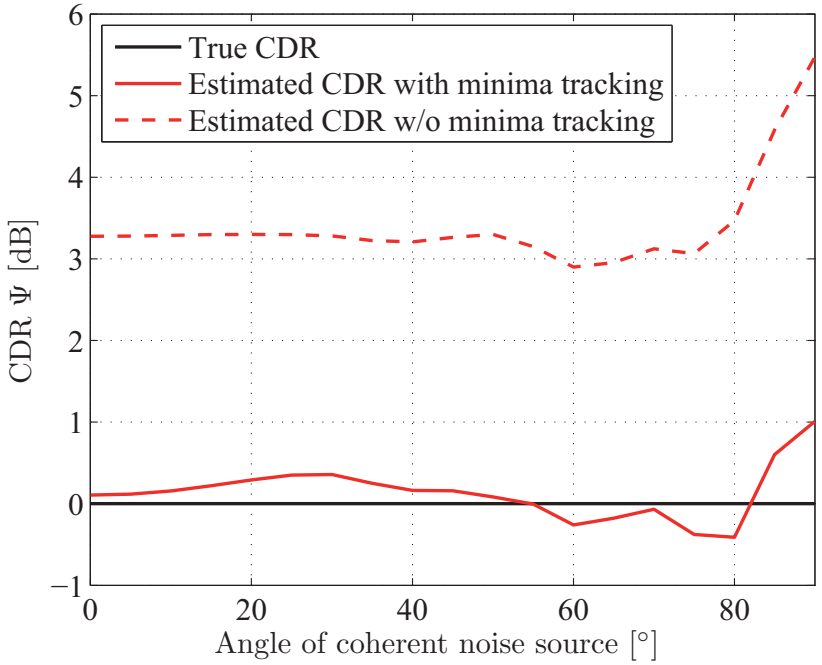


Figure 2.39: True and estimated CDR for varying angles θ of the coherence noise source. 90° indicates frontal direction.

2.4 Summary

The acoustic properties of the most relevant environments for this thesis where hearing aids and mobile phones are mostly used were discussed. The analysis of the sound field coherence and typical values for acoustic parameters such as RT and DRR were given based on measurements of room impulse responses and recordings of background noise. Furthermore, efficient algorithms to estimate the most important acoustic parameters, i.e., short-term noise field coherence, DRR and RT were discussed and evaluated. Finally, an efficient algorithm to estimate the binaural cues as well as a novel noise field classification algorithm were presented. In summary, the novel aspects are:

- an elaborate analysis of the acoustic environment based on measurements and recordings in realistic scenarios including:
  - design of mobile phone mock-up devices and discussions on microphone positions,
  - evaluation of power level differences of speech and noise between the microphones,
  - verification of common assumption of homogeneous noise fields,
  - investigation of practical values for frequency-dependent RT and DRR,
  - proof that room reverberation is audible for mobile phone conversations even in the hand-held position based on listening tests.
- a new generalized coherence model for mixed diffuse and coherent noise fields.
- the review and evaluation of a binaural coherence model which takes the head shadowing effects into account.
- an improved method for the onset time detection of a RIR which is more accurate than common approaches.
- novel models to approximate the frequency-dependency of the RT.
- an extended method to estimate the frequency-dependent RT blindly from a reverberant speech signal.
- derivation of a new blind method to estimate the DRR which is superior to state-of-the-art approaches.
- a improved method to estimate the binaural cues from a reverberant speech signal.
- a robust noise field classification algorithm which is capable to determine the CDR even in segments of speech presence.

This chapter has given all prerequisites for a deep understanding and design of advanced speech enhancement algorithms which are considered in the remainder of this thesis.

3

Joint Dereverberation and Noise Reduction

Speech signals captured by the microphones of a speech communication device are often distorted by interfering noise sources as well as room reverberation. Such degradations may reduce the listening comfort and the speech intelligibility. The target of any speech enhancement algorithm should be a reduction of unwanted background noise and room reverberation while ensuring that the occurring speech distortions are as low as possible.

Among the wide variety of speech enhancement algorithms, a predominant principle is the frequency domain spectral subtraction technique [Bol79], which generally requires a short-term PSD estimate of the interfering signal $\widehat{\Phi}_{\text{int}}(\lambda, \mu)$. Based on this estimate, spectral weighting gains $G(\lambda, \mu)$ are calculated and applied to the degraded input spectral magnitudes by

$$\widehat{S}(\lambda, \mu) = X(\lambda, \mu) \cdot G(\lambda, \mu). \tag{3.1}$$

The enhanced time-domain signal $\widehat{s}(k)$ is obtained by using the *Inverse Fast Fourier Transform* (IFFT) and overlap-add [Cro80]. This general structure of spectral weighting-based speech enhancement which operates in the frequency domain is depicted in Fig. 3.1.

In the context of speech dereverberation and noise reduction, the interfering PSDs are denoted by $\widehat{\Phi}_{\text{rev}}^{(\text{late})}(\lambda, \mu)$ and $\widehat{\Phi}_{\text{nn}}(\lambda, \mu)$, respectively[1]. When it comes to noisy and reverberant signals, an overall interfering PSD is introduced and denoted by $\widehat{\Phi}_{\text{int}}(\lambda, \mu)$. In this chapter, dereverberation and noise reduction are considered independently first, followed by discussions how to efficiently combine different algorithms and estimation techniques to allow for a joint reduction of reverberation and background noise.

[1] Please note that in literature the term *Late Reverberant Spectral Variance* (LRSV) is used alternatively to the short-term late reverberant speech PSD.

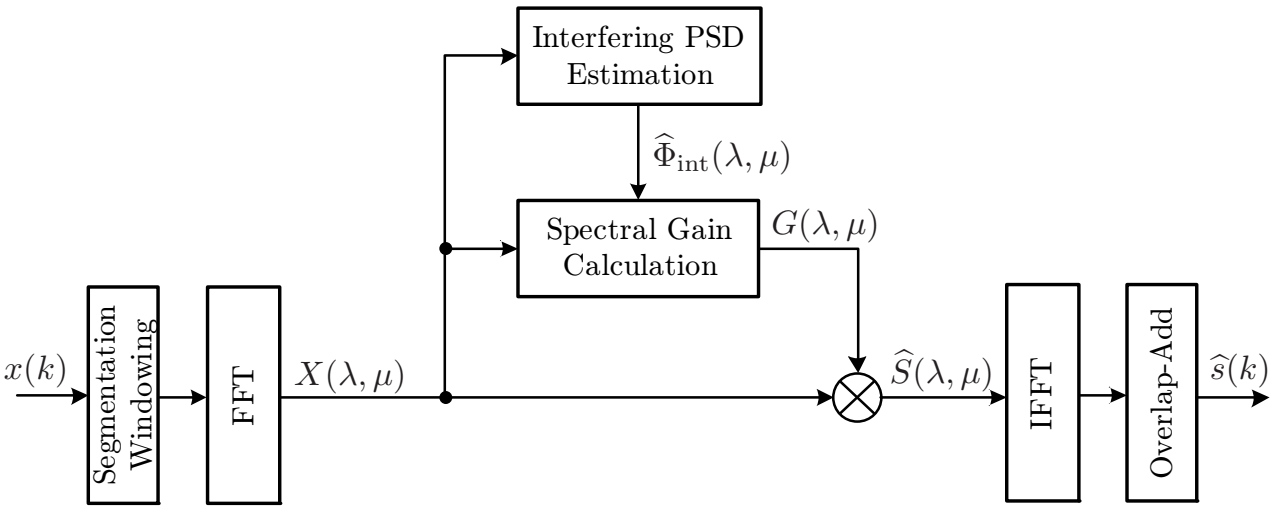


Figure 3.1: Principle of single-channel speech enhancement based on spectral weighting in the frequency domain.

3.1 Dereverberation

In speech communication systems, room reverberation often leads to a degradation of speech quality and intelligibility. This especially applies for hands-free devices, binaural telephone headsets, and digital hearing aids. Especially in the context of hearing aids and *Cochlear Implants* (CI), reverberation can greatly reduce the localization performance and speech intelligibility. Studies in [See11] have shown that normal hearing participants showed no localization impairments due to reverberation down to a DRR of −8dB, whereas the localization was already affected at positive DRRs between 0 and +10 dB for people with a hearing loss.

The effects of room reverberation can generally be categorized into two distinct perceptual components: overlap-masking and coloration. Late reverberation causes mainly overlap-masking effects, whereas the early reflections are known to cause a coloration of the anechoic speech signal, cf. [NLT89]. It is well-known that very early reflections, i.e., a few milliseconds after the onset time of the RIR, can lead to an increase in intelligibility since they can be partially integrated with the direct speech signal, cf. [BSP03, WRDK11]. In [SJSV10], artificial very early reverberation was added to a transcoded speech signal in order to reduce the amount of perceived speech distortions and to increase the intelligibility, i.e., higher *Speech Transmission Index* (STI) [IEC03]. Hence, the very early reflections should not be removed from the reverberant speech signal.

Many contributions have been made in the past to reduce the effects due to reverberation, cf., [Leb99, NG05, Hab07, LV09b, NG10]. Since a joint suppression of both early *and* late reverberation is quite challenging, several (single- and multi-microphone) two-stage algorithms are proposed in the literature. The authors in [WW06] present an inverse filtering algorithm which maximizes the kurtosis of the *Linear Prediction* (LP) residual signal for the reduction of early reverberation, followed by a spectral subtraction rule that reduces long-term reverberation. A similar

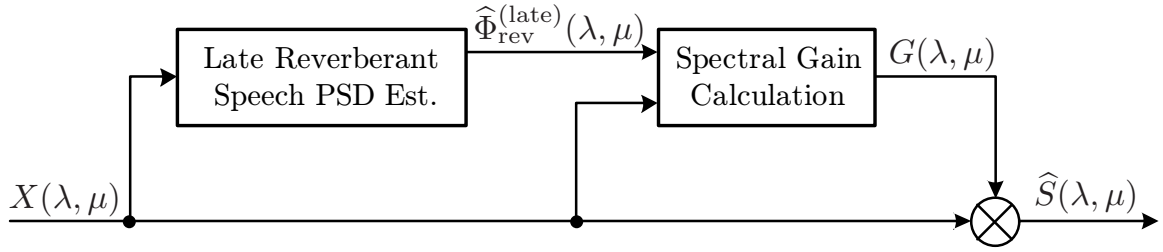


Figure 3.2: Principle of single-channel dereverberation based on spectral weighting in the frequency domain.

approach is described in [GHN08] where spatio-temporal averaging is combined with a spectral subtraction algorithm.

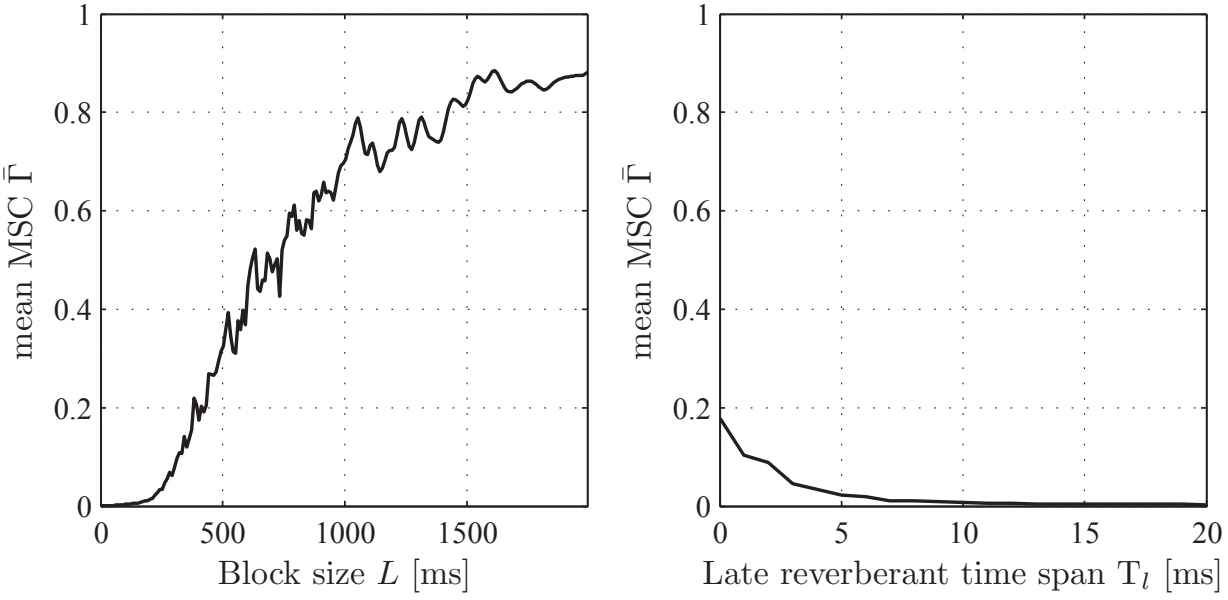
This section discusses two major dereverberation techniques: Single-channel algorithms which estimate the short-term PSD of the late reverberant speech are introduced and extended in Sec. 3.1.1. This includes methods which rely on statistical reverberation models as well as approaches that exploit inter-frame correlation. A key aspect is that all required acoustic parameters such as the RT and DRR are estimated blindly from the noisy and reverberant input signals. Section 3.1.2 gives an overview of dual-channel methods which exploit the different coherence properties of both speech and reverberation. Furthermore, alternative algorithms which take advantage of the discrete model of speech production are briefly discussed in Sec. 3.1.3.

Further approaches are not considered in this thesis. Please refer to, e.g., [NG10] and the references therein.

3.1.1 Estimation of Late Reverberant Speech PSD

The generalized signal model of single-channel speech dereverberation by spectral weighting is shown in Fig. 3.2 which requires a reliable estimate of the late reverberant speech PSD $\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)$. All algorithms in literature which are based on this principle assume that the direct speech plus a few early reflections and the late reverberant speech are uncorrelated. However, this statement is only true for small block sizes as shown by the following experiment. A RIR of a lecture room is decomposed into its direct speech components plus early reverberation as well as its late reverberant speech components according to Eq.(2.20) with $\mathrm{T}_l = 100\,\mathrm{ms}$. Figure 3.3 shows in Subfigure (a) the mean MSC (averaged over all frequency bins and frames) between the direct speech components and the late reverberant speech component over a varying block size for the MSC calculation. It can be seen that the direct and late reverberant speech are only uncorrelated for small block sizes, e.g., the considered 20 ms.

Subfigure (b) shows the results for a fixed block size of 20 ms and a varying late reverberant time span T_l. From this figure it can be concluded that for a small block size, the correlation can be neglected for $\mathrm{T}_l > 10\,\mathrm{ms}$.



(a) (b)

Figure 3.3: Correlation between direct and late reverberant speech. (a) mean MSC over block size with $\mathrm{T}_l = 100\,\mathrm{ms}$, (b) mean MSC over the late reverberant time span.

3.1.1.1 Estimation using Statistical Reverberation Models

An efficient dereverberation algorithm based on a statistical model of late reverberation (see [Pol88] and Sec. 2.1.3.3) has been proposed first in [Leb99] and was later refined in [LBD01]. The basic idea is to estimate the PSD of the late reverberant speech components and to formulate a weighting rule that aims to suppress late reverberant speech components while leaving the direct speech and early reflections unaltered.

This subsection briefly describes an improved single-channel algorithm based on [LBD01] which utilizes a generalized statistical model of the room impulse response according to [HGC09] (see Sec. 2.1.3.3). This generalization allows to use the algorithm also in situations where the source is located within the critical distance. Possible extensions to binaural outputs are discussed in the application Sec. 4.1.

As discussed in Sec. 2.1.3, a room impulse response can be decomposed into the two distinctive components $h_{\mathrm{early}}(k)$ and $h_{\mathrm{late}}(k)$ according to Eq.(2.20). Hence, the reverberant signal $x(k)$ can be decomposed into its early and late reverberant speech components $x_{\mathrm{early}}(k)$ and $x_{\mathrm{late}}(k)$ by

$$x(k) = \underbrace{\sum_{n=0}^{\mathrm{T}_l f_s - 1} s(k-n)h(n)}_{x_{\mathrm{early}}(k)} + \underbrace{\sum_{n=\mathrm{T}_l f_s}^{\mathrm{T}_r f_s} s(k-n)h(n)}_{x_{\mathrm{late}}(k)}, \tag{3.2}$$

where the corresponding DFT spectra are termed $X_{\text{early}}(\lambda,\mu)$ and $X_{\text{late}}(\lambda,\mu)$, respectively. An estimate of the late reverberant speech PSD can be obtained by means of a simple statistical model for the room impulse response. Based on Eq.(2.28), it can be shown that the late reverberant component $x_{\text{late}}(k)$ (or $X_{\text{late}}(\lambda,\mu)$) can be modeled as an uncorrelated noise process (see Fig. 3.3(a) and [LBD01]). An estimator for the short-term PSD of the late reverberant speech is given in [LBD01] by

$$\widehat{\Phi}_{\text{rev}}^{(\text{late})}|_{\text{LB}}(\lambda,\mu) = e^{-2\rho \text{T}_l} \cdot \widehat{\Phi}_{xx}(\lambda - N_l, \mu), \tag{3.3}$$

with the PSD $\widehat{\Phi}_{xx}(\lambda,\mu)$ of the reverberant speech obtained by recursive smoothing (using Eq.(2.35) with $\alpha^{(\text{xx})}$) and N_l the number of frames corresponding to T_l. From Eq.(3.3) it can be seen that the estimator requires knowledge about the decay rate or reverberation time which was assumed to be frequency-independent in the original implementation, i.e., $\rho = \text{const}\,\forall\,\mu$. In the experiment section it will be shown how this affects the estimation accuracy.

The major limitation of this estimator is the usage of the statistical model which is only valid when the direct path energy is smaller than the energy of all reflections (DRR $< 0\,$dB, within the critical distance). Based on the generalized statistical RIR model Eq.(2.30), an improved estimator for the late reverberant speech PSD can be expressed by [HGC09]

$$\begin{aligned}\widehat{\Phi}_{\text{rev}}^{(\text{late})}|_{\text{HB}}(\lambda,\mu) = (1-\kappa(\mu)) \cdot e^{-2\rho(\mu)\text{T}_l} \cdot \widehat{\Phi}_{\text{rev}}^{(\text{late})}|_{\text{HB}}(\lambda-1,\mu) \\ + \kappa(\mu) e^{-2\rho(\mu)\text{T}_l} \cdot \widehat{\Phi}_{xx}(\lambda - N_l, \mu).\end{aligned} \tag{3.4}$$

The constant $\kappa(\mu)$ $(0 \leq \kappa(\mu) \leq 1)$ is inversely proportional to the direct-to-reverberant energy ratio [HGC09] and the decay rate is now dependent on the frequency. In contrast to Eq.(2.25), the DRR must be given in linear scale in this equation. $\widehat{\Phi}_{xx}(\lambda,\mu)$ is calculated as in Eq.(3.3). Please note that for the special case $\kappa(\mu) = 1\,\forall\,\mu$, the estimator in Eq.(3.4) reduces to the approach of [LBD01] given by Eq.(3.3).

Recently, a new estimator has been proposed in [EH11] which is based on the decaying cosine RIR model given by Eq.(2.33).

3.1.1.2 Estimation Exploiting Long-Term Correlation

The algorithm proposed in [EH09, EH10][2] exploits the property that reverberation increases the correlation between successive speech samples and successive DFT coefficients, respectively. The basic idea is to approximate the late reverberant speech components by a weighted sum of J previous DFT coefficients which are spaced P frames apart according to

$$\widetilde{X}_{\text{late}}(\lambda,\mu) = \sum_{j=0}^{J} c_j(\mu) S(\lambda - \Delta - jP, \mu), \tag{3.5}$$

[2] The author would like to thank Jan Erkelens for helpful discussions and for providing a MATLAB reference implementation.

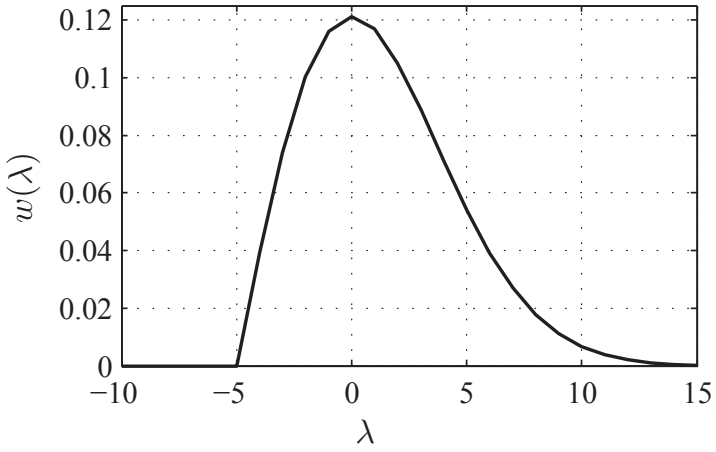


Figure 3.4: Rayleigh distribution used as approximation of equalized impulse response with $\nu = -5$ as in [WW06] over frame index λ.

with frequency-dependent coefficients $c_j(\mu)$ and interval Δ which is introduced to skip early reverberation. In the practical implementation where $S(\lambda, \mu)$ is not given, Eq.(3.5) is predicted by

$$\widehat{X}_{\text{late}}(\lambda, \mu) = \sqrt{B} \sum_{j=0}^{J} \hat{c}_j(\mu) \widehat{S}(\lambda - \Delta - jP, \mu). \tag{3.6}$$

The correction factor B is introduced to compensate the bias of the estimate and $\widehat{S}(\lambda, \mu)$ are the previously enhanced DFT coefficients. The PSD of the late reverberant speech $\widehat{\Phi}^{(\text{late})}_{\text{rev}}|_{\text{EK}}(\lambda, \mu)$ is computed by recursive smoothing of $\widehat{X}_{\text{late}}(\lambda, \mu)$ over time. All further steps how to estimate the unknown quantities in Eq.(3.6), i.e., B and $\hat{c}_j(\mu)$ are described in [EH09, EH10]. Even though this estimator does not require explicit knowledge of the RT or DRR, the choice of the prediction order is somehow related to the RT.

3.1.1.3 Alternative Approaches

In [WW05, WW06] it is assumed that the power spectrum of the late reverberant speech components is a smoothed and shifted version over time of the inverse-filtered speech using the true inverse RIR. The dereverberated speech signal is estimated by convolving the reverberant signal by an approximation of an equalized impulse response $w(\lambda)$ with frame index λ. This is given by a Rayleigh distribution shape of $w(\lambda)$ with fixed parameters, independent of frequency. A spectral subtraction rule is presented which uses this inverse-filtered speech as the late reverberant interfering components. The estimated late reverberant speech PSD reads [WW06]

$$\widehat{\Phi}^{(\text{late})}_{\text{rev}}|_{\text{WW}}(\lambda, \mu) = \gamma^{(\text{WW})} w(\lambda - N_l) * X(\lambda, \mu), \tag{3.7}$$

with N_l as in Eq.(3.3) and scaling constant $\gamma^{(\text{WW})}$. The shape of the equalized impulse response is approximated by a Rayleigh distribution as

$$w(\lambda) = \begin{cases} \frac{\lambda+\nu}{\nu^2} \exp\left(\frac{-(\lambda+\nu)^2}{\nu^2}\right) & \text{for } \lambda < -\nu \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

with spreading parameter ν ($\nu < N_l$), which is chosen heuristically.

A further approach presented in [FK07] states that, due to the long smearing effects of room reverberation, the PSD of the late reverberant speech components of current frame λ is given by the sum of the filtered versions of the previous L_{RT} frames (L_{RT} is the number of frames corresponding to T_{60}). This, however, requires an estimate of the reverberation time. The estimated PSD of the late reverberant speech is given by [FK07]

$$\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}|_{\mathrm{FK}}(\lambda,\mu) \approx \sum_{l=0}^{L_{\mathrm{RT}}} |\gamma^{(\mathrm{FK})}(\lambda,\mu)|^2\, |X(\lambda-l)|^2, \tag{3.9}$$

where the coefficients of the late reverberant speech $\gamma^{(\mathrm{FK})}(\lambda,\mu)$ are estimated by

$$\gamma^{(\mathrm{FK})}(\lambda,\mu) = \frac{X(\lambda,\mu)X^*(\lambda-L_{\mathrm{RT}},\mu))}{|X(\lambda-L_{\mathrm{RT}},\mu)|^2}. \tag{3.10}$$

An additional recursive smoothing over time with constant $\alpha^{(\mathrm{FK})}$ is employed.

3.1.1.4 Spectral Weighting Rules

The weights for the suppression of late reverberation can be calculated, e.g., by a spectral magnitude subtraction rule

$$G_{\mathrm{SS}}(\lambda,\mu) = 1 - \frac{1}{\sqrt{\eta(\lambda,\mu)}}, \tag{3.11}$$

where the required *a posteriori Signal-to-Interference Ratio* (SIR) is calculated by

$$\eta(\lambda,\mu) = \frac{|X(\lambda,\mu)|^2}{\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)}. \tag{3.12}$$

A lower bound $G_{\min}$ is applied to all weighting gains to counter overestimation of $\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)$.

Additionally, the Wiener filter rule is considered and the *a priori* SIR is estimated by means of the *Decision-Directed Approach* (DDA) [EM84] according to

$$\xi(\lambda,\mu) = \alpha^{(\mathrm{DDA})} \frac{|\widehat{S}(\lambda-1,\mu)|^2}{\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda-1,\mu)} + (1-\alpha^{(\mathrm{DDA})}) \cdot \max(\eta(\lambda,\mu)-1,0)) \tag{3.13}$$

with smoothing factor $\alpha^{(\mathrm{DDA})}$. The Wiener filter is then expressed by

$$G_{\mathrm{WF}}(\lambda,\mu) = \frac{\xi(\lambda,\mu)}{1+\xi(\lambda,\mu)}. \tag{3.14}$$

Alternatively, an MMSE log-spectral amplitude estimator (MMSE-LSA) as proposed in [Hab07, Hab10] can be employed. The reader is referred to, e.g., [Loi07, VM06] for further spectral weighting rules.

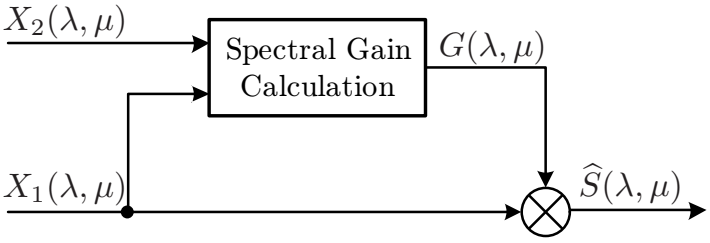


Figure 3.5: Principle of dual-channel coherence-based dereverberation with a single-channel output in the frequency domain.

3.1.2 Dereverberation Using the Coherence Function

The principle of coherence-based dereverberation algorithms is to exploit the different coherence properties of the desired speech and room reverberation. The low coherence between the two microphones of the interfering signals (diffuse background noise and reverberation) is used to estimate the (direct) speech PSD and to remove all non-coherent signal parts while keeping the coherent parts unaffected. Since only the direct speech shows a high coherence among sensors as shown later, this approach also reduces (non-coherent) early reverberation. At the same time, very early reverberation which is coherent between the microphones is not altered.

A further advantage is that no estimation of room acoustic parameters (e.g., T_{60}, DRR) is required and that *a priori* information about the sound field can significantly improve the effectiveness of the algorithm. Please note that this class of algorithms is capable of reducing both (non-coherent) room reverberation and diffuse background noise.

The usage of any coherence-based speech enhancement algorithm generally assumes that the source-microphone distance is smaller than the critical distance. Therefore, the speech signals captured by the two microphones $x_{1|2}(k)$ are mutually correlated, i.e., the MSC between the two microphone signals is close to one. This assumption can be fulfilled mostly for hearing devices and close talking telephone devices. Furthermore, the algorithm requires a short-term estimation of all required PSDs with a block length much smaller than the length of the room impulse response (see Sec. 2.3.1). The general structure is shown in Fig. 3.5, where the spectral weighting gains $G(\lambda,\mu)$ are determined based on the two input channels.

In contrast to the decomposition of the reverberant signal in Eq.(3.2), we will now consider a division into its direct components ($\mathrm{T}_d < 2\,\mathrm{ms}$)[3] and reverberation components ($\mathrm{T}_d \geq 2\,\mathrm{ms}$). For the sake of simplicity, the decomposition is given for the monaural case only, as an extension for each of the two (possibly binaural) channels can be performed in the same manner. The decomposed input signal $x(k)$ can be

[3]The boundary is chosen such that the onset time plus 2 ms marks the direct speech components without early reflections but including very early reverberation.

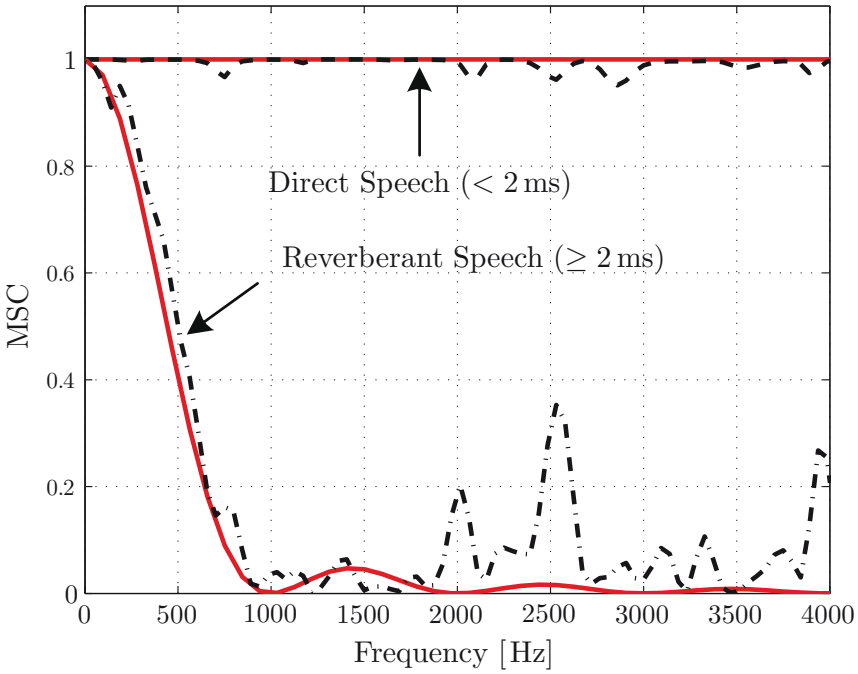


Figure 3.6: MSC of direct speech and speech reverberation (without head). Theoretical curves (solid) and measured curves (dashed) for the parameters: $d_{\mathrm{LM}} = 0.17\,\mathrm{m}$, $\mathrm{T}_{60} = 0.81\,\mathrm{s}$.

expressed by

$$x(k) = \underbrace{\sum_{n=0}^{\mathrm{T}_d \cdot f_\mathrm{s}-1} s(k-n)h(n)}_{x_{\mathrm{direct}}(k)} + \underbrace{\sum_{n=\mathrm{T}_d \cdot f_\mathrm{s}}^{\mathrm{L_r} \cdot \mathrm{f_s}} s(k-n)h(n)}_{x_{\mathrm{reverb}}(k)}, \tag{3.15}$$

where the time span of the direct sound (including sound propagation) is given by T_d (see also Eq.(2.25)). While in Sec. 3.1.1.1 the early speech component $x_{\mathrm{early}}(k)$ was the target signal, now the direct speech component $x_{\mathrm{direct}}(k)$ is the target signal.

Let us now regard the MSC of the two components as illustrated in Fig. 3.6, where the curves have been generated as follows. First, two measured room impulse responses (without dummy head) have been decomposed manually into direct and reverberation components. Afterwards, speech data of 8 s duration from the NTT database has been convolved with each of the RIRs resulting in separate direct and reverberation signals for each channel. Finally, the MSC between the two channels (left and right) has been calculated for both components and is plotted in the range $1-4\,\mathrm{kHz}$. The upper dashed line in Fig. 3.6 shows the MSC of the direct speech component ($C_{x_{\mathrm{direct},1}x_{\mathrm{direct},2}}(\Omega)$) while the lower dashed curve shows the MSC of the reverberation speech component ($C_{x_{\mathrm{reverb},1}x_{\mathrm{reverb},2}}(\Omega)$). The solid lines give the corresponding theoretical coherence function. As a high amount of direct speech is assumed, the theoretical coherence for the direct speech is one for all frequencies. The lower solid line gives the theoretical curve for an ideal diffuse sound field according to Eq.(2.6) with $d_{\mathrm{mic}} = 0.17\,\mathrm{m}$. As seen from the figure, the assumptions having made about

the coherence of direct speech and speech reverberation are valid. Since the reverberation components received by the microphones can be represented by two additive, uncorrelated noise sources, the terms noise and reverberation components are used interchangeably in the following.

Based on the early work in [Dan68], a first practical realization of a coherence-based algorithm for the purpose of speech dereverberation was presented in [ABB77]. The spectral weights are determined directly from the estimated auto- and cross-PSDs. In [ABB77], two alternative spectral weighting gains are calculated by

$$G_{\mathrm{AL}}^{(\mathrm{I})}(\lambda,\mu) = \frac{|\widehat{\Phi}_{x_1x_2}(\lambda,\mu)|}{\frac{1}{2}\left(\widehat{\Phi}_{x_1x_1}(\lambda,\mu) + \widehat{\Phi}_{x_2x_2}(\lambda,\mu)\right)}, \tag{3.16}$$

$$G_{\mathrm{AL}}^{(\mathrm{II})}(\lambda,\mu) = \frac{|\widehat{\Phi}_{x_1x_2}(\lambda,\mu)|}{\sqrt{\widehat{\Phi}_{x_1x_1}(\lambda,\mu)\cdot\widehat{\Phi}_{x_2x_2}(\lambda,\mu)}}, \tag{3.17}$$

where the hat-operator $\{\hat{\cdot}\}$ indicates an estimate of the auto- and cross-PSD terms using Eqs.(2.35),(2.36),(2.37). A similar approach using the magnitude squared coherence weighting gains is proposed in [Pei92].

An improved coherence-based spectral weighting rule was initially derived in [MB03] for the purpose of speech denoising in diffuse noise fields. In [JV10, JSEV10], the dereverberation performance was evaluated and an extension to a binaural output system which also takes the head shadowing into account was proposed. In the remainder this novel method is introduced and extended.

A common framework for speech enhancement is based on the optimal *Minimum Mean Square Error* (MMSE) criterion, cf., [VM06]. It turns out that the optimal weighting gains are given by the Wiener solution

$$G(\lambda,\mu) = \frac{\Phi_{ss}(\lambda,\mu)}{\Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu)}, \tag{3.18}$$

where $\Phi_{ss}(\lambda,\mu)$ denotes the auto-PSD of the original (undisturbed) signal and $\Phi_{nn}(\lambda,\mu)$ the auto-PSD of the additive noise component. As discussed previously, the term $\Phi_{nn}(\lambda,\mu)$ is also referred to the auto-PSD of the speech reverberation component.

For calculating the optimal postfilter coefficients in multi-microphone systems, several approaches have been presented in the past. They all have in common that the estimation procedure is optimized for a specific sound field model. A well-known technique by Zelinski assumes a perfectly incoherent sound field and hence, uncorrelated noise at different sensors [Zel88]. Since this assumption does not hold in real sound fields, an improved approach was presented by McCowan in [MB03] who suggested to use a model of the coherence for a spherically isotropic (diffuse) sound field (Eq.(2.6)). Assuming the same noise PSD across sensors, i.e., homogeneous noise field, as well

as time-aligned signals, the auto- and cross PSDs read

$$\Phi_{x_1x_1}(\lambda,\mu) = \Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu) \tag{3.19}$$
$$\Phi_{x_2x_2}(\lambda,\mu) = \Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu) \tag{3.20}$$
$$\Phi_{x_1x_2}(\lambda,\mu) = \Phi_{ss}(\lambda,\mu) + \Gamma_{n_1n_2}(\Omega)\,\Phi_{nn}(\lambda,\mu), \tag{3.21}$$

where $\Gamma_{n_1n_2}(\Omega)$ is a model of the noise field coherence. Since the employed models are real-valued, e.g., using Eq.(2.6), the cross-PSD $\Phi_{x_1x_2}(\lambda,\mu)$ is necessarily real-valued in the derivation. In the practical implementation, where $\Phi_{x_1x_2}(\lambda,\mu)$ is estimated from the input signals, the imaginary part of the complex value is discarded as in [MB03] which does not alter the performance.

The average PSD of the two channels can be expressed by the arithmetic mean

$$\Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu) = \frac{1}{2}\left(\Phi_{x_1x_1}(\lambda,\mu) + \Phi_{x_2x_2}(\lambda,\mu)\right) \tag{3.22}$$

which can be rearranged to

$$\Phi_{nn}(\lambda,\mu) = \frac{1}{2}\left(\Phi_{x_1x_1}(\lambda,\mu) + \Phi_{x_2x_2}(\lambda,\mu)\right) - \Phi_{ss}(\lambda,\mu). \tag{3.23}$$

By reordering Eq.(3.21) to

$$\Phi_{nn}(\lambda,\mu) = \frac{\Phi_{x_1x_2}(\lambda,\mu) - \Phi_{ss}(\lambda,\mu)}{\Gamma_{n_1n_2}(\Omega)} \tag{3.24}$$

and by combining it with Eq.(3.23) leads to an estimate of the original (undistorted) signal auto-PSD [MB03, JV10, JSEV10]

$$\widehat{\Phi}_{ss}^{(\mathrm{I})}(\lambda,\mu) = \frac{\widehat{\Phi}_{x_1x_2}(\lambda,\mu) - \frac{1}{2}\,\Gamma_{n_1n_2}(\Omega)\left(\widehat{\Phi}_{x_1x_1}(\lambda,\mu) + \widehat{\Phi}_{x_2x_2}(\lambda,\mu)\right)}{1 - \Gamma_{n_1n_2}(\Omega)}. \tag{3.25}$$

Since the denominator should not be negative, a maximum threshold $\Gamma_{\max}$ for the coherence function has to be applied to ensure that $1 - \Gamma_{n_1n_2}(\Omega) \geq 0$ holds for the denominator. The resulting spectral weights of the Wiener filter given by Eq.(3.18) can now be calculated with Eqs.(3.25),(3.22) as

$$G_{\mathrm{MC}}^{(\mathrm{I})}(\lambda,\mu) = \frac{\widehat{\Phi}_{ss}^{(\mathrm{I})}(\lambda,\mu)}{\frac{1}{2}\left(\widehat{\Phi}_{x_1x_1}(\lambda,\mu) + \widehat{\Phi}_{x_2x_2}(\lambda,\mu)\right)}. \tag{3.26}$$

As an alternative to the arithmetic mean of the two noise PSDs (Eq.(3.22)) in the derivation of Eqs.(3.25),(3.26), it is proposed to use the geometric mean instead. This averaging has also been found beneficial in dual-channel noise PSD estimation, cf. [DE96]. The geometric mean is more robust towards large differences of the two auto PSDs, i.e., in the case $\Phi_{x_1x_1}(\lambda,\mu) \ll \Phi_{x_2x_2}(\lambda,\mu)$ or $\Phi_{x_2x_2}(\lambda,\mu) \ll \Phi_{x_1x_1}(\lambda,\mu)$, and ensures larger spectral weighting gains and hence, less speech attenuation.

The insertion of the geometric mean

$$\Phi_{ss}(\lambda,\mu)+\Phi_{nn}(\lambda,\mu)=\sqrt{\Phi_{x_1x_1}(\lambda,\mu)\cdot\Phi_{x_2x_2}(\lambda,\mu)} \tag{3.27}$$

leads to

$$\widehat{\Phi}_{ss}^{(\mathrm{II})}(\lambda,\mu)=\frac{\widehat{\Phi}_{x_1x_2}(\lambda,\mu)-\Gamma_{n_1n_2}(\Omega)\sqrt{\widehat{\Phi}_{x_1x_1}(\lambda,\mu)\cdot\widehat{\Phi}_{x_2x_2}(\lambda,\mu)}}{1-\Gamma_{n_1n_2}(\Omega)}, \tag{3.28}$$

and

$$G_{\mathrm{MC}}^{(\mathrm{II})}(\lambda,\mu)=\frac{\widehat{\Phi}_{ss}^{(\mathrm{II})}(\lambda,\mu)}{\sqrt{\widehat{\Phi}_{x_1x_1}(\lambda,\mu)\cdot\widehat{\Phi}_{x_2x_2}(\lambda,\mu)}}. \tag{3.29}$$

The spectral weights are further confined by a lower threshold $G_{\min}$ for robustness against underestimation errors of $\widehat{\Phi}_{ss}(\lambda,\mu)$ and to control the amount by which reverberation is attenuated.

The crucial point is now to select a suitable model for the sound field coherence $\Gamma_{n_1n_2}(\Omega)$ in Eqs.(3.25),(3.28). For an ideal diffuse sound field with a line-of-sight between two microphones, the optimal solution is the model given by Eq.(2.6). However, when it comes to binaural signal processing where no line-of-sight between the microphones can be assumed, this model is not appropriate. Since the head-shadowing has a severe impact on the coherence, it is proposed to use the coherence model for a binaural spherically isotropic sound field as described in Sec. 2.2.1.1 and App. B.

Since the main field of operation are binaural hearing aids, an extensive discussion on the application of this class of algorithms is given in Sec. 4.1.

3.1.3 Alternative Approaches

Several algorithms for speech dereverberation based on a *discrete model of speech production* have been proposed in the past [JV09a]. They are based on a simplified model consisting of an excitation source and a time-varying vocal tract filter, cf. [VM06]. The corresponding model parameters are estimated by means of *Linear Prediction* (LP) techniques. In order to reduce the effect of room reverberation, the spectral envelope as well as the excitation signal can be modified.

Early studies in [YM00] apply an adaptive time-domain weighting function to the LP residual. This should emphasize regions with a high *Signal-to-Reverberant Energy Ratio* (SRR) and attenuate low SRR regions. A different approach in [GMF01] exploits the kurtosis of the LP residual, which is an indicator for the peakedness. While it has a more Gaussian distribution (smaller kurtosis) in reverberant environments, the kurtosis becomes larger with decreasing reverberation times. An adaptive filter is designed to maximize the kurtosis and hence, to minimize the effect of reverberation. In [WW06] this method is, with some modifications, combined with a

spectral subtraction dereverberation method. [PdLN11] investigates the optimal parameters and defines an improved stopping criterion for the adaptive filter. A further method named *Spatiotemporal averaging Method for Enhancement of Reverberant Speech* (SMERSH) is described in [GNW04]. The objective is to reduce unwanted peaks in the LP residual by averaging the residual signal between consecutive cycles of opening and closing of the glottis (larynx cycle) while excluding the segments around the glottal closure instances (GCI) [GNW04]. Estimation of the GCI is performed and the LP residual is multiplied in the time-domain with a cosine window having the length of one larynx cycle. Afterwards, averaging over the nearest neighboring cycles is carried out. This technique is performed on voiced speech only where a pulse-like excitation is assumed. Since it leaves unvoiced speech unaffected, a further improvement was presented in [TGGN07]. An equalization filter is applied to perform the equivalent operation of temporal averaging on unvoiced and silence speech. In [GB99] a wavelet clustering algorithm is applied to the LP residual. The basic idea is to cluster the multiple channel signals according to their wavelet extrema to obtain a single residual signal. This is synthesized to obtain a dereverberated signal. The same authors suggest in [GB01] to perform a rough estimate of the room impulse response for each channel in a multi-channel approach. A weighting function is computed for each channel by applying a matched filter type operation. Each weighted residual signal is then aligned and added. In [JV09b] the application of the postfilter algorithm used in CELP speech codecs for the purpose of speech dereverberation was discussed. It is shown that in case of a reverberant signal, the amplitudes of the unwanted peaks in the excitation signal are attenuated and that this approach is capable of reducing early reverberation.

In experiments conducted in this thesis and in [JV09a], the dereverberation performance of this class of algorithms was found to be very limited compared to the abovementioned concepts. Since most of the source-model based algorithms introduce also a high amount of audible speech distortions, they will no longer be considered in the remainder of this work.

3.1.4 Performance Evaluation

In order to evaluate the performance of the discussed dereverberation concepts, experiments with different (binaural) room impulse responses from the AIR database were carried out. To cover a wide scenario, 11 different RIRs from the stairway, lecture room and office were convolved with a sequence of 3 min. speech material from the TSP database. Before evaluation, the levels of the processed signals are normalized to -26 dBov using the ITU-T Rec. P.56 speech voltmeter [ITU93]. Silence periods have been removed before evaluation using the VAD of the AMR-WB speech codec [3GP04c].

3.1.4.1 Estimation Accuracy: Late Reverberant Speech PSD Estimators

This section compares the performance of the five presented algorithms to estimate the PSD of the late reverberant speech:

- LB: using a statistical model (Sec. 3.1.1.1, [LBD01]),
- HB: using a generalized stat.model (Sec. 3.1.1.1, [HGC09]),
- EK: exploiting inter-frame correlation (Sec. 3.1.1.2, [EH10]),
- WW: inverse filtering approach (Sec. 3.1.1.3, [WW06]),
- FK: weighted sum approach (Sec. 3.1.1.3, [FK07]).

The performance is evaluated in terms of the log-error distortion measure between the estimated PSD $\widehat{\Phi}_{\text{rev}}^{(\text{late})}(\lambda, \mu)$ and a reference $\Phi_{\text{rev}}^{(\text{late})}(\lambda, \mu)$.

As introduced for the evaluation of noise PSD estimation algorithms in [GH11], the measures are separated into overestimations and underestimations. The overall error is given by the sum of the individual measures. This procedure has the great advantage to distinguish between overestimations which, when applied to speech enhancement, mainly cause speech distortions and underestimations which mainly result in remaining spurious time-frequency bursts which are perceived as musical noise. The corresponding measures are given by

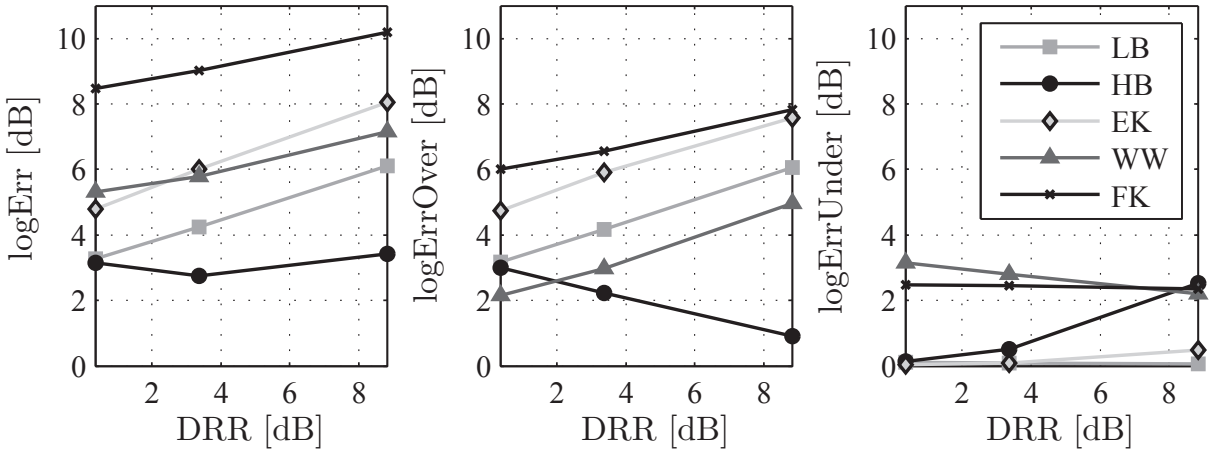
$$\text{logErrOver} = \frac{1}{K\,M} \sum_{\lambda=1}^{K} \sum_{\mu=1}^{M} \left| \min \left(0, 10 \log_{10} \left[\frac{\Phi_{\text{rev}}^{(\text{late})}(\lambda, \mu)}{\widehat{\Phi}_{\text{rev}}^{(\text{late})}(\lambda, \mu)} \right] \right) \right|, \tag{3.30}$$

$$\text{logErrUnder} = \frac{1}{K\,M} \sum_{\lambda=1}^{K} \sum_{\mu=1}^{M} \max \left(0, 10 \log_{10} \left[\frac{\Phi_{\text{rev}}^{(\text{late})}(\lambda, \mu)}{\widehat{\Phi}_{\text{rev}}^{(\text{late})}(\lambda, \mu)} \right] \right), \tag{3.31}$$

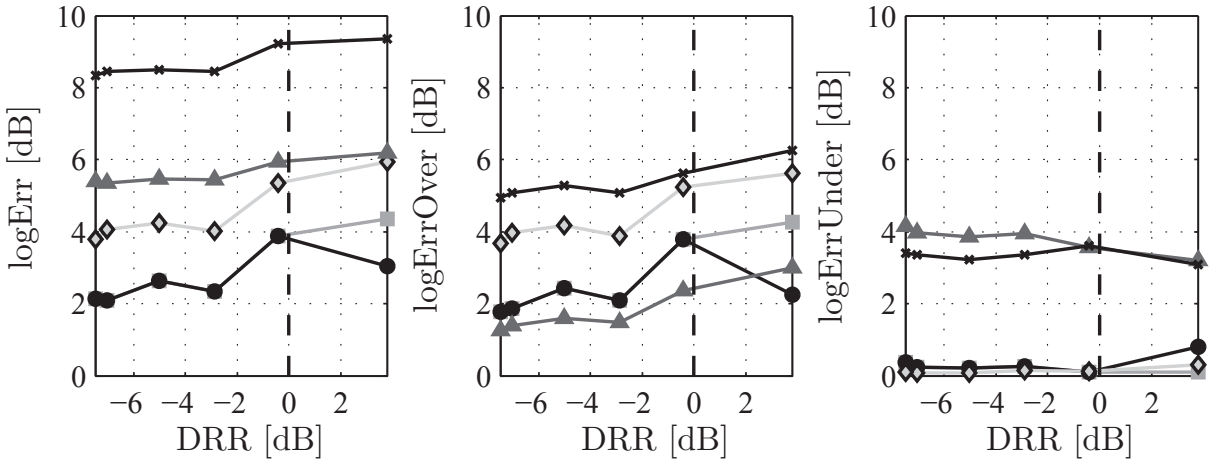
$$\text{logErr} = \text{logErrOver} + \text{logErrUnder}, \tag{3.32}$$

with total number of frames K. The ideal PSD $\Phi_{\text{rev}}^{(\text{late})}(\lambda, \mu)$ is obtained using the true reverberant periodograms $X_{\text{late}}(\lambda, \mu)$ smoothed over frames λ with smoothing factor $\alpha^{(\text{ideal})}$. Relevant settings are given in Table 3.1.

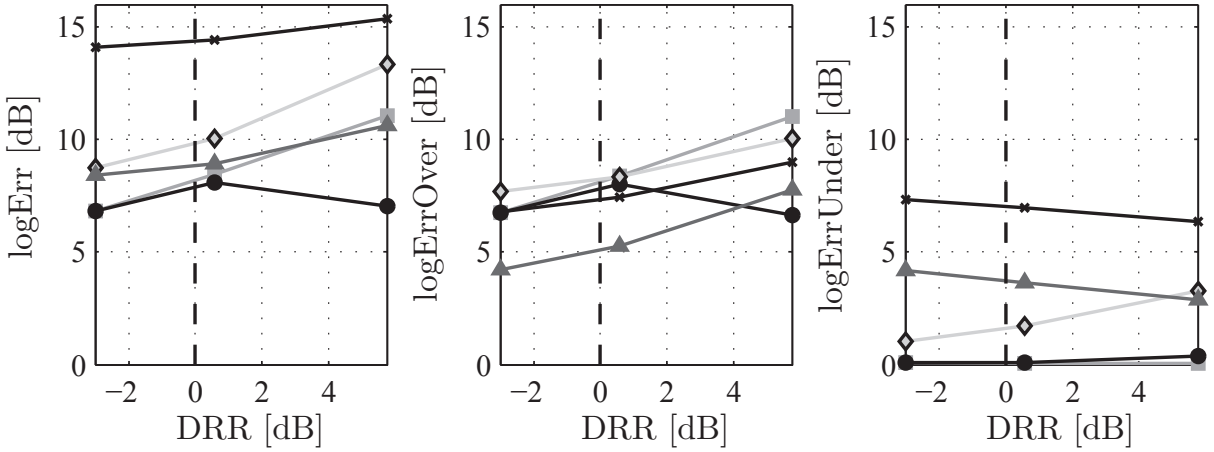
The results are shown in Fig. 3.7 for three different scenarios: stairway, lecture room and office. Plotted are: (left) overall estimation errors, (middle) overestimations, (right) underestimations. All results are obtained with ideal, frequency independent RT and DRR values, where applicable. The main result is that, among the compared five methods, the estimator HB shows the best overall performance. The difference between LB and HB is that, when the DRR becomes positive, strong overestimations occur for LB, which is in accordance with the findings in [HGC09]. The FK method shows the highest estimation errors and result in a high degree of overestimations, which results in a high amount of audible speech distortions. In contrast to that, the WW algorithm leads to less overestimation errors but the resulting high underestimations are audible in terms of disturbing musical tones. For the evaluation, the first 30 s and hence, the transition period after initialization has been neglected. An analysis of this period has shown that the EK algorithm requires, depending on the parameters and prediction order, $10 - 30$ s for the adaptation. All other algorithms



(a) Stairway ($T_{60} = 0.82\,s$).



(b) Lecture room ($T_{60} = 0.81\,s$).



(c) Office ($T_{60} = 0.51\,s$).

Figure 3.7: Performance comparison of five different late reverberant speech PSD estimators in three different rooms. Plotted are the overall estimation errors (left), overestimations (middle) and underestimations (right).

Table 3.1: Main simulation parameters of late reverberant speech PSD estimation algorithms.

Algorithm	Parameter description	Setting
all	Smoothing factor used for ideal PSD calculation	$\alpha^{(\text{ideal})} = 0.9$
LB/HB	Smoothing factor used for reverberant speech PSD	$\alpha^{(\text{xx})} = 0.9$
LB/HB	Late reverberant time span	$\mathrm{T}_l = 80\,\text{ms}$
EK	Prediction interval (in frames)	$P = 2$
EK	Prediction order (in frames)	$J = 30$
EK	No. of frames back that late rev. is assumed to start	$\Delta = 2$
WW	Spreading parameter (Rayleigh distribution)	$\nu = 5$
WW	Attenuation factor	$\gamma^{(\text{WW})} = 0.35$
FK	Smoothing factor	$\alpha^{(\text{FK})} = 0.85$

behave quite similar and show a much shorter period of only a few seconds. For applications where a very low computational complexity is required and small distortions of the processed speech signal can be tolerated, the WW method is proposed to be used since it requires no knowledge about RT and DRR.

In the following experiments, the HB algorithm is evaluated in more detail. First, the benefit of a frequency-dependent RT estimation compared to a frequency-independent estimation is shown. For this experiment, the estimation error over frequency was measured in the corridor location, where for each frequency bin the average over consecutive frames was considered. In contrast to the previous experiments, the required RT is estimated blindly from the reverberant input signal (see Sec. 2.1.3.1) frequency-independent and frequency-dependent in three subbands using the proposed approximation model (RT Model 2, Eq.(2.39)). In average, over frequency only an improvement of 0.2 dB was observed when estimating the RT frequency-dependent. However, when regarding the estimation error plot over frequency in Fig. 3.8, a significant improvement for the higher frequency range can be observed. The results are in accordance with the RT over frequency shown in Fig. 2.27 (d), where the red solid line indicates high overestimations of the RT for higher frequencies if a frequency independent value for the RT is used. Hence, it is suggested to perform a frequency-dependent RT estimation in order to reduce the amount of speech distortions. Additionally, the influence of a frequency-dependent blind DRR estimation (using Eq.(2.45)) compared to the averaged (over frequency), frequency-independent blind DRR estimate (using Eq.(2.46)) was evaluated. It turned out that the usage of the frequency-dependent DRR estimator results only in a small decrease of the estimation error. Thus, due to a more efficient implementation, the frequency-independent method is used in the remainder of this work.

In the last experiment, the influence of possible estimation errors of the RT and DRR on the estimation accuracy of the HB algorithm is regarded. The left subfigure in Fig. 3.9 shows the logarithmic error over the RT estimation error in percent, where

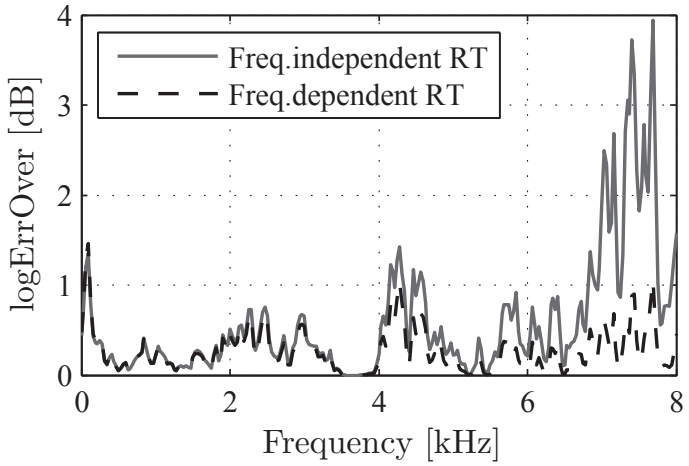


Figure 3.8: Influence of frequency-dependent RT estimation compared to frequency-independent RT estimation on the overestimation error for the HB algorithm (Corridor: BT phone, $d_{\mathrm{LM}} = 0.1\,\mathrm{m}$, $\mathrm{T}_{60} = 1.12\,\mathrm{s}$, $\mathrm{DRR} = 12.71\,\mathrm{dB}$).

0% indicates the use of the true RT. It can be seen that underestimations of the RT (negative percent values) result in underestimations whereas overestimations of the RT lead to a strong overestimation of the late reverberant speech PSD. The corresponding effects of the DRR estimation accuracy is shown in the right subfigure where the opposed effect can be observed, i.e., DRR underestimations lead to PSD overestimations and DRR overestimations to PSD underestimations. From this experiment it can be concluded that the parameter estimation has a significant influence on the estimation accuracy of the late reverberant speech PSD estimator and it has to be ensured that RT overestimations and DRR underestimations have to be limited to the greatest possible extent.

The evaluations in this section have shown that the HB algorithm gives the best performance among state-of-the-art estimators of the PSD of the late reverberant speech signal. It was demonstrated that taking the proposed frequency-dependent blind RT estimation procedure introduced in Sec. 2.3.3.1 into account is beneficial in terms of computational complexity and lowers the estimation error. Besides, the DRR estimator proposed in Sec. 2.3.4 is capable to estimate reliable values of the DRR and is beneficial for an integration into the HB method. By means of the proposed modifications, this estimator gives very reliable results and has the great advantage that all required acoustic parameters are estimated blindly from the reverberant input signals.

The same tendency was observed when evaluating the dereverberation performance in terms of objective speech quality measures. Such results are given later in the application chapter.

3.1.4.2 Dereverberation Performance: Coherence-Based Algorithms

For an objective evaluation of the discussed coherence-based dereverberation concepts, the non-intrusive measurement based on the SRMR is employed [FC08, FZC10].

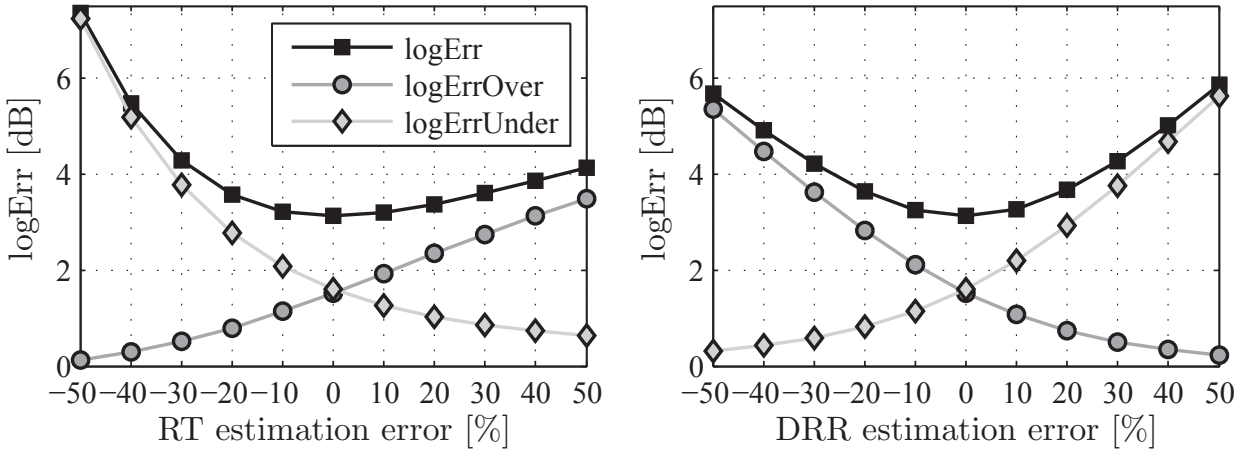


Figure 3.9: Influence of RT and DRR estimation errors (left and right) on the estimation accuracy of the HB method. The parameter estimation error is given in percent where 0% indicates perfect estimation. The results are obtained using RIRs of the stairway hall.

Table 3.2: Main simulation parameters of coherence-based dereverberation algorithms.

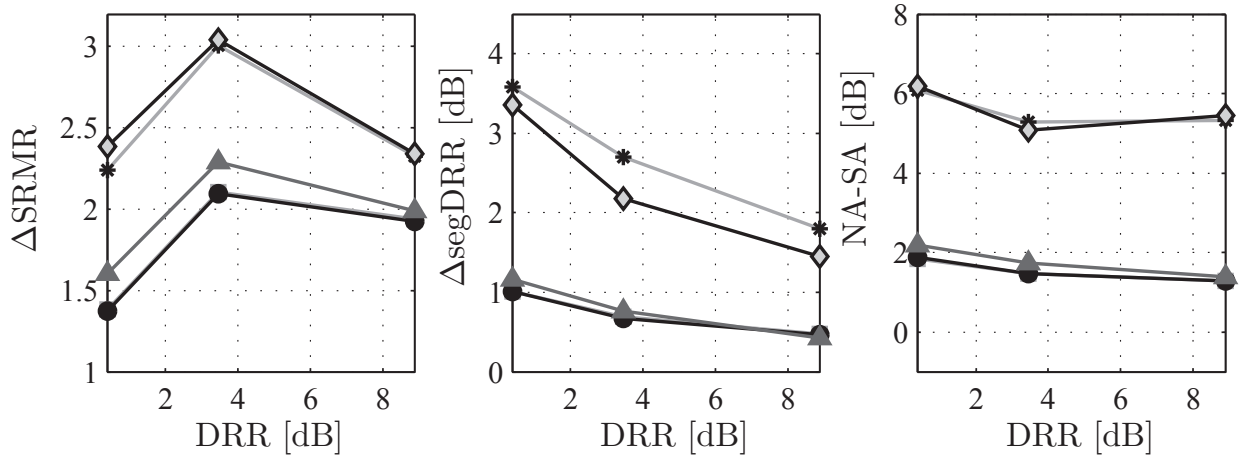
Algorithm	Parameter description	Setting
all	Smoothing factor	$\alpha^{(\mathrm{PSD})} = 0.9$
MC	Threshold	$\Gamma_{\max} = 0.99$

It is calculated by means of a gammatone filterbank analysis of temporal envelopes of the speech signal and shows a good correlation with subjective ratings of the overall speech quality and intelligibility. The considered ΔSRMR indicates an improvement compared to the reverberant speech if the value is positive. Furthermore, the increase in DRR (ΔsegDRR) as well as the noise attenuation (NA) minus speech attenuation (SA) measure (NA-SA) is used. Please refer to App. C for a detailed definition of the quality measures. All results were confirmed by informal listening tests. Main simulation settings are given in Table 3.2.

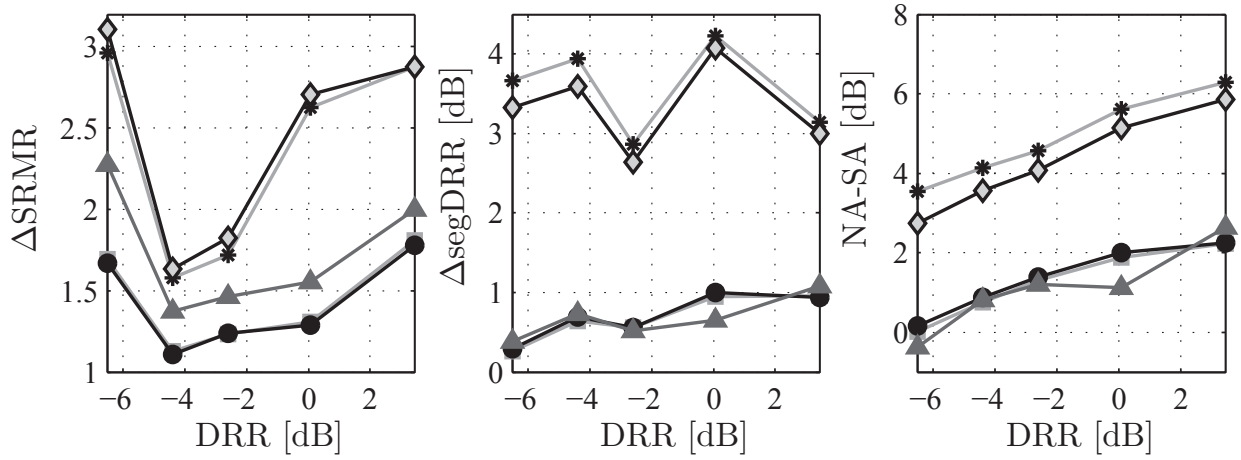
The results of the five considered coherence-based dereverberation algorithms:

- AL(I): Eq.(3.16) [ABB77],
- AL(II): Eq.(3.17) [ABB77],
- ZE: incoherent noise field; Eq.(3.26) with $\Gamma_{n_1 n_2}(\Omega) = 0$ [Zel88],
- MC(I): diffuse noise field and arithmetic averaging;
  Eq.(3.26) with $\Gamma_{n_1 n_2}(\Omega) = \mathrm{sinc}\,(\Omega f_s d_{\mathrm{mic}}/c)$ [MB03],
- MC(II/*Proposed*): diffuse noise field and geometric averaging;
  Eq.(3.29) with $\Gamma_{n_1 n_2}(\Omega) = \mathrm{sinc}\,(\Omega f_s d_{\mathrm{mic}}/c)$,

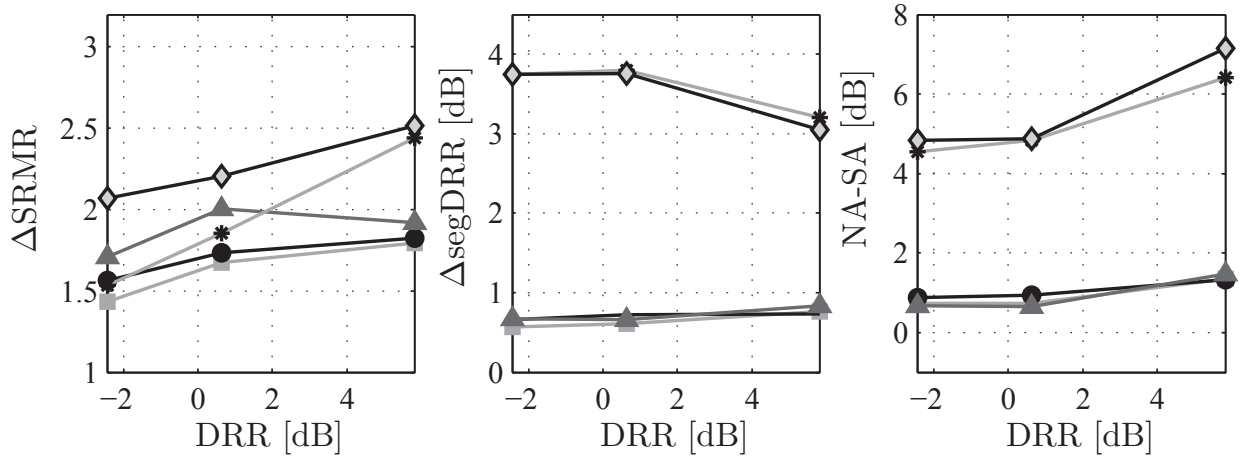
are shown in Fig. 3.10. From this experiment it can be concluded that the AL(I),



(a) Stairway ($T_{60} = 0.82\,s$).



(b) Lecture room ($T_{60} = 0.81\,s$).



(c) Office ($T_{60} = 0.51\,s$).

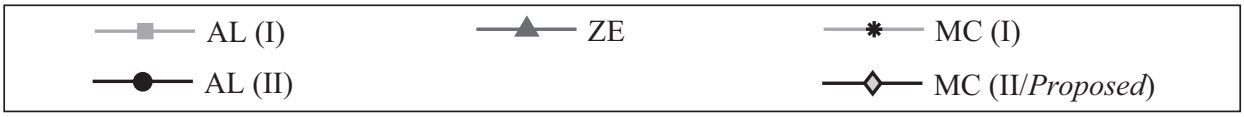


Figure 3.10: Performance comparison of coherence-based dereverberation algorithms.

AL(II) and ZE algorithm show a quite similar performance. In terms of the employed coherence model, the usage of the diffuse sound field model (MC) instead of the incoherent sound field model (ZE) increases the performance significantly. The MC algorithms MC(I) and MC(II) give the best results in terms of reverberation suppression and speech distortions. In average, the proposed geometric mean of the noise PSDs (MC(II)) leads to a better SRMR performance while achieving a lower speech attenuation.

3.2 Noise Reduction

Algorithms for the reduction of background noise are nowadays essential components in many speech communication systems. Most mobile phones and hearing aids have integrated single- or multi-microphone algorithms to enhance the speech quality in adverse environments. As for the speech dereverberation, noise reduction algorithms that are based on spectral weighting techniques require knowledge of the PSD of the interfering background noise.

In this section two novel dual-channel noise PSD estimators are presented which are derived explicitly for the application in hearing aids and mobile phones, respectively. Additionally, a new spectral weighting rule for mobile phone applications is introduced.

3.2.1 Estimation of Background Noise PSD

During the last decades, different single- and multi-microphone short-term noise PSD estimators have been proposed. A predominant single-channel algorithm is based on *Minimum Statistics* (MS) [Mar01a] which tracks the minimum of the PSD of the noisy input signal over a large window with a duration of typically 1.5 s. Disadvantages of this and related approaches are a high computational complexity, memory consumption, and the difficulties in estimating non-stationary noise. A low complexity single-channel MMSE-based algorithm that is also capable of tracking non-stationary noise has recently been proposed in [HHJ10]. The algorithm was further improved in [GH11] and is denoted as *Speech Presence Probability* (SPP) approach in the following.

Multi-channel noise PSD estimators for systems with two or more microphones have not been studied very intensively. In [DE96], a dual-channel spectral subtraction algorithm is proposed which uses the left and right signals of a binaural hearing aid system to estimate the noise PSD. However, this algorithm assumes uncorrelated noise between the different microphones which results in an underestimation of the noise PSD in realistic conditions [Ham02]. A further binaural estimator based on [DE96] was proposed in [KPB09] which is explicitly designed for binaural hearing aids. An alternative dual-channel approach [CK11] is based on a GSC beamformer and uses basically the output of the noise canceler for the noise PSD estimation.

A further overview of state-of-the-art algorithms and comparison results can be found, e.g., in [MSK97, TTM+11, HJN+11].

3.2.1.1 Noise PSD Estimation Based on Power Level Difference

The motivation for the novel PLD-based noise PSD estimator, which is termed as *Power Level Difference Noise Estimator* (PLDNE), is given by measurements of mobile devices in the BT microphone configuration (see Sec. 2.2.2). The main idea is to exploit explicitly the power level differences of the desired speech signal between the two microphones [JHN+12, Her11]. Hence, this algorithm is only suitable for applications where a certain power level difference of the desired signal between the two microphones exists, e.g., mobile phones where a secondary microphone is placed on the top of the device.

Two important assumptions are the existence of a homogeneous noise field, as well as a sufficient attenuation of the desired speech signal between the two microphones of, e.g., 10 dB.

In a first step, the normalized difference of the power spectral density $0 \leq \Delta\Phi_{\text{PLDNE}}(\lambda,\mu) \leq 1$ of the noisy input is calculated for every frequency bin μ by

$$\Delta\Phi_{\text{PLDNE}}(\lambda,\mu) = \left| \frac{\widehat{\Phi}_{x_1x_1}(\lambda,\mu) - \widehat{\Phi}_{x_2x_2}(\lambda,\mu)}{\widehat{\Phi}_{x_1x_1}(\lambda,\mu) + \widehat{\Phi}_{x_2x_2}(\lambda,\mu)} \right|. \tag{3.33}$$

All PSD values are calculated by recursive smoothing over time with constant $\alpha^{(\text{PLDNE})}$.

The idea behind the subsequent noise PSD estimation is as follows. In case of background noise-only periods, $\Delta\Phi_{\text{PLDNE}}(\lambda,\mu)$ will be close to zero as the input power levels are almost equal. If the value lies below a threshold $\phi_{\min}$, the noise PSD estimate is determined directly from the primary input signal $x_1(k)$ by recursive smoothing with $\alpha^{(\text{X1})}$ by

$$\begin{aligned} \widehat{\Phi}_{nn}(\lambda,\mu) = {} & \alpha^{(\text{X1})} \cdot \widehat{\Phi}_{nn}(\lambda-1,\mu) + (1-\alpha^{(\text{X1})}) \cdot |X_1(\lambda,\mu)|^2, \\ & \text{if } \Delta\Phi_{\text{PLDNE}}(\lambda,\mu) < \phi_{\min}. \end{aligned} \tag{3.34}$$

Regarding the noise-free case, the auto-PSD at $x_1(k)$ will be larger than at $x_2(k)$ according to Fig. 2.19 and thus, $\Delta\Phi_{\text{PLDNE}}(\lambda,\mu)$ becomes close to one. As a consequence, the updating of the noise estimate will be stopped if the difference is larger than a threshold $\phi_{\max}$, i.e.,

$$\begin{aligned} \widehat{\Phi}_{nn}(\lambda,\mu) = {} & \widehat{\Phi}_{nn}(\lambda-1,\mu) \\ & \text{if } \Delta\Phi_{\text{PLDNE}}(\lambda,\mu) > \phi_{\max}. \end{aligned} \tag{3.35}$$

In between these two extremes, a noise estimation using the secondary input signal $x_2(k)$ is used as approximation with smoothing constant $\alpha^{(\text{X2})}$ according to

$$\widehat{\Phi}_{nn}(\lambda,\mu) = \alpha^{(\text{X2})} \cdot \widehat{\Phi}_{nn}(\lambda-1,\mu) + (1-\alpha^{(\text{X2})}) \cdot |X_2(\lambda,\mu)|^2. \tag{3.36}$$

The approximation is feasible since the highly attenuated speech components in $x_2(k)$, which lead to a low SNR, can be neglected. In situations with babble noise, it is beneficial to combine the PLDNE algorithm with further single- or dual-channel noise PSD estimators, e.g., [Mar01a, HHJ10, JNK+11] instead of keeping the last estimate in Eq.(3.35).

3.2.1.2 Noise PSD Estimation Based on Coherence

In the following, a new generalized dual-channel noise PSD estimator which uses knowledge of the noise field coherence [JNK+11] is derived. It turns out that the approach of [DE96] can be seen as a special case for an uncorrelated background noise assumption. The novel algorithm has a low computational complexity and can be combined with different speech enhancement systems. The derivation is related to the discussions in Sec. 3.1.2, where an estimator for the *speech PSD* based on the noise field coherence was derived and incorporated in a Wiener filter rule for the reduction of diffuse background noise and reverberation. The main advantage of the following approach is a *noise PSD* estimate for versatile application in any spectral noise reduction rule.

Based on Eqs.(3.19),(3.20),(3.21), the arithmetic mean of the input PSDs can be expressed by Eq.(3.27) which can be rearranged to

$$\Phi_{ss}(\lambda,\mu) = \frac{1}{2}\left(\Phi_{x_1x_1}(\lambda,\mu) + \Phi_{x_2x_2}(\lambda,\mu)\right) - \Phi_{nn}(\lambda,\mu). \tag{3.37}$$

By reordering Eq.(3.21) to

$$\Phi_{ss}(\lambda,\mu) = \Phi_{x_1x_2}(\lambda,\mu) - \Phi_{nn}(\lambda,\mu)\,\Gamma_{n_1n_2}(\Omega) \tag{3.38}$$

and by combining it with Eq.(3.37) leads to an estimate of the background noise PSD

$$\widehat{\Phi}_{nn}^{(\mathrm{I})}(\lambda,\mu) = \frac{\frac{1}{2}\left(\widehat{\Phi}_{x_1x_1}(\lambda,\mu) + \widehat{\Phi}_{x_2x_2}(\lambda,\mu)\right) - \widehat{\Phi}_{x_1x_2}(\lambda,\mu)}{1-\Gamma_{n_1n_2}(\Omega)}. \tag{3.39}$$

Again, the alternative derivation with the geometric mean PSD can be employed which leads to [JNK+11]

$$\widehat{\Phi}_{nn}^{(\mathrm{II})}(\lambda,\mu) = \frac{\sqrt{\widehat{\Phi}_{x_1x_1}(\lambda,\mu)\cdot\widehat{\Phi}_{x_2x_2}(\lambda,\mu)} - \widehat{\Phi}_{x_1x_2}(\lambda,\mu)}{1-\Gamma_{n_1n_2}(\Omega)}. \tag{3.40}$$

For both estimators, the usage of the theoretical ideal diffuse sound field according to Eq.(2.6) for $\Gamma_{n_1n_2}(\Omega)$ is suggested. When it comes to the application in a binaural context, the more accurate binaural noise field model introduced in Sec. 2.2.1.1 should be employed. It can be seen that for the special case of an uncorrelated noise field, i.e., $\Gamma_{n_1n_2}(\Omega) = 0$, the estimator where the geometric mean PSD is utilized, reduces to the approach of [DE96].

3.2.2 Spectral Weighting Rules

Based on the estimated PSD of the interfering background noise, several spectral weighting rules can be employed. The most important single-channel spectral subtraction and Wiener filter rules were already introduced in Sec. 3.1.1.4.

A new dual-channel method which is introduced in the following is motivated by the PLD algorithm initially proposed in [YAR09]. Here, an alternative calculation of the spectral gains is derived which leads to a much better noise reduction performance [JHN+12]. The algorithm also exploits the power level difference between both microphones and is therefore explicitly suitable for BT mobile phones in the *Hand-Held Position* (HHP).

It is again assumed that the power levels are equal for noise whereas speech results in a higher PSD at microphone $x_1(k)$. The auto-PSDs of the input signals are given by

$$\Phi_{x_1x_1}(\lambda,\mu) = \Phi_{s_1s_1}(\lambda,\mu) + \Phi_{n_1n_1}(\lambda,\mu), \tag{3.41}$$
$$\Phi_{x_2x_2}(\lambda,\mu) = \Phi_{s_2s_2}(\lambda,\mu) + \Phi_{n_2n_2}(\lambda,\mu). \tag{3.42}$$

By introducing a relative transfer function $H_{12}(\lambda,\mu)$ of the desired speech signal between the two microphones, the auto-PSD at the secondary microphone can be expressed by

$$\Phi_{s_2s_2}(\lambda,\mu) = |H_{12}(\lambda,\mu)|^2 \cdot \Phi_{s_1s_1}(\lambda,\mu) \tag{3.43}$$
$$\Phi_{x_2x_2}(\lambda,\mu) = |H_{12}(\lambda,\mu)|^2 \cdot \Phi_{s_1s_1}(\lambda,\mu) + \Phi_{n_2n_2}(\lambda,\mu). \tag{3.44}$$

Two difference equations for the auto-PSD of the noisy input and the noise-only signals are introduced as

$$\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu) = \Phi_{x_1x_1}(\lambda,\mu) - \Phi_{x_2x_2}(\lambda,\mu), \tag{3.45}$$
$$\Delta\Phi_{nn}(\lambda,\mu) = \Phi_{n_1n_1}(\lambda,\mu) - \Phi_{n_2n_2}(\lambda,\mu). \tag{3.46}$$

The power level difference of the noisy input signal can thus be expressed as

$$\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu) = \Phi_{s_1s_1}(\lambda,\mu)(1 - |H_{12}(\lambda,\mu)|^2) + \Delta\Phi_{nn}(\lambda,\mu). \tag{3.47}$$

Due to the assumption of a homogeneous noise field, the difference $\Delta\Phi_{nn}(\lambda,\mu)$ can be neglected, i.e., $\Delta\Phi_{nn}(\lambda,\mu) \approx 0$. Hence, the equation for the PLD reads

$$\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu) \approx (1 - |H_{12}(\lambda,\mu)|^2) \cdot \Phi_{s_1s_1}(\lambda,\mu). \tag{3.48}$$

The final spectral weighing rule is the Wiener filter equation

$$G(\lambda,\mu) = \frac{\Phi_{s_1s_1}(\lambda,\mu)}{\Phi_{s_1s_1}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu)}. \tag{3.49}$$

In analogy to [YAR09], nominator and denominator in Eq.(3.49) are expanded by $1 - |H_{12}(\lambda,\mu)|^2$. Finally, by using Eq.(3.48), the weighting rule reads

$$G^{(\mathrm{PLD})}(\lambda,\mu) = \frac{\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu)}{\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu) + \gamma(1 - |H_{12}(\lambda,\mu)|^2) \cdot \Phi_{nn}(\lambda,\mu)}. \tag{3.50}$$

To counteract possible overestimations of the transfer function $H_{12}(\lambda,\mu)$, the correction factor γ is introduced [YAR09]. In the case of speech absence, $\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu)$ will

be zero and hence, the gains will be zero, too. When there is pure speech the right part of the denominator of Eq.(3.50) will be zero. Thus the gains $G^{(\mathrm{PLD})}(\lambda,\mu)$ will turn to one.

The required transfer function $H_{12}(\lambda,\mu)$ is derived from the cross-PSD of the noisy input $\Phi_{x_1x_2}(\lambda,\mu)$. In [YAR09], the cross-PSD is expressed by

$$\Phi_{x_1x_2}(\lambda,\mu) = H_{12}(\lambda,\mu)\cdot\Phi_{x_1x_1}(\lambda,\mu) + \Phi_{n_1n_2}(\lambda,\mu), \tag{3.51}$$

and the transfer function is given by

$$H_{12}(\lambda,\mu) = \frac{\Phi_{x_1x_2}(\lambda,\mu) - \Phi_{n_1n_2}(\lambda,\mu)}{\Phi_{x_1x_1}(\lambda,\mu) - \Phi_{nn}(\lambda,\mu)}. \tag{3.52}$$

The required cross-PSD of the background noise $\Phi_{n_1n_2}(\lambda,\mu)$ is calculated in [YAR09] from the first 400 ms where no speech activity is assumed.

In contrast to Eq.(3.51), in the proposed implementation the cross-PSD is correctly expressed by

$$\Phi_{x_1x_2}(\lambda,\mu) = H_{12}(\lambda,\mu)\cdot\Phi_{s_1s_1}(\lambda,\mu) + \Phi_{n_1n_2}(\lambda,\mu). \tag{3.53}$$

By incorporating the coherence of the noise field $\Gamma_{n_1n_2}(\Omega)$, the cross-PSD reads with

$$\Phi_{s_1s_1}(\lambda,\mu) = \Phi_{x_1x_1}(\lambda,\mu) - \Phi_{nn}(\lambda,\mu) : \tag{3.54}$$

$$\Phi_{x_1x_2}(\lambda,\mu) = H_{12}(\lambda,\mu)\cdot(\Phi_{x_1x_1}(\lambda,\mu) - \Phi_{nn}(\lambda,\mu)) + \Gamma_{n_1n_2}(\Omega)\cdot\Phi_{nn}(\lambda,\mu). \tag{3.55}$$

Hence, the proposed transfer function is given by

$$H_{12}(\lambda,\mu) = \frac{\Phi_{x_1x_2}(\lambda,\mu) - \Gamma_{n_1n_2}(\Omega)\cdot\Phi_{nn}(\lambda,\mu)}{\Phi_{x_1x_1}(\lambda,\mu) - \Phi_{nn}(\lambda,\mu)}. \tag{3.56}$$

With Eq.(3.56), the computation of the transfer function does not require an additional calculation of the noise cross-PSD anymore and allows the algorithm to cope with non-stationary noise and changing SNR conditions compared to [YAR09]. In the practical implementation, the power level difference is proposed to be calculated by

$$\Delta\Phi_{\mathrm{PLD}}(\lambda,\mu) = \max\left\{\Phi_{\mathrm{x_1x_1}}(\lambda,\mu) - \Phi_{\mathrm{x_2x_2}}(\lambda,\mu), 0\right\}, \tag{3.57}$$

which prevents speech distortions if the assumption of a homogeneous noise field is violated, e.g., due to an interfering talker. Furthermore, a lower threshold $G_{\min}$ for the weighting function is employed. In the practical realization, all required PSDs are estimated by means of recursive smoothing. The employed sound field coherence model $\Gamma_{n_1n_2}(\Omega)$ can be chosen freely as in the coherence-based noise PSD estimator.

Please note that in case when one microphone is covered by a finger of the phone user, the difference $\Delta\Phi_{nn}(\lambda,\mu)$ cannot be neglected as in Eq.(3.48). In this case, the background noise exhibits a certain power level difference between the microphones as well and no noise reduction can be performed. Since this does not cause additional audible speech attenuation, the influence of a possible microphone covering is negligible.

Table 3.3: Main simulation parameters of background noise PSD estimation algorithms.

Algorithm	Parameter description	Setting
all	Constant used for ideal PSD	$\alpha^{(\text{ideal})} = 0.9$
PLDNE	Thresholds	$\phi_{\min} = 0.2$, $\phi_{\max} = 0.8$
PLDNE	Smoothing factors	$\alpha^{(\text{PLDNE})} = 0.9$, $\alpha^{(\text{X1})} = 0.8$, $\alpha^{(\text{X2})} = 0.9$

3.2.3 Performance Evaluation

3.2.3.1 Estimation Accuracy: Background Noise PSD Estimators

The evaluation section of the novel noise PSD estimators is subdivided into two separate experiment paragraphs since both algorithms are explicitly developed for specific acoustic conditions.

a) Noise PSD Estimation Based on Power Level Difference
This first paragraph compares the performance of the novel PLD-based noise PSD estimator (PLDNE) in the mobile-phone BT configuration (see Fig. 2.15). As a reference, three state-of-the-art single-channel noise estimators, which work on the primary signal $x_1(k)$ only, are considered:

- MS: Minimum Statistics [Mar01a],
- MMSE: MMSE-based noise tracker [HHJ10],
- SPP: Speech Presence Probability-based [GH11].

It has to be mentioned that the coherence-based estimator presented in Sec. 3.2.1.2 was mainly developed for binaural hearing aids with a dual-channel signal processing. Hence, this algorithm is evaluated in separate simulations. Besides, the reference algorithms use only a single-channel input because no alternative dual-channel method is known from literature which is capable of achieving a similar performance than the considered state-of-the-art single-channel approaches.

The performance is rated in terms of the symmetric segmental logarithmic estimation error between the ideal noise PSD $\Phi_{nn}(\lambda, \mu)$ and the estimated noise PSD $\widehat{\Phi}_{nn}(\lambda, \mu)$. The ideal noise PSD is obtained using the true noise periodograms smoothed over time λ with smoothing factor $\alpha^{(\text{ideal})}$. In order to rate over- and underestimations separately, Eqs.(3.30),(3.31) are applied [GH11]. The main simulation parameters are shown in Table 3.3 and the results for pub and traffic noise are depicted in Fig. 3.11.

It can be seen that the novel algorithm shows a better overall performance to the MMSE and SPP algorithm and outperforms MS for all SNR conditions significantly. Regarding the overall performance and low computational complexity of the PLDNE algorithm, this method is a key component for a novel speech enhancement system for mobile phones.

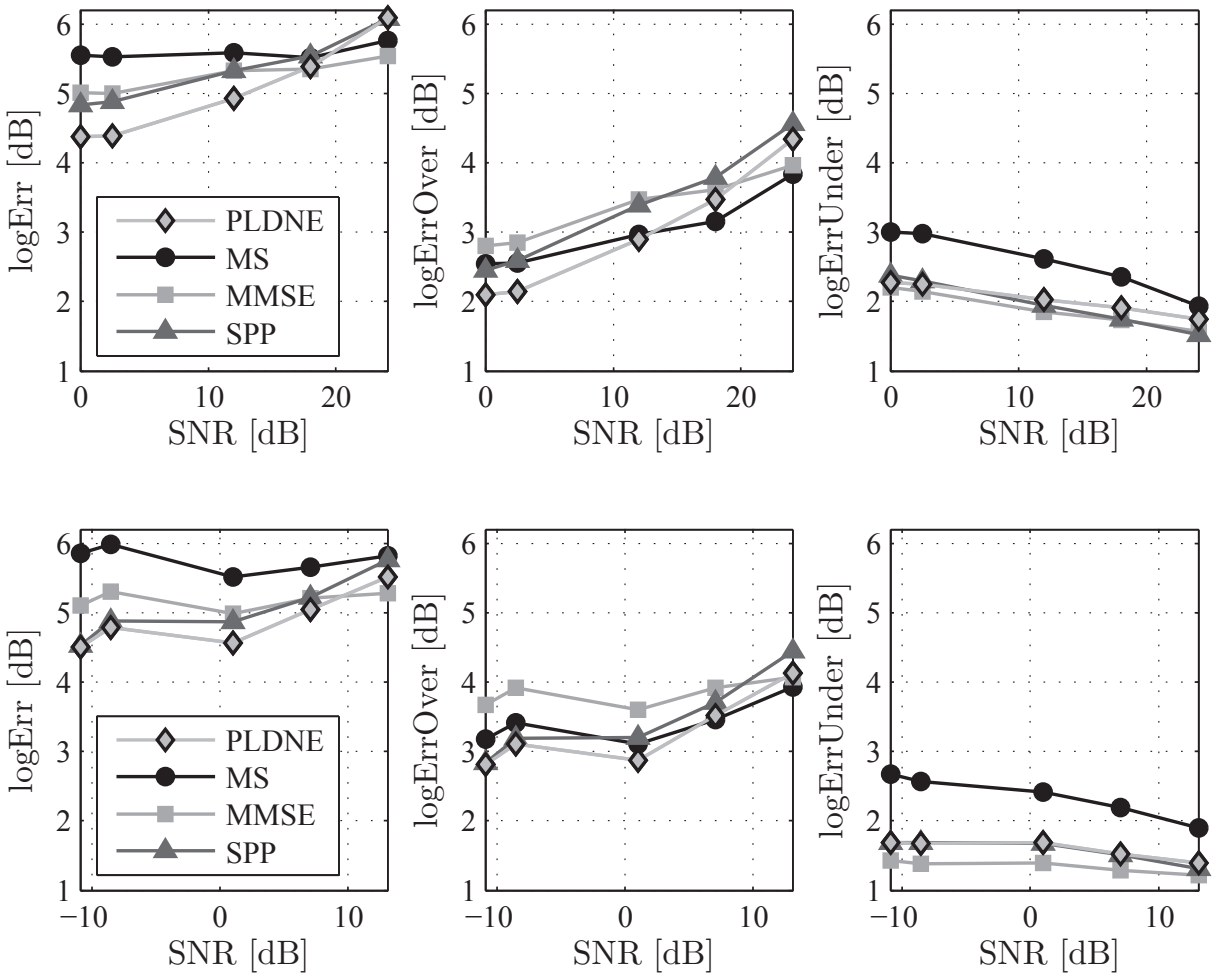


Figure 3.11: Noise PSD estimation error for two different noise sources: (top) pub noise, (bottom) traffic noise. Plotted are the overall estimation errors (left), overestimations (middle) and underestimations (right).

b) Noise PSD Estimation Based on Coherence

The second paragraph evaluates the proposed generalized coherence-based noise PSD estimator (GCoh) for the applications in binaural hearing aids. The new method with both arithmetic and geometric PSD averaging ((GCoh I) and (GCoh II)) is compared to the reference noise PSD estimator (Coh) [DE96], as well as to the two single-channel noise PSD estimators MS and MMSE. For the sake of brevity, only the overall logarithmic estimation error (Eq.(3.32)) is considered.

Since the proposed generalized concept of the noise PSD estimator is capable of employing arbitrary coherence models (see Eqs.(3.39),(3.40)), we investigate the algorithm under two special cases assuming an uncorrelated noise field as in [DE96] and an ideal diffuse noise field using Eq.(2.6).

First, generated dual-channel signals are computed using the approach of [HCG08], where predefined spatial coherence constraints and hence, different microphone distances d_{mic}, can be employed. Speech samples from the TSP speech database [Kab02] are degraded with additive noise from the ETSI background noise database [ETS09]. Such simulation ensures reproducible results and the employment of objective evalu-

Table 3.4: Main simulation parameters of coherence-based background noise PSD estimation algorithms.

Algorithm	Parameter description	Setting
Coh	Smoothing factors	$\alpha = 0.9, \alpha_{\mathrm{DD}} = 0.98, \alpha_{\mathrm{nn}} = 0.9$
GCoh	Smoothing factors	$\alpha = 0.9, \alpha_{\mathrm{DD}} = 0.98, \alpha_{\mathrm{nn}} = 0.9$

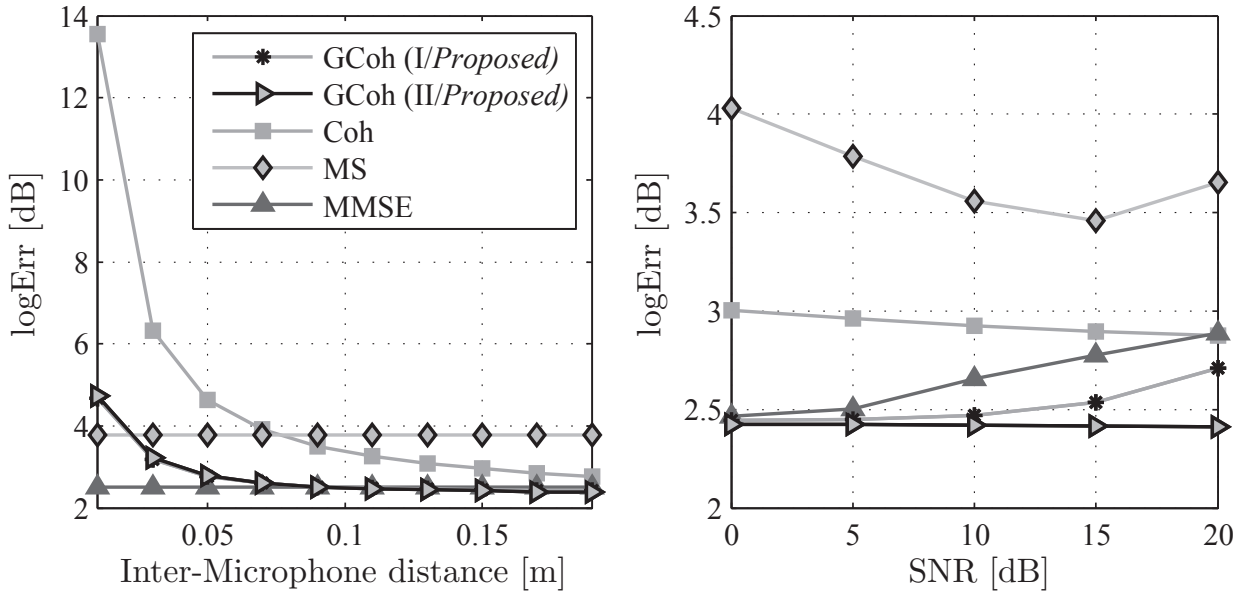


Figure 3.12: Noise PSD estimation error for generated pub noise: (left) fixed SNR of 5 dB over microphone distance, (right) fixed inter-microphone distance $d_{\mathrm{mic}} = 0.1\,\mathrm{m}$ over the SNR.

ation measures, especially as the coherence of realistic noise fields such as a cocktail-party or office environment can be modeled by a diffuse noise field.

Further simulation parameters are listed in Table 3.4.

The left subfigure in Fig. 3.12 shows the logErr at a varying inter-microphone distance and fixed SNR of 5 dB using generated background noise signals. The results in terms of fixed inter-microphone distance of 0.10 m over a varying SNR are illustrated in the right subfigure. It can clearly be seen that the proposed GCoh estimators outperform the dual-channel estimator Coh in terms of a lower estimation error. Since the algorithm Coh assumes uncorrelated noise, the estimation error is reduced for larger microphone distances where the correlation of the noise field becomes lower for lower frequencies. However, even the proposed algorithm requires a microphone distance larger than a specific minimum which is determined by the real coherence characteristics. The proposed algorithm outperforms MS and MMSE for inter-microphone distances $> 2\,\mathrm{cm}$ and $> 10\,\mathrm{cm}$, respectively. The curves show similar results as the MMSE approach for low SNR conditions and a better performance for high SNR conditions, with the expense of an additional microphone, but with the additional benefit of a lower computational complexity. In a [JNK+11] the computational complexity was shown to be $15 - 20\,\%$ lower for the GCoh method compared to the MMSE

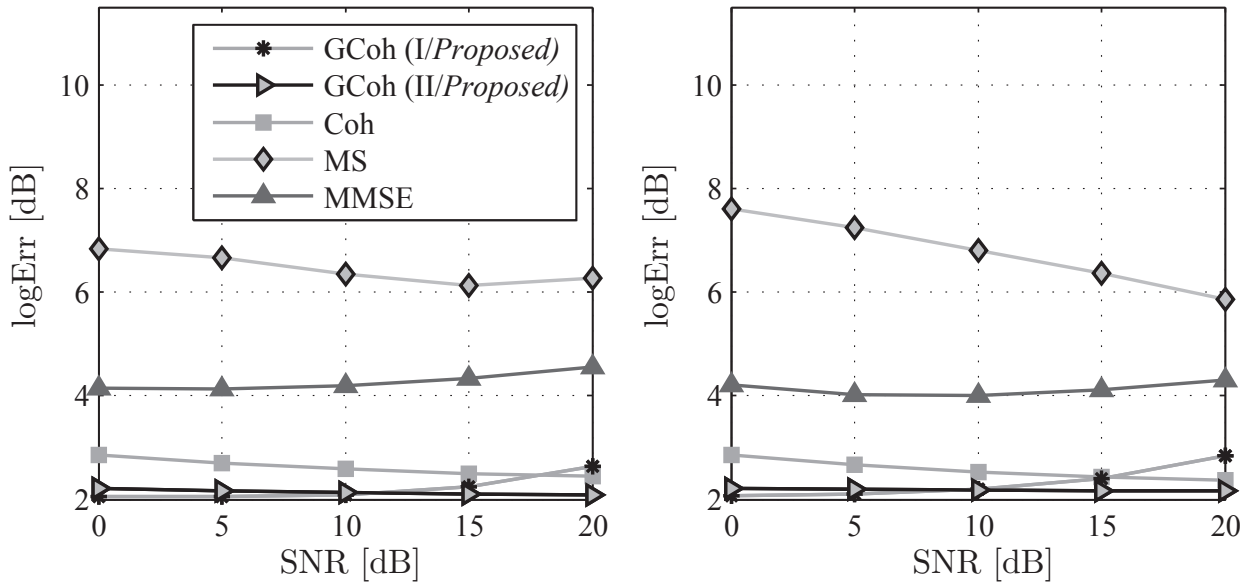


Figure 3.13: Noise PSD estimation error for two different recorded noise sources from the ETSI database [ETS09] at a fixed distance $d_{\text{mic}} = 0.15\,\text{m}$: (left) cafeteria noise, (right) kindergarten noise.

algorithm.

To verify the results of the generated background noise with real recordings, binaural babble noise from the ETSI background noise database [ETS09] is used. The signals are measured with two microphones of a dummy head at a fixed inter-microphone $d_{\text{mic}} = 0.15\,\text{m}$. This results are given in Fig. 3.13 for cafeteria noise (left) and kindergarten noise (right). Due to the inter-microphone spacing of 15 cm, the GCoh algorithms outperforms all other methods, independent of the input SNR. Regarding the choice of the averaging, the use of the geometric mean (GCoh II) is favorable, which is consistent to the findings in Sec. 3.1.4.2. A deeper analysis has shown that the coherence-based noise PSD estimator mainly causes overestimations in the low frequency region where both noise and speech are highly correlated (see Fig. 2.3 (right)). Please note that coherence-based noise estimators work in general not for purely coherent noise sources and show a larger estimation variance at lower frequencies.

3.2.3.2 Noise Reduction Performance

The PLD spectral weighting rule is evaluated for the target application of BT mobile phones used in the *Hand-Held Position* (HHP). Hence, for the required noise PSD estimator, the PLDNE algorithm (see Sec. 3.2.1.1) is taken into account. We investigate the PLD spectral weighting rule assuming an ideal diffuse noise field with $d_{\text{mic}} = 0.1\,\text{m}$ in Eq.(3.56). The speech and noise signals are recorded with the dual-microphone BT mock-up as illustrated in Fig. 2.15 (right). Background noise signals are recorded inside an acoustic chamber using the standardized multi-loudspeaker procedure described in [ETS09] and speech signal are emitted by the dummy head mouth. The performance of the PLD weighting rule (PLD (*Proposed*)) using Eq.(3.56) is compared with the original implementation by [YAR09] (PLD (Original)) using Eq.(3.52) and

Table 3.5: Main simulation parameters of noise reduction algorithms.

Algorithm	Parameter description	Setting
all	Gain thresholds	$G_{\min} = 0.1$
PLD	Overestimation factor	$\gamma = 4$
WF	DDA smoothing factor	$\alpha^{(\mathrm{DDA})} = 0.98$

also a single-channel (SC) Wiener filter with decision-directed approach for the a priori SNR calculation (Wiener). All algorithms use the PLDNE noise estimator. Furthermore, the *Dual-Microphone Spectral Subtraction* (DMSS) algorithm, which is distinctly developed for BT mobile phones [GCNL00] is evaluated. Two common spectral subtraction approaches provide a rough speech and noise estimate for each channel by using the other channel respectively. In a following step these estimates are used by a third spectral subtraction stage which results in the enhanced output. See Table 3.5 for the main simulation settings. The parameters for the DMSS algorithm are chosen as suggested in [GCNL00].

The noise reduction performance is determined by means of the noise attenuation minus speech attenuation (NA-SA) measure, where higher values indicate an improvement. The *Speech Intelligibility Index* (SII) [ANS07] was calculated from the noisy as well as the enhanced signal. An SII higher than 0.75 indicates a good communication system and values below 0.45 correspond to a poor system. The averaged results for traffic and pub noise are shown in Fig. 3.14.

From the plots, it can be concluded that the proposed PLD spectral weighting rule outperforms related approaches in terms of a much higher noise reduction performance and a marginal increase in speech intelligibility. The modifications of the original PLD implementation also result in a high performance gain. All results are consistent with the subjective listening impression where the highest amount of musical tones was observed for the DMSS algorithm. Since for babble noise the major frequency components of the noise signal lie in the same regions as those of the desired speech signal, this scenario can be seen as the most difficult one. However, all experiments have also been conducted with train station noise where the same tendency has been observed.

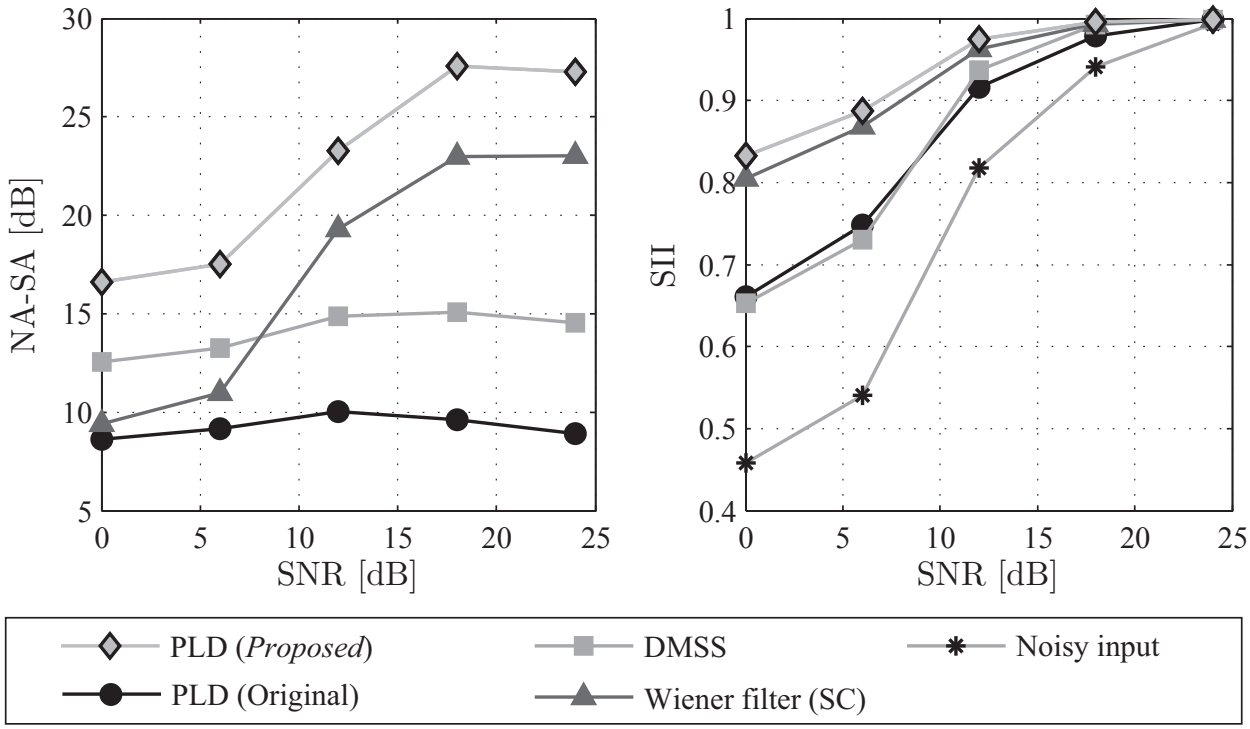


Figure 3.14: Simulation results of different weighting rules: (left) noise reduction performance, (right) influence on intelligibility. NA-SA: noise attenuation minus speech attenuation, SII: speech intelligibility index. Averaged over traffic and pub noise. All algorithms use the PLDNE noise estimator.

3.3 Joint Dereverberation and Noise Reduction

Until now the reduction of room reverberation and background noise was considered independently. Since in realistic enclosures room reverberation and background noise can occur simultaneously, the speech enhancement algorithm should be capable of reducing both interfering signals jointly.

The considered signal model is given by

$$x_m(k) = s_m(k) * h_m(k) + n_m(k), \tag{3.58}$$

where the reverberant speech is degraded by additional background noise $n_m(k)$. The resulting *Signal-to-Interference Ratio* (SIR) is defined as the ratio of the (desired) direct speech power to the sum of early reverberation, late reverberation and background noise power. The motivation for the subsequent experiments in Secs. 3.3.1, 3.3.2 is to show how reverberation and background noise affects the accuracy of RT, DRR and interfering PSD estimation algorithms. Concepts how to tackle the occurring limitations are presented in Sec. 3.3.3.

3.3.1 Influence of Noise on RT and DRR Estimation

Most dereverberation algorithms are based on the spectral subtraction principle and require knowledge of the RT and DRR in order to estimate the PSD of the late reverberant speech. Hence, it is of special interest to evaluate the estimation accuracy in noisy and reverberant conditions.

Figure 3.15 illustrates the influence of additional cafeteria and kindergarten noise from the ETSI background noise database on the RT estimation accuracy using the ML approach which is discussed in more detail in Sec. 2.3.3. Since the impulse responses of these rooms are not available, we use RIRs from the AIR database of the lecture room and corridor as an approximation. Please note that anechoic speech signals from the TSP database are first convolved with RIRs taken from the AIR database. Afterwards, the reverberant signals are summed with background noise from the ETSI database according to the signal model in Eq.(3.58).

It can be seen in the left subfigure that the cafeteria noise causes a severe RT overestimation for SNRs $< 7\,\mathrm{dB}$. In contrast to that, the influence of the kindergarten noise can be neglected since it mainly consists of transient disturbances, which are harmless to the estimator and small diffuse components. The right subfigure depicts the results for the corridor location in the HHP, where a large underestimation is expected due to the high DRR. However, even in this configuration, a vast overestimation can be observed for negative SNRs. It is proposed to perform a pre-denoising in low SNR conditions before estimating the RT, even though a pre-denoising might result in an underestimation.

A similar experiment evaluates the influence of background noise on the DRR estimation accuracy using the dual-channel approach presented in Sec. 2.3.4. It turns

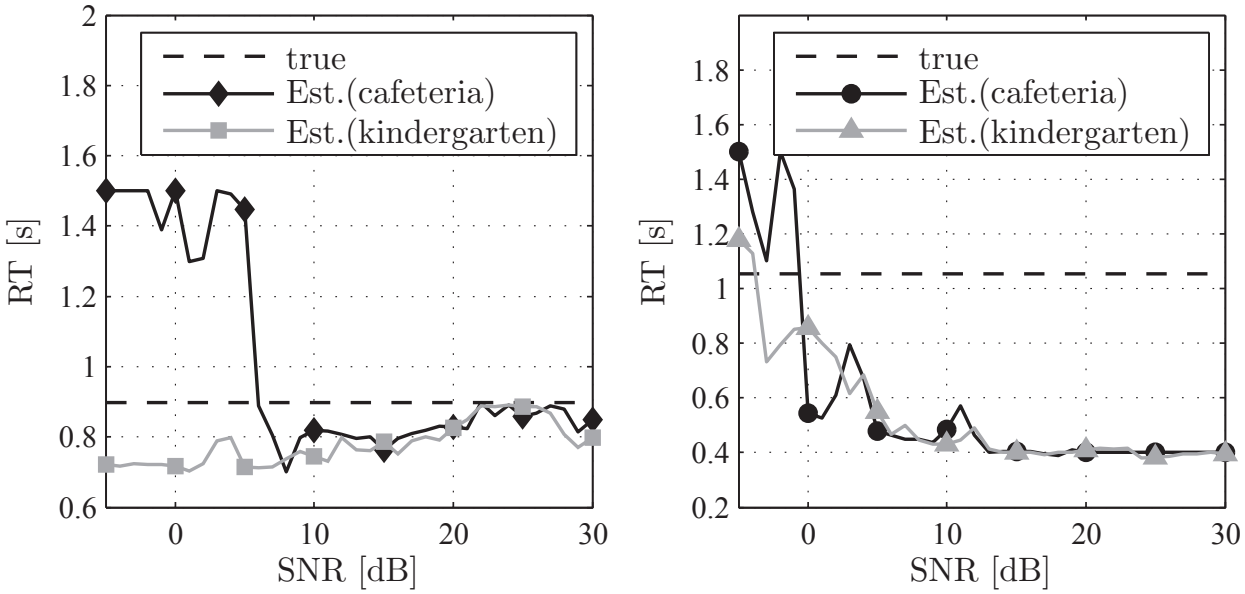


Figure 3.15: Influence of background noise on RT estimation performance: (left) lecture room, (right) corridor (HHP).

out that in noisy and reverberant conditions, the estimator mainly underestimates the true DRR as shown in Fig. 3.16 (left). Especially the diffuse noise components of both noise types increase the reverberant energy components and hence, result in an underestimated DRR. The effect is less distinctive for very high DRR for the corridor/HHP impulse response given in the right subfigure. Here, even a slight overestimation was observed. Please note that in additional experiments with WGN, the accuracy tends to underestimate the true DRR, independent of the DRR. In conclusion, it is shown by this and further experiments that diffuse background noise mainly causes underestimations, whereas background noise which is coherent between the microphones such as an interfering talker, results in overestimations. A possible pre-denoising of the input signals for the DRR estimator has to take this into account since, e.g., a strong reduction of (the incoherent noise components of) diffuse noise would result in an overestimation. Thus, as a compromise it is suggested to avoid any pre-processing before estimating the DRR.

3.3.2 Influence of Reverberation on Noise PSD Estimation

In this section it will be shown that additional room reverberation leads to high overestimations of state-of-the-art background noise PSD estimators. The ideal background noise PSD is obtained from the noisy speech (w/o reverberation) and is compared to the estimated noise PSD from the noisy and reverberant input signals using MS and MMSE.

Figure 3.17 shows the increase in the logERR due to room reverberation over the DRR, exemplarily for the lecture room and cafeteria noise. The error ΔlogERR is defined as the difference of the estimation error given the noisy speech

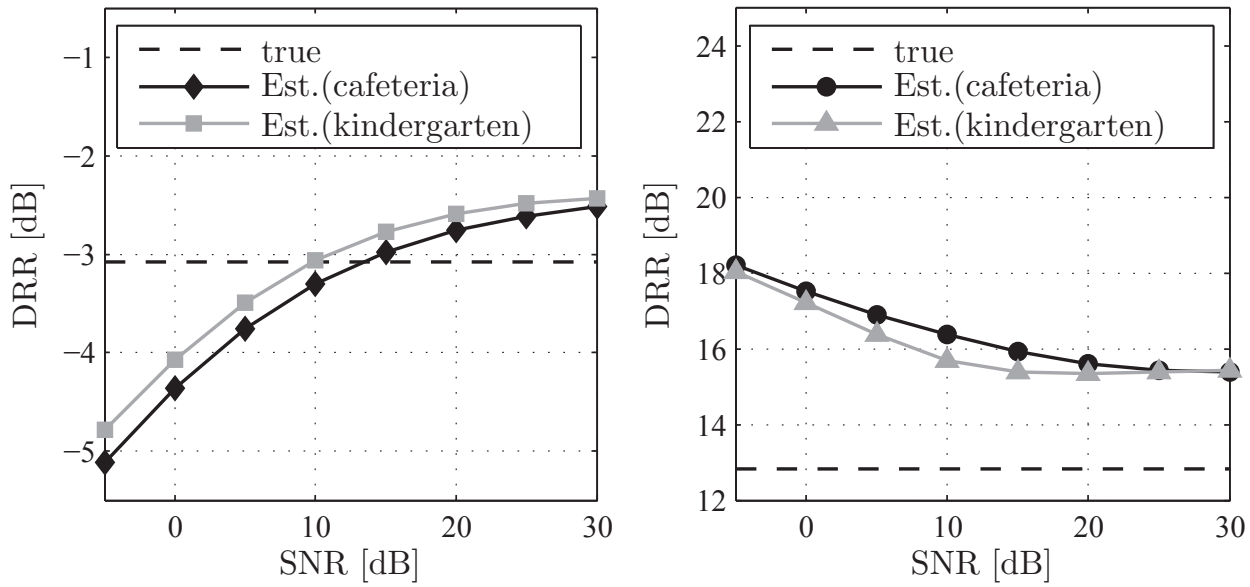


Figure 3.16: Influence of diffuse background noise on DRR estimation performance: (left) lecture room, (right) corridor (HHP).

$x(k) = s(k) + n(k)$ and given the noisy and reverberant speech $x(k) = s(k) * h(k) + n(k)$. The ideal noise PSD is always referred to the background noise signal only. Hence, positive values indicate overestimations due to additional reverberation. It is obvious that additional reverberation leads to overestimations of the background noise especially for negative DRR values even though the additional reverberation does not amplify the background noise components or influences the SNR. The same tendency was observed for the late reverberant speech PSD estimators when the input signal was noisy and reverberant. Hence, the two estimators cannot be regarded independently and a summation of both estimates leads to a severe overestimation and a high amount of speech distortions.

Possible solutions, e.g., for MS are the extension of the tracking window length to a multiple of the reverberation time. This can be explained with the principle of MS which tries to track the minimum of a long tracking window and assumes that the PSD will decline down to the noise floor. For strong reverberation the PSD will never decline down to the noise floor and a high overestimation is caused. The extension of the tracking window length, however, results in an increased memory consumption and the capability to track changes in the background noise will highly be degraded.

3.3.3 Proposed Concept

An intuitive approach for a reduction of reverberation and background noise is a subsequent application of individual speech enhancement algorithms as depicted by Fig. 3.18. For the sake of brevity, only the single-channel case is illustrated in the figures and please note that the novel DRR estimator requires a dual-channel input signal.

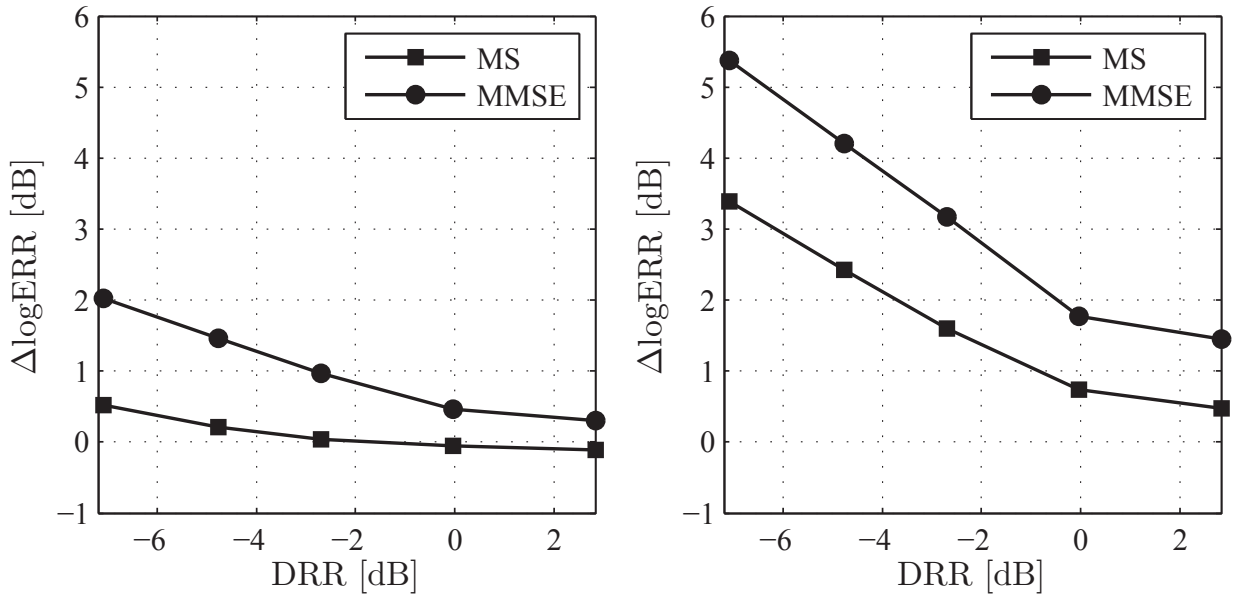


Figure 3.17: Influence of reverberation on background noise PSD estimation performance using cafeteria noise and RIRs from the lecture room: (left) 0 dB SNR, (right) 10 dB SNR.

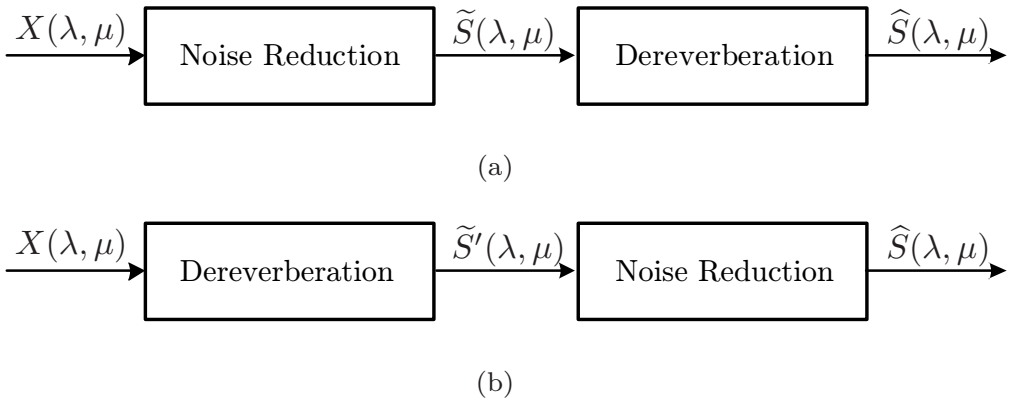


Figure 3.18: Joint dereverberation and noise reduction concepts: (a)/(b) subsequent noise reduction and dereverberation.

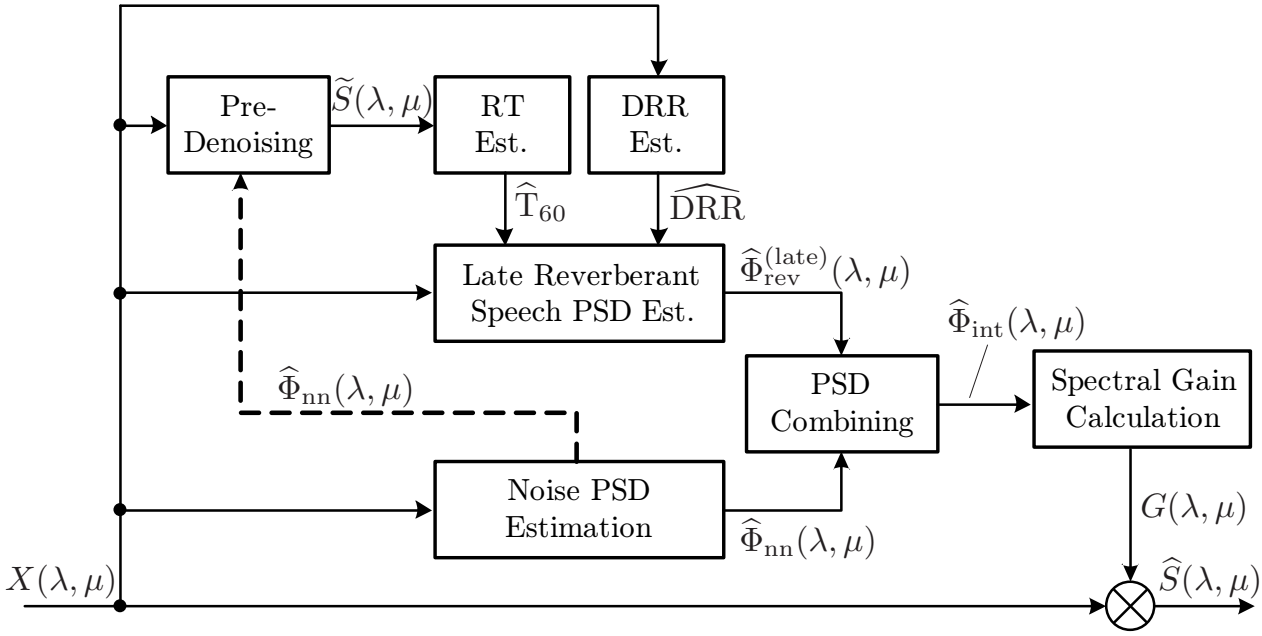


Figure 3.19: Joint dereverberation and noise reduction concepts: PSD combining and pre-denoising.

Following the previous discussion, these concepts are error-prone. On the one hand, a pre-denoising (Fig. 3.18 (a)) might attenuate the speech signal due to overestimations of the background noise PSD. On the other hand, the required estimates for the dereverberation algorithms, i.e., RT and DRR, might be influenced by a pre-denoising as well. The alternative concept of a dereverberation following a subsequent noise reduction (Fig. 3.18 (b)) has the major limitation that, again, the estimations might be biased.

The idea of an alternative concept which is proposed in Fig. 3.19 is to estimate the noise and late reverberant speech PSD individually and to combine the two estimates in different ways. Moreover, only the RT estimation module takes advantage of a pre-denoised signal. The combination of the interfering PSDs $\widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu)$ and $\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)$ to the overall PSD $\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu)$ can be performed, e.g., by

$$\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu) + \widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu), \tag{3.59a}$$

$$\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \frac{1}{2}\left(\widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu) + \widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)\right), \tag{3.59b}$$

$$\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \sqrt{\widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu)\cdot\widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)}, \tag{3.59c}$$

$$\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \min\left\{\widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu), \widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)\right\}, \tag{3.59d}$$

$$\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \max\left\{\widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu), \widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)\right\}. \tag{3.59e}$$

The choice of the combination method should be selected carefully, dependent on properties the incorporated noise PSD estimator. Since both estimators are biased by reverberation and background noise, respectively, taking the maximum of both estimators as in Eq.(3.59e) is beneficial and lowers the occurring overestimations compared to the intuitive summation using Eq.(3.59a).

In general it is not fully understood to what extent reverberation and/or background noise should be removed in order to increase the intelligibility. In this thesis, it is stated that in very noisy conditions with a negative SNR, the effects of reverberation can be neglected. This was confirmed by listening experiments with a varying SNR and different background noise types. Even a pre-denoising for the RT estimation leads to high overestimations and hence, if SNR $< 0\,\mathrm{dB}$, the PSD combination block takes only the dominant $\widehat{\Phi}_{\mathrm{nn}}(\lambda, \mu)$ into account.

An evaluation and the application of the discussed concept for a joint dereverberation and noise reduction to binaural hearing aids is given in Sec. 4.1.

3.4 Reduction of Musical Noise

3.4.1 Smoothing of Spectral Weights

The direct application of spectral gains to the noisy and reverberant DFT coefficients can lead to various artifacts such as speech distortions or musical noise. Overestimations of the background noise PSD and late reverberant speech PSD cause audible speech distortions and musical tones. These can be reduced by applying a high spectral floor to the gains which, however, reduces the enhancement performance. In order to counteract musical noise which are mainly caused by underestimations of the interfering PSD, different methods are possible:

- Smoothing of the spectral weights in the cepstral domain over quefrency [Ger10],
- Smoothing of the spectral weights in the frequency domain over frequency [EV09].

In [JSEV10] we have shown that a gain smoothing over frequency proposed first in [EV09] can greatly improve the perceived speech quality for a speech dereverberation algorithm. The main idea of [EV09] is to reduce the annoying musical tones especially in low SIR regions requiring a reliable and robust detector. In contrast to [EV09, JSEV10], it is proposed that the gain smoothing should only be applied in a frequency region specified by $f_{\mathrm{smooth}}^{(\mathrm{min})} \leq f \leq f_{\mathrm{smooth}}^{(\mathrm{max})}$, e.g., $4\,\mathrm{kHz} \leq \mathrm{f} \leq 7\,\mathrm{kHz}$.

In order to obtain a good indication whether a frame contains speech or not, the power ratio between the pre-enhanced signal $X(\lambda, \mu) \cdot G(\lambda, \mu)$ and the input signal $X(\lambda, \mu)$ is calculated for each frame λ as [EV09]

$$\zeta(\lambda) = \frac{\sum_{\mu=0}^{L-1} |G(\lambda, \mu) \cdot X(\lambda, \mu)|^2}{\sum_{\mu=0}^{L-1} |X(\lambda, \mu)|^2}. \tag{3.60}$$

If the frame mainly contains anechoic speech (high SIR), the power of the processed frame is equal or only slightly lower to the power of the input frame, i.e., $\zeta(\lambda) \approx 1$.

By contrast, the speech enhancement system is supposed to strongly attenuate the input signal in low SIR conditions, resulting in a power ratio $\zeta(\lambda) \approx 0$. Based on $\zeta(\lambda)$, the magnitudes of the weighting gains $G(\lambda, \mu)$ of frame λ are adaptively smoothed over frequency μ using a moving average window. The odd window length $N_s(\lambda)$ is set to [EV09]

$$N_s(\lambda) = \begin{cases} 1, & \text{if} \quad \zeta(\lambda) \geq \zeta_{\text{thr}}(\lambda) \\ 2 \cdot \text{round}\left(\left(1 - \frac{\zeta(\lambda)}{\zeta_{\text{thr}}}\right) \cdot \Lambda\right) + 1, & \text{else.} \end{cases} \tag{3.61}$$

The function round$(\cdot)$ rounds the element to the nearest integer and Λ is a scaling factor that determines the maximum degree of smoothing.

In order to detect only low SIR regions, a threshold ζ_{thr} is required that controls the trade-off between speech distortions and musical noise reduction. The term $1 - \frac{\zeta(\lambda)}{\zeta_{\text{thr}}}$ provides a soft-decision that states the reliability of the low SIR detection. Equation (3.61) ensures that the more reliable a low SIR frame was detected, the longer the window length resulting in stronger smoothing of the weighting gains. Applying a moving average window of length $N_s(\lambda)$ is equivalent to a linear filtering with the impulse response $H_s(\lambda, \mu)$ as follows:

$$H_s(\lambda, \mu) = \begin{cases} \frac{1}{N_s(\lambda)}, & \text{if} \quad \mu < N_s(\lambda) \\ 0 & \text{else} \end{cases}, \tag{3.62}$$

where $\mu \in \{0, 1, .., M-1\}$. Please refer to [EV09] where the magnitude responses of $H_s(\lambda, \mu)$ are shown for different values of N_s. Within the smoothing procedure, the weighting gain magnitudes are convolved over frequency μ by the filter $H_s(\lambda, \mu)$ in every frame λ:

$$\widetilde{G}(\lambda, \mu) = G(\lambda, \mu) * H_s(\lambda, \mu), \tag{3.63}$$

where the smoothed gains are termed $\widetilde{G}(\lambda, \mu)$. This postfilter is incorporated in the two novel dual-channel speech enhancement systems for hearing aids and mobile phones, which are presented in Secs. 4.1.3 and 4.2.1, respectively.

3.4.2 Psychoacoustic Weighting

As an alternative to a smoothing over frequency, a psychoacoustically motivated spectral weighting rule can be employed. The considered method was initially developed to reduce artifacts in noise reduction systems [GJV98] and used in a combination with acoustic echo control [GMJV02]. In [JV11] it was shown that this concept is also applicable for the purpose of speech dereverberation. The main idea is to reduce reverberation only in such frequency components which are not masked by the speech signal. Thus, no complete dereverberation is desired but to preserve a low level natural sounding reverberation which reduces the amount of musical tones and other artifacts significantly.

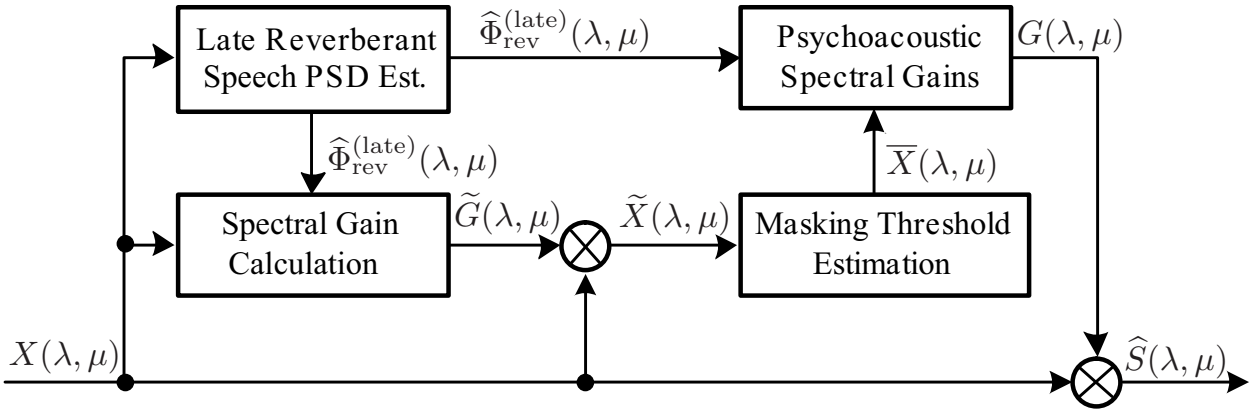


Figure 3.20: Block diagram of the considered single-channel speech dereverberation system using a psychoacoustically-motivated spectral weighting rule.

The main idea is to perform a pre-dereverberation of the input spectra and to calculate the masking thresholds from the pre-enhanced signal. Based on an estimate of the late reverberant speech PSD and the masking threshold, the final spectral weights are calculated and applied to the reverberant signal. The overall block diagram of the new system is depicted in Fig. 3.20. The processing steps are as follows:

1. Estimate the late reverberant speech PSD $\widehat{\Phi}_{\text{rev}}^{(\text{late})}(\lambda,\mu)$ using the method described above Eq.(3.4).

2. Calculate preliminary spectral gains $\widetilde{G}(\lambda,\mu)$ by means of the spectral subtraction rule (Eq.(3.11)).

3. Compute a pre-dereverberated signal by

$$\widetilde{X}(\lambda,\mu) = X(\lambda,\mu)\cdot\widetilde{G}(\lambda,\mu). \tag{3.64}$$

4. Estimate masking threshold $\overline{X}(\lambda,\mu)$ based on $\widetilde{X}(\lambda,\mu)$ using the ISO model [ISO93].

5. Calculate psychoacoustic weighting gains $G(\lambda,\mu)$ by [GJV98]:

$$G(\lambda,\mu) = \min\left(\sqrt{\frac{\overline{\text{X}}(\lambda,\mu)}{\widehat{\Phi}_{\text{rev}}^{(\text{late})}(\lambda,\mu)}} + \zeta_{\text{int}}, 1\right), \tag{3.65}$$

   with an interference attenuation factor ζ_{int}.

6. Perform the final dereverberation by applying the psychoacoustic gains to the reverberant input spectra by

$$\widehat{S}(\lambda,\mu) = X(\lambda,\mu)\cdot G(\lambda,\mu). \tag{3.66}$$

The application of this concept for recordings taken in the German parliament is discussed in Sec. 4.3. Related psychoacoustic dereverberation approaches are discussed, e.g., in [Tsi11].

3.5 Summary

This chapter has presented an overview and evaluations of state-of-the-art algorithms for independent dereverberation and noise reduction using one or two microphones. Furthermore, novel and improved algorithms are presented which are explicitly developed for the two considered applications, i.e., exploit the noise field coherence and the PLD between the two microphone signals. Moreover, novel concepts for joint reduction of reverberation and background noise are introduced and discussed. Additionally, strategies to reduce the undesired musical tones, which usually occur in spectral weighting-based speech enhancement, are presented.

The major contributions of this chapter are:

- an extensive evaluation of late reverberant speech PSD estimators including:
  - discussions on the necessity of a frequency-dependent RT and DRR estimation,
  - investigations of the required RT and DRR estimation accuracy,
  - proposal of an improved method which operates even within the critical distance and estimates the frequency-dependent RT as well as the DRR blindly.
- comparison and improvement of coherence-based dereverberation approaches:
  - adoption and improvement of methods which were developed for noise reduction for the purpose of speech dereverberation,
  - integration of more accurate models of the noise field coherence.
- derivation of two new estimators of the background noise PSD:
  - a method for mobile phone applications which exploits the power level differences of speech and interfering source signals,
  - a method for binaural hearing aids based on the noise field coherence,
  - extensive evaluations of both approaches and proof of lower estimation error compared to existing methods.
- proposal of a new spectral weighting rule for mobile phones exploiting the PLD.
- presentation of a concept for joint dereverberation and noise reduction:
  - investigation of drawbacks of consecutive dereverberation and noise reduction,
  - proposal of an interlaced concept combining estimators of the late reverberant speech and background noise PSD as well as blind methods for RT and DRR estimation.
- discussions on the reduction of musical tones.

With the presented algorithms and concepts, a significant improvement of speech quality and listening comfort can be obtained even under adverse acoustic conditions.

4

Applications

4.1 Speech Enhancement for Binaural Hearing Aids

The major drawback of current speech enhancement algorithms for hearing aids is that most of these techniques were developed for systems with a single output channel given one or possibly multiple input channels. Therefore, they are only suitable for bilateral processing, which means that the devices at the left and right ear are independently performing monaural enhancement without taking spatial information into account. Several studies have shown that unsynchronized bilateral processing degrades the ability for sound localization [JSEV10] and that hearing impaired persons localize sounds better without their independent bilateral hearing aids than with them [vdBKM+05]. This can be explained by the fact that the binaural cues, which are the basis for human sound localization, are not preserved. This comprises mostly the *Interaural Level Difference* (ILD) and *Interaural Time Difference* (ITD), cf. [Bla96]. Thus, it is advantageous to perform binaural instead of bilateral processing, especially as an appropriate wireless data-link between both hearing aids can be assumed, cf. [HCE+05].

An important class of binaural noise reduction algorithms, besides spectral subtraction techniques, is the *Multichannel Wiener Filter* (MWF), cf. [DM02]. An advanced concept is the *Speech Distortion Weighted Multichannel Wiener Filter* (SDW-MWF) proposed in [DSWM05, DSWM07], which employs a minimum mean-squared-error estimate of the clean speech from noisy reference signals by exploiting the correlation properties of both speech and noise. A special parameter allows to balance the trade-off between distortions and noise reduction. This method has further been extended in order to allow for a preservation of the binaural cues, e.g., in [CMW11b, Cor11]. A comparison of the SDW-MWF [DSWM07] to a beamformer with binaural post-filter and a spatial prediction approach is given in [MHD11]. It turns out that the SDW-MWF leads to the best results in terms of SNR and *Speech Reception Threshold* (SRT) improvement. One major limitation of MWF approaches is the requirement of a noise-robust VAD in order to update the noise correlation matrix. Any misdetection, especially the classification of speech segments as noise, leads to strong artifacts

and audible distortions. The problem has been tackled partly in [CMW11a, Cor11], where a VAD-robust SDW-MWF is proposed. Alternatively, the recently proposed estimation procedure for the noise correlation matrix [HG12] or the procedure as described in [MHA11] could be employed, which is however, out of the scope of this thesis. Further approaches for binaural noise reduction are, not limited to, binaural *Blind Source Separation* (BSS) strategies as proposed in [HKLP08, RZK10, RPF+10] and binaural beamformer as presented in [LV06, HKLP08]. A binaural spectral subtraction concept, which does not aim to preserve the binaural cues, is published in [DE96].

Binaural dereverberation is addressed, e.g., in [Pei92, Wit01, WH03] and advanced concepts such as [JSEV10] already lead to good results. However, no speech enhancement system for binaural hearing aids exists so far which allows for a reduction of both early and late reverberation as well as background noise and explicitly preserves the binaural cues.

This section consists of two major parts: First, in Secs. 4.1.1, 4.1.2 the necessity of binaural instead of bilateral processing is studied. It will be shown exemplarily with three known dereverberation algorithms how bilateral signal processing affects the source localization. This comprises possible extensions of monaural algorithms to a binaural output. In the second part of this section, a novel two-stage binaural speech enhancement system is proposed in Sec. 4.1.3 which does not alter the binaural cues. All needed acoustic quantities are estimated blindly from the noisy and reverberant speech signals, i.e., RT and DRR. Finally, a short discussion on the binaural data-link is given.

4.1.1 Extension of Monaural Algorithms to Binaural Output

An extension of monaural speech enhancement algorithms to a binaural output is not trivial. A discussion of fixed and adaptive beamforming with binaural output can be found, e.g., in [DRZ97, WGDZ97]. The authors in [LV06] propose a binaural noise reduction system consisting of a binaural superdirective beamformer and two identical postfilters. A comprehensive study how binaural noise reduction algorithms can preserve binaural cues can be found in [vdB08].

The general concept of a bilateral spectral subtraction in the DFT domain without the exchange of any control information is illustrated in Fig. 4.1 (top). Each processing unit calculates spectral gains which are applied independently to the disturbed input spectra by

$$\widehat{S}_l^{(\mathrm{bil})}(\lambda,\mu) = X_l(\lambda,\mu)\cdot G_l(\lambda,\mu) \tag{4.1a}$$

$$\widehat{S}_r^{(\mathrm{bil})}(\lambda,\mu) = X_r(\lambda,\mu)\cdot G_r(\lambda,\mu). \tag{4.1b}$$

It is obvious that by using this kind of unsynchronized processing, the binaural cues may be severely degraded.

In order to ensure unaffected source localization, it is important to preserve the binaural cues to a certain extent. A preservation of the binaural cues can be ensured

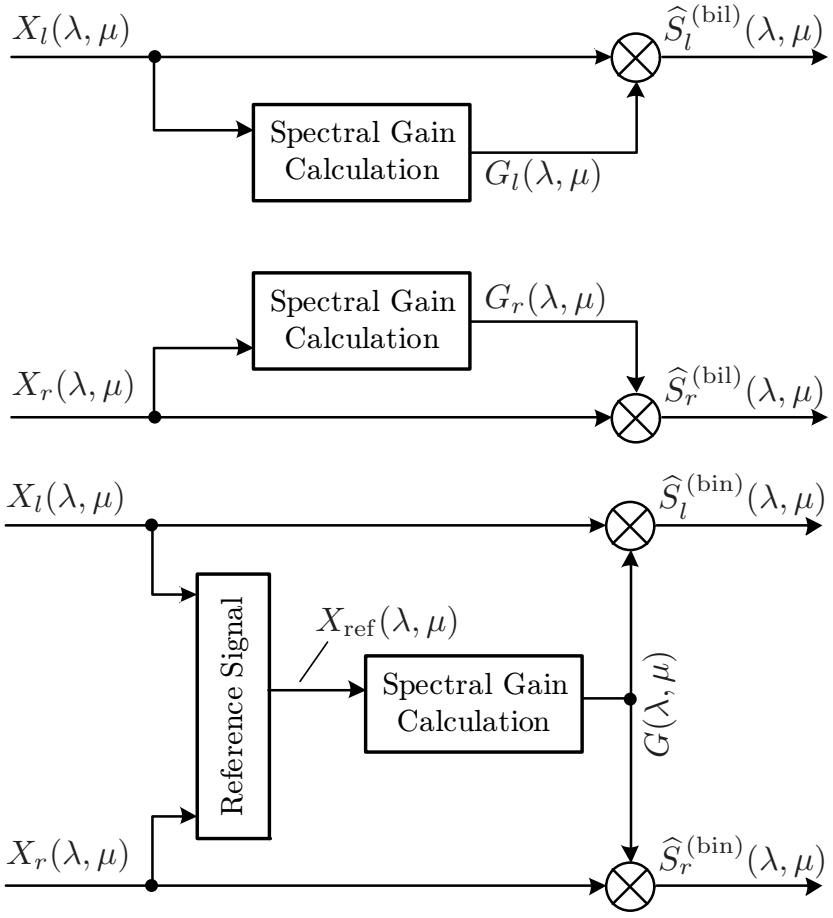


Figure 4.1: Block diagrams of (top) bilateral spectral subtraction and (bottom) proposed extension of monaural to binaural spectral subtraction, both operating in the frequency domain.

basically in two different ways. The first method would be to reconstruct the binaural cues after the processing using a binaural postfilter. An overview about binaural reproduction suitable for the application to BSS can be found in [WPK08]. The problem of binaural cue preservation in the context of binaural artificial bandwidth extension has been addressed, e.g., in [LV09a]. The other method, which is considered here, is to incorporate the cue preservation into the processing algorithm. One promising method is to apply the same spectral weighting gains to each of the two input channels, which inherently preserves the ILD cues, cf. [Pei92, LV06, JV10, JSEV10]. The ITD is also not affected since the algorithm keeps the phase of the input signals and the same algorithmic delay exist among the two channels. Techniques and theoretical analyses on the binaural cue preservation of the MWF is given, e.g., in [Cor11, vdB08] and references therein.

In order to calculate the spectral gains out of the two channels, different methods are possible. In this thesis, a reference signal is calculated from the average of both time-aligned signals ($X_l'(\lambda,\mu)$, $X_r'(\lambda,\mu)$) according to

$$X_{\mathrm{ref}}(\lambda,\mu) = \frac{1}{2}\cdot\left[X_l'(\lambda,\mu) + X_r'(\lambda,\mu)\right]. \tag{4.2}$$

The estimation of the time delays is performed by means of the *Generalized Cross-*

Correlation with Phase Transform (GCC-PHAT) as described in [KC76] and is incorporated in the reference signal block. The weighting gains $G(\lambda, \mu)$, which are determined from $X_{\text{ref}}(\lambda, \mu)$ only, are applied to the disturbed (non time-aligned) input spectra by

$$\widehat{S}_l^{(\text{bin})}(\lambda, \mu) = X_l(\lambda, \mu) \cdot G(\lambda, \mu) \tag{4.3a}$$

$$\widehat{S}_r^{(\text{bin})}(\lambda, \mu) = X_r(\lambda, \mu) \cdot G(\lambda, \mu). \tag{4.3b}$$

This concept is shown in Fig. 4.1 (bottom). Alternatively, two separate weighting gains can be computed and a combination of the two gains, e.g., by taking the minimum, maximum or average, can be applied to both channels as discussed in [TGM11].

4.1.2 Influence of Bilateral Dereverberation on Binaural Cues

In this subsection, the influence of bilateral, i.e., independent dereverberation on the binaural cues of the desired speech signal is investigated. The objective of any enhancement algorithm with respect to the binaural cues should be to preserve the absolute mean values while ensuring very low fluctuations from frame to frame. The minimum audible changes for ILD (0.5 dB) and ITD (10 μs) should not be exceeded, cf. [Har99]. Several studies have shown that the human auditory system is capable of re-learning the spatial information when receiving altered binaural cues. However, since the adaptation typically takes a week or more, an adjustment to rapid changes is impossible [vdB08, WZ09]. Exemplarily, we restrict the analysis to changes of ILD. An extension to ITD can be done by switching to an ITD estimator instead of the ILD estimator.

To quantify the impact of independent dereverberation of both channels on the binaural cues, three dereverberation algorithms will be used for an independent single-channel enhancement. The first two considered dereverberation algorithms: SMERSH and CELP are both based in the discrete model of speech production (see Sec. 3.1.3). The third algorithm (LB) is based on an estimate of the late reverberant speech PSD in combination with a spectral subtraction rule (see Secs. 3.1.1, 3.1.1.4). The basis for this investigation lies in the observation that the DRR is highly dependent on the azimuth angle as shown in Fig. 2.14.

The influence on the binaural cues will be investigated as follows. From the dual-channel input signals, the binaural cues are estimated from the reverberant speech signals before processing. Afterwards, a bilateral dereverberation (independently without any data-link between left and right processing units or synchronization) is performed with the described algorithms. Finally, the binaural cues are estimated again and compared to the cues before processing. For all binaural cue estimation tasks, the cue selection procedure of Eq.(2.54) is used with an adaptive threshold. The evaluation is carried out with speech files from the NTT database that are convolved with BRIRs measured in a stairway hall at different azimuth angles, all in the presence of a dummy head. The investigation focuses on frame-by-frame fluctuations

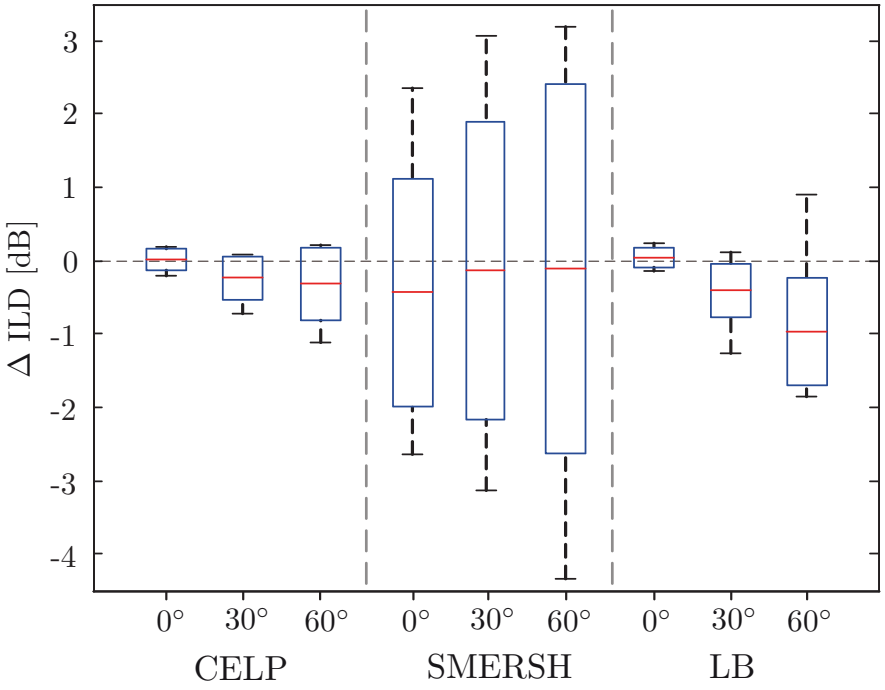


Figure 4.2: Differences in ILD estimation compared to reverberant speech (stairway hall: $d_{\mathrm{LM}} = 1\,\mathrm{m}$, $\mathrm{T}_{60} = 0.82\,\mathrm{s}$) for the three considered azimuth angles of the source. The results are equally weighted for all frequency bands above 1.5 kHz .

of the ILD cues, which is an important issue since most algorithms perform enhancement of short speech frames which causes a different degree of enhancement per frame. Therefore, we calculate the ILD frame-wise for each of the 24 (non-uniform) subbands over all frames to measure the variance in ILD estimation. Finally, the results are averaged over all bands above 1.5 kHz. The ILD difference in each frame is denoted by $\Delta\mathrm{ILD}(\lambda)$.

The results for three different azimuth angles are depicted in Fig. 4.2. The boxes represent the variance from the mean value (horizontal red line inside the box) and the end of the whiskers represent minimum and maximum of $\Delta\mathrm{ILD}(\lambda)$.

It can be seen from Fig. 4.2 that all tested algorithms cause high variations in the binaural cues as shown exemplarily here for the ILD and hence, influence the source localization. All algorithms show the lowest influence for the frontal direction (0°) and distort the cues most severely for sources from aside. The most significant increase in ILD fluctuations occurs for the SMERSH algorithm. Even though the CELP postfilter leads to the smallest variations among the tested approaches, moderate changes in ILD are still audible for sidewise sources. In terms of the dereverberation performance, the spectral subtraction technique (LB) shows the highest amount of reverberation reduction. Since all algorithms exhibit changes in binaural ILD cues that are mostly above the minimum audible difference, we can conclude that bilateral dereverberation has a clearly perceivable influence on the source localization.

Table 4.1: Influence of bilateral dereverberation on binaural cues: results of the listening test.

Simulation setup	No preference	Bilateral dereverberation Fig. 4.1(a)	Binaural dereverberation Fig. 4.1(b)
$\theta = 0°$, $d = 2\,\mathrm{m}$	5.9 %	11.7 %	82.4 %
$\theta = 60°$, $d = 1\,\mathrm{m}$	11.7 %	23.5 %	64.8 %
Average	**8.8 %**	**17.6 %**	**73.6 %**

4.1.2.1 Influence of Bilateral Processing by a Listening Test

The degradation of the binaural cues due to bilateral dereverberation has also been investigated with an informal listening experiment. During the test with 17 experienced listeners, three different signals were presented to the participants: the reverberant speech, the processed signal using a bilateral dereverberation algorithm (A) and the processed signal after binaural dereverberation (B).

The test signals (A) are processed using the LB algorithm by applying the structure in Fig. 4.1 (top). The binaural signals (B) are generated using the same algorithm in the binaural configuration which means that the same spectral weighting gains are applied to each channel (see Fig. 4.1 (bottom)).

For each of the sentences, the listeners were asked to judge the overall speech quality as well as the audible modifications in the interaural time and level differences (compared to the provided reverberant speech). The listeners could choose between 'A sounds better than B', 'B sounds better than A' and 'no preference'. The samples could be played ad libitum before the probands had to make their judgments. The reverberant signals are generated using binaural room impulse responses of a stairway hall at different azimuth angles and distances. To ensure high quality audio and to avoid distortions due to the headphone, a calibrated combination of a HEAD acoustics PEQ V digital equalizer and Sennheiser HD600 headphone was used. The test took place in a low-reverberant studio booth having a high sound isolation of 42 dB against outside noise.

The results for two different azimuth angles are stated in Table 4.1. It can be seen that for both setups most participants preferred the binaural dereverberation method (B) over the bilateral algorithm (A). This corresponds to the objective evaluation results conducted in the beginning of the section.

4.1.3 Binaural Speech Enhancement

This section presents a novel two-stage binaural speech enhancement system which allows for a joint reduction of early and late reverberation as well as diffuse and coherence background noise.

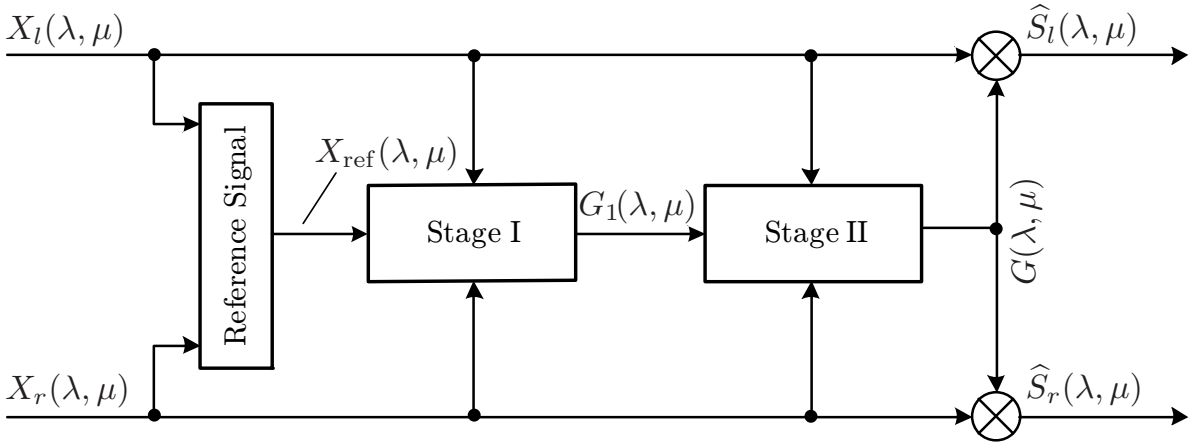


Figure 4.3: Simplified schematic diagram of the proposed two-stage binaural cue preserving speech enhancement algorithm.

The cascade of the two stages as depicted in Fig. 4.3 is mainly motivated by the fact that each stage requires different properties for the input signal in terms of the DRR and diffusiveness of the background noise. The basic idea of such a combination is that Stage I of the algorithm mainly reduces the late reverberant and background noise components, while the subsequent Wiener filter in Stage II attenuates all non-coherent signal components. This results in an efficient reduction of both early and late reverberation as well as coherent and diffuse background noise. Due to the algorithmic structure, the binaural cues are not affected.

The first stage comprises estimates of the late reverberant speech PSD (mainly controlled by parameter T_l) and background noise PSD, which are combined and used to calculate spectral weighting gains which allow to reduce late reverberation as well as coherent background noise components. The second stage attenuates the residual interferences after the first stage.

After the first processing step, the DRR is increased since late reverberation is attenuated while keeping the direct and early speech component unaffected. This first stage does not influence the coherence between both channels since the same spectral weights are applied to both channels. The second stage estimates the (direct) speech PSD, which requires a high DRR in order to reduce estimation errors. Thus, it is beneficial to increase the DRR in the previous step. The second stage attenuates all non-coherent parts, hence early and late reverberation, a great DRR increase can be expected as well. Consequently, a reversed order of the two stages would be less effective, as confirmed by our experiments. Please note that the actual pre-enhancement of Stage I is carried out in Stage II and for the sake of clarity, the required time-aligned spectra $X_l'(\lambda,\mu)$ and $X_r'(\lambda,\mu)$, which are also computed in the reference signal unit, are not sketched.

It has to be mentioned that the computation of the reference signal $X_{\mathrm{ref}}(\lambda,\mu)$, which can also be seen as a *Delay-and-Sum Beamformer* (DSB), already performs a reduction of background noise and reverberation. Therefore, the resulting reverberation time and hence, the estimated PSD of the late reverberant speech is only an approximation of the estimates directly from the input signals. Since the DSB provides only

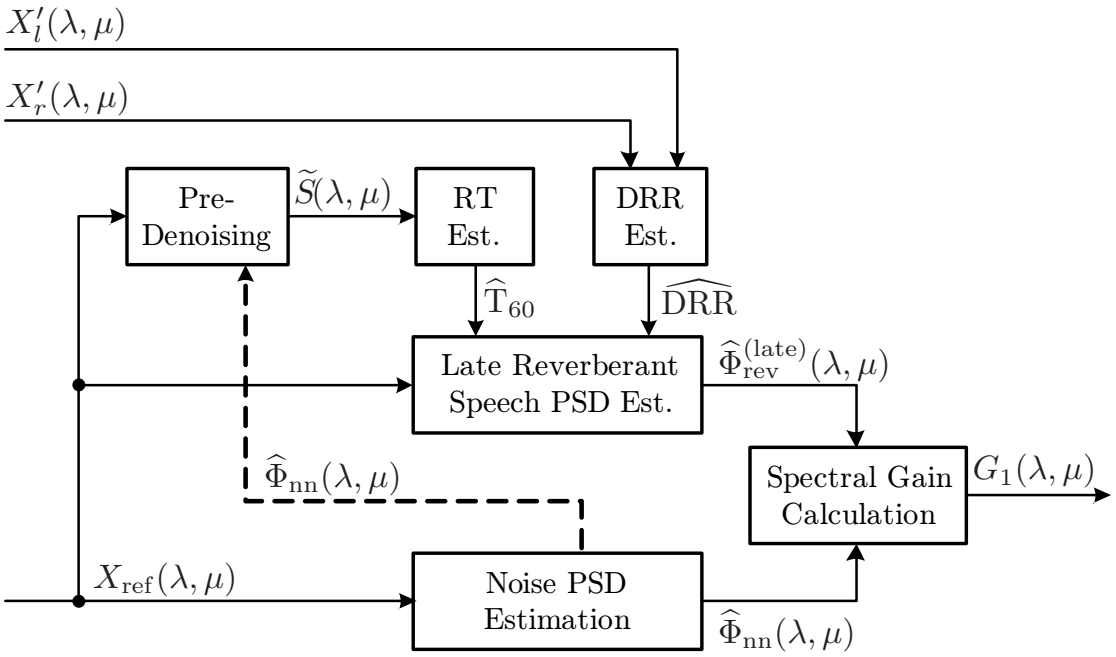


Figure 4.4: Schematic diagram of Stage I.

a small amount of reverberation reduction this approximation is still feasible as a small variation in the estimated reverberation time is not critical.

In order to fulfill the special requirements of hearing aids with open fitting such that the algorithmic delay is below 10 ms [SM02], a filtering in the time-domain by means of a filter-bank equalizer is also possible, cf. [LV08b].

4.1.3.1 Stage I: Reduction of Late Reverberation and Background Noise

In the first stage, two independent estimates for the late reverberant speech PSD as well as the background noise PSD are calculated from the single-channel reference DFT coefficients $X_{\mathrm{ref}}(\lambda,\mu)$ as illustrated in Fig. 4.4. The late reverberant speech PSD estimation is based on the HB algorithm introduced in Sec. 3.1.1.1 which allows for a robust estimation even within the critical distance. The noise PSD estimation is carried out by means of the SPP algorithm which was found to be the best currently available single-channel noise PSD estimator as discussed in Sec. 3.2.3.1. Alternatively, the proposed coherence-based noise PSD estimator (see Sec. 3.2.1.2) could be employed if diffuse noise without additional coherent interfering sources can be guaranteed.

Following the discussions in Sec. 3.3.3, the two PSD estimates are combined by taking the maximum of each individual frequency bin using Eq.(3.59e). The spectral gains $G_1(\lambda,\mu)$ are computed based on the overall interfering PSD $\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu)$ and the Wiener filter with DDA according to Eq.(3.14).

Since the late reverberant speech PSD estimator requires reliable knowledge of the RT and DRR, the frequency-dependent RT estimator presented in Sec. 2.3.3 as well as the DRR estimator discussed in Sec. 2.3.4 are employed. Please note that the RT

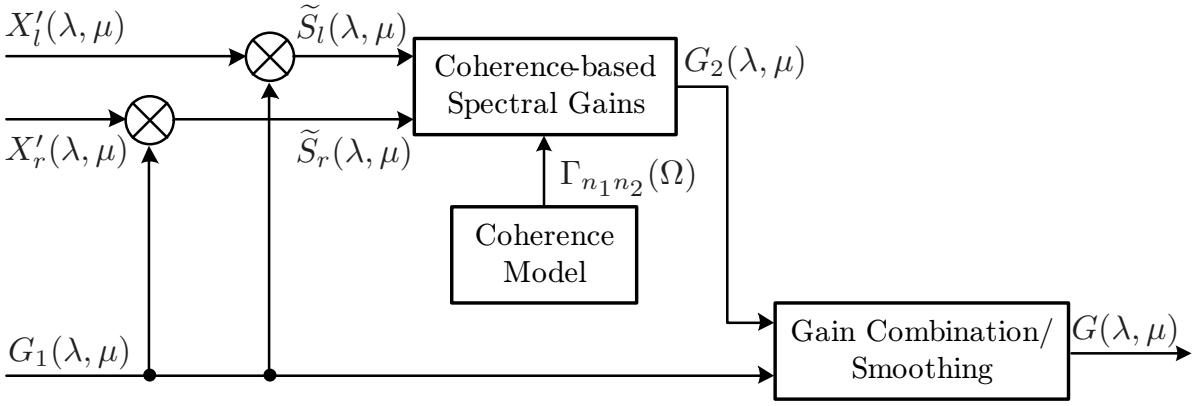


Figure 4.5: Schematic diagram of Stage II.

estimation is performed on a pre-denoised signal, obtained by applying intermediate spectral gains $G_{nn}(\lambda,\mu)$, using $\widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu)$ and a Wiener filter with DDA, to the reference signal:

$$\widetilde{S}(\lambda,\mu) = X_{\mathrm{ref}}(\lambda,\mu) \cdot G_{nn}(\lambda,\mu). \tag{4.4}$$

The overall gains are denoted by $G_1(\lambda,\mu)$ which are passed over to the second stage. Please note that these gains could already be applied to the disturbed input spectra if the second stage is not employed.

4.1.3.2 Stage II: Speech Enhancement Based on Sound Field Coherence

The motivation for a second processing step is that the spectral subtraction rule described in the previous section aims at reducing late reverberation and background noise only and thus, early and residual late reverberation remains. The subsequent coherence-based algorithm exploits the low coherence of the sound field between different microphones to estimate the (direct) speech PSD and to remove all non-coherent signal parts while keeping the coherent parts unaffected. Since only the direct speech shows a high coherence between the microphones, this approach also reduces early reverberation. A further advantage is that no estimation of room acoustic parameters (e.g., T_{60}) is required and that *a priori* information about the sound field can significantly improve the effectiveness of the algorithm.

The general structure is illustrated in Fig. 4.5. First, intermediate enhanced signals are computed by applying the gains of Stage I to the time-aligned input spectra according to

$$\widetilde{S}_l(\lambda,\mu) = X'_l(\lambda,\mu) \cdot G_1(\lambda,\mu) \tag{4.5a}$$

$$\widetilde{S}_r(\lambda,\mu) = X'_r(\lambda,\mu) \cdot G_1(\lambda,\mu). \tag{4.5b}$$

From these signals, the PSD of the clean speech components $\widehat{\Phi}_{ss}(\lambda,\mu)$ is estimated by applying Eq.(3.28) and by using an appropriate model for the noise field coherence $\Gamma_{n_1 n_2}(\Omega)$. The spectral weights $G_2(\lambda,\mu)$ are then obtained by Eq.(3.29).

Finally, the weighting gains of the two stages $G_1(\lambda,\mu)$ and $G_2(\lambda,\mu)$ are confined by lower thresholds $G_{\min}^{(1|2)}$ and are multiplied to obtain the final weights $G(\lambda,\mu)$:

$$G(\lambda,\mu) = G_1(\lambda,\mu)\cdot G_2(\lambda,\mu). \tag{4.6}$$

This corresponds to a subsequent application of the weighting gains of the two stages. For reducing the amount of musical tones, spectral smoothing over frequency of the magnitudes $G(\lambda,\mu)$ is performed using the smoothing procedure outlined in Sec. 3.4.1. Finally, the weighting gains $G(\lambda,\mu)$, which are confined by $G_{\min}^{(\mathrm{cmb})}$, are applied to the input spectra by

$$\widehat{S}_l(\lambda,\mu) = X_l(\lambda,\mu)\cdot G(\lambda,\mu) \tag{4.7a}$$

$$\widehat{S}_r(\lambda,\mu) = X_r(\lambda,\mu)\cdot G(\lambda,\mu). \tag{4.7b}$$

The crucial point is now to select a suitable model for the sound field coherence $\Gamma_{n_1 n_2}(\Omega)$ in Eq.(3.28). For an ideal diffuse sound field with line-of-sight between two microphones, the optimal solution is the model in Eq.(2.6). However, when it comes to binaural signal processing where no line-of-sight between the microphones can be assumed, this model is not appropriate. Since the head-shadowing has a severe impact on the coherence, we propose to use the semi-analytical coherence model for a binaural spherically isotropic sound field as described in Sec. 2.2.1.1 and App. B.

Please note that in the practical implementation, the approximation of the coherence model based on the sum of Gaussians is beneficial as presented in [JV10, JSEV10]. Alternatively, pre-calculated look-up tables can be stored.

4.1.4 Performance Evaluation

The performance evaluation is subdivided into separate experiments for the dereverberation and noise reduction performance as well as the overall enhancement performance with noisy and reverberated input signals. Furthermore, the influence of head shadowing on the algorithm performance and the chosen noise field coherence model is given. The main simulation settings are listed in Table 4.2. The settings for the DRR estimator are chosen according to Table 2.6.

Based on further simulations with a desired talker from various azimuth angles θ, it is shown in [JV10] that the employed time-alignment ensures the same speech enhancement performance for all incidence angles. Sources which are not in the range of $-30^\circ \leq \theta \leq +30^\circ$ (see Fig. 2.1 (b)) cannot be enhanced without any compensation of the different time-delays of arrival between the microphones.

4.1.4.1 Dereverberation Performance

This subsection gives the results of three different binaural dereverberation algorithms. The two-stage system (Two-Stage) is evaluated and the performance is compared to a processing employing the individual stages (Stage I and Stage II) only.

Table 4.2: Main simulation parameters of proposed two-stage speech enhancement system.

Parameter description	Setting
Smoothing factors	$\alpha^{(\mathrm{xx})} = 0.9$, $\alpha^{(\mathrm{PSD})} = 0.8$
Late reverberant time span	$\mathrm{T}_l = 80\,\mathrm{ms}$
Threshold	$\Gamma_{\max} = 0.99$
Gain Smoothing factors	$\zeta_{\mathrm{thr}} = 0.2$, $\Lambda = 25$
Gain thresholds	$G_{\min}^{(1)} = 0.1$, $G_{\min}^{(2)} = 0.3$, $G_{\min}^{(\mathrm{cmb})} = 0.2$
Smoothing freq. range	$4\,\mathrm{kHz} \leq \mathrm{f} \leq 7\,\mathrm{kHz}$

For a fair comparison of the dereverberation performance, the noise reduction is deactivated such that only the late reverberant speech PSD estimate in Stage I is used, i.e., $\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \widehat{\Phi}_{\mathrm{rev}}^{(\mathrm{late})}(\lambda,\mu)$. Furthermore, no additive background noise was added to the reverberant signal.

The results for the lecture and office room in terms of the enhancement of the objective measures: SRMR, segDRR and SII are given over different input DRRs in Fig. 4.6. It turns out that among the compared methods, the proposed two-stage system shows the highest amount of reverberation reduction in terms of SRMR and segDRR improvement. It outperforms the two individual stages for all tested scenarios.

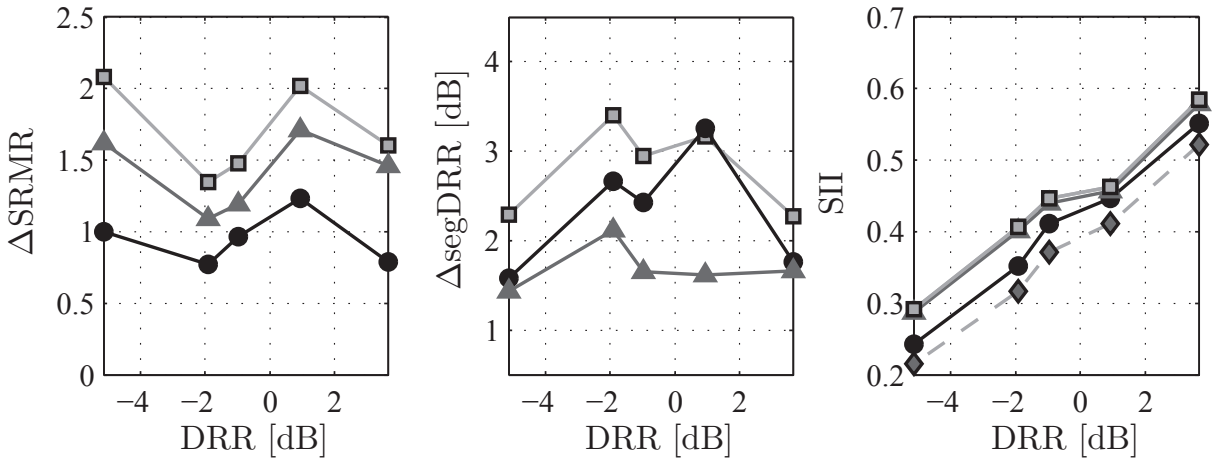
The SII improvements are almost equivalent for the two-stage system and Stage II only, which means that the highest amount of intelligibility improvement is caused by the coherence-based processing. This tendency of the objective measurements was verified by informal listening tests. It could be observed that the first stage greatly reduces the late reverberant tail but no early reverberation. The coherence-based algorithm (Stage II only) made the processed signal sound more "clearly" (reduction in coloration). This effect was in particularly audible for the two-stage system, where a reduction of both coloration and overlap-masking results in the best listening comfort with the lowest amount audible distortions among all tested approaches. The same tendency was observed for the NA-SA measure as well as for the *Bark Spectral Distortion* (BSD), cf. [JSEV10].

4.1.4.2 Noise Reduction Performance

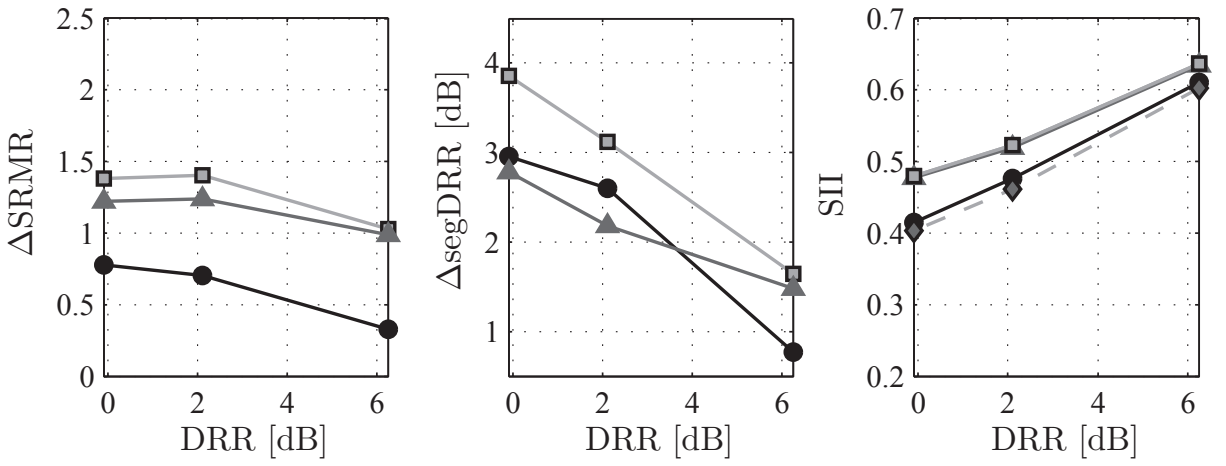
The purpose of the second experiment is to assess the noise reduction performance only. Hence, for a fair comparison, only the background noise PSD estimator in Stage I was activated, i.e., $\widehat{\Phi}_{\mathrm{int}}(\lambda,\mu) = \widehat{\Phi}_{\mathrm{nn}}(\lambda,\mu)$. For the objective comparison, the instrumental measures NA-SA, ΔPESQ and SII are chosen.

In addition to the proposed binaural system, the binaural SDW-MWF with an adaptive smoothing procedure which updates the correlation matrix in segments of speech absence [MHD11] is employed[1]. We use an *ideal VAD* determined from the clean

[1]The author would like to thank Daniel Marquardt for helpful discussions and for providing a MATLAB reference implementation.



(a) Lecture room ($T_{60} = 0.69\,s$).



(b) Office ($T_{60} = 0.51\,s$).

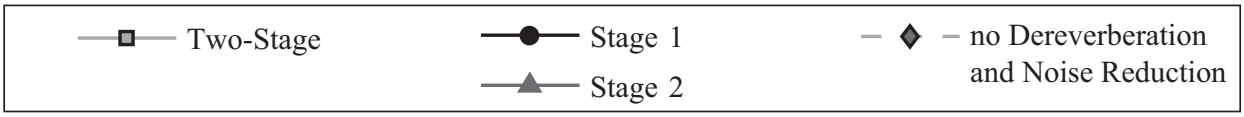


Figure 4.6: Dereverberation performance of two-stage system(w/o background noise).

speech signal. In a real system VAD algorithms as in [CMW11a, Cor11, JKAO09] or the proposed dual-channel VAD used for the DRR estimator have to be employed. This integration is, however, out of the scope of this work.

The results over a SNR range from -5 to 20 dB using binaural pub noise recordings from the ETSI background noise database [ETS09] are illustrated in Fig. 4.7. As for the dereverberation performance, it can be observed that both individual stages perform a good noise reduction and that the combined system leads to the best results, as expected. For negative signal-to-noise ratios, the performance of all five algorithms drops but as the SII measures still shows an increase of approx. 0.2 for the two-stage system, an improved intelligibility is obtained. The SDW-MWF shows the

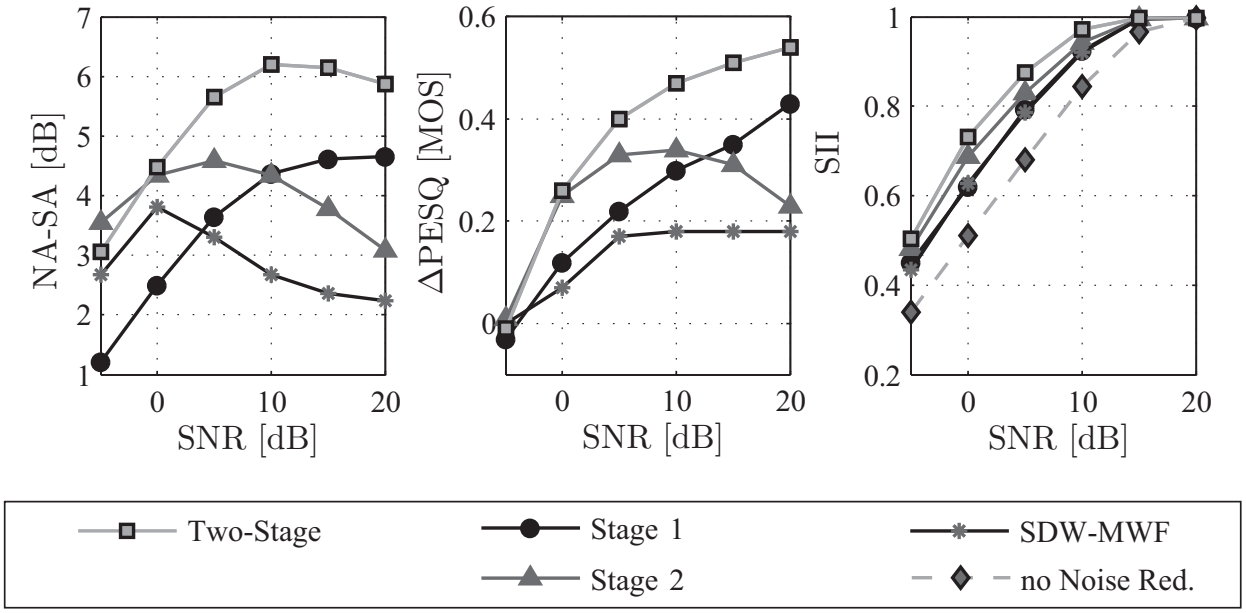


Figure 4.7: Noise reduction performance of two-stage system using recorded pub noise at different SNRs (w/o reverberation).

lowest performance in this diffuse noise situation as it is mainly capable of reducing interfering point sources from aside the desired source. Similar results were observed for cafeteria and kindergarten noise.

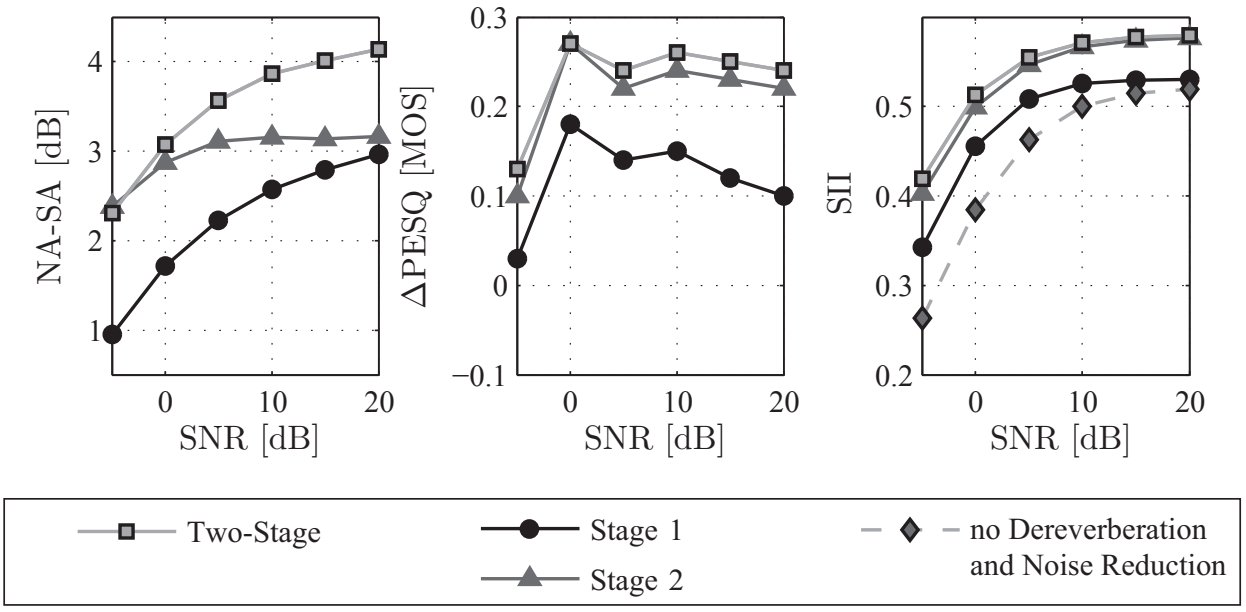


Figure 4.8: Dereverberation and noise reduction performance of two-stage system using recorded pub noise at different SNRs and RIRs of the lecture room.

4.1.4.3 Joint Dereverberation and Noise Reduction Performance

The last experiment deals with the enhancement of input signals which are disturbed by additive background noise *and* room reverberation. For this purpose, the reverberant signals used in Sec. 4.1.4.1 were degraded with additional background noise as in Sec. 4.1.4.2. The complete algorithm with both noise PSD and late reverberant speech PSD estimators is used and all required acoustic parameters (frequency-dependent RT and DRR) are estimated blindly.

The objective scores are given in Fig. 4.8, exemplarily for the lecture room and pub noise over a varying SNR. The same findings as for the dereverberation and noise reduction alone can be made: The two-stage algorithm is beneficial in all tested conditions and it can be concluded that with the proposed speech enhancement system a significant reduction of all considered interferences even under adverse acoustic conditions can be achieved. Please note that the same trend was observed for other RIRs and different background noise types.

4.1.4.4 Influence of Head Shadowing

Finally, the advantage of the novel binaural coherence model in comparison to the free-field diffuse model is shown exemplarily for the coherence-based Stage II. The plot in Fig. 4.9 depicts the measured *Speech Attenuation* (SA) over frequency by using two different coherence models. It can be seen that in the low frequency range the employment of the binaural coherence model can greatly reduce the resulting speech distortions by up to 2 dB. This corresponds to Fig. 2.9 where it is illustrated that in binaural applications the proposed binaural coherence model is more accurate than the frequently used free-field diffuse model.

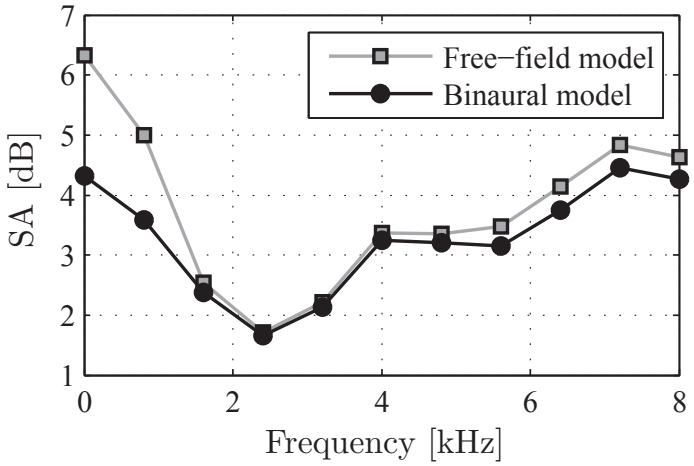


Figure 4.9: Illustration of effects of novel binaural coherence model on speech attenuation (SA) using Stage II and a reverberant input signal (lecture room RIR, w/o background noise).

4.1.5 Binaural Wireless Data-Link

The availability of a wireless data-link between two hearing aids allows for the integration of binaural speech enhancement algorithms. By means of an appropriate data exchange, the algorithms can exploit spatial information and, most important, preserve the binaural cues. This is a great advantage compared to bilateral signal processing and overcomes several limitations especially as state-of-the-art hearings aids are already capable of transmitting data from one hearing aid to the other device at a total data-rate of 212 kbit/s, cf. [Cor11]. Therefore it is very important to study the impact of the capacity of a wireless data-link on the performance of speech enhancement algorithms. The authors in [RV06] and [SdB09] investigate the optimal trade-off between the required bit-rate and the noise attenuation provided by a binaural beamformer. In [Cor11] a possible reduction of the bitrate required for the binaural data-link of the SDW-MWF algorithm is discussed.

An elaborate discussion on the wireless data-link with all its consequences for speech enhancement algorithms and the battery capacity is out of the scope of this thesis. In the following, it will only be investigated exemplarily how the binaural data-link influences the coherence of the desired (direct speech) signals reaching the left and right ear and how this affects the speech enhancement performance. In order to model the digital transmission from one ear to the other in the time-domain, we consider the simple case of a transmission with A-law quantization only, cf. [VM06], as well as the usage of two different speech codecs: G.722 [ITU88] and AMR-WB [3GP04b]. The transmission model is depicted in Fig. 4.10. The left device receives the transmitted (quantized and/or coded, i.e., encoder and decoder) signal $x_r^{(\mathrm{T})}(k)$ from the right device and performs the binaural processing based on the signals $x_l(k)$ and $x_r^{(\mathrm{T})}(k)$ to obtain an enhanced signal $\hat{s}_l(k)$, and vice versa for the right device. Transmission errors and the influence of delay are not considered.

Coherence-based speech enhancement algorithms, e.g., Stage II of the proposed binaural system, exploit the high coherence of the desired signal and the low coherence

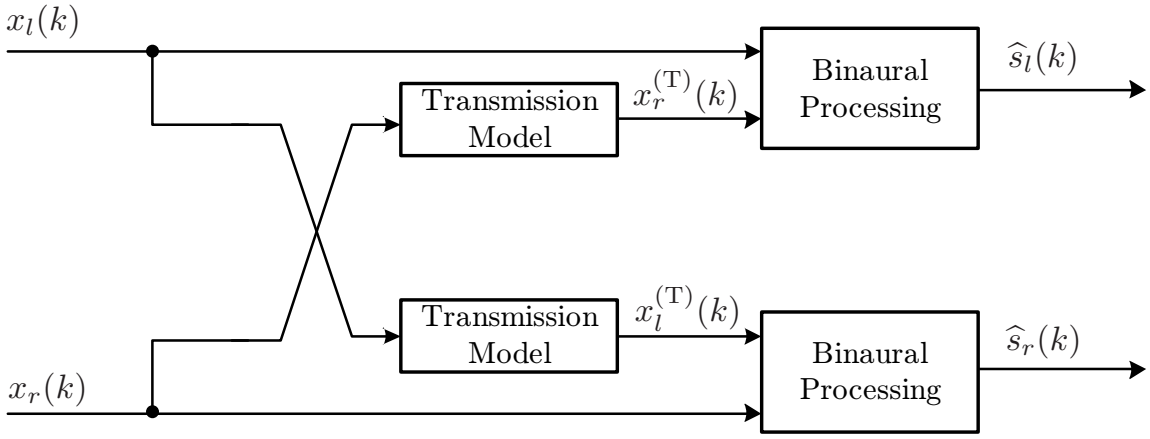


Figure 4.10: Binaural wireless data-link transmission model.

of the interference. Therefore, it is required that the coherence of the desired signal is not affected by the data-link. Figure 4.11 (left) shows the MSC between the unquantized signal $x_l(k)$ and the quantized signal $x_r^{(\mathrm{T})}(k)$ in dependency of the bit-rate at a sampling frequency of $f_s = 16\,\mathrm{kHz}$. It can be observed that the coherence is greatly affected by the transmission, especially at lower bit-rates. To quantify the effects on the speech enhancement performance, the signals $x_l(k)$ and $x_r^{(\mathrm{T})}(k)$ are used as the input signals of the Stage II algorithm. Figure 4.11 (right) shows the occurring speech distortions in terms of the SA measure over frequency. It can be observed that at least a data-rate of 128 kbit/s is required when transmitting the signals with this simple transmission model. The decrease of the coherence for the lower data-rates in Fig. 4.11 (left) has a severe impact on the speech attenuation. We conclude that the transmission loss has to be taken into account and as a consequence that the performance of coherence-based algorithms is lowered due to the quantization noise.

To take into account more appropriate transmission models, experiments were conducted in the same manner using the G.722 as well as the AMR-WB speech codec. The G.722 codec is capable of transmitting audio and speech signals in the frequency range of $50 - 7000\,\mathrm{Hz}$ at fixed bit-rates of 48 kbit/s, 56 kbit/s and 64 kbit/s (only Mode 1 with 64 kbit/s is considered). Within the *Sub-Band Adaptive Differential Pulse Code Modulation* (SB-ADPCM) concept (see [VM06]), the full frequency band is split into two subbands (higher and lower) using a *Quadrature Mirror Filter* (QMF) filterbank and the signals in each subband are encoded using ADPCM. The lower subband is coded with 6 bits ($\widehat{=}\,48\,\mathrm{kbit/s}$) and the higher subband with 2 bits only ($\widehat{=}\,16\,\mathrm{kbit/s}$). The AMR-WB codec is capable of processing the same audio bandwidth at various bit-rates using the *Algebraic Code Excited Linear Prediction* (ACELP) principle, cf. [VM06]. However, only the lower frequency range of $50 - 6400\,\mathrm{Hz}$ is encoded by means of waveform coding. The higher range of $6400 - 7000\,\mathrm{Hz}$ is processed by means of an artificial bandwidth extension with or without side information. In terms of the algorithmic latency, only the G.722 is favorable for the integration into a binaural wireless transmission system. The AMR-WB codec is used as a reference only.

The influence of transcoding with the G.722 (64 kbit/s) as well as the AMR-WB

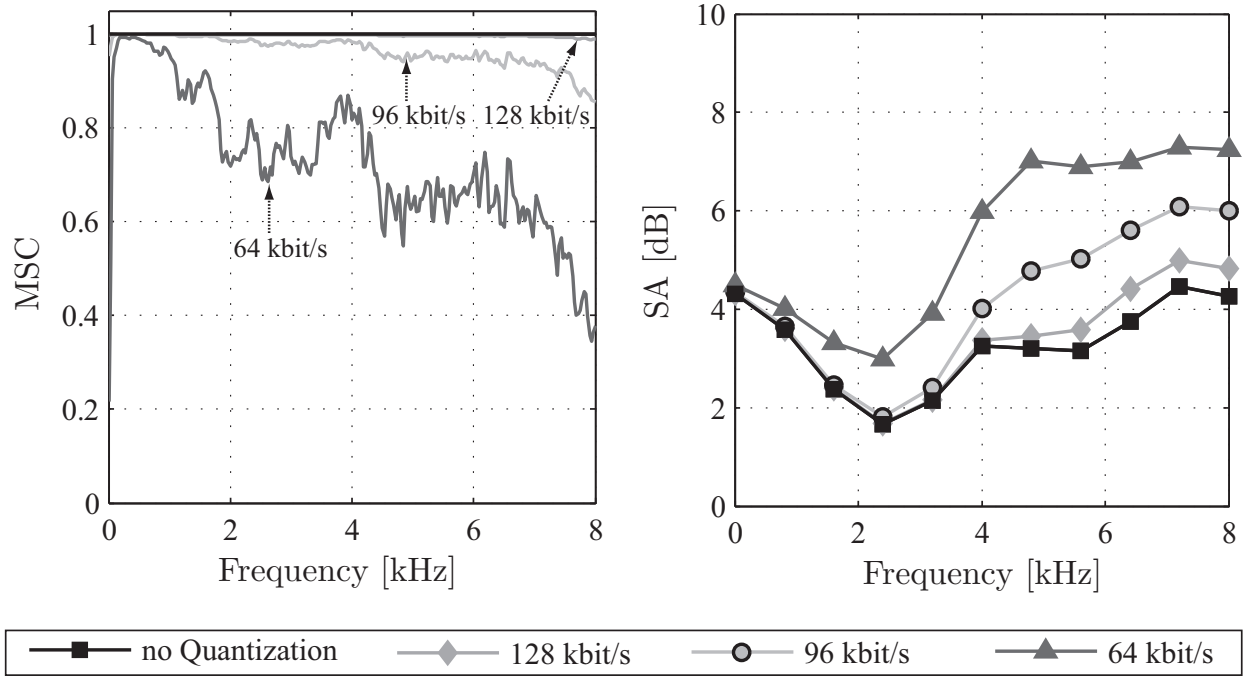


Figure 4.11: Influence of single-sided A-law quantization on MSC of direct speech signal (left) and dereverberation performance (right).

(23.05 kbit/s, 12.65 kbit/s) on MSC and SA is depicted in Fig. 4.12. In the left figure it can be seen that the transcoding with the G.722 leaves the MSC unaffected in the lower frequency range, whereas a small decline can be observed for the higher frequency range. This can be explained with the internal processing in two different subbands with different bit-rates. The AMR-WB codec results in a MSC of nearly zero in the frequency range above 6.4 kHz since the employed artificial bandwidth extension diminishes the correlation of the two signals significantly. From the right plot it can be concluded that only the G.722 is capable of maintaining a sufficient speech quality when transmitting one signal via a wireless data-link.

Furthermore, it has to be considered that the wireless transmission has a significant effect on the operating time of the hearing aids due to the limited battery capacity. Hence, a more intelligent data-link which controls the amount of data which needs to be transmitted and more efficient speech codecs that are adapted to the application demands are necessary to save battery power and to limit the introduced artifacts, see also [HKLP08].

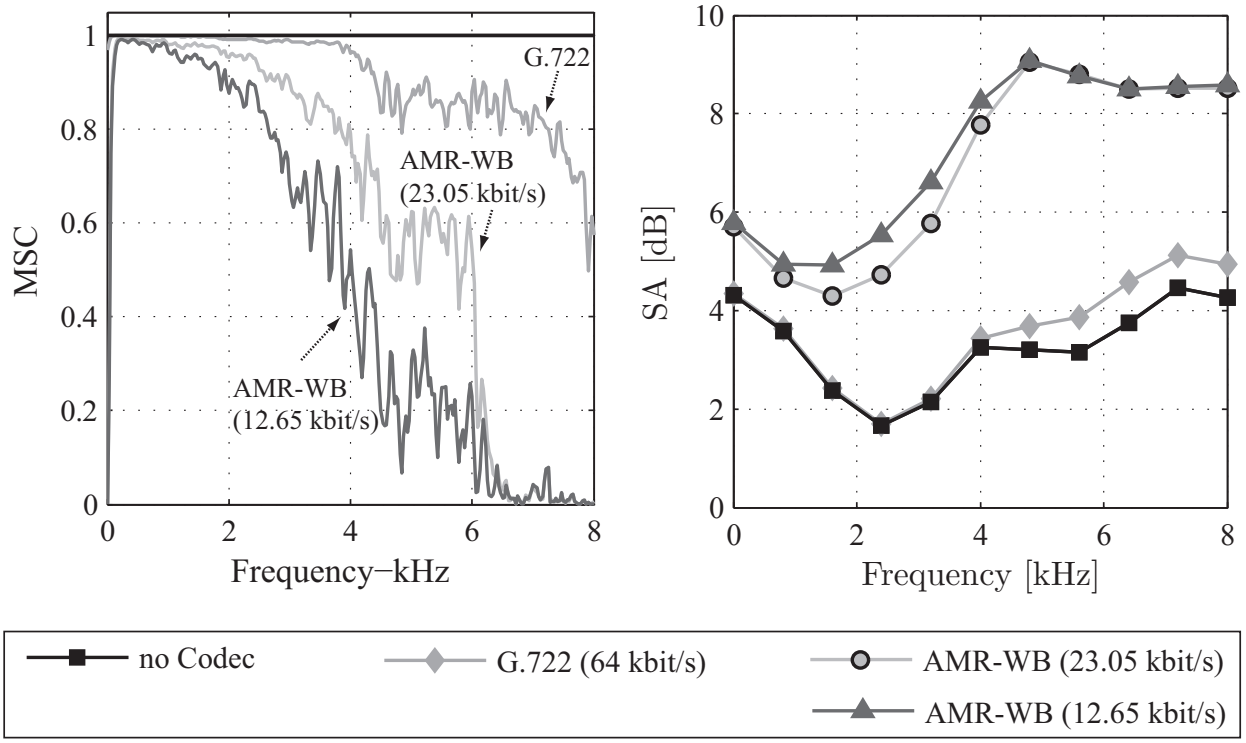


Figure 4.12: Influence of single-sided transmission using G.722 and AMR-WB on MSC of direct speech signal (left) and dereverberation performance (right).

4.2 Speech Enhancement for Dual-Microphone Mobile Phones

In this section, the application of the discussed speech enhancement algorithm for dual-microphone mobile phones is discussed briefly. In order to employ such algorithms, a secondary microphone can be placed either next to the common primary microphone on the bottom of the device or on top of the device (see Fig. 2.15). An analysis of the acoustical environment based on recordings with a dual-microphone mock-up phone mounted on a dummy head was presented in Sec. 2.2.2. In contrast to related dual-channel noise reduction systems such as [GLBF03], the new approach allows for a scalable extension of an existing single-channel noise suppression system by exploiting a secondary microphone channel for a more robust noise PSD estimation and improved spectral weighing rule.

For further discussion and evaluation results of state-of-the-art noise PSD estimators and spectral weighing rules for both types of microphone alignments, i.e., *Bottom-Bottom* (BB) and *Bottom-Top* (BT), please refer to [Her11]. Here, we restrict the discussion to the most promising dual-microphone BT configuration for mobile phones and the hand-held position.

4.2.1 Speech Enhancement System

The novel speech enhancement system which operates in the short-term Fourier domain is depicted in Fig. 4.13. Based on the two input signals $X_1(\lambda, \mu)$ and $X_2(\lambda, \mu)$, a noise PSD estimate $\widehat{\Phi}_{\mathrm{nn}}(\lambda, \mu)$ is calculated using the proposed PLDNE algorithm presented in Sec. 3.2.1.1. From the input signals and the noise PSD estimate, spectral weighting gains $G'(\lambda, \mu)$ are calculated by means of the PLD gain algorithm (see Sec. 3.2.2). For reducing the amount of musical tones, spectral smoothing over frequency of the magnitudes $G'(\lambda, \mu)$ is performed using the smoothing procedure outlined in Sec. 3.4.1. The final gains $G(\lambda, \mu)$ are then applied to the primary input signal.

Please note that no additional estimator for the late reverberant speech PSD is employed since the PLD system is already capable of reducing room reverberation as shown later. In order to fulfill the delay requirements for mobile phones, cf. [DB08], a filtering in the time-domain by means of a filter-bank equalizer is also possible, cf. [LV08b].

4.2.2 Performance Evaluation

Since an evaluation of the PLD system in terms of the noise reduction performance and estimation accuracy has already been performed in Sec. 3.2.2, we restrict the following experiments to the dereverberation performance only. The input signals for the dual-channel algorithm were obtained by convolving a speech signal with

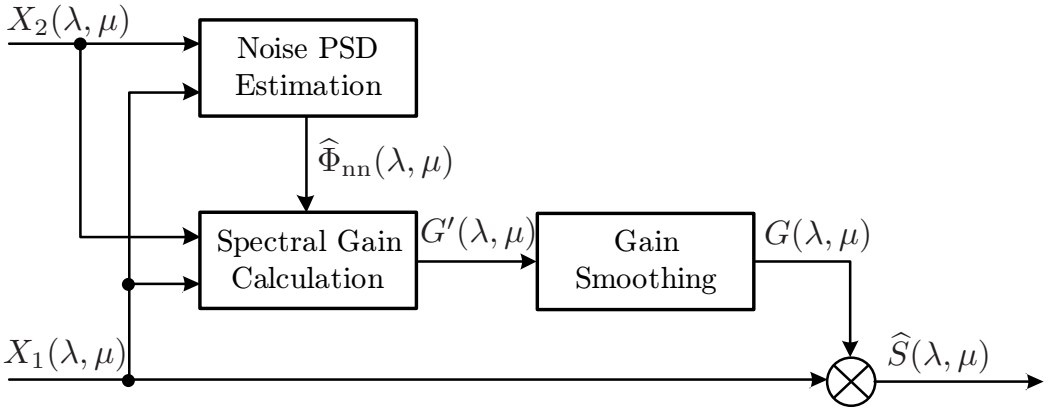


Figure 4.13: Schematic diagram of the proposed dual-channel noise reduction system for mobile phones.

Table 4.3: Simulation results of dereverberation performance of PLD system. The DRR values are determined from the impulse responses used to generate the reverberant input signals.

Room	input DRR (top mic.)	input DRR (bottom mic.)	ΔSRMR
Office	0.94 dB	14.92 dB	0.10
Stairway	2.68 dB	15.72 dB	0.16
Corridor	0.28 dB	13.96 dB	0.47

two different RIRs measured with the BT mock-up phone and a dummy head. The first RIR was measured between artificial mouth and the bottom microphone and the second RIR between mouth and the top microphone. The input DRR values of the RIRs measured at the two microphones and the results in terms of SRMR improvement are listed in Table 4.3. Since the direct path energy is already very high compared to the reverberant energy, only moderate improvements in terms of the SRMR measure could be obtained. In the subjective listening impression, the long reverberant tail was completely removed and the listening comfort was greatly improved after the processing.

A further experiment with a noisy and reverberant signal showed the same tendency. Hence, we conclude that the proposed dual-channel speech enhancement for mobile phones is capable of reducing background noise and room reverberation.

Furthermore, a possible cross-talk between the (hand-held mode) loudspeaker and the secondary microphone was investigated and the measurement-based simulations have shown that this coupling does not cause any audible distortions. Additionally, the algorithm is robust towards different time-delays of arrival of the two microphone signals regarding the typical dimensions of a mobile phone and is also robust towards a possible unintended covering of the top microphone by the user.

Figure 4.14: Pictures of the German Bundestag[2]. (left) Audience where the lines mark possible sound propagation paths from the loudspeaker system to the microphones at the lectern; (right) Speaker at lectern with two microphones.

4.3 Dereverberation for Recordings Taken in the German Parliament

Along with the two main application areas, a case study is presented to show how dereverberation can be applied to enhance *single-channel* speech recordings such as taken in the German parliament[1]. Based on the psychoacoustically-motivated dereverberation concept (see Sec. 3.4.2), a significant improvement in terms of the perceived quality is obtained in comparison to a conventional dereverberation approach [JV11]. Since time-varying changes of the acoustical environment are negligible, all required acoustical parameters such as RT and DRR, are determined in an off-line procedure.

Room reverberation is usually caused by reflections of the emitted source, e.g., a speaker which stands far away from another speaker in an enclosure. In contrast to that, in a parliament discussion, the speaker is located at a lectern and the speech is captured by microphones at a small distance. This speech signal is processed and emitted by a loudspeaker system to the audience. In very large rooms, this signal is then reflected on the walls and fed back into the microphones with a certain sound propagation delay. In order to avoid instability and overshoots, a feedback cancellation is employed which usually consists of a notch filter or an adaptive filter with a short filter length. However, the captured reverberation remains if no further countermeasures are employed. This acoustic situation is illustrated for the German parliament in the left subfigure of Fig. 4.14 where simplified sound propagation paths are marked by the dashed lines. A speaker standing at the lectern is shown in the right subfigure where the short distance to the capturing microphones can clearly be seen. Even though two microphones are mounted on the lectern, only one microphone is used in this setup since the second one is used only as a microphone breakdown replacement.

[1]The author would like to thank Steven Rösler from the German Bundestag parliament television for providing the audio recordings.

[2]Photographic material provided by the digital image service of the German Bundestag. (c) Werner Schüring (left) and Thomas Trutschel/photothek.net (right).

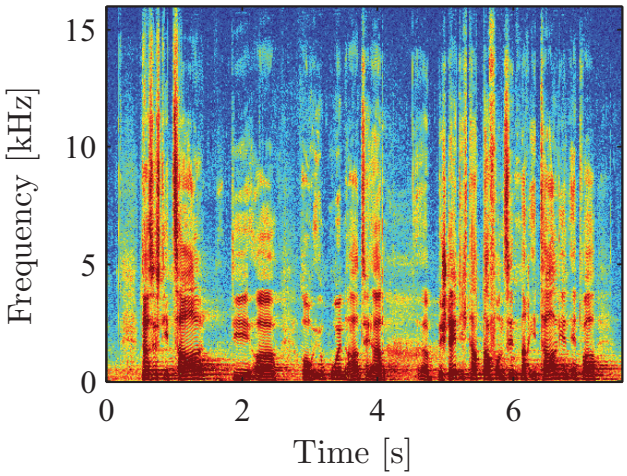


Figure 4.15: Spectrogram of a recording without processing ($x(k)$).

In the spectrogram of a short recording in Fig. 4.15, the smearing over time due to reverberation can clearly be seen. Since the acoustical scenario does not change, apart from small movements of the speaker and the audience, all required acoustical parameters for the considered speech enhancement algorithm such as RT and DRR have to be estimated only once. This off-line procedure is carried out blindly from recorded data since no acoustical RIR measurements are available.

In this application scenario a modified ML approach based on [LYJV10] is used in a preceding off-line procedure. The RT is estimated in speech offset periods only which are determined by a VAD. For all obtained speech offset segments which are larger than 200 ms, the ML procedure is applied and the results are averaged. From this estimation procedure, an average reverberation time of 0.86 s was obtained in the considered scenario, using 45 min. of speech material. The DRR is determined off-line by measuring the energy drop of the signal after a sharp speech offset by manual segmentation. The resulting DRR was measured as 18 dB, which indicates that the source is located within the critical distance.

As employed in the speech enhancement system for binaural hearing aids, we take advantage of the late reverberant speech PSD estimator HB ([HGC09]) as discussed in Sec. 3.1.1.1. Here, only the single-channel case is considered and the extension by the psychoacoustic-motivated weighting rule discussed in Sec. 3.4.2 is used.

4.3.1 Performance Evaluation

In a first step, the above mentioned off-line procedure to determine the RT and DRR was carried out. In a second step, the single-channel recordings were processed with and without the psychoacoustical extension of the conventional dereverberation algorithm using Eq.(3.65) and Eq.(3.11), respectively. The corresponding time-domain signals are termed $\tilde{x}(k)$ and $\hat{x}(k)$ (Please refer to Fig. 3.20 for the signal naming of the intermediate signals). Further important simulation parameters are listed in Table 4.4.

Table 4.4: Main simulation parameters.

Parameter	Settings
Sampling frequency	$f_s = 32\,\text{kHz}$
Frame length	$L = 640$ (20 ms)
FFT length	$M = 1024$ (including zero-padding)
Frame overlap	50% (Hann window)
Smoothing factor	$\alpha^{(\text{xx})} = 0.9$
Reverberation time	$\text{T}_{60} = 0.86\,\text{s}$
Late reverberant time span	$\text{T}_l = 0.1\,\text{s}$
DRR	$\text{DRR} = 18\,\text{dB}$
Interference attenuation factor	$\zeta_{\text{int}} = -15\,\text{dB}$
Gain threshold	$G_{\min} = 0.1$

Table 4.5: Dereverberation performance in terms of the non-intrusive quality measure SRMR.

	Pre-dereverberated $\widetilde{x}(k)$	Dereverberated $\widehat{x}(k)$
ΔSRMR	0.7	0.83

Since neither the room impulse response nor any anechoic reference signal is available, only non-intrusive objective quality measures can be used. Table 4.5 shows the results in terms of the SRMR difference measure where a high improvement can be observed. The psychoacoustic extension even leads to an increase in SRMR compared to the pre-dereverberated signal. The corresponding spectrograms of the enhanced signals are shown in Fig. 4.16. It can be seen that due to the psychoacoustic weighting (right subfigure), random fluctuations, i.e., musical tones, could be reduced significantly. The reduction of musical tones and further artifacts was confirmed by the subjective listening impression.

In future applications, the developed algorithm might be used for a post-processing of recorded data, e.g., for news broadcast or as a plug-in for video players such as the VLC. Besides, the enhanced recordings can be used for archival storage and documentation. Audio and video demonstrations are available online[2].

[2]Download link: `http://www.ind.rwth-aachen.de/~bib/jeub11b`

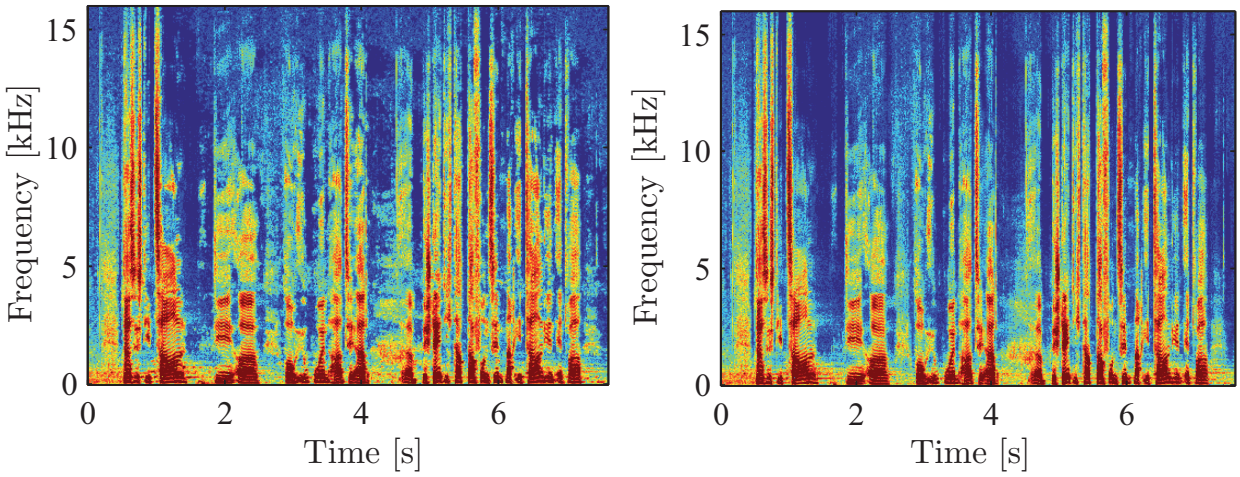


Figure 4.16: Spectrogram of processed speech: (left) with pre-dereverberation only ($\widetilde{x}(k)$); (right) with dereverberation using psychoacoustic weighting ($\widehat{x}(k)$).

4.4 Summary

In this chapter, the application of the introduced concepts and algorithms for dereverberation and noise reduction were discussed. Two main examples for speech enhancement applied to binaural hearing aids, dual-microphone mobile phones were given. Furthermore, a case study was presented to show how dereverberation can be applied for a processing of recorded speech in a large plenary hall.

The new aspects of this chapter are briefly:

- a novel and efficient speech enhancement system for binaural hearing aids including discussions on:
  - the extension of monaural algorithms to binaural output and the proposal of a new concept based on a reference signal,
  - investigations how bilateral dereverberation influences the binaural cues based on objective measures and a listening experiment,
  - proposal of a two-stage speech enhancement system where all required acoustic parameters are estimated blindly from the reverberant and noisy input signals which does not alter the binaural cues,
  - separate experiments on noise reduction, dereverberation as well as joint noise reduction and dereverberation performance to show the superiority compared to related approaches in terms of objective quality measures and subjective listening impression,
  - proof that the usage of an accurate model of the noise field coherence greatly reduces the occurring speech distortions,
  - investigations on the wireless data-link.
- proposal of a generalized speech enhancement scheme for dual-microphone mobile phones which is capable to reduce background noise and room reverberation and shows a better performance than state-of-the-art methods.
- case study on the application of speech dereverberation for enhancing speech recordings taken in the German parliament
  - adaptation of a psychoacoustically-motivated spectral weighting rule known from noise reduction.

This chapter has demonstrated the effectiveness of the proposed algorithms and concepts in important and realistic applications. All methods have the potential to be integrated into future binaural hearing aids and dual-microphones mobile phones.

5

Summary

Within this work, new algorithms and concepts enabling a joint reduction of room reverberation *and* environmental background noise for speech communication systems under adverse acoustic conditions were presented. The new strategies are based on measurements and recordings in realistic acoustical environments keeping in mind the specific acoustic conditions for the two main applications, i.e., binaural hearing aids and dual-microphone mobile phones. Along with an acoustic environment analysis, the main outcomes of this thesis are two novel speech enhancement systems which are adopted in particular to the specific application scenarios.

For binaural hearing aid applications, a new two-stage algorithm is proposed. It is based on a coherence model which takes the shadowing effects of the head into account as well as two combined estimators for the late reverberant speech *Power Spectral Density* (PSD) and background noise PSD. A key aspect is that all required acoustic parameters such as the *Reverberation Time* (RT) and *Direct-to-Reverberation Energy Ratio* (DRR) are estimated blindly from the noisy and reverberant input signals. The second algorithm, developed for dual-microphone mobile phones, exploits explicitly the *Power Level Difference* (PLD) of speech and all interfering sources and comprises a new noise PSD estimator and spectral weighting rule.

Analysis and Models of the Acoustic Environment

In the beginning of this thesis, the acoustic properties of the occurring reverberant and background noise conditions were evaluated. For this investigation, the coherence function plays an important role. Different existing models for important noise fields were discussed and a new model for a mixed coherent and diffuse noise field was derived which can be used, e.g., for noise field classification algorithms. In the context of binaural hearing aids, it is especially important to take the occurring shadowing effects of the head on the coherence into account.

The considered speech communication devices have to deal with different acoustic situations which are, in case of reverberation, mainly determined by the source-microphone distance. It was shown that the DRR and RT are highly dependent on the frequency for larger rooms and for most situations, the source can be assumed

to be within the so-called critical distance. The common assumption of a homogeneous and diffuse background noise field was verified by recordings in realistic acoustic environments.

Furthermore, two important estimators which allow for a blind (frequency-dependent) estimation of the RT and DRR were presented and evaluated. Moreover, it was demonstrated based on a listening experiment that room reverberation can lead to a decrease in listening comfort even for hand-held telephony under certain conditions which is usually assumed to be negligible.

As an additional outcome, a large database of all measured room impulse responses, the so-called *Aachen Impulse Response* (AIR) database, was developed and is available online.

Joint Dereverberation and Noise Reduction

In the second main chapter, dereverberation and noise reduction were considered independently first, followed by discussions how to efficiently combine different algorithms and estimation techniques to allow for a joint reduction of reverberation and background noise. This comprises also single-channel algorithms which have been extended to two input and two output channels.

The dereverberation section has started with an evaluation and improvement of different classes of state-of-the-art algorithms which require different properties for the input signal in terms of the DRR. First, single-channel methods which estimate the late reverberant speech PSD were introduced. In the evaluations, it turned out that an algorithm which is based on a generalized model of the *Room Impulse Response* (RIR) shows the best performance. The algorithm has been extended by means of two blind estimators of the required parameters: RT and DRR. By further incorporating a frequency-dependent RT estimate, the estimation error could be reduced by up to 2.5 dB in the higher frequency range. Second, dual-channel algorithms which exploit the coherence of the two input signals were introduced and improved. To ensure a low speech attenuation, it is beneficial in this case to take accurate models of the noise field coherence into account. In terms of diffuse noise fields, a performance gain of up to 4 dB (NA-SA) was obtained my means of an appropriate coherence model. The use of a model which takes the shadowing of the head into account can even further reduce the speech attenuation in the lower frequency range by 2 dB (SA) in binaural applications.

For the reduction of background noise, two novel short-term noise PSD estimators were introduced. At first, a new generalized expression of a known dual-channel estimator is derived which allows to incorporate different coherence models. In doing so, the logarithmic estimation error could be reduced by 1 dB in average and up to 8 dB in the lower frequency range. For inter-microphone distances larger than 0.15 m, it outperforms state-of-the-art single-channel algorithms such as *Minimum Statistics* (MS) and a *Speech Presence Probability* (SPP)-based method. Second, a new concept which exploits the PLD of speech and noise was used to derive a low-complexity dual-channel noise PSD estimator. Also with this new concept, improvements compared to

related approaches could be achieved. Moreover, a modified spectral weighting rule is derived which is also based on the considered power level differences and the noise field coherence. The algorithm requires only a low computational complexity and can efficiently be implemented using first order IIR filters for the auto- and cross-PSD estimation. Experiments have shown that the novel algorithm is capable of reducing unwanted background noise as well as reverberation. Compared to related dual-channel approaches, a performance gain of $5-8\,\mathrm{dB}$ (NA-SA) was measured even under adverse acoustic conditions.

In the sequel, the combination of dereverberation *and* noise reduction was examined. First, it was shown that background noise can have a severe impact on the RT and DRR estimation accuracy and that reverberation increases the estimation error of background noise PSD estimation algorithms. Based on these investigations, an advanced concept how to jointly reduce reverberation and background noise by means of an interlaced combination of different algorithms and estimation techniques was proposed. The new structure allows now to estimate the required RT and DRR even under adverse background noise conditions.

Applications

The introduced concepts for dereverberation and noise reduction have been adopted and extended to the two main application scenarios:

- The first application example regards binaural *hearing aids* where a novel two-stage algorithm is presented which efficiently reduces early reverberation, late reverberation as well as additive background noise. The algorithm operates in the frequency domain and consists of two components: The first stage of the algorithm is based on a statistical model of the RIR and comprises a spectral weighting rule which depends on the short-term PSD of the late reverberant speech and background noise. In a second stage, the residual reverberation and noise is attenuated by a dual-channel Wiener filter which is based on a new coherence model taking into account head shadowing.

  Furthermore, a novel smoothing procedure of the spectral gains over frequency is integrated to reduce musical tones. The overall binaural input-output structure does not affect the most important binaural cues, i.e., *Interaural Time Difference* (ITD) and *Interaural Level Difference* (ILD), and hence, keeps the localization ability. This was motivated by investigations how state-of-the-art dereverberation algorithms influence the binaural cues in bilateral processing. It has further been shown that the majority of listeners prefer binaurally over bilaterally processed signals.

  Modern binaural hearing aids are capable of a wireless data exchange between both devices. This allows to employ sophisticated speech enhancement algorithms which can exploit spatial information and preserve the binaural cues. We have investigated the influence of a wireless data-link on the coherence and the attenuation of the desired speech components. Based on a low-complexity transmission with A-law quantization and two speech codecs, i.e., G.722 and AMR-WB, it was

shown that the coherence can be greatly affected by a wireless transmission. For the AMR-WB and low data-rates, the speech attenuation was increased up to 4 dB. It is suggested that, at least for the considered configuration, the G.722 at a data-rate of 64 kbit/s can ensure an unaffected processing using the abovementioned binaural two-stage system.

- The second example of use are dual-microphone *mobile phones* where the algorithm explicitly exploits the special acoustic characteristics of the *Bottom-Top* (BT) microphone configurations. By taking into account the different power level differences of speech, noise and reverberation, an effective concept is proposed which is advantageous compared to state-of-the-art methods. Related dual-channel methods were outperformed by $10 - 20$ dB (NA-SA), depending on the input SNR. The performance gain due to a secondary microphone can only be fully exploited in the bottom-top configuration which should be the design target for any future mobile communication device.

In an additional case study, we have demonstrated how the concept of speech dereverberation can be used to enhance single-channel speech recordings taken in large plenary halls such as the German parliament. Based on a new psychoacoustically-motivated dereverberation concept, which was adopted from noise reduction, the listening comfort and intelligibility could be increased significantly. In future applications, the developed algorithm might be used for a post-processing of recorded data, e.g., for news broadcast.

In conclusion it can be stated that this thesis has demonstrated how to efficiently tackle the problem of room reverberation *and* environmental background noise even under adverse acoustic conditions. The emerging technologies allow to jointly reduce early *and* late reverberation as well as additive background noise and are capable of improving the listening comfort and the speech intelligibility for hearing aid and mobile phone users.

A

AIR Database and Acoustic Measurements

A.1 Speech and Background Noise Databases

Table A.1: Speech and background noise databases.

Database	Description	Reference
NTT	Speech database (21 languages, $f_s = 16\,\text{kHz}$)	[NC94]
TSP	Speech database (english speakers, $f_s = 48\,\text{kHz}$)	[Kab02]
ETSI	Background noise database	[ETS09]
	(Stereo and binaural in real rooms, $f_s = 48\,\text{kHz}$)	
NOISEX-92	Background noise database	[VS93]
	($f_s = 19.98\,\text{kHz}$)	

A.2 Acoustic Measurement System

Table A.2: Measurement equipment.

Device	Description
RME Multiface II	Soundcard
RME Octamic	Microphone amplifier
Genelec 8130A	Loudspeaker
	(RIR measurements and four-loudspeaker noise generation)
Genelec 7050B	Subwoofer (used for four-loudspeaker noise generation)
Beyerdynamic MM1	Measurement microphone
	(RIR measurements and mock-up phone)
HEAD acoustics HMS II.3	Dummy head with mouth simulator
HEAD acoustics HMS II.5	Dummy head w/o mouth simulator
HEAD acoustics HHP III	Dummy head handset positioner
HEAD acoustics PEQ V	Headphone equalizer (Listening tests)
Sennheiser HD 600	Stereo headphones (Listening tests)

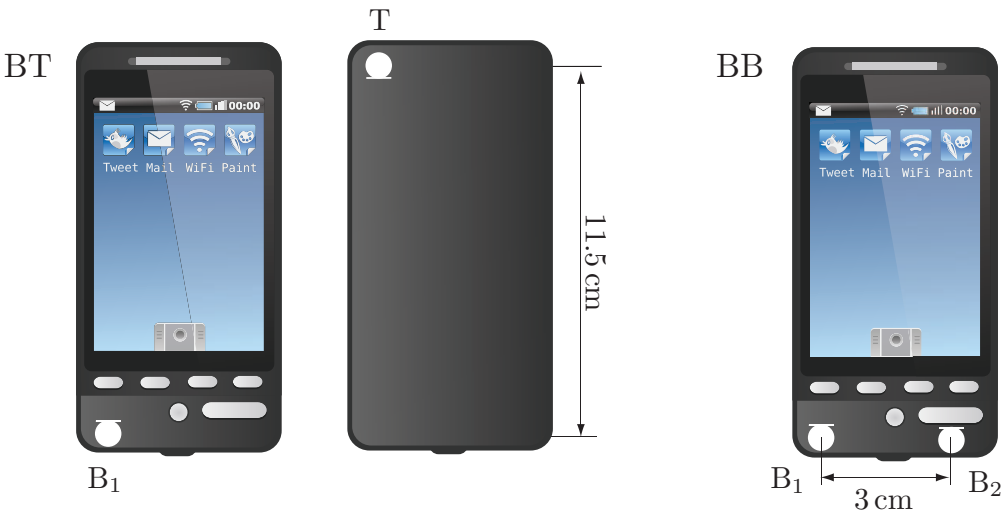


Figure A.1: Illustration of mobile-phone with the two considered microphone positions. (a) bottom-top (BT) alignment, (b) bottom-bottom (BB) alignment.

A.3 AIR Database

Table A.3: Room impulse response databases.

Database	Description	Reference
AIR	Binaural and mobile phone RIRs in different rooms[1] ($f_s = 48\,\text{kHz}$)	[JSV09]
MARDY	Eight-channel RIRs in acoustic lab ($f_s = 48\,\text{kHz}$)	[WGH+06]
Oldenburg	Binaural RIRs and background noise recordings (three microphone hearing aids, $f_s = 48\,\text{kHz}$)	[KEA+09]
ITU G.191	Single-channel and stereo RIRs in different rooms ($f_s = 32/48\,\text{kHz}$)	[ITU09]
Nierrhein	Single-channel RIRs in different rooms and cars ($f_s = 16\,\text{kHz}$)	[Kit10]

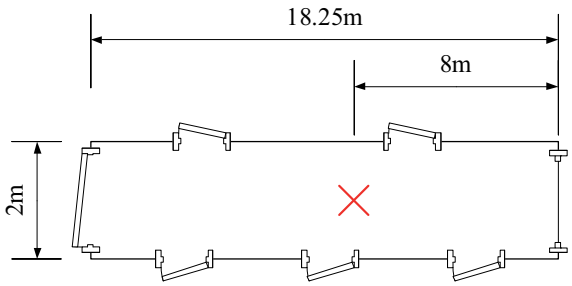
Corridor

Binaural

Dist. [m]	RT [s]	DRR [dB]
×	×	×

Mobile Phone

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	0.98/1.34	12.78/6.51
BT (HHP)	1.12	12.71



[1] Please note that the various measurements have been conducted over a time-span of 3 years where the acoustical environment might have been changed due to new furniture and the recordings at different positions in the room. Download link: `http://www.ind.rwth-aachen.de/air`

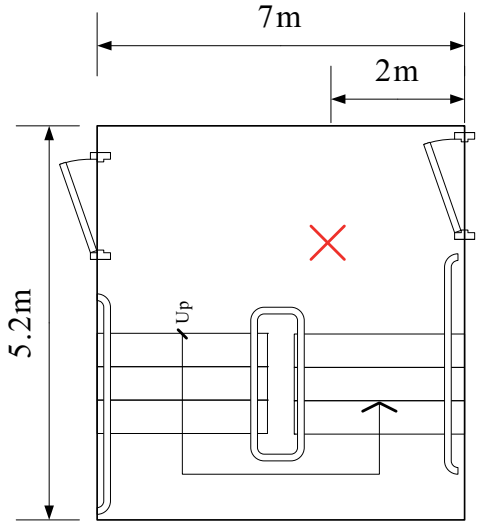
Stairway

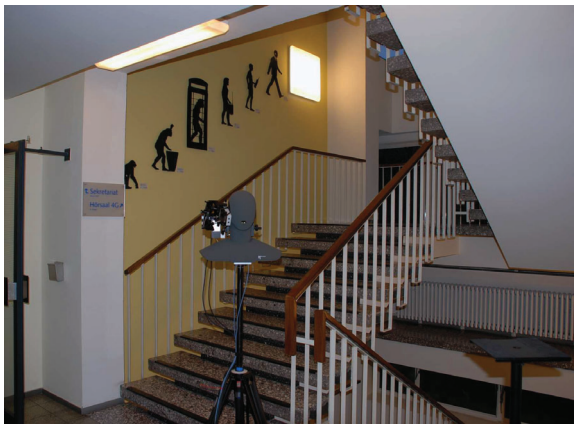
*Binaural**

Dist. [m]	RT [s]	DRR [dB]
1	0.72	9.58
2	0.83	3.98
3	0.90	0.51
	ø0.82	

Mobile Phone (different location)

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	1.31/1.52	10.87/4.7
BT (HHP)	0.89	13.75





* All binaural RIRs measured with dummy head at different azimuth angles $(-90, 15, ..., +90^{\circ})$

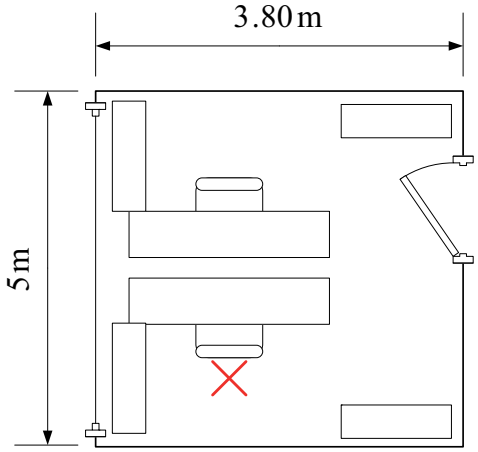
Office Room

*Binaural**

Dist. [m]	RT [s]	DRR [dB]
1	0.45	8.08
2	0.53	2.54
3	0.57	−1.45
	ø0.51	

Mobile Phone

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	0.4/0.52	12.27/5.28
BT (HHP)	0.52	13.18



* All binaural RIRs measured with and without dummy head

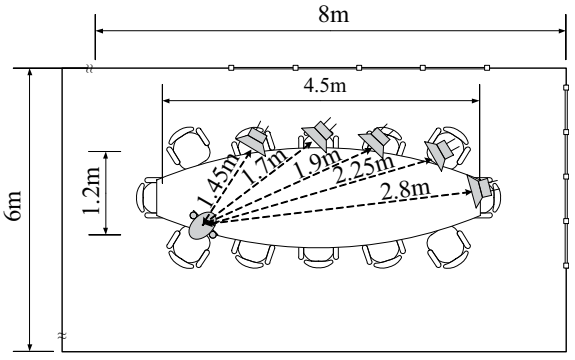
Meeting Room

*Binaural**

Dist. [m]	RT [s]	DRR [dB]
1.45	0.27	8.67
1.7	0.28	6.71
1.9	0.27	6.27
2.25	0.29	6.53
2.8	0.29	6.83
	ø0.28	

Mobile Phone

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	0.16/0.24	14.21/9.34
BT (HHP)	✗	✗



* All binaural RIRs measured with and without dummy head

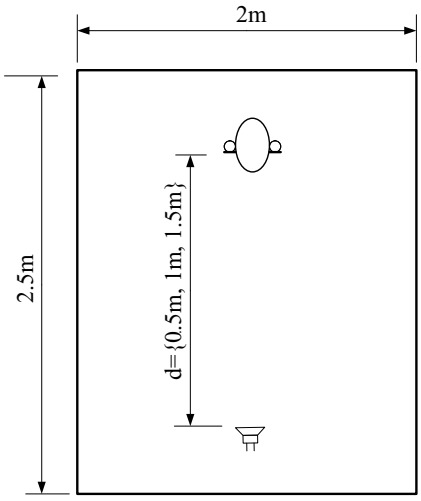
Booth

*Binaural**

Dist. [m]	RT [s]	DRR [dB]
0.5	0.17	9.17
1.0	0.18	4.60
1.5	0.28	3.24
	Ø0.21	

Mobile Phone

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	×	×
BT (HHP)	×	×



* All binaural RIRs measured with and without dummy head

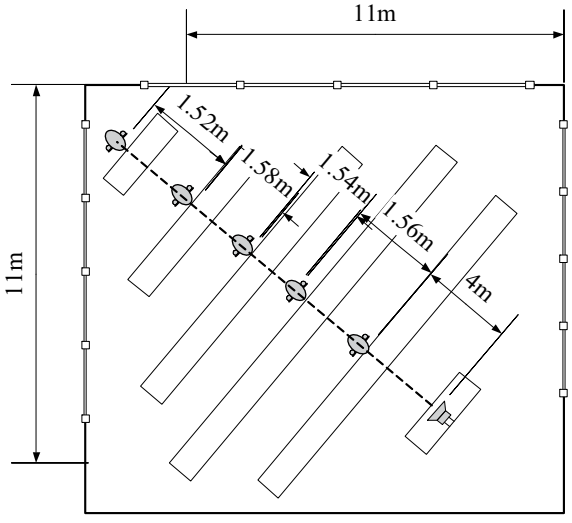
Lecture Room

*Binaural**

Dist. [m]	RT [s]	DRR [dB]
2.25	0.73	6.74
4	0.69	1.75
5.56	0.83	−1.45
7.1	0.85	−3.64
8.68	0.86	−6.38
10.2	0.87	−3.83
	ø 0.81	

Mobile Phone

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	0.15/0.57	14.49/11.18
BT (HHP)	×	×



*All binaural RIRs measured with and without dummy head

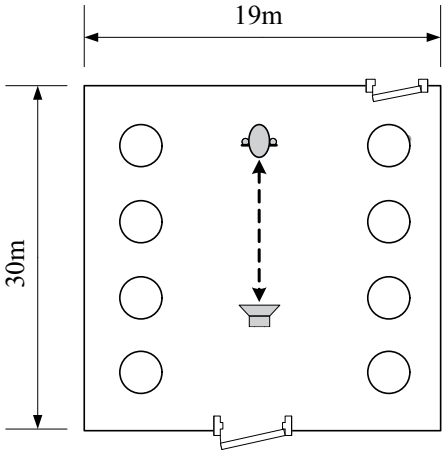
Aula Carolina (Aachen, Germany)

*Binaural**

Dist. [m]	RT [s]	DRR [dB]
1	4.01	8.16
2	4.30	3.45
3**	4.82	0.36
5	5.16	0.53
10	7.37	−4.7
15	5.93	−2.11
20	6.66	−3.83
	ø5.47	

Mobile Phone

Device	RT [s]	DRR [dB]
BB (HHP/HFRP)	×	×
BT (HHP)	×	×



* All binaural RIRs measured with dummy head
** At 3 m distance including different azimuth angles $(-90, 45, ..., +90°)$

B

Binaural Coherence of Noise Fields

A semi-analytical signal processing model for the binaural coherence of homogeneous isotropic noise fields is presented in this section. The model is based on the original work by [Dör98] and has been republished in [JDV11]. It is derived from a simplified geometrical model of the human head, where the shadowing between the left and right ear is modeled by two non-reflecting circular plates. Based on Kirchhoff's diffraction theory, it is shown how the corresponding coherence is calculated. This model can be used as part of various binaural signal processing algorithms, such as speech enhancement for digital hearing aids or binaural speech transmission systems. In the experiments in Sec. 2.2.1.1, it was confirmed that the proposed theoretical model shows a good match with the coherence obtained from measurements in a highly reverberant environment. A MATLAB reference implementation is available online[1].

B.1 Geometric Head Diffraction Model

A simplified geometric model for the complex head geometry according to Fig. B.1 is used in the following. It is assumed that the two microphones M_1 and M_2 are placed at distance d_{mic} next to the pinna of each side. The resulting shadowing is modeled by two non-reflecting circular plates (P_1 and P_2) with radius r_{head} and distance d_{head}. We further define a punctual sound source Q by its position vector $\underline{r}_q$ or angles θ (azimuth) and φ_{q} (elevation) in the far-field and H_1 and H_2 the transfer functions between Q and the two microphones M_1 and M_2. The position vector of M is denoted by $\underline{r}_{\mathrm{m}}$. The point of origin is marked with 0 in the figures. Besides, any point on the plate is given either by position vector $\underline{r}$ or $r = ||\underline{r}||$ and φ_r. All further variables will be introduced below successively. Assuming two omnidirectional microphones, the spherically isotropic coherence can be calculated for a homogeneous noise field by the integration over all possible directions of incident of a directional sound source. Since this procedure requires knowledge about the transfer functions $H_{1|2}$, a derivation based on optical principles will be given in the following.

[1] Download link: `http://www.ind.rwth-aachen.de/~bib/jeub11`

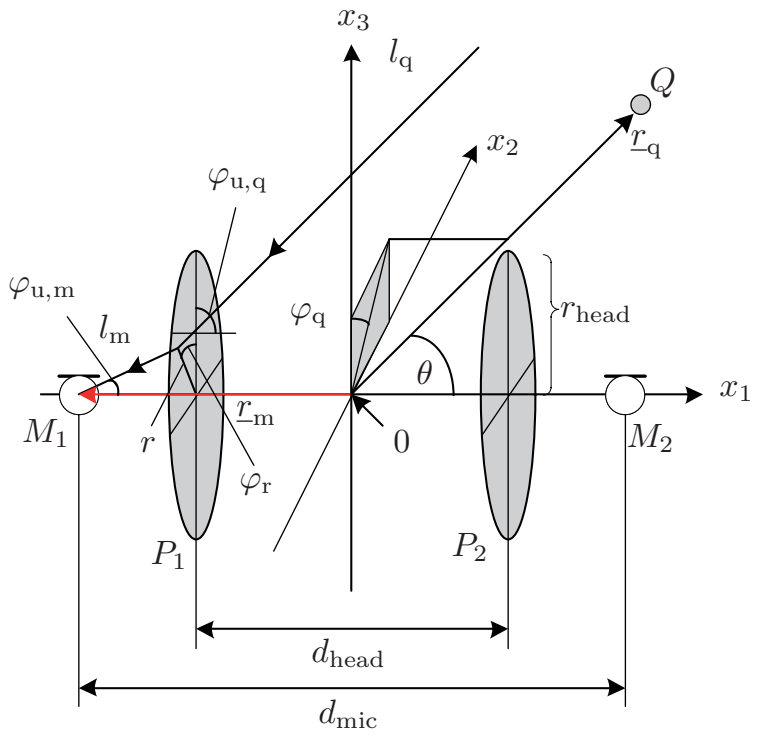


Figure B.1: Simplified geometrical model of the human head with two circular plates $P_{1|2}$. Microphones are denoted by $M_{1|2}$.

B.2 Kirchhoff's Diffraction Theory

Kirchhoff's diffraction theory was initially developed to explain optical phenomena in terms of diffraction. However, this general theory can also be applied to sound waves. Consider a monochromatic wave which propagates from a punctual source Q through the opening of a screen (pinhole) S_1 as depicted in Fig. B.2 (a), where M_1 is the point at which the disturbance is to be determined. The distances between opening and Q and opening and M_1 are denoted by l_q and l_m respectively. In the following we assume that the opening is small compared to the distance of both Q and M_1 from the obstacle and show how the corresponding transfer function between Q and M_1, i.e., $H_\mathrm{qm}(f)$ is derived. Based on the Kirchhoff diffraction theory, the potential at location M_1 can be expressed by the *integral theorem of Helmholtz and Kirchhoff* as

$$\phi(\underline{r}_\mathrm{m}) = \frac{1}{4\pi} \int\limits_S \left[\phi(\underline{r}) \frac{\partial}{\partial u} \left(\frac{e^{-j\beta l_\mathrm{m}}}{l_\mathrm{m}} \right) - \frac{e^{-j\beta l_\mathrm{m}}}{l_\mathrm{m}} \frac{\partial \phi(\underline{r})}{\partial u} \right] \mathrm{dS} \tag{B.1}$$

where $\int_S \ldots \mathrm{dS}$ gives the integration over a non-reflecting surface S which encloses the point M_1, $\frac{\partial}{\partial u}$ denotes differentiation along the inward normal $\underline{u}$ to the surface of integration and $l_\mathrm{m} = ||\underline{l}_\mathrm{m}||$ is the distance of the element dS from M_1. The wave number is denoted by $\beta = \frac{2\pi f}{c}$. The surface S is formed by three partial surfaces which together form a closed surface, the opening S_1, a portion S_2 at the backside of the obstacle and a large sphere, centered at M_1, S_3. This decomposition allows for

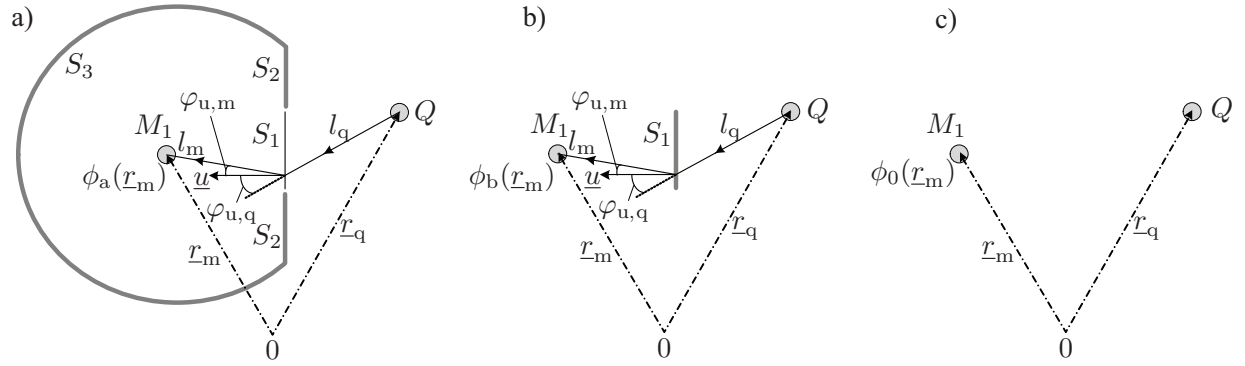


Figure B.2: Illustration of Kirchhoff's diffraction theory and Babinet's principle. (a) pinhole arrangement with potential $\phi_{\mathrm{a}}(\underline{r}_{\mathrm{m}})$ at microphone M_1, (b) inverse arrangement with obstacle and corresponding potential $\phi_{\mathrm{b}}(\underline{r}_{\mathrm{m}})$, (c) free-field arrangement with potential $\phi_0(\underline{r}_{\mathrm{m}})$ Eq.(B.3) as the sum of $\phi_{\mathrm{a}}(\underline{r}_{\mathrm{m}})$ and $\phi_{\mathrm{b}}(\underline{r}_{\mathrm{m}})$. The surface S_1 corresponds to plate P_1 in Fig.B.1.

some important simplifications as shown later. So far the wave propagation between Q and M_1 was disturbed by a pinhole. Now in order to calculate the potential and hence, the transfer function with an obstacle between Q and M_1, Eq.(B.1) could be applied for the case depicted in Fig. B.2 (b). However, since this is inherently difficult, the *Babinet principle* is applied in the following, cf. [BW99]. This theorem gives the relation between the free-field potential $\phi_0(\underline{r}_{\mathrm{m}})$ (Fig. B.2 (c)) and a superposition of the potential $\phi_{\mathrm{a}}(\underline{r}_{\mathrm{m}})$ for the pinhole arrangement (Fig. B.2 (a)) with the potential $\phi_{\mathrm{b}}(\underline{r}_{\mathrm{m}})$ of the complementary arrangement (Fig. B.2 (b)) according to

$$\phi_0(\underline{r}_{\mathrm{m}}) \;=\; \phi_{\mathrm{a}}(\underline{r}_{\mathrm{m}}) + \phi_{\mathrm{b}}(\underline{r}_{\mathrm{m}}). \tag{B.2}$$

Therefore, the potential $\phi_{\mathrm{a}}(\underline{r}_{\mathrm{m}})$ is calculated first, followed by the use of Eq.(B.2) to obtain $\phi_{\mathrm{b}}(\underline{r}_{\mathrm{m}})$. The corresponding free-field potential (without obstacle) at the microphone reads [BW99]:

$$\phi_0(\underline{r}_{\mathrm{m}}) \;=\; C\,\frac{e^{-j\beta||\underline{r}_{\mathrm{q}}-\underline{r}_{\mathrm{m}}||}}{||\underline{r}_{\mathrm{q}}-\underline{r}_{\mathrm{m}}||} \tag{B.3}$$

with constant C. A further difficulty is that the values of $\phi_{\mathrm{a}}(\underline{r}_{\mathrm{m}})$ and $\frac{\partial}{\partial u}$ on the partial surfaces S_1, S_2 and S_3 are never known exactly. Therefore, the following approximations are made which are referred to as the *Kirchhoff boundary conditions*. For S_1 it is assumed that the rim of the opening can be neglected and hence, that the potential will not considerably differ from the values obtained in the absence of the plate (free-field). Hence, it can be written

$$\phi_{\mathrm{a}}^{(S_1)}(\underline{r}) = \phi_0(\underline{r}) = C\,\frac{e^{-j\beta l_{\mathrm{q}}}}{l_{\mathrm{q}}}\,, \tag{B.4}$$

$$\frac{\partial\phi_{\mathrm{a}}^{(S_1)}(\underline{r})}{\partial u} = C\,\frac{e^{-j\beta l_{\mathrm{q}}}}{l_{\mathrm{q}}}\left(-j\beta-\frac{1}{l_{\mathrm{q}}}\right)\cos\varphi_{\mathrm{u,q}}\,. \tag{B.5}$$

Furthermore, the potential and hence, the derivative vanish on S_2, i.e., $\phi_{\mathrm{a}}^{(S_2)}(\underline{r}) = 0$ and $\frac{\partial\phi_{\mathrm{a}}^{(S_2)}(\underline{r})}{\partial u} = 0$. Additionally, the integral over S_3 will vanish by letting the radius

increase indefinitely (see [BW99] for details). With such simplifications and

$$\frac{\partial}{\partial u}\left(\frac{e^{-j\beta l_\mathrm{m}}}{l_\mathrm{m}}\right) = \frac{e^{-j\beta l_\mathrm{m}}}{l_\mathrm{m}}\left(-j\beta - \frac{1}{l_\mathrm{m}}\right)\cdot\left(-\cos\varphi_{\mathrm{u,m}}\right), \tag{B.6}$$

the potential for the pinhole arrangement (Fig. B.2 (a)) can be given with Eq.(B.1) as

$$\begin{aligned}\phi_\mathrm{a}(\underline{r}_\mathrm{m}) = \frac{jfC}{2c}\int\limits_{S_1}\frac{e^{-j\beta(l_\mathrm{q}+l_\mathrm{m})}}{l_\mathrm{q}\,l_\mathrm{m}}\Bigg[\left(1-\frac{jc}{2\pi f\,l_\mathrm{m}}\right)\cos\varphi_{\mathrm{u,m}} \\ + \left(1-\frac{jc}{2\pi f\,l_\mathrm{q}}\right)\cos\varphi_{\mathrm{u,q}}\Bigg]\,\mathrm{dS},\end{aligned} \tag{B.7}$$

which is known as the *Fresnel-Kirchhoff diffraction formula*. The desired transfer function between Q and M_1 arises from the potential by means of a normalization as

$$H_\mathrm{qm}(\Omega) = l_0\cdot\phi_\mathrm{b}(\underline{r}_\mathrm{m}), \tag{B.8}$$

where l_0 denotes a scaling factor such that $||H_\mathrm{qm}|| = 1$ holds in the free-field. Finally, by means of Eqs.(B.2), (B.3), (B.7), this transfer function reads

$$\begin{aligned}H_\mathrm{qm}(\Omega) = \frac{e^{-j\frac{\Omega f_\mathrm{s}}{c}||\underline{r}_\mathrm{q}-\underline{r}_\mathrm{m}||}}{||\underline{r}_\mathrm{q}-\underline{r}_\mathrm{m}||}\,l_0 \\ - \frac{j\Omega f_\mathrm{s} l_0}{4\pi c}\int\limits_{S_1}\frac{e^{-j\frac{\Omega f_\mathrm{s}}{c}(l_\mathrm{q}+l_\mathrm{m})}}{l_\mathrm{q}\,l_\mathrm{m}}\Bigg[\left(1-\frac{jc}{\Omega f_\mathrm{s}\,l_\mathrm{q}}\right)\cos\varphi_{\mathrm{u,q}} \\ + \left(1-\frac{jc}{\Omega f_\mathrm{s}\,l_\mathrm{m}}\right)\cos\varphi_{\mathrm{u,m}}\Bigg]\mathrm{dS},\end{aligned} \tag{B.9}$$

The geometric interpretation of l_q and l_m will be given in the next section.

B.3 Binaural Coherence Model

For each microphone, the transfer function is calculated and, due to the symmetry, the diffraction at the corresponding nearest plate is taken into account. First, it is considered that the angle θ lies in the range $0 \le \theta < \frac{\pi}{2}$ (see Fig. B.1). According to Eq.(B.9), the transfer function between Q and M_1 is given by

$$\begin{aligned}H_1(\Omega) = \frac{e^{-j\frac{\Omega f_\mathrm{s}}{c}||\underline{r}_\mathrm{q}-\underline{r}_\mathrm{m}||}}{||\underline{r}_\mathrm{q}-\underline{r}_\mathrm{m}||}\,l_0 \\ - \frac{j\,\Omega\,f_\mathrm{s}\,l_0}{4\pi c}\int\limits_{P_1}\frac{e^{-j\frac{\Omega f_\mathrm{s}}{c}(l_\mathrm{q}+l_\mathrm{m})}}{l_\mathrm{q}\,l_\mathrm{m}} \\ \cdot\left[\left(1-\frac{jc}{\Omega f_\mathrm{s} l_\mathrm{q}}\right)\cos\varphi_{\mathrm{u,q}} + \left(1-\frac{jc}{\Omega f_\mathrm{s} l_\mathrm{m}}\right)\cos\varphi_{\mathrm{u,m}}\right]\mathrm{dP}.\end{aligned} \tag{B.10}$$

As previously $\int_{P_1} \ldots \mathrm{dP}$ indicates the integration over a surface, here, of plate P_1. $\underline{r}_\mathrm{q}$ and $\underline{r}_\mathrm{m}$ are the position vectors, such that the distance between source and microphone is given by $||\underline{r}_\mathrm{q} - \underline{r}_\mathrm{m}||$. According to Fig. B.1, it can be written

$$||\underline{r}_\mathrm{q} - \underline{r}_\mathrm{m}|| \;=\; r_\mathrm{q} + \frac{d_\mathrm{mic}}{2}\,\cos\theta, \tag{B.11}$$

where $r_\mathrm{q} = ||\underline{r}_\mathrm{q}||$ is the distance from Q to the point of origin. This equation holds since the distance from the source is assumed large compared to the distance of the microphones, i.e., $r_q \gg d_{mic}$. The distance l_q in Eqs.(B.9) and (B.10), which corresponds to the distance of the sound wave from the source to the point specified by (r, ϕ_r) of the plate P_1, reads

$$l_\mathrm{q} \;=\; r_\mathrm{q} + \frac{d_\mathrm{head}}{2}\,\cos\theta - r\,\sin\theta\,\cos(\varphi_\mathrm{r} - \varphi_\mathrm{q})\,. \tag{B.12}$$

The distance between this specific point on the plate and microphone M_1 can be expressed by

$$l_\mathrm{m} \;=\; \sqrt{r^2 + \left(\frac{d_\mathrm{mic} - d_\mathrm{head}}{2}\right)^2}. \tag{B.13}$$

Additionally, the incident and emergent angles in Eq.(B.10) can be written according to Fig. B.1 due to the far-field assumption as

$$\cos\varphi_\mathrm{u,q} \;=\; \cos\theta \quad \text{and} \quad \cos\varphi_\mathrm{u,m} \;=\; \frac{d_\mathrm{mic} - d_\mathrm{head}}{2\,l_\mathrm{m}}. \tag{B.14}$$

Besides, the distances l_q and $||\underline{r}_\mathrm{q} - \underline{r}_\mathrm{m}||$ can be replaced by r_q, unless for such arguments in the exponential function. Taking further into account that $1 - \frac{jc}{\Omega\, f_\mathrm{s}\, l_\mathrm{q}} \approx 1$, the transfer function can be expressed by means of polar coordinates with $\mathrm{dP} = \mathrm{r}\cdot\mathrm{d}\varphi_\mathrm{r}\cdot\mathrm{dr}$ and after rearranging as

$$\begin{aligned} H_1(\Omega,\theta) = {} & \frac{e^{-j\frac{\Omega f_\mathrm{s}}{c}\,r_\mathrm{q}}}{r_\mathrm{q}/l_0} \cdot \Bigg[e^{-j\frac{\Omega f_\mathrm{s}}{c}\,\frac{d_\mathrm{mic}}{2}\cos\theta} \\ & - \frac{j\,\Omega f_\mathrm{s}}{4\pi c} \int_0^{r_\mathrm{head}} \int_0^{2\pi} e^{-j\frac{\Omega f_\mathrm{s}}{c}\left(\frac{d_\mathrm{head}}{2}\cos\theta - r\sin\theta\cos\tilde{\varphi} + l_\mathrm{m}\right)} \\ & \cdot \left[\cos\theta + \left(1 - \frac{jc}{\Omega\, f_\mathrm{s}\, l_\mathrm{m}}\right)\frac{d_\mathrm{mic} - d_\mathrm{head}}{2 l_\mathrm{m}}\right]\frac{r}{l_\mathrm{m}}\,\mathrm{d}\tilde{\varphi}\,\mathrm{dr}\Bigg], \end{aligned} \tag{B.15}$$

where l_m is dependent on r as in (B.13). The derivation of Eq.(B.9) was performed by means of Babinet's principle where the diffraction at the obstacle was replaced by a pinhole. There it was assumed that the potential field inside the pinhole (surface S_1 in Fig. B.2 (a)) is equal to the free-field potential, i.e., $\phi_0(\underline{r}_\mathrm{m})$. However, this assumption does not hold if the source is located at the same side as the obstacle

(or opening). This can be explained since in this case, in the inverse arrangement the surfaces S_2 and S_3 would block the line-of-sight between source and the opening. Hence, in our geometrical approximation of the head as two plates, Eq.(B.9) is valid only for $0 \leq \theta < \frac{\pi}{2}$, where for the special case $\theta \approx \frac{\pi}{2}$ the equation is also only an approximation. Therefore, Eq.(B.9) cannot be used to calculate the transfer function between source and M_2 for $0 \leq \theta < \frac{\pi}{2}$. Since in this case Q and M_2 are located at the same side of the plate, no diffraction occurs. Hence, the frequency response corresponds to the free-field condition for $0 \leq \theta < \frac{\pi}{2}$, obtained from Eqs.(B.3), (B.11) and the same normalization as in Eq.(B.9) by

$$H_2(\Omega,\theta) \quad = \quad \frac{e^{-j\frac{\Omega f_\mathrm{s}}{c} r_\mathrm{q}}}{r_\mathrm{q}/l_0} \cdot \; e^{j\frac{\Omega f_\mathrm{s}}{c}\frac{d_\mathrm{mic}}{2}\cos\theta}. \tag{B.16}$$

For the case $\frac{\pi}{2} < \theta \leq \pi$, the opposite effect occurs, i.e., M_2 is blocked by P_2 while a free-field condition can be assumed between source Q and M_1. Due to symmetry, the transfer function can be hence formulated as

$$\begin{aligned} H_1(\Omega,\theta) &= H_2(\Omega,\pi-\theta) \\ H_2(\Omega,\theta) &= H_1(\Omega,\pi-\theta) \end{aligned} \qquad \text{for} \quad \frac{\pi}{2} < \theta \leq \pi\,. \tag{B.17}$$

The coherence definition of Eq.(2.3) can, under the assumption of two omnidirectional microphones, generally be expressed as

$$\Gamma_{x_1x_2}(\Omega) = \frac{\int\limits_0^{\pi/2} \left(H_1(\Omega,\theta)\, H_2^*(\Omega,\theta) + H_2(\Omega,\theta)\, H_1^*(\Omega,\theta)\right) \sin\theta \,\mathrm{d}\theta}{\int\limits_0^{\pi/2} \left(|H_1(\Omega,\theta)|^2 + |H_2(\Omega,\theta)|^2\right) \sin\theta \,\mathrm{d}\theta}. \tag{B.18}$$

Since H_1 and H_2 are independent of φ_q, the double integral simplifies to the integration over θ.

With Eq.(B.17), (B.18) reads

$$= \quad \frac{2\int\limits_0^{\pi/2} \mathrm{Re}\left\{\mathrm{H}_1(\Omega,\theta)\, \mathrm{H}_2^*(\Omega,\theta)\right\} \sin\theta \,\mathrm{d}\theta}{\int\limits_0^{\pi/2} \left(|H_1(\Omega,\theta)|^2 + |H_2(\Omega,\theta)|^2\right) \sin\theta \,\mathrm{d}\theta}. \tag{B.19}$$

Finally, with Eqs.(B.19), (B.15) and (B.16), the coherence can be calculated. For a detailed derivation of Eq.(B.19), the reader is referred to, e.g., [Kut09]. It can be observed that the coherence is independent of scaling factor l_0 and distance r_q since the prefactor $\frac{1}{r_\mathrm{q}/l_0}\, e^{-j\frac{\Omega f_\mathrm{s}}{c}\, r_\mathrm{q}}$ eliminates. Since a closed-form solution of the integral in Eq.(B.15) cannot be obtained, the coherence given by Eq.(B.19) has to

be solved numerically. One solution is to calculate the integral by summation. The intervals Δr and $\Delta\tilde{\varphi}$ in Eq.(B.15) have to be chosen such that the corresponding surface elements $\Delta S = r \cdot \Delta\tilde{\varphi} \cdot \Delta r$ are small compared to the sound wavelength λ. Here, it is proposed that the maximum length of every surface element should be one-tenth of the wavelength which results for a given head radius r_{head} in

$$N_r = \frac{r_{\mathrm{head}}}{\Delta r} = \frac{r_{\mathrm{head}}}{\lambda/10} = 10\,\frac{r_{\mathrm{head}}\,f}{c} \tag{B.20}$$

intervals for the summation over r. Similarly, the summation over $\tilde{\varphi}$ in Eq.(B.19) requires

$$N_{\tilde{\varphi}} = \frac{2\pi}{\Delta\tilde{\varphi}} = \frac{2\pi\,r_{\mathrm{head}}}{\lambda/10} = 10\,\frac{2\pi\,r_{\mathrm{head}}\,f}{c} \tag{B.21}$$

intervals. The Integrals $\int_0^{\pi/2}\ldots\,\mathrm{d}\theta$ are calculated by summation over $N_\theta = 36$ intervals of length $\Delta\theta = \frac{\pi/2}{N_\theta}$.

In order to calculate the binaural coherence for a 2D noise field, where the noise sources are distributed in the same horizontal plane as the head, the $\sin\theta$-terms in Eq.(B.19) have to be disregarded.

C

Objective Quality Measures

- **SII**

  The *Speech Intelligibility Index* (SII) [ANS07] is a measure to determine the intelligibility of a speech communication system. An SII higher than 0.75 indicates a good communication system and values below 0.45 correspond to a poor system.

- **SRMR**

  The non-intrusive *Speech to Reverberation Modulation energy Ratio* (SRMR) [FC08, FZC10] was developed to evaluate speech dereverberation algorithms and has the great advantage that not reference signal such as the clean or anechoic signal needs to be available. The method is based on a modulation spectral representation which is obtained by means of a gammatone filterbank analysis of the temporal envelopes of the speech signal. Figure C.1 shows the effects of room reverberation on the SRMR measure, plotted over the DRR.

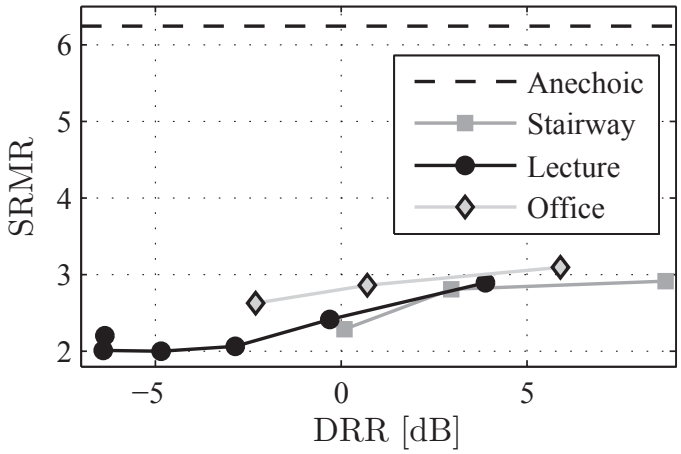


Figure C.1: Influence of reverberation on SRMR measure using a reverberant speech signal. The horizontal dashed line represents the SRMR value of the anechoic signal.

- **NA-SA**

  The segmental *Noise Attenuation* (NA) and *Speech Attenuation* (SA) is calculated independent for each frame with the same length and overlap as the speech enhancement algorithm itself. The overall values are determined by averaging over all considered frames according to

$$\mathrm{SA} = \frac{1}{\mathcal{C}(K_s)} \sum_{l \in K_s} \left(10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} s_d^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \widetilde{s}_d^2(k + l \cdot L_\mathrm{F})} \right) \right) \tag{C.1}$$

$$\mathrm{NA} = \frac{1}{\mathcal{C}(K_n)} \sum_{l \in K_n} \left(10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} n_\mathrm{int}^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \widetilde{n}_\mathrm{int}^2(k + l \cdot L_\mathrm{F})} \right) \right), \tag{C.2}$$

  where K_n denote the overall number of frames and K_s are the frames with speech activity, determined by a VAD. $\mathcal{C}(\cdot)$ are the number of elements and L_F denoted the block length. The difference between noise and speech attenuation (NA-SA) is a good indicator of the overall performance and values grater than 0 dB indicate an efficient speech enhancement while ensuring low speech distortions compared to the amount of attenuated interfering source, see also [Gus99].

- **segDRR**

  As for the NA-SA measure, the segDRR is calculated for each segment and the filtered direct speech as well as the filtered reverberant speech are used to calculate the DRR before and after processing. The difference can be regarded as the increase in DRR and hence, positive values indicate an improved speech quality.

- **PESQ**

  The *Perceptual Evaluation of Speech Quality* (PESQ) score [ITU01] was initially developed for the evaluation of speech codecs and is also widely used for the assessment of speech enhancement algorithms. Figure C.2 shows the effects of room reverberation on the PESQ score, plotted over the DRR using the direct speech signal as the required reference signal.

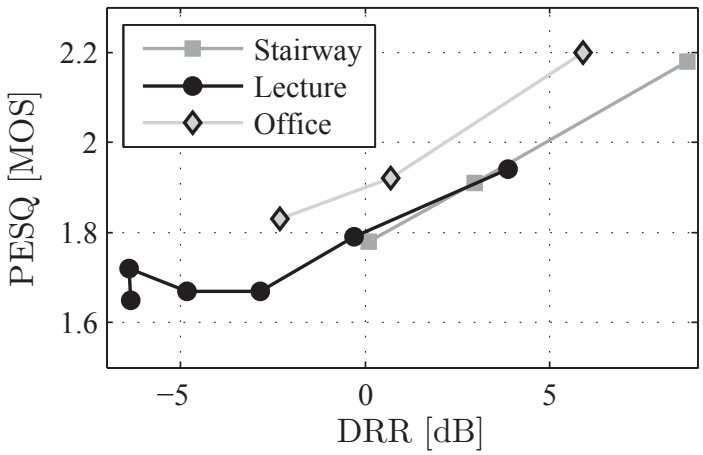


Figure C.2: Influence of reverberation on PESQ score using a reverberant speech signal.

Bibliography

[3GP04a] 3GPP TS 26.071. *Adaptive Multi-Rate (AMR) speech codec; General description.* 3GPP, 2004.

[3GP04b] 3GPP TS 26.171. *Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description.* 3GPP, 2004.

[3GP04c] 3GPP TS 26.194. *Adaptive Multi-Rate - Wideband speech codec, Voice Activity Detector.* V6.0.0, 2004.

[AB79] J. Allen and D. Berkley. "Image method for efficiently simulating small-room acoustics". *Journal of the Acoustical Society of America (JASA)*, vol. 65, no. 4, pp. 943–950, 1979.

[ABB77] J. Allen, D. Berkley, and J. Blauert. "Multimicrophone signal-processing technique to remove room reverberation from speech signals". *Journal of the Acoustical Society of America (JASA)*, vol. 62, no. 4, pp. 912–915, 1977.

[ACV86] W. Armbrüster, R. Czarnach, and P. Vary. "Adaptive noise cancellation with reference input - possible applications and theoretical limits". *Proc. European Signal Processing Conference (EUSIPCO)*, The Hague, The Netherlands, 1986.

[ANS07] ANSI S3.5-1997. *Methods for the Calculation of the Speech Intelligibility Index.* ANSI, r2007 edition, 2007.

[Ant08] C. Antweiler. "Multi-channel system identification with perfect sequences". R. Martin, U. Heute, and C. Antweiler, editors, *Advances in Digital Speech Transmission.* Wiley, 2008.

[BD98] C. Brown and R. Duda. "A structural model for binaural sound synthesis". *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, 1998.

[BH99] A. Bronkhorst and T. Houtgast. "Auditory distance perception in rooms". *Nature*, vol. 397, pp. 517–520, 1999.

[Bla96] J. Blauert. *Spatial Hearing - The Psychophysics of Human Sound Localization.* MIT Press, Cambridge, USA, rev. edition, 1996.

[Bol79] S. Boll. "Suppression of acoustic noise in speech using spectral subtraction". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[BSP03] J. Bradley, H. Sato, and M. Picard. "On the importance of early reflections for speech in rooms". *Journal of the Acoustical Society of America (JASA)*, vol. 113, no. 6, pp. 3233–3244, 2003.

[BW99] M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Cambridge University Press, 7 edition, 1999.

[BW08] M. Buck and A. Wolf. "Model-based dereverberation of single-channel speech signals". *Proc. German Annual Conference on Acoustics (DAGA)*, Dresden, Germany, 2008.

[CBH06] J. Chen, J. Benesty, and Y. Huang. "Time delay estimation in room acoustic environments: An overview". *EURASIP Journal on Applied Signal Processing*, vol. 2006, 2006.

[CG95] J.-H. Chen and A. Gersho. "Adaptive postfiltering for quality enhancement of coded speech". *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 59–71, 1995.

[CK11] M.-S. Choi and H.-G. Kang. "A two-channel noise estimator for speech enhancement in a highly nonstationary environment". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 905–915, 2011.

[CKN73] G. Carter, C. Knapp, and A. Nuttall. "Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing". *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 4, pp. 337–344, 1973.

[CMW11a] B. Cornelis, M. Moonen, and J. Wouters. "Binaural voice activity detection for MWF-based noise reduction in binaural hearing aids". *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011.

[CMW11b] B. Cornelis, M. Moonen, and J. Wouters. "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.

[Cor11] B. Cornelis. *Design and evaluation of noise reduction techniques for binaural hearing aids*. PhD thesis, K.U. Leuven, Leuven, Belgium, 2011.

[Cro80] R. Crochiere. "A weighted overlap-add method of short-time fourier analysis/synthesis". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[Dan68] L. Danilenko. *Binaurales Hören im nichtstationären diffusen Schallfeld*. PhD thesis, RWTH Aachen University, Aachen, Germany, 1968.

[DB08] P. Degry and C. Beaugeant. "Solution to speech intelligibility improvement in mobile phones". *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2008.

[DDP08] G. Defrance, L. Daudet, and J.-D. Polack. "Finding the onset of a room impulse response: Straightforward?". *Journal of the Acoustical Society of America (JASA) - Express Letters*, vol. 124, no. 4, Oct. 2008.

[DE96] M. Dörbecker and S. Ernst. "Combination of two channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation". *Proc. European Signal Processing Conference (EUSIPCO)*, Trieste, Italy, 1996.

[Dil01] H. Dillon. *Hearing Aids*. Thieme, Stuttgart, 2001.

[DM02] S. Doclo and M. Moonen. "GSVD-based optimal filtering for single and multimicrophone speech enhancement". *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[Dör98] M. Dörbecker. *Mehrkanalige Signalverarbeitung zur Verbesserung akustisch gestörter Sprachsignale am Beispiel elektronischer Hörhilfen*. PhD thesis, RWTH Aachen University, Aachen, Germany, 1998.

[DRZ97] J. Desloge, W. Rabinowitz, and P. M. Zurek. "Microphone-array hearing aids with binaural output. I. Fixed-processing systems". *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, 1997.

[DSWM05] S. Doclo, A. Spriet, J. Wouters, and M. Moonen. "Speech distortion weighted multi-channel wiener filtering techniques for noise reduction". J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, chapter 9. Springer, 2005.

[DSWM07] S. Doclo, A. Spriet, J. Wouters, and M. Moonen. "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction". *Speech Communication*, vol. 49, pp. 636–656, 2007.

[EH09] J. Erkelens and R. Heusdens. "Single-microphone late-reverberation suppression in noisy speech by exploiting long-term correlation in the DFT domain". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[EH10] J. Erkelens and R. Heusdens. "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1746–1765, 2010.

[EH11] J. Erkelens and R. Heusdens. "A statistical room impulse response model with frequency dependent reverberation time for single-microphone late reverberation suppression". *Proc. Conference of the Int. Speech Communication Association (INTERSPEECH)*, 2011.

[Elk01] G. Elko. "Spatial coherence functions for differential microphones in isotropic noise fields". M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 4. Springer, 2001.

[EM84] Y. Ephraim and D. Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[ETS09] ETSI 202 396-1. *Speech and multimedia Transmission Quality (STQ); Part 1: Background noise simulation technique and background noise database*, 03 2009. V1.2.3.

[EV09] T. Esch and P. Vary. "Efficient musical noise suppression for speech enhancement systems". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[FC08] T. Falk and W.-Y. Chan. "A non-intrusive quality measure of dereverberated speech". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, 2008.

[FK07] K. Furuya and A. Kataoka. "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1579–1591, 2007.

[FM04] C. Faller and J. Merimaa. "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence". *Journal of the Acoustical Society of America (JASA)*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[FZC10] T. Falk, C. Zheng, and W.-Y. Chan. "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[GB99] S. Griebel and M. Brandstein. "Wavelet transform extrema clustering for multi-channel speech dereverberation". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Pocono Manor, USA, 1999.

[GB01] S. Griebel and M. Brandstein. "Microphone array speech dereverberation using coarse channel modeling". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, 2001.

[GCNL00] H. Gustaffson, I. Claesson, S. Nordholm, and U. Lindgren. "Dual microphone spectral subtraction". Technical report, Department of Telecommunications and Signal Processing, University of Karlskrona/Ronneby, Sweden, 2000.

[Ger10] T. Gerkmann. *Statistical Analysis of Cepstral Coefficients and Applications in Speech Enhancement*. PhD thesis, Ruhr-Universität Bochum, Bochum, Germany, 2010.

[GH11] T. Gerkmann and R. Hendriks. "Noise power estimation based on the probability of speech presence". *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2011.

[GHN08] N. Gaubitch, E. Habets, and P. Naylor. "Multimicrophone speech dereverberation using spatiotemporal and spectral processing". *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, Seattle, USA, 2008.

[GJV98] S. Gustafsson, P. Jax, and P. Vary. "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, 1998.

[GKL+09] B. Geiser, H. Krüger, H. Löllmann, P. Vary, D. Zhang, H. Wan, H. Li, and L. Zhang. "Candidate proposal for ITU-T super-wideband speech and audio coding". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[GLBF03] A. Guerin, R. Le Bouquin, and G. Faucon. "A two-sensor noise reduction system: Applications for hands-free car kit". *EURASIP Journal on Applied Signal Processing*, , no. 11, pp. 1125–1134, 2003.

[GLJ+12] N. Gaubitch, H. Löllmann, M. Jeub, T. Falk, P. Naylor, P. Vary, and M. Brookes. "Performance comparison of algorithms for blind reverberation time estimation from speech". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, 2012.

[GM90] B. Glasberg and B. Moore. "Derivation of auditory filter shapes from notched-noise data.". *Hearing Research*, vol. 47, pp. 103–138, 1990.

[GMF01] B. Gillespie, H. Malvar, and D. Florencio. "Speech dereverberation via maximum-kurtosis subband adaptive filtering". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, 2001.

[GMJV02] S. Gustafsson, R. Martin, P. Jax, and P. Vary. "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction". *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[GNW04] N. Gaubitch, P. Naylor, and D. Ward. "Multi-microphone speech dereverberation using spatio-temporal averaging". *Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 2004.

[Gus99] S. Gustafsson. *Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction.* PhD thesis, RWTH Aachen University, Aachen, Germany, 1999.

[Hab07] E. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement.* Phd thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, June 2007.

[Hab10] E. Habets. "Speech dereverberation using statistical reverberation models". P. Naylor and N. Gaubitch, editors, *Speech Dereverberation*, chapter 3. Springer, 2010.

[Ham02] V. Hamacher. "Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, USA, 2002.

[Har99] W. M. Hartmann. "How we localize sound". *Physics Today*, pp. 24–29, Nov. 1999.

[HCE+05] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. "Signal processing in high-end hearing aids: State of the art, challenges, and future trends". *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 2915–2929, 2005.

[HCG08] E. Habets, I. Cohen, and S. Gannot. "Generating nonstationary multisensor signals under a spatial coherence constraint". *Journal of the Acoustical Society of America (JASA)*, vol. 124, no. 5, pp. 2911–2917, 2008.

[Her11] C. Herglotz. "Dual-channel reduction of low-frequency noise". Diploma thesis, RWTH Aachen University, Germany, 2011.

[HG12] R. Hendriks and T. Gerkmann. "Noise correlation matrix estimation for multi-microphone speech enhancement". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 223–233, 2012.

[HGC09] E. Habets, S. Gannot, and I. Cohen. "Late reverberant spectral variance estimation based on a statistical model". *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[HHJ10] R. Hendriks, R. Heusdens, and J. Jensen. "MMSE based noise PSD tracking with low complexity". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010.

[HJN+11] C. Herglotz, M. Jeub, C. Nelke, C. Beaugeant, and P. Vary. "Evaluation of single- and dual-channel noise power spectral density estimation algorithms for mobile phones". *Proc. German Conference on Speech Signal Processing (ESSV)*, Aachen, Germany, 2011.

[HK06] R. Huber and B. Kollmeier. "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[HKLP08] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder. "Binaural signal processing in hearing aids: Technologies and algorithms". R. Martin, U. Heute,

and C. Antweiler, editors, *Advances in Digital Speech Transmission*. Wiley, 2008.

[HNS+10a] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda. "Estimating direct-to-reverberant energy ratio based on spatial correlation model segregating direct sound and reverberation". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010.

[HNS+10b] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda. "Evaluating estimation of direct-to-reverberation energy ratio using D/R spatial correlation matrix model". *Proc. Int. Congress on Acoustics (ICA)*, Sydney, Australia, 2010.

[HNS+11] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda. "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, 2011.

[HYN07] T. Hidaka, Y. Yamada, and T. Nakagawa. "A new definition of boundary point between early reflections and late reverberation in room impulse responses". *Journal of the Acoustical Society of America (JASA)*, vol. 122, pp. 326–332, 2007.

[IEC03] IEC 60268-16. *Sound system equipment Part 16: Objective rating of speech intelligibility by speech transmission index*. IEC, 2003.

[ISO93] ISO/IEC 11172-3:1993. *Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s - Part 3: Audio*. ISO/IEC, 1993.

[ITU88] ITU-T Rec. G.722. *7 kHz audio-coding within 64 kbit/s*. ITU, Geneva, Switzerland, 1988.

[ITU93] ITU-T Rec. P.56. *Objective measurement of active speech level*. ITU, Geneva, Switzerland, 1993.

[ITU96] ITU-T Rec. P.58. *Head and Torso Simulator for Telephonometry*. ITU, Geneva, Switzerland, 1996.

[ITU00] ITU-T Rec. P.340. *Transmission characteristics and speech quality parameters of hands-free terminals*. ITU, Geneva, Switzerland, 2000.

[ITU01] ITU-T Rec. P.862. *Perceptual evaluation of speech quality (PESQ)*. ITU, Geneva, 2001.

[ITU03] ITU-R Rec. BS.1284-1. *General methods for the subjective assessment of sound quality*. ITU, Geneva, Switzerland, 2003.

[ITU07] ITU-T Rec. P.64. *Determination of sensitivity/frequency characteristics of local telephone systems*. ITU, Geneva, Switzerland, 2007.

[ITU09] ITU-T Rec. G.191. *Software tools for speech and audio coding standardization*. ITU, Geneva, Switzerland, March 2009.

[JDV11] M. Jeub, M. Dörbecker, and P. Vary. "A semi-analytical model for the binaural coherence of noise fields". *IEEE Signal Processing Letters*, vol. 18, no. 3, March 2011.

[Jef48] L. Jeffress. "A place theory of sound localization". *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, Feb. 1948.

[JHN+12] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary. "Noise reduction for dual-microphone mobile phones exploiting power level differences". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[JKAO09] M. Jeub, D. Kolossa, R. Astudillo, and R. Orglmeister. "Performance analysis of wavelet-based voice activity detection". *Proc. German Annual Conference on Acoustics (DAGA)*, Rotterdam, The Netherlands, 2009.

[JLV10] M. Jeub, H. Löllmann, and P. Vary. "Blind dereverberation for hearing aids with binaural link". *Proc. ITG Conference on Speech Communication*, Bochum, Germany, 2010.

[JNBV11] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary. "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals". *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011.

[JNK+11] M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary. "Robust dual-channel noise power spectral density estimation". *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011.

[JR00] F. Jacobsen and T. Roisin. "the coherence of reverberant sound fields". *Journal of the Acoustical Society of America (JASA)*, vol. 108, no. 1, July 2000.

[JSEV10] M. Jeub, M. Schäfer, T. Esch, and P. Vary. "Model-based dereverberation preserving binaural cues". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[JSK+10] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary. "Do we need dereverberation for hand-held telephony?". *Proc. Int. Congress on Acoustics (ICA)*, Sydney, Australia, 2010.

[JSV09] M. Jeub, M. Schäfer, and P. Vary. "A binaural room impulse response database for the evaluation of dereverberation algorithms". *Proc. Int. Conference on Digital Signal Processing (DSP)*, Santorini, Greece, 2009.

[JV09a] M. Jeub and P. Vary. "Dereverberation of speech signals based on the discrete model of speech production". *Proc. German Conference on Speech Signal Processing (ESSV)*, Dresden, Germany, 2009.

[JV09b] M. Jeub and P. Vary. "Enhancement of reverberant speech using the CELP postfilter". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[JV10] M. Jeub and P. Vary. "Binaural dereverberation based on a dual-channel wiener filter with optimized noise field coherence". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010.

[JV11] M. Jeub and P. Vary. "On the application of psychoacoustically-motivated dereverberation for recordings taken in the German parliament". *Proc. German Conference on Speech Signal Processing (ESSV)*, Aachen, Germany, 2011.

[Kab02] P. Kabal. "TSP speech database". Technical report, Department of Electrical & Computer Engineering, McGill University, Montreal, Quebec, Canada, 2002.

[Kat08] J. Kates. *Digital Hearing Aids*. Plural Publishing, Berlin, 2008.

[KC76] C. Knapp and G. Carter. "The generalized correlation method for estimation of time delay". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[KEA+09] H. Kayser, S. Ewert, J. Anemller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses". *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[Kit10] A. Kitzig. "Niederrhein university room impulse response package (NRU-RIR)". Technical report, Niederrhein University, Krefeld, Germany, June 2010.

[KPB09] A. Kamkar-Parsi and M. Bouchard. "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 521–533, 2009.

[Kut09] H. Kuttruff. *Room Acoustics*. Spon Press, Oxon, 2009.

[LB86] I. Lindevald and A. Benade. "Two-ear correlation in the statistical sound fields of rooms". *Journal of the Acoustical Society of America (JASA)*, vol. 80, no. 2, pp. 661–664, 1986.

[LBD01] K. Lebart, J. Boucher, and P. Denbigh. "A new method based on spectral subtraction for speech dereverberation". *Acta Acustica United with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[LC08] Y.-C. Lu and M. Cooke. "Binaural distance perception based on direct-to-reverberant energy ratio". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, 2008.

[LC10] Y.-C. Lu and M. Cooke. "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[Leb99] K. Lebart. *Speech Dereverberation applied to Automatic Speech Recognition and Hearing Aids*. PhD thesis, L'universite de Rennes, Rennes, France, 1999.

[Loi07] P. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[LV06] T. Lotter and P. Vary. "Dual-channel speech enhancement by superdirective beamforming". *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–14, 2006.

[LV08a] H. Löllmann and P. Vary. "Estimation of the reverberation time in noisy environments". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, 2008.

[LV08b] H. Löllmann and P. Vary. "Low delay filter-banks for speech and audio processing". E. Hänsler and G. Schmidt, editors, *Speech and Audio Processing in Adverse Environments*, chapter 2, pp. 13–61. Springer, Berlin, August 2008.

[LV09a] L. Laaksonen and J. Virolainen. "Binaural artificial bandwidth extension (B-ABE) for speech". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[LV09b] H. Löllmann and P. Vary. "Low delay noise reduction and dereverberation for hearing aids". *EURASIP Journal on Applied Signal Processing*, vol. 1, 2009.

[LV11] H. Löllmann and P. Vary. "Estimation of the frequency dependent reverberation time by means of warped filter-banks". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[LYJV10] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary. "An improved algorithm for blind reverberation time estimation". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, 2010.

[Mar95] R. Martin. *Freisprecheinrichtung mit mehrkanaliger Echokompensation und Störgeräuschreduktion*. PhD thesis, RWTH Aachen University, Aachen, Germany, 1995.

[Mar01a] R. Martin. "Noise power spectral density estimation based on optimal smoothing and minimum statistics". *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[Mar01b] R. Martin. "Small microphone arrays with postfilter". M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 12. Springer, 2001.

[MB03] I. McCowan and H. Bourlard. "Microphone array post-filter based on noise field coherence". *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[MHA11] J. Marin-Hurtado and D. Anderson. "Robust non-VAD implementation of multichannel wiener filter for binaural noise reduction". *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 341–344, 2011.

[MHD11] D. Marquardt, V. Hohmann, and S. Doclo. "Performance comparison of binaural beamforming and mwf-based noise reduction algorithms for hearing aids". *Proc. German Annual Conference on Acoustics (DAGA)*, Düsseldorf, Germany, 2011.

[MM01] S. Müller and P. Massarani. "Transfer-function measurement with sweeps". *Journal of the Audio Engineering Society (JAES)*, vol. 49, no. 6, pp. 443–471, 2001.

[MSK97] J. Meyer, K. Simmer, and K. Kammeyer. "Comparison of one- and two-channel noise-estimation techniques". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, London, UK, 1997.

[NC94] NTT-Corporation. "Multi-lingual speech database for telephonometry", 1994.

[NG05] P. Naylor and N. Gaubitch. "Speech dereverberation". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005.

[NG10] P. Naylor and N. Gaubitch, editors. *Speech Dereverberation*. Springer, London, 2010.

[NLT89] A. Nábělek, T. Letowski, and F. Tucker. "Reverberant overlap- and self-masking in consonant identification". *Journal of the Acoustical Society of America (JASA)*, vol. 86, no. 4, pp. 1259–1265, 1989.

[PAG95] R. Patterson, M. Allerhand, and C. Gigure. "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform". *Journal of the Acoustical Society of America (JASA)*, vol. 98, no. 4, pp. 1890–1894, Oct. 1995.

[PdLN11] T. Prego, A. de Lima, and S. Netto. "Perceptual improvement of a two-stage algorithm for speech dereverberation". *Proc. Conference of the Int. Speech Communication Association (INTERSPEECH)*, 2011.

[Pei92] J. Peissig. *Binaurale Hörgerätestrategien in komplexen Störschallsituationen*. PhD thesis, Universität Göttingen, Göttingen, Germany, 1992.

[Pie78] A. Piersol. "Use of coherence and phase data between two receivers in evaluation of noise environments". *Journal of Sound and Vibration*, vol. 56, no. 2, pp. 215–228, 1978.

[Pol88] J.-D. Polack. *La transmission de l'energie sonore dans les salles*. PhD thesis, Universite du Maine, Le Mans, France, 1988.

[RJW+03] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien, C. Lansing, and A. Feng. "Blind estimation of reverberation time". *Journal of the Acoustical Society of America (JASA)*, vol. 114, no. 5, pp. 2877–2892, 2003.

[RPF+10] K. Reindl, P. Prokein, E. Fischer, Y. Zheng, and W. Kellermann. "Combining monaural beamforming and blind source separation for binaural speech enhancement in multi-microphone hearing aids". *Proc. ITG Conference on Speech Communication*, Bochum, Germany, 2010.

[RV06] O. Roy and M. Vetterli. "Rate-constrained beamforming for collaborating hearing aids". *Int. Symposium on Information Theory (ISIT)*, Seattle, USA, 2006.

[RWS09] A. Raake, M. Wältermann, and S. Spors. "Which wideband speech codec? Quality impact due to room-acoustics at send side and presentation method". *Proc. Audio Engineering Society (AES) Conference*, New York, USA, 2009.

[RZK10] K. Reindl, Y. Zheng, and W. Kellermann. "Speech enhancement for binaural hearing aids based on blind source separation". *Proc. 4th Int. Symp.on Communications, Control, and Signal Proc. (ISCCSP)*, Limassol, Cyprus, 2010.

[Sab21] W. Sabine. *Collected Papers on Acoustics*. Peninsula Pub, 1993 (Originally 1921).

[Sch62] M. Schroeder. "Frequency-correlation functions of frequency responses in rooms". *Journal of the Acoustical Society of America (JASA)*, vol. 34, no. 12, pp. 1819–1823, December 1962.

[Sch65] M. Schroeder. "New method of measuring reverberation time". *Journal of the Acoustical Society of America (JASA)*, vol. 37, no. 3, pp. 409–412, 1965.

[SdB09] S. Srinivasan and A. den Brinker. "Rate-constrained beamforming in binaural hearing aids". *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[See11] B. Seeber. "Binaural hearing in reverberant space with cochlear implants". *Conf. on Implantable Auditory Prostheses*, p. 23, Pacific Grove, CA, USA, 2011.

[SJSV10] M. Schäfer, M. Jeub, B. Sauert, and P. Vary. "Reverberation-based post-processing for improving speech intelligibility". *Proc. Int. Congress on Acoustics (ICA)*, Sydney, Australia, 2010.

[Sla98] M. Slaney. "Auditory toolbox". Technical report, Interval Research Corporation, Palo Alto, USA, 1998.

[SM02] M. Stone and B. Moore. "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production". *Ear and Hearing*, vol. 23, no. 4, pp. 325–338, 2002.

[SS07] R. Stewart and M. Sandler. "Statistical measures of early reflections of room impulse responses". *Proc. of Int. Conference on Digital Audio Effects (DAFx)*, Bordeaux, France, 2007.

[TAV10] A. Telle, C. Antweiler, and P. Vary. "Der perfekte Sweep - Ein neues Anregungssignal zur adaptiven Systemidentifikation zeitvarianter akustischer Systeme". *Proc. German Annual Conference on Acoustics (DAGA)*, Berlin, Germany, 2010.

[TGGN07] M. Thomas, N. Gaubitch, J. Gudnason, and P. Naylor. "A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging". *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2007.

[TGM11] A. Tsilfidis, E. Georganti, and J. Mourjopoulos. "Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[Tsi11] A. Tsilfidis. *Signal Processing Methods for Enhancing Speech and Music Signals in Reverberant Environments.* PhD thesis, University of Patras, Patras, Greece, 2011.

[TTM+11] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin. "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[Ush10] J. Usher. "An improved method to determine the onset timings of reflections in an acoustic impulse response". *Journal of the Acoustical Society of America (JASA) - Express Letters*, vol. 127, no. 4, 2010.

[Van94] J. Vanderkooy. "Aspects of MLS measuring systems". *Journal of the Audio Engineering Society (JAES)*, vol. 42, no. 4, pp. 219–231, 1994.

[vdB08] T. van den Bogeart. *Preserving binaural cues in noise reduction algorithms for hearing aids.* PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2008.

[vdBKM+05] T. van den Bogeart, T. Klasen, M. Moonen, L. van Deun, and J. Wouters. "Horizontal localization with bilateral hearing aids: Without is better than with". *Journal of the Acoustical Society of America (JASA)*, vol. 119, no. 1, pp. 515–526, 2005.

[VHH98] P. Vary, U. Heute, and W. Hess. *Digitale Sprachsignalverarbeitung.* Teubner, Stuttgart, Germany, 1998.

[VM06] P. Vary and R. Martin. *Digital Speech Transmission. Enhancement, Coding and Error Concealment.* Wiley&Sons, Chichester, 2006.

[VS93] A. Varga and H. 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, no. 3, pp. 247–251, 1993.

[Wel67] P. Welch. "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms". *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[Wen09] J. Wen. *Reverberation: Models, Estimation and Applications*. PhD thesis, Imperial College London, London, UK, 2009.

[WGDZ97] D. Welker, J. Greenberg, J. Desloge, and P. Zurek. "Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system". *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 543–551, 1997.

[WGH+06] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor. "Evaluation of speech dereverberation algorithms using the MARDY database". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, 2006.

[WH03] T. Wittkopp and V. Hohmann. "Strategy-selective noise reduction for binaural digital hearing aids". *Speech Communication*, vol. 39, pp. 111–138, 2003.

[WHN08] J. Wen, E. Habets, and P. Naylor. "Blind estimation of reverberation time based on the distribution of signal decay rates". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, 2008.

[Wit01] T. Wittkopp. *Two-channel noise reduction algorithms motivated by models of binaural interaction*. PhD thesis, Universität Oldenburg, Oldenburg, Germany, 2001.

[WPK08] S. Wehr, H. Puder, and W. Kellermann. "Blind source separation and binaural reproduction with hearing aids: An overview". *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2008.

[WRDK11] A. Warzybok, J. Rennies, S. Doclo, and B. Kollmeier. "Influence of early reflections on speech intelligibility under different noise conditions". *Forum Acusticum*, Aalborg, Denmark, 2011.

[WW05] M. Wu and D. Wang. "A two-stage algorithm for enhancement of reverberant speech". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA, 2005.

[WW06] M. Wu and D. Wang. "A two-stage algorithm for one-microphone reverberant speech enhancement". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[WZ09] B. Wright and Y. Zhang. "A review of the generalization of auditory learning". *Phil. Trans. R. Soc. B*, vol. 364, pp. 301–311, 2009.

[YAR09] N. Yousefian, A. Akbari, and M. Rahmani. "Using power level difference for near field dual-microphone speech enhancement". *Applied Acoustics*, vol. 70, pp. 1412 – 1421, 2009.

[YM00] B. Yegnanarayana and P. Murthy. "Enhancement of reverberant speech using LP residual signal". *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[Zel88] R. Zelinski. "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms". *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New York, USA, 1988.