SPEECH DEREVERBERATION IN NOISY ENVIRONMENTS USING TIME-FREQUENCY DOMAIN SIGNAL MODELS

Enthallung von Sprachsignalen unter Einfluss von Störgeräuschen mittels Signalmodellen im Zeit-Frequenz Bereich

Der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg
zur
Erlangung des Doktorgrades

Doktor-Ingenieur (Dr.-Ing.)
vorgelegt von
Sebastian Braun
aus Eichstätt
Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung: 28.02.2018

Vorsitzender des Promotionsorgans: Prof. Dr.-Ing. Reinhard Lerch
Gutachter: Prof. Dr. ir. Emanuël A. P. Habets
Dr. Patrick A. Naylor
Acknowledgments

This doctoral thesis is part of the work that I conducted as a researcher in the spatial audio group of Prof. Habets at the International Audio Laboratories Erlangen, a joint institution of the Friedrich-Alexander Universität Erlangen Nürnberg (FAU) and the Fraunhofer institute for integrated circuits (IIS).

I would like to express my highest gratitude to my supervisor and mentor Prof. Emanuël Habets for offering me the chance to pursue my PhD, all the knowledge and motivation he gave me, and our fruitful discussions. From him I learned a lot about research, scientific writing, but also engineering. A special thanks goes also to Dr. Patrick Naylor for his deep interest in my work, being the second assessor and his helpful comments on the thesis.

I am thankful to Dr. Fabian Küch, Manfred Lutzky and Dr. Frederik Nagel from Fraunhofer IIS, to enable the funding of my research position. I am grateful and happy that during and also after writing this thesis, I could contribute to exciting Fraunhofer IIS projects, especially the UROC project managed by Dr. Anthony Lombard, where I also learned a lot as an engineer. Further thanks and greetings go to Dr. Cornelia Falch and Dr. Markus Kallinger, who had a large influence that I took the path to pursue a PhD.

I thank everyone I collaborated with during my PhD for their efforts, time, discussions and countless emails: Prof. Sharon Gannot, Ofer Schwartz, Boaz Schwarz, Prof. Simon Doclo, Dr. Daniel Marquardt, Prof. Jesper Jensen, Adam Kuklasinski, Dr. Ramón Fernandez Astudillo, Dr. Oliver Thiergart, Dr. Konrad Kowalczyk, Daniel Jarrett, Johannes Fischer, Matteo Torcoli, João F. Santos and Wei Zhou. A special thanks to Prof. Tiago Falk for hosting me for a short-term research visit at INESC in Montréal.

Furthermore, I would like to thank all colleagues at AudioLabs and Fraunhofer IIS for the nice atmosphere, good chats, coffee and lunch breaks, and social events, creating one of the best places to work. Special thanks to have the best office colleagues Maja and Maria, and all members of the spatial audio group, Soumitro, Oli, Moritz, Youssef, Axel, Olli and Konrad. I also want to thank Nils Werner, Stefan Balke and Fabian Stöter for developing and sharing their awesome web tools, that I used extensively, and the AudioLabs administrative team Stefan Turowski, Elke Weiland, Day-See Riechmann and Tracy Harris for taking care and helping with countless things.

Finally, I am very grateful to my family and friends for their support during my PhD.
Abstract

Reverberation is the sum of reflected sound waves and is present in any conventional room. Speech communication devices such as mobile phones in hands-free mode, tablets, smart TVs, teleconferencing systems, hearing aids, voice-controlled systems, etc. use one or more microphones to pick up the desired speech signals. When the microphones are not in the proximity of the desired source, strong reverberation and noise can degrade the signal quality at the microphones and can impair the intelligibility and the performance of automatic speech recognizers. Therefore, it is a highly demanded task to process the microphone signals such that reverberation and noise are reduced. The process of reducing or removing reverberation from recorded signals is called dereverberation.

As dereverberation is usually a completely blind problem, where the only available information are the microphone signals, and as the acoustic scenario can be non-stationary, dereverberation is one of the most challenging tasks in speech enhancement. While in theory perfect dereverberation can be achieved by inverse filtering under some conditions and with knowledge of the room impulse response (RIR), in practice the blind identification of the RIR is not sufficiently accurate and robust in time-varying and noisy acoustic conditions. Therefore, successful dereverberation methods have been developed in the time-frequency domain that often relax the problem to partial dereverberation, where mainly the late reverberation tail is reduced. Although in the recent years some robust and efficient methods have been proposed that can reduce the late reverberation tail to some extent, it is still challenging to obtain a dereverberated signal with high audio quality, without speech distortion and artifacts using real-time processing techniques with minimal delay. In this thesis, we focus on robust dereverberation methods for online processing as required in real-time speech communication systems.

To achieve dereverberation, two main aspects can be exploited: temporal and spatial information. Firstly, reverberation introduces correlation over time and extends the duration of phonemes or sound events. By exploiting temporal correlation, filters can be derived to extract the desired speech signal or to reduce the reverberation. Secondly, by using multiple microphones, spatial information can be exploited to distinguish between the coherent direct sound and the reverberation, which has a spatially diffuse property. To extract the coherent sound, spatial filters, also known as beamformers, can be used that combine the microphone signals such that only sound from a certain direction is extracted, whereas sound from other directions and diffuse sound components are suppressed. In this thesis, a variety of signal models is exploited to model reverberation using temporal and spatial aspects. All considered signal models are defined in the short-time Fourier transform (STFT) domain, which is widely used in
many speech and audio processing techniques, therefore allowing simple integration with other existing techniques. In particular, we utilize a narrowband moving average model, a narrowband multichannel autoregressive model, and a spatial coherence based model. For each of these three signal models, a method for dereverberation and noise reduction is proposed.

The first main contribution is a single-channel estimator of the late reverberation power spectral density (PSD), which is required to compute a Wiener filter reducing reverberation and noise. The proposed reverberation PSD estimator is based on a narrowband moving average model using relative convolutive transfer functions (RCTFs). In contrast to other single-channel reverberation PSD estimators, the proposed estimator explicitly models time-varying acoustic conditions and additive noise, and requires no prior information on the room acoustics like the reverberation time or the direct-to-reverberation ratio (DRR).

The second main contribution is a multichannel reverberation PSD estimator based on the spatial coherence, where the reverberation is modeled as an additive diffuse sound component with a time-invariant spatial coherence. In the multichannel case, the desired signal can be estimated by a multichannel Wiener filter (MWF) that requires the reverberation PSD. To mitigate speech distortion and artifacts, a generalized method to control the attenuation of reverberation and noise at the output of a MWF independently is proposed. As there exists a wide variety of such single- and multichannel reverberation PSD estimators, an extensive overview, comparison and benchmark of state-of-the-art estimators is provided. As a cure for a common weakness of all reverberation PSD estimators, a bias compensation for high DRRs is proposed.

The third main contribution is an online solution for dereverberation and noise reduction based on a narrowband multichannel autoregressive (MAR) signal model for time-varying acoustic environments. Using this model, the late reverberation is predicted from previous reverberant speech samples using the MAR coefficients, and is then subtracted from the current reverberant signal. A main novelty of this approach is a parallel estimation structure, that allows to obtain causal estimates of time-varying MAR coefficients in noisy environments. In addition, a method to control the amount of reverberation and noise reduction independently is proposed.

In the last part of this thesis, the three proposed dereverberation systems are compared using objective measures, a listening test, and an automatic speech recognition system. It is shown that the proposed algorithms efficiently reduce reverberation and noise, and can be directly applied in speech communication devices. The theoretical overview and the evaluation shows that each dereverberation method has different strengths and limitations. By considering these algorithms as representatives of their dereverberation class, useful insights and conclusions are provided that can help for the choice of a dereverberation method for a specific application.
Kurzfassung

Akustischer Hall ist die Summe von reflektierten Schallwellen und ist jedem üblichen Raum vorhanden. Geräte zur Sprachkommunikation, z. B. Mobiltelefone im Freisprechmodus, Tablets, smart TVs, Telekonferenzsysteme, Hörgeräte, sprachgesteuerte Systeme etc. benutzen ein oder mehrere Mikrofone, um die gewünschten Sprachsignale aufzunehmen. Wenn sich die Mikrofone nicht in Nähe der gewünschten Schallquelle befinden, können starker akustischer Hall und andere Störgeräusche die Qualität der Mikrofonsignale verringern und die Sprachverständlichkeit sowie die Erkennungsraten von automatischen Spracherkennern stark verschlechtern. Daher ist es eine sehr aktuelle und gefragte Aufgabe, akustischen Hall und Störgeräusche in den Mikrofonsignalen zu reduzieren. Der Prozess, akustischen Nachhall in aufgenommenen Mikrofonsignalen zu reduzieren oder entfernen, wird als Enthallung bezeichnet.


Zur Enthallung können zwei Hauptaspekte ausgenutzt werden: zeitliche und räumliche Information. Erstens erzeugt Hall zeitliche Korrelation und verlängert die Dauer von Phonemen oder akustischen Ereignissen. Diese zeitliche Korrelation kann ausgenutzt werden, um Filter herzuleiten, die das gewünschte Sprachsignal extrahieren oder den Hall reduzieren können. Zweitens kann mit mehreren Mikrofonen räumliche Information ausgenutzt werden, um zwischen dem kohärenten Direktschall und dem räumlich diffusen Nachhall zu unterscheiden. Um die kohärente Schallkomponente zu extrahieren, werden Beamformer oder räumliche Filter benutzt, welche die Mikrofonsignale so filtern und kombinieren, dass nur der Schall aus einer bestimmten Richtung extrahiert wird und der Schall aus anderen Richtungen und diffuser

Der erste Hauptbeitrag ist ein einkanaliger Schätzer der spektralen Leistungsichte, eng. power spectral density (PSD), der Signalkomponente des späten Nachhalls, welche benötigt wird um einen spektralen Wiener filter zur Reduktion von Hall und Störgeräuschen zu berechnen. Der vorgeschlagene PSD Schätzer basiert auf einem schmalbandigen endlichen Filtermodell, das relative konvolutive Transferfunktionen verwendet. Im Gegensatz zu anderen einkanaligen PSD Schätzern der Hallkomponente modelliert der entwickelte Schätzer explizit zeit-variante akustische Bedingungen sowie additive Störgeräusche, und benötigt keine Information über die Raumakustik, wie z. B. die Nachhallzeit oder das Direkt-zu-Nachhall Energieverhältnis, eng. direct-to-reverberation ratio (DRR).


Im letzten Teil dieser Dissertation werden die drei entwickelten Enthallungssysteme mittels objektiven Maßen, einem Hörversuch und einem automatischen Spracherkennungssystem verglichen. Es wird gezeigt, dass die entwickelten Algorithmen den akustischen Nachhall und Störgeräusche effektiv reduzieren können und direkt in Geräten für Sprachkommunikation implementiert werden können. Die theoretische und experimentelle Evaluierung zeigt die unterschiedlichen
# Acronyms, Symbols, and Notation

## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGC</td>
<td>automatic gain control</td>
</tr>
<tr>
<td>AR</td>
<td>autoregressive</td>
</tr>
<tr>
<td>ASR</td>
<td>automatic speech recognition</td>
</tr>
<tr>
<td>CD</td>
<td>cepstral distance</td>
</tr>
<tr>
<td>CDR</td>
<td>coherent-to-diffuse ratio</td>
</tr>
<tr>
<td>CTF</td>
<td>convolutive transfer function</td>
</tr>
<tr>
<td>DDR</td>
<td>direct-to-diffuse ratio</td>
</tr>
<tr>
<td>DFT</td>
<td>discrete Fourier transform</td>
</tr>
<tr>
<td>DNR</td>
<td>diffuse-to-noise ratio</td>
</tr>
<tr>
<td>DOA</td>
<td>direction-of-arrival</td>
</tr>
<tr>
<td>DRR</td>
<td>direct-to-reverberation ratio</td>
</tr>
<tr>
<td>EDC</td>
<td>energy decay curve</td>
</tr>
<tr>
<td>EM</td>
<td>expectation-maximization</td>
</tr>
<tr>
<td>ESPRIT</td>
<td>estimation of signal parameters via rotational invariance techniques</td>
</tr>
<tr>
<td>FIR</td>
<td>finite impulse response</td>
</tr>
<tr>
<td>fwSSIR</td>
<td>frequency-weighted segmental signal-to-interference ratio</td>
</tr>
<tr>
<td>GUI</td>
<td>graphical user interface</td>
</tr>
<tr>
<td>IIR</td>
<td>infinite impulse response</td>
</tr>
<tr>
<td>IR</td>
<td>interference reduction</td>
</tr>
<tr>
<td>iSNR</td>
<td>input signal-to-noise ratio</td>
</tr>
<tr>
<td>LCMV</td>
<td>linearly constrained minimum variance</td>
</tr>
<tr>
<td>LLR</td>
<td>log-likelihood ratio</td>
</tr>
<tr>
<td>LPC</td>
<td>linear predictive coding</td>
</tr>
<tr>
<td>MA</td>
<td>moving average</td>
</tr>
<tr>
<td>MAP</td>
<td>maximum a posteriori</td>
</tr>
<tr>
<td>MAR</td>
<td>multichannel autoregressive</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-frequency cepstral coefficient</td>
</tr>
<tr>
<td>MFCCD</td>
<td>Mel-frequency cepstral coefficient distance</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>MIMO</td>
<td>multiple-input-multiple-output</td>
</tr>
<tr>
<td>MINT</td>
<td>multiple-input/output inverse theorem</td>
</tr>
<tr>
<td>ML</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>MLE</td>
<td>maximum likelihood estimator</td>
</tr>
<tr>
<td>MMSE</td>
<td>minimum mean square error</td>
</tr>
<tr>
<td>MOS</td>
<td>mean opinion score</td>
</tr>
<tr>
<td>MSE</td>
<td>mean-squared error</td>
</tr>
<tr>
<td>MTF</td>
<td>multiplicative transfer function</td>
</tr>
<tr>
<td>MUSHRA</td>
<td>multi-stimulus test with hidden reference and anchor</td>
</tr>
<tr>
<td>MUSIC</td>
<td>multiple signal classification</td>
</tr>
<tr>
<td>MVDR</td>
<td>minimum variance distortionless response</td>
</tr>
<tr>
<td>MWF</td>
<td>multichannel Wiener filter</td>
</tr>
<tr>
<td>NR</td>
<td>noise reduction</td>
</tr>
<tr>
<td>PDF</td>
<td>probability density function</td>
</tr>
<tr>
<td>PESQ</td>
<td>perceptual evaluation of speech quality</td>
</tr>
<tr>
<td>PMWF</td>
<td>parametric multichannel Wiener filter</td>
</tr>
<tr>
<td>PSD</td>
<td>power spectral density</td>
</tr>
<tr>
<td>QAreverb</td>
<td>quality assessment of reverberated speech</td>
</tr>
<tr>
<td>RC</td>
<td>reduction control</td>
</tr>
<tr>
<td>RCTF</td>
<td>relative convolutive transfer function</td>
</tr>
<tr>
<td>RIR</td>
<td>room impulse response</td>
</tr>
<tr>
<td>RLS</td>
<td>recursive least-squares</td>
</tr>
<tr>
<td>RMS</td>
<td>root-mean square</td>
</tr>
<tr>
<td>RR</td>
<td>reverberation reduction</td>
</tr>
<tr>
<td>RTF</td>
<td>relative transfer function</td>
</tr>
<tr>
<td>SCD</td>
<td>speaker change detection</td>
</tr>
<tr>
<td>SDI</td>
<td>speech distortion index</td>
</tr>
<tr>
<td>SIR</td>
<td>signal-to-interference ratio</td>
</tr>
<tr>
<td>SPP</td>
<td>speech presence probability</td>
</tr>
<tr>
<td>SRMR</td>
<td>signal-to-reverberation-modulation ratio</td>
</tr>
<tr>
<td>STFT</td>
<td>short-time Fourier transform</td>
</tr>
<tr>
<td>TLS-ESPRIT</td>
<td>total least-squares ESPRIT</td>
</tr>
<tr>
<td>UCA</td>
<td>uniform circular array</td>
</tr>
<tr>
<td>ULA</td>
<td>uniform linear array</td>
</tr>
<tr>
<td>WER</td>
<td>word error rate</td>
</tr>
<tr>
<td>WF</td>
<td>Wiener filter</td>
</tr>
</tbody>
</table>
Symbols

\( a, b, c \) | scalars
---|---
\( \mathbf{a}, \mathbf{b}, \mathbf{c} \) | column vectors
\( \mathbf{A}, \mathbf{B}, \mathbf{C} \) | matrices
\( \mathbf{I}_{A \times B} \) | (possibly truncated) identity matrix with \( A \) rows and \( B \) columns
\( \mathbf{1}_{A \times B} \) | matrix of ones with \( A \) rows and \( B \) columns
\( \mathbf{0}_{A \times B} \) | matrix of zeros with \( A \) rows and \( B \) columns
\( (\cdot)^* \) | complex conjugate
\( (\cdot)^T \) | transpose
\( (\cdot)^H \) | conjugate (Hermitian) transpose
\( \hat{\cdot} \) | estimated quantity
\( e \) | Euler’s number
\( j \) | imaginary unit
\( \mathcal{E}\{\cdot\} \) | expectation
\( \mathcal{N}(\mu, \Phi) \) | normal (Gaussian) distribution with mean vector \( \mu \) and covariance \( \Phi \)
\( \mathcal{R}\{\cdot\} \) | real part
\( \mathcal{I}\{\cdot\} \) | imaginary part
\( * \) | convolution operator
\( \delta \) | Kronecker-Delta function
\( \partial \) | differential operator
\( \otimes \) | Kronecker product
\( \text{diag}\{\cdot\} \) | main diagonal elements of a matrix returned as column vector
\( \text{tr}\{\cdot\} \) | trace
\( \text{vec}\{\cdot\} \) | stacks the columns of a matrix into a vector
\( \| \cdot \|_F \) | Frobenius norm

Notation

\( c \) | speed of sound
\( D \) | number of frames considered as early reflections
\( f_s \) | sampling frequency
\( i \) | plane wave index
\( I \) | number of plane waves
\( k \) | frequency index
\( K \) | number of frequencies, DFT length
\( L \) | filter length
\( m \) | microphone index
\( M \) | number of microphones
\( n \) | time frame index
Glossary of Acronyms, Symbols, and Notation

\( t \)  
 discrete time index

\( T_{60} \)  
 reverberation time

\( \Gamma(k) \)  
 spatial coherence

\( \Gamma_r(k) \)  
 spatial coherence matrix of the reverberation

\( h(t) \)  
 room impulse response

\( \phi(k, n) \)  
 power spectral density

\( \Phi(k, n) \)  
 PSD matrix

\( R^{(m)}(k, n) \)  
 reverberation sound component in the STFT domain

\( r(k, n) \)  
 multichannel reverberation sound component in the STFT domain

\( s(t) \)  
 speech signal

\( S_{d,1}(k, n) \)  
 plane wave signal in the STFT domain

\( s_d(k, n) \)  
 vector of \( I \) plane wave signals in the STFT domain

\( S_D(k, n) \)  
 early speech signal in the STFT domain

\( s(k, n) \)  
 multichannel early speech signal in the STFT domain

\( v(t) \)  
 noise signal

\( V_{n}(k, n) \)  
 noise signal in the STFT domain

\( v(k, n) \)  
 multichannel noise component in the STFT domain

\( x(t) \)  
 reverberant speech signal

\( X_{m}(k, n) \)  
 reverberant speech signal in the STFT domain

\( x(k, n) \)  
 multichannel reverberant speech signal in the STFT domain

\( y(t) \)  
 microphone signal

\( Y_{m}(k, n) \)  
 microphone signal in the STFT domain

\( y(k, n) \)  
 multichannel microphone signal in the STFT domain
# Contents

Acknowledgments ......................................................... i

Abstract ................................................................. iv

Kurzfassung ............................................................... vii

Glossary of Acronyms, Symbols, and Notation .......................... ix

List of Figures ........................................................... xvii

List of Tables ............................................................. xix

1 Introduction ............................................................. 1
  1.1 Problem description and motivation ................................ 1
      1.1.1 Physical and perceptual description of reverberation ...... 1
      1.1.2 Room acoustical measures .................................. 3
      1.1.3 Applications of dereverberation ........................... 5
  1.2 Classes of dereverberation methods ................................. 6
  1.3 Challenges and open issues ....................................... 10
  1.4 Outline and main contributions .................................. 11
      1.4.1 Structure of this thesis ................................... 11
      1.4.2 List of publications ...................................... 12

2 STFT domain signal models for dereverberation ..................... 15
  2.1 STFT filterbank .................................................. 16
  2.2 Multiplicative transfer function model ........................... 17
  2.3 Narrowband moving average model ................................ 17
  2.4 Narrowband multichannel autoregressive model ................... 19
  2.5 Temporal exponential decay model ................................ 19
  2.6 Spatial coherence based model ................................... 21
      2.6.1 Spatial description sound fields .......................... 21
      2.6.2 Signal model ............................................. 22
      2.6.3 Validity of the spherical diffuse model for reverberation 23
  2.7 Discussion of the reverberation models ........................... 24
      2.7.1 Relations between the models ............................. 24
2.7.2 Underlying models for MA and MAR filter coefficients ........................................ 25
2.8 Conclusion .................................................................................................................. 26

3 Spectral and spatial reverberation suppression ............................................................ 27
3.1 Single-channel dereverberation using spectral enhancement ........................................ 27
3.1.1 Signal model .......................................................................................................... 28
3.1.2 MMSE estimation of early speech signal ................................................................. 28
3.2 Multichannel dereverberation using spatial filtering ................................................... 29
3.2.1 Signal model .......................................................................................................... 29
3.2.2 MMSE estimation of desired signal ....................................................................... 30
3.2.3 Parameter estimation of the MWF ......................................................................... 31
3.3 Reduction control for the MWF .................................................................................. 35
3.3.1 Definition of the target signal ................................................................................. 35
3.3.2 Derivation of the proposed filter ........................................................................... 35
3.3.3 Analysis for the rank-one case .............................................................................. 36
3.3.4 Experimental validation ....................................................................................... 37
3.3.5 Conclusion .............................................................................................................. 38
3.4 Summary ..................................................................................................................... 38

4 Single-channel late reverberation PSD estimation ......................................................... 41
4.1 Problem formulation .................................................................................................... 41
4.2 State-of-the-art single-channel methods ..................................................................... 42
4.2.1 Using the exponential decay model ...................................................................... 42
4.2.2 Using a narrowband moving average model .......................................................... 44
4.3 Late reverberation PSD estimation using relative convolutive transfer functions .... 44
4.3.1 Markov model for relative CTFs ............................................................................ 44
4.3.2 Kalman filter to estimate RCTF coefficients ......................................................... 45
4.4 Performance evaluation ............................................................................................... 47
4.4.1 Setup ...................................................................................................................... 47
4.4.2 CTF filter length and reverberation time ................................................................. 48
4.4.3 Performance in time-varying conditions ................................................................. 49
4.4.4 Speech enhancement performance .......................................................................... 51
4.5 Summary ..................................................................................................................... 52

5 Multichannel late reverberation PSD estimation ......................................................... 53
5.1 Problem formulation .................................................................................................... 55
5.2 State-of-the-art reverberation PSD estimators ............................................................ 56
5.2.1 Spatial coherence based direct PSD estimators ...................................................... 56
5.2.2 Spatial coherence based indirect PSD estimators .................................................. 62
5.2.3 Temporal model based PSD estimators .................................................................. 63
5.3 PSD matrix based least-squares method with blocking .............................................. 63
5.4 Discussion of multichannel reverberation PSD estimators ......................................... 65
5.5 Bias compensation ...................................................................................................... 66
5.6 Evaluation of reverberation PSD estimators ........................................... 67
  5.6.1 Acoustic setup and simulation parameters .................................... 67
  5.6.2 Signal generation and performance measures ................................. 68
  5.6.3 Performance evaluation for a single source .................................... 70
  5.6.4 Evaluation for the multi-wave model ........................................... 75
5.7 Summary ......................................................................................... 78

6 MIMO reverberation cancellation using a multichannel autoregressive model .... 79
  6.1 Signal model and problem formulation ............................................. 80
    6.1.1 Multichannel autoregressive reverberation model with additive noise .. 80
    6.1.2 Problem formulation .................................................................. 81
  6.2 State-space modeling of stochastic MAR coefficients ......................... 81
  6.3 MAR online dereverberation in the noise-free case .............................. 82
    6.3.1 Noise-free signal model ............................................................. 82
    6.3.2 State-of-the-art RLS estimation of the MAR coefficients ............... 83
    6.3.3 Kalman filter to estimate MAR coefficients ................................. 84
    6.3.4 Parameter estimation for Kalman filter ...................................... 84
    6.3.5 Relation to the RLS algorithm ................................................... 86
  6.4 MAR online dereverberation and noise reduction ................................. 87
    6.4.1 Noisy reverberant signal model and problem formulation .............. 87
    6.4.2 State-of-the-art MAP-EM method ............................................. 88
    6.4.3 MMSE estimation by alternating minimization using Kalman filters ... 89
  6.5 Reduction control ........................................................................... 94
  6.6 Discussion ..................................................................................... 95
  6.7 Evaluation ..................................................................................... 96
    6.7.1 Experimental setup ................................................................... 96
    6.7.2 Results for the dereverberation methods assuming no noise .......... 97
    6.7.3 Results for the noise and reverberation reduction methods .......... 98
  6.8 Summary ....................................................................................... 102

7 Evaluation and comparison of proposed dereverberation methods .............. 105
  7.1 Evaluation for speech enhancement ................................................... 105
    7.1.1 Setup .................................................................................... 105
    7.1.2 Experimental results .................................................................. 106
    7.1.3 Listening test .......................................................................... 110
  7.2 Evaluation for automatic speech recognition ...................................... 116
    7.2.1 REVERB challenge setup ....................................................... 116
    7.2.2 Automatic speech recognition results ....................................... 117
  7.3 Summary ....................................................................................... 118

8 Conclusion and outlook ....................................................................... 119
  8.1 Conclusion ................................................................................... 119
  8.2 Future work .................................................................................. 121
A  Signal energy ratio definitions for generating simulated signals 123

B  Performance measures 125
   B.1  Objective measures for speech enhancement 125
      B.1.1  Signal-based speech enhancement measures 125
      B.1.2  Signal-based EDC estimation 127
   B.2  Logarithmic PSD estimation error 128

C  Appendix to Chapter 6: Computation of residual noise and reverberation 129

Bibliography 131
### List of Figures

1.1 Recording a speaker in a reverberant room. ........................................... 2
1.2 A RIR and description of its parts. ......................................................... 2
1.3 Spectrograms of the direct sound and reverberant speech signal. ................. 3
1.4 Processing chain for typical hands-free communication scenario. .................. 5
1.5 Relations between reverberation models and enhancement classes. ................. 10

2.1 Modification of a microphone signal in the STFT domain. ........................... 16
2.2 Narrowband moving average reverberation model for multiple channels. ...... 18
2.3 Narrowband multichannel autoregressive reverberation model. .................... 18
2.4 Energy of a RIR at three different \( \frac{1}{3} \)-octave bands. .............................. 20
2.5 Plane wave arriving at two microphones. ............................................... 21
2.6 Spatial coherence of late reverberation and of a diffuse sound field. ............ 23

3.1 Structure of the general MWF with required estimation blocks. ................. 32
3.2 Structure of the decomposed rank-one MWF with required estimation blocks. ... 32
3.3 Typical microphone array geometries. .................................................... 33
3.4 Wiener gain for controlled Wiener filter and limited Wiener filter. ............... 37
3.5 Reverberation reduction and noise reduction for the MWF with reduction control. 38

4.1 Exponential model based late reverberation PSD estimators. ....................... 43
4.2 Structure of the RCTF based dereverberation method. ............................... 47
4.3 Absolute PSD estimation error in octave bands with different reverberation times. 49
4.4 PSD estimation error for various reverberation times and distances. ............. 49
4.5 Spectrograms of the late reverberation PSDs. ......................................... 50
4.6 Estimation error of late reverberation PSD for varying iSNR. ...................... 51
4.7 Segmental absolute estimation error of the late reverberation PSD. ............... 51

5.1 Classification of multichannel reverberation PSD estimators. ..................... 54
5.2 Reverberation PSD bias compensation depending on the DRR. ..................... 67
5.3 Room setup with circular array. .............................................................. 68
5.4 PSD error for artificial stationary sound field over iSNR. ......................... 70
5.5 PSD error for artificial stationary sound field over DRR. ........................... 71
5.6 PSD error for stationary sound field over steering error with DDR = 25 dB. .... 71
5.7 PSD error for stationary sound field over steering error with DDR = 0 dB. ....... 71
List of Figures

5.8 PSD error using simulated RIRs with $T_{60} = 600$ ms. .................................................. 72
5.9 PSD error using simulated RIRs with varying $T_{60}$. .................................................. 72
5.10 PSD error depending on the bin-wise DRR. ................................................................. 73
5.11 Speech distortion vs. interference reduction. ............................................................... 74
5.12 Improvement of perceptual measures. ....................................................................... 74
5.13 Speech distortion vs. interference reduction without and with bias compensation. .... 75
5.14 Improvement of perceptual measures without and with bias compensation .......... 75
5.15 PSD estimation error of stationary diffuse noise for multiple plane waves. ........... 76
5.16 Reverberation PSD estimation error for two simultaneously active speakers. .... 77
5.17 Speech distortion vs. interference reduction without and with bias compensation. 75
5.18 Improvement of perceptual measures without and with bias compensation. ......... 75
5.19 PSD estimation error of stationary diffuse noise for multiple plane waves. ........... 76
5.20 Reverberation PSD estimation error for two simultaneously active speakers. .... 77

6.1 Multichannel autoregressive signal model assuming no noise. ................................. 83
6.2 Generative multichannel autoregressive model. .......................................................... 88
6.3 State-of-the-art sequential noise reduction and dereverberation structure [92]. ....... 89
6.4 Proposed parallel dual Kalman filter structure. ........................................................... 90
6.5 Proposed structure to control the reduction of noise and reverberation. ................. 95
6.6 Special cases of the dual-Kalman framework. .............................................................. 96
6.7 Segmental objective measures for dynamic scenario. .................................................... 97
6.8 Objective measures of dual-Kalman for varying microphone number. ................. 98
6.9 Objective measures of dual-Kalman for varying filter length $L_{AR}$. ...................... 99
6.10 Simulated room setup for moving sources. ................................................................. 100
6.11 Segmental measures for a moving source. ................................................................. 101
6.12 Noise reduction and reverberation reduction for varying control parameters. ....... 102
7.1 Objective measures for varying the reduction control. .............................................. 107
7.2 Objective measures in different acoustic conditions. ................................................. 108
7.3 EDC in two different acoustic conditions. ................................................................. 108
7.4 Simulated array geometries. ....................................................................................... 109
7.5 Influence of the used number of microphones. ........................................................... 109
7.6 Objective measures using different arrays for dual-Kalman and MWF. .................. 110
7.7 GUI for the first session of the listening test. ............................................................... 111
7.8 GUI for the first session of the listening test. ............................................................... 112
7.9 Results of the listening test. ......................................................................................... 114
7.10 Perceived amount of reverberation per acoustic condition. ..................................... 115
7.11 Overall quality per acoustic condition. ..................................................................... 115
7.12 Word error rates of the ASR system from the REVERB challenge. ..................... 117

B.1 Estimated EDC using direct speech and noisy reverberant speech signals. ........... 128
## List of Tables

2.1 Deterministic or statistical modeling of the MA filter coefficients. . . . . . . . . . . 25
2.2 Deterministic or stochastic modeling of the MAR coefficients. . . . . . . . . . . . 26

4.1 Reverberation times per octave band and DRR of measured RIRs. . . . . . . . . 48
4.2 Speech enhancement measures using different late reverberation PSD estimators. 52

5.1 Classification and properties of reverberation PSD estimators. . . . . . . . . . 65

6.1 Objective measures for stationary pink noise and babble noise. . . . . . . . . . . 100

7.1 Statistical significance in terms of p-value using a t-test. . . . . . . . . . . . . . 114
7.2 Correlation of subjective score with objective measures. . . . . . . . . . . . . . 116
7.3 Correlation of WER with objective measures. . . . . . . . . . . . . . . . . . . . . 117
CHAPTER 1

Introduction

1.1 Problem description and motivation

From the rapidly growing market for speech communication systems and voice-operated systems, a wide variety of devices emerged like mobile phones, laptops, tablets, smart TVs, teleconferencing systems, hearing aids, voice operated systems in, medical applications, smart homes, smart loudspeakers, speech-to-text converters, etc. In most of these applications, it is desired that the users can move around freely, such that the microphones of these devices are located at a certain distance to the speech sources. Therefore, the received signals can be degraded by reverberation and noise, which can impair the speech quality and, in severe cases, the speech intelligibility [1].

1.1.1 Physical and perceptual description of reverberation

The phenomenon reverberation is defined as the following physical process: A sound source emits its sound waves in all directions with a certain radiation pattern. The emitted sound is reflected by furniture and the walls of a room. Finally, the sum of the direct sound and all reflections, that arrive at different times and with different amplitudes, is received by a microphone or by the human ear. The number of times a sound wave was reflected on an acoustic boundary surface before it arrives at the receiver is called the reflection order. The sum of all reflections is known as reverberation [2].

A schematic scenario, where some microphones are used to record a speaker in a reverberant room is shown in Figure 1.1. In addition to the desired direct sound, the microphones pick up also reflected the sound waves from the walls and sound from other noise sources, e.g. a fan, air conditioning, or ambient background noise.

In modern devices, the audio signals are digitally recorded, stored, processed and transmitted. In the digital domain, the reverberation process is described by a linear time-discrete filtering process with input sequence \( s(t) \) and output sequence \( x(t) \), where \( t \) is the discrete time sample index. A reverberant signal \( x(t) \) is then described by convolution of the source signal \( s(t) \) and the room impulse response (RIR) \( h(t) \) as

\[
x(t) = \sum_{t'=-\infty}^{\infty} h(t') s(t - t').
\] (1.1)
CHAPTER 1. INTRODUCTION

Figure 1.1: Recording a speaker in a reverberant room. The microphones pick up the direct sound, reflections and other noise sources.

Figure 1.2: A RIR and description of its parts.

When measuring RIRs and using them to model or simulate reverberant signals, commonly finite impulse response (FIR) filters are used for $h(t)$. From a perceptual point of view, a good rule of thumb for the necessary length $L_h$ of the finite RIR $h(t)$ is the reverberation time $T_{60}$ times the sampling frequency, because 60 dB is approximately the limit of the perceivable dynamic range of a single sound event for the human ear. Therefore, in practice the sum in (1.1) is finite if $h(t)$ is bounded in the interval $t = [0, L_h - 1]$. An exemplary RIR is shown in Figure 1.2.

The RIR is often separated into three parts, which are perceived differently, namely the direct sound, early reflections and late reverberation as shown in Figure 1.2. For speech signals, the separation between early and late reverberation is typically defined around 50 ms after the direct sound [2]. Early reflections do not harm or can even improve the speech intelligibility, as they are not perceived as separate sound events, and the human auditory system integrates their energy with the direct sound [3]. However, the perceptual effect of early reflections is often described as coloration or change in timbre, which may impair the signal quality. Reflections that arrive later than approximately 50 ms after the direct sound are not anymore integrated with the direct sound by the human auditory system and are perceived as separate sound. Furthermore, the density of the reflections increases with the time of arrival after the direct sound, such that after about 50 ms, the reflection density is so high that a noise-like decaying reverberation
1.1. **PROBLEM DESCRIPTION AND MOTIVATION**

tail is perceived [4, 5]. If the late reverberation tail, which originates from previously spoken phonemes, overlaps with current phonemes, the speech intelligibility can deteriorate due to masking effects [6–8]. This effect is called *overlap masking* and can be observed in Figure 1.3, where we see spectrograms of the direct sound component of a speech signal on top, and a spectrogram of the same speech signal in a reverberant environment below. In the spectrogram of the reverberant signal, we can observe a temporal smearing of the direct sound due to the reverberation.

The human hearing system has extraordinary abilities to focus on certain sounds in complex sound scenes, like a desired speaker, while blending out all other sounds that are not of interest such as environmental noise, undesired sound sources like other speakers, fans, etc., and also heavy reverberation. If a complex sound scene is recorded with a single microphone and reproduced somewhere else via headphones or loudspeakers, the important spatial and spectral cues are lost [9], which greatly improve the intelligibility and the ability to focus on certain sounds. In complex sound scenes with multiple sources like reverberation and noise, the human ability to extract the desired information from a recorded sound can be severely impaired. Therefore, the research field of speech enhancement is trying to process recorded audio signals for certain applications such that only sound sources of interest, e.g. the direct sound of a desired speaker, is extracted, while all other sound components such as reverberation and undesired noise are suppressed.

The typical dereverberation problem is *blind*, that means that the only available information are the microphone signals. A dereverberation approach is not completely blind if some additional information is given, e.g. the position of the speaker, information about the room like the reverberation time, or even the RIR.

1.1.2 **Room acoustical measures**

The reverberation characteristics for a single source-microphone position combination are fully described by the RIR $h(t)$. From the RIR, the following measures can be derived, which describe the room acoustics and the reverberation properties:
CHAPTER 1. INTRODUCTION

- The energy decay curve (EDC) describes the decay of the sound energy in a reverberant room in dB. It is given by the Schroeder integral \([10]\) for discrete time samples by

\[
\text{EDC}(t) = 10 \log_{10} \sum_{t' = t}^{\infty} h^2(t').
\] (1.2)

The EDC is typically normalized to the first sample at \(t = 0\) such that it yields a maximum of 0 dB.

- The reverberation time \(T_{60}\) describes the time of the free sound energy decay of 60 dB \([2]\). If a stationary signal is played in a room and switched off at time \(t\), the time until the energy dropped 60 dB below the level of the signal at time \(t\) is called \(T_{60}\). Therein, the EDC of a RIR measured in the farfield is assumed to have linear decay on the logarithmic scale. As the EDC is often not linear at the beginning due to high energy contributions of the direct sound and early reflections, and at the end due to measurement noise, the \(T_{60}\) is usually computed by extrapolating the decay between -5 dB and -35 dB to the 60 dB decay \([11]\).

- The direct-to-reverberation ratio (DRR) is the energy ratio between the direct sound and the reverberation, and is given by

\[
\text{DRR} = \frac{\sum_{t = 0}^{t_d} h^2(t)}{\sum_{t = t_d + 1}^{L_{h} - 1} h^2(t)}.
\] (1.3)

where \(t_d\) denotes the time index of the direct sound peak.

- The clarity index \(C_{50}\) is the early-to-late reverberation energy ratio defined as \([2]\)

\[
C_{50} = 10 \log_{10} \frac{\sum_{t = 0}^{t_{50}} h^2(t)}{\sum_{t = t_{50} + 1}^{L_{h} - 1} h^2(t)},
\] (1.4)

where \(t_{50}\) denotes the time index 50 ms after the direct sound peak. The motivation behind this measure is that the transition between early and late reverberation is usually defined as the time 50 ms after the direct sound peak.

- The critical distance is defined as the distance from the sound source, where the DRR given in (1.3) is one, i.e. when there is an equal energy balance between direct sound and reverberation. By assuming that the reverberation has the properties of a diffuse sound field, the critical distance can be computed by \([2]\)

\[
d_{\text{crit}} = \frac{1}{4} \sqrt{\frac{\gamma_{\text{dir}} A_{\text{absorb}}}{\pi}},
\] (1.5)

where \(A_{\text{absorb}}\) is the equivalent absorption surface of the room in m\(^2\) and \(\gamma_{\text{dir}}\) is the directivity factor of the source.
1.1. PROBLEM DESCRIPTION AND MOTIVATION

1.1.3 Applications of dereverberation

In the following, typical applications, where dereverberation is useful or required are described.

**Speech communication** The main applications for which several solutions are proposed in this thesis are speech communication scenarios using distant microphones from the speaker, such as in hands-free or distant-talking mode for devices used for communications, or hearing aids. In telecommunication, the delay introduced by processing and transmitting the audio signals has to be sufficiently short: the typically accepted delay ranges up to 150 ms [12]. Therefore, speech enhancement algorithms used to process the signals before transmission have to return the audio data in real-time up to a short processing delay. **Batch processing** algorithms are therefore not suitable for this task, as they require some amount of data, typically in the range of several seconds, before they can start processing. All methods proposed in this thesis are **online** processing algorithms, which can deliver an output signal with a short algorithmic delay of about 10-30 ms.

In some cases, the microphone array has a special structure, like in the cases of binaural hearing aids or spherical microphone arrays. Although we contributed methods designed for these special applications in [13,14] for binaural hearing aids, and in [15] for spherical microphone arrays, we do not consider such special cases explicitly in this thesis, but consider rather general applicable algorithms. In general speech communications setups as shown in Figure 1.4, there can be further interferences at the near-end, such a loudspeaker emitting the signal from the communication partner at the far-end. In addition, it is desired to transmit the audio signal with an equal long-term level to enhance the listening comfort, and to keep the signal level at the optimal working point for quantization of audio transmission codecs [16]. Therefore, several processing blocks are required [17], such as acoustic echo control, noise reduction and dereverberation, and automatic gain control (AGC) as shown in Figure 1.4. In [18] we have proposed a method for automatic spatial gain control as an extension of a spatial filter that is designed for dereverberation and noise reduction. This thesis deals with dereverberation and noise reduction methods that could be used in a similar communication setup as shown in Figure 1.4.

**Figure 1.4**: Typical hands-free communication scenario and required processing chain on the near-end side.
CHAPTER 1. INTRODUCTION

Automatic speech recognition  A research field with growing popularity and importance is automatic speech recognition (ASR). The performance of modern ASR systems has improved rapidly over the recent years. Good performance can be achieved in conditions with in little to moderately reverberant environments when using close-talking microphones, where commercial systems can achieve word error rates (WERs) in the range of 1-3%. However, the WER drops severely in reverberant conditions when using distant microphones [19, 20]. Although the robustness of modern ASR systems can be improved by training in reverberant conditions, multi-condition training is very time- and resource-consuming. The ASR performance can still be further improved by pre-processing the signals using dereverberation algorithms as shown in [20]. In the ideal case, a good dereverberation pre-processing would make training for multiple acoustic conditions obsolete.

Music production  A different application for dereverberation from speech communication is recording of music performances in acoustically suboptimal or undesired locations. Sometimes a room with the desired acoustical properties is not available, or the location of the performance is given by outer circumstances. In this case, dereverberation software can be useful to remove the reverberation from the recording, and artificial reverberation can be added later in the mix if desired. Compared to dereverberation for speech enhancement, dereverberation software for music is a quite recent development, since the quality requirements in music production are much higher than in speech communication applications. Furthermore, the signal properties of music signals are much more diverse than the properties of human speech. Examples for commercially available dereverberation software are Acon Digital DeVerbate¹, Sonible Freiraum², SPL Deverb³, Dyvision ReverbRemover⁴, Zynaptic Unveil⁵ and Accusonus ERA-R⁶. Although the methods proposed in this thesis could be adapted for the processing of music signals, this thesis considers only speech applications.

1.2 Classes of dereverberation methods

In following, an exhaustive overview of dereverberation classes is provided. Note that we define the classes depending on the operation that is performed to obtain the dereverberated signal. Although some classes are strongly connected to certain signal models, some classes can share signal models. Dereverberation algorithms can be categorized into the following classes:

1. Acoustic channel inversion or equalization: Algorithms of this class identify the acoustic system and then equalize it (cf. [1] and the references therein). Given a perfect estimate of the acoustic system described by a FIR, perfect dereverberation can be achieved by applying the multiple-input/output inverse theorem (MIIN) [21] (i.e., by applying a multichannel equalizer). However, this approach is not robust against estimation errors of

¹ http://acondigital.com/products/deverberate/
² http://www.sonible.com/de/freiraum/
⁴ http://www.dyvision.co.uk/reverbremover.html
⁵ http://www.zynaptiq.com/unveil/
⁶ https://accusonus.com/products/era-r
1.2. CLASSES OF DEREVERBERATION METHODS

the acoustic impulse responses. As a consequence, this approach is also sensitive to changes in the room, and to position changes of the microphones and sources. For a single source, more robust equalizers were developed in [22, 23]. Additive noise is usually not taken into account, although first attempts have been made recently in [24]. However, available methods for blind estimation of long time domain RIRs in the presence of noise are not sufficiently robust and accurate in practice [25,26]. Furthermore, in the presence of multiple sources, multiple RIRs, i.e. a multiple-input-multiple-output (MIMO) system, have to be blindly identified, which is theoretically not impossible but extremely challenging [27].

2. Joint system identification and source estimation: In [28] a particle filtering framework was proposed, in order to identify all-pole acoustic channel responses and to estimate the speech signal modeled as a time-varying autoregressive (AR) process. In this model, the additive noise is filtered by the same channel response as the speech, which is not realistic and might be a problem in noisy scenarios. Due to the complexity of the system, realistic results for long acoustic channels could not be shown.

More recently, multichannel system identification methods using FIR reverberation models have been developed that jointly estimate the acoustic system and the speech signal [29,30]. The desired speech signal is then obtained inherently in these approaches. Furthermore, additive noise is properly taken into account in these models. The method in [30] estimates the RIRs in the frequency domain using a variational Bayesian framework. Due to the time domain FIR model, long frames are required to model long RIRs, otherwise only the direct path and early reflections can be identified. In [29] an approximation of the finite RIR model in the short-time Fourier transform (STFT) domain is used, namely the convolutive transfer function (CTF) model [31], which ignores correlation between STFT frequency bins. The method uses a recursive expectation-maximization (EM) framework, where the speech signal is estimated in the E-step by a Kalman filter. Due to the scaling ambiguity of the first CTF coefficient, no solution has been found yet to correctly identify this coefficient. Therefore, this method cannot remove the early reflections. Although the method is computationally rather complex, promising results can be achieved.

3. Multichannel linear prediction based system identification and equalization: Algorithms of this class, e.g. [32–35], describe the reverberant signals using an AR or multichannel autoregressive (MAR) model. In the multichannel case, the model predicts the reverberation in one microphone channel from previous samples of all channels, while consequently, the prediction error corresponds to the desired speech signal. Early approaches based on this model were in the time domain: The approach proposed in [32] estimates the clean speech signal of a single-source based on multichannel linear prediction by enhancing the linear prediction residual of the clean speech. In [36] a multichannel linear prediction filter is computed from pre-whitened and time aligned microphone signals. In [37] multi-step long-term linear prediction is used to estimate only the late reverberation, and the desired speech signal is then estimated by spectral subtraction. As these time domain approaches have still suffer from the problem of whitening the speech signal too much, a pre-whitening of the signal has to be used to alleviate this effect.
Promising results have been obtained by recent methods based on this model formulated in a subband domain such as the STFT domain. In [33–35] the reverberant signal is modeled as an MAR process and the regression coefficients are estimated from the observations. The clean speech is then estimated using the regression coefficients. Advantages of the MAR model are that it is also valid in the presence of multiple sources [34, 35], and the regression coefficients are more robust against position changes in the room than, e.g., FIR models [38]. Unfortunately, linear prediction based dereverberation algorithms are computationally complex and sensitive to noise, because in the presence of additive noise, the autoregressive signal cannot be observed directly. Systems that remove the noise prior to linear prediction based dereverberation have been proposed in [39, 40]. In Chapter 6, an online solution for joint dereverberation and noise reduction based on the MAR model is proposed.

4. **Spectral enhancement**: Exclusively spectral filters are typically single-channel approaches that modify the signal in the short-time spectral domain. As in these approaches, a real-valued filter gain is applied to each frequency bin of the signal, only the signal magnitude is modified, while the phase remains unchanged. In the short-time spectral domain, the reverberation can be modeled as an independent additive component. Therefore, only the late reverberation is properly modeled and can be suppressed using these techniques, e.g., as proposed in [41, 42] or more recently in [43, 44]. The early reflections that arrive within the same frame as the direct sound cannot be removed. One of the most popular spectral filters, the Wiener filter (WF), is presented in Section 3.1.

The quality and performance of spectral enhancement methods directly depend on the estimation of the power spectral densities (PSDs), which are in the case of reverberation highly time-varying. Chapters 4 and 5 deal with the estimation of the late reverberation PSD using different models for the single- and multichannel case. Spectral enhancement approaches generally enjoy large popularity, because an extension to joint reverberation and noise suppression can be straightforward using existing noise estimation algorithms, and well studied simple techniques to control artifacts can be employed.

5. **Spatial filtering**: Spatial filters, commonly known as beamformers, weight and combine multiple microphone signals to achieve a certain spatial pickup pattern. Typically, this spatial pattern is a beam pointing towards the source of interest, i.e., it aims at extracting the direct sound while suppressing other directions and noise [45]. Whereas in many applications spatial filters are time-invariant or computed only depending on the steering vector, a higher performance can be achieved using signal-dependent spatial filters. In [46] a parametric sound field model in the STFT domain was proposed, where at each frequency the desired sound is modeled as one or multiple plane waves and the reverberation is modeled as a diffuse sound field with a time-varying PSD and a time-invariant spatial coherence. If these parameters can be estimated almost instantaneously in each frame, the filter can quickly adapt to changes in the sound field. As these filters typically depend on the direction-of-arrival (DOA) of the plane waves and the second-order statistics of the signal components, they provide an optimal tradeoff between dereverberation and
noise reduction. Their dereverberation performance is highly dependent on the estimation accuracy of the reverberation PSD, which is a challenging task, because the direct sound and reverberation can generally not be observed separately.

In contrast to spectral filters, spatial filters can be designed distortionless with respect to the desired signal, such as the minimum variance distortionless response (MVDR) beamformer. Some widely used spatial filters such as the multichannel Wiener filter (MWF) are generalized forms of single-channel spectral filters and can under some assumptions be decomposed into a distortionless spatial filter and a spectral post-filter [47]. In Section 3.2 the MWF is introduced and Chapter 5 deals with multichannel estimation of the late reverberation PSD.

6. **Subband multi-frame filtering:** In contrast to spectral or spatial enhancement approaches where only a single time frame is filtered per frequency, recently subband filtering approaches based on second-order statistics (e.g. multi-frame MVDR or Wiener filters) have been proposed for noise reduction [48–50]. These filters are designed to directly obtain an optimal estimate of the desired signal by minimizing a cost-function. This subband filtering framework has been used for dereverberation in [51] using the CTF model. Whereas promising results can be achieved in theory, the estimation of the highly time-varying inter-frame correlation coefficients is challenging, which in practice introduces strong speech distortion and artifacts that are difficult to control. Additive noise has not been considered in [51].

7. **Cepstral enhancement:** As the convolution of the acoustic RIR with the speech signal in the time domain translates to an addition in the cepstral domain, several attempts have been made to develop dereverberation methods in the cepstral domain [52–55]. A popular and simple approach is cepstral mean subtraction [56] as typically used in the front-end processing of state-of-the-art ASR systems [57]. However, cepstral domain approaches have three major disadvantages: i) These approaches are capable of only reducing early reflections, which are in most applications a minor problem than the late reverberation tail. ii) Within this model it is unknown how to model additive signal components such as noise or multiple speakers. As soon as there is any additive signal component present, the additive model in the cepstral domain is violated. iii) The real cepstrum is simple to compute, but is only related to the magnitude or a minimum-phase system response. The complex cepstrum can encapsulate mixed-phase system responses, but can be obtained only using computationally complex and error sensitive methods such as polynomial rooting [58, 59]. iv) It is very challenging to control perceptual artifacts introduced by modifying the cepstrum and to avoid modifying speech related cepstral coefficients [60], which makes these approaches less popular.

8. **LPC residual processing:** linear predictive coding (LPC) that is popular in the field of speech coding, can be used to separate the speech spectral envelope from the excitation signal, i.e. the LPC residual [61]. A class of dereverberation methods operates on the LPC residual by assuming that the reverberation effects are mainly present in the LPC residual,
while the spectral envelope remains rather unaffected [62, 63]. Some methods have been proposed to reduce the reverberation effects from the LPC residual and then to synthesize the dereverberated signal using the extracted speech spectral envelope [55, 63, 64]. A major problem of this framework is that additive noise and multiple sources violate the signal model. Furthermore, impulsive speech signals cannot be modeled by the time domain LPC model.

Figure 1.5 shows the relations between the underlying signal models and the different enhancement methods. The class numbers are indicated below each class. It is important to note that some of these techniques can be combined. The probably most popular technique combination is a spatial filter with spectral post-filter, as this combination is obtained naturally by the MWF under rank-one assumption of the desired signal. Another example of a popular combination is to use cepstral mean subtraction [56] for reducing early reflections after any dereverberation technique that reduces only late reverberation. This combination is commonly used in state-of-the-art ASR systems.

### 1.3 Challenges and open issues

Although dereverberation has been a topic of research for many decades, the problem is still far from being solved completely. The facts that the problem is completely blind and that accurate modeling of the reverberation is highly complex make the dereverberation problem one of the most challenging tasks in speech enhancement.

In the last 5-10 years, significant progress has been made in the research community. There exist some practical and robust methods that can suppress late reverberation at the expense of some speech distortion. However, the tradeoff between reverberation reduction and speech distortion of most state-of-the-art methods is often biased towards the reduction. Conversely speaking, when designing algorithms for minimal speech distortion, the possible amount of reverberation reduction is very limited.

The aim of this thesis is therefore to develop dereverberation methods with the following requirements:
1. The algorithms should be able to provide sufficient reverberation reduction and low speech distortion.

2. The algorithms should be designed for online processing, while keeping the processing delay as short as possible.

3. The algorithms should be able to adapt sufficiently fast to changes in the acoustic scene.

4. The algorithms should be robust against slowly time-varying additive noise and should be able to reduce both reverberation and noise.

5. The algorithms should be generalizable to multiple sources, or still be robust in the presence of multiple sources.

The number of state-of-the-art methods fulfilling the requirements 2 - 5 simultaneously is very limited, while in terms of the first requirement there is always room for improvement, and state-of-the-art solutions, especially methods using a small number of microphones, leave sufficient room for improvement [20].

1.4 Outline and main contributions

The structure of this thesis is outlined in Section 1.4.1, where the main aspects and novel contributions to the research field are emphasized. An overview of the publications that make up the main content of the thesis is provided in Section 1.4.2.

1.4.1 Structure of this thesis

Chapter 2 provides an extensive overview of all reverberation signal models in the short-time spectral domain that have been successfully used for dereverberation. A discussion about relations between and advantages of the signal models is included. In Chapter 3 single- and multichannel filters for reverberation suppression are introduced, and a novel generalized method to control the residual reverberation and noise at the output of these filters independently is proposed. The filters presented in Chapter 3 depend on estimates of the reverberation PSD. In Chapters 4 and 5, single- and multichannel reverberation PSD estimators are proposed. While the proposed multichannel reverberation PSD estimator described in Section 5.3 was one of the first direct reverberation PSD estimators based on a spatial coherence model, the development of several other PSD estimators based on the same model, which are included in the state-of-the-art review, were published chronologically after our publication. In addition, an extensive overview and experimental comparison between all considered reverberation PSD estimators is presented in Chapter 5. In Chapter 6 a novel method for reverberation cancellation based on a multichannel autoregressive model is proposed. In total, three very different dereverberation methods based on completely different signal models have been proposed in this thesis. In Chapter 7 an extensive experimental comparison between these three developed dereverberation methods is conducted. The evaluation is based on objective measures, a subjective listening test, and results from an ASR system. By considering the three dereverberation methods as representatives
for each signal model or enhancement method, new insights on the usefulness and effectivity of methods based on these models are provided, which will help in further research in the field of speech dereverberation.

1.4.2 List of publications

The following publications contain the main contributions of this thesis and are closely related to the chapters and sections.


- **Section 5.2.1.1:** "Maximum likelihood estimation of the late reverberant power spectral density in noisy environments" - O. Schwartz, S. Braun, S. Gannot and E. A. P. Habets, *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.


- **Section 4.3:** "Late reverberation PSD estimation for single-channel dereverberation using relative convolutive transfer functions" - S. Braun, B. Schwartz, S. Gannot and E. A. P. Habets, *In International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.


- **Section 3.3:** "Linear prediction based online dereverberation and noise reduction using alternating Kalman filters", - S. Braun, E. A. P. Habets, *to appear in IEEE/ACM Transaction on Speech, Audio and Language Processing.*
In addition to the publications above, the following publications contributed significant insights to the algorithms developed in the course of this thesis, but are not directly related to the chapters and are not presented in this thesis.


CHAPTER 2

STFT domain signal models for dereverberation

In contrast to mathematical models that try to describe a physical phenomenon as accurately as possible, acoustical signal models used in speech enhancement are often very simple and coarse models based on severely simplified assumptions. Classical speech enhancement tasks such as noise reduction and dereverberation usually are completely blind problems, where all necessary information has to be estimated exclusively from the microphone signals. Therefore, simpler signal models with a reduced number of unknown parameters leading to less complex mathematical solutions often work better in practice than more exact, but complex models.

Many speech enhancement techniques, especially noise reduction techniques [65], are designed in or operate in a time-frequency domain or subbands. The short-time Fourier transform (STFT) [66] is a transform that can be interpreted as a filterbank, which is widely used in speech enhancement. As in typical far-field communication scenarios reverberation and noise are usually present simultaneously, it is desired to design techniques that can reduce reverberation and noise jointly, while dereverberation techniques not considering noise can suffer from instability or other issues in noisy conditions. The majority of established noise reduction techniques operate in the STFT domain typically using time frames of 10 – 30 ms length, such that short-term stationarity within one STFT frame can be assumed for all sound components, especially for the speech component [65,67]. To simplify the integration of knowledge from existing noise estimation and reduction techniques into joint dereverberation and noise reduction techniques, it is desirable to design joint methods in the same domain.

The design of dereverberation algorithms in the STFT domain has a second advantage: In very reverberant environments, the length $L_h$ of the room impulse response (RIR) is in the range of several thousand taps, which results in extremely challenging estimation problems. The problem becomes even harder if the acoustic scenario, i.e. the RIR $h(t)$, changes over time. By using the STFT with overlapping frames of e.g., 50% or 75%, the temporal resolution is drastically reduced. Therefore, also the required filter length to describe a RIR in the STFT domain is reduced, which can simplify blind estimation problems by a large amount.

To design dereverberation methods in the STFT domain, we require a model to describe the reverberation in this domain. Within this thesis, several STFT domain reverberation signal models are used, which are presented as an overview in the remainder of this chapter. In Section 2.1, the STFT filterbank is introduced, and it is shown how the signals are processed.
in the STFT domain. In Sections 2.2 – 2.6, several models are introduced that show, how the reverberant signals can be described in the STFT domain. In Section 2.7, the signal models are discussed and relations are shown.

2.1 STFT filterbank

Within this thesis, we use a STFT filterbank representation of the acoustical signals. The STFT representation \( Y(k, n) \) of the time-domain signal \( y(t) \) is given by [66]

\[
Y(k, n) = \sum_{t=-\infty}^{\infty} y(t) w_a(t - nN_{\text{hop}}) e^{-j2\pi \frac{k}{K} f_s(t-nN_{\text{hop}})},
\]

(2.1)

where \( w_a(t) \) is the analysis window, \( N_{\text{hop}} \) is the hop-size or frame shift (in the filterbank interpretation considered as downsampling factor) corresponding to a time interval of \( T_{\text{hop}} = N_{\text{hop}} f_s \), \( K \) is the discrete Fourier transform (DFT) length, \( f_s \) is the sampling frequency, \( k \) and \( n \) are the frequency and time frame indices, respectively, and \( j = \sqrt{-1} \) is the imaginary unit.

The time domain signal can be re-synthesized from the STFT representation \( Y(k, n) \) using the inverse STFT as

\[
y(t) = \sum_{t=-\infty}^{\infty} w_s(t - nN_{\text{hop}}) \frac{1}{K} \sum_{k=0}^{K-1} Y(k, n) e^{j2\pi \frac{k}{K} f_s(t-nN_{\text{hop}})},
\]

(2.2)

where \( w_s(t) \) is the synthesis window. By choosing the analysis and synthesis windows appropriately, the time domain signal \( y(t) \) can be perfectly reconstructed [68]. Note that the analysis and synthesis windows \( w_a(t) \) and \( w_s(t) \) are of finite length and can be chosen identically. To avoid circular convolution effects, the number of frequencies \( K \) can be chosen larger than the length of the analysis and synthesis windows. In this case, the signal is zero-padded before the DFT analysis.

In STFT domain processing schemes, one or multiple microphone signals \( Y(k, n) \) are modified to obtain the processed signal \( Z(k, n) \). The processing is often independent for each frequency. The final output signal in the time domain \( z(t) \) is then obtained from \( Z(k, n) \) using (2.2) [68]. The STFT domain processing procedure is illustrated in Figure 2.1 for a single microphone.
signal $y(t)$. Note that also multiple microphone signals can be used to generate either a single or multiple output signals. When used in real-time processing systems, the STFT introduces an algorithmic delay of the window length, since also with overlapping frames, we have to wait until one window frame of data is received, before one output block can be computed.

### 2.2 Multiplicative transfer function model

The DFT of convolution, e.g. of the reverberant signal $x(t)$ given in (1.1) transforms to multiplication of the DFTs of $h(t)$ and $s(t)$ [69]. However, if the analysis window is of finite length, there are crossband filtering effects between the DFTs of $h(t)$ and $s(t)$. If these crossband filtering effects are neglected and if the RIR $h(t)$ is shorter than the STFT analysis window length, the convolution between the RIR $h(t)$ and the speech signal $s(t)$ given in (1.1) can be described as a multiplication in the frequency domain, i.e.

$$X(k, n) = H(k, n)S(k, n),$$

(2.3)

where $X(k, n)$, $H(k, n)$ and $S(k, n)$ are the STFT domain representations of the reverberant signal $x(t)$, the RIR $h(t)$ and the source signal $s(t)$. The model (2.3) is called multiplicative transfer function (MTF) model [70].

In reverberant environments, the effective length of the RIRs is typically in the range of 0.5 s to 1 s, or in extreme cases even longer. To use the MTF model for enhancement in reverberant environments, unusually long analysis windows would be required. However, the use of such long windows contradicts a common assumption in many speech enhancement algorithms like noise reduction, which assume that the speech signal $s(t)$ is stationary within one analysis frame. Short-term stationarity of non-reverberant speech can be typically assumed within analysis windows of 20 to 30 ms length. Furthermore, long analysis windows would increase the processing delay, which is desired to be as low as possible in real-time speech communication applications.

Some dereverberation classes are based on the MTF assumption, such as the cepstral enhancement, LPC residual processing, but also acoustic channel inversion and equalization suffer from the same problem when they are implemented using block processing. When using shorter window lengths than the RIR length due to practical reasons, these methods usually are unable to reduce the late part of the reverberation. Therefore, we do not consider the MTF model further in this thesis.

### 2.3 Narrowband moving average model

In contrast to the MTF model, when the STFT analysis window is shorter than the effective length of the RIR, the convolution (1.1) for a RIR of finite length is can be described in the STFT domain as a sum of $K$ crossband convolutions [71, 72]

$$X(k, n) = \sum_{k'=0}^{K-1} \sum_{\ell=0}^{L} H_{\ell}(k, k', n) S(k', n - \ell),$$

(2.4)
where $H_\ell(k, k', n)$ for $\ell = \{0, 1, \ldots, L\}$ is the time-frequency response of the RIR. Note that in general we assume that $H_\ell(k, k', n)$ is time-varying, such that there exists a set of $L + 1$ filter coefficients at each frame index $n$. The crossband filters take the crosstalk between the frequency bands due to non-ideal subband filtering of the STFT filterbank into account [73]. By assuming that the crossband RIR coefficients are zero $H_\ell(k, k', n) = 0$ for $k \neq k'$, the reverberant signal is described by an independent convolution, i.e. a finite impulse response (FIR) or all-zero filtering operation, per frequency band $k$. Under this assumption, the model is also called convolutive transfer function (CTF) model [74] and is given by

$$X(k, n) = \sum_{\ell=0}^{L_{\text{MA}}} H_\ell(k, n) S(k, n - \ell), \quad (2.5)$$

where $H_\ell(k, n)$ for $\ell = \{0, 1, \ldots, L\}$ are the time-varying filter coefficients and $L_{\text{MA}}$ is the CTF length. If we consider multiple microphone channels $M$, we denote (2.5) as

$$x(k, n) = \sum_{\ell=0}^{L_{\text{MA}}} h_\ell(k, n) S(k, n - \ell), \quad (2.6)$$

where the vectors $x(k, n) = [X_1(k, n), \ldots, X_M(k, n)]^T$ and $h_\ell(k, n) = [H_{\ell}^{(1)}(k, n), \ldots, H_{\ell}^{(M)}(k, n)]^T$ contain the reverberant signals and the filter coefficients at each microphone. Note that in the single-channel case, the microphone index $m \in \{1, \ldots, M\}$ is omitted. If the speech signal $S(k, n)$ is assumed as an independent random variable over the time frames $n$, the reverberant signal vector $x(k, n)$ given by (2.6) is generated by a moving average process. Figure 2.2 shows the generative process of the moving average (MA) reverberation model.

An advantage of the narrowband MA model\(^\dagger\) compared to the time domain FIR model (1.1) is that the required filter lengths $L_{\text{MA}}$ are relatively short, typically covering a range of 10 to 50 frames, whereas the effective length $L_h$ of RIRs in the time domain typically is in the range of several thousand samples.

\(^\dagger\)Note that (2.5) is strictly speaking not a MA process without the assumption that $S(k, n)$ is temporally uncorrelated. However, this assumption is typically made in speech enhancement methods.
2.4 Narrowband multichannel autoregressive model

In contrast to the narrowband MA model presented in the previous section, the reverberant signal can be described as a multichannel autoregressive (MAR) process. In this model it is assumed that the reverberation in the current frame can be predicted from previous reverberant signal frames using a linear prediction filter, also known as room regression coefficients [39], whereas the early speech signal that contains only direct sound and early reflections is the innovation, which cannot be predicted from past frames. Alternatively, we interpret the reverberant multichannel signal to be generated by a multiple-input-multiple-output (MIMO) infinite impulse response (IIR) filtering system, more specifically, an all-pole filter. Using the MAR model, the reverberant microphone signals are given by

\[ x(k,n) = \sum_{\ell=D}^{L_{AR}} C_\ell(k,n) x(k,n - \ell) + s(k,n), \quad (2.7) \]

where the matrices \( C_\ell(k,n) \) of size \( M \times M \) contain the MAR coefficients, which are related to the room, the vector \( s(k,n) = [S_D^{(1)}(k,n), \ldots, S_D^{(M)}(k,n)]^T \) contains the desired early speech at each microphone \( S_D^{(m)}(k,n) \), \( D \geq 1 \) is the prediction delay, \( L_{AR} > D \) is the number of past frames that are required to predict the reverberant signals \( x(k,n) \), and \( M \) is the number of microphones. The choice of the delay \( D \) determines how many early reflections are contained in the early speech signals \( s(k,n) \). As the speech signal vector \( s(k,n) \) is the unpredictable signal component, it is assumed to be an independently distributed random variable.

Note that we generally assume that the room regression coefficients \( C_\ell(k,n) \) can be time-varying as indicated by the dependence on the frame index \( n \). Since the MAR model does not model a speech source signal or the direct sound, like the MA model in Section 2.3, the MAR model is also valid for multiple active speakers, which can be contained in the early speech signal vector \( s(k,n) \). Figure 2.3 shows the generative MAR process of the reverberant multichannel speech signals. Note that in the MAR model, the reverberant signal \( X_m(k,n) \) at channel \( m \) depends on previous frames of all channels \( m \in \{1, \ldots, M\} \) (due to the matrix-vector multiplications in (2.7)), whereas in the MA model shown in Figure 2.2, the reverberation is modeled independently across the microphone channels (see (2.5)).

2.5 Temporal exponential decay model

A popular and widely used model in speech enhancement is based on the assumption that the RIR can be modeled as exponentially decaying noise, known as Polack’s model [4]. In [75] this model was generalized to decouple the energy of the direct sound from the exponentially decaying reverberation tail. This decoupling takes into account that within the critical distance, the energy of the direct sound can be much stronger than the energy of the reverberation tail and would therefore severely violate the exponential decay model.

The derivation of the model starts from the time domain convolution equivalent in the STFT domain given in (2.4). The room coefficients \( H_\ell(k,k',n) \) are assumed to be zero-mean Gaussian
random variables with $\mathcal{E}\{|H_\ell(k,k',n)|^2\} = 0$ for $k \neq k'$. Therefore, crossband filtering effects are neglected, and we arrive at the formulation of the narrowband MA model given in (2.5). Furthermore, the room coefficients $H_\ell(k,k',n)$ are modeled as exponentially decaying noise [75]

$$H_\ell(k,k',n) \equiv H_\ell(k,n) = \begin{cases} B_{\text{dir}}(k,n) & \ell = 0 \\ B_{\text{rev}}(k,n,\ell) e^{-\alpha_{60}(k,n)\ell N_{\text{hop}}} & \ell \geq 1, \end{cases} \quad (2.8)$$

where $B_{\text{dir}}(k,n)$ and $B_{\text{rev}}(k,n,\ell)$ are zero-mean mutually independent and identically distributed complex Gaussian random variables, and the decay rate $\alpha_{60}(k,n)$ is linked to the frequency dependent reverberation time $T_{60}(k,n)$ by [4, 41, 75]

$$\alpha_{60}(k,n) = \frac{3 \ln(10)}{T_{60}(k,n)f_s}. \quad (2.9)$$

As a consequence, the power spectral density (PSD) of the room coefficients $\phi_h(k,\ell) = \mathcal{E}\{|H(\ell,k)|^2\}$, also known as the spectral envelope, is given by

$$\phi_h(k,n,\ell) = \begin{cases} \phi_{\text{dir}}(k,n) & \ell = 0 \\ \phi_{\text{rev}}(k,n,\ell) e^{-2\alpha_{60}(k,n)\ell N_{\text{hop}}} & \ell \geq 1, \end{cases} \quad (2.10)$$

where $\phi_{\text{dir}}(k,n) = \mathcal{E}\{|B_{\text{dir}}(k,n)|^2\}$ and $\phi_{\text{rev}}(k,n) = \mathcal{E}\{|B_{\text{rev}}(k,n)|^2\}$ are the PSDs of the room coefficient components.

In Figure 2.4, the energy of a RIR measured in an acoustic lab with $T_{60} = 630$ ms at 4 m distance is shown for three different $1/3$-octave bands. First, the RIR was transformed into the STFT domain using Hann windows of 32 ms length and a frame shift of 16 ms, and then the energy per $1/3$-octave band was computed. We can observe that the energy decay roughly follows an exponential function, i.e. a linear decay in the logarithmic domain. However, the model is not exact, and is problematic especially at the first few frames that are close to the direct sound.

Note that unlike in [75], we assume a possibly time-varying RIR $H_\ell(k,n)$, which implies that also the PSDs of the direct sound and the reverberation, $\phi_{\text{dir}}(k,n)$ and $\phi_{\text{rev}}(k,n)$ respectively, and the reverberation time $T_{60}(k,n)$ are time-varying.
2.6 Spatial coherence based model

In Section 2.6.1, some prerequisites to describe spatial sound fields, i.e., sound fields sampled by microphones at different locations, are introduced, before the spatial coherence model is introduced in Section 2.6.2.

2.6.1 Spatial description sound fields

The spatial coherence between two microphone signals $Y_m(k, n)$ and $Y_{m'}(k, n)$ is the normalized cross-PSD, or the relative transfer function (RTF) between the microphone pair, and is given by \[ \Gamma_{m,m'}(k) = \frac{\mathcal{E}\{Y_m(k, n)Y_{m'}^*(k, n)\}}{\sqrt{\mathcal{E}\{|Y_m(k, n)|^2\}\mathcal{E}\{|Y_{m'}(k, n)|^2\}}}. \] \hspace{1cm} (2.11)

If the spatial sound field is a plane wave as shown in Figure 2.5, the sensors are omnidirectional, and the propagation between the microphones is lossless, the spatial coherence or RTF between the microphones is given by \[ \Gamma_{pw}^{m,m'}(k) = e^{j\lambda(k)d_{m,m'}\cos\theta_i(k)}, \] \hspace{1cm} (2.12)

where $\theta_i(k, n)$ is the direction-of-arrival (DOA) of the $i$-th plane wave, $d_{m,m'}$ is the distance between the microphones, and $\lambda(k) = 2\pi f_s K c$ is the spatial frequency with $K$, $f_s$ and $c$ being the DFT length, the sampling frequency, and the speed of sound, respectively. We can see that for a plane wave, the spatial coherence mainly depends on the DOA and the microphone distance. Note that for a plane wave, we have always $|\Gamma_{pw}^{m,m'}(k)| = 1$.

If the sound field is spatially uncorrelated, which is the case e.g. for microphone self-noise, the spatial coherence is zero.

If a sound field is spatially homogeneous and isotropic, i.e., diffuse, there exist several analytic expressions to describe the coherence. In a spherically diffuse sound field, an infinite number of plane waves arrives equally distributed from all directions [78]. The spherical diffuse field coherence for omnidirectional microphones is given by [76]

\[ \Gamma_{m,m'}^{diff}(k) = \text{sinc} \left( \lambda(k)d_{m,m'} \right), \] \hspace{1cm} (2.13)
where sinc(·) = \frac{\sin(·)}{·}. In this case, the generally complex spatial coherence is real-valued. Note that there also exist other diffuse sound field expressions, i.e. the cylindrical diffuse field, which can be described analytically by a Bessel function [76]. However, for directional microphones [79] or in hearing aid setups [80], the spatial diffuse coherence \( \Gamma_{\text{diff}}^{m,m'}(k) \) is more complex to describe.

### 2.6.2 Signal model

The model described in the following is based on the spatial coherence and is therefore only suitable for multichannel setups with \( M > 1 \). Per time-frequency bin, the multichannel reverberant signal component \( x(k,n) \) is expressed as the sum of \( I \) desired sound components \( S_{d,i}(k,n) \), \( i \in \{1, \ldots, I\} \) and the late reverberation \( r(k,n) = [R^{(1)}(k,n), \ldots, R^{(M)}(k,n)]^T \). The reverberant signal vector is given by

\[
x(k,n) = \sum_{i=1}^{I} a_i(k,n) S_{d,i}(k,n) + r(k,n)
\]

where \( a_i(k,n) = [A_{i,1}(k,n), \ldots, A_{i,M}(k,n)]^T \) are the RTFs from the \( i \)-th desired sound source at the reference microphone \( S_{d,i}(k,n) \) to all \( M \) microphones, the matrix \( A(k,n) = [a_1(k,n), \ldots, a_I(k,n)] \) with dimensions \( M \times I \) contains the \( I \) RTF vectors in its columns, and the vector \( s_d(k,n) = [S_{d,1}(k,n), \ldots, S_{d,I}(k,n)]^T \) contains the desired sound components.

In many cases, the desired sound components \( S_{d,i}(k,n) \) correspond to the direct sound at the reference microphone under the farfield assumption, i.e. they are modeled as plane waves [81]. However, \( S_{d,i}(k,n) \) can also correspond to the direct sound with some early reflections [82] at the reference microphone. Therefore, we generally refer to \( S_{d,i}(k,n) \) as the desired sound or target signals [83, 84]. If the desired sound components are modeled as plane waves, the RTFs of the desired sound for omnidirectional microphones are given by

\[
A_{i,m}(k,n) = e^{j\lambda(k)d_m \cos \theta_i(k,n)},
\]

where \( \theta_i(k,n) \) is the DOA of the \( i \)-th plane wave, \( d_m = d_{1,m} - d_{1,\text{ref}} \) is the signed distance between the \( m \)-th microphone and the reference microphone, and \( \lambda(k) = \frac{2\pi k f_s}{c} \) is the spatial frequency. Note that the element of the RTF vector \( a_i(k,n) \) corresponding to the reference microphone is one, and the reference microphone can be chosen arbitrarily.

We assume that the desired sound components \( S_{d,i}(k,n) \) and the reverberation \( r(k,n) \) are zero-mean Gaussian random variables and are mutually uncorrelated. This assumption holds if the STFT frame length is much shorter than the reverberation time, such that the main part of the reverberation tail originating from the direct sound in frame \( n \) ends up in later frames. Consequently, the PSD matrix of the reverberant signal vector \( \Phi_x(k,n) = \mathcal{E}\{x(k,n)x^H(k,n)\} \) is given by

\[
\Phi_x(k,n) = A(k,n)\Phi_d(k,n)A(k,n) + \Phi_r(k,n),
\]

where \( \Phi_d(k,n) = \mathcal{E}\{S_{d,1}(k,n)S_{d,1}^H(k,n)\} \) and \( \Phi_r(k,n) = \mathcal{E}\{r(k,n)r^H(k,n)\} \).
2.6. SPATIAL COHERENCE BASED MODEL

where the PSD matrices of the plane waves and the reverberation are given by

$$\Phi_d(k,n) = E\{s_d(k,n)s_d^H(k,n)\} \quad \text{and} \quad \Phi_r(k,n) = E\{r(k,n)r^H(k,n)\},$$

respectively.

The reverberation $r(k,n)$ is assumed to have a spatially homogenous and stationary spatial coherence described by the spatial coherence matrix $\Gamma_r(k)$. Therefore, the reverberation PSD matrix can be expressed as a scaled coherence matrix by

$$\Phi_r(k,n) = \phi_r(k,n)\Gamma_r(k), \quad (2.17)$$

where $\phi_r(k,n)$ is the time-varying late reverberation PSD. Furthermore, a widely used additional assumption is that the reverberant sound field is isotropic, which means that the reverberation $r(k,n)$ can be described by a diffuse noise field with the spatial coherence function given in (2.13). Note that under the diffuse field assumption, the time-invariant coherence matrix $\Gamma_r(k)$ is determined only by the microphone array configuration, but is independent of the signal.

2.6.3 Validity of the spherical diffuse model for reverberation

In [85] it was found that the spherical diffuse field assumption is valid for the late reverberation tail and frequencies above the Schroeder frequency given by

$$f_{\text{Schroeder}} = 2000 \sqrt{\frac{T_{60}}{V_{\text{room}}}} \text{ Hz}, \quad (2.18)$$

where $V_{\text{room}}$ is the volume of the room in $m^3$ and the $T_{60}$ is given in s.

To show how well the spherical diffuse field assumption (2.13) for late reverberation holds, we convolved a 10 s long white noise signal with the late part of RIRs measured with omnidirectional microphones in a room with $T_{60} = 800$ ms to generate the late reverberation signals $R^{(m)}(k,n)$ artificially. The early part of the RIRs until 50 ms after the direct sound peak was set to zero. Figure 2.6 shows as grey solid line the real part of the spatial coherence between two microphone pairs with spacings of 4.5 cm and 9 cm, computed from the late reverberation using (2.11), where the expectation operators were approximated by long-term temporal averaging. The black dashed line shows the theoretical diffuse coherence function given by (2.13), which
depends on the microphone spacing. We can observe that the real part of the late reverberation coherence for this room on average fits the spherical diffuse field coherence.

Although in most considered dereverberation methods based on this model, the coherence matrix \( \mathbf{\Gamma}_r(k) \) is assumed to be given by (2.13), the method in [86] also allows to estimate this matrix from the data, which could be advantageous when the reverberant sound field differs from a theoretical diffuse field.

### 2.7 Discussion of the reverberation models

All presented STFT domain reverberation models have in common that each frequency bin is modeled independently without any interaction between neighboring frequency bins. In the following, relations between some models are pointed out, and different assumptions or underlying models for room coefficients are discussed.

#### 2.7.1 Relations between the models

Under some assumptions, we show that there exists a relation between the moving average model (2.6) and the autoregressive model (2.7) by invoking the multiple-input/output inverse theorem (MINT) [21]. Let us first assume a stationary acoustic scenario, such that the MA and MAR coefficients, \( \mathbf{h}_\ell(k) \) and \( \mathbf{C}_\ell(k) \) respectively, are independent of the time frame index \( n \). Remember that at each frequency \( k \), the MA model assumes \( M \) causal filters of finite order \( L \) with the non-zero coefficients \( \mathbf{h}_\ell(k) \), \( \ell = \{0, 1, \ldots, L\} \), and \( \mathbf{h}_\ell(k) = 0 \) for \( \ell < 0 \) \& \( \ell > L \). Furthermore, we assume only a single source.

Let us first introduce the MIMO filter coefficients \( \tilde{\mathbf{C}}_\ell(k) \) that contain \( \mathbf{C}_\ell(k) \) as

\[
\tilde{\mathbf{C}}_\ell(k) = \begin{cases} 
\mathbf{I}_{M \times M}, & \text{for } \ell = 0 \\
-\mathbf{C}_\ell(k), & \text{for } D \leq \ell \leq L \\
0_{M \times M} & \text{elsewhere.}
\end{cases}
\]  

(2.19)

where \( \mathbf{I}_{M \times M} \) is the identity matrix. If we now re-arrange the MAR model (2.7) to solve for \( \mathbf{s}(k, n) \) and use (2.19), we obtain

\[
\mathbf{s}(k, n) = \mathbf{x}(k, n) - \sum_{\ell=0}^{L} \mathbf{C}_\ell(k) \mathbf{x}(k, n-\ell) \\
= \sum_{\ell=0}^{L} \tilde{\mathbf{C}}_\ell(k) \mathbf{x}(k, n-\ell).
\]  

(2.20)

Under the condition that the \( M \) MA filters \( \mathbf{h}_\ell(k) \) do not share common zeros in the \( z \)-plane, and according to the MINT [21] and the multi-step linear prediction theory [37, 87], there exist filter coefficients \( \tilde{\mathbf{C}}_\ell(k) \) fulfilling

\[
\sum_{\ell=0}^{D-1} \mathbf{h}_\ell(k) = \sum_{\ell'=0}^{L} \tilde{\mathbf{C}}_{\ell'}(k) \mathbf{h}_{\ell'-\ell}(k).
\]  

(2.21)
2.7. DISCUSSION OF THE REVERBERATION MODELS

Table 2.1: Methods based on the MA model using either deterministic or statistical modeling of the filter coefficients.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model for $h_\ell(k,n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwartz et. al [29]</td>
<td>deterministic</td>
</tr>
<tr>
<td>Braun et. al [88], Section 4.3</td>
<td>first-order Markov model</td>
</tr>
<tr>
<td>Habets et. al [75, 89], Section 4.2.1</td>
<td>stochastic model with deterministic decay shape</td>
</tr>
</tbody>
</table>

From (2.21) we can see that the MAR coefficients $C_\ell(k)$ can be part of an equalizer to the MA filters $h_\ell(k)$ removing the late reverberation after $D$ frames. If we now insert the MA model (2.6) for $x(k,n)$ into the MIMO MAR filtering process given by (2.20), the early speech signal vector $s(k,n)$ is given by

$$s(k,n) = \sum_{\ell=0}^{D-1} h_\ell(k) S(k, n-\ell). \quad (2.22)$$

This means that under the assumed conditions, the early speech signal vector $s(k,n)$ in the MAR model is equal to the first $D$ summands of the MA model given by (2.6).

Unlike all other considered models, the spatial coherence based model does not exploit temporal information across multiple time frames. This results in two advantages of spatial coherence model: i) The model can be generally used to model other possibly non-stationary sound components such as babble noise, which have a certain time-invariant spatial coherence, irrespectively of their temporal structure across multiple time frames. ii) Given instantaneous parameter estimators, the parameters of the spatial coherence model can be computed instantaneously in each frame, which allows a fast adaption to changes in the acoustic scenario.

In contrast to the spatial coherence model, temporal correlation based models such as the narrowband MA and MAR models can discriminate between reverberation and other temporally uncorrelated sound components.

It is interesting to note that the exponential decay model presented in Section 2.5 can be seen as a special case of the narrowband MA model (2.5). The exponential decay model makes a deterministic assumption about the temporal envelope of the filter coefficients $H_\ell(k,n)$ across $\ell$, whereas the general narrowband MA model makes no such assumptions. While in the exponential decay model the coefficients $H_\ell(k,n)$ are modeled as strictly uncorrelated random noise, the Markov model can model possible correlations of $H_\ell(k,n)$ between the coefficients itself (across $\ell$) and additionally the evolution across time $n$.

2.7.2 Underlying models for MA and MAR filter coefficients

The filter coefficients of the narrowband MA model presented in Section 2.3, which includes also the exponential decay model in Section 2.5 as a special case, can be modeled as deterministic or as stochastic variables. Table 2.1 provides an overview of the different models for the coefficients as used in different dereverberation methods. In the multichannel method proposed in [29], the MA filter coefficients are estimated in the M-step of an expectation-maximization (EM) algorithm as a deterministic variable. In the single-channel method proposed in Section 4.3, the MA filter coefficients are modeled stochastically by a first-order Markov model, which explicitly
models gradually time-varying filter coefficients. In the methods proposed in [75, 89] using the model presented in Section 2.5, the MA filter coefficients are modeled as zero-mean Gaussian random variables, wherein their variance is assumed to follow a deterministic exponential decay. The decay envelope depends deterministically on the reverberation time $T_{60}(k,n)$ and on the PSDs $\phi_{\text{dir}}(k,n)$ and $\phi_{\text{rev}}(k,n)$. Therefore, this model can be considered as a stochastic model using deterministic side information.

Methods based on the narrowband MAR model presented in Section 2.4 such as [33, 90] and following publications [34, 35] were initially proposed by assuming the MAR coefficients as deterministic and time-invariant. In following publications [39, 91–93], recursive online solutions based on the same deterministic and time-invariant model were proposed, which might be able to track slowly time-varying MAR coefficients to some extent in practice. However, methods based on such deterministic models can suffer from stability issues in quickly time-varying acoustic conditions as shown in [39, 94]. In Chapter 6 we propose a stochastic model for the MAR coefficients, in particular a first-order Markov model, to explicitly account for gradually time-varying acoustic environments. An overview of the MAR coefficient models is given in Table 2.2.

### 2.8 Conclusion

An overview of several substantially different reverberation signal models in the STFT domain was presented. It has to be kept in mind that these models were developed in the context of speech dereverberation and are based on certain assumptions. One common assumption is the independence of the frequency bins. The dereverberation methods presented in the remainder of this thesis are based on these signal models, but also include additive noise. For convenience and clarity, at the beginning of each chapter, the currently considered signal model will be shortly reviewed.
CHAPTER 3

Spectral and spatial reverberation suppression

The most popular approaches used in speech enhancement are spectral and spatial suppression. Therein, the undesired sound component, in our case the sum of reverberation and noise, is modeled as an additive stochastic variable independent of the desired sound component, which is in our case the direct sound or the early speech. In the short-time Fourier transform (STFT) domain with typically short window lengths compared to the effective room impulse response (RIR) length, the late reverberation tail ends up in different time frames than the desired sound, which justifies the assumption of independence between the desired and undesired sound components in the same frame.

Multichannel spatio-spectral suppression techniques such as the multichannel Wiener filter (MWF) contain single-channel spectral suppression methods such as the Wiener filter (WF) as special cases by considering only one microphone. However, parameter estimators that rely on spatial information cannot be used in the single-channel case. As in this thesis, we also define the desired signal differently in the single- and multichannel cases, we treat both as separate cases. In Section 3.1 a standard single-channel spectral suppression method, namely the WF is presented, and in Section 3.2 the MWF is derived. Methods to estimate the most important parameter for dereverberation required to compute the WF and MWF, i.e. the late reverberation power spectral density (PSD), is discussed extensively in Chapters 4 and 5. In Section 3.3, we propose a generalized method to control the residual reverberation and noise at the output of a MWF without making any assumptions on the rank of the PSD matrices.

3.1 Single-channel dereverberation using spectral enhancement

Spectral suppression techniques popular for noise reduction can be used to jointly suppress reverberation and noise [1]. Commonly used filters for this task are the spectral WF [96] or similar methods, e.g. the short-time spectral amplitude (STSA) [97] and log-spectral amplitude (STSA) [98] estimators. In this thesis, we restrict ourselves to the WF, although the STSA and LSA estimators could be used equivalently. In Section 3.1.1 the single-channel signal model is presented, and in Section 3.1.2 the WF for suppression of reverberation and noise is derived.
3.1.1 Signal model

In single-channel spectral enhancement methods, the observed signal is described as the sum of uncorrelated signal components \([99]\). As typically some early reflections arrive within the same time frame as the direct sound, it is common to define the early speech signal as the desired signal, and the late reflections that arrive in other time frames as the undesired sound component. In this way, the assumption of independent sound components can be justified. The reverberant signal is split into early reflections and late reverberation, where the parameter \(D \geq 1\) defines the start time of the late reverberation in frames. The microphone signal is then described by

\[
Y(k,n) = S_D(k,n) + R_D(k,n) + V(k,n),
\]

where \(S_D(k,n)\) is the desired signal containing the direct sound and early reflections, \(R_D(k,n)\) is the late reverberation, \(X(k,n)\) is the reverberant speech signal, and \(V(k,n)\) is additive noise. Note that the separation time between early reflections and late reverberation \(D\) is defined in accordance with the narrowband moving average (MA) model described in Section 2.3. The parameter \(D\) is usually user defined and is indicated by the subscript of \(S_D(k,n)\) and \(R_D(k,n)\).

By assuming all three sound components in (3.1) as mutually uncorrelated zero-mean Gaussian random variables, the PSD of the microphone signal \(\varphi_Y(k,n) = \mathcal{E}\{|Y(k,n)|^2\}\) is given by

\[
\varphi_Y(k,n) = \varphi_{S_D}(k,n) + \varphi_{R_D}(k,n) + \varphi_V(k,n),
\]

where \(\varphi_{S_D}(k,n) = \mathcal{E}\{|S_D(k,n)|^2\}\), \(\varphi_{R_D}(k,n) = \mathcal{E}\{|R_D(k,n)|^2\}\) and \(\varphi_V(k,n) = \mathcal{E}\{|V(k,n)|^2\}\) are the PSDs of the individual signal components.

3.1.2 MMSE estimation of early speech signal

The desired signal \(S_D(k,n)\) can be estimated in the minimum mean square error (MMSE) sense using the filter \(G(k,n)\) by the minimization problem

\[
G_{WF}(k,n) = \arg \min_G \mathcal{E}\{|S_D(k,n) - GY(k,n)|^2\}
\]

and its estimate is obtained by

\[
\hat{S}_D(k,n) = G_{WF}(k,n)Y(k,n).
\]

The resulting Wiener filter \([96,99]\) is then given by

\[
G_{WF}(k,n) = \frac{\varphi_{S_D}(k,n)}{\varphi_{S_D}(k,n) + \varphi_{R_D}(k,n) + \varphi_V(k,n)} = \frac{\xi_D(k,n)}{\xi_D(k,n) + 1},
\]
where $\xi_D(k,n) = \frac{\phi_{SD}(k,n)}{\phi_{RD}(k,n) + \phi_V(k,n)}$ is the a priori signal-to-interference ratio (SIR). A popular method to estimate the a priori SIR $\xi_D(k,n)$ is the decision-directed approach [97], which is a weighting between the maximum likelihood estimate using the a posteriori SIR $\gamma_D(k,n) = \frac{|Y(k,n)|^2}{\phi_{RD}(k,n) + \phi_V(k,n)}$ and the previously estimated instantaneous a priori SIR $\tilde{\xi}_D(k,n) = \frac{|\hat{S}_{D}(k,n)|^2}{\phi_{RD}(k,n) + \phi_V(k,n)}$ with (3.4). The decision-directed estimate of the a priori SIR is given by [97]

$$\hat{\xi}_D(k,n) = \beta_{dd} G_{WF}^2(k,n-1) \gamma_D(k,n-1) + (1 - \beta_{dd}) \max\{\gamma_D(k,n) - 1, 0\},$$

(3.7)

where $\beta_{dd}$ is the decision-directed weighting parameter.

Assuming knowledge of the noise PSD $\phi_V(k,n)$, which can be estimated during periods of speech absence [100, 101], the remaining challenge is to estimate the late reverberation PSD $\phi_R(k,n)$. Methods to estimate $\phi_R(k,n)$ are discussed in Chapter 4.

### 3.2 Multichannel dereverberation using spatial filtering

In this section, we derive the widely used MWF as a spatial filter for dereverberation. To derive this filter, it is important to define the desired signal according to the application. The frequency-domain MWF [46, 102] is typically derived using the spatial coherence reverberation model presented in Section 2.6. Within this thesis, we restrict ourselves to the MWF, while there exists a variety of other spatial filters such as the linearly constrained minimum variance (LCMV) [103], or combinations of spatial filters with different spectral postfilters [104–106], which would result in a lot of possibilities.

Furthermore, spatial filters can be derived in different domains, e.g. in the spherical harmonic domain by using spherical microphone arrays [15]. For applications such as binaural hearing aids, a spatial filter yielding multiple output signals can be designed, while the spatial coherence of the interference has to be preserved [13]. However, within this thesis, we focus on the general form of the MWF in the spatial domain. After presenting the signal model in Section 3.2.1, the MWF is derived in Section 3.2.2, and in Section 3.2.3 we discuss how the parameters for the MWF can be estimated except for the late reverberation PSD that will be discussed in Chapter 5.

#### 3.2.1 Signal model

To derive a spatial filter, we use the spatial coherence model as presented in Section 2.6. The microphone signals containing the reverberant signal and additive noise are therefore written in the vector $\mathbf{y}(k,n) = [Y_1(k,n), \ldots, Y_M(k,n)]^T$ and are given by

$$\mathbf{y}(k,n) = \mathbf{A}(k,n)\mathbf{s}_d(k,n) + \mathbf{r}(k,n) + \mathbf{v}(k,n),$$

(3.8)

where the vector $\mathbf{s}_d(k,n) = [S_{d,1}(k,n), \ldots, S_{d,I}(k,n)]^T$ contains the $I$ desired signal components at the reference microphone, the matrix $\mathbf{A}(k,n) = [\mathbf{a}_1(k,n), \ldots, \mathbf{a}_I(k,n)]$ contains the relative
transfer functions (RTFs) of the desired sound, the vector \( \mathbf{x}_d(k,n) \) denotes the desired sound at the microphones, the vector \( \mathbf{r}(k,n) \) denotes the reverberation, and the vector \( \mathbf{v}(k,n) \) denotes additive noise.

By assuming all components in (3.8) mutually uncorrelated, and using the spatial coherence model for the reverberation component \( \mathbf{r}(k,n) \) given by (2.17), the PSD matrix of the microphone signals \( \Phi_y(k,n) = \mathcal{E}\{\mathbf{y}(k,n)\mathbf{y}^H(k,n)\} \) is given by

\[
\Phi_y(k,n) = \mathbf{A}(k,n)\Phi_d(k,n)\mathbf{A}^H(k,n) + \phi_r(k,n)\Gamma_r(k) + \Phi_v(k,n),
\]

(3.9)

where \( \Phi_d(k,n) = \mathcal{E}\{\mathbf{s}_d(k,n)\mathbf{s}_d^H(k,n)\} \) and \( \Phi_v(k,n) = \mathcal{E}\{\mathbf{v}(k,n)\mathbf{v}^H(k,n)\} \) are the PSD matrices of the desired sound and the noise, respectively, which are defined similarly as \( \Phi_y(k,n) \), \( \phi_r(k,n) \) is the reverberation PSD and \( \Gamma_r(k) \) is the spatial coherence of the reverberation, which is usually assumed to be the spherical diffuse field coherence given in (2.13). Note that \( \Phi_d(k,n) \) is a diagonal matrix if the desired sound components are mutually uncorrelated. Further, we introduce the interference PSD matrix \( \Phi_{in}(k,n) = \Phi_r(k,n) + \Phi_v(k,n) \) as the sum of reverberation and noise.

If explicitly only a single desired sound component per time and frequency is assumed \( (I = 1) \), the index \( i \) is omitted and the PSD matrix model (3.9) becomes

\[
\Phi_y(k,n) \bigg|_{I=1} = \phi_d(k,n)\mathbf{a}(k,n)\mathbf{a}^H(k,n) + \phi_r(k,n)\Gamma_r(k) + \Phi_v(k,n),
\]

(3.10)

where \( \phi_d(k,n) = \mathcal{E}\{|\mathbf{s}_d(k,n)|^2\} \) and the index \( i \) is dropped for \( I = 1 \), i.e. \( \mathbf{s}_d(k,n) \equiv \mathbf{s}_{d,1}(k,n) \).

### 3.2.2 MMSE estimation of desired signal

In the case of multiple desired signal components, the desired target signal is defined as their sum, i.e.,

\[
\mathbf{S}_d(k,n) = \sum_{i=1}^{I} \mathbf{s}_{d,i}(k,n) = \mathbf{1}_{I\times1}^T\mathbf{s}_d(k,n),
\]

(3.11)

where \( \mathbf{1}_{I\times1} = [1,1,\ldots,1]^T \) is a vector of ones with length \( I \).

To estimate the target signal given by (3.11), a spatial filter is applied to the microphone signals such that

\[
\hat{\mathbf{S}}_d(k,n) = \mathbf{g}^H(k,n)\mathbf{y}(k,n).
\]

(3.12)

An estimated target signal \( \hat{\mathbf{S}}_d(k,n) \) can be obtained using the MWF proposed in [46]. The filter minimizes the reverberation and noise while extracting the target signal in the MMSE sense, i.e.,

\[
\mathbf{g}_{\text{MWF}}(k,n) = \arg\min_{\mathbf{g}} \mathcal{E}\{\|\mathbf{1}_{I\times1}^T\mathbf{s}_d(k,n) - \mathbf{g}^H\mathbf{y}(k,n)\|^2\}.
\]

(3.13)
3.2. MULTICHANNEL DEREVERBERATION USING SPATIAL FILTERING

The solution is the general or multi-wave MWF, and is given by

\[
g_{\text{MWF}}(k,n) = \left[ A(k,n)\Phi_d(k,n)A_H(k,n) + \Phi_{\text{in}}(k,n) \right]^{-1} A(k,n)\Phi_d(k,n) 1_{I \times 1}. \tag{3.14}
\]

Note that in the case of a single plane wave \( I = 1 \), the matrix \( A(k,n) \) reduces to the column vector \( a(k,n) \). By using Woodbury’s matrix identity [107, Section 2.7.3], the MWF for \( I = 1 \) can be split into a purely spatial filter, the minimum variance distortionless response (MVDR) beamformer \( g_{\text{MVDR}}(k,n) \) [45], and a purely spectral filter, the Wiener postfilter \( G_{\text{WPF}}(k,n) \), i.e.

\[
g_{\text{MWF}}|_{I=1} = \left[ aa^H\phi_d + \Phi_{\text{in}} \right]^{-1} a\phi_d \tag{3.15a}
\]

\[
= \begin{pmatrix}
\Phi_{\text{in}}^{-1} a \\
\Phi_{\text{in}}^{-1} a
\end{pmatrix}
\begin{pmatrix}
\Phi_d \\
\phi_d
\end{pmatrix}^{-1} 
\begin{pmatrix}
G_{\text{MVDR}} \\
G_{\text{WPF}}
\end{pmatrix} \tag{3.15b}
\]

where the frequency and time indices are omitted for better readability. This form is also referred to as single-wave MWF, and is also known as rank-one MWF, since for \( I = 1 \) the desired microphone signal PSD matrix \( A\Phi_dA_H \) is of rank-one.

Advantages of the form (3.15b) are that the spatial filter \( g_{\text{MVDR}}(k,n) \) is independent of the plane wave PSD \( \phi_d(k,n) \), and that the spectral Wiener postfilter \( G_{\text{WPF}}(k,n) \) can be estimated using established techniques from single-channel spectral filtering methods such as the decision-directed approach [97].

In the next section, we discuss how the parameters required to compute the MWF can be estimated.

3.2.3 Parameter estimation of the MWF

The computation of the general multi-wave MWF given by (3.14) differs from the computation of the single-wave MWF given by (3.15): In the general form (3.14), the PSD matrix of the desired sound has to be computed explicitly, whereas in the single-wave form (3.15), the MVDR does not require the desired sound PSD, and the Wiener postfilter can be computed without directly estimating the desired sound PSD \( \phi_d(k,n) \) by using the decision-directed approach. Both forms of the MWF require estimates of the RTFs and the interference PSD matrix \( \Phi_{\text{in}}(k,n) = \phi_i(k,n)\Gamma_i(k) + \Phi_v(k,n) \), i.e. the reverberation PSD \( \phi_i(k,n) \), the coherence matrix \( \Gamma_i(k) \), and the noise PSD matrix \( \Phi_v(k,n) \). The structures of the general multi-wave MWF and the rank-one MWF are illustrated in Figures 3.1 and 3.2, where the signal flow and the required parameter estimation blocks are shown. In Sections 3.2.3.1 - 3.2.3.4 the reader is referred to established approaches to estimate the required parameters, while the multichannel estimation of the late reverberation PSD is discussed extensively in Chapter 5.

3.2.3.1 Estimation of the RTFs

It is popular to model the desired sound components \( S_{d,i}(k,n) \) at the reference microphones as plane waves, as in this case the RTFs \( A(k,n) \), are given by complex phase factors depending only
on the narrowband direction-of-arrivals (DOAs) $\theta_i(k, n)$ as given in (2.15). In some applications, the DOA can be either assumed to be fixed or known, whereas in other cases there is no such information available, and the DOAs have to be estimated from the observed signals. Note that in our signal model, we generally assume independent DOAs per time-frequency bin.

In the case of a single DOA and $M = 2$ microphones, the DOA can be estimated directly from the phase difference between the two microphones. An extension to an arbitrary number of microphones $M$ has been proposed in [108]. Most state-of-the-art DOA estimation techniques exploit special array geometries such as the ones shown in Figure 3.3. State-of-the-art narrowband DOA estimators, e.g. root-MUSIC [109] or estimation of signal parameters via rotational invariance techniques (ESPRIT) [110], generally can estimate multiple simultaneous DOAs per time-frequency bin, and are often designed for uniform linear arrays (ULAs). But also for other common array shapes suitable DOA estimators have been developed, such as beamspace
3.2. MULTICHANNEL DEREVERBERATION USING SPATIAL FILTERING

![Figure 3.3: Typical microphone array geometries: a) uniform linear array, b) non-uniform linear array, c) uniform circular array](image)

root-MUSIC [111] for ULAs. For more information on DOA estimation, the reader is referred to [112,113].

When choosing the array geometry, one has to keep in mind that narrowband DOA estimators can only deliver useful results below the spatial aliasing frequency $f_{\text{alias}}$ that depends on the smallest microphone distance $d_{\text{min}}$ [114]. Spatial aliasing occurs, when the resolution of the spatial sampling of the sound field is too low. The spatial aliasing frequency is given by

$$f_{\text{alias}} = \frac{c}{2d_{\text{min}}}, \quad (3.16)$$

where $c$ is the speed of sound.

If the desired sound $S_d(k,n)$ models the early speech signal component, an estimate of the early speech RTFs is required. Suitable RTF estimators have been proposed for example in [82,115–118].

### 3.2.3.2 Estimation of the interference PSD matrix

By assuming the additive noise to be stationary or slowly time-varying, the noise PSD matrix can be estimated during speech absence using established approaches such as the speech presence probability (SPP) based method proposed in [101,119] or other methods proposed in [100,120]. In this thesis, we assume to have prior knowledge of the noise PSD matrix $\Phi_\nu(k,n)$.

The coherence matrix $\Gamma_r(k)$ can be computed from the microphone array geometry by using the diffuse spatial coherence model given in (2.13). The reverberation PSD $\phi_r(k,n)$ can be estimated in numerous ways using single- and multichannel approaches that are discussed and compared in detail in Section 5.

### 3.2.3.3 Desired signal PSD estimation for multiple desired signals

If the desired sound components $S_{d,i}(k,n)$ are independent, the PSD matrix

$$\Phi_d(k,n) = \text{diag} \{\phi_{d,1}(k,n), \ldots, \phi_{d,I}(k,n)\} \quad (3.17)$$

is a diagonal matrix with the real-valued PSDs $\phi_{d,i}(k,n) = \mathcal{E}\{|S_{d,i}(k,n)|^2\}$ on the main diagonal and zero elsewhere, where the operator $\text{diag} \{\cdot\}$ creates a diagonal matrix with its arguments on the main diagonal. Given an estimate of the interference PSD matrix $\Phi_\text{in}(k,n)$, the PSD
matrix of the desired signals at the microphones \( \Phi_{\text{xd}}(k,n) = A(k,n)\Phi_d(k,n)A^H(k,n) \) can be estimated by

\[
\hat{\Phi}_{\text{xd}}(k,n) = \hat{\Phi}_y(k,n) - \hat{\Phi}_{\text{in}}(k,n),
\]

(3.18)

where the microphone PSD matrix \( \Phi_y(k,n) \) can be estimated from the input signal vectors using first-order recursive smoothing by

\[
\hat{\Phi}_y(k,n) = \alpha \hat{\Phi}_y(k,n-1) + (1-\alpha)\mathbf{y}(k,n)\mathbf{y}^H(k,n),
\]

(3.19)

and \( \alpha \) is a smoothing factor close to 1. Note that it has to be ensured that the matrix \( \hat{\Phi}_{\text{xd}}(k,n) \) is positive semi-definite. A least-squares estimate of the desired PSDs is obtained by [81]

\[
\hat{\phi}_d,1, \ldots, \hat{\phi}_d,I = (A^HA)^{-1}A^H\text{vec}\left\{\hat{\Phi}_{\text{xd}}\right\},
\]

(3.20)

where the operator \( \text{vec}\{\cdot\} \) stacks the \( B \) columns of a \( A \times B \) matrix into a single column vector of size \( AB \times 1 \), and the stacked matrix \( A \) of size \( M^2 \times I \) is defined as \( A = [\text{vec}\{a_1a_1^H\}, \ldots, \text{vec}\{a_Ia_I^H\}] \).

### 3.2.3.4 Estimation of the Wiener postfilter for a single desired signal

By computing the MWF in the decomposed MVDR-postfilter form given in (3.15b) as shown in Figure 3.2, the MVDR filter \( g_{\text{MVDR}}(k,n) \) can be computed using only knowledge of the RTF vector \( a(k,n) \), which is constructed using knowledge of the DOA per frequency, and the interference matrix \( \Phi_{\text{in}}(k,n) \). In analogy to (3.6), we can reformulate the Wiener postfilter \( G_{\text{WPF}}(k,n) \) in (3.15b) in terms of the a priori SIR at the output of the MVDR beamformer \( \xi_{\text{MVDR}}(k,n) = \phi_{\text{in},\text{MVDR}}(k,n) \), where \( \phi_{\text{in},\text{MVDR}} = (a^H\Phi_{\text{in}}^{-1}a)^{-1} = g_{\text{MVDR}}^H\Phi_{\text{in}}g_{\text{MVDR}} \) is the residual interference at the MVDR output. The Wiener postfilter is then computed by

\[
G_{\text{WPF}}(k,n) = \frac{\xi_{\text{MVDR}}(k,n)\xi_{\text{MVDR}}(k,n)+1}{\xi_{\text{MVDR}}(k,n)},
\]

(3.21)

where the SIR \( \xi_{\text{MVDR}}(k,n) \) can be estimated using the decision-directed approach [97] in analogy to (3.7). The decision-directed a priori SIR at the MVDR output is then estimated by

\[
\hat{\xi}_{\text{MVDR}}(k,n) = \beta_{\text{dd}}G_{\text{WPF}}^2(k,n-1)\gamma_{\text{MVDR}}(k,n-1) + (1-\beta_{\text{dd}})\max\{\gamma_{\text{MVDR}}(k,n)-1, 0\}, \tag{3.22}
\]

where the a posteriori SIR at the MVDR output is given by

\[
\gamma_{\text{MVDR}}(k,n) = \frac{|g_{\text{MVDR}}^H(k,n)\mathbf{y}(k,n)|^2}{\phi_{\text{in},\text{MVDR}}(k,n)}.
\]

(3.23)

In contrast to the general multi-wave MWF shown in Figure 3.1, the decomposed single-wave MWF shown in Figure 3.2 requires no explicit estimation block for the desired signal PSD, which is implicitly estimated in the SIR estimation block for the postfilter.
3.3 Reduction control for the MWF

In practice estimation errors of the parameters required to compute the WF or MWF often lead to the fact that the undesired signal components cannot be suppressed below the perceivable threshold. Therefore, in some cases it is desirable to have control over the amount of residual reverberation and noise, which can also help to mitigate artifacts such as speech distortion and musical tones [121].

There exist some approaches for a time-domain Wiener filter introducing a parameter to control the residual noise [122,123]. In the class of single-channel spectral enhancement methods, [124] proposes a method to control the amount of residual noise, and in [125] a similar method is proposed for two interfering sound components in the context of joint noise reduction and echo cancellation. A multichannel method for partial noise reduction for hearing aids is proposed in [126]. Additional control over the trade-off between speech distortion and noise reduction is provided by the parametric multichannel Wiener filter (PMWF) [127]. In typical applications of the PMWF, the trade-off parameter is set to a fixed empirically determined value [102,128], or is heuristically controlled [129]. Existing solutions, however, either do not focus on directly controlling the amount of residual noise, or are only single-channel methods.

In [83] we proposed a method to control the amount of noise reduction for a parametric MWF. We present the method here only for the conventional MWF, but also show how to generalize the reduction control to multiple interferences, in this case reverberation and noise.

3.3.1 Definition of the target signal

Let us first define a new target signal \( Z_{MWF}(k,n) \), which contains the desired signal \( S_d(k,n) \) defined in (3.11) and attenuated reverberation and noise, as

\[
Z_{MWF}(k,n) = S_d(k,n) + \beta_r e_{ref}^T r(k,n) + \beta_v e_{ref}^T v(k,n),
\]

(3.24)

where \( e_{ref} \) is a \( M \times 1 \) vector containing one at the element of the reference microphone and zero elsewhere, and \( \beta_r \) and \( \beta_v \) are the reverberation and noise attenuation factors, respectively.

Our aim is to obtain an estimate \( \hat{Z}_{MWF}(k,n) \) of the target signal given by (3.24) using the spatial filter \( g_Z(k,n) \) as

\[
\hat{Z}_{MWF}(k,n) = g_Z^H(k,n)y(k,n).
\]

(3.25)

3.3.2 Derivation of the proposed filter

We employ the mean-squared error (MSE) criterion to derive the proposed filter, i.e.

\[
g_Z(k,n) = \arg \min_g \mathcal{E}\{|Z_{MWF}(k,n) - g^H(k,n)y(k,n)|^2\}.
\]

(3.26)

The solution to (3.26) is given by

\[
g_Z(k,n) = \Phi_y^{-1}(k,n) \left[ A(k,n) \Phi_d(k,n) A^H(k,n) + \beta_r \phi_r(k,n) \Gamma_r(k) + \beta_v \Phi_v(k,n) \right] e_{ref}.
\]

(3.27)
By reformulating the standard MWF (3.14) in a similar form as (3.27) to
\[
g_{MWF}(k,n) = \Phi_y^{-1}(k,n)A(k,n)\Phi_d(k,n)A_H(k,n)e_{ref},
\]
for the case \(\beta_r = \beta_v = \beta\), we can rewrite the proposed filter in terms of \(g_{MWF}(k,n)\) as
\[
g_Z(k,n) = (1 - \beta)g_{MWF}(k,n) + \beta e_{ref}.
\]
We can see that in this case, the output of the proposed filter \(g_Z(k,n)\) is a weighting between output of the standard MWF \(g_{MWF}(k,n)\) and the input signal at the reference microphone.

Note that in the single-channel case, we can equivalently compute the WF given by (3.6) with reduction control by
\[
G_Z(k,n) = \max\{G(k,n), \beta_{lim}\},
\]
where \(\beta_{lim}\) is the lower bound. Despite being simple to compute, the lower bounded Wiener postfilter (3.33) has three disadvantages compared to the proposed controlled MWF (3.27):

1. The lower bound (3.33) can only applied using the rank-one assumption, i.e. for \(I = 1\), but is not applicable for \(I > 1\).

2. In the multichannel case, the MVDR provides already some amount of reduction. Therefore, the lower bound \(\beta_{lim}\) in (3.33) applied to the Wiener postfilter \(G_{WF}(k,n)\) in (3.15b) is not directly related to the amount of desired attenuation as \(\beta\), since the attenuation of the MVDR also has to be taken into account.

3. Noise and reverberation cannot be controlled separately using (3.33), and only the postfilter can be controlled but not the MVDR spatial filter. In contrast, using the proposed method, even for the general MWF without rank-one assumption direct control is achieved.
3.3. REDUCTION CONTROL FOR THE MWF

Figure 3.4: Wiener gain for controlled Wiener filter and limited Wiener filter for $\beta = \beta_{\text{lim}}$.

For the single-channel case $M = 1$ we analyze the reduction of the controlled Wiener filter (3.32) and for the Wiener filter limited as in (3.33). The Wiener gains for both filters are shown in Figure 3.4. We can see that the controlled Wiener filter is a soft limiting function of the a priori SIR $\xi(k, n)$, whereas the lower bounded Wiener filter is a hard limiting function. An advantage of the hard limiting of $G_{\text{lim}}(k, n)$ is the slightly higher robustness against musical tones in low SIR regions, where the gain function is limited by the lower bound $\beta_{\text{lim}}$, and the gain $G_{\text{lim}}(k, n)$ is constant. However, an optimal estimate of the target signal with controlled residual noise and reverberation given in (3.24) is obtained using the controlled WF.

3.3.4 Experimental validation

In this section, the concept of reduction control for a MWF is validated. We simulate a reverberant signal by convolving a 15 s anechoic speech signal with RIRs simulated using the image method [131]. We simulated a source 2 m in front of a linear array of $M = 4$ microphones with equidistant spacing of 2.5 cm in a shoebox room with $T_{60} = 0.5$ s. The sampling frequency was 16 kHz, and the STFT parameters were a square-root Hann window of 32 ms length, a frame shift of $T_{\text{hop}} = 16$ ms, and a discrete Fourier transform (DFT) length of $K = 1024$ samples.

The parameters required to compute the single-source MWF given by (3.27) were computed as close to ideal oracle estimates: the RTF vector $\mathbf{a}(k, n)$ was computed using knowledge of the source DOA and the array geometry using (2.15); the diffuse coherence matrix $\mathbf{\Gamma}_r(k)$ was computed by (2.13); the stationary noise PSD matrix $\Phi_{\nu}(k) = \mathcal{E}\{\mathbf{v}(k, n)\mathbf{v}^H(k, n)\}$ was estimated using the known noise signals and by approximating the expectation operator by an arithmetic average across all time frames. The time-varying PSDs of the reverberation and the desired signal, $\phi_r(k, n)$ and $\phi_d(k, n)$ respectively, were computed as oracle estimates from the known signals $S_d(k, n)$ and $r(k, n)$ by recursive averaging, i.e.

\[
\bar{\phi}_r(k, n) = \alpha \bar{\phi}_r(k, n - 1) + (1 - \alpha) \frac{1}{M} \mathbf{r}^H(k, n)\mathbf{r}(k, n)
\]

\[
\bar{\phi}_d(k, n) = \alpha \bar{\phi}_d(k, n - 1) + (1 - \alpha) |S_d(k, n)|^2,
\]

where $\alpha = e^{-T_{\text{hop}}/\tau}$ is an exponential recursive averaging factor with the time constant chosen here as $\tau = 50$ ms. The direct sound at the reference microphone $S_d(k, n)$ was obtained by convolving only the direct path of the simulated RIR with the speech signal, whereas the reverberation signals $\mathbf{r}(k, n)$ were obtained by convolving windowed RIRs with the speech signal containing only the late reverberation RIRs 48 ms after the direct sound.
The reverberation reduction (RR) and noise reduction (NR) at the output of the MWF with reduction control given by (3.25) were computed by

\[
RR(n) = \frac{\sum_k |g_H^R(k,n)r(k,n)|^2}{\sum_k |e_{ref}r(k,n)|^2}
\]

\[
NR(n) = \frac{\sum_k |g_H^V(k,n)v(k,n)|^2}{\sum_k |e_{ref}v(k,n)|^2}
\]

(3.36) (3.37)

In Figure 3.5 the RR and NR is shown for five different settings of the attenuation factors \(\beta_r\) and \(\beta_v\). For \(\beta_r = \beta_v = 0\) there is no reduction control and the output is equal to the conventional MWF given in (3.14). The range of the RR and NR values is limited to -20 dB for visualization purposes. We can observe that the RR and NR values reach the desired attenuation factors during speech absence. For the settings shown by the red dashed line, we can observe that no RR is achieved, while the NR still yields the defined value during speech absence. The opposite can be observed for the yellow dashed line, where no NR and 7 dB RR is achieved.

3.3.5 Conclusion

We proposed a method to independently control the residual reverberation and noise at the output of a general MWF. The user-defined attenuation of reverberation and noise is achieved during speech absence, while during speech presence an optimal tradeoff in the MMSE sense is achieved. It was shown experimentally that if close to optimal parameter estimates are available, the attenuation of the noise and reverberation components can be controlled independently.

3.4 Summary

In this chapter we derived single- and multichannel Wiener filters for joint reduction of reverberation and noise. These filters are dependent on several parameters that have to be estimated.
from the observed signals. In the single-channel case, the reverberation and noise PSDs are required, which are then used for the decision-directed estimation of the a priori SIR. In the multichannel case, we have to estimate the noise PSD matrix, the RTFs of the plane waves, and the reverberation PSD. Established estimators for all parameters except the reverberation PSD were reviewed in this chapter, while the reverberation PSD, which determines the dereverberation performance of these filters to a large extent, is discussed in the next two chapters for the single- and multichannel case. In addition, we proposed a novel method to control the attenuation of reverberation and noise independently for an MWF in any general form. In contrast to limiting the spectral gain of the Wiener postfilter of the decomposed MWF, the proposed approach does not rely on this decomposition.
CHAPTER 4

Single-channel late reverberation PSD estimation

As presented in Section 3.1, in the single-channel case the dereverberated signal can be estimated using a spectral enhancement technique such as the Wiener filter, which requires in addition to the noise power spectral density (PSD) an estimate of the late reverberation PSD. In this chapter, we present single-channel estimators for the late reverberation PSD. In Section 4.1, the signal model with additive noise is presented and the problem is formulated. In Section 4.2, state-of-the-art single-channel reverberation PSD estimators are reviewed. In Section 4.3, a novel single-channel reverberation PSD estimator is proposed. The estimators are evaluated in Section 4.4, and the chapter is concluded in Section 4.5.

4.1 Problem formulation

The single-channel Wiener filter (WF) for joint dereverberation and noise reduction as presented in Section 3.1 is based on the additive signal model (3.1) in the short-time Fourier transform (STFT) domain

\[ Y(k,n) = \underbrace{S_D(k,n) + R_D(k,n)}_{X(k,n)} + V(k,n), \]

where \( S_D(k,n) \) is the early speech signal, \( R_D(k,n) \) is the late reverberation, and \( V(k,n) \) is additive noise. Given knowledge of the PSDs of the late reverberation and noise, \( \phi_{R_D}(k,n) \) and \( \phi_V(k,n) \) respectively, the early speech signal \( S_D(k,n) \) can be estimated using a Wiener filter as described in Section 3.1.

Using the convolutive transfer function (CTF) approximation, i.e. the narrowband moving average reverberation model (2.5), all frequencies are assumed to be independent, such that the early speech and late reverberation signals can be described by possibly time-varying narrowband filter coefficients \( H_\ell(k,n) \) with \( \ell \in \{0, L_{MA}\} \) and the anechoic signal \( S(k,n) \), i.e.,

\[ S_D(k,n) = \sum_{\ell=0}^{D-1} H_\ell(k,n)S(k,n-\ell) \]

(4.1)

\[ R_D(k,n) = \sum_{\ell=D}^{L_{MA}} H_\ell(k,n)S(k,n-\ell), \]

(4.2)
where the number frames \( D \) denotes the separation between early and late reverberation with \( 1 \leq D < L_{\text{MA}} \).

All single-channel late reverberation PSD estimators presented in this chapter can be derived based on the CTF model. In Section 4.2.1 two state-of-the-art estimators assuming an exponential decaying envelope of the CTF coefficients \( H_t(k,n) \) are presented. In Section 4.2.2 a state-of-the-art method that estimates the CTF coefficients by exploiting inter-frame correlation between the estimated signal \( \hat{S}(k,n) \) and the reverberant signal \( \hat{X}(k,n) \), in order to obtain an the late reverberation PSD. In Section 4.3 we introduce a novel formulation using relative convolutive transfer functions (RCTFs) and propose to estimate the RCTF coefficients using a Kalman filter.

4.2 State-of-the-art single-channel methods

4.2.1 Using the exponential decay model

There exist different estimation structures based on the exponential decay model presented in Section 2.5. Due to their signal flow structure, they are called the forward [75] and backward [89] estimators. The forward estimator requires only the noise-free, reverberant signal PSD of past frames \( \phi_X(k,n) = \mathcal{E}\{|X(k,n)|^2\} \) and the direct-to-reverberation ratio, whereas the backward estimator requires the delayed enhanced signal PSD \( \phi_{S_D}(k,n) \).

4.2.1.1 Forward estimator

Given the exponential decay model for the room impulse response (RIR) given in (2.10), the late reverberation PSD can be estimated by [75]

\[
\hat{\phi}_{RD}(k,n) = [1 - \kappa(k)] e^{-2\alpha_{60}(k)N_{\text{hop}}D} \hat{\phi}_{RD}(k,n - D) + \kappa(k) e^{-2\alpha_{60}(k)N_{\text{hop}}D} \hat{\phi}_X(k,n - D),
\]

(4.3)

where the reverberation decay constant \( \alpha_{60}(k) \) is defined in (2.9), and \( \kappa(k,n) = \frac{\phi_{\text{rev}}(k,n)}{\phi_{\text{dir}}(k,n)} \). If the source-microphone distance is larger than the critical distance, where \( \phi_{\text{rev}}(k,n) \geq \phi_{\text{dir}}(k,n) \), it is often assumed that \( \kappa(k,n) \) is upper bounded at \( \kappa(k,n) = 1 \). As a consequence, the first term in (4.3) can be neglected and the estimator becomes equal to Lebart’s estimator [41]. The ratio \( \kappa(k,n) \) can also be computed depending on the direct-to-reverberation ratio (DRR) by [75]

\[
\kappa(k,n) = \frac{1 - e^{-2\alpha_{60}(k)N_{\text{hop}}D}}{\text{DRR}(k,n) e^{-2\alpha_{60}(k)N_{\text{hop}}D}},
\]

(4.4)

and the DRR is defined in (1.3). It is important to note that here the reverberation time \( T_{60}(k) \), and therefore also \( \alpha_{60}(k) \), are assumed to be stationary within the duration of the decay. Note that an estimate of the reverberant noise-free reverberant signal PSD \( \hat{\phi}_X(k,n) \) is required, which can e.g. be obtained by \( \hat{\phi}_X(k,n) = \max\{\phi_Y(k,n) - \phi_V(k,n), 0\} \).
4.2. STATE-OF-THE-ART SINGLE-CHANNEL METHODS

4.2.1.2 Backward estimator

The method proposed in [89] is only valid beyond the critical distance (farfield assumption), where \( \phi_{\text{rev}}(k, n) \geq \phi_{\text{dir}}(k, n) \), and is called the backward estimator: it is a recursive expression, that depends on the PSD of the previously estimated desired signal \( \phi_{SD}(k, n) \), and is given by

\[
\hat{\phi}_{RD}(k, n) = e^{-2\alpha_0(k)N_{\text{hop}}} \hat{\phi}_{RD}(k, n-1) + e^{-2\alpha_0(k)N_{\text{hop}}D} \hat{\phi}_{SD}(k, n-D),
\]  

(4.5)

where \( \hat{\phi}_{SD}(k, n) = |\hat{S}_D(n)|^2 \) is obtained using the Wiener filter (3.4).

The structures of the forward and backward estimators based on the exponential decay model are illustrated in Figure 4.1. Note that in the forward estimator, information is passed only forward, whereas the backward estimator has a recursive structure, where the estimated early speech signal is delayed and fed back to the late reverberation PSD estimator. An advantage of the backward estimator is that it does not require the additional step to estimate the noise-free reverberant PSD \( \phi_X(k, n) \). Insights in the performance of the forward and backward estimators are given in Section 4.4.

\[
Y(n) \quad \text{Wiener filter} \quad \hat{S}_D(n)
\]

(a) Forward estimator

\[
Y(n) \quad \text{Wiener filter} \quad \hat{S}_D(n)
\]

(b) Backward estimator

Figure 4.1: Exponential model based late reverberation PSD estimators.
4.2.2 Using a narrowband moving average model

A model based on CTFs taking inter-frame correlations into account was proposed in [132]. After estimating the noise-free reverberant signal \( X(k, n) \), the CTF filter coefficients are obtained for each \( \ell \in \{ D, D + 1, \ldots, L_{MA} \} \) by

\[
H_\ell(k, n) = \arg \min_H \mathcal{E}\{ |X(k, n) - HS(k, n - \ell)|^2 \}.
\]

(4.6)

Note that the cost-functions are optimized separately for each \( \ell \), such that there is no globally optimal set of coefficients for all \( \ell \) obtained. By approximating the desired anechoic signal by the enhanced signal (3.4), i.e. \( S(k, n) \approx \hat{S}_D(k, n) \), we obtain the solution

\[
\hat{H}_\ell(k, n) = \frac{\langle \hat{X}(k, n)\hat{S}_D^*(k, n - \ell) \rangle}{\langle |\hat{S}_D(k, n - \ell)|^2 \rangle},
\]

(4.7)

where \( \langle \cdot \rangle \) denotes a time averaging operator. Note that here also an estimate of the noise-free reverberant signal \( \hat{X}(k, n) \) is required, that can be obtained using a Wiener filter and knowledge of the noise PSD. A heuristic modification of (4.7) is proposed in [132] to improve the convergence speed and robustness. The late reverberation PSD is obtained as the smoothed power of the estimated late reverberation signal by skipping overlapping frames and summing only every \( P \)-th contribution, i.e.,

\[
\hat{R}_\text{erk} D(k, n) = \sqrt{B} \sum_{\ell=D}^{L_{MA}/P} \hat{H}_{\ell P}(k, n)\hat{S}_D(k, n - \ell P),
\]

(4.8)

where \( L_{MA} \) is an integer multiple of \( P \), \( \hat{H}_{\ell P}(k, n) \) is a modified version of (4.7) including a normalization for faster convergence, and \( B \) is a bias correction factor, both described detailed in [132]. The late reverberation PSD is finally obtained by

\[
\phi_{\text{erk} D}(k, n) = \langle |\hat{R}_\text{erk} D(k, n)|^2 \rangle.
\]

(4.9)

4.3 Late reverberation PSD estimation using relative convolutive transfer functions

In this section, a late reverberation PSD estimator based on a relative convolutive transfer function (RCTF) formulation is proposed [88]. In Section 4.3.1, the RCTFs are introduced and modeled by a first-order Markov model. In Section 4.3.2, a Kalman filter to estimate the RCTFs is proposed, which can then be used to estimate the late reverberation PSD.

4.3.1 Markov model for relative CTFs

Following the definition from (4.1) with \( D = 1 \), we have \( S_1(n) = H_0(n)S(n) \). The frequency index \( k \) is omitted in the following for brevity, wherever possible. To describe the reverberation
as a function of $S_1(n)$ which contains the direct-path and some early reflections, we introduce the RCTF with respect to $H_0(k,n)$ as

$$H_{\ell,0}(k,n) = \frac{H_\ell(k,n)}{H_0(k,n)}, \quad (4.10)$$

and define the RCTF vector without $H_{0,0}(k,n) = 1$ as $h_0(n) = [H_{1,0}(k,n), \ldots, H_{LMA,0}(k,n)]^T$, where $(\cdot)^T$ denotes the transpose operator. Using (4.10), we can reformulate the microphone signal (3.1) as

$$Y(n) = s^T(n-1)h_0(n) + S_1(n) + V(n), \quad (4.11)$$

where the vector $s_1(n) = [S_1(n), \ldots, S_1(n-L_{MA}+1)]^T$ and $E(n)$ is the signal that cannot be predicted using $s_1(n-1)$. We assume that the prediction error $E(k,n) \sim \mathcal{N}(0, \Phi_E(n))$ is a zero-mean complex Gaussian random variable and uncorrelated over time, i.e. $\mathbb{E}\{E(n)E^*(n-\ell)\} = 0, \forall \ell \neq 0$. This assumption hold well for the STFT coefficients of non-reverberant speech and a wide variety of noise types that typically have short to moderate temporal correlation in the time domain, and is widely used in speech processing methods [29, 132, 133].

To account for the time-varying reverberant conditions due to slow movements of the speaker or the microphone, or other changes in the room, we propose to use a first-order Markov model for the RCTF coefficients

$$h_0(n) = h_0(n-1) + w(n), \quad (4.12)$$

where $w(n)$ is a white noise process taking the uncertainty of $h_0(n)$ over time into account. We assume that $w(n) \sim \mathcal{N}(0, \Phi_w)$ and that $w(n)$ and $E(n)$ are uncorrelated.

### 4.3.2 Kalman filter to estimate RCTF coefficients

The well-known Kalman filter [134] is an adaptive solution to estimate the coefficients $h_0(n)$ by considering (4.12) and (4.11) as state and measurement equations, respectively. The Kalman filter is optimal if the Gaussian assumptions on $E(n)$ and $w(n)$ hold, and if $s_1(n-1)$ is known. The Kalman filter prediction and update equations to estimate $h_0(n)$ are given in Algorithm 4.1 in lines 5-6 and 9-11, where the error covariance matrix

$$P(n) = \mathbb{E}\left\{\left[\tilde{h}_0(n) - h_0(n)\right] \left[\tilde{h}_0(n) - h_0(n)\right]^H\right\}. \quad (4.13)$$

After having obtained the estimate $\tilde{h}_0(n)$ from the Kalman filter, an estimate of the late reverberation PSD following the late reverberation model (4.2) and using the RCTF formulation (4.10) is obtained by

$$\tilde{\phi}_{RD}(n) = \left\langle \left|\tilde{R}_D(n)\right|^2 \right\rangle = \left\langle \sum_{\ell=D}^{L_{MA}} \tilde{H}_{\ell,0}(n) S_1(n-\ell) \right\rangle^2. \quad (4.14)$$

Note that $R_1(n)$ contains all reverberation except the first frame, whereas $R_D(n)$ contains the late reverberation starting at frame $D$. 

---

**Algorithm 4.1**

1. Initialize $h_0(0)$ and $P(0)$.
2. Predict: $\tilde{h}_0(n) = h_0(n-1) + w(n)$.
3. Update: $h_0(n) = \tilde{h}_0(n) + P(n)Y(n) - E(n)$.
4. Update $P(n)$.

---

**Equation (4.14)**

$$\tilde{\phi}_{RD}(n) = \left\langle \left|\tilde{R}_D(n)\right|^2 \right\rangle = \left\langle \sum_{\ell=D}^{L_{MA}} \tilde{H}_{\ell,0}(n) S_1(n-\ell) \right\rangle^2. \quad (4.14)$$

Note that $R_1(n)$ contains all reverberation except the first frame, whereas $R_D(n)$ contains the late reverberation starting at frame $D$. 

---

**Notes**

- **Equation (4.10)**
  - $H_{\ell,0}(k,n) = \frac{H_\ell(k,n)}{H_0(k,n)}$.

- **Equation (4.11)**
  - $Y(n) = s^T(n-1)h_0(n) + S_1(n) + V(n)$.

- **Equation (4.12)**
  - $h_0(n) = h_0(n-1) + w(n)$.
Algorithm 4.1 Proposed algorithm

1: Initialize:
2: \( \hat{P}(0) = I; \hat{h}_0(0) = 0; \hat{s}_1(0) = 0; \hat{\psi}_E(0) = 0 \)
3: for each \( k \) and \( n \) do
4: 1) Kalman prediction step:
5: \( \hat{P}(n|n-1) = \hat{P}(n-1) + \Phi_w(n) \)
6: \( E(n|n-1) = Y(n) - \hat{s}_1^T(n-1)\hat{h}_0(n-1) \)
7: 2) Estimate prediction error PSD using (4.16), (4.17), (4.15).
8: 3) Kalman update step:
9: \( \hat{k}(n) = \hat{P}(n|n-1)\hat{s}_1^T(n-1) \left[ \hat{s}_1^T(n-1)\hat{P}(n|n-1)\hat{s}_1(n-1) + \hat{\phi}_E(n) \right]^{-1} \)
10: \( \hat{P}(n) = [I - k(n)\hat{s}_1^T(n-1)] \hat{P}(n|n) \)
11: \( \hat{h}_0(n) = \hat{h}_0(n-1) + k(n)E(n|n-1) \)
12: 4) Estimate desired signal using (4.14), (3.7), (3.6) and (3.4).
13: 5) Estimate auxiliary signal:
14: \( \hat{\phi}_{R_1}(n) = \left( \hat{s}_1^T(n)\hat{h}_0(n) \right)^2 \)
15: Estimate \( \xi_1(n) \) using the decision-directed approach (3.7).
16: \( \hat{S}_1(n) = \frac{\xi_1(n)}{\xi_1(n)+1} Y(n) \)
17: \( \hat{s}_1(n) = [\hat{S}_1(n), \ldots, \hat{S}_1(n-L_{MA}+1)]^T \)

To compute the Kalman filter and (4.14), we need to estimate the unknowns \( s_1(n-1), \phi_E(n) \) and \( \Phi_w(n) \). The elements of the vector \( s_1(n-1) \) can be estimated by using an additional Wiener filter similar to (3.4) and (3.6), that estimates \( s_1(n) \) by using the signal-to-interference ratio (SIR) \( \xi_1(n) = \frac{\phi_{S_1}(n)}{\phi_{S_1}(n)+\phi_V(n)} \) with the PSD \( \phi_{R_1}(n) = E\{ |R_1(n)|^2 \} \).

The PSD of the Kalman prediction error \( \phi_E(n) = \phi_{S_1}(n) + \phi_V(n) \) can be estimated by assuming that the RCTF coefficients \( \hat{h}_0(n) \) vary only slowly over time and therefore \( \hat{h}_0(n-1) \). We propose to estimate the PSD of the prediction error using a weighted sum of the a priori (predicted) estimate of the current frame and the a posteriori (updated) estimate of the last frame, i.e.,

\[
\hat{\phi}_E(n) = \beta_E \hat{\psi}_E(n-1) + (1 - \beta_E) \hat{\psi}_E(n|n-1),
\]

(4.15)

where \( \hat{\psi}_E(n) \) and \( \hat{\psi}_E(n|n-1) \) denote the estimated a posteriori and a priori error variances, respectively, and \( 0 < \beta_E < 1 \) is a weighting factor. The a posteriori error variance \( \psi_E(n) = E\{ |E(n)|^2 |Y(n), s_1(n-1), h_0(n) \} \) is conditioned only on signal model components contained in \( Y(n) \) as given by (4.11), and can be estimated by

\[
\hat{\psi}_E(n) = \left| Y(n) - \hat{s}_1^T(n-1)\hat{h}_0(n) \right|^2.
\]

(4.16)

In contrast, the a priori error variance is conditioned on the RCTF coefficients of the previous frame \( h_0(n-1) \), defined as \( \psi_E(n|n-1) = E\{ |E(n)|^2 |Y(n), s_1(n-1), h_0(n-1) \} \), and can be estimated by

\[
\hat{\psi}_E(n|n-1) = |Y(n) - \hat{s}_1^T(n-1)\hat{h}_0(n-1)|^2 = |E(n|n-1)|^2.
\]

(4.17)
4.4 Performance evaluation

After presenting the experimental setup, the relation between the reverberation time and the RCTF length of the proposed estimator is investigated. Then, we analyze the performance of the proposed and the state-of-the-art estimators in time-varying acoustic conditions. Finally, we investigate the performance of a WF using the estimated PSDs.

4.4.1 Setup

A test signal was created by convolving measured room impulse responses with a male speaker signal of 21 s length. The RIRs were measured in a laboratory with variable acoustics at Bar-Ilan University, Israel. The room was set up for three different reverberation times, where in each setting, two positions at different distances were measured. The $T_{60}$ in octave bands and the DRR for each source distance are shown in Table 4.1. Pink noise was added with variable input signal-to-noise ratio (iSNR) as defined in Appendix A. The sampling frequency was 16 kHz and the STFT was implemented with a $K = 1024$ point discrete Fourier transform (DFT), a frame shift of $T_{\text{hop}} = 16$ ms, and Hann windows of 32 ms length. The recursive smoothing factor for time averaging was $\alpha = e^{-\frac{T_{\text{hop}}}{\tau}}$ with a time constant of $\tau = 20$ ms. The weighting factor and the process noise variance of the RCTF Markov model were chosen empirically as $\beta_E = 0.3$ and $\sigma_w^2(n)$.


\[ \sigma_w^2 = 10^{-4}, \] respectively. The overlap-skip factor for the method presented in Section 4.2.2 was \( P = 2; \) other values yielded worse results as found in [132].

The stationary noise PSD \( \phi_V(k,n) \) was estimated in advance during periods of speech absence. For evaluation purposes, we defined the true reverberation signal \( R_D(k,n) \) as the anechoic signal convolved with the tail of the reverberation starting 48 ms after the direct path. The delay \( D = 3 \) was set accordingly.

### 4.4.2 CTF filter length and reverberation time

From the CTF model it may seem reasonable that there is an optimal choice of the CTF filter length \( L_{MA} \) depending on the (frequency dependent) reverberation time. The signed logarithmic PSD estimation error is described in the Appendix B.2 and is computed for all time-frequency bins, where late reverberation was present. Here, we computed the absolute logarithmic PSD estimation error in octave bands for iSNR = 20 dB. Figure 4.3 shows the mean absolute error for two reverberation times using different CTF lengths \( L_{MA} \). The time span in ms corresponding to \( L_{MA} \) frames is indicated on the top x-axis. For example for the used hop size of \( T_{hop} = 16 \) ms, a filter length of \( L_{MA} = 15 \) covers a time span of \( 15 \times T_{hop} = 240 \) ms. Although a clear linear relationship between the optimal filter length \( L_{MA} \) and the subband \( T_{60}(k) \) is not given for all frequency bands in this experiment, we found that on average the optimal filter length corresponds to about \( 0.75 T_{60}(k) \).

The influence of the reverberation time and the source distance is shown in Figure 4.4. The RCTF based estimator is computed with two settings: i) choosing \( L_{MA} \) in each octave band depending on the reverberation time given in Table 4.1, where the length \( L_{MA} \) corresponds to 0.75 \( T_{60} \), and ii) with a fixed setting of \( L_{MA} = 15 \) for all frequencies, corresponding to 240 ms. The exponential model based estimators were also computed using the corresponding reverberation time for each octave band. The exponential model based forward estimator was once computed for the farfield assumption \( \kappa = 1 \), and once with the true \( \kappa \) that was computed using knowledge of the DRR by (4.4). Figure 4.4 shows the PSD, where the error means are indicated by various symbols (+, ·, ◦, ×), and the upper and lower confidence intervals of 68% are indicated by the whisker bars. Therefore, an error with zero mean and minimal variance is desired, while a positive error means overestimation and a negative error means underestimation.

We can observe that the true \( \kappa \) for the exponential forward estimator reduces the variance for the smaller distances, where the source might be within the critical distance, while at the large distances the results are identical to \( \kappa = 1 \) as the farfield assumption holds. The exponential
4.4. PERFORMANCE EVALUATION

![Figure 4.3](image1.png)

**Figure 4.3:** Absolute PSD estimation error in octave bands for RIRs with different reverberation times.

![Figure 4.4](image2.png)

**Figure 4.4:** PSD estimation error for various reverberation times and distances at iSNR = 20 dB.

The **backward** estimator tends to overestimate the PSD. The proposed RCTF based estimator with $L_{MA}$ corresponding to $0.75T_{60}$ yields a mean error close to 0 dB with the smallest variance in most conditions, and performs very similar to the exponential forward estimator with true $\kappa$. The performance loss by simply using a fixed length of, e.g., $L_{MA} = 15$, is rather small, such that in practice a fixed $L_{MA}$ can work in a wide range of acoustic conditions. Therefore, in further experiments, we use $L_{MA} = 15$.

### 4.4.3 Performance in time-varying conditions

In this experiment, we simulate a time-varying condition by changing the position of a speaker. For the first 9 s a male speaker is talking from the first position in the room with $T_{60} = 630$ ms at 2 m distance, and after 9 s he switches to the second position at 4 m distance. Pink noise was added with an iSNR of 20 dB. Figure 4.5 depicts an excerpt for the first 9 s of the true late reverberation PSD and the estimated PSDs using the methods under test. We can observe a tendency to overestimation for the exponential forward and backward estimators, and a tendency
to underestimation for the CTF model based estimator proposed by Erkelens [132], while the RCTF estimator seems to match the true PSD best.

Figure 4.6 depicts the logarithmic PSD estimation error in terms of mean and confidence intervals between the estimated and the true late reverberation PSD for various iSNRs. It can be observed that the exponential forward estimator yields the lowest variance for moderate and high iSNRs, but the mean error of the Erkelens and the RCTF estimators are closer to zero compared to the slight overestimating tendency of the exponential forward and backward estimators. The variance of the proposed estimator is slightly smaller than the variance of the Erkelens estimator. At low iSNRs the performance of all algorithms degrades. The exponential forward estimator is less robust in low iSNRs compared to the other estimators as its variance increases drastically.

To show the convergence behavior of the proposed algorithm, we computed the segmental logarithmic error of the estimated late reverberation PSDs at 20 dB iSNR in $T_{60} = 630$ ms. Figure 4.7 depicts on top the spectrogram of the microphone signal, where the position change is indicated by the red vertical line, and on bottom the absolute segmental error for all estimators. The proposed algorithm initially converges faster than the Erkelens estimator and yields on average the lowest absolute error during speech activity. The position change is tracked quickly by all estimators.
4.4. PERFORMANCE EVALUATION

![Graph showing PSD error vs. iSNR] (Figure 4.6) Estimation error of late reverberation PSD for varying iSNR, $T_{60} = 630$ ms.

![Spectrogram of microphone signal and segmental absolute estimation error] (Figure 4.7) Spectrogram of microphone signal (top) and segmental absolute estimation error of the late reverberation PSD (bottom).

4.4.4 Speech enhancement performance

Furthermore, we evaluate the performance of the Wiener filter given by (3.4) for dereverberation and noise reduction using the late reverberation PSD estimates discussed in this chapter. Table 4.2 shows the cepstral distance (CD), log-likelihood ratio (LLR), perceptual evaluation of speech quality (PESQ), frequency-weighted segmental signal-to-interference ratio (fwSSIR), quality assessment of reverberated speech (QAreamb) and Mel-frequency cepstral coefficient distance (MFCCD) as presented in Section B.1.1. The results for the objective measures are averaged over all iSNRs and all acoustic conditions shown in Table 4.1. Positive values indicate an improvement compared to the unprocessed microphone signal. It may seem surprising that the proposed estimator, which yields on average the best PSD error in terms of mean bias and variance, does not achieve consistently the best objective measures. However, PESQ, fwSSIR, QAreamb and MFCCD mainly reflect the interference suppression, which is higher when the reverberation PSD is slightly overestimated. Therefore, the exponential forward and backward estimators yield slightly better results in terms of these measures. In terms of CD and LLR, which are also related to the overall signal quality, the proposed RCTF yields equal results to the exponential forward estimator. The estimator proposed by Erkelens yields significantly worse results. The results for the exponential forward and backward estimators might still degrade, if
Table 4.2: Speech enhancement measure improvements using a Wiener filter with different late reverberation PSD estimators for iSNR = \{0, 10, \ldots, 50\} dB and six acoustic conditions.

<table>
<thead>
<tr>
<th>PSD estimator</th>
<th>$\Delta$CD</th>
<th>$\Delta$LLR</th>
<th>$\Delta$PESQ</th>
<th>$\Delta$fwSSIR [dB]</th>
<th>$\Delta$QAreverb</th>
<th>$\Delta$MFCCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCTF</td>
<td>0.58</td>
<td>0.12</td>
<td>0.34</td>
<td>8.56</td>
<td>4.85</td>
<td>0.07</td>
</tr>
<tr>
<td>Erkelens [135]</td>
<td>0.26</td>
<td>0.06</td>
<td>0.28</td>
<td>7.94</td>
<td>4.18</td>
<td>0.02</td>
</tr>
<tr>
<td>Exp forward [75]</td>
<td>0.58</td>
<td>0.12</td>
<td>0.39</td>
<td>8.98</td>
<td>5.43</td>
<td>0.09</td>
</tr>
<tr>
<td>Exp backward [89]</td>
<td>0.50</td>
<td>0.10</td>
<td>0.40</td>
<td>8.50</td>
<td>6.17</td>
<td>0.09</td>
</tr>
</tbody>
</table>

the reverberation time and the DRR is estimated and not known in advance, while the RCTF and Erkelens estimators used a fixed CTF length.

4.5 Summary

This chapter dealt with the estimation of the late reverberation PSD, which is required to compute a WF for dereverberation and noise reduction. In the single-channel case, only temporal information can be exploited. We have proposed a novel RCTF formulation based on a narrow-band moving average (MA) reverberation model. This model is more general than an established model assuming an exponentially decaying envelope of the CTF coefficients. Using the RCTF formulation, a Kalman filter was derived to estimate the RCTF coefficients and then to obtain the late reverberation PSD. Within the Kalman filter, the RCTF coefficients are assumed as time-varying and are modeled by a first-order Markov model. In contrast to methods based on the well-known exponential decay model, the proposed model is more general, includes additive noise, and does not require estimates of the reverberation time and DRR. The proposed RCTF based late reverberation PSD estimator outperforms a method based on a related CTF formulation, which estimates the CTF coefficients individually and not as a joint set like the Kalman filter. The proposed RCTF based estimator achieves similar results to methods based on the exponential decay model, without using prior knowledge of the reverberation time and DRR.
CHAPTER 5

Multichannel late reverberation PSD estimation

As presented in Section 3.2, the multichannel Wiener filter (MWF) and related beamformer-postfilter structures have been widely used for joint dereverberation and noise reduction [47, 82, 84, 106, 136, 137]. Within this framework, the late reverberation is modeled in the short-time Fourier transform (STFT) domain as an additive diffuse sound field with a time-varying power spectral density (PSD) and a time-invariant spatial coherence as discussed in Section 2.6. As the diffuse spatial coherence can be calculated analytically for known microphone array geometries, the remaining challenge is to obtain an estimate of the late reverberation PSD, which directly affects the performance of the MWF and hence the quality of the dereverberated signal. As the late reverberation PSD is highly time-varying, it is challenging to obtain an accurate estimate thereof.

To the best of our knowledge, the first multichannel methods to estimate the late reverberation PSD were proposed in [15, 105]. While early attempts for coherence based dereverberation have been already made decades ago in [138], in recent years advances have been made by exploiting the proposed spatial coherence model for dereverberation by estimating the coherent-to-diffuse ratio (CDR) as proposed in [139–141]. Although not in the context of dereverberation, methods to estimate the direct sound PSD in a diffuse noise field [104], or the diffuse sound PSD [142–145] have been proposed. In the past years, a multitude of estimators for the late reverberation or diffuse sound PSD has been developed, which assume that the reverberant sound field can be described by the direct sound propagating as plane waves in a time-varying diffuse field. Existing late reverberation PSD estimators can be divided into four classes as shown in Figure 5.1. The first three classes use the spatial coherence reverberation model. The first two classes are direct PSD estimators, whereas the third class comprises indirect PSD estimators, which require an additional step to obtain the reverberation PSD. In contrast to the first three classes, the fourth class is based on temporal reverberation models. Figure 5.1 shows the classification of the reverberation PSD estimators with indication of the class numbers, and references to methods within each class.

Estimators of the first class model the reverberation as a diffuse sound field with known spatial coherence and block the direct sound utilizing direction-of-arrival (DOA) information. In general, there can be multiple plane waves with different DOAs per time and frequency bin. The blocking operation simplifies the estimation procedure as the resulting (in some methods multiple) signals
CHAPTER 5. MULTICHANNEL LATE REVERBERATION PSD ESTIMATION

![Diagram of multichannel reverberation PSD estimators](image)

**Figure 5.1:** Classification of multichannel reverberation PSD estimators. The classes are indicated as numbers. Estimators marked with * were developed within this thesis.

After the blocking operation contain only filtered diffuse sound and noise. In [47, 151], the error PSD matrix of the blocking output signals is minimized, whereas in [84] and [147], maximum likelihood estimators (MLEs) are derived given the blocked, or blocked and additionally filtered, signals. In [147] the solution is obtained using the Newton method, whereas in [84] the solution is obtained by a root-finding procedure. Thiergart et al. developed several spatial filters to extract the diffuse sound while blocking the direct sound. In [81] a spatial filter is derived that maximizes the diffuse-to-noise ratio (DNR) at its output, while in [146] a linearly constrained minimum variance (LCMV) beamformer is proposed with a novel constraint set to block the direct sound and to extract the diffuse sound.

The spatial coherence based estimators of the second class use no blocking of the direct sound, therefore the unknown PSDs of direct sound and reverberation have to be estimated jointly. In contrast to the first class, these estimators assume only a single direct sound wave per time and frequency. In [152] a closed-form MLE for the direct sound and reverberation PSDs is presented without taking additive noise into consideration. The method presented in [148] obtains the MLE of the direct and reverberation PSDs using the Newton method. In [86], a batch expectation-maximization (EM) algorithm to estimate the direct and reverberation PSDs in the maximum likelihood (ML) sense is presented, where unlike in all other methods considered in this thesis, this method also estimates the spatial coherence matrix of the reverberation. In [149], the direct and reverberation PSDs are estimated jointly in the least-squares sense by minimizing the Frobenius norm of an error matrix.
Estimators of the third class are considered as indirect PSD estimators based on the spatial coherence model: Rather than estimating the reverberation PSD directly, the estimate is obtained by first estimating the CDR, and then estimating the reverberation PSD. Although there exists a multitude of CDR estimators [139, 140, 150], we constrain ourselves to the best performing CDR estimator reported in [150] to limit the number of algorithms under test.

Estimators of the fourth class utilize temporal reverberation models and make no assumption on the spatial coherence. In this class, the reverberation is described either using Polack’s model [41, 75], or using a narrowband moving average model [88], also known as convolutive transfer function (CTF) model (see Section 2.3). However, when the estimated late reverberation PSD is used in the MWF for dereverberation, an assumption on the spatial coherence of the late reverberation is still required.

In [153] a power based CDR estimation method using directional microphones is proposed. As we exclusively consider omnidirectional microphones in this work, and therefore a direct comparison is not possible, we do not further consider this estimator.

All of the reverberation PSD estimators of these four classes can be used equivalently in the established MWF or similar beamformer-postfilter structures [106] for dereverberation. However, the properties and performance of this vast variety of PSD estimators are unclear and have been never compared in a unified framework. In this chapter, we provide an overview and comparison of the current state-of-the-art reverberation PSD estimators as a guideline for an application-dependent choice.

In Section 5.1, we review the general signal model and formulate the problem of estimating the reverberation PSD. Section 5.2 reviews state-of-the-art estimators and Section 5.3 presents the proposed estimator [151]. It should be noted that the proposed estimator in Section 5.3 was one of the first estimators of the reverberation PSD for dereverberation using an MWF, which triggered the publication of some estimators included in Section 5.2 by other researchers. The relations and differences between the estimators are discussed in Section 5.4. A common weakness of the spatial coherence based estimators is a systematic bias at high direct-to-reverberation ratios (DRRs). Therefore, we propose a bias compensation method depending on the DRR in Section 5.5. An exhaustive experimental evaluation for a multitude of estimators based on the single-wave model is presented in Section 5.6, where we analyze the error of the estimated PSDs as well as the resulting performance of the spatial filter using the PSD estimates. An evaluation of the proposed PSD estimators for the general multi-wave model is conducted in Section 5.6.4.

5.1 Problem formulation

Given the signal model for the MWF in (3.8), we recall the model of the microphone signal PSD matrix with the spatial coherence model of the reverberation $r(k,n)$ described in Section 2.6 for the multi-wave and single-wave case, i.e.,

$$
\Phi_y(k,n) = A(k,n)\Phi_d(k,n)A^H(k,n) + \phi_r(k,n)\Gamma_r(k) + \Phi_v(k,n)
$$

$$
\Phi_y(k,n) \bigg|_{l=1} = \phi_d(k,n)a(k,n)a^H(k,n) + \phi_r(k,n)\Gamma_r(k) + \Phi_v(k,n).
$$
Our aim is to obtain an estimate of the reverberation PSD $\phi_r(k,n)$ given knowledge of the propagation matrix $A(k,n)$ or the vector $a(k,n)$, the reverberation spatial coherence matrix $\Gamma_r(k)$, and the noise PSD matrix $\Phi_v(k,n)$. To improve the readability of the equations in the remainder of this chapter, the dependency on the frequency and time indices $(k,n)$ is omitted in equations wherever possible.

## 5.2 State-of-the-art reverberation PSD estimators

### 5.2.1 Spatial coherence based direct PSD estimators

As shown in Figure 5.1, the coherence based direct reverberation PSD estimators comprise blocking based methods (Sections 5.2.1.1 and 5.2.1.2) and non-blocking based methods (Sections 5.2.1.3 - 5.2.1.5). These estimators are exclusively based on the spatial coherence model (2.17) without any assumptions on the temporal structure of the reverberation. Therefore, these estimators mainly depend on the reverberation coherence matrix $\Gamma_r(k)$ and on the relative transfer functions (RTFs) $A(k,n)$ or $a(k,n)$. As in Section 3.2, we further assume prior knowledge or an available estimate of the noise PSD matrix $\Phi_v(k,n)$.

#### 5.2.1.1 MLE using blocking output signals

For $I < M$ desired directional signals, the blocking based methods use a set of $J = M - I$ signals, which are generated by canceling the desired sound from the microphone signals using a blocking matrix. The $J$-dimensional signal vector $u(k,n)$ is obtained as

$$u = B^H y,$$

(5.1)

where the blocking matrix $B(k,n)$ of dimension $M \times J$ has to fulfill the constraint

$$B^H A = 0_{J \times 1}.$$  

(5.2)

Possible choices for the blocking matrix are discussed in [145,151,154]. In this work, we use the eigenspace blocking matrix given by [154]

$$B = [I_{M \times M} - A(A^H A)^{-1} A^H] I_{M \times J},$$

(5.3)

where $I_{M \times J}$ is a truncated identity matrix. As a consequence of using (3.8) with (5.1) and (5.2), it follows that the PSD matrix of the blocking output signals $u(k,n)$ is given by

$$\Phi_u = B^H \Phi_y B$$

$$= B^H \frac{\phi}{\Theta_{J \times J}} B + \phi_r B^H \Gamma_r B + B^H \frac{\Phi_v}{\Phi_v} B.$$  

(5.4)

Note that $\tilde{\Gamma}_r(k,n)$ and $\tilde{\Phi}_v(k,n)$ denote the reverberation coherence and noise PSD matrix after applying the blocking matrix.
The methods presented in [147] and [84] both start from the assumption that the elements of the microphone signal vector \( y(k,n) \) are zero-mean complex Gaussian random variables
\[
y(k,n) \sim \mathcal{N}(0, \Phi_y).
\] (5.5)

In [147], the ML problem is solved by an iterative Newton method, whereas in [84] a filtered version of the signals \( u(k,n) \) is used and the ML problem is solved by a root-finding method. Although both [84] and [147] were proposed by assuming only a single desired sound component per frequency \((I = 1)\), an extension to a general \( I \) can be achieved by defining the blocking matrix with multiple constraints as in (5.2).

**Solution using root-finding method** [84] In the ML estimator using a root finding method, the blocking output signals \( u(k,n) \) are filtered to diagonalize \( \tilde{\Phi}_v(k,n) \) in (5.4). Specifically, a whitening matrix \( D(k,n) \) of dimension \( J \times J \) defined as the Cholesky factor of the inverse of \( \tilde{\Phi}_v(k,n) \), i.e.,
\[
\tilde{\Phi}_v^{-1}(k,n) = D(k,n)D^H(k,n),
\]
yielding
\[
\Phi_u = D^HBD^H \quad \text{(5.6)}
\]
and its PSD matrix is given by
\[
\Phi_u = D^H\Phi_u D = \phi_r \Gamma_r + \mathbf{I}, \quad \text{(5.7)}
\]
with \( \Phi_u = D^H\tilde{\Phi}_v D = D^H\tilde{\Phi}_v D \).

As a result of the described whitening of \( \tilde{\Phi}_v(k,n) \), the matrices \( \Phi_u(k,n) \) and \( \Gamma_r(k,n) \) can be diagonalized using the same unitary matrix \( C(k,n) \), i.e.,
\[
\Phi_u = C\Lambda_u C^H, \quad \Gamma_r = C\Lambda_r C^H, \quad \text{(5.8)}
\]
where the orthonormal columns of \( C(k,n) \) are the eigenvectors, and where \( \Lambda_u(k,n) \) and \( \Lambda_r(k,n) \) are diagonal matrices containing the eigenvalues of \( \Phi_u(k,n) \) and \( \Gamma_r(k,n) \), respectively. Due to (5.7), these eigenvalues are related as \( \lambda_{u,j} = \phi_r \lambda_{r,j} + 1 \), where \( \lambda_u(k,n) \) and \( \lambda_r(k,n) \) denote the \( j \)-th eigenvalue of \( \Phi_u(k,n) \) and \( \Gamma_r(k,n) \), respectively.

Given the filtered blocking output signals \( \Phi_u(k,n) \) in (5.6), the ML estimate of \( \phi_r \) is given by
\[
\hat{\phi}_r = \arg \max_{\phi_r} \log f(\Phi_u; \phi_r), \quad \text{(5.9)}
\]
where \( f(\Phi_u; \phi_r) \) is the complex Gaussian probability density function (PDF) of the filtered and blocked signals \( u(k,n) \) that depends on \( \phi_r(k,n) \). By setting the derivative of the log-likelihood function to zero and exploiting the diagonal structure of the involved matrices (for more details, see [84]), we obtain the polynomial
\[
p(\phi_r) = \sum_{j=1}^{J} \left( \frac{\phi_r - g_j - 1}{\lambda_{r,j}} \right) \prod_{\ell=1}^{J} \left( \frac{\phi_r + 1}{\lambda_{r,\ell}} \right)^2, \quad \text{(5.10)}
\]
CHAPTER 5. MULTICHANNEL LATE REVERBERATION PSD ESTIMATION

where \( g_j(k,n) \) denotes the \( j \)-th diagonal element of \( \mathbf{C}^\mathsf{H} \mathbf{D}^\mathsf{H} \hat{\Phi}_u \mathbf{D} \). It has been shown in [84] that the root of the polynomial \( p(\phi_r) \) yielding the highest value of the likelihood function (5.9) is the ML estimate \( \hat{\phi}_r(k,n) \). An estimate of \( \Phi_u(k,n) \) can be obtained by using recursive averaging, i.e.,

\[
\hat{\Phi}_u(k,n) = \alpha \hat{\Phi}_u(k,n-1) + (1 - \alpha) \mathbf{u}(k,n)\mathbf{u}^\mathsf{H}(k,n),
\]

where \( 0 < \alpha < 1 \) is a forgetting factor.

**Solution using Newton’s method** [147] To solve the ML estimation problem

\[
\hat{\phi}_r = \arg \max_{\phi_r} \log f(\mathbf{u}; \phi_r),
\]

Newton’s method may be used to derive an iterative search (c.f. [155]) by the update rule

\[
\phi_r^{(\ell+1)} = \phi_r^{(\ell)} - \frac{\mathcal{D}(\phi_r^{(\ell)})}{\mathcal{H}(\phi_r^{(\ell)})},
\]

where \( \ell \) denotes the iteration index, and \( \mathcal{D}(\phi_r) \) and \( \mathcal{H}(\phi_r) \) are the gradient and the Hessian of the log-likelihood given by

\[
\mathcal{D}(\phi_r) = \frac{\partial \log f(\mathbf{u}; \phi_r)}{\partial \phi_r},
\]

\[
\mathcal{H}(\phi_r) = \frac{\partial^2 \log f(\mathbf{u}; \phi_r)}{\partial \phi_r^2}.
\]

As shown in [147], the gradient is equal to

\[
\mathcal{D}(\phi_r) = J \text{tr} \left( \Phi_u^{-1} \mathbf{u} \mathbf{u}^\mathsf{H} - \mathbf{I} \right) \Phi_u^{-1} \frac{\partial \Phi_u}{\partial \phi_r},
\]

where \( \text{tr} \{ \cdot \} \) denotes the trace operator, and \( \frac{\partial \Phi_u}{\partial \phi_r} = \mathbf{B}^\mathsf{H} \mathbf{\Gamma}_r \mathbf{B} \). The Hessian matrix is equal to

\[
\mathcal{H}(\phi_r) = -J \text{tr} \left( \Phi_u^{-1} \frac{\partial \Phi_u}{\partial \phi_r} \Phi_u^{-1} \mathbf{u} \mathbf{u}^\mathsf{H} \Phi_u^{-1} \frac{\partial \Phi_u}{\partial \phi_r} + (\Phi_u^{-1} \mathbf{u} \mathbf{u}^\mathsf{H} - \mathbf{I}) \Phi_u^{-1} \frac{\partial \Phi_u}{\partial \phi_r} \Phi_u^{-1} \frac{\partial \Phi_u}{\partial \phi_r} \right). \tag{5.16}
\]

As shown in [147], the Newton update (5.12) can be computed efficiently by re-arranging (5.15) and (5.16), using an eigenvalue decomposition of \( \mathbf{B}^\mathsf{H} \mathbf{\Gamma}_r \mathbf{B} \) and exploiting the resulting diagonal matrices. In practice, the matrix \( \mathbf{u}(k,n)\mathbf{u}^\mathsf{H}(k,n) \) is substituted by the smoothed version \( \hat{\Phi}_u(k,n) \). The Newton iterations are initialized with \( \phi_r^{(0)}(k,n) = \epsilon^{\frac{1}{J}} \text{tr} \left\{ \hat{\Phi}_u(k,n) \right\} \), where \( \epsilon \) is a small positive value. The Newton algorithm is stopped if the estimate at iteration \( \ell = \ell_{\text{stop}} \) reaches a predefined lower or upper bound, or if a convergence threshold is reached, and the estimate is obtained by \( \hat{\phi}_r(k,n) = \phi_r^{(\ell_{\text{stop}})}(k,n) \).

**5.2.1.2 Diffuse beamformers**

Thiergart et al. developed several beamformers [46,81,146] that aim at extracting the reverberation, modeled as diffuse sound, while blocking the desired sound. As our preliminary experiments unveiled almost identical performance across those beamformers in terms of reverberation PSD estimation, we present the most elegant here: The beamformer proposed in [146] minimizes the
5.2. STATE-OF-THE-ART REVERBERATION PSD ESTIMATORS

noise under the linear constraints of blocking the desired sound and not distorting the diffuse sound on average, i.e.,

\[
g_r = \arg \min_{g} g^H \Phi_v g \quad (5.17a)
\]

subject to

\[
g_r^H A = 0 \quad (5.17b)
\]

\[
g_r^H \gamma_1 = 1, \quad (5.17c)
\]

where \(\gamma_1(k)\) is the first column of \(\Gamma_r(k)\). While the constraint (5.17b) blocks the direct sound, the constraint (5.17c) aims at capturing the sound from all directions (besides the spatial null). The analytic solution to (5.17) is equal to an LCMV filter [146].

The late reverberation PSD can then be estimated by subtracting the PSD of the filtered noise components from the PSD of the filtered input signals normalized by the filtered diffuse coherence [81], i.e.

\[
\hat{\phi}_r = \max \left\{ \frac{g_r^H \Phi_y g_r - g_r^H \Phi_v g_r}{g_r^H \Gamma_r g_r}, 0 \right\}, \quad (5.18)
\]

where the max\{\cdot\} operation is introduced to avoid negative PSD estimates. The input PSD matrix is recursively estimated using

\[
\hat{\Phi}_y(k,n) = \alpha \hat{\Phi}_y(k,n-1) + (1 - \alpha) y(k,n) y^H(k,n). \quad (5.21)
\]

5.2.1.3 Joint MLE of reverberation and signal PSDs using Newton’s method [148]

A ML method to jointly estimate \(\phi_d(k,n)\) and \(\phi_r(k,n)\) is presented under the assumption that the diffuse coherence matrix \(\Gamma_r(k)\) is known.

By defining \(p(k,n) = [\phi_d(k,n), \phi_r(k,n)]^T\) as the unknown parameter set, the ML estimate of \(p(k,n)\) given \(y(k,n)\) can be found by Newton’s method [155] using

\[
p^{(\ell+1)} = p^{(\ell)} - \mathcal{H}^{-1}(p^{(\ell)}) \delta(p^{(\ell)}), \quad (5.19)
\]

where \(\delta(p)\) is the gradient of the log-likelihood, and \(\mathcal{H}(p)\) is the corresponding Hessian matrix, i.e.,

\[
\delta(p) \equiv \frac{\partial \log f(y; p)}{\partial p}, \quad \mathcal{H}(p) \equiv \frac{\partial^2 \log f(y; p)}{\partial p \partial p^T}, \quad (5.20)
\]

where \(f(y; p)\) is the complex Gaussian PDF of the microphone signal vector. The gradient \(\delta(p) \equiv [\delta_d(p), \delta_r(p)]^T\) is a 2-dimensional vector with elements

\[
\delta_p(p) = M \text{ tr} \left( (\Phi_y^{-1} y y^H - \mathbf{1}) \Phi_y^{-1} \frac{\partial \Phi_y}{\partial \phi_p} \right), \quad (5.21)
\]

where \(p \in \{d,r\}\), \(\frac{\partial \Phi_y}{\partial \phi_d} = a a^H\), and \(\frac{\partial \Phi_y}{\partial \phi_r} = \Gamma_r\).
The Hessian is a symmetric $2 \times 2$ matrix:

$$
\mathcal{H} (p) = \begin{bmatrix} 
\mathcal{H}_{ss} (p) & \mathcal{H}_{rs} (p) \\
\mathcal{H}_{sr} (p) & \mathcal{H}_{rr} (p) 
\end{bmatrix}
$$

(5.22)

with the elements

$$
\mathcal{H}_{pq} (p) = -M \operatorname{tr} \left\{ \Phi^{-1} \frac{\partial \Phi}{\partial \phi_q} \Phi^{-1} \frac{\partial \Phi}{\partial \phi_p} + (\Phi^{-1} \Phi - \mathbf{I}) \left( \Phi^{-1} \frac{\partial \Phi}{\partial \phi_p} \Phi^{-1} \frac{\partial \Phi}{\partial \phi_q} - \Phi^{-1} \frac{\partial \Phi}{\partial \phi_q} \Phi^{-1} \frac{\partial \Phi}{\partial \phi_p} \right) \right\},
$$

(5.23)

where $p, q \in \{d, r\}$.

In practice, the matrix $\mathbf{y} \mathbf{y}^H$ in (5.21) and (5.23) is replaced by the smoothed version $\hat{\Phi}_y$. The algorithm is initialized with $\phi (0) = \epsilon \mathbf{1} \mathbf{1}^T \mathcal{H} \{ \hat{\Phi}_y \}$. The Newton algorithm is stopped if the estimates at iteration $\ell$ reach a predefined lower or upper bound, or if a convergence threshold is reached.

### 5.2.1.4 Joint MLE of reverberation and signal PSDs using the EM method [86]

The EM presented method is a batch processing algorithm that provides estimates of the desired sound PSD $\phi_d(k, n)$, the RTF vector $\mathbf{a}(k, n)$, the late reverberation PSD $\phi_r(k, n)$ and the late reverberation coherence matrix $\Gamma_r(k)$. For consistency with the other methods, we assume that $\mathbf{a}(k, n)$ is known and is therefore not estimated within the EM procedure. In Section 5.2.1.4, we describe the method proposed in [86]. In Section 5.2.1.4, the method is modified by assuming prior knowledge of the coherence matrix $\Gamma_r(k)$ to investigate the effect of estimating $\Gamma_r(k)$.

**ML-EM with unknown reverberation coherence matrix**  The desired and diffuse sound components are concatenated in the hidden data vector

$$
\mathbf{q}(k, n) \triangleq \left[ S_d(k, n) \quad \mathbf{r}^T(k, n) \right]^T.
$$

(5.24)

Using this definition, the microphone signal model (3.8) for $I = 1$ can be rewritten as

$$
\mathbf{y}(k, n) = \mathbf{H}(k, n) \mathbf{q}(k, n) + \mathbf{v}(k, n),
$$

(5.25)

where the matrix $\mathbf{H}(k, n) \triangleq [\mathbf{a}(k, n), \mathbf{I}_{M \times M}]$. The desired parameter set is

$$
\mathbf{\theta}(k) = \{ \overline{\phi}_d(k), \overline{\phi}_r(k), \Gamma_r(k) \},
$$

(5.26)

where $\overline{\phi}_d(k) = [\phi_d(k, 1), \ldots, \phi_d(k, N)]^T$ and $\overline{\phi}_r(k) = [\phi_r(k, 1), \ldots, \phi_r(k, N)]^T$, with $N$ being the number of frames. By concatenating the hidden data vectors of all time frames $1, \ldots, N$ to $\overline{\mathbf{q}}(k) = [\mathbf{q}^T(k, 1), \ldots, \mathbf{q}^T(k, N)]^T$, and defining $\overline{\mathbf{y}}(k)$ similarly, the conditional expectation of the log-likelihood function can be deduced as

$$
Q \left( \mathbf{\theta}; \mathbf{\theta}^{(\ell)} \right) = \mathcal{E} \left\{ \log f(\overline{\mathbf{y}}(k), \overline{\mathbf{q}}(k); \mathbf{\theta}) \ \bigg| \ \overline{\mathbf{y}}(k); \ \mathbf{\theta}^{(\ell)} \right\},
$$

(5.27)

where $\mathbf{\theta}^{(\ell)}$ is the parameter-set estimate at iteration $\ell$. 

---

Note: The content has been formatted to maintain readability and coherence, with proper use ofMathJax for mathematical expressions. The document is now in a standard text format suitable for further analysis or translation.
For implementing the E-step, it is sufficient to estimate 
\[ \hat{q}(k, n) \triangleq \mathbb{E}\{q(k, n)|y(k, n); \theta^{(t)}\} \]
and \( \hat{\Psi}_{q}(k, n) \triangleq \mathbb{E}\{q(k, n)q^H(k, n)|y(k, n); \theta^{(t)}\} \) being the first- and second-order statistics of the hidden-data given the measurements, respectively. Assuming that \( y(k, n) \) and \( q(k, n) \) in (5.25) are zero-mean Gaussian random vectors, \( q(k, n) \) can be estimated by the optimal linear estimator
\[ \hat{q} = \Phi_q^{(t)} H^H \left( \Phi_y^{(t)} \right)^{-1} y \] (5.28)
with
\[ \Phi_q^{(t)}(k, n) = \begin{bmatrix} \phi_d^{(t)}(k, n) & 0_{1 \times M} \\ 0_{M \times 1} & \phi_r^{(t)}(k, n) \Gamma_r^{(t)}(k) \end{bmatrix}, \] (5.29)
and \( \Phi_y^{(t)} = H\Phi_q^{(t)} H^H + \Phi_v \). The matrix
\[ \hat{\Psi}_{q}(k, n) \triangleq \begin{bmatrix} |S_d(k, n)|^2 & S_d(k, n)r^H(k, n) \\ S_d^*(k, n)r(k, n) & r(k, n)r^H(k, n) \end{bmatrix}, \] (5.30)
can be obtained by
\[ \hat{\Psi}_{q} = \hat{q} q^H + \Phi_q^{(t)} - \Phi_q^{(t)} H^H \left( \Phi_y^{(t)} \right)^{-1} H\Phi_q^{(t)}. \] (5.31)

Maximizing \( Q(\theta; \theta^{(t)}) \) with relation to the problem parameters constitutes the M-step, i.e.

1. \( \phi_d^{(t+1)}(k, n) = |S_d(k, n)|^2 \) (5.32)
2. \( \Gamma_r^{(t+1)}(k) = \frac{1}{N} \sum_{n=1}^{N} \frac{r(k, n)r^H(k, n)}{\phi_r^{(t)}(k, n)} \) (5.33)
3. \( \phi_r^{(t+1)}(k, n) = \frac{1}{M} \text{tr} \left\{ \frac{r(k, n)r^H(k, n)}{\Gamma_r^{(t+1)}(k)} \right\} \) (5.34)

The EM iterations are initialized with \( \phi_d^{(0)}(k, n) = \epsilon_s \frac{1}{M} \text{tr} \left\{ \hat{\Phi}_y(k, n) \right\} \) and \( \phi_r^{(0)}(k, n) = \epsilon_r \frac{1}{M} \text{tr} \left\{ \hat{\Phi}_y(k, n) \right\} \), where \( \epsilon_s > \epsilon_r \), and \( \Gamma_r^{(0)}(k) \) is initialized with the theoretical diffuse field coherence given by (2.13).

**ML-EM with known reverberation coherence matrix** By assuming that the diffuse field model for \( \Gamma_r(k) \) given by (2.13) holds, the method described in Section 5.2.1.4 needs to be modified only by omitting (5.33) in the M-step and using the a priori known spatial coherence matrix instead.

**5.2.1.5 Joint least-squares estimation of reverberation and signal PSDs [149]**
By matching \( \hat{\Phi}_y(k, n) \), which can be estimated from the microphone signals, and its model given in (3.9), the problem at hand can be formulated as a system of \( M^2 \) equations in two
unknown variables [149]. Since there are more equations than variables, the vector $p(k,n) = [\phi_d(k,n), \phi_r(k,n)]^T$ that minimizes the total squared error can be found by minimizing the squared Frobenius norm as

$$\hat{p} = \arg\min_p \|\hat{\Phi}_y - (\phi_d a a^H + \phi_r \Gamma_r + \Phi_v)\|_F^2,$$  

(5.35)

where $\|\cdot\|_F^2$ denotes the Frobenius norm, and $\Phi_{LS}(k,n)$ is the error matrix. Following some algebraic steps, the cost-function in (5.35) can be written as

$$\|\Phi_{LS}\|_F^2 = p^T A_{LS} p - 2 b_{LS}^T p + C_{LS},$$  

(5.36)

where $C_{LS}(k,n)$ is independent of $p(k,n)$, and $A_{LS}(k,n)$ and $b_{LS}(k,n)$ are defined as

$$A_{LS} \equiv \begin{bmatrix} (a^H a)^2 & a^H \Gamma_r a \\ a^H \Gamma_r a & \text{tr} \left\{ \Gamma_r^H \Gamma_r \right\} \end{bmatrix},$$  

(5.37)

and

$$b_{LS} \equiv \begin{bmatrix} \Re \left\{ a^H (\hat{\Phi}_y - \Phi_v) a \right\} \\ \Re \left\{ \text{tr} \left\{ (\hat{\Phi}_y - \Phi_v) \Gamma_r^H \right\} \right\} \end{bmatrix}.$$  

(5.38)

Since the cost-function $\|\Phi_{LS}(k,n)\|_F^2$ in (5.36) has a quadratic form, setting its gradient with respect to $p(k,n)$ to zero yields

$$\hat{p}(k,n) = A_{LS}^{-1}(k,n) b_{LS}(k,n).$$  

(5.39)

### 5.2.2 Spatial coherence based indirect PSD estimators

While all methods in Section 5.2.1 directly estimate the late reverberation PSD, this section deals with indirect estimators. Within this class of estimators, we focus on methods using an estimate of the CDR to estimate the PSD of the diffuse sound, i.e. late reverberation. These estimators rely on the fact that the desired signal $S_d(k,n)$ is fully coherent across all microphones and the coherence of the reverberation is smaller than one.

The CDR as defined in [139, 150] for the microphone pair $m, m' \in \{1, \ldots, M\}$ is given by

$$\text{CDR}_{m,m'}(k,n) = \frac{\phi_d(k,n) |A_m(k,n)| |A_{m'}(k,n)|}{\phi_r(k,n)},$$  

(5.40)

where $A_m(k,n)$ is the $m$-th element of $a(k,n)$. The CDR can be estimated using various methods, e.g. [150, 156]. To limit the number of estimators under test, we restrict ourselves to the "proposed 2" CDR estimator described in [150], which was reported to perform best across the
5.3. PSD MATRIX BASED LEAST-SQUARES METHOD WITH BLOCKING

Considered CDR estimators. By also compensating for the additive noise as proposed in [79], the CDR estimate is obtained by

$$\hat{\text{CDR}}_{m,m'} = 1 - \frac{r_d^{(m,m')} \cos \left( \arg \left( r_d^{(m,m')} \right) \right)}{\Gamma^{(m,m')} - \Gamma_d^{(m,m')}} \left| \left( \Gamma_d^{(m,m')} \right)^* \left[ \Gamma^{(m,m')} - \Gamma_x^{(m,m')} \right] \right| \Re \left\{ \left( \Gamma_d^{(m,m')} \right)^* \Gamma_x^{(m,m')} \right\} - 1, \quad (5.41)$$

where

$$\Gamma_{d}^{(m,m')} = \frac{A_m A_{m'}^*}{|A_m| |A_{m'}|}$$

and

$$\Gamma_{x}^{(m,m')} = \frac{\Phi_{x}(m,m')}{\sqrt{\Phi_{x}(m,m') \Phi_{x}^{*}(m',m')}}$$

with

$$\Phi_{y}(m,m') = \Phi_{y}^{(m,m')} - \Phi_v^{(m,m')}.$$

To take all microphones into account, we average the CDR estimate across all microphone pairs [104, 142]

$$\hat{\text{CDR}}(k,n) = \frac{1}{|\mathcal{M}|} \sum_{m,m' \in \mathcal{M}} \frac{\hat{\text{CDR}}_{m,m'}(k,n)}{|A_{m}(k,n)| |A_{m'}(k,n)|}, \quad (5.42)$$

where the set \( \mathcal{M} \) contains all microphone pair combinations. Given an estimate of the CDR, and exploiting the diffuse homogeneity, the late reverberation PSD is obtained by [151]

$$\hat{\phi}_s(k,n) = \frac{1}{\hat{\text{CDR}}(k,n)} \frac{\hat{\Phi}_y(k,n) - \Phi_v(k,n)}{\hat{\Phi}_v(k,n) \hat{\Phi}_v^{*}(k,n)} \right\} \hat{\text{CDR}}(k,n) + 1. \quad (5.43)$$

If the desired sound \( S_d(k,n) \) is modeled as a plane wave such that \( |A_m(k)| = 1, \ \forall m \), then (5.40) – (5.43) can be simplified.

5.2.3 Temporal model based PSD estimators

Instead of modeling the late reverberation as a diffuse sound field, estimators within the fourth class exploit the temporal structure of reverberation, without taking any dependence between the microphone channels into account.

Several temporal model based late reverberation PSD estimators have been presented already in Section 4. In detail we consider as representatives the exponential decay model based forward estimator presented in Section 4.2.1 given by (4.3), and the relative convolutive transfer function (RCTF) based estimator described in Section 4.3. In the multichannel case, the temporal model based estimators are computed for each microphone \( m = \{1, \ldots , M\} \) independently. Following the assumption that the spatial coherence of the reverberation is spatially homogenous, the final multichannel PSD estimate is obtained by averaging all microphone-specific estimates \( \hat{\phi}_s^{(m)}(k,n) \) as [86]

$$\hat{\phi}_s(k,n) = \frac{1}{M} \sum_{m=1}^{M} \hat{\phi}_s^{(m)}(k,n), \quad (5.44)$$

where \( \hat{\phi}_s^{(m)}(k,n) \) is computed as described either in Section 4.2.1 or 4.3 for each microphone \( m \).

5.3 PSD matrix based least-squares method with blocking

In [143] a blocking based noise PSD estimation method was proposed. The proposed reverberation PSD estimator uses the blocking procedure of \( I \) plane waves as described in Section 5.2.1.1
by (5.1) and (5.3) yielding the blocked PSD matrix $\Phi_u(k, n)$ given by (5.4). In [47, 151], the error matrix between the estimated PSD matrix $\hat{\Phi}_u(k, n)$ and its model is defined as

$$\Phi_e = \hat{\Phi}_u - \left[ \hat{\Phi}_v + \phi_r \hat{\Gamma}_r \right], \quad (5.45)$$

Similar as in [47, 157], we assume the real and imaginary elements of $\Phi_e(k, n)$ as independent zero-mean Gaussian random variables with equal variance. This is however not the case for the diagonal elements which are strictly real-valued. Therefore, we define an operator $V$ that creates a vector containing all real elements and all off-diagonal imaginary elements of a complex matrix $\Phi$ of size $M \times M$ as

$$V\{\Phi\} = \left[ \Re\{\Phi(1,1)\}, \Re\{\Phi(p,p')\}, \ldots, \Im\{\Phi(1,2)\}, \Im\{\Phi(q,q')\}, \ldots \right]^T, \quad (5.46)$$

where $p, p' \in \mathbb{N}\{1, \ldots, M\}$ and $q, q' \in \mathbb{N}\{1, \ldots, M\}$ with $q \neq q'$. The column vector $V\{\Phi\}$ is of length $2(M^2 - M)$. Using this operator we define the error vector $V\{\Phi_e(k, n)\}$. The PDF of this error vector can be modeled as a multivariate Gaussian distribution with zero mean and covariance $\sigma^2 I$ as

$$f_V(\{\Phi_e(k, n)\}) = \frac{1}{(\sqrt{2\pi\sigma})^{2M^2 - M}} \exp\left( -\frac{(m - \phi_r n)^T(m - \phi_r n)}{2\sigma^2} \right) \quad (5.47)$$

where $m = V\{\hat{\Phi}_r\}$ and $n = V\{\hat{\Gamma}_r\}$. By maximizing the log-likelihood function $\log(f)$, we obtain the least-squares solution for $n^T n \neq 0$

$$\hat{\phi}_r = \max \left\{ 0, \left( n^T n \right)^{-1} n^T m \right\}, \quad (5.48)$$

where the max\{\cdot\} operation is included to ensure that the estimated PSD is positive also in the presence of estimation errors. Although we excluded the imaginary diagonal elements in this ML solution, we obtain the mathematically equivalent solution derived in [47] as a least-squares problem by minimizing the Frobenius norm of the error matrix, i.e.

$$\hat{\phi}_r = \arg\min_{\phi_r} \| \Phi_e \|_F^2. \quad (5.49)$$

The solution to (5.49) is given by

$$\hat{\phi}_r = \frac{\tr \left\{ \hat{\Gamma}_r^H \left( \hat{\Phi}_u - \hat{\Phi}_v \right) \right\}}{\tr \left\{ \hat{\Gamma}_r^H \hat{\Gamma}_r \right\}}, \quad (5.50)$$

where $\tr \{ \cdot \}$ denotes the trace operator. Note that (5.48) and (5.50) are mathematically equivalent. An estimate of $\Phi_u(k, n)$ can be obtained by using recursive averaging, i.e., $\hat{\Phi}_u(k, n) = \alpha \hat{\Phi}_u(k, n-1) + (1 - \alpha) u(k, n) u^H(k, n)$, where $\alpha$ denotes the forgetting factor.
5.4 Discussion of multichannel reverberation PSD estimators

An overview of the different classes of estimators along with important properties is shown in Table 5.1. The first two columns indicate the section numbers and acronyms of the PSD estimation methods that will be used for future reference. Discriminative properties of the methods are whether they

- exploit a spatial coherence model,
- require prior knowledge of the spatial coherence of the reverberation $\Gamma_r(k),$
- exploit a temporal structure,
- additionally/inherently deliver an estimate of the desired sound PSD $\phi_d(k,n),$
- are online or batch processing methods, and the type of solution (closed-form, iterative, recursive, or polynomial rooting.).

All considered estimators that exploit the spatial coherence model depend on the RTFs $a_i(k,n).$

It is worthwhile to note that the estimator pairs Blocking PSD LS (Section 5.3) and PSD LS (Section 5.2.1.5), and Blocking ML Newton (Section 5.2.1.1) and ML Newton (Section 5.2.1.3), use the same mathematical solution methods, the least squares method or the Newton method, respectively. The main difference is that the former estimators use a blocking of the desired sound, whereas the latter jointly estimate late reverberation and desired sound PSDs.

Whereas the non-blocking spatial coherence based estimators ML Newton, ML EM est. coh., ML EM diff. coh. necessarily produce an estimate of the desired sound PSD $\phi_d(k,n)$ as a byproduct, in the methods Blocking ML root, PSD LS, a method to estimate the desired sound PSD $\phi_d(k,n)$ independently of $\phi_r(k,n)$ is proposed, i.e. it is not necessary to compute $\hat{\phi}_d(k,n)$ to obtain $\hat{\phi}_r(k,n).$ It would be also possible to estimate $\phi_d(k,n)$ from the CDR estimator using an additional step.

All methods except the two ML EM versions are suitable for online processing, i.e. no future information is required to compute $\hat{\phi}_r(k,n)$ at time frame $n.$ The methods Blocking ML
Newton, ML Newton and ML EM use iterative solution methods, while Blocking ML root uses a polynomial rooting solution, which are all computationally highly demanding and much more expensive than the other online solutions. As shown in Section 4.4.3, the recursive RCTF estimator requires a short convergence phase of 1-2 s before providing accurate estimates, whereas the coherence based estimators and the LRSV practically instantaneously deliver useful estimates without delay.

Note that all spatial coherence based methods (Section 5.2.1 and 5.2.2) can also be used to estimate the PSD of non-reverberation related sound fields, e.g. non-stationary diffuse noise like babble noise. However, the spatial coherence based estimators cannot discriminate between reverberation originating from a speech signal and other diffuse sounds, if the reverberation and the other diffuse components have a similar spatial coherence. Methods exploiting temporal structures of reverberation (Exp. fw and RCTF) can discriminate between reverberation and other sound fields. Consequently, this means that these methods are not suitable to estimate the PSD of general diffuse sound fields.

Furthermore, all spatial coherence based methods require prior knowledge or estimates of the RTFs of the desired sound $a(k, n)$, whereas the temporal model based reverberation PSD estimators Exp. fw and RCTF are independent per microphone channel, but require some temporal information like the $T_{60}$ or the CTF length $L_{MA}$.

### 5.5 Bias compensation

As will be shown in Section 5.6, all spatial coherence based estimators under test are severely biased in high direct-to-diffuse ratio (DDR) conditions. Overestimation of the reverberation PSD is especially harmful to the audio quality as it causes speech distortion when used for dereverberation. Therefore, we propose a simple compensation method using the correction factor $c_d(k, n) = f_c(\hat{\phi}_d, \hat{DDR})$ as a function of the estimated DDR. A similar compensation method was proposed in [158] in the context of noise reduction.

As a proof of concept and without claiming optimality, we fit an exponential function to the mean logarithmic PSD estimation error of the three coherence based estimators depending on the estimated DDR as shown in Figure 5.2, where the logarithmic PSD estimation error is defined in Appendix B.2 in (B.8). We approximate the error using the function

$$c_d(DDR) = a \cdot e^{b \log_{10} DDR}, \quad (5.51)$$

where the bias function $c_d$ is obtained in dB and the DDR is estimated by $\hat{DDR} = \hat{\phi}_d - \hat{\phi}_r - \hat{\phi}_v$. By using MATLAB’s `fit()` function, the exponential function (5.51) was fitted to the average error of the three coherence based estimators within the range $\hat{DDR} = [-20, 20]$ dB as shown in Figure 5.2 as an example. The such obtained values are $a = 2.735$ and $b = 0.0928$. Figure 5.2 shows the used error data and the fitted curve $c_d(\hat{DDR})$. 

5.6 Evaluation of reverberation PSD estimators

In this section, we evaluate the performance of the reverberation PSD estimators for single and multiple sources. Sections 5.6.1 - 5.6.2 discuss the used simulation parameters, signal generation and performance measures. Since not all estimators are based on a multi-wave model, in Sections 5.6.3.1 - 5.6.3.4 experiments for a single source are presented and in Sections 5.6.4.1 - 5.6.4.2 experiments for multiple sources are presented. Firstly, in Section 5.6.3.1 we evaluate the PSD estimation performance of stationary diffuse noise PSD using only spatial coherence based estimators, to evaluate their performance when the signals perfectly fit the model. The ML-EM with unknown coherence (Section 5.2.1.4) is omitted from this first evaluation, as in this case the data perfectly fits the coherence model. Secondly, an evaluation using speech signals and simulated RIRs using the image method [131] is conducted in Section 5.6.3.2. Thirdly, in Section 5.6.3.3 and 5.6.3.4, measured RIRs for a single source are used to confirm the results in realistic environments. In Section 5.6.4.1, we investigate the performance of PSD estimators based on a multi-wave model using stationary noise signals that perfectly fit the signal model. In Section 5.6.4.2 the multi-wave estimators are evaluated in the presence of multiple speakers using measured RIRs.

5.6.1 Acoustic setup and simulation parameters

In all simulations and measurements, a uniform circular array with a radius of 4.5 cm and \( M = 6 \) omnidirectional microphones was used. The microphone and source positions for the simulated RIRs are shown in Figure 5.3, where in most experiments only source 1 was active. In Sections 5.6.3.1 and 5.6.4.1 the desired sound components \( S_{d,i}(k,n) \) was generated as a plane wave, while \( r(k,n) \) was generated as a stationary diffuse field. In all other experiments, the reverberant signals were generated by convolving speech signals with simulated or measured RIRs. The RIRs were simulated in a shoebox room of variable size and \( T_{60} \) using the image method [131]. Furthermore, RIRs were measured in a large empty meeting room of size 12.8 ×
10.5 × 3 m with $T_{60} \approx 800$ ms. The source positions relative to the array were similar for the simulated and measured rooms, and are shown in Figure 5.3.

The signals were sampled with a sampling frequency of $f_s = 16$ kHz, and analyzed using an STFT using square-root Hann windows of length 32 ms, a frame shift of $T_{\text{hop}} = 16$ ms, and a $K = 1024$ point discrete Fourier transform (DFT). The noise PSD matrix $\Phi_v(k,n)$ and the RTF vector $a(k,n)$ of the desired sound were assumed to be known in advance.

The noise PSD matrix was computed during speech absence and the steering vector was obtained from the direct sound peak of the RIRs. The recursive smoothing parameter for estimating the PSD matrices was set to $\alpha = e^{-\frac{T_{\text{hop}}}{\tau}}$ with a time constant of $\tau = 50$ ms. Algorithm specific parameters were chosen as follows: $T_{60}(k)$ was set to the fullband reverberation time for each room, $\epsilon = 0.01$, $\epsilon_d = 0.5$, and $\epsilon_r = 0.1$. As found optimal in Section 4.4.2, the CTF length $L_{MA}$ was chosen correspondingly to $0.75T_{60}$, i.e. $L_{MA} = 38$. To match the late reverberation modeled by the spatial coherence model, for the temporal model based estimators we chose the start time of the late reverberation after $D = 1$ frame.

### 5.6.2 Signal generation and performance measures

For a fair evaluation, the definition of the ground truth, i.e. the oracle PSD $\phi_r(k,n)$, is very important as it significantly influences the results. For the temporal model based PSD estimators, the start time of the undesired reverberation measured from the direct sound peak can be user defined by the delay $D$ given in frames. One frame delay for the used STFT parameters corresponds to $T_{\text{hop}} = 16$ ms. For the spatial coherence model, we do not have the freedom to define the start time of the late reverberation: The model rather assumes that all reverberation that is uncorrelated with the direct sound in the current frame has the assumed coherence. To find a fair match between the temporal and spatial coherence model, we define that the oracle late reverberation starts 16 ms after the direct sound peak, and we accordingly use $D = 1$ for the temporal model based estimators.

In Sections 5.6.3.1 and 5.6.4.1, idealized signals are used that fit the spatial coherence model as exact as possible. In these experiments, all signal components were generated using stationary white noise in the time domain: the direct sound components at the reference microphone $S_d(k,n) = \sum_{i=1}^I S_{d,i}(k,n)$ with PSD $\phi_d(k,n)$, the diffuse sound component $r(k,n)$ by imposing the diffuse long-term coherence given by (2.13) on white noise signals using the method proposed in [159], and the additive noise component $v(k,n)$. To obtain the test signals, the three stationary signal components of 10 s length were summed up depending on the input signal-to-noise ratio (iSNR) and the DDR as defined in Appendix A.
In Sections 5.6.3.2 – 5.6.3.4 and 5.6.4.2, the reverberant signals were generated by convolving anechoic speech signals with simulated and measured RIRs. The time-domain representation of the oracle reverberation component for evaluation purposes \( r(k, n) \) was obtained by convolution of anechoic speech signals with windowed RIRs containing only the late part of the reverberation, starting 16 ms after the direct sound peak. The time-domain representation of the additive noise \( v(k, n) \) was pink noise, in order to maintain an approximately constant iSNR per frequency band to the speech. The anechoic speech signals were three male and three female speech signals from [160] with a total length of 2 min.

The steering vectors \( a_i(k, n) \) were computed using the direct part of the RIRs. The time-domain representations of the direct sound signals \( x_d(k, n) \) and the \( i \)-th direct sound source at the reference microphone \( S_{d,i}(k, n) \) were generated by convolving the RIRs with white noise as measurement signal. The steering vectors are then obtained in the least-squares sense by

\[
\hat{a}_i(k, n) = \frac{\sum_{n=1}^{N} x_d(k, n) S_{d,i}^*(k, n)}{\sum_{n=1}^{N} |S_{d,i}(k, n)|^2}. \tag{5.53}
\]

The oracle reverberation PSD is computed as the spatially averaged instantaneous power of the late reverberation, i.e.

\[
\overline{\phi}_r(k, n) = \frac{1}{M} r^H(k, n) r(k, n). \tag{5.54}
\]

For the evaluation of the estimated desired speech signals in Sections 5.6.3.3 and 5.6.3.4, in addition to using the theoretical diffuse coherence given by (2.13), we also used the oracle coherence matrix of the late reverberation \( \Gamma_r(k) \), where the \((m, m')\)-th element was computed by

\[
\Gamma_r^{(m, m')}(k) = \frac{\sum_{n=1}^{N} R^{(m)}(k, n) (R^{(m')}(k, n))^*}{\sqrt{\left(\sum_{n=1}^{N} |R^{(m)}(k, n)|^2\right) \left(\sum_{n=1}^{N} |R^{(m')}(k, n)|^2\right)}}, \tag{5.55}
\]

where \( N \) is the total number of frames.

The oracle desired signal for evaluation purposes is defined as the direct sound at the reference microphone, and was obtained by convolving only the windowed direct path of the reference microphone RIR with the anechoic speech signal.

The estimation accuracy of the various PSD estimators is quantified using the bin-wise logarithmic error introduced in Appendix B.2 by (B.8) that is averaged over all time-frequency bins, where late reverberation is present. In the following figures, the mean is represented by various symbols (+, ⋆, o, ×, etc.) and upper and lower confidence intervals are indicated as the semi-standard deviations by whisker bars. The performance of the MWF computed using the PSD estimates is evaluated using the speech enhancement measures cepstral distance (CD), log–likelihood ratio (LLR), perceptual evaluation of speech quality (PESQ) and frequency-weighted segmental signal-to-interference ratio (fwSSIR) that are described in Appendix B.1.1 requiring
the target signal and the processed signal. Furthermore, we compute the segmental interference reduction (IR) and speech distortion index (SDI) \[77\] as

\[
IR = \frac{1}{T} \sum_{n \in T} 10 \log_{10} \frac{\sum_{t=(n-1)T}^{nT} s_{rv}^2(t)}{\sum_{t=(n-1)T}^{nT} s_{rv,MWF}^2(t)}
\]

\[
SDI = \frac{1}{T} \sum_{n \in T} \frac{\sum_{t=(n-1)T}^{nT} (s_{d,MWF}(t) - s_d(t))^2}{\sum_{t=(n-1)T}^{nT} s_d^2(t)},
\]

(5.56)
(5.57)

where \(s_{rv}(t)\) and \(s_{rv,MWF}(t)\) are the time-domain representations of late reverberation plus noise at the reference microphone and at the MWF output, respectively, \(s_d(t)\) and \(s_{d,MWF}(t)\) are the time-domain representations of direct sound at the reference microphone and the MWF output, \(t\) is the sample index, \(T\) is the number of samples corresponding to a segment of 20 ms and the set \(T\) contains only time segments where speech is active, determined by an ideal voice activity detector.

5.6.3 Performance evaluation for a single source

In this section, the performance of all PSD estimators under restriction of the single-wave model \((I = 1)\) is evaluated in the presence of a single active source.

5.6.3.1 Evaluation of spatial coherence based PSD estimators for stationary diffuse noise

For a stationary diffuse field the results of the logarithmic PSD estimation error is shown in Figure 5.4 over varying iSNRs with a fixed DDR at 10 dB. It can be observed that at higher iSNRs, all estimators have a small variance, and their error means are close to 0 dB, which means a small estimation error. Blocking ML Newton is performing slightly worse in medium iSNRs compared to the other estimators due to a small increase in its variance. At low iSNRs the estimators Blocking ML root, Blocking ML Newton, and ML Newton are the most robust as mainly their variance increases, but the mean stays close to 0 dB. ML EM and CDR are very robust in terms of their variance in low iSNRs, but their mean increases, so they become biased to positive values.

![Figure 5.4: Mean and standard deviation of log error for artificial stationary sound field with DDR = 10 dB.](image-url)
5.6. EVALUATION OF REVERBERATION PSD ESTIMATORS

Figure 5.5: Mean and standard deviation of log error for artificial stationary sound field with iSNR = 15 dB.

Figure 5.5 shows the dependency of the log error on the DDR, while keeping the iSNR fixed at 15 dB. All coherence based estimators show a large error at high positive DDRs. This can be explained by the fact that if the direct sound component dominates the observed signal, it becomes difficult to accurately estimate the comparatively weak diffuse sound component.

Figures 5.6 and 5.7 show the influence of DOA estimation errors for two DDRs (25 dB and 0 dB, respectively) at a iSNR of 15 dB. In this simulation, the steering vector $\mathbf{a}(k,n)$ was computed with an angular offset, while the actual source position was kept constant. From these results it can be observed that an offset in the steered DOA increases the variance and the mean shifts to positive values (overestimation). At high DDR, the impact of steering errors is very prominent (see Figure 5.6), whereas at DDR = 0 dB and lower, the DOA error influence is minor (Figure 5.7). However, fortunately, at high DRRs the underlying assumptions of the typical signal models used for steering vector estimation are matched more accurately. This can mitigate the influence of steering vector estimation errors at high DRR in practice.

Figure 5.6: Mean and standard deviation of log error for artificial stationary diffuse field for DOA steering offset with iSNR = 15 dB and DDR = 25 dB.

Figure 5.7: Mean and standard deviation of log error for artificial stationary diffuse field for DOA steering offset with iSNR = 15 dB and DDR = 0 dB.
In this section we showed that the coherence based PSD estimators work well at high iSNR, low DDR, and low steering errors, but they exhibit weaknesses in low iSNR, high DDR, and in the presence of steering errors.

5.6.3.2 Evaluation of PSD estimators for late reverberation

Figure 5.8 shows the log error obtained using simulated RIRs in a shoebox room with \( T_{60} = 0.6 \) s and a source distance of 2.5 m for varying iSNR. In contrast to Section 5.6.3.1, this experiment includes also the temporal model based PSD estimators. The trends and the relative behavior between the coherence based estimators are the same as in the controlled stationary experiment in Figure 5.4, but the variances are larger and the means are slightly biased to positive values due to model mismatches. The temporal model based estimator RCTF shows a different behavior with a lower mean log error compared to the spatial coherence based estimators. Therefore, the RCTF estimator yields less overestimation, which is beneficial in terms of speech distortion. The Blocking ML Newton, BF LCMV, ML Newton and RCTF estimators are more robust against overestimation at low iSNRs while the Blocking PSD LS, PSD LS, and CDR are the most sensitive in low iSNR. The error of the ML EM with diffuse coherence and the ML EM with estimated coherence yield the same variance, but surprisingly, the mean of ML EM with diffuse coherence is slightly closer to zero dB.

Figure 5.9 shows the log error for varying \( T_{60} \) and a constant source distance at iSNR of 15 dB. We can observe that the log error decreases towards higher reverberation times. The reason is
that the DRR increases for decreasing $T_{60}$, such that the trends are similar as in Figure 5.5. If the distance would be varied with the $T_{60}$ such that the DRR stays constant, the error distributions would remain also approximately constant.

In Figure 5.10, the mean log error is grouped depending on the true bin-wise (local) DRR in steps of 5 dB using the data from all iSNR conditions shown in Figure 5.8. It is interesting to note that the temporal model based estimators Exp. fw and RCTF yield a different behavior than the coherence based PSD estimators. The RCTF estimator is more robust against overestimation at high DRRs, which we expect to result in less speech distortion.

The trends shown in the previous section for the coherence based estimators can be confirmed when used to estimate the late reverberation PSD. The RCTF estimator yields more underestimation and less overestimation than all other tested estimators.

### 5.6.3.3 Performance of the spatial filter using the late reverberation PSD estimates

In this experiment, the performance of the MWF using the various PSD estimates is evaluated using measured impulse responses (c.f. Section 5.6.1). As there are less significant differences between the spatial coherence based estimators observable at higher iSNRs (c.f. Figure 5.8), we present results here only for iSNR = 15 dB. In this experiment, the oracle late reverberation PSD is used either with the diffuse coherence (2.13) or with the oracle late reverberation coherence (5.55) to investigate the mismatch effect of the diffuse field model.

Figure 5.11 shows the IR vs. the SDI as computed by (5.56) and (5.57). The optimal point lies in the lower right corner. The best performance is obtained by using the oracle PSD with oracle reverberation coherence, which has a clear advantage over the oracle PSD with theoretical diffuse coherence matrix. The estimators performing closest to the oracle PSD are the temporal model-based methods LRSV and CTF. Among the coherence-based methods, the PSD LS, ML EM diff. coh., ML Newton and Blocking ML root perform slightly better than Blocking PSD LS, BF LCMV and CDR. The Blocking ML Newton has a lower SDI at the price of less IR, while the ML EM est. coh. surprisingly performs worse with a low IR.

The improvement of CD, LLR, PESQ and fwSSIR compared to the unprocessed reference microphone is shown in Figure 5.12 (higher values are better). The oracle PSD with oracle reverberation coherence yields the best results, followed by the oracle PSD with theoretical diffuse coherence. Most estimators yield rather close results, while the Blocking ML Newton and the ML EM variants show less improvement. Interestingly, for the considered scenario, the ML EM with estimated coherence yields much less interference reduction and lower improvement of perceptual measures compared to using the ML EM with diffuse coherence. This stands in
contrast to the results achieved using the oracle reverberation coherence, and means that the estimated coherence by the ML EM is not accurate enough, which even deteriorates the results.

The best performing estimators in terms of the perceptually motivated speech enhancement measures shown in Figure 5.12 are Blocking PSD LS, Blocking ML root, BF LCMV, ML Newton, PSD LS, CDR and RCTF. The differences between the objective measures of these estimators are minimal and not significant. By comparing Figures 5.11 and 5.12, it can be observed that estimators with higher IR yield higher PESQ and fwSSIR improvements, whereas estimators with less speech distortion yield slightly better CD and LLR improvements.

Note that the available objective measures in Figure 5.12 can predict the reverberation reduction quite reliably, but are only moderately related to the subjective signal quality. The most reliable measures of the signal quality are the CD and LLR as shown in Section 7.1.3 and [20]. However, also the speech distortion shown in Figure 5.11 is an additional indicator for the quality. Informal subjective listening tests confirmed that some estimators produce very similar results, while others can sound very different\(^1\) as represented in Figure 5.12. Perceptual differences between the estimators are more prominent at lower iSNRs, while at iSNR > 25 dB they become perceptually almost indistinguishable (see Figure 5.8). The tradeoff between speech

\(^1\)Sound examples can be found at www.audiolabs-erlangen.de/resources/2017-COMPARE-PSD-ESTIMATORS
5.6. EVALUATION OF REVERBERATION PSD ESTIMATORS

Figure 5.13: Speech distortion vs. interference reduction without and with bias compensation, iSNR = 15 dB

Figure 5.14: Improvement of perceptual measures without and with bias compensation, iSNR = 15 dB

distortion and interference reduction (see Figure 5.11) is clearly audible, which can be a guide on which estimator to choose depending on subjective preference and application.

5.6.3.4 Effect of bias compensation for coherence based estimators

In this section, we evaluate the bias compensation method for selected coherence-based estimators. The bias compensation function shown in Figure 5.5 was trained using RIRs simulated by the image method [131], while for the following evaluation, the measured RIRs with different speech and noise data were used.

The results for some selected coherence based estimators without and with the proposed bias compensation are shown in Figures 5.13 and 5.14. We can see in Figure 5.13 that the bias compensation method proposed in Section 5.5 greatly reduces the speech distortion close to the speech distortion obtained using the oracle PSD, while sacrificing only a small amount of interference reduction. The CD, LLR, PESQ and fwSSIR shown in Figure 5.14 can be improved by using bias compensation. Informal subjective listening tests confirmed that the speech distortion can be reduced by the proposed bias compensation method.

5.6.4 Evaluation for the multi-wave model

In the presence of multiple sources, we assume a general number of plane waves $I$ per frequency bin. Only the blocking based reverberation PSD estimators presented in Sections 5.2.1.1 and 5.3 are based on the general sound field model with an arbitrary number of plane waves $I$. In this section, the proposed Blocking PSD least-squares estimator (Section 5.3) and the diffuse beamformer ($BF LCMV$, Section 5.2.1.2) are compared by modeling multiple plane waves $I$. For
reference, also show results for some temporal model based estimators presented in Section 4 are shown, although the underlying models assume only a single active source.

In Section 5.6.4.1 we simulate multiple stationary plane wave signals propagating in a stationary diffuse noise field, and aim at estimating the diffuse noise PSD, similar as in Section 5.6.3.1. In Section 5.6.4.2, we investigate the performance of the PSD estimators in the presence of multiple speakers using estimated DOAs in a scenario multiple speakers and measured RIRs.

### 5.6.4.1 Stationary diffuse field with multiple plane waves

The performance of the *PSD blocking LS* and the *BF LCMV* estimators is first evaluated in an ideal scenario with stationary signals matching the signal model, similar as in Section 5.6.3.1: We generated three stationary plane wave signals arriving from different directions with equal power as white noise signals. The source positions are shown in Figure 5.3. These plane waves are propagating in a stationary diffuse noise field generated using the method proposed in [161]. Uncorrelated sensor noise was added with iSNR = 50 dB. The directions of the sources $\theta_i$, and therefore the matrix $A(k,n)$, are known in this experiment. The two PSD estimators are computed by successively taking one, two or all three known DOAs into account. A fourth DOA is additionally taken into account pointing to a direction, where no source is active, having the maximum angular distance to all active DOAs.

Figure 5.15 shows the logarithmic PSD estimation error for successively taking $I = \{1, \ldots, 4\}$ directions into account. We can observe that the error decreases if the first $I = 3$ directions, where sources are active, are taken into account. By additionally blocking the fourth direction, where no source is active, the mean error does not change significantly, but the variance increases slightly. The variance increase is larger for the *BF LCMV* estimator than for *PSD blocking LS*. For the *BF LCMV*, the variance increase from $I = 3$ to $I = 4$ can lie in the reduced degrees of freedom to minimize the noise. Similarly, by blocking $I = 4$ instead of $I = 3$ directions in the *PSD blocking LS*, the noise at the output of the blocking matrix can increase.

### 5.6.4.2 Multiple speakers in reverberant environment

In this experiment, two simultaneously active speakers at position 1 and 2 in Figure 5.3 are simulated using measured RIRs in the room with $T_{60} = 800$ ms. The DOAs are estimated using
5.6. EVALUATION OF REVERBERATION PSD ESTIMATORS

![Graph showing log error vs. iSNR for different PSD estimators.]

**Figure 5.16:** Reverberation PSD estimation error for two simultaneously active speakers in a measured room with $T_{60} = 800$ ms.

![Graph comparing oracle PSD, Blocking PSD LS, BF LCMV, Exp. fw, and RCTF.]

**Figure 5.17:** Speech enhancement results using the MWF for two active speakers in a measured room with $T_{60} = 800$ ms and iSNR = 15 dB.

beamspace root-MUSIC [111]. Either $I = 1$ or $I = 2$ narrowband DOAs are estimated and used to construct the propagation matrix $A(k, n)$.

Figure 5.16 shows the logarithmic PSD estimation error for the Blocking PSD LS and BF LCMV estimators with $I = \{1, 2\}$, and the temporal model based estimators Exp. fw and RCTF. We can observe that the mean of the log estimation error shifts towards 0 dB at higher iSNRs for the Blocking PSD LS and BF LCMV estimators for increasing $I$. In analogy to Figure 5.4, the BF LCMV estimator yields a larger variance than the Blocking PSD LS estimator. The temporal model based estimators Exp. fw and RCTF yield a bias in the mean error and have a slightly larger variance compared to the spatial coherence based estimators due to the model mismatch regarding the number of sources. The performance gain in Figure 5.16 by blocking $I = 2$ instead of $I = 1$ DOAs is rather small. Because of the spectral sparsity of the speech, also estimating $I = 1$ DOA per frequency can achieve good results in the presence of multiple speakers.

The objective measures for the MWF using either one or both DOA estimates are shown in Figure 5.17. The MWF using the temporal model based PSD estimators Exp. fw and RCTF was computed with $I = 2$ estimated DOAs. We can observe that all objective measures except fwSSIR improve by using $I = 2$ instead of $I = 1$. The proposed Blocking PSD LS estimator for $I = 2$ yields the best results for CD, LLR, and fwSSIR, while the PESQ improvement is on par with the RCTF estimator. It is interesting to note that although the performance of the temporal model based estimators Exp. fw and RCTF degrades in the presence of multiple
sources, their performance is still close to the coherence based estimators in terms of LLR and PESQ as shown in Figure 5.17.

5.7 Summary

In this chapter, we provided an extensive overview and categorization of multichannel reverberation PSD estimators, proposed a reverberation PSD estimator, discussed advantages and disadvantages, and provided an extensive experimental comparison. A major class are the spatial coherence based estimators using the same signal model and assumptions as the MWF for dereverberation and noise reduction presented in Section 3.2. For the tested scenarios, the temporal model based estimators using optimal parameters yielded the best tradeoff between speech distortion and interference reduction. The spatial coherence based estimators suffer from a tendency to overestimate the reverberation PSD in high DRR, while differences between most coherence based estimators were minor. However, the ML EM and the Blocking ML Newton estimators performed below average. Computationally complex iterative solutions such as the Newton or ML algorithms yield results below average in some conditions, as they can be trapped in local minima, and strongly depend on the initialization. Some estimators have different strengths and weaknesses in terms of robustness against noise, or the tradeoff between speech distortion and interference reduction.

We proposed a simple mapping as a cure for the common weakness of a bias towards overestimation in high DRR. Using this compensation, the coherence based estimators can yield similar results to the temporal model based estimators. Furthermore, while the signal model of the spatial coherence based estimators can be generalized, such that they can be used to estimate the PSD of any noise field with a known stationary coherence, the temporal model based estimators match temporal characteristics of reverberation and show a different behavior. Furthermore, it was shown that the performance in the presence of multiple speakers can be improved for the proposed spatial coherence based PSD estimator as well as other estimators based on a multi-wave signal model.

Due to the insignificant performance differences in most experiments, besides the temporal model estimators, which depend on additional parameters like the $T_{60}$ or the CTF length, a clear preference for any spatial coherence based PSD estimator is hard to suggest. However, by taking performance, complexity and processing delay into account, the PSD LS would be a reasonable choice, whereas in the presence of multiple sources, the blocking based estimators BF LCMV or Blocking PSD LS may be favored choices.
The dereverberation methods presented in this chapter are based on the narrowband multichannel autoregressive (MAR) model discussed in Section 2.4. The first methods based on this model were multichannel linear prediction based approaches in the short-time Fourier transform (STFT) domain [33,35,90]. In this framework, the late reverberation at each microphone in the current frame is predicted using past frames of all reverberant signals using the MAR prediction coefficients, often termed as room regression coefficients. The estimated late reverberation signal is then subtracted from the current reverberant signal. In [34], a generalized framework for the multichannel linear prediction methods called blind impulse response shortening was presented.

Advantages of methods based on the MAR model are that they are valid for multiple sources, they directly estimate a dereverberation filter of finite length, the required filters are relatively short, and these methods are suitable as pre-processing techniques for further multichannel processing such as spatial filtering. A great challenge of the MAR signal model is the integration of additive noise, which has to be removed in advance [40,90] without destroying the autoregressive relation between neighboring frames of the reverberant signal. As the first solutions based on the multichannel linear prediction framework were batch algorithms, further efforts have been made to develop online algorithms, which are suitable for real-time processing [91–93,162,163]. However, the reduction of additive noise in an online solution using the MAR model has been considered only in [92] to the best of our knowledge.

After introducing the signal model in Section 6.1, we propose to model the MAR coefficients by a first-order Markov model as described in Section 6.2. In contrast to the deterministic modeling of the MAR coefficients used in the state-of-the-art methods, the stochastic model is suitable for time-varying acoustic environments. In Section 6.3 we derive a Kalman filter to estimate the MAR coefficients modeled by the first-order Markov process by assuming no additive noise at the microphones. It is shown that this proposed method is a generalization of state-of-the-art recursive least-squares (RLS) methods [91–93,162,163]. By also considering additive noise in Section 6.4, a method for dereverberation and noise reduction using two alternating Kalman filters is proposed, which also uses the first-order Markov model for the MAR coefficients. This method is a generalization of the single Kalman filter proposed in Section 6.3 to the noisy case.
The proposed solution has several advantages to state-of-the-art solutions: Firstly, in contrast to the sequential signal and autoregressive (AR) parameter estimation methods used for noise reduction presented in [164,165], we propose a parallel estimation structure to estimate the MAR coefficients and the noise-free reverberant signals. This parallel structure allows a fully causal estimation chain as opposed to state-of-the-art sequential structures, where the noise reduction stage relies on outdated MAR coefficients [92]. Secondly, in the proposed method we assume randomly time-varying MAR coefficients, instead of a time-invariant linear filter and a time-varying non-linear filter like in the expectation-maximization (EM) based algorithm proposed in [92]. Thirdly, our proposed algorithm is an adaptive algorithm that converges over time, and does not require multiple iterative computation steps per time frame in contrast to [92]. As an optional extension, in Section 6.5 we propose a method to independently control the amount of residual reverberation and noise at the output for the proposed framework. After a short discussion in Section 6.6, the proposed method is evaluated and compared to state-of-the-art methods in Section 6.7, and the chapter is summarized in Section 6.8.

6.1 Signal model and problem formulation

In this section, the model for the noisy microphone signals with the MAR reverberation model is introduced, and the problem to obtain the desired early speech signal is formulated.

6.1.1 Multichannel autoregressive reverberation model with additive noise

We assume an array of $M$ microphones with arbitrary directivity and arbitrary geometry. The multichannel microphone signals are given in the STFT domain by $Y_m(k,n)$ for $m \in \{1, \ldots, M\}$, where $k$ and $n$ denote the frequency and time indices, respectively. In vector notation, the microphone signals can be written as $y(k,n) = [Y_1(k,n), \ldots, Y_M(k,n)]^T$. Using the MAR reverberation model given by (2.4) in Section 2.4, the microphone signals consisting of the reverberant speech and additive noise are given by

$$y(k,n) = \sum_{\ell=\bar{D}}^{L_{AR}} C_\ell(k,n)x(k,n-\ell) + s(k,n) + v(k,n), \quad (6.1)$$

where the vector $x(k,n)$ contains the reverberant speech at each microphone, $v(k,n)$ is additive noise, the vector $s(k,n) = [S^{(1)}_D(k,n), \ldots, S^{(M)}_D(k,n)]^T$ contains the desired early speech at each microphone $S^{(m)}_D(k,n)$, and the $M \times M$ matrices $C_\ell(k,n)$, $\ell \in \{\bar{D}, \bar{D}+1, \ldots, L_{AR}\}$ contain the MAR coefficients predicting the late reverberation component $r(k,n)$ from past frames of $x(k,n)$. The desired early speech vector $s(k,n)$ is the innovation in the autoregressive process $x(k,n)$. The innovation $s(k,n)$ is also known as the prediction error in the linear prediction terminology. The choice of the delay $D \geq 1$ determines how many early reflections are contained in the early speech signals $s(k,n)$, and should be chosen depending on the amount of overlap between STFT frames, such that there is little to no correlation between the direct sound contained in $s(k,n)$
6.2. STATE-SPACE MODELING OF STOCHASTIC MAR COEFFICIENTS

and the late reverberation \( r(k, n) \). The length \( L_{AR} > D \) is the number of past frames that are used to predict the reverberant signal.

For simpler handling of equations, we introduce a matrix notation to describe the observed signal vector (6.1). For the sake of a more compact notation, the frequency index \( k \) is omitted in the remainder of this chapter. Let us first define the quantities

\[
X(n) = I_M \otimes \begin{bmatrix} x^T(n - L_{AR} + D) & \ldots & x^T(n) \end{bmatrix}
\]

(6.2)

\[
c(n) = \text{Vec} \left\{ \begin{bmatrix} C_{L_{AR}}(n) & \ldots & C_D(n) \end{bmatrix}^T \right\},
\]

(6.3)

where \( I_M \) is the \( M \times M \) identity matrix, \( \otimes \) denotes the Kronecker product, and the operator \( \text{Vec}\{\cdot\} \) stacks the columns of a matrix sequentially into a vector. Consequently, \( c(n) \) is a column vector of length \( L_c = M^2(L_{AR} - D + 1) \) and \( X(n) \) is a sparse matrix of size \( M \times L_c \). Using the definitions (6.2) and (6.3) with the signal model (6.1), the observed signal vector is given by

\[
y(n) = X(n - D)c(n) + s(n) + v(n).
\]

6.1.2 Problem formulation

If estimates of the reverberant signal vector \( x(n) \) and MAR coefficients \( c(n) \) can be obtained, the early speech signals can be estimated by applying the estimated MAR coefficients as a finite multiple-input-multiple-output (MIMO) filter to the estimated reverberant signals \( \hat{x}(n) \), i.e.

\[
\hat{s}(n) = \hat{x}(n) - \overbrace{X(n - D)\hat{c}(n)}^{r(n)}.
\]

(6.5)

The methods presented in Section 6.3 assume that there is no additive noise \( v(n) \) such that (6.5) can be computed directly using the microphone signals with \( \hat{x}(n) = y(n) \). The methods presented in Section 6.4 try obtain the estimated reverberant speech \( \hat{x}(n) \) by removing the noise from \( y(n) \).

6.2 State-space modeling of stochastic MAR coefficients

In previous works [35,90,166], the MAR coefficients were assumed to be stationary, which does not hold in time-varying acoustic scenarios. Furthermore, preliminary experiments have shown that the MAR coefficients \( c(n) \) can be time-varying even in spatially stationary scenarios due to model errors of the STFT domain model (6.1).

To model the uncertainty of the MAR coefficient vector across time, we assume that \( c(n) \) consists of independent complex random variables. A similar model for acoustic filter coefficients has been successfully used in the context of echo cancellation [167] and dereverberation using a finite impulse response (FIR) model [30]. We describe the MAR coefficient state vector by a first-order Markov model

\[
c(n) = A(n) c(n - 1) + w(n),
\]

(6.6)
where the matrix $A(n)$ models the propagation of the state vector across time and $w(n) \sim \mathcal{N}(0_{L_c \times 1}, \Phi_w(n))$ is a zero-mean complex Gaussian perturbation noise with covariance matrix $\Phi_w(n)$. Further, we assume that $w(n)$ and $s(n)$ are mutually uncorrelated. Note that the noise $w(n)$ determines the uncertainty of $c(n)$ across time, i.e. when the noise is zero, the MAR coefficients are stationary.

### 6.3 MAR online dereverberation in the noise-free case

In this section, we show how the MAR coefficients can be estimated online in the noise-free case. In Section 6.3.1, the signal model is re-formulated without additive noise. After reviewing the state-of-the-art solution based on RLS in Section 6.3.2, a generalized Kalman filter solution using the stochastic state-space model introduced in Section 6.2 is proposed. In Section 6.3.4 parameter estimators for the Kalman filter are presented and in Section 6.3.5 the relation to the RLS algorithm is discussed.

**6.3.1 Noise-free signal model**

We assume that the microphone signals contain no additive noise, i.e. $v(k,n) = 0_{M \times 1}$. As a consequence, the microphone signal vector $y(k,n)$ is given by

$$y(n) = Y(n-D)c(n) + s(n),$$

where the matrix $Y(n)$ of size $M \times L_c$ is similarly defined as (6.2) by

$$Y(n) = I_M \otimes [y^T(n-L_{AR}+D), \ldots, y^T(n)],$$

and $c(n)$ is given by (6.3). The MAR coefficient vector $c(n)$ is described by the first-order Markov model given in (6.6).

We assume that the desired early speech vector $s(n) \sim \mathcal{N}(0_{M \times 1}, \Phi_s(n))$ is a circularly complex zero-mean Gaussian random variable with the respective covariance matrix $\Phi_s(n) = \mathcal{E}\{s(n)s^\dagger(n)\}$. Furthermore, we assume that $s(n)$ is uncorrelated across time. These assumptions hold well for the STFT coefficients of non-reverberant speech that typically have short to moderate temporal correlation in the time domain, and are widely used in speech processing methods [29, 132, 133].

According to (6.5), an estimate of the desired signal vector is obtained in the noise-free case by a linear MIMO filtering operation of the microphone signals, i.e.

$$\hat{s}(n) = y(n) - Y(n-D)\hat{c}(n).$$

The assumed noise-free signal model is depicted in Figure 6.1. The MAR coefficients can be estimated using the RLS algorithm as shortly reviewed in Section 6.3.2, or a Kalman filter can be used as will be proposed in Section 6.3.3.
6.3. MAR ONLINE DEREVERBERATION IN THE NOISE-FREE CASE

The figure shows a multichannel autoregressive signal model assuming no noise with Markov modeling of the regression coefficients.

6.3.2 State-of-the-art RLS estimation of the MAR coefficients

By assuming the MAR coefficients as a deterministic variable instead of using the statistical model described in Section 6.2, the MAR coefficients can be estimated in the minimum mean square error (MMSE) sense by

$$\hat{c}(n) = \arg\min_c E\{\|y(n) - Y(n-D)c(n)\|^2\}.$$  (6.10)

A recursive solution to the problem given in (6.10) can be obtained using the RLS algorithm as proposed in [39, 91, 93]. The RLS algorithm is a special case of the Kalman filter that is presented in the next section, under the condition that the spatial covariance matrix $\Phi_s(n)$ is a scaled identity matrix. This would imply that the early speech signal at the microphones $s(n)$ is spatially uncorrelated. Considering typical compact microphone arrays, where the direct sound and early reflections arrive within the same time frame at the microphones, this is a very coarse and unrealistic assumption. The relation between the RLS and the Kalman filter is discussed in Section 6.3.5.

In [39] the power spectral density (PSD) of the early speech contained on the main diagonal of $\Phi_s(n)$ is approximated by the reverberant input PSD, whereas in [91, 93] a spectral subtraction based estimator for the early speech PSD that is based on the exponential decay model that was introduced in Section 2.5. The used estimator based on the exponential decay model was already reviewed in Section 4.2.1. Note that this estimator requires knowledge of the reverberation time and in some cases also the direct-to-reverberation ratio (DRR). In Section 6.3.4.2, we derive a maximum likelihood (ML) estimator for the early speech covariance matrix $\Phi_s(n)$ that can not only be used for the proposed Kalman filter, but also for the RLS algorithm. In [39] an additional control mechanism called speaker change detection (SCD) is proposed to keep the RLS stable when the acoustic scene changes. In [91] no such control mechanism is used, but the MAR coefficients are estimated by using an additional delay to avoid instability issues in dynamic scenarios.
6.3.3 Kalman filter to estimate MAR coefficients

By considering (6.7) and (6.6) as observation and state equations, respectively, an optimal estimate of the MAR coefficient vector is obtained by minimizing the mean-squared error (MSE)

\[ \mathcal{E}\{|c(n) - \hat{c}(n)|^2\}. \quad (6.11) \]

The covariance matrix of the state vector error is defined as

\[ \Phi_{\Delta c}(n) = \mathcal{E}\{|c(n) - \hat{c}(n)||c(n) - \hat{c}(n)|^H\}. \quad (6.12) \]

We denote the Kronecker-Delta function by \( \delta(n) \). Given that we have a dynamic system described by the state model (6.6), the observation model (6.7), and that the assumptions

\[
\begin{align*}
    w(n) &\sim \mathcal{N}(0_{L_c \times 1}, \Phi_w(n)) \quad (6.13) \\
    s(n) &\sim \mathcal{N}(0_{M \times 1}, \Phi_s(n)) \quad (6.14) \\
    \mathcal{E}\{w(n)w^H(n-j)\} &= \Phi_w(n)\delta(n-j) \quad (6.15) \\
    \mathcal{E}\{s(n)s^H(n-j)\} &= \Phi_s(n)\delta(n-j) \quad (6.16) \\
    \mathcal{E}\{w(n)s^H(n-j)\} &= 0_{L_c \times M}, \quad (6.17)
\end{align*}
\]

are fulfilled, the well-known Kalman filter [134] is a suitable recursive estimator of \( c(n) \) that minimizes (6.11) given all currently available observations \( \{y(0), y(1), \ldots, y(n)\} \) for \( n \to N \), where \( N \gg 0 \). The Kalman filter prediction and update equations are given by

\[
\begin{align*}
    \hat{\Phi}_{\Delta c}(n|n-1) &= A(n)\hat{\Phi}_{\Delta c}(n-1)A^H(n) + \Phi_w(n) \quad (6.18) \\
    \hat{c}(n|n-1) &= A(n)\hat{c}(n-1) + \Phi_w(n) \quad (6.19) \\
    e(n) &= y(n) - Y(n-D)\hat{c}(n|n-1) \quad (6.20) \\
    K(n) &= \hat{\Phi}_{\Delta c}(n|n-1)Y^H(n-D) \left[ Y(n-D)\hat{\Phi}_{\Delta c}(n|n-1)Y^H(n-D) + \Phi_s(n) \right]^{-1} \quad (6.21) \\
    \hat{\Phi}_{\Delta c}(n) &= [I_{L_c} - K(n)Y(n-D)]\hat{\Phi}_{\Delta c}(n|n-1) \quad (6.22) \\
    \hat{c}(n) &= \hat{c}(n|n-1) + K(n)e(n), \quad (6.23)
\end{align*}
\]

where the vector \( e(n) \) is called the prediction error and the matrix \( K(n) \) is called the Kalman gain, which minimizes the trace of the error matrix \( \Phi_{\Delta c}(n) \). If the state of the Kalman filter at \( n = 0 \) is unknown, we propose to initialize with the values \( \hat{\Phi}_{\Delta c}(0) = I_{L_c} \) and \( \hat{c}(0) = 0_{L_c \times 1} \).

By reviewing (6.9) and (6.20), we observe that the prediction error \( e(n) \) is an estimate of the desired signals \( s(n) \) given the predicted MAR coefficients \( \hat{c}(n|n-1) \).

6.3.4 Parameter estimation for Kalman filter

The Kalman filter requires the covariance of the state vector perturbation noise \( \Phi_w(n) \), the state propagation matrix \( A(n) \), and the covariance matrix \( \Phi_s(n) \). In following we propose suitable estimators for these generally time-varying parameters.
6.3.4.1 Dynamic state modeling

Similarly as in [168] we want to model slowly time-varying acoustic conditions by the first-order Markov model given by (6.6). As the transitions of the MAR coefficients over time are unknown, we choose $A(n) = I_{Lc}$. By assuming the elements of $w(n)$ uncorrelated and identically distributed, the uncertainty of the MAR coefficients is determined by the scalar noise variance $\phi_w(n)$, such that $\Phi_{w}(n) = \phi_w(n)I_{Lc}$. We propose to estimate the variance $\phi_w(n)$ depending on the change of the MAR coefficients between subsequent frames using

$$\hat{\phi}_w(n) = \frac{1}{L_c} \mathcal{E}\{\|\hat{c}(n) - \hat{c}(n-1)\|_2^2\} + \eta,$$

where $\eta$ is a small positive number to model the continuous variability of the MAR coefficients if the difference between subsequent estimated coefficients is zero.

6.3.4.2 Estimation of the spatial covariance matrix

The estimation accuracy of the spatial covariance matrix $\Phi_s(n)$ greatly effects the quality of the estimated desired signals $\hat{s}(n)$. Therefore, we derive an ML estimator and an approximation thereof, that can be computed given the currently available data.

According to the signal model (6.7) only the last $L_{AR}$ observations are required to model the current observed signal. The conditional probability distribution of $y(n)$ at frame $n$ is $f\{y(n) | \Theta(n)\}$, where we define the parameter set $\Theta(n) = \{y(n), y(n-1), \ldots, y(n-L_{AR}), c(n)\}$. While the unconditioned probability distribution of $y(n)$ is dependent on previous time frames, the conditional probability given the previous $L_{AR}$ time frames is independent. The conditional mean and covariance of $y(n)$ given $\Theta(n)$ are

$$\mu_{y|\Theta} = \mathcal{E}\{y(n) | \Theta(n)\} = Y(n-D)c(n) \quad (6.25)$$

and

$$\mathcal{E}\{[y(n) - \mu_{y|\Theta}] [y(n) - \mu_{y|\Theta}]^H | \Theta(n)\} = \Phi_s(n), \quad (6.26)$$

respectively. As the desired signal vector $s(n)$ is Gaussian distributed, the probability density function (PDF) of $y(n)$ is a Gaussian conditional likelihood function with mean (6.25) and covariance (6.26) and is given by

$$f \{y(n) | \Theta(n)\} = \frac{1}{\pi^M |\Phi_s(n)|} \exp[-(y(n)-Y(n-D)c(n))^H \Phi_s^{-1}(n) [y(n)-Y(n-D)c(n)]]. \quad (6.27)$$

By assuming short-time stationarity of the PDF of $y(n)$ within $N$ frames, and as the conditioned PDF is independent across time, the joint PDF at frame $n$ is given by

$$f \{y(n-N+1), \ldots, y(n)\} = \prod_{\ell=n-N+1}^{n} f \{y(\ell) | \Theta(\ell)\}. \quad (6.28)$$
By maximizing the log-likelihood function of (6.28) subject to \( \Phi_s(n) \), the covariance at frame \( n \) is given by

\[
\Phi_{s,\text{ML}}(n) = \frac{1}{N} \sum_{\ell=n-N+1}^{n} [y(\ell) - Y(\ell - D)c(\ell)] [y(\ell) - Y(\ell - D)c(\ell)]^H. \tag{6.29}
\]

We can compute (6.29) using the estimates \( \{\hat{c}(n-N+1), \ldots, \hat{c}(n-1)\} \). As we do not have yet the estimate \( \hat{c}(n) \) at frame \( n \) available when we need to compute (6.29), we approximate it by its predicted estimate \( \hat{c}(n) \approx c(n|n-1) \) in (6.29). Under this assumption and by using (6.9) and (6.20), an estimate of (6.29) is given by

\[
\hat{\Phi}_{s,\text{ML}}(n) = \frac{1}{N} \left( \sum_{\ell=n-N+1}^{n-1} \hat{s}(\ell)\hat{s}^H(\ell) + e(n)e^H(n) \right). \tag{6.30}
\]

For practical reasons we replace the arithmetic average in (6.30) by an exponentially weighted average, such that we can compute the estimate \( \hat{\Phi}_{s,\text{ML}}(n) \) recursively, i.e.,

\[
\hat{\Phi}_s(n) = \alpha \hat{\Phi}_{s,\text{POS}}(n-1) + (1-\alpha)e(n)e^H(n), \tag{6.31}
\]

where the recursive a posteriori ML estimate of \( \Phi_{s,\text{ML}}(n) \) given \( \hat{s}(n) \) is given by

\[
\hat{\Phi}_{s,\text{POS}}(n) = \alpha \hat{\Phi}_{s,\text{POS}}(n-1) + (1-\alpha)\hat{s}(n)\hat{s}^H(n), \tag{6.32}
\]

and \( \alpha \) is an exponential recursive smoothing factor with a time constant corresponding to the duration of \( N \) frames.

Although it is proposed in some solutions in [34] that the spatial covariance matrix \( \Phi_s(n) \) is diagonal, this is typically not the case due to the strong inter-microphone coherence of the desired signals. If only a single sound source is present, this matrix is of rank one.

### 6.3.5 Relation to the RLS algorithm

It is well known that the RLS algorithm has a strong similarity to the Kalman filter [169], although they are significantly different from the theoretical point of view. The equations of the proposed Kalman filter (6.18) – (6.23) become identical to the RLS algorithm proposed in [91], if \( \Phi_s(n) \) is assumed to be a scaled identity matrix, \( A(n) = I_{Lc}, \ w(n) = 0_{Lc \times 1} \), and equation (6.18) changes to

\[
\Phi_{\Delta c}(n|n-1) = \frac{1}{\alpha_{\text{RLS}}} \hat{\Phi}_{\Delta c}(n-1), \tag{6.33}
\]

where \( \alpha_{\text{RLS}} \) is a forgetting factor that controls the convergence. The RLS algorithm is then described by (6.33) and (6.20) - (6.23), where \( \hat{\Phi}_{\Delta c}(n|n-1) \) is replaced by \( \Phi_{\Delta c}^{\text{RLS}}(n|n-1) \). Note that in the RLS algorithm, the covariance matrix \( \Phi_{\Delta c}(n) \) denotes the covariance of the MAR coefficients \( c(n) \) instead of the error covariance matrix.

The major advantages of the proposed Kalman filter over the RLS algorithm are:

- time varying stochastic modeling of MAR coefficients,
• modeling of the spatial covariance matrix $\Phi_s(n)$,
• adaptively controlled updating of $\hat{\Phi}_{\Delta c}(n|n-1)$,
• guaranteed stability and controllability of the Kalman filter under the given assumptions,
• ML based estimation of $\Phi_s(n)$ without requiring additional information like the reverberation time.

In contrast, by exploiting the diagonality of $\Phi_s(n)$, the RLS has a lower computational complexity by a factor of $M^2$ and does not require an $M \times M$ matrix inversion.

### 6.4 MAR online dereverberation and noise reduction

In this section, we propose a solution to jointly estimate the reverberant signals and the MAR coefficients from the noisy observations, and then to estimate the desired signals by (6.5). In Section 6.4.1 the noisy signal model is restated and the problem formulated. In Section 6.4.2 a state-of-the-art solution assuming deterministic MAR coefficients is reviewed and in Section 6.4.3 an alternating minimization algorithm using two Kalman filters with the stochastic MAR coefficient model presented in Section 6.2 is proposed.

#### 6.4.1 Noisy reverberant signal model and problem formulation

As already given in (6.4) the microphone signals with additive noise are described by

$$y(n) = X(n-D)c(n) + s(n) + v(n),$$

where the vector $d(n)$ contains the early speech plus noise signals. We assume that the desired early speech vector $s(n) \sim \mathcal{N}(0_{M \times 1}, \Phi_s(n))$, the noise vector $v(n) \sim \mathcal{N}(0_{M \times 1}, \Phi_v(n))$ and therefore also $d(n) \sim \mathcal{N}(0_{M \times 1}, \Phi_d(n))$ are circularly complex zero-mean Gaussian random variables with the respective covariance matrices $\Phi_s(n) = \mathcal{E}\{s(n)s^H(n)\}$, $\Phi_v(n) = \mathcal{E}\{v(n)v^H(n)\}$ and $\Phi_d(n) = \mathcal{E}\{d(n)d^H(n)\}$. Furthermore we assume that $s(n)$ and $v(n)$ are uncorrelated across time, and that $s(n)$, $v(n)$ and $r(n)$ are mutually uncorrelated. These assumptions hold well for the STFT coefficients of non-reverberant speech and a wide variety of noise types that typically have short to moderate temporal correlation in the time domain, and are widely used in speech processing methods [29, 132, 133]. The MAR coefficient vector $c(n)$ is given by the first-order Markov model (6.6).

We introduce an alternative notation using the stacked vectors

$$x(n) = \begin{bmatrix} x^T(n-L_{AR}+1) & \cdots & x^T(n) \end{bmatrix}^T$$

$$s(n) = \begin{bmatrix} 0_{1 \times M(L_{AR}-1)} & s^T(n) \end{bmatrix}^T,$$

$$r(n) = \begin{bmatrix} r^T(n) \end{bmatrix}^T, d(n) = \begin{bmatrix} d^T(n) \end{bmatrix}^T,$$
CHAPTER 6. MIMO REVERBERATION CANCELLATION

88

+ s(n) (n) + w(n) y(n) − D z−1

A

Figure 6.2: Generative model of the reverberant signal, multichannel autoregressive coefficients and noisy observation.

indicated as underlined variables, which are column vectors of length $ML$, and the propagation and observation matrices

$$F(n) = \begin{bmatrix} 0_{M(LAR-1)\times M} & I_{M(LAR-1)} \\ C_{LAR}(n) & \cdots & C_D(n) & 0_{M \times M(D-1)} \end{bmatrix} \quad (6.37)$$

$$H = \begin{bmatrix} 0_{M \times M(LAR-1)} & I_M \end{bmatrix} \quad (6.38)$$

respectively, where the $ML_{AR} \times ML_{AR}$ propagation matrix $F(n)$ contains the MAR coefficients $C_\ell(n)$ in the bottom $M$ rows, and $H$ is a $M \times ML_{AR}$ selection matrix. Using (6.37) and (6.38), we can write (6.35) and (6.34) as

$$\mathbf{x}(n) = F(n)\mathbf{x}(n-1) + \mathbf{s}(n) \quad (6.39)$$

$$\mathbf{y}(n) = H\mathbf{x}(n) + \mathbf{v}(n). \quad (6.40)$$

Note that (6.34) and (6.40) are equivalent using different notations. Figure 6.2 shows the generation process of the observed signals and the underlying (hidden) processes of the reverberant signals and the MAR coefficients.

In the next two sections, a state-of-the-art method and a novel method will be proposed to solve the problem defined in Section 6.1.2 by jointly estimating the noise-free reverberant signal vector $\mathbf{x}(n)$ and the MAR coefficients $c(n)$, such that we can obtain the desired early speech signal vector by (6.5).

6.4.2 State-of-the-art MAP-EM method

In [92] a method is proposed, that estimates the MAR coefficients $c(n)$ using a Bayesian approach based on maximum a posteriori (MAP) estimation and the noise-free desired signals $\mathbf{x}(n)$ are then estimated using an EM algorithm. Unlike the proposed method that will be presented in Section 6.4.3, the MAP-EM assumes the MAR coefficients to be deterministic and is not using the stochastic model (6.6). The block diagram shown in Figure 6.3 the signal flow of the MAP-EM method. All necessary parameters are estimated within the EM procedure.
Since the noise reduction stage receives delayed MAR coefficients, \( c(n) \) can be assumed only slowly time-varying. The algorithm is online, but the EM procedure requires about 20 iterations per frame to converge. The noise statistics are estimated within the EM algorithm.

6.4.3 MMSE estimation by alternating minimization using Kalman filters

The stacked reverberant speech signal vector \( \mathbf{x}(n) \) and the MAR coefficient vector \( c(n) \) (which is encapsulated in \( \mathbf{F}(n) \)) can be estimated in the MMSE sense by minimizing the cost-function

\[
J(x, c) = \mathcal{E}\{\|\mathbf{x}(n) - \hat{\mathbf{F}}(n)\hat{\mathbf{x}}(n-1) + \hat{s}(n)\|_2^2\}. \tag{6.41}
\]

To simplify the estimation problem (6.41) in order to obtain a closed-form solution, we resort to an alternating minimization technique [170], which minimizes the cost-function for each variable separately, while keeping the other variable fixed and using the available estimated value. The two sub-cost-functions, where the respective other variable is assumed as fixed, are given by

\[
J_c(c(n)|\hat{x}(n)) = \mathcal{E}\{\|c(n) - \hat{c}(n)\|_2^2\}, \tag{6.42}
\]

\[
J_x(\hat{x}(n)|c(n)) = \mathcal{E}\{\|\hat{x}(n) - \hat{\mathbf{x}}(n)\|_2^2\}. \tag{6.43}
\]

Note that to solve (6.42) at frame \( n \), it is sufficient to know the delayed stacked vector \( \hat{x}(n-D) \) to construct \( \hat{\mathbf{x}}(n-D) \), since the signal model (6.34) at time frame \( n \) depends only on past values of \( \mathbf{x}(n) \) with \( D \geq 1 \). Therefore we can state for the given signal model \( J_c(c(n)|\hat{x}(n)) = J_c(c(n)|\hat{x}(n-D)) \).

By now replacing the deterministic dependencies of the cost-functions (6.42) and (6.43) on \( \hat{x}(n) \) and \( c(n) \) by the available estimates, we naturally arrive at the alternating minimization procedure for each time step \( n \):

1) \( \hat{c}(n) = \arg \min_c J_c(c(n)|\hat{x}(n-D)) \tag{6.44} \)

2) \( \hat{x}(n) = \arg \min_{\hat{x}} J_x(\hat{x}(n)|\hat{c}(n)) \tag{6.45} \)
CHAPTER 6. MIMO REVERBERATION CANCELLATION

3) Reverberation estimation

1) AR coefficient estimation (Kalman filter)

2) Noise reduction (Kalman filter exploiting reverberant AR signal model)

\[ \hat{x}(n) - D \]

\[ \hat{c}(n) \]

\[ \hat{r}(n) \]

\[ \hat{s}(n) \]

\[ \Phi_v(n) \]

\[ y(n) \]

\[ \tilde{x}(n) \]

\[ z^{-D} \]

\[ \hat{r}(n) \]

\[ \hat{c}(n) \]

\[ \hat{x}(n-D) \]

Figure 6.4: Proposed parallel dual Kalman filter structure. The three-step procedure ensures that all blocks receive current parameter estimates without delay at each time step \( n \). For the grey noise estimation block, there exist several suitable solutions, which are beyond the scope of this thesis.

The ordering of solving (6.44) before (6.45) is essential if the coefficients \( c(n) \) are time-varying. Although convergence of the global cost-function (6.41) to the global minimum is not guaranteed, it converges to local minima if (6.42) and (6.43) decrease individually. For the given signal model, (6.44) and (6.45) can be solved using the Kalman filter [134].

The resulting procedure to estimate the desired signal vector \( s(n) \) by (6.5) results in the following three steps, which are also outlined in Figure 6.4:

1. Estimate the MAR coefficients \( c(n) \) from the noisy observed signals and delayed noise-free signals \( \tilde{x}(n') \) for \( n' \in \{1, \ldots, n-D\} \), which are assumed to be deterministic and known. In practice, these signals are replaced by the estimates \( \hat{x}(n') \) obtained from the second Kalman filter in Step 2.

2. Estimate the reverberant microphone signals \( \tilde{x}(n) \) by exploiting the autoregressive model. This step is considered as noise reduction stage. Here, the MAR coefficients \( c(n) \) are assumed to be deterministic and known. In practice, the MAR coefficients are obtained as the estimate \( \hat{c}(n) \) from Step 1. The obtained Kalman filter is similar to the Kalman smoother used in [40].

3. From the estimated MAR coefficients \( \hat{c}(n) \) and from delayed versions of the noise-free signals \( \tilde{x}(n) \), an estimate of the late reverberation \( \hat{r}(n) \) can be obtained. The desired signal is then obtained by subtracting the estimated reverberation from the noise-free signal using (6.5).

The noise reduction stage requires the second-order noise statistics as indicated by the grey estimation block in Figure 6.4. As there exist sophisticated methods to estimate second-order noise statistics, e.g. [100, 101, 119], further investigation of the noise statistics estimation is beyond the scope of this thesis, and we assume the noise statistics to be known.
6.4. MAR ONLINE DEREVERBERATION AND NOISE REDUCTION

The proposed structure overcomes the causality problem of commonly used sequential structures for AR signal and parameter estimation \(^92, 164\), where each estimation step requires a current estimate from each other. Such state-of-the-art sequential structures are illustrated in Figure 6.3 for the given signal model, where in this case the noise reduction stage receives delayed MAR coefficients. This would be suboptimal in the case of time-varying coefficients \(c(n)\).

In contrast to related state-parameter estimation methods \([164, 165]\), our desired signal is not the state variable but a signal obtained from both state estimates (6.5).

6.4.3.1 Optimal estimation of MAR coefficients

Given knowledge of the delayed reverberant signals \(x(n)\) that are estimated as shown in Figure 6.4, we derive a Kalman filter to estimate the MAR coefficients in this section.

Kalman filter for MAR coefficient estimation Let us assume, we have knowledge of the past reverberant signals contained in the matrix \(X(n - D)\). In the following, we consider (6.6) and (6.34) as state and observation equations, respectively. Given that \(w(n)\) and \(d(n)\) are circularly complex zero-mean Gaussian random variables and are mutually uncorrelated, we can obtain an optimal sequential estimate of the MAR coefficient vector by minimizing the trace of the error matrix

\[
\Phi_{\Delta e}(n) = E\{[(c(n) - \hat{c}(n))(c(n) - \hat{c}(n))]^H\}. \tag{6.46}
\]

The solution is obtained using the well-known Kalman filter equations \([94, 134]\)

\[
\hat{\Phi}_{\Delta e}(n|n - 1) = A\hat{\Phi}_{\Delta e}(n - 1)A^H + \Phi_w(n) \tag{6.47}
\]

\[
\hat{c}(n|n - 1) = A\hat{c}(n - 1) \tag{6.48}
\]

\[
e(n) = y(n) - X(n - D)\hat{c}(n|n - 1) \tag{6.49}
\]

\[
K(n) = \hat{\Phi}_{\Delta e}(n|n - 1)X^H(n - D) \left[ X(n - D)\hat{\Phi}_{\Delta e}(n|n - 1)X^H(n - D) + \Phi_d(n) \right]^{-1} \tag{6.50}
\]

\[
\hat{\Phi}_{\Delta e}(n) = [I_{L_e} - K(n)X(n - D)]\hat{\Phi}_{\Delta e}(n|n - 1) \tag{6.51}
\]

\[
\hat{c}(n) = \hat{c}(n|n - 1) + K(n)e(n), \tag{6.52}
\]

where \(K(n)\) is called the Kalman gain and \(e(n)\) is the prediction error. Note that the prediction error is an estimate of the early speech plus noise vector \(d(n)\) using the predicted MAR coefficients, i.e. \(e(n) = d(n|n - 1)\). For the case \(v(n) = 0_{M \times 1}\) this obtained Kalman filter is identical to the Kalman filter derived in Section 6.3.3 with \(\Phi_d(n) = \Phi_s(n)\) and \(X(n) = Y(n)\).

Parameter estimation The matrix \(X(n - D)\) containing only delayed frames of the reverberant signals \(x(n)\) is estimated using the second Kalman filter described in Section 6.4.3.2.

We assume \(A = I_{L_e}\) and the covariance of the uncertainty noise \(\Phi_w(n) = \phi_w(n)I_{L_e}\), where we propose to estimate the scalar variance \(\phi_w(n)\) by \([167]\)

\[
\hat{\phi}_w(n) = \frac{1}{L_e}\|\hat{c}(n) - \hat{c}(n - 1)\|^2 + \eta, \tag{6.53}
\]
and \( \eta \) is a small positive number to model the continuous variability of the MAR coefficients if the difference between subsequent estimated coefficients is zero.

In analogy to Section 6.3.4.2, the covariance \( \Phi_d(n) \) can be estimated in the ML sense given the conditional PDF \( f(y(n) \mid \hat{\Theta}(n)) \), where \( \hat{\Theta}(n) = \{ \hat{x}(n-1), \ldots, \hat{x}(n-L) \} \) are the currently available parameter estimates at frame \( n \). By assuming stationarity of \( \Phi_d(n) \) within \( N \) frames, the a priori ML estimate given the currently available information is obtained by

\[
\hat{\Phi}_d^{\text{ML}}(n) = \frac{1}{N} \left( \sum_{\ell=n-N+1}^{n-1} \hat{d}(n-\ell)\hat{d}^H(n-\ell) + e(n)e^H(n) \right),
\]

where \( \hat{d}(n) = y(n) - \hat{X}(n-D)\hat{c}(n) \) and \( e(n) = d(n|n-1) \) is the predicted speech plus noise signal, since \( \hat{c}(n) \) is not yet available.

In practice, the arithmetic average in (6.54) can be replaced by a recursive average, yielding the recursive ML estimate

\[
\hat{\Phi}_d(n) = \alpha \hat{\Phi}_d^{\text{pos}}(n-1) + (1 - \alpha)e(n)e^H(n),
\]

where the recursive a posteriori covariance estimate given \( \hat{d}(n) \), which can be computed only for the previous frame, is obtained by

\[
\hat{\Phi}_d^{\text{pos}}(n) = \alpha \hat{\Phi}_d^{\text{pos}}(n-1) + (1 - \alpha)\hat{d}(n)\hat{d}^H(n),
\]

and \( \alpha \) is a recursive averaging factor.

### 6.4.3.2 Optimal noise reduction

Given knowledge of the current MAR coefficients \( c(n) \) that are estimated as shown in Figure 6.4, we derive a second Kalman filter to estimate the stacked noise-free reverberant signal vector \( \bar{x}(n) \) in this section.

**Kalman filter for noise reduction** By assuming the MAR coefficients \( c(n) \), respectively the matrix \( F(n) \), as given and by considering the stacked reverberant signal vector \( \bar{x}(n) \) containing the latest \( L \) frames of \( x(n) \) as state variable, we consider (6.39) and (6.40) as state and observation equations. Due to the assumptions on \( s(n) \) and (6.36), the stacked vector \( \bar{s}(n) \) is also a zero-mean Gaussian random variable and its covariance matrix \( \Phi_{\bar{s}}(n) = E\{\bar{s}(n)\bar{s}^H(n)\} \) contains \( \Phi_s(n) \) in the lower right corner and is zero elsewhere.

Given that \( \bar{s}(n) \) and \( v(n) \) are circularly complex zero-mean Gaussian random variables, which are mutually uncorrelated and uncorrelated over time, we can obtain an optimal sequential estimate of \( \bar{x}(n) \) by minimizing the trace of the error matrix

\[
\Phi_{\Delta\bar{x}}(n) = E \{ [\bar{x}(n) - \bar{x}(n)] [\bar{x}(n) - \bar{x}(n)]^H \}.
\]

(6.57)
The standard Kalman filtering equations to estimate the state vector $\hat{x}(n)$ are given by the predictions

$$\hat{\Phi}_{\Delta x}(n|n-1) = F(n)\hat{\Phi}_{\Delta x}(n-1)F^H(n) + \Phi_v(n)$$

and updates

$$K_x(n) = \hat{\Phi}_{\Delta x}(n|n-1)H^H \left[ H\hat{\Phi}_{\Delta x}(n|n-1)H^H + \Phi_v(n) \right]^{-1}$$

$$e_x(n) = y(n) - H\hat{x}(n|n-1)$$

$$\hat{\Phi}_{\Delta x}(n) = [I_{ML} - K_x(n)H] \hat{\Phi}_{\Delta x}(n|n-1),$$

$$\hat{x}(n) = \hat{x}(n|n-1) + K_x(n)e_x(n)$$

where $K_x(n)$ and $e_x(n)$ are the Kalman gain and the prediction error of the noise reduction Kalman filter. The estimated noise-free reverberant signal vector at frame $n$ is encapsulated in the state vector and given by $\hat{s}(n) = H\hat{x}(n)$.

**Parameter estimation** The noise covariance matrix $\Phi_v(n)$ is assumed to be known in advance. For stationary or slowly time-varying noise, it can be estimated from the microphone signals during speech absence e.g. using the methods proposed in [100, 101, 119].

Further, we have to estimate $\Phi_s(n)$, i.e. the desired speech covariance matrix $\Phi_s(n)$. In order to reduce musical tones arising from the noise reduction procedure performed by the Kalman filter, we use a decision-directed approach [97] to estimate the current speech covariance matrix $\Phi_s(n)$, which is in this case a weighting between the a posteriori estimate $\hat{\Phi}_s^{\text{pos}}(n) = \mathcal{E}\{s(n)s^H(n)|\hat{s}(n)\}$ at the previous frame and the a priori estimate $\hat{\Phi}_s^{\text{pri}}(n) = \mathcal{E}\{s(n)s^H(n)|y(n),\hat{r}(n)\}$ at the current frame. The decision-directed estimate is given by

$$\hat{\Phi}_s(n) = \beta_{dd}\hat{\Phi}_s^{\text{pos}}(n-1) + (1 - \beta_{dd})\hat{\Phi}_s^{\text{pri}}(n),$$

where $\beta_{dd}$ is the decision-directed weighting parameter. To reduce musical tones, the parameter is typically chosen to put more weight on the previous a posteriori estimate.

The recursive a posteriori ML estimate is obtained by

$$\hat{\Phi}_s^{\text{pos}}(n) = \alpha\hat{\Phi}_s^{\text{pos}}(n-1) + (1 - \alpha)s(n)s^H(n),$$

where $\alpha$ is a recursive averaging factor.

To obtain the a priori estimate $\hat{\Phi}_s^{\text{pri}}(n)$, we derive a multichannel Wiener filter (MWF), i.e.

$$W_{\text{MWF}}(n) = \arg\min_W \mathcal{E}\{||s(n) - W^Hy(n)||_2^2\}.$$  

By inserting (6.39) in (6.40), we can rewrite the observed signal vector as

$$y(n) = s(n) + H\Phi(n)x(n-1) + v(n),$$
Algorithm 6.1 Proposed algorithm per frequency band $k$

1: **Initialize:** \( \hat{c}(0) = 0, \hat{x}(0) = 0, \hat{\Phi}_{\Delta c}(n) = I_{Lc}, \hat{\Phi}_{\Delta x}(n) = I_{MLAR} \)

2: **for** each $n$ **do**

3: Estimate the noise covariance \( \Phi_v(n) \), e.g. using [101]

4: \( X(n-D) \leftarrow \hat{x}(n-1) \)

5: Compute \( \hat{\Phi}_w(n) = \phi_w(n)I_{Lc} \) using (6.53)

6: Obtain \( \hat{c}(n) \) by calculating (6.47)-(6.49), (6.54), (6.50)-(6.52)

7: \( F(n) \leftarrow \hat{c}(n) \)

8: \( \Phi_s(n) \leftarrow \hat{\Phi}_s(n) \) using (6.64)

9: Obtain \( \hat{x}(n) \) by calculating (6.59)-(6.62)

10: Estimate the desired signal using (6.5)

where all three components are mutually uncorrelated. Note that estimates of all components of the late reverberation \( r(n) \) are already available at this point. An instantaneous estimate of \( \hat{\Phi}_s(n) \) using an MMSE estimator given the currently available information is then obtained by

\[
\hat{\Phi}_s^{\text{pri}}(n) = W_{\text{MWF}}^H(n) y(n) y^H(n) W_{\text{MWF}}(n).
\]

(6.68)

The MWF filter matrix is given by

\[
W_{\text{MWF}}(n) = \Phi^{-1}_y(n) \left[ \Phi_y(n) - \Phi_r(n) - \Phi_v(n) \right],
\]

(6.69)

where \( \Phi_y(n) \) and \( \Phi_r(n) \) are estimated using recursive averaging from the signals \( y(n) \) and \( \hat{r}(n) \), similar to (6.65).

### 6.4.3.3 Algorithm overview

The complete algorithm is outlined in Algorithm 6.1. The initialization of the Kalman filters is uncritical. The initial convergence phase could be improved if good initial estimates of the state variables are available, but the algorithm always converged and stayed stable in practice.

Although the proposed algorithm is perfectly suitable for real-time processing applications, the computational complexity is quite high. The complexity depends on the number of microphones $M$, the filter length $L_{AR}$ per frequency, and the number of frequency bands.

### 6.5 Reduction control

In some applications, it is beneficial to have independent control over the reduction of the undesired sound components such as reverberation and noise. Therefore, we show how to compute an alternative output signal $z(n)$, where we have control over the reduction of reverberation and noise.

The desired controlled output signal is given by

\[
z(n) = s(n) + \beta_r r(n) + \beta_v v(n),
\]

(6.70)
6.6 DISCUSSION

where $\beta_v$ and $\beta_r$ are attenuation factors of the reverberation and noise. By re-arranging (6.70) using (6.34), and by replacing unknown variables by the available estimates, we can compute the desired controlled output signals by

$$\hat{z}(n) = \beta_v y(n) + (1 - \beta_v)\hat{x}(n) - (1 - \beta_r)\hat{r}(n).$$

(6.71)

Note that for $\beta_v = \beta_r = 0$, the output $\hat{z}(n)$ is identical to the early speech estimate $\hat{s}(n)$, and for $\beta_v = \beta_r = 1$, the output $\hat{z}(n)$ is equal to $y(n)$.

Typically, speech enhancement algorithms have a trade-off between the amount of interference reduction and artifacts such as speech distortion or musical tones. To reduce audible artifacts in periods where the MAR coefficient estimation Kalman filter is adapting fast and exhibits a high prediction error, we use the estimated error covariance matrix $\hat{\Phi}_{\Delta e}(n)$ given by (6.51) to adaptively control the reverberation attenuation factor $\beta_r$. If the error of the Kalman filter is high, we like the attenuation factor $\beta_r$ to be close to one. We propose to compute the reverberation attenuation factor at time frame $n$ by the heuristically chosen mapping function

$$\beta_r(n) = \max \left( \frac{1}{1 + \mu_r L_c \text{tr} \left\{ \hat{\Phi}_{\Delta e}(n) \right\}^{-1}}, \beta_{r,\min} \right),$$

(6.72)

where the fixed lower bound $\beta_{r,\min}$ limits the allowed reverberation attenuation, and the factor $\mu_r$ controls the attenuation depending on the Kalman error.

The structure of the proposed system with reduction control is illustrated in Figure 6.5. The noise estimation block is omitted here as it can be also integrated into the noise reduction block.

6.6 Discussion

The proposed dual-Kalman structure is a generalizing framework for several previously proposed methods. In the case of no additive noise $v(n)$, the dual-Kalman naturally reduces to the single-Kalman filter proposed in Section 6.3.3. Further, the single-Kalman is a generalization of the RLS algorithm. However, the assumption required for the RLS of zero spatial correlation of
6.7 Evaluation

In this section, we evaluate the state-of-the-art and proposed methods using the experimental setup described in Section 6.7.1. The dereverberation methods assuming no noise, namely the RLS and the proposed single-Kalman presented in Section 6.3, are evaluated in Section 6.7.2. The methods for dereverberation and noise reduction presented in Section 6.4 are evaluated in Section 6.7.3.

6.7.1 Experimental setup

The reverberant signals were generated by convolving room impulse responses (RIRs) with anechoic speech signals from [160]. We used two different kinds of RIRs: measured RIRs in an acoustic lab with variable acoustics at Bar-Ilan University, Israel, or simulated RIRs in a shoebox room using the image method [131] for moving sources. In the case of moving sources, the simulated RIRs facilitate the evaluation, as it is possible to additionally generate RIRs containing only direct sound and early reflections to obtain the target signal for evaluation.

In the simulated and measured cases, we used a linear microphone array with up to \( M = 4 \) omnidirectional microphones with inter-microphone spacings \( \{11, 7, 14\} \) cm. Note that in all experiments except in Section 6.7.3.1, only 2 microphones with spacing 11 cm are used. The speech samples were taken from [160]. Either stationary pink noise or recorded babble noise was added to the reverberant signals with a certain input signal-to-noise ratio (iSNR). We used a sampling frequency of 16 kHz, and the STFT parameters were a square-root Hann window of 32 ms length, a frame shift of \( T_{\text{hop}} = 16 \text{ ms} \), and a discrete Fourier transform (DFT) length of \( K = 1024 \) samples. The delay separating early and late reverberation was set to \( D = 2 \) frames.

The recursive averaging factor was \( \alpha = e^{-\frac{T_{\text{hop}}}{\tau}} \) with \( \tau = 25 \text{ ms} \), the decision-directed weighting factor was \( \beta_{\text{dd}} = 0.98 \), and we chose \( \eta = 10^{-4} \). We present results without reduction control (RC), i.e. \( \beta_{v} = \beta_{r} = 0 \), and with RC using different settings for \( \beta_{v} \) and \( \beta_{r_{\text{min}}} \), where we chose \( \mu_{r} = -10 \text{ dB} \) in (6.72).

For evaluation, the target signals were generated as the direct speech signal with early reflections up to 32 ms after the direct sound peak (corresponds to a delay of \( D = 2 \) frames). The
processed signals are evaluated in terms of the cepstral distance (CD) [171], perceptual evaluation of speech quality (PESQ) [172], frequency-weighted segmental signal-to-interference ratio (fwSSIR) [173], where reverberation and noise are considered as interference, and the normalized signal-to-reverberation-modulation ratio (SRMR) [174]. These measures have been shown to yield reasonable correlation with the perceived amount of reverberation and overall quality in the context of dereverberation [20, 175]. The CD reflects more the overall quality and is sensitive to speech distortion, while PESQ, fwSSIR and SRMR are more sensitive to reverberation/interference reduction. We present only results for the first microphone as all other microphones show the same behavior.

### 6.7.2 Results for the dereverberation methods assuming no noise

In acoustic conditions with sufficiently high iSNR, we can assume that no noise is present such that the single-Kalman and RLS methods presented in Section 6.3 are applicable. To evaluate the performance of the Kalman vs. RLS, and the performance of the proposed ML speech covariance estimator independently, the RLS is implemented in three variants: 1) using the originally proposed estimator for $\Phi_s(n)$ based on spectral subtraction requiring the reverberation time, without speaker change detection (SCD), termed $RLS + SS$; 2) the same implementation as 1) with SCD [39], termed $RLS-SCD + SS$; and 3) the RLS using the proposed ML early speech covariance estimator (6.31) derived in Section 6.3.4.2, termed $RLS-SCD + MLE$. As the matrix $\Phi_s(n)$ is required to be diagonal in the RLS, we used an identity matrix scaled by the mean of the main diagonal of $\hat{\Phi}_s(n)$. The RLS forgetting factor was set to $\alpha_{RLS} = 0.99$ as proposed in [91].

We simulated a test signal using a measured RIR with $T_{60} = 630$ ms and added pink noise with a high iSNR of 50 dB. For the first 15 s, the first speaker is talking. After the first speaker stopped, the second speaker starts talking from a different position. The processed signals are

![Figure 6.7: Segmental objective measures for dynamic scenario using measured RIRs with $T_{60} = 630$ ms, iSNR = 50 dB, $L_{AR} = 15$.](image-url)
CHAPTER 6. MIMO REVERBERATION CANCELLATION

6.7 Results for the noise and reverberation reduction methods

In this section, the performance of the MAR based methods for dereverberation and noise reduction presented in Section 6.4 is investigated. The single-Kalman method presented in Section 6.3.3 is included as well as reference method. In the first two experiments, specific parameters of the dual-Kalman are investigated.

6.7.3.1 Dependence on number of microphones

We investigated the performance of the proposed dualKalman algorithm depending on the number of microphones $M$. The desired signal consisted of a single active speech source, that switched position after 9 s using measured RIRs with $T_{60} = 630$ ms and stationary pink noise with $iSNR = 15$ dB. Figure 6.8 shows CD, PESQ, fwSSIR and SRMR for a varying number of microphones $M$. The measures for the noisy reverberant input signal are indicated as light grey dashed line, and the SRMR of the target signal, i.e. the early speech, is indicated as dark grey dash-dotted line. For $M = 1$, the CD is larger than for the input signal, which indicates an overall quality deterioration, whereas PESQ, fwSSIR and SRMR still improve over the input, i.e. reverberation and noise are reduced. The performance in terms of all measures increases by increasing the number of microphones.

6.7.3.2 Dependence on filter length

In this experiment, we investigate the effect of the filter length $L_{AR}$ using measured RIRs with different reverberation times. As in the first experiment, the speaker position switched after 9 s evaluated in terms of CD, fwSSIR and PESQ in 50% overlapping segments of 2 s length. The improvement of the measures compared to the unprocessed reference microphone signal is shown in Figure 6.7. We can observe that all methods improve the measures, except the RLS without SCD. The RLS + SS without SCD (black line) drops to a negative improvement for all measures, which indicates that it becomes unstable after the position change. This problem could be solved by using the RLS with SCD (grey dashed line). We can observe that the performance of the RLS-SCD can be improved by using the proposed early speech ML covariance estimator (blue line) instead of the spectral subtraction based estimator based on the exponential decay model. The performance of the Kalman filter (orange line) is very similar to the RLS with SCD, but the Kalman filter stays stable even in dynamic scenarios without any further control mechanisms.
6.7. EVALUATION

and stationary pink noise was added with iSNR = 15 dB. Figure 6.9 shows the improvement of the objective measures compared to the unprocessed microphone signal. Positive values indicate an improvement for all relative measures, where \( \Delta \) denotes the improvement. For the used STFT parameters, the reverberation times \( T_{60} = \{480, 630, 940\} \) ms correspond to filter lengths \( L_{AR} = \{30, 39, 58\} \) frames. We can observe that the best CD, PESQ and fwSSIR values depend on the reverberation time, but the optimal values are obtained at around 25% of the corresponding length of the reverberation time. In contrast, the SRMR monotonously grows with increasing \( L_{AR} \). It is worthwhile to note that the reverberation reduction becomes more aggressive with increasing \( L_{AR} \). If the reduction is too aggressive by choosing \( L_{AR} \) too large, the desired speech is distorted as the \( \Delta CD \) indicates with negative values.

6.7.3.3 Comparison with state-of-the-art methods

In this experiment, we compare the proposed dual-Kalman algorithm to the single-Kalman proposed in Section 6.3.3 and to the MAP-EM method reviewed in Section 6.4.2 for two noise types in varying iSNRs. The desired signal consisted of a single active speech source, that switched position after 9 s using measured RIRs with \( T_{60} = 630 \) ms. Either stationary pink noise or recorded babble noise was added with varying iSNRs. Table 6.1 shows the improvement of the objective measures compared to the unprocessed microphone signal in stationary pink noise and in babble noise, respectively. Note that although the babble noise is not short-term stationary, we used a stationary long-term estimate of the noise covariance matrix \( \Phi_v(n) \), which is realistic to obtain as an estimate in practice.

It can be observed that the proposed algorithm either without or with RC yields the best measures in all conditions. The RC provides a trade-off control between interference reduction and distortion of the desired signal or other artifacts. The CD as an indicator for overall quality is consistently better with RC, whereas the other measures, which majorly reflect the amount of interference reduction, consistently achieve slightly higher results without RC in stationary noise. In babble noise at low iSNRs, the dual-Kalman with RC yields higher PESQ than without RC. This indicates that the RC can help to improve the quality by masking artifacts in challenging iSNR conditions and in the presence of noise covariance estimation errors. In high iSNR conditions, the performance of the dual-Kalman becomes similar to the performance of the single-Kalman as expected.
6.7.3.4 Tracking of moving speakers

A moving source was simulated using simulated RIRs in a shoebox room with $T_{60} = 500$ ms based on the image method [131,176], using the setup shown in Figure 6.10: For the first 8 s, the desired source was at position A. During the time interval [8,13] s it moved with constant velocity from position A to B, where it stayed then for the rest of the time. Positions A and B were 2 m apart and are shown in the simulated room in Figure 6.10. There were no periods of speech absence. Figure 6.11 shows the segmental improvement of CD, PESQ, fwSSIR and SRMR for this dynamic scenario. In this experiment, the target signal for evaluation is generated by simulating the wall reflections only up to the second reflection order.

We observe in Figure 6.11 that all measures decrease during the movement, while after the speaker has reached position B, the measures reach high improvements again. The convergence
of all methods behaves similar, while the dual-Kalman without and with RC perform best. During the moving time period, the MAP-EM yields sometimes higher fwSSIR and SRMR, but at the price of much worse CD and PESQ. The reduction control improves the CD, such that the CD improvement always stays positive, which indicates that the RC can reduce speech distortion and artifacts. It is worthwhile to note that even if the reverberation reduction can become less effective during movement of the speech source, the dual-Kalman algorithm did not become unstable, and the improvements of PESQ, fwSSIR and SRMR were always positive, and the ΔCD was always positive by using the RC. This was also verified using real recordings with moving speakers.

### 6.7.3.5 Evaluation of reduction control

In this section, we evaluate the performance of the RC in terms of the reduction of noise and reverberation by the proposed system. In Appendix C it is shown how the residual noise and reverberation signal components after processing by the dual-Kalman filter system with RC, \( z_v(n) \) and \( z_r(n) \) respectively, can be computed. The noise reduction and reverberation reduction measures are then given by

\[
NR(n) = \frac{\sum_k ||z_v(k,n)||_2^2}{\sum_k ||v(k,n)||_2^2} \tag{6.73}
\]

\[
RR(n) = \frac{\sum_k ||z_r(k,n)||_2^2}{\sum_k ||r(k,n)||_2^2}. \tag{6.74}
\]

In this experiment, we simulated a scenario with a single speaker at a stationary position using measured RIRs in the acoustic lab with \( T_{60} = 630 \) ms. In Figure 6.12, five different
settings for the attenuation factors are shown: No reduction control ($\beta_v = \beta_r, \text{min} = 0$), a moderate setting with $\beta_v = \beta_r, \text{min} = -7$ dB, reducing either only reverberation or only noise by 7 dB, and a stronger attenuation setting with $\beta_v = \beta_r, \text{min} = -15$ dB. We can observe that the noise reduction measure yields the desired reduction levels only during speech pauses. The reverberation reduction measure shows that a high reduction is only achieved during speech absence. This does not mean that the residual reverberation is more audible during speech presence, as the direct sound of the speech perceptually masks the residual reverberation. During the first 5 seconds, we can observe the reduced reverberation reduction caused by the adaptive reverberation attenuation factor (6.72), as the Kalman filter error is high during the initial convergence.

### 6.8 Summary

In this chapter, a dereverberation method based on a narrowband MAR reverberation signal model was presented. Using this model, the multichannel early speech signal is obtained by subtracting the predicted reverberation using previous reverberant signals from the current reverberant signal. The prediction of the reverberation is obtained using MAR coefficients. Therefore, the dereverberation process can be described as a MIMO reverberation canceller.

Whereas in the noise-free case, the MAR coefficients can be estimated directly from the observed signals using a single Kalman filter, in the general noisy case, the reverberant signals have to estimated as well by reducing the noise. We have proposed an alternating minimization algorithm that is based on two interacting Kalman filters to jointly estimate the MAR coefficients and the reverberant signal, in order to finally reduce the noise and reverberation from each microphone signal. The so-called dual-Kalman framework is a generalized framework for the single-Kalman method to estimate the MAR coefficients that does not consider additive noise. Further, the single-Kalman algorithm is a generalization of state-of-the-art RLS methods. An essential part of the proposed algorithm is the ML estimator for the speech covariance matrix,
which was shown to outperform existing estimators and can also improve the performance of the RLS method. Both proposed solutions using recursive Kalman filters are suitable for online processing applications. In contrast to state-of-the-art methods, within the proposed Kalman filters the MAR coefficients are modeled by a first-order Markov model, in order to explicitly take the uncertainty of time-varying coefficients into account. The effectiveness and superior performance to similar online methods were demonstrated in various experiments including non-stationary babble noise and moving sources. In addition, a method to control the reduction of noise and reverberation independently was proposed to mask possible artifacts, and to adjust the output signal to perceptual requirements and taste.
CHAPTER 7

Evaluation and comparison of proposed dereverberation methods

In this thesis, three substantially different dereverberation systems were proposed:

(i) the single-channel Wiener filter (WF) based dereverberation (Section 3.1) using the late reverberation power spectral density (PSD) estimator using the relative convolutive transfer function (RCTF) model (Section 4.3), in the following referred to as WF,

(ii) the multichannel Wiener filter (MWF) based dereverberation (Section 3.2.2) with the blocking PSD least-squares reverberation PSD estimator proposed in Section 5.3 and the bias compensation described in Section 5.5 (MWF),

(iii) the reverberation canceller based on the multichannel autoregressive (MAR) model presented in Section 6.4.3 (dual-Kalman).

For both WF and MWF we use the reduction control as proposed in Section 3.3.

In this chapter, we present an extensive experimental comparison between these three methods demonstrating the advantages and disadvantages. The acoustic setup and processing parameters are described in Section 7.1.1. At first, an experiment to find the optimal parameters for reduction control for each algorithm is conducted in Section 7.1.2.1. Using the optimal parameters for each algorithm, results for various acoustic scenarios are presented in Section 7.1.2.2. In addition to the evaluation using objective speech enhancement measures, the results of a subjective listening test are presented in Section 7.1.3, and the improvement by using the dereverberation algorithms as a front-end processing for an automatic speech recognition (ASR) system are presented in Section 7.2.

7.1 Evaluation for speech enhancement

7.1.1 Setup

The test data was generated using a female and a male speech signal from [160] with total length of 40 s. Six different acoustic conditions were generated using the room impulse responses (RIRs) measured in the acoustic lab at Bar-Ilan University with three reverberation times of $T_{60} = \{480, 630, 940\}$ ms each with source distances of 2 m and 4 m. A linear array of omnidirectional
CHAPTER 7. COMPARISON OF DEREVERBERATION METHODS

Microphones with spacings 4 - 6 - 4 cm was used. Either stationary pink noise or recorded babble noise was added to the reverberant signals with various input signal-to-noise ratios (iSNRs) as defined in Appendix A. Additionally, in Section 7.1.2.3 simulated RIRs by the image method [131] were used. The target signal for evaluation was the direct sound obtained by applying a window around the direct sound peak of the RIRs. The signals were processed with a sampling frequency of $f_s = 16$ kHz, using a short-time Fourier transform (STFT) with square-root Hann windows of length 32 ms, a frame shift of $T_{\text{hop}} = 16$ ms, and a $K = 1024$ point discrete Fourier transform (DFT). The stationary noise PSD (matrix) was assumed to be known and computed in advance as a long-term arithmetic average over periods of speech absence.

Although we have shown in Sections 4.4.2 and 6.7.3 that convolutive transfer function (CTF) length $L_{\text{MA}}$ and the MAR filter length $L_{\text{AR}}$ of the WF and dual-Kalman methods depend on the reverberation time, we used the fixed parameters $L_{\text{MA}} = 20$ and $L_{\text{AR}} = 15$ in this evaluation to show the performance of the systems in realistic conditions, as it might be difficult to set these parameters adaptively in practice. For the MWF, we used the single-wave model with $I = 1$, and therefore used the decomposed spatial filter - postfilter form. For linear arrays, the narrowband direction-of-arrivals (DOAs) were estimated for each time-frequency bin using TLS-ESPRIT [110], while for uniform circular arrays (UCAs) beamspace root-MUSIC [111] was used. For all methods, we evaluated the output of the reduction control with a specified attenuation factor, which was chosen equal for reverberation and noise $\beta_r = \beta_v = \beta$.

Note that the dual-Kalman used only two microphones (the pair with the largest spacing), whereas the WF used only one microphone and the MWF used all four microphones.

### 7.1.2 Experimental results

In this section, we evaluate and compare the proposed algorithms using the measures described in Appendix B.

#### 7.1.2.1 Reduction control

In this experiment, the effects of the reduction control for the three dereverberation algorithms by changing the attenuation factor for noise and reverberation are investigated. Figure 7.1 shows the objective measures for varying attenuation factors for noise and reverberation. In general we observe a consistent ranking of dual-Kalman, MWF and WF from highest to lowest performance. We can observe that the dual-Kalman yields a generally lower cepstral distance (CD), log-likelihood ratio (LLR) and higher perceptual evaluation of speech quality (PESQ) than MWF and WF for all attenuation limits. For measures mainly related to the reverberation and noise reduction like frequency-weighted segmental signal-to-interference ratio (fwSSIR), quality assessment of reverberated speech (QAreverb) the optimal attenuation limits are the lowest values. However, for measures closer related to the overall quality like CD, LLR, Mel-frequency cepstral coefficient distance (MFCCD), and partly PESQ, there exist "sweet spots" for the attenuation factors. The sweet spots for the dual-Kalman are generally at lower attenuation limits than for the Wiener filters. Therefore, the dual-Kalman can achieve higher suppression with better quality. The MWF consistently yields a slightly higher fwSSIR than the WF and the
dual-Kalman, while the dual-Kalman is still better than the WF for lower attenuation limits. QAverb shows similar performance for the dual-Kalman and WF, which both outperform the MWF. The MFCCD shows a large gap between the single-channel WF and both other methods, where again the dual-Kalman consistently performs best.

In the following experiments, the attenuation factors are chosen as the on average optimal values from Figure 7.1 individually for each method, namely $\beta = -18$ dB for the dual-Kalman, $\beta = -15$ dB for the MWF, and $\beta = -13$ dB for the WF.

Overall we can state that the dual-Kalman is the most powerful method as it achieves the best perceived quality and the least speech distortion with the most or equal reverberation and noise suppression with only two microphones. The possible range of noise and reverberation suppression is larger than for the MWF and WF.

7.1.2.2 Performance in multiple acoustic conditions

The performance in different acoustic conditions using the measured RIRs with three different reverberation times described in detail in Table 4.1 is shown in Figure 7.2 in terms of the speech enhancement measures.

We can observe that for all measures except fwSSIR, the dual-Kalman achieves the best results, while the WF yields the smallest improvements. In terms of fwSSIR, the dual-Kalman and MWF achieve quite similar results. This indicates a remarkable performance of the dual-Kalman as it only uses two microphones compared to the 4-channel MWF. Whereas the WF can achieve in some conditions achieves improvements of PESQ, fwSSIR and QAverb similar to the MWF, it fails to improve the MFCCD, and shows only minor improvements of the CD for some conditions.

A deeper insight in the characteristics of the dereverberation methods can be gained from the energy decay curve (EDC) before and after processing. In Appendix B.1.2, a method to estimate the EDC from the direct sound and the processed or unprocessed signals is described.
Note that in this procedure, it is assumed that the microphone signals are processed by linear time-invariant systems, which is actually not the case for the dereverberation systems under test. The EDCs are shown for two acoustic conditions in Figure 7.3. We can observe that the EDC of the dual-Kalman output achieves the fastest initial decay, whereas the WF reduces the least early reflections, but achieves slightly higher suppression of the late reverberation. In contrast to the WF, the dual-Kalman and MWF clearly achieve also some suppression of the early reflections.

7.1.2.3 Dependence on microphone array geometry

In spatial filtering approaches like the MWF, the microphone array geometry, the number of microphones, and even the source direction relative to the array orientation have a large impact on the performance [77]. As shown in Chapter 6, the performance of the dual-Kalman also
increases with the number of microphones. However, it is yet unclear how the performance of dual-Kalman and MWF relate depending on the array geometry and the number of microphones.

To investigate the dependence of the array geometry, we simulated RIRs in a shoebox room using the image method [131], in order to compare multiple exactly co-located arrays. We simulated three different arrays with $M = 8$ microphones as shown in Figure 7.4: two uniform linear arrays (ULAs) with $d = 2$ cm spacing in broadside and endfire orientation to the source, and a UCA with radius 5 cm. The first microphone of all three arrays used as reference microphone was at the same location, shown as black dot. The source was at 3.3 m distance in the indicated direction from the reference microphone and $T_{60} = 600$ ms.

As first experiment, we investigate the influence of the number of microphones using the ULA in broadside orientation. We evaluated the performance of the dual-Kalman and the MWF by successively using more microphones from $M = \{1, \ldots, 8\}$. While the dual-Kalman can be computed for $M = 1$ without modification, the MWF reduces the single-channel WF, where the parameters have to be estimated differently, i.e. we used the RCTF based reverberation PSD estimator. Figure 7.5 shows the improvement of CD, PESQ and fwSSIR for this experiment. We can observe that the performance of both methods increases with $M$. The CD of the dual-Kalman increases faster for small $M$, while the MWF saturates at larger $M$. The PESQ curves behave similar for both methods, while the fwSSIR of the dual-Kalman is slightly higher than the MWF. The dual-Kalman with $M = 1$ yields significantly worse results as in the multichannel case, and yields worse results than the single-channel WF.

As second experiment, both methods were computed using seven different (sub-)array setups from the arrays shown in Figure 7.4: (i-ii) using $M = 2$ neighboring microphones of both ULAs with $d = 2$ cm, (iii-iv) using the $M = 2$ outer microphones of both ULAs with $d = 16$ cm, and (v-vii) using all $M = 8$ microphones of the three arrays. Figure 7.6 shows the improvement of CD, PESQ and fwSSIR for the dual-Kalman and MWF using the different arrays or subarrays. The influence of the ULA orientation on the dual-Kalman is rather small, whereas the
CHAPTER 7. COMPARISON OF DEREVERBERATION METHODS

Figure 7.6: Objective measures using different microphone arrays for dual-Kalman (DK) and MWF. Simulation of a rectangular room with $T_{60} = 600$ ms and iSNR = 15 dB.

MWF heavily depends on the array orientation. For $M = 2$, a larger spacing increases the performance, where the dual-Kalman yields a large performance gain, and the MWF only a smaller gain. Note that although a large microphone spacing is beneficial for the low frequency performance of spatial filters, a too large spacing introduces spatial aliasing at high frequencies, which is problematic especially for narrowband DOA estimators. While spatial aliasing is not problematic for the spacing of $d = 2$ cm, where the spatial aliasing frequency given by (3.16) is $f_{\text{alias}} = 8.57$ kHz, and therefore lies above the Nyquist frequency\(^1\), for a spacing of $d = 16$ cm the spatial aliasing frequency $f_{\text{alias}} = 1.07$ kHz can be problematic. The circular array achieves a much higher performance than the linear array due to the higher spatial diversity. It is interesting to note that the performance of the MWF for the endfire ULA is relatively low, whereas for the UCA the MWF achieves clearly the best performance in terms CD and fwSSIR. For $M = 8$, the dual-Kalman outperforms the MWF, although it has to be noted that the computational complexity of the dual-Kalman is very high for such a large number of microphones. However, it is worthwhile to note that the dual-Kalman with $M = 2$ and a large microphone spacing achieved a performance that is higher than the MWF with a 8-channel ULA, and even outperformed the MWF using a ULA with $M = 8$.

7.1.3 Listening test

The signal based speech enhancement measures described in Appendix B.1.1 can reliably indicate trends for improvement within a class of similar enhancement methods or processing with different parameters, but coherence across substantially different enhancement methods is sometimes less reliable as found in [175]. Therefore, to confirm the results of the objective measures and to test their reliability across multiple dereverberation classes, a listening test was conducted. The participants were asked for the perceived amount of reverberation and the overall quality. In Section 7.1.3.1 the design and the conditions of the listening test are presented, and in Section 7.1.3.2 the results and the correlation with the speech enhancement measures are evaluated.

\(^1\)The Nyquist frequency given by $f_s/2$ determines the highest frequency that can be captured.
7.1.3.1 Test design

The listening test was designed as a multi-stimulus test similar to a multi-stimulus test with hidden reference and anchor (MUSHRA) [177]. As in the case of dereverberation it is arguable if the target signal (direct sound) has the highest quality in the perceptual sense, the standard MUSHRA is not suitable for our purpose. We rather want to assess how well the reverberation is reduced and if the signal quality is improved or distorted. The test was divided into two sessions, where in the first session the participants were asked to rate the perceived amount of reverberation and in the second session they were asked to rate the overall quality. The design of the two sessions is described in the following.

Perceived reverberation  In the first session the participants were asked: "Please rate the perceived amount of reverberation, without taking into account other aspects like distortion, artifacts, naturalness, etc." The reference signal was the direct sound. For each condition, 6 signals had to be rated:

- The hidden reference (direct sound at the reference microphone)
- The early reflections signal
- The reverberant noisy reference microphone signal
- Three processed signals by the dual-Kalman, MWF and WF.

As the participants were asked for the perceived amount of reverberation, the hidden reference should be rated with the highest number of points. In this session, there was no defined anchor,
although it is expected that the unprocessed microphone signal is rated with the highest amount of reverberation. The inclusion of the early reflections signal serves as an indicator, how close the algorithms can get to this signal, as the dereverberation methods mainly can reduce only the late reverberation. The graphical user interface (GUI) with the rating scale for the first session is shown in Figure 7.7.

**Overall quality** In the second session the participants were asked: "Please rate the overall quality for each item with respect to the reference. Take into account every aspect like perceptual similarity, naturalness, distortion, artifacts, reverberation characteristics, additive noise, etc.” The reference signal was the direct sound. For each condition, the following 6 signals had to be rated:

- The hidden reference (direct sound)
- The reverberant noisy microphone signal
- The 3.5 kHz lowpass filtered microphone signal (as a lower anchor)
- Three processed signals by the **dual-Kalman**, **MWF** and **WF**.

The hidden reference should be rated with the highest ranking. The 3.5 kHz lowpass anchor serves as a sanity check: if an algorithm produces worse quality than this anchor, it cannot be considered as useful. The GUI with the rating scale for the second session is shown in Figure 7.8.
Evaluation scale  For evaluation, we used a differential score [178] with respect to the unprocessed reference microphone signal, such that the score of the microphone signal is always zero. In this way, the scores are scaled for each condition, which can contain different amounts of reverberation. Therefore, an improvement over the unprocessed microphone signal is indicated by a positive score, and a deterioration by a negative score.

With this test design, we can test i) how close we can get in terms of reverberation reduction to the early reflection signal, ii) if the overall quality improves or deteriorates compared to the reverberant signal, and iii) if the quality becomes worse than the low-pass filtered reverberant signal.

Test conditions  The presented signals were excerpts of 7-11 s length from the signals used in Section 7.1.2.2. The excerpts started at least 5 s after the signal beginning to skip the initial convergence phase of adaptive algorithms. We tested 12 conditions comprising 3 rooms × 2 distances × 2 speech signals (male and female). Pink noise was added with an iSNR of 30 dB. The microphone signals were processed by the dual-Kalman (2 channels using the widest spaced microphone pair with 14 cm), the MWF (4-channel linear array with spacing 4 - 6 - 4 cm) and the WF (1 channel). The webMUSHRA software [179] was used for presentation to the listeners. The signals were presented in an acoustically treated control room typical for recording studios via open headphones (Stax SRS-507). The signals were root-mean square (RMS) normalized. Before each session, a training for one condition was conducted to familiarize the participants with the task and the signals. The participants could freely adjust the volume during the training.

20 participants took part in the listening test, 3 female and 17 male expert and non-expert listeners. None of the participants reported abnormal hearing and their age ranged from 24-39 years with an average of 31.1 years. All participants found the reference signal reliably in all conditions. One session took the participants between 15 to 30 minutes.

7.1.3.2 Listening test results

The overall results for the perceived amount of reverberation and overall quality are shown in Figure 7.9, where the red bars indicate the median, the blue box indicates the upper and lower quartiles, and the whiskers indicate the data range not considered as outliers, which are shown as red ”+” markers.

We can observe that the direct sound was always rated highest, as it was part of the task. The reverberation reduction of the three methods was rated quite similar, where the WF reduced slightly less reverberation compared to dual-Kalman and MWF. Although there is still a significant gap to the early speech signal, a clear reverberation reduction compared to the unprocessed reverberant signal is achieved for all methods. Figure 7.9 (b) shows that all three methods improve the subjective overall quality over the reverberant signal and the low-pass anchor. The statistical significance between the three algorithms is shown in Table 7.1 in terms of the p-value using a paired Student’s t-test [180]. If the p-value is smaller than 0.05, the null-hypothesis that two datasets come from distributions with equal means is typically rejected. The p-values for
the perceived amount of reverberation shown as black numbers in Table 7.1 indicate that the dual-Kalman and MWF have a statistically significant higher reverberation reduction that the WF, whereas between the dual-Kalman and MWF there is no statistical significance as the p-value is much larger than 0.05. The p-values for the overall quality shown as grey numbers in Table 7.1 indicate that the means of the three dereverberation methods are significantly different. Therefore, we have a clear ranking in terms of overall quality, where the dual-Kalman was rated best and the WF was rated worst.

The listening test results per acoustic condition are shown in Figures 7.10 and 7.11. Groups with statistically non-significantly different means, i.e. with p-values $p > 0.05$ are indicated by a rectangular bracket with an equality sign (=). In Figure 7.10 we can observe that in the longest reverberation time $T_{60} = 940$ ms, the reverberation reduction of the dual-Kalman is equal or below the MWF. In terms of overall quality shown in Figure 7.11, the dual-Kalman yields always the best results except in the longest $T_{60}$, where it is on par with the MWF. The reason for the slightly reduced performance of the dual-Kalman in longer $T_{60}$ is that the MAR filter length $L_{AR}$ was fixed and not set according to the acoustic scenario.

It is known from previous studies [20, 175] that the prediction of subjective performance measures using objective measures is challenging, especially in the context of dereverberation. Therefore, we test how well the objective measures used in this thesis correlate with the subjective ratings. Table 7.2 shows the Pearson correlation coefficient [180] between the improvements of the objective measures CD, LLR, PESQ, fwSSIR, QA_reverb and MFCCD, and the median of the subjective ratings on the differential scale for each condition as shown in Figure 7.9. Most
7.1. EVALUATION FOR SPEECH ENHANCEMENT

Objective measures show a high correlation with the perceived amount of reverberation, where $\Delta CD$, $\Delta LLR$, $\Delta wSSIR$ show a high correlation factor above 0.9, but $\Delta QAreverb$ and $\Delta SRMR$ are only moderately correlated. For the overall quality, none of the tested measures yields a
reliably high correlation above 0.9, but the best predictors for overall quality are as known from other studies $\Delta CD, \Delta LLR, \Delta PESQ$, although a high reliability is not given. One reason for the difficulty in predicting the overall quality could be that the overall quality is a highly subjective rating, which can vary a lot across different listeners depending on personal preferences. In the context of dereverberation, the $\Delta SRMR$ shows even a negative correlation with the overall quality. Although QAreverb and signal-to-reverberation-modulation ratio (SRMR) have been developed to predict the perceived amount of reverberation in speech signals, and they have been shown to deliver good predictions for unprocessed reverberant signals, their application to processed signals using dereverberation algorithms has to be handled with care and was not reliable in our experiments. The correlation results shown in Table 7.2 are roughly comparable with the results presented in the REVERB challenge [20] for the participating two- and eight-channel systems, while the one-channel systems showed significantly different correlation with the subjective ratings.

### 7.2 Evaluation for automatic speech recognition

As it is known that reverberation is harmful to the performance of ASR systems [19, 20], the word error rate (WER) of ASR systems can also be used to evaluate the performance of dereverberation algorithms. In the REVERB challenge [20], a dataset and framework to evaluate speech enhancement algorithms with focus on dereverberation was provided. We used the REVERB framework to evaluate the proposed dual-Kalman, MWF and WF dereverberation systems.

#### 7.2.1 REVERB challenge setup

The REVERB challenge framework comprises development and evaluation data sets with 3 simulated room conditions using measured RIRs with a speech signal database [181] and added recorded noise, and a real room condition reverberant speech signals were recorded directly in noisy conditions. The rooms in the simulated conditions had reverberation times of $T_{60} = \{300, 600, 700\}$ ms, and the room for real recordings had $T_{60} = 700$ ms. In each room two, source distances were measured, labelled as near and far, i.e. $\{0.5, 2\}$ m for the simulated rooms and $\{1, 2.5\}$ m for the real room, respectively. An 8-channel uniform circular microphone array with radius 10 cm was used for recording. The signals were pre-processed using the dereverberation and noise reduction algorithms under test, and the processed signal was then fed into the default REVERB challenge ASR system using the Hidden Markov Model Toolkit (HTK) [57] trained on clean speech. In our experiments, we used the default REVERB challenge ASR setup without modification.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>$\Delta CD$</th>
<th>$\Delta LLR$</th>
<th>$\Delta PESQ$</th>
<th>$\Delta fwSSIR$</th>
<th>$\Delta QAreverb$</th>
<th>$\Delta SRMR$</th>
<th>$\Delta MFCCD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rev. reduction</td>
<td>0.93</td>
<td>0.92</td>
<td>0.89</td>
<td>0.91</td>
<td>0.54</td>
<td>0.46</td>
<td>0.86</td>
</tr>
<tr>
<td>Overall quality</td>
<td>0.69</td>
<td>0.72</td>
<td>0.73</td>
<td>0.62</td>
<td>0.10</td>
<td>-0.29</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 7.2: Pearson correlation of subjective differential score with improvements of objective measures.
7.2. EVALUATION FOR AUTOMATIC SPEECH RECOGNITION

We evaluated the dual-Kalman with $M = 2$ microphones, the MWF for $M = 2$ and $M = 8$, and the single-channel WF with obviously $M = 1$ using this framework. In contrast to the official 2-channel setup of the REVERB challenge, which used the first two adjacent microphones spaced 4.1 cm, we used microphones 1 and 5 to have a larger spacing of 20 cm. The processing parameters were as follows: The sampling rate was $f_s = 16$ kHz, and the STFT analysis filterbank used $K = 512$ point DFT with square-root Hann windows of length 32 ms and a frame shift of $T_{\text{hop}} = 16$ ms. The noise PSD (matrix) was estimated using the speech presence probability-based approach proposed in [101]. The algorithm-specific parameters were set as described in Section 7.1.1. For the MWF, one DOA per time-frequency bin was estimated using beamspace root-MUSIC [111].

7.2.2 Automatic speech recognition results

The WER results of the development dataset are shown in Figure 7.12. For the real room conditions there is no clean signal available. We can observe large WER improvements in all reverberant conditions except for room1 sim near, which contains the least amount of reverberation. In this condition, only the 8-channel MWF achieves a slight improvement, the 2-channel MWF and dual-Kalman neither improve nor deteriorate the WER, and the WF slightly deteriorates the results. In the strongly reverberant conditions, the WER can be reduced by more than 40%. In general, the 8-channel MWF with the circular array achieves the best results, followed by the 2-channel dual-Kalman and the 2-channel MWF. Similar as in the previous experiments in Sections 7.1.2 and 7.1.3, the single-channel WF yields the lowest performance. According to the results obtained in Section 7.1.2.3, we can expect to achieve equal or better results than the MWF using the dual-Kalman with $M = 8$, but this setup was not computed due to its high computational complexity.

When designing front-end processors for ASR systems, it is often unclear, how to tune the dereverberation and noise reduction systems to achieve the best ASR performance, since the subjectively or objectively measured signal quality does not necessarily correlate with the ASR performance.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>$\Delta \text{CD}$</th>
<th>$\Delta \text{LLR}$</th>
<th>$\Delta \text{PESQ}$</th>
<th>$\Delta \text{fwSSIR}$</th>
<th>$\Delta \text{SRMR}$</th>
<th>$\Delta \text{MFCCD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta \text{WER}$</td>
<td>0.83</td>
<td>0.88</td>
<td>0.18</td>
<td>0.86</td>
<td>0.80</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 7.3: Pearson correlation of WER improvement with objective measure improvements.
performance. Therefore, we also computed the CD, LLR, PESQ, fwSSIR, SRMR, and MFCCD for the unprocessed and processed signals. Table 7.3 shows the Pearson correlation coefficients between the improvement of the speech enhancement measures and the WER improvement for the REVERB development set. It is not surprising to see that the WER is highly correlated with the MFCCD, which therefore is a useful performance predictor for ASR front-end processing systems. Also CD, LLR, fwSSIR, SRMR yield rather high correlation with the WER. Surprisingly, PESQ shows very low correlation for the used data, whereas for unprocessed reverberant speech, higher correlation has been shown between PESQ and WER [182].

7.3 Summary

In this chapter, the three proposed methods for dereverberation and noise reduction were compared in a number of experiments. These methods are

1. the multiple-input-multiple-output (MIMO) reverberation canceller based on the narrow-band MAR model using two alternating Kalman filters (Section 6.4) with reduction control (Section 6.5), referred to as dual-Kalman,

2. the MWF (Section 3.2.1) with reduction control (Section 3.3) using the blocking based reverberation PSD least-squares estimator (Section 5.3) with bias compensation (Section 5.5), referred to as MWF, and

3. the single-channel WF (Section 3.1.2) with reduction control (Section 3.3) using the RCTF based reverberation PSD estimator (Section 4.3), referred to as WF.

The multichannel approaches performed consistently better than the single-channel WF in terms of objective and subjective evaluation measures and ASR performance. We showed that the dual-Kalman has several advantages: the 2-channel dual-Kalman achieved the highest signal quality with an approximately equal amount of reverberation suppression to the MWF with a 4-channel linear array. The dual-Kalman is much less dependent on the array geometry and orientation than the MWF. While multichannel algorithms can generally achieve higher performance with large microphone spacings, in contrast to the dual-Kalman, the MWF performance is constrained by spatial aliasing issues at larger microphone spacings. The 2-channel dual-Kalman outperformed the MWF using an 8-channel linear array. Only when using multi-dimensional array geometries such as a UCA or significantly more microphones, the MWF can achieve comparable or higher performance than the dual-Kalman.

A listening test revealed that the 2-channel dual-Kalman yields a higher overall quality than the 4-channel MWF with the same amount of reverberation reduction. We found that the CD, LLR, PESQ and fwSSIR are consistently useful measures across multiple dereverberation methods showing a high correlation with subjective results. We showed that dereverberation and noise reduction could significantly increase the performance of an ASR system in reverberant conditions by pre-processing the signals. The MFCCD was found to be a good predictor for the ASR performance.
When speech communication devices such as mobile phones, hearing aids, and voice-controlled systems are used in distant talking scenarios in reverberant and noisy environments, the speech intelligibility, quality, and the automatic speech recognition performance are impaired. Dereverberation and noise reduction methods for such applications should be robust in time-varying environments, work with sufficiently small processing delay, and provide a good tradeoff between speech distortion and interference reduction.

8.1 Conclusion

An introduction to the dereverberation problem was given in Chapter 1. In this thesis, we proposed three different methods for dereverberation in the short-time Fourier transform (STFT) domain based on different signal model classes for single- and multi-channel setups. An overview of the used reverberation signal models is given in Chapter 2. All proposed methods are suitable for real-time frame-by-frame processing, and can quickly adapt to changes in the acoustic scene, which allows a direct application in speech communication systems.

The first two methods presented in Chapter 3 are using spectral and spatial Wiener filters (WFs), where the reverberation is modeled as an uncorrelated additive sound component in each time-frequency bin independently. A method to control the reduction of reverberation and noise independently at the output of a general multichannel Wiener filter (MWF) is proposed, where no assumption on the rank of the desired signal power spectral density (PSD) matrix is required.

A method to estimate the late reverberation PSD required for the spectral single-channel WF is presented in Chapter 4. The PSD estimation method is based on a narrowband moving average (MA) model using relative convolutive transfer functions (RCTFs). A single-channel WF is used to suppress the late reverberation and noise, which requires the PSDs of these signal components. A Kalman filter is used to estimate the RCTFs, which are then used to obtain an estimate of the late reverberation PSD. The RCTFs are explicitly modeled as time-varying random variables to account for time-varying acoustic conditions. The proposed RCTF based PSD estimator outperforms a state-of-the-art convolutive transfer function (CTF) based PSD estimator, and performs similar to late reverberation PSD estimators based on an exponential
decay model. In contrast to the exponential decay model, the proposed estimator does not require prior knowledge of the reverberation time and the direct-to-reverberation ratio (DRR).

A method to estimate the late reverberation PSD required for the MWF is presented in Chapter 5. The method is based on a spatial coherence model, where the desired signal is estimated using a MWF. The reverberation is modeled as a diffuse sound component with a time-varying PSD and a fixed spatial coherence, whereas the desired sound is often modeled as independent plane waves per time-frequency bin. The reverberation PSD is estimated by blocking the desired sound and estimating the reverberation PSD from an error PSD matrix using a least-squares procedure. A wide variety of multichannel reverberation PSD estimators is available in the literature that can be used in the MWF framework. Therefore, an extensive overview and benchmark comparing up to 10 different reverberation PSD estimators is provided. The comparison shows the strengths and weaknesses of the different solutions and underlying models. To mitigate a common weakness, namely the overestimation of the reverberation PSD at high DRR, we proposed a bias compensation method using a trained mapping function. The bias compensation could significantly reduce the speech distortion, while sacrificing only a small amount of interference reduction. Differences between the PSD estimators were mostly small or even non-significant, where slight trends for the trade-off between speech distortion and interference reduction could be observed. The most significant differences could be observed between the spatial coherence based estimators and temporal model based estimators, while some coherence based estimators performed below average. While the temporal model based estimators using optimal parameters yielded the best tradeoff between speech distortion and interference reduction, the coherence based estimators were slightly overestimating and therefore showed a tendency to more speech distortion. However, this problem could be mitigated by the bias compensation. We showed that in the presence of multiple sources, it is advantageous to use a multi-wave signal model. However only the blocking based subclass of the spatial coherence based estimators is generalizable to the multi-wave model with minor modifications. While in the presence of multiple sources using a multi-wave model increased the performance of the blocking based estimators, the tested temporal model estimators, which are independent of the modeled plane waves, still provided similarly useful results.

The third dereverberation method presented in Chapter 6 is a multiple-input-multiple-output (MIMO) reverberation canceller based on a narrowband multichannel autoregressive (MAR) signal model that removes reverberation and noise using a novel dual-Kalman filtering structure. In contrast to existing techniques based on this model, the proposed solution is an online processing method, explicitly models time-varying acoustic conditions, and integrates the MAR coefficient estimation and noise reduction stages in a causal manner, which is important in time-varying scenarios. The proposed dual-Kalman solution includes a single-Kalman and the state-of-the-art recursive least-squares (RLS) solution for the noise-free case as special cases. A state-of-the-art expectation-maximization (EM) method for dereverberation and noise reduction using a suboptimal structure for the time-varying case is outperformed in stationary and non-stationary acoustic conditions.

An evaluation presented in Chapter 7 comparing the three dereverberation algorithms showed a clear ranking of the methods. While the multichannel methods yielded consistently
better results than the single-channel method, both multichannel methods achieved similar results in terms of reverberation reduction. However, the dual-Kalman achieved a higher overall quality while using a smaller number of microphones. In contrast to the MWF, the dual-Kalman is rather robust against the array geometry and orientation, and requires no information on the array geometry and array calibration. For certain array geometries and a larger number of microphones, the MWF can achieve similar results to the dual-Kalman, with a lower computational complexity. In contrast to the dual-Kalman, the array geometry plays an essential role for the performance of the MWF.

The dereverberation algorithms developed in this thesis are shown to be suitable for use in typical noisy and slowly time-varying acoustic scenarios. Compared to state-of-the-art, the robustness, reduction performance, and quality could be substantially improved, while useful insights into properties and advantages of different STFT domain signal models for dereverberation were provided.

8.2 Future work

In future work, the complexity of the dual-Kalman method could be reduced to keep the complexity lower for a larger number of microphones. Its performance could be further improved by an adaptive selection of the MAR filter length. A cascade of the dual-Kalman with a spatial filtering approach such as the MWF can yield a system with further improved performance. In [35, 183, 184] the multichannel linear prediction framework has already been combined with spatial and spectral post-processors. However, it is an open question how the spatial filter after the dual-Kalman reverberation canceller can be designed optimally. Further work could be carried out by inherently combining multiple signal models, such as the spatial coherence model with temporal models such as the narrowband MA or MAR models, to develop novel dereverberation methods. As the vast majority of successful dereverberation methods mainly reduces only late reverberation, a major open research topic is the robust reduction of early reflections in time-varying noisy environments. Furthermore, the proposed dereverberation methods could be integrated into models for more complex signal processing tasks that usually rely on non-reverberant mixing models such as source localization [185] or source separation [186].
Appendix A

Signal energy ratio definitions for generating simulated signals

In the simulations presented in this thesis, the reverberant speech signals $x_m(t)$ and the noise $v_m(t)$ were generated independently. The noise was added to the reverberant signals with a defined ratio called input signal-to-noise ratio (iSNR). The $m$-th microphone signal is computed by

$$y_m(t) = x_m(t) + v_m(t).$$  \hfill (A.1)

Prior to the addition (A.1), the noise was scaled such that the iSNR, given by

$$\text{iSNR} = \frac{\sigma_x^2}{\sigma_v^2},$$  \hfill (A.2)

where $\sigma_x^2$ and $\sigma_v^2$ are the variances of $x_m(t)$ and $v_m(t)$, respectively, yields a defined value (typically given in dB). In the case of $M > 1$, all noise signals were scaled by the same factor as the first microphone channel.

In Section 5.6, we generated artificially composed sound fields consisting of stationary direct and diffuse sound components, denoted by $s_m(t)$ and $r_m(t)$. When the artificial noisy microphone signal $y_m(t)$ was generated by

$$y_m(t) = s_m(t) + r_m(t) + v_m(t),$$  \hfill (A.3)

prior to the addition, the signals $s_m(t)$ and $r_m(t)$ were scaled to a specific direct-to-reverberation ratio (DRR) given by

$$\text{DDR} = \frac{\sigma_s^2}{\sigma_r^2},$$  \hfill (A.4)

where $\sigma_s^2$ and $\sigma_r^2$ are the variances of $s_m(t)$ and $r_m(t)$, respectively.
Appendix B

Performance measures

In this chapter, the used measures for evaluation are presented. In Section B.1 signal based objective measures are presented that measure the signal quality or the enhancement from the processed signal, where most measures require knowledge of a target signal. In Section B.2 an error measure to evaluate the estimation accuracy of a power spectral density (PSD) is proposed.

B.1 Objective measures for speech enhancement

To assess speech enhancement algorithms that reduce reverberation and noise, it is often not straightforward to measure the dereverberation performance using room acoustical measures (also known as system based measures) described in the Section 1.1.2. These system based measures require the room impulse response (RIR), which can be measured only for linear time-invariant systems. However, typical short-time based processing methods may be non-linear and time-variant systems. Furthermore, we might want to assess the reduction of reverberation and noise jointly. For this reason, there exists a variety of signal based objective measures, which try to predict certain properties from a signal, which are related to the reduction of certain sound components or to perceptual aspects. These measures can be classified as intrusive measures, which require a target reference signal, and non-intrusive measures, which do not require a target reference signal.

In [20,175,187] investigations on the performance of such objective measures for the assessment of dereverberation algorithms have been presented. Unfortunately, a single measure does not exist, which can predict the subjective quality and effectiveness of dereverberation algorithms reliably for a wide range of conditions and algorithms. Therefore, we use a variety of measures predicting certain signal enhancement related aspects.

In addition, we propose a robust method to estimate the energy decay curve (EDC) from the direct sound and processed or unprocessed reverberant signals.

B.1.1 Signal-based speech enhancement measures

The following signal based measures have been widely used in the context of dereverberation to predict certain aspects from the processed and the target signals:
1. **Frequency-weighted segmental signal-to-interference ratio (fwSSIR) [173]:** The fwSSIR requires the target signal and the signal considered as interference to compute their power ratio using a Mel-frequency weighting. The amplitude of the interference signal is obtained by amplitude subtraction of the target signal from the input signal. The measure has been shown to yield high correlation with reverberation suppression and overall quality.

2. **Perceptual evaluation of speech quality (PESQ) [172]:** This is a measure to predict perceptual similarity of a degraded signal to a target signal, originally designed for assessment of telephony and audio coding applications.

3. **Quality assessment of reverberated speech (QAreverb) [188]:** This measure was developed to quantify the perceived amount of reverberation in speech signals. The amount of reverberation is predicted based on the estimated reverberation time $T_{60}$, the room spectral variance $\sigma_h^2$ and the DRR given by

$$Q_R = \frac{T_{60} \sigma_h^2}{\text{DRR}}. \quad (B.1)$$

These three parameters are estimated from the direct sound and the processed signal. In contrast to [188] we do not use a mapping to mean opinion score (MOS) as the provided mapping was trained with reverberant speech, but not with dereverberated speech.

4. **Normalized signal-to-reverberation-modulation ratio (SRMR) [174]:** A non-intrusive measure predicting the amount of reverberation only from the processed test signal. The measure exploits a energy ratio between different modulation frequency bands. The SRMR is given by

$$\text{SRMR} = \frac{\sum_k E_{\text{low}}(k)}{\sum_k E_{\text{high}}(k)}, \quad (B.2)$$

where $E_{\text{low}}(k)$ is the normalized modulation energy in low modulation frequency bands containing mainly speech energy, and $E_{\text{high}}(k)$ in higher modulation frequency bands containing mainly energy added by reverberation.

5. **Cepstral distance (CD) [173]:** This distance measure in the cepstrum domain mainly correlates well with speech distortion and overall quality, but also reverberation reduction.

6. **Log-likelihood ratio (LLR) [173]:** The measure is based on a statistical similarity test using an all-pole speech production model estimated by linear predictive coding (LPC). The measure is correlated with overall quality, amount of reverberation and speech distortion.

7. **Mel-frequency cepstral coefficient distance (MFCCD) [90]:** This is the squared normalized distance between the first 13 Mel-frequency cepstral coefficients (MFCCs) of test signal and target signal. This measure is primarily useful to predict the usefulness of dereverberation algorithms as pre-processing for automatic speech recognition (ASR) systems.

The intrusive measures fwSSIR, PESQ, QAreverb, CD, LLR and MFCCD compare the processed signal to the target signal, i.e. the direct sound or the early speech signal. The non-
B.1. OBJECTIVE MEASURES FOR SPEECH ENHANCEMENT

Intrusive measure SRMR does not require the target signal. While fwSSIR, PESQ, QAreverb are computed only for segments, where the target signal is active, CD, LLR and MFCCD are computed for all segments. In most cases, we compute the improvement of these measures compared to the unprocessed reference microphone signal, denoted by a \( \Delta \). The improvement is computed for each measure such positive values also indicate that the performance got better. If \( Q(\cdot) \) is the function computing the measure, then for measures where higher values are better, i.e. fwSSIR, PESQ, QAreverb and SRMR, the improvement is computed by

\[
\Delta Q = Q(\text{processed}) - Q(\text{unprocessed}).
\] (B.3)

For distance measures, where lower values are better, i.e. CD, LLR and MFCCD, the improvement is given by

\[
\Delta Q = Q(\text{unprocessed}) - Q(\text{processed}).
\] (B.4)

B.1.2 Signal-based EDC estimation

By comparing the EDC of unprocessed and processed RIRs, we can gain direct insight how the reverberation energy envelope is reduced and shaped by a dereverberation algorithm. Unfortunately, the dereverberated RIR can only be computed for linear time-invariant dereverberation systems, which is not the case for most block-based processing methods. Nevertheless, we can estimate the RIR from the direct sound and processed signals using speech signals by assuming that the signals were processed by a linear system. In this way, we can estimate a processed RIR at the reference microphone, that was obtained on average by the dereverberation system under test. From the estimated RIRs, we can finally compute the EDCs.

Given the direct sound signal \( s_d(t) \) and the processed signal \( z(t) \), we assume that \( z(t) \) was generated by the average response \( h(t) \) as

\[
z(t) = h(t) \ast s_d(t),
\] (B.5)

where \( \ast \) denotes convolution. We take overlapping segments of 2 s length of the processed and direct signals with a shift of 200 ms. By taking the discrete Fourier transform (DFT) of these segments, we obtain the frequency domain signals \( S_{d}(k,n) \) and \( Z(k,n) \), where \( k \) and \( n \) denote the frequency and time frame indices. The processed signal in the frequency domain is then given by

\[
Z(k,n) = H(k) S_{d}(k,n),
\] (B.6)

where \( H(k) \) is the DFT of \( h(t) \). We can estimate \( H(k) \) in the least-squares sense by

\[
\hat{H}(k) = \frac{\sum_{n \in T} Z(k,n) S_{d}^*(k,n)}{\sum_{n \in T} |S_{d}(k,n)|^2},
\] (B.7)

where the set \( T \) contains only frames, where the direct signal is present. Finally, we can compute the average processed RIR \( \hat{h}(t) \) as the inverse DFT of \( \hat{H}(k) \) and obtain the EDC using (1.2).
APPENDIX B. PERFORMANCE MEASURES

Figure B.1: Estimated EDC using direct speech and noisy reverberant speech signals in room with $T_{60} = 630$ ms and iSNR = 15 dB.

The following short experiment shows that estimation method of the EDC approximately holds also in the presence of additive noise. A reverberant and noisy speech signal of 15 s length was generated using a measured RIR with $T_{60} = 630$ ms and additive pink noise with iSNR = 15 dB. Figure B.1 shows the estimated EDC obtained from the reverberant noisy signal and the direct sound signal using (B.7) as black line. From the first 120 ms of the estimated EDC, the linear decay curve was extrapolated using a least-squares fit and is shown as dashed grey line. We can observe that the extrapolated linear decay crosses the $-60$ dB decay quite accurately at the reverberation time of the room, while the noise affects only the later tail of the EDC. Therefore, the EDC estimation procedure can give reasonable results also when there is residual noise left in the processed signals.

B.2 Logarithmic PSD estimation error

To evaluate estimated time- and frequency-dependent PSDs we employ the logarithmic PSD error given by

$$e(k,n) = 10 \log_{10} \frac{\hat{\phi}(k,n)}{\bar{\phi}(k,n)},$$

where $\hat{\phi}(k,n)$ is the estimated PSD and $\bar{\phi}$ is the reference PSD. The error $e(k,n)$ directly reflects over- and underestimation as positive and negative values in dB, respectively. The log error is analyzed statistically in terms of its mean $\mu_e$ and the lower and upper semi-variance [189]

$$\sigma_{e,1}^2 = \frac{1}{|T_1|} \sum_{(k,n) \in T_1} (e(k,n) - \mu_e)^2, \quad T_1 : e(k,n) \leq \mu_e \quad \text{(B.9a)}$$

$$\sigma_{e,u}^2 = \frac{1}{|T_u|} \sum_{(k,n) \in T_u} (e(k,n) - \mu_e)^2, \quad T_u : e(k,n) > \mu_e \quad \text{(B.9b)}$$

where the sets of time-frequency bins $T_1$ and $T_u$ contain all bins below or above the mean, respectively, where a value in $\bar{\phi}(k,n)$ is present. Therefore, a log error with zero mean and small semi-standard deviations $\sigma_{e,1}$ and $\sigma_{e,u}$ is desired. In figures showing the log error, the mean is represented by symbols (circle, square, etc.) and the semi-standard deviations are indicated by whisker bars.
Appendix C

Appendix to Chapter 6:
Computation of residual noise and reverberation

To compute residual power of noise and reverberation at the output of the proposed system, we need to propagate these signals through the system. By propagating only the noise at the input $v(n)$ through the dual-Kalman system instead of $y(n)$ as in Figure 6.4, we obtain the output $\hat{s}_v(n)$, which is the residual noise contained in $\hat{s}(n)$. By also taking the reduction control (RC) into account, the residual contribution of the noise $v(n)$ in the output signal $z(n)$ is $z_v(n)$. By inspecting (6.59), (6.61) and (6.63), the noise is fed through the noise reduction Kalman filter by the equation

$$
\hat{v}(n) = F(n)\hat{v}(n-1) + K_x(n) [v(n) - HF(n)\hat{v}(n-1)] \\
= K_x(n)v(n) + [F(n) - K_x(n)HF(n)]\hat{v}(n-1),
$$

(C.1)

where $\hat{v}(n)$ is the residual noise vector of length $ML$, similarly defined as (6.35), after noise reduction. The output after the dereverberation step is obtained by

$$
\hat{s}_v(n) = H\hat{v}(n) - HF(n)\hat{v}(n-1) .
$$

(C.2)

With RC, the residual noise is given in analogy to (6.71) by

$$
z_v(n) = \beta_v v(n) + (1 - \beta_v)\hat{v}(n) - (1 - \beta_r)\hat{v}(n|n-1).
$$

(C.3)
The calculation of the residual reverberation $z_r(n)$ is more difficult. To exclude the noise from this calculation, we first feed the oracle reverberant noise-free signal vector $x(n)$ through the noise reduction stage:

$$\tilde{x}(n) = \mathbf{F}(n)\tilde{x}(n-1) + \mathbf{K}_x(n)[x(n) - \mathbf{H}\mathbf{F}(n)\tilde{x}(n-1)]$$
$$= \mathbf{K}_x(n)x(n) + [\mathbf{F}(n) - \mathbf{K}_x(n)\mathbf{H}\mathbf{F}(n)]\tilde{x}(n-1),$$  \hspace{1cm} (C.4)

where $\tilde{x}(n) = \mathbf{H}\tilde{x}(n)$ is the output of the noise-free signal vector $x(n)$ after the noise reduction stage. According to (6.71) the output of the noise-free signal vector after dereverberation and RC is obtained by

$$z_x(n) = \beta_v x(n) + (1 - \beta_v)\tilde{x}(n) - (1 - \beta_r)\tilde{r}(n)$$ \hspace{1cm} (C.5)

where $\tilde{r}(n) = \tilde{X}(n-D)\tilde{c}(n)$ and the matrix $\tilde{X}(n)$ is obtained using $\tilde{x}(n)$ in analogy to (6.2).

Now let us assume that the noise-free signal vector after the noise reduction $\tilde{x}(n)$ and the noise-free output signal vector after dereverberation and RC $z_x(n)$ are composed as

$$\tilde{x}(n) \approx s(n) + r(n)$$ \hspace{1cm} (C.6)
$$z_x(n) \approx s(n) + z_r(n),$$ \hspace{1cm} (C.7)

where $z_r(n)$ denotes the residual reverberation in the RC output $z(n)$. By using (C.6) and knowledge of the oracle desired signal vector $s(n)$, we can compute the reverberation signal

$$r(n) = \tilde{x}(n) - s(n).$$ \hspace{1cm} (C.8)

From the difference of (C.6) and (C.7) and using (C.8), we can obtain the residual reverberation signals as

$$z_r(n) = r(n) - \left[\frac{\tilde{x}(n) - z_x(n)}{r(n) - z_r(n)}\right].$$ \hspace{1cm} (C.9)

Now we can analyze the power of residual noise and reverberation at the output and compare it to their respective power at the input.
Bibliography


Bibliography


Bibliography


[176] [Online]. Available: https://github.com/ehabets/Signal-Generator


