Model-Based Speech Enhancement
Exploiting Temporal and Spectral Dependencies

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

vorgelegt von

Diplom-Ingenieur

Thomas Esch

aus Mechernich

Berichter: Universitätsprofessor Dr.-Ing. Peter Vary
Universitätsprofessor Dr.-Ing. Rainer Martin

Tag der mündlichen Prüfung: 30. Januar 2012

AACHENER BEITRÄGE ZU DIGITALEN NACHRICHTENSYSTEMEN

Herausgeber:

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

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

1. Auflage Aachen:
Wissenschaftsverlag Mainz in Aachen
(Aachener Beiträge zu digitalen Nachrichtensystemen, Band 32)
ISSN 1437-6768
ISBN 3-86130-359-0

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

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

Gedruckt auf chlorfrei gebleichtem Papier

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

Acknowledgments

This thesis was written during my time as research assistant at the *Institute of Communication Systems and Data Processing (IND)* at the *Rheinisch-Westfälische Technische Hochschule Aachen (RWTH Aachen University)*. Many results and achievements of this work would not have been possible unless the support of many people. It is my great pleasure to take the opportunity to thank them here.

First of all, I would like to express my deep and sincere gratitude to my supervisor Prof. Dr.-Ing. Peter Vary. His continuous support, precious advices and fruitful discussions have significantly contributed to the success of this thesis. I am also indebted to Prof. Dr.-Ing. Rainer Martin for being the co-supervisor of this thesis and showing much interest in the obtained results.

Furthermore, I want to thank all my current and former colleagues as well as the permanent staff at the IND for providing a pleasant and enjoyable working environment. In particular, I am grateful to Dr.-Ing. Christiane Antweiler, Dipl.-Ing. Moritz Beermann, Dipl.-Math. Annika Böttcher, Dipl.-Ing. Tobias Breddermann, Dipl.-Ing. Benedikt Eschbach, Dipl.-Ing. Bernd Geiser, Dipl.-Ing. Florian Heese, M. Sc. Marco Jeub, Dr.-Ing. Hauke Krüger, Dr.-Ing. Heinrich Löllmann, Dr.-Ing. Helge Lüders, Dipl.-Ing. Christoph Nelke, Dipl.-Ing. Matthias Pawig, Dipl.-Ing. Matthias Rüngeler, Dipl.-Ing. Bastian Sauert, Dipl.-Ing. Magnus Schäfer, Dipl.-Ing. Thomas Schlien, Dr.-Ing. Laurent Schmalen, Dipl.-Ing. Birgit Schotsch, Dr.-Ing. Aulis Telle. Of course, I would also like to express my appreciation to all the students who made significant contributions to this work.

Special thanks go to Nokia Research Center in Tampere (Finland) for providing funding and for the good collaboration.

I owe my loving thanks to my family, in particular, my parents Walter and Doris Esch as well as my sister Christina for their personal support over the years. Moreover, I am thankful to my parents-in-law Josef and Gertrud Siepen.

Last but not least, I am deeply grateful to my wonderful wife Monika. Thank you for your love, encouragement, patience, and understanding.

Cologne, February 2012 Thomas Esch

Abstract

Mobile telephony has become an integral part of everyday life for billions of people around the world. The exchange of information via speech is nowadays possible from almost all places at anytime. However, even though the vision of permanent reachability and connectivity has been realized in the meantime nearly worldwide, there is still room for improvements when it comes to the transmission of speech under noisy conditions. The performance of any speech communication system may significantly deteriorate when the speech signal is disturbed by ambient interferences such as traffic noise or office noise, possibly leading to a poor speech quality and intelligibility.

In this thesis, a novel model-based speech enhancement system is presented which performs single-channel noise reduction of degraded speech signals. In contrast to state-of-the-art noise suppression techniques, the developed algorithms explicitly exploit *temporal and spectral dependencies* of speech and noise signals. To account for the temporal correlation, a *modified Kalman filter* is derived in the frequency domain. As main novelties, the proposed solution performs complex-valued prediction of speech and noise DFT coefficients and uses SNR-dependent MMSE estimators which are adapted to measured statistics of the input signal. In order to incorporate the spectral dependencies of speech signals, a new wideband speech enhancement system is presented which utilizes techniques known from *artificial bandwidth extension*. The developed method re-uses the processed and enhanced signal from lower frequencies to improve the results of a conventional noise suppression technique at higher frequencies. As additional part, this work proposes effective countermeasures to reduce the occurrence of *musical noise* and provides a novel solution for the suppression of *rapidly time-varying harmonic noise*.

All developed speech enhancement techniques within this thesis are thoroughly evaluated by means of instrumental measurements and auditory judgments. It turns out that the proposed algorithms achieve distinctly better results compared to state-of-the-art approaches with respect to noise attenuation and speech distortions. The novel model-based system is not restricted to the application in mobile phones. It can be used in addition to improve the speech quality of hands-free devices, conferencing systems or digital hearing aids.

Contents

Notations, Symbols & Abbreviations

Mathematical Notation

In this thesis, the following conventions are used: capital bold letters refer to matrices, e.g., $\mathbf{X}$, vectors are written in bold letters, e.g., $\mathbf{x}$ and scalars are not bold, e.g., x. Quantities in the time domain are usually written in lower-case letters, e.g., $x(k)$ while quantities in the frequency domain are written in upper-case letters, e.g., $X(\mu)$. Estimated quantities are labeled by a hat, e.g., $\hat{x}$.

List of Principal Symbols

Latin Symbols

$\mathbf{A}$	speech transition matrix
A	magnitude of short-time speech DFT coefficient
$\mathbf{a}_{\mathrm{hb}}$	feature vector representing spectral envelope of high band signal
a_κ	κ-th AR coefficient of speech model
$\hat{A}_{\mathrm{prop}}$	magnitude estimate of short-time speech DFT coefficient in propagation step
$\mathbf{a}_{\mathrm{wb}}$	feature vector representing spectral envelope of wideband signal
$\mathbf{B}$	noise transition matrix
B	Minimum Statistics bias correction factor
b	Bark band index
b_{hb}	dimension of $\mathbf{a}_{\mathrm{hb}}$
B_{l}	lower frequency bin limit
b_{lb}	dimension of $\mathbf{x}_{\mathrm{lb}}$
b_τ	τ-th AR coefficient of noise model
B_{u}	upper frequency bin limit
b_{wb}	dimension of $\mathbf{a}_{\mathrm{wb}}$
$c_{\mathrm{dlb},j}$	j-th cepstral coefficient in $\mathbf{x}_{\mathrm{dlb}}$
$c_{\mathrm{lb},j}$	j-th cepstral coefficient in $\mathbf{x}_{\mathrm{lb}}$
D	differential signal in frequency domain

d	differential signal in time domain (speech and noise prediction)
$d_{\mathrm{hb}}^{\mathrm{LSD}}$	log spectral distortion in high band
$d_{\mathrm{lb}}^{\mathrm{LSD}}$	log spectral distortion in low band
D_{MS}	Minimum Statistics search window length
d_s	differential signal in time domain (only speech prediction)
$\mathbb{E}\{\cdot\}$	expectation operator
e	Euler number
e_{lb}	low band speech excitation signal
E_N	noise excitation signal in frequency domain
e_n	noise excitation signal in time domain
$\mathbf{E}_{\mathrm{prop}}^N$	vector containing noise prediction errors in frequency domain (propagation step)
E_{prop}^N	noise prediction error in frequency domain (propagation step)
$\mathbf{e}_{\mathrm{prop}}^n$	vector containing noise prediction errors in time domain (propagation step)
e_{prop}^n	noise prediction error in time domain (propagation step)
$\mathbf{E}_{\mathrm{prop}}^S$	vector containing speech prediction errors in frequency domain (propagation step)
E_{prop}^S	speech prediction error in frequency domain (propagation step)
$\mathbf{e}_{\mathrm{prop}}^s$	vector containing speech prediction errors in time domain (propagation step)
e_{prop}^s	speech prediction error in time domain (propagation step)
E_S	speech excitation signal in frequency domain
e_s	speech excitation signal in time domain
$\mathbf{E}_{\mathrm{up}}^N$	vector containing noise estimation errors in frequency domain (update step)
E_{up}^N	noise estimation error in frequency domain (update step)
$\mathbf{e}_{\mathrm{up}}^n$	vector containing noise estimation errors in time domain (update step)
e_{up}^n	noise estimation error in time domain (update step)
$\mathbf{E}_{\mathrm{up}}^S$	vector containing speech estimation errors step in frequency domain (update step)
E_{up}^S	speech estimation error in frequency domain (update step)
$\mathbf{e}_{\mathrm{up}}^s$	vector containing speech estimation errors in time domain (update step)
e_{up}^s	speech estimation error in time domain (update step)
e_{wb}	wideband speech excitation signal
$\exp(\cdot)$	exponential function
f_0	fundamental frequency of harmonic noise
f_{s}	sampling frequency
G	spectral weighting gain
G_{h}	spectral weighting gain for harmonic noise reduction
G_{hb}	overall spectral weighting gain in high band
$G_{\mathrm{hb}}^{\mathrm{bwe}}$	spectral weighting gain in high band determined by ABWE techniques
$G_{\mathrm{hb}}^{\mathrm{conv}}$	spectral weighting gain of conventional noise reduction in high band
$G_{\mathrm{hb}}^{\mathrm{opt}}$	optimum spectral weighting gain in high band
$\mathbf{g}_N$	unit vector of dimension M_{K}
$\mathbf{g}_n$	unit vector of dimension M_{K}

G_{p}	spectral weighting gain of postprocessing technique
$G_{\mathrm{P},N}$	noise prediction gain
$G_{\mathrm{P},N,\mathrm{max}}$	maximum noise prediction gain for a given autocorrelation length L'_{AC}
$G_{\mathrm{P},S}$	speech prediction gain
$G_{\mathrm{P},S,\mathrm{max}}$	maximum speech prediction gain for a given autocorrelation length L_{AC}
G_{s}	spectral weighting gain for stationary noise reduction
$\mathbf{g}_S$	unit vector of dimension N_{K}
$\mathbf{g}_s$	unit vector of dimension N_{K}
h	differential entropy
H_{MA}	impulse respond corresp. to moving average window of length L_{MA}
$\mathbf{h}_N$	unit vector of dimension M_{K}
$\mathbf{h}_n$	unit vector of dimension M_{K}
$\mathbf{h}_S$	unit vector of dimension N_{K}
$\mathbf{h}_s$	unit vector of dimension N_{K}
$\mathbf{I}$	identity matrix
I	mutual information
$\mathrm{Im}\{\cdot\}$	imaginary part operator
J	Kullback Leibler distance
j	imaginary unit
k	discrete time index at $f_{\mathrm{s}} = 8\,\mathrm{kHz}$
k'	discrete time index at $f_{\mathrm{s}} = 16\,\mathrm{kHz}$
$\mathbf{K}^N$	Kalman filter gain for noise signal in frequency domain
$\mathbf{k}^n$	Kalman filter gain for noise signal in time domain
$\mathbf{K}^S$	Kalman filter gain for speech signal in frequency domain
$\mathbf{k}^s$	Kalman filter gain for speech signal in time domain
L_{AC}	length of speech autocorrelation function
L'_{AC}	length of noise autocorrelation function
L_{F}	frame length
L_{FS}	frame shift size
L_{MA}	length of moving average window
$\log_x(\cdot)$	logarithmic function to base x
M_{B}	number of Bark bands
M_{F}	FFT length at full frequency resolution
M'_{F}	FFT length at reduced frequency resolution
M_{GM}	number of GMM mixture components
M_{K}	model order of noise process
N	short-time DFT coefficient of noise signal
$\mathbf{n}$	state vector containing noise samples
n	noise signal
N_{bwe}	number of HMM states
N_{C}	number of MFCCs used in $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{x}_{\mathrm{dlb}}$
N_{h}	short-time DFT coefficient of harmonic noise signal
n_{h}	harmonic noise signal
N_{hb}	short-time DFT coefficient of high band noise signal
$\mathbf{n}_{\mathrm{hb}}$	vector containing noise samples in high band

N_{lb}	short-time DFT coefficient of low band noise signal
$\mathbf{n}_{\mathrm{lb}}$	vector containing noise samples in low band
n_{lb}	noise signal in low band
N_{K}	model order of speech process
$\hat{\mathbf{N}}_{\mathrm{prop}}$	vector containing noise estimates in frequency domain (propagation step)
$\hat{N}_{\mathrm{prop}}$	noise estimate in frequency domain (propagation step)
$\hat{\mathbf{n}}_{\mathrm{prop}}$	vector containing noise estimates in time domain (propagation step)
$\hat{n}_{\mathrm{prop}}$	noise estimate in time domain (propagation step)
$\hat{\mathbf{N}}_{\mathrm{up}}$	vector containing noise estimates in frequency domain (update step)
$\hat{N}_{\mathrm{up}}$	noise estimate in frequency domain (update step)
$\hat{\mathbf{n}}_{\mathrm{up}}$	vector containing noise estimates in time domain (update step)
$\hat{n}_{\mathrm{up}}$	noise estimate in time domain (update step)
N_{s}	short-time DFT coefficient of stationary noise signal
n_{s}	stationary noise signal
$P(\cdot)$	probability mass function
$p(\cdot)$	probability density function
$\mathbf{P}^{N}_{\mathrm{prop}}$	noise error covariance matrix in frequency domain (propagation step)
$\mathbf{P}^{n}_{\mathrm{prop}}$	noise error covariance matrix in time domain (propagation step)
$\mathbf{P}^{S}_{\mathrm{prop}}$	speech error covariance matrix in frequency domain (propagation step)
$\mathbf{P}^{s}_{\mathrm{prop}}$	speech error covariance matrix in time domain (propagation step)
$\mathbf{P}^{N}_{\mathrm{up}}$	noise error covariance matrix in frequency domain (update step)
$\mathbf{P}^{n}_{\mathrm{up}}$	noise error covariance matrix in time domain (update step)
$\mathbf{P}^{S}_{\mathrm{up}}$	speech error covariance matrix in frequency domain (update step)
$\mathbf{P}^{s}_{\mathrm{up}}$	speech error covariance matrix in time domain (update step)
R	magnitude of short-time noisy DFT coefficient
$\mathrm{Re}\{\cdot\}$	real part operator
S	short-time DFT coefficient of speech signal
$\mathbf{s}$	state vector containing speech samples
s	speech signal
$\hat{S}_{\mathrm{h}}$	estimated short-time speech DFT coefficient after harmonic noise reduction
S_{hb}	short-time DFT coefficient of speech signal in high band
s_{hb}	speech signal in high band
$\mathcal{S}_i$	i-th HMM state
$\hat{\mathbf{S}}_{\mathrm{lb}}$	vector containing short-time DFT coefficients of estimated low band speech signal
S_{lb}	short-time DFT coefficient of speech signal in low band
s_{lb}	speech signal in low band
SNR_b	SNR in b-th Bark band
$\hat{S}_{\mathrm{p}}$	estimated short-time speech DFT coefficient after postprocessing
$\hat{s}_{\mathrm{p}}$	estimated speech signal after postprocessing
$\hat{\mathbf{S}}_{\mathrm{prop}}$	vector containing speech estimates in frequency domain (propagation step)
$\hat{S}_{\mathrm{prop}}$	speech estimate in frequency domain (propagation step)
$\hat{\mathbf{s}}_{\mathrm{prop}}$	vector containing speech estimates in time domain (propagation step)

$\hat{s}_{\text{prop}}$	speech estimate in time domain (propagation step)
$\hat{\mathbf{S}}_{\text{up}}$	vector containing speech estimates in frequency domain (update step)
$\hat{S}_{\text{up}}$	speech estimate in frequency domain (update step)
$\hat{\mathbf{s}}_{\text{up}}$	vector containing speech estimates in time domain (update step)
$\hat{s}_{\text{up}}$	speech estimate in time domain (update step)
s_{wb}	wideband speech signal
$\mathbf{x}_{\text{dlb}}$	observation vector representing disturbed envelope of low band signal
$\mathbf{X}_{\text{lb}}$	sequence of observation vectors $\mathbf{x}_{\text{lb}}$
$\mathbf{x}_{\text{lb}}$	observation vector representing spectral envelope of low band signal
Y	short-time DFT coefficient of noisy signal
$\mathbf{y}$	vector containing noisy samples
y	noisy signal
Y_{hb}	short-time DFT coefficient of noisy signal in high band
y_{hb}	noisy signal in high band
Y_{lb}	short-time DFT coefficient of noisy signal in low band
y_{lb}	noisy signal in low band
$\hat{Y}_{\text{prop}}$	noisy estimate in frequency domain (propagation step)
$\hat{y}_{\text{prop}}$	noisy estimate in time domain (propagation step)

Greek Symbols

α_{DD}	decision-directed smoothing factor
α_{hb}	cross-fading factor in high band
$\bar{\alpha}_{\text{hb}}$	averaged cross-fading factor in high band
$\alpha_{\text{hb}}^{\text{ref}}$	reference cross-fading factor in high band
$\alpha_{\text{K}}^{\text{DD}}$	decision-directed smoothing factor in update step
α_{MS}	Minimum Statistics smoothing factor
$\Gamma(\cdot)$	Gamma function
γ	a posteriori SNR
γ_{h}	a posteriori SNR required for harmonic noise reduction
γ_{hb}	a posteriori SNR in high band
$\gamma_{\text{hb}}^{\text{opt}}$	optimum a posteriori SNR in high band
γ_{K}	a posteriori SNR in update step of Kalman filter
γ_{s}	a posteriori SNR required for stationary noise reduction
δ	model parameter in generalized Gamma PDF
ϵ_0	constant required for adaptive bandwidth resolution solution
ϵ_{B}	threshold required for adaptive bandwidth resolution solution
θ	model parameter in generalized Gamma PDF
ϑ	phase of short-time noisy DFT coefficient
λ	frame index
μ	frequency index
μ'	frequency index in subsampled frequency domain
ξ	a priori SNR
ξ_{h}	a priori SNR required for harmonic noise reduction

ξ_{hb}	a priori SNR in high band
$\xi_{\mathrm{hb}}^{\mathrm{opt}}$	optimum a priori SNR in high band
$\bar{\xi}_{\mathrm{lb}}^{\mathrm{opt}}$	averaged optimum a priori SNR in low band
ξ_{K}	a priori SNR in update step of Kalman filter
ξ_{s}	a priori SNR required for stationary noise reduction
ρ	model parameter in generalized Gamma PDF
σ_D^2	power of differential signal in frequency domain
$\sigma_{d_s}^2$	power of differential signal in time domain
$\sigma_{E_N}^2$	power of noise excitation signal in frequency domain
$\sigma_{e_n}^2$	power of noise excitation signal in time domain
$\sigma_{E_{\mathrm{prop}}^N}^2$	power of noise prediction error signal in frequency domain
$\sigma_{e_{\mathrm{prop}}^n}^2$	power of noise prediction error signal in time domain
$\sigma_{E_{\mathrm{prop}}^S}^2$	power of speech prediction error signal in frequency domain
$\sigma_{e_{\mathrm{prop}}^s}^2$	power of speech prediction error signal in time domain
$\sigma_{E_S}^2$	power of speech excitation signal in frequency domain
$\sigma_{e_s}^2$	power of speech excitation signal in time domain
σ_N^2	power spectral density of noise signal
σ_n^2	power of noise signal
$\sigma_{\mathrm{h},N}^2$	power spectral density of harmonic noise signal
$\sigma_{n_{\mathrm{lb}}}^2$	variance of observation vector in low band
$\sigma_{\mathrm{s},N}^2$	power spectral density of stationary noise signal
σ_S^2	power spectral density of speech signal
σ_s^2	power of speech signal
$\hat{\boldsymbol{\Sigma}}_Y^2$	matrix containing estimates of noisy power spectral densities
$\hat{\boldsymbol{\sigma}}_Y^2(\mu)$	μ-th row of $\hat{\boldsymbol{\Sigma}}_Y^2$
σ_Y^2	power spectral density of noisy signal
σ_y^2	power of noisy signal y
$\hat{\boldsymbol{\Sigma}}_{Y,\mathrm{mod}}^2$	modified version of $\hat{\boldsymbol{\Sigma}}_Y^2$
$\hat{\boldsymbol{\sigma}}_{Y,\mathrm{mod}}^2(\mu)$	modified version of $\hat{\boldsymbol{\sigma}}_Y^2(\mu)$
ϕ	phase of short-time speech DFT coefficient
$\hat{\phi}_{\mathrm{prop}}$	phase estimate of short-time speech DFT coefficient in propagation step
χ	scaling factor required for postfilter of spectral weighting gains
ψ	power ratio required for low SNR detection
ψ_{T}	modified version of ψ dependent on ψ_{thr}
ψ_{thr}	threshold required for postfilter of spectral weighting gains

Other Symbols

!	factorial operator
$*$	convolution operator
$\angle\{\cdot\}$	phase operator
$\lvert\cdot\rvert$	magnitude operator

List of Abbreviations

ABWE	Artificial Bandwidth Extension
AK	A Priori Knowledge
AMR	Adaptive Multi-Rate
AR	Autoregressive
AWGN	Additive White Gaussian Noise
CCR	Comparison Category Rating
DFT	Discrete Fourier Transform
EM	Expectation Maximization
ETSI	European Telecommunications Standards Institute
FIR	Finite Impulse Response
FFT	Fast Fourier Transform
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IFFT	Inverse Fast Fourier Transform
IIR	Infinite Impulse Response
ITU-T	International Telecommunication Union -Telecommunication Standardization Sector
LMS	Least Mean Square
LP	Linear Prediction
LPC	Linear Predictive Coding
LS	Least-Square
LSD	Log Spectral Distortion
LSA	Log Spectral Amplitude
LTP	Long Term Prediction
MAP	Maximum A Posteriori
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MS	Minimum Statistics
PDF	Probability Density Function
PMF	Probability Mass Function
PF	Postfilter
PSD	Power Spectral Density
QMF	Quadrature Mirror Filter
SNR	Signal-to-Noise-Ratio
STFT	Short-Time Fourier Transform
STSA	Short-Time Spectral Amplitude
VAD	Voice Activity Detection
VQ	Vector Quantizer
WGN	White Gaussian Noise
WMOPS	Weighted Million Operations per Second
ZCR	Zero Crossing Rate

1

Introduction

Since time immemorial, speech is one of the most important communication forms of humanity. While in former times conversations were possible only face-to-face, the invention of the telephone was a breakthrough into the era of telecommunication. Since that time, people are able to make long-distance calls and to communicate with other persons from all around the world. In the course of time, the telecommunication sector has become a very important asset in economy. With the progress in technology and the increasing demand for permanent reachability and connectivity, the exchange of information via speech is feasible nowadays from anywhere at anytime. The use of mobile phones has become an inherent part in everyday life.

In order to ensure a high transmission quality in mobile telephone networks, digital signal processing plays a very important role. The increasing computational performance of technical platforms allows the realization of more and more sophisticated and complex algorithms in mobile phones. An effective and efficient concatenation of all parts within the transmission chain from the acoustic front-end to the radio link leads to a satisfying end-to-end speech quality in the ideal case. However, the quality as well as the intelligibility of any speech communication system may significantly deteriorate if the input signal is degraded by ambient interferences such as echoes, background noises and reverberation. A possible degradation may have severe influence on the required listening effort. Depending on the *Signal-to-Noise-Ratio* (SNR), interferences make a conversation uncomfortable or even impossible in the worst case. In order to cope with such acoustic environments, speech enhancement algorithms are meanwhile implemented in many digital speech communication systems. These algorithms aim at reducing echoes, background noises or reverberation by means of digital signal processing without affecting the speech signal. In literature, a large number of different solutions can be found for speech enhancement. In general, the approaches can be separated into two main classes: *single* and *multi-sensor* (microphone) systems. While multi-sensor speech enhancement systems can additionally exploit spatial properties of the environment, e.g., by beamforming or adaptive noise cancellation, single-sensor systems are restricted to one microphone signal and usually rely on *A Priori Knowledge* (AK) of speech and interference.

This thesis covers single-sensor speech enhancement in noisy environments where the speech signal is disturbed by additive background noises like traffic noise, office

noise or 'babble noise' of other speakers. Therefore, the terms speech enhancement, noise reduction and noise suppression are used interchangeably from now on. A typical application example is depicted in Fig. 1.1. A person is standing in a noisy environment and wants to communicate with another person using a mobile phone. The noisy signal y, consisting of the clean speech signal s and the environmental noise n, is captured by the phone's microphone. Before the signal is transmitted over the radio channel, noise suppression is applied resulting in an estimate $\hat{s}$ of the clean speech signal at the output. Afterwards, the enhanced signal $\hat{s}$ is further processed by speech coding, channel coding and modulation and is finally transmitted to the far-end person. If the noise suppression works successfully, the background noise is effectively reduced, leading to a much more comfortable listening condition. The utilization of noise reduction algorithms is not only limited to mobile phones. Further application areas are, e.g., hands-free devices, conferencing systems, digital hearing aids and speech recognition systems.

1.1 Related Works

In literature, a large variety of different approaches can be found for the purpose of single-channel noise reduction. Overviews can be found, e.g., in [VM06, VHH98, BCHC09, HS06, Loi07, Ben07, Dav02]. The first approaches go back to the year 1965.

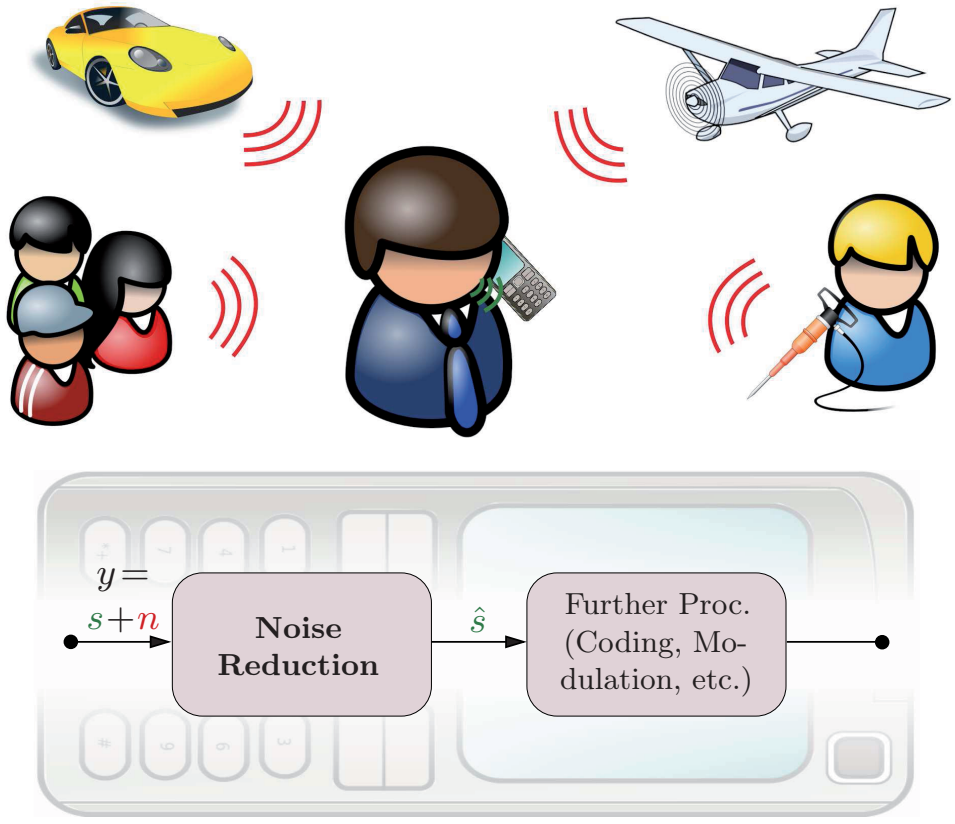

Figure 1.1: Application example for noise reduction by means of a mobile phone.

In [Sch65], the first patent on spectral subtraction was published and was realized as an analog circuit. However, noise reduction did not gather momentum until the digital age. *Digital Signal Processors* (DSPs) paved the way to develop and implement sophisticated algorithms in the digital domain. The first digital techniques can be found in [Bol79, LO79, MM80] and are based on the spectral subtraction approach and the Wiener filtering method. In principle, all solutions can be divided into two categories: *transformation techniques* and *model-based* approaches.

The transformation techniques transform the noisy input signal in an adequate domain where speech and noise can be separated in a simpler way by exploiting appropriate signal characteristics. One of the most utilized domains is the *Fourier domain* due to the efficient implementation of the *Discrete Fourier Transform* (DFT). In this domain, a statistical estimation framework is usually used, which applies individual adaptive weighting gains to either the complex-valued noisy DFT coefficients or the real-valued noisy magnitudes where the phase of the input signal is maintained. In order to derive the weighting gains, in most cases a specific distortion measure is minimized relying on a mathematical cost function like *Minimum Mean Square Error* (MMSE), *Maximum A Posteriori* (MAP) or *Maximum Likelihood* (ML). Moreover, certain assumptions on the statistics of the speech and noise signals are made. While a Gaussian model is often used for the noise signal, the distribution of the speech signal is typically modeled either as Gaussian or super-Gaussian. Distinguished solutions in this area can be found, e.g., in [EM84, EM85, Var85, Mar05, LV05, EHHJ07, BKM08]. The application of a DFT provides a uniform resolution in the frequency domain. In order to adjust the spectral resolution with respect to psychoacoustical criteria, e.g., wavelet-based transforms [GEH98, LGO$^+$96, SB97] or allpass transformed DFT filter-banks [HS08, Chapter 2] can be used. These approaches make non-uniform time-frequency resolutions possible which can be adapted for instance to the well-known Bark scale [ZF90]. Another technique, which belongs to this important category of noise suppression algorithms, is the so-called *subspace approach*. Within the subspace, it is assumed that the noisy signal can be represented as speech-plus-noise subspace and noise-only subspace. Hence, the objective is to eliminate the noise-only subspace while reconstructing the speech signal from the remaining speech-plus-noise subspace. Common subspace transforms are, e.g., the Karhunen-Loève transform as well as the singular value decomposition. Techniques based on subspace decomposition are proposed, e.g., in [DBC91, EVT95, HL03].

In contrast to the transformation techniques, model-based approaches *additionally* take into account models of the human speech production process or the human auditory system in order to further improve the noise reduction performance. The model-based approaches include, e.g., psychoacoustically-based techniques [TPM93, GMJV02], the application of *Hidden Markov Models* (HMMs) [EMJ89, Eph92] and Kalman filtering techniques [PB87, WC98, Pud02].

1.2 Structure of this Thesis

In the first part of this work, a model-based speech enhancement system is proposed which focuses on the exploitation of *temporal and spectral dependencies* of speech as well as noise signals. In the second part, two well-known problems of state-of-the-art noise reduction techniques are addressed, namely the reduction of *musical noise* and speech enhancement in highly *non-stationary noise environments*. The remainder of this thesis is divided into 5 chapters as follows.

Chapter 2 provides a brief overview of noise suppression techniques working in the frequency domain. The basic principles of statistically-based noise reduction algorithms are outlined which are required in the sequel of this thesis. After introducing the applied analysis-synthesis structure for the transformation into the frequency domain, several methods for noise *Power Spectral Density* (PSD) estimation and SNR estimation are presented. Finally, a short literature survey of state-of-the-art weighting rules is given relying on different optimization criteria and different statistical assumptions concerning the *Probability Density Functions* (PDFs) of the speech as well as the noise signal.

In *Chapter 3*, a novel model-based speech enhancement system is presented which exploits *temporal correlation* of speech and noise signals. The proposed scheme is based on a Kalman filter structure that is applied in the frequency domain to the complex-valued input DFT coefficients by using a two step approach. In the first step, information from previous, enhanced speech and noise DFT coefficients is exploited to perform estimates of the current DFT coefficients. In general, the predictions in this first step are erroneous resulting in non-zero prediction errors. Thus, a second step is applied in which the first estimates of speech and noise are updated by using adequate statistical weighting rules, amongst others SNR-dependent MMSE estimators which are explicitly adapted to (measured) statistics of the speech prediction error signal. In addition to instrumental measurements, the results of an informal listening test are presented in order to investigate the potential of the proposed model-based noise reduction technique.

In *Chapter 4*, *spectral dependencies* of speech signals are analyzed and it is shown how to benefit from these dependencies for the purpose of wideband speech enhancement (50 Hz – 7 kHz). As a novel feature, techniques known from *Artificial Bandwidth Extension* (ABWE) are used in this chapter to improve the spectral estimation process in the high band (4 kHz – 7 kHz). Therefore, the spectral dependencies between low band (50 Hz – 4 kHz) and high band are exploited. While a conventional noise suppression technique is used in the low band, a joint approach is applied for the speech enhancement in the high band. Based on a trained *Hidden Markov Model* (HMM), features from the processed (enhanced) low band signal are extracted and used to estimate subband energies of the high band speech signal. The resulting weighting gains determined from these energy estimates are adaptively combined with conventional gains for the high band. The performance of the proposed noise reduction technique is evaluated by means of instrumental measurements as well as auditory judgments. In addition to the wideband speech enhancement system, this chapter provides an

information theoretic view on ABWE under noisy conditions. A performance bound is formulated and the influence of noise reduction prior to ABWE is investigated by real entropy measurements.

Chapter 5 covers additional methods for quality improvements. When a noise suppression system is applied in a real-world scenario, the user often has to cope with a tradeoff between noise attenuation, speech distortions and the occurrence of annoying musical noise artifacts in the processed output signal. This chapter provides helpful techniques to be used in order to achieve a good compromise between these three aspects. In the first part, two different *musical noise countermeasures* are presented which can be applied to an arbitrary noise reduction system in a postprocessing stage. Method A performs adaptive spectral smoothing of the weighting gains relying on a low input SNR detector. In contrast, Method B is based on noise suppression with adaptive frequency resolution where the resolution is lower during speech pauses in order to reduce the tonality of the residual noise. The second part of this chapter deals with speech enhancement in *non-stationary noise environments* which is still a very challenging problem. A noise reduction system is presented which is capable of tracking and suppressing rapidly time-varying harmonic noise as well as stationary noise. In a first stage, the harmonic noise power is estimated and attenuated using a modified Minimum Statistics approach which performs frequency warping according to the harmonic's fundamental frequency. A conventional noise estimation technique is applied in a second stage in order to reduce the remaining random components of the noise spectrum. Instrumental measurements and informal listening tests are carried out in order to evaluate the performance of the proposed methods.

Chapter 6 summarizes the developed speech enhancement techniques and gives some concluding remarks on the main results of this thesis.

In the *Appendix*, particular equations are derived and certain assumptions made within the thesis are further analyzed. Moreover, the computational complexity as well as the memory requirements of the proposed Kalman filter approach are evaluated and the applied instrumental measurements are presented in detail.

Parts of the results of this thesis are presented in the following references published by the author: [EV11, HEV11, ERHV10b, ERHV10a, EHGV10, JSEV10, HEGV10, EV09, KSE+09, EV08c, EV08b, EV08a, Esc06]. These references are highlighted by underlines in the following, i.e., [_____].

2

Statistical Noise Suppression Techniques

Noise reduction is covered in literature for more than 30 years. Depending on the environment, the application, the number of available microphones, the source signals and the type of noise, the solutions look very different. In order to perform single-channel speech enhancement in communication systems, a widely used method is the so-called *spectral decomposition* of the noisy input signal using *statistical noise suppression techniques*. Thereby, at first an appropriate block-based analysis system is applied in order to segment the microphone signal into overlapping frames. Utilizing the *Short-Time Fourier Transform* (STFT), the resulting segments are transformed into the frequency domain where the respective spectral STFT coefficients are modified by a gain function. This gain function aims to minimize a specific distortion measure between clean and estimated speech signal usually in terms of magnitude and phase modifications. In most cases, the corresponding weighting rules rely on a mathematical cost function like *Minimum Mean Square Error* (MMSE), *Maximum A Posteriori* (MAP) or *Maximum Likelihood* (ML) as well as statistical characteristics of the speech and the noise signal. The resulting weighting gains are determined for each frame and frequency bin and their absolute values vary between 0 and 1. They should be high, i.e., near 1, in good *Signal-to-Noise-Ratio* (SNR) conditions in order to minimize speech distortions and low, i.e., near 0, if the current noisy STFT coefficients contain no speech or only weak speech components. In addition to appropriate statistical models and adequate distortion measures, a statistical noise reduction method usually also requires knowledge about the noise *Power Spectral Density* (PSD) and the frequency domain input SNR. As both entities are in general not known a priori, they have to be estimated and updated during runtime. Eventually, after weighting the input STFT coefficients according to the specific statistical estimator, the processed spectrum is transformed back into the time domain.

This chapter provides the basic principles of noise reduction in the frequency domain which are required in the sequel of this thesis. It gives a general overview about statistical noise reduction techniques including different statistical models for speech and noise, different cost functions and state-of-the-art approaches for the estimation of the required noise PSD and input SNR. In order to get a more detailed insight

into statistical noise suppression techniques, the author refers to the literature, e.g., [VM06, Ben07, BCHC09, HS06, Lim83].

The remainder of this chapter is organized according to the signal flow of a typical statistical noise reduction system depicted in Fig. 2.1. After introducing the underlying signal model, the analysis-synthesis structure which is applied in this thesis is outlined including the transformation into the frequency domain. Thereafter, different techniques for the estimation of the noise power as well as the input SNR are presented. Eventually, several gain calculation rules using different cost functions and different *Probability Density Functions* (PDFs) to model the statistics of speech and noise are comprised in detail.

2.1 Problem Formulation

A clean speech signal $s(k)$ is assumed to be degraded by an additive noise signal $n(k)$. The resulting noisy signal $y(k)$ picked up by the microphone is given by:

$$y(k) = s(k) + n(k), \tag{2.1}$$

where k is the discrete time index. Speech and noise signals are supposed to be uncorrelated and can be described by zero-mean random processes. The aim of any noise suppression system is then to estimate the clean speech signal having access only to the noisy microphone signal $y(k)$. It is desirable to attenuate the noise signal as much as possible while keeping the distortions of the speech signal as low as possible at the same time. The resulting estimate at the output is denoted by $\hat{s}(k)$.

In the following, the functionality of each block in Fig. 2.1 between noisy input and processed output signal is described in detail incorporating state-of-the-art examples for the respective topics.

2.2 Analysis and Synthesis

In the derivation of most statistical noise reduction algorithms, speech and noise is often assumed to be stationary. In this case, the resulting filter coefficients would be fixed during runtime and the noise suppression could easily be realized using simple *Finite Impulse Response* (FIR) or *Infinite Impulse Response* (IIR) filters. However, noise and in particular speech can be highly non-stationary as the production of human speech follows a time-varying process. Especially plosive sounds arise from dynamically transient changes of the vocal tract leading to the fact that some properties of speech signals as, e.g., energy or correlation, can be assumed to be stationary or at least quasi-stationary only for short time segments of $10 - 100\,\text{ms}$ [RS78].

To account for the temporal changes of speech and noise, the processing of the noisy input signal is performed framewise and the filter coefficients are updated continuously.

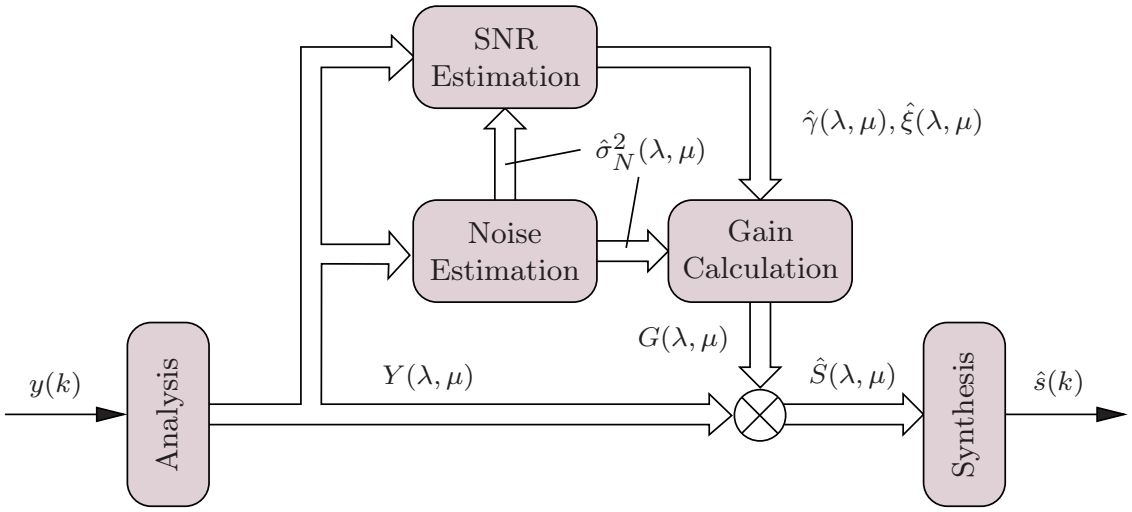

Figure 2.1: System block diagram of a conventional noise suppression system working in the frequency domain.

For this purpose, an analysis-synthesis system with perfect reconstruction is applied as depicted in Fig. 2.2.

At first, the noisy input signal $y(k)$ is segmented into overlapping frames of length L_{F} with a frame shift size L_{FS}, cf. Fig. 2.2(a). The choice of the frame length determines the temporal resolution of the analysis-synthesis system. As the temporal resolution directly affects the spectral resolution, each spectral noise suppression system has to cope with a tradeoff between both. On the one hand, a high spectral resolution is desirable in order to preserve speech components as, e.g., pitch harmonics. However, on the other hand, for the enhancement of highly dynamic speech parts (plosives, onsets and offsets), it is better to work with a high temporal resolution. Typical sizes for the frame length L_{F} are within the range from 5 ms to 40 ms [VM06, PLW10]. The frame overlap commonly varies between 50% and 75% [Ben07].

In order to counteract the *spectral leakage effect*, the samples in each frame are multiplied by a tapered window. Frequently used window functions are, e.g., the Hann window, the Hamming window or the Blackman window [OS98]. After segmentation, windowing and if necessary zero padding, the noisy short-time segments are transformed into the frequency domain using a short-time *Discrete Fourier Transform* (DFT) of length M_{F}. The spectrum of the noisy input signal is given by:

$$Y(\lambda,\mu) = R(\lambda,\mu)\cdot \mathrm{e}^{j\vartheta(\lambda,\mu)} \tag{2.2}$$

$$= S(\lambda,\mu) + N(\lambda,\mu) \tag{2.3}$$

$$= A(\lambda,\mu)\cdot \mathrm{e}^{j\phi(\lambda,\mu)} + N(\lambda,\mu), \tag{2.4}$$

where $R(\lambda,\mu)$ and $\vartheta(\lambda,\mu)$ are the noisy magnitude and the corresponding phase at frame λ and frequency bin μ. $S(\lambda,\mu)$ and $N(\lambda,\mu)$ represent the complex-valued spectral DFT coefficients of speech and noise and $A(\lambda,\mu)$ and $\phi(\lambda,\mu)$ denote the spectral amplitude and phase of the clean speech signal.

Transforming the noisy input signal into the spectral domain is a widely accepted technique for speech enhancement as it strongly corresponds to the processing taking

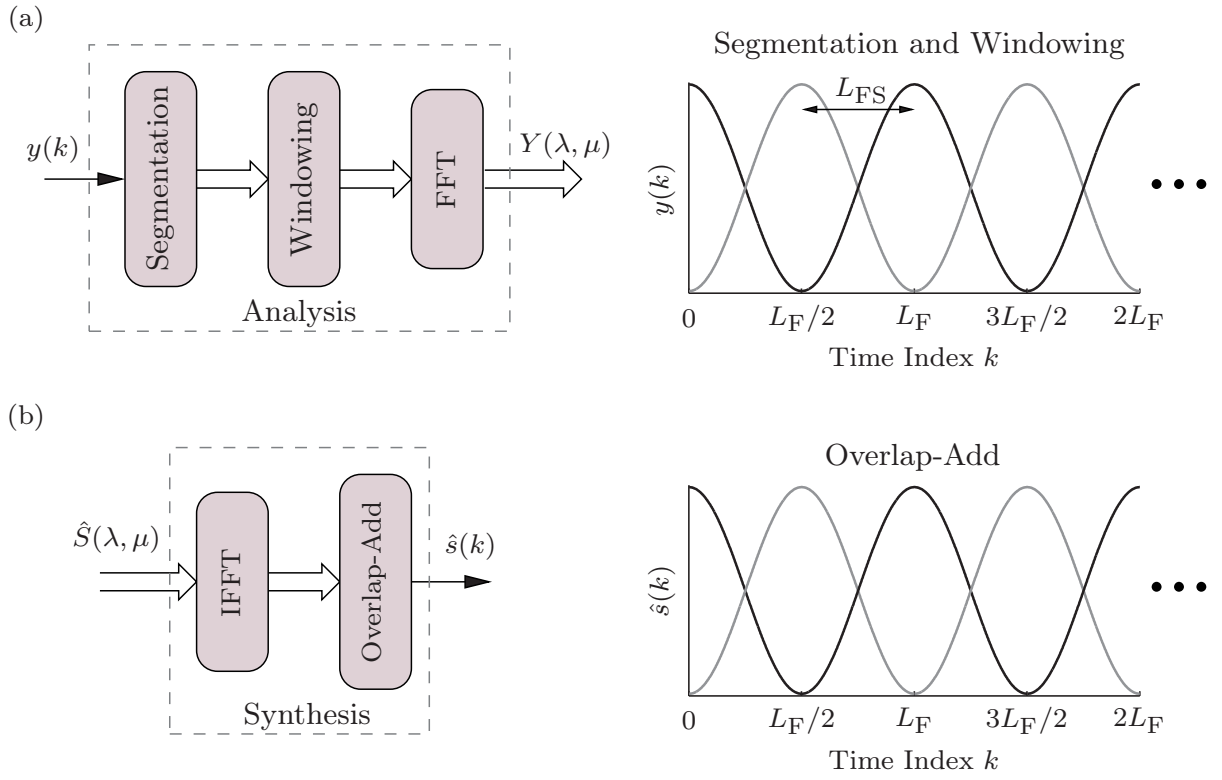

Figure 2.2: Analysis-synthesis system: insights into (a) analysis block including segmentation, windowing and FFT, (b) synthesis block including IFFT and overlap-add.

place in the human auditory system [ZF90]. In addition, the weighting gains can be designed independently for each frequency bin and an efficient realization of the Fourier transform is possible using the *Fast Fourier Transform* (FFT).

Using the noisy input spectrum $Y(\lambda, \mu)$, the short-time noise PSD as well as the frequency domain SNR are estimated and the weighting gains are determined, see Fig. 2.1. The actual spectral weighting is performed by multiplying the noisy spectrum $Y(\lambda, \mu)$ by weighting gains $G(\lambda, \mu)$ resulting in estimates $\hat{S}(\lambda, \mu)$ of the clean speech DFT coefficients $S(\lambda, \mu)$ according to:

$$\begin{aligned} \hat{S}(\lambda, \mu) &= G(\lambda, \mu) \cdot Y(\lambda, \mu) \\ &= G(\lambda, \mu) \cdot R(\lambda, \mu) \cdot \mathrm{e}^{j\vartheta(\lambda, \mu)}. \end{aligned} \tag{2.5}$$

The weighting gains can be complex-valued. However, as the human ear is rather insensitive w.r.t. phase distortions [WL82], most estimators only modify the spectral magnitudes and use the phases $\vartheta(\lambda, \mu)$ of the noisy input signal for reconstruction. As mentioned before, the absolute values of $G(\lambda, \mu)$ lie in the range between 0 and 1. In order to obtain the enhanced signal also in the time domain, all operations applied in the analysis are reversed in a subsequent synthesis. As shown in Fig. 2.2(b), an *Inverse Fast Fourier Transform* (IFFT) and overlap-add are utilized for this purpose [GL84].

2.3 Noise PSD Estimation

A crucial component of any practical speech enhancement system is the estimation of the short-term noise PSD $\sigma_N^2(\lambda, \mu)$. Especially in non-stationary noise environments an accurate tracking of $\sigma_N^2(\lambda, \mu)$ becomes very challenging and requires adaptive methods which are able to estimate the noise power also during speech activity. If the noise PSD is overestimated, the suppression of the input signal might be too strong leading to speech distortions and a loss in intelligibility. In contrast, high noise levels might remain as consequence of a noise underestimation.

Various single-channel noise PSD estimation algorithms can be found in literature. A comparison of some state-of-the-art estimators is presented in [TTM+11]. The most prominent techniques are briefly outlined in the following.

2.3.1 Voice Activity Detection

One of the first methods known in literature for noise PSD estimation is based on *Voice Activity Detection* (VAD), e.g., [MM80] and [vC89]. The estimate $\hat{\sigma}_N^2(\lambda, \mu)$ is obtained by smoothing the noisy periodogram over time in noise-only frames and keeping this estimate fixed during speech activity. In order to decide whether the current noisy DFT coefficients contain speech or not, a hypothesis-based framework with final threshold decision is applied usually. However, especially in low SNR conditions, it remains very difficult to achieve an accurate VAD [VM06]. Moreover, noise is often non-stationary. Thus, the accuracy of this estimation technique is limited and performs well only in situations with a moderate noise level. A comparison of *International Telecommunication Union -Telecommunication Standardization Sector* (ITU-T) and *European Telecommunications Standards Institute* (ETSI) voice activity detection approaches can be found, e.g., in [BCR01].

2.3.2 Minimum Statistics

In contrast to a VAD driven estimator, the *Minimum Statistics* algorithm is able to update the estimated noise PSD also during speech activity [Mar01]. The approach relies on two assumptions:

- Speech and noise are statistically independent and
- the power of the noisy signal often decays to the power level of the noise signal (e.g., in speech pauses).

Based on these assumptions it is possible to track the minimum of the short-term noisy PSD. As this minimum is always smaller or equal to the mean noise power, a bias correction is necessary.

In a first step, the noisy periodogram $|Y(\lambda,\mu)|^2$ is recursively smoothed over time where $|\cdot|$ represents the magnitude operator. The smoothed signal power $\hat{\sigma}_Y^2(\lambda,\mu)$ is given by:

$$\hat{\sigma}_Y^2(\lambda,\mu) = \alpha_{\mathrm{MS}}(\lambda,\mu)\cdot\hat{\sigma}_Y^2(\lambda-1,\mu) + (1-\alpha_{\mathrm{MS}}(\lambda,\mu))\cdot|Y(\lambda,\mu)|^2, \tag{2.6}$$

with $\alpha_{\mathrm{MS}}(\lambda,\mu)\in[0,1]$ denoting a frame and frequency-dependent smoothing factor. Afterwards, the minimum $\hat{\sigma}_{Y,\min}^2(\lambda,\mu)$ of the most recent D_{MS} values is tracked for each frequency bin separately by a sliding time window according to:

$$\hat{\sigma}_{Y,\min}^2(\lambda,\mu) = \min_{\tilde{\lambda}\in[\lambda-D_{\mathrm{MS}}+1,\lambda]} \hat{\sigma}_Y^2(\tilde{\lambda},\mu). \tag{2.7}$$

The duration of the time window for the minimum search should be equal to approximately 1.5 seconds [Mar01]. Eventually, the minimum value is multiplied by a bias correction factor $B(\lambda,\mu)$, which is mainly dependent on the variance of the noisy input periodogram. The final noise PSD estimate is given by:

$$\hat{\sigma}_N^2(\lambda,\mu) = B(\lambda,\mu)\cdot\hat{\sigma}_{Y,\min}^2(\lambda,\mu). \tag{2.8}$$

Minimum Statistics shows good estimation results in stationary and slowly changing noise conditions. However, due to the large window length D_{MS}, it is not able to track a sudden rise in noise energy leading to an underestimation of the noise power in this case. The concept of noise estimators based on Minimum Statistics has been further studied in order to enhance the noise tracking capabilities, e.g., in [PC08, TJ09].

2.3.3 Minima Controlled Recursive Averaging

The *Minima Controlled Recursive Averaging* approach performs noise estimation by using a weighted recursion which averages previous noise PSD estimates over time [CB02]. The final estimate is given by:

$$\hat{\sigma}_N^2(\lambda+1,\mu) = \tilde{\alpha}_{\mathrm{MC}}(\lambda,\mu)\cdot\hat{\sigma}_N^2(\lambda,\mu) + (1-\tilde{\alpha}_{\mathrm{MC}}(\lambda,\mu))\cdot|Y(\lambda,\mu)|^2. \tag{2.9}$$

The frame and frequency-dependent weighting factor $\tilde{\alpha}_{\mathrm{MC}}(\lambda,\mu)$ is adjusted by the estimated speech presence probability $\hat{p}'(\lambda,\mu)$ according to:

$$\tilde{\alpha}_{\mathrm{MC}}(\lambda,\mu) = \alpha_{\mathrm{MC}} + (1-\alpha_{\mathrm{MC}})\cdot\hat{p}'(\lambda,\mu), \tag{2.10}$$

where $\alpha_{\mathrm{MC}}\in[0,1]$ denotes a further smoothing parameter. The probability $\hat{p}'(\lambda,\mu)$ is again determined recursively as follows:

$$\hat{p}'(\lambda,\mu) = \alpha_{\mathrm{p}}\cdot\hat{p}'(\lambda-1,\mu) + (1-\alpha_{\mathrm{p}})\cdot I\mathrm{p}(\lambda,\mu), \tag{2.11}$$

with $0\leq\alpha_{\mathrm{p}}\leq 1$. The function $I_{\mathrm{p}}(\lambda,\mu)$ states an indicator for speech presence and is set to 1 if the ratio $P_{\mathrm{r}}(\lambda,\mu)$ between current noisy energy $\hat{\sigma}_Y^{2'}(\lambda,\mu)$ and the minimum

of the local energy $\hat{\sigma}^{2'}_{Y,\min}(\lambda,\mu)$ within a specific time window exceeds a predefined threshold δ_{MC}. Otherwise, if $P_{\mathrm{r}}(\lambda,\mu)$ is lower than δ_{MC}, the function is set to 0, i.e.:

$$I_{\mathrm{p}}(\lambda,\mu) = \begin{cases} 1, & \text{if} \quad P_{\mathrm{r}}(\lambda,\mu) \geq \delta_{\mathrm{MC}} \\ 0, & \text{if} \quad P_{\mathrm{r}}(\lambda,\mu) < \delta_{\mathrm{MC}} \end{cases}, \tag{2.12}$$

with $P_{\mathrm{r}}(\lambda,\mu) = \frac{\hat{\sigma}^{2'}_{Y}(\lambda,\mu)}{\hat{\sigma}^{2'}_{Y,\min}(\lambda,\mu)}$. The energy $\hat{\sigma}^{2'}_{Y}(\lambda,\mu)$ is obtained by smoothing the squared magnitudes of the noisy DFT coefficients over time and frequency. The minimum of the local energy $\hat{\sigma}^{2'}_{Y,\min}(\lambda,\mu)$ is determined using a simplified version of the Minimum Statistics algorithm, see Sec. 2.3.2.

In obtaining the current noise PSD estimate, the minima controlled recursive averaging algorithm suffers from the same problems as Minimum Statistics. In order to determine the minimum $\hat{\sigma}^{2'}_{Y,\min}(\lambda,\mu)$, a large time window is used making it difficult to track non-stationarities of noise signals. An improved algorithm is proposed in [Coh03]. Here, two iterations of smoothing and minimum tracking are applied and a bias compensation factor is introduced resulting in the possibility to use smaller lengths for the search window. In [KC09], the approach is extended by a second order conditional MAP criterion.

2.3.4 MMSE Based Noise PSD Tracking

In this recently published algorithm [HHJ10], the noise PSD is derived from an MMSE estimate of the squared noise magnitude resulting in the following conditional expectation:

$$\hat{\sigma}^2_N(\lambda,\mu) = \mathbb{E}\{|N(\lambda,\mu)|^2 \,\big|\, |Y(\lambda,\mu)|\}, \tag{2.13}$$

where $\mathbb{E}\{\cdot\}$ represents the expectation operator.

In the derivation, it is assumed that speech and noise DFT coefficients exhibit a complex Gaussian PDF. The solution according to Eq. 2.13 requires an estimate of the input SNR which is obtained by using the current noisy periodogram and the noise PSD estimates from previous frames. Computing the expectation in Eq. 2.13 leads to an unbiased estimator for $\sigma^2_N(\lambda,\mu)$ in theory. In practice however, the estimate of the input SNR might introduce a bias which can be reduced by performing an additional bias compensation. A final smoothing operation across time is applied at the end in order to reduce the variance of the noise PSD estimates.

According to [TTM+11], this MMSE based noise PSD tracking algorithm shows a good noise PSD tracking performance also in challenging noise environments. Moreover, the computational complexity can be reduced by a factor of 8 compared to the Minimum Statistics approach [HHJ10].

2.4 Signal-to-Noise-Ratio Estimation

In addition to the estimation of the short-term noise PSD, most statistical noise suppression techniques require estimates of the *a posteriori SNR* $\gamma(\lambda, \mu)$ and *a priori SNR* $\xi(\lambda, \mu)$. The a posteriori SNR is defined as the ratio between the noisy periodogram and the short-term noise PSD according to:

$$\gamma(\lambda, \mu) = \frac{|Y(\lambda, \mu)|^2}{\sigma_N^2(\lambda, \mu)}. \tag{2.14}$$

If an estimate of $\sigma_N^2(\lambda, \mu)$ is available, the a posteriori SNR can easily be measured. Much more difficult to determine is the a priori SNR $\xi(\lambda, \mu)$ as it requires an estimate of the unknown short-term speech PSD $\sigma_S^2(\lambda, \mu)$ as well. Assuming speech and noise to be uncorrelated, it can be expressed dependent on the a posteriori SNR as follows [VM06]:

$$\xi(\lambda, \mu) = \frac{\sigma_S^2(\lambda, \mu)}{\sigma_N^2(\lambda, \mu)} = \mathbb{E}\{\gamma(\lambda, \mu) - 1\}. \tag{2.15}$$

For the estimation of $\xi(\lambda, \mu)$ many approaches can be found in literature. In the following, two well-known types of a priori SNR estimators are presented: the *decision-directed* approach and a method based on *Generalized Autoregressive Conditional Heteroscedasticity* (GARCH) models.

2.4.1 Decision-Directed Approach

The decision-directed approach [EM84] is widely accepted in literature and contributes to an improved subjective quality of the enhanced speech signal. For estimating the a priori SNR, this approach linearly combines estimates from previous frames with an instantaneous SNR realization relying on the a posteriori SNR according to:

$$\hat{\xi}(\lambda, \mu) = \alpha_{\mathrm{DD}} \cdot \frac{|\hat{S}(\lambda - 1, \mu)|^2}{\hat{\sigma}_N^2(\lambda - 1, \mu)} + (1 - \alpha_{\mathrm{DD}}) \cdot \max\big(\hat{\gamma}(\lambda, \mu) - 1, 0\big), \tag{2.16}$$

where $\max(\cdot, \cdot)$ returns the maximum of its two arguments. The smoothing factor α_{DD} adjusts the tradeoff between noise reduction and speech distortions and typically lies in the range $0.9 \leq \alpha_{\mathrm{DD}} \leq 0.99$. In this work, α_{DD} is set to 0.98 according to [EM84].

2.4.2 Estimation Based on GARCH Models

GARCH models are known from financial applications where they are used, e.g., to model the time-varying volatility of stocks. As speech signals in the frequency domain also show 'variability clustering' and a 'heavy tail behavior' which means that large amplitudes tend to follow large amplitudes and low amplitudes tend to follow low

amplitudes while the phase can hardly be predicted, the use of GARCH models for speech variance estimation is proposed in [Coh04].

In [Coh05a], a two step speech PSD estimator is presented. In the first step, the variance estimate of frame λ is propagated in time from the information that is available at frame $\lambda - 1$. Therefore, the following GARCH model is used:

$$\hat{\sigma}^2_{S,1}(\lambda,\mu) = \sigma^2_{\min} + \rho_{\mathrm{G}} \cdot \hat{\sigma}^2_{S,2}(\lambda-1,\mu) + \delta_{\mathrm{G}} \cdot (\hat{\sigma}^2_{S,1}(\lambda-1,\mu) - \sigma^2_{\min}), \tag{2.17}$$

where $\sigma^2_{\min}$ denotes a lower bound, ρ_{G} a moving average parameter and δ_{G} an autoregressive parameter with respect to the constraints $\sigma^2_{\min} > 0$, $\rho_{\mathrm{G}} \geq 0$, $\delta_{\mathrm{G}} \geq 0$ and $\rho_{\mathrm{G}} + \delta_{\mathrm{G}} < 1$.

In the second step, the first estimate $\hat{\sigma}^2_{S,1}(\lambda,\mu)$ is updated by using the additional information $Y(\lambda,\mu)$ of the current frame. Therefore, the following conditional expectation is derived:

$$\hat{\sigma}^2_{S,2}(\lambda,\mu) = \mathbb{E}\{|S(\lambda,\mu)|^2 \,|\, \mathcal{H}_1(\lambda,\mu), \hat{\sigma}^2_{S,1}(\lambda,\mu), Y(\lambda,\mu)\}, \tag{2.18}$$

where $\mathcal{H}_1(\lambda,\mu)$ denotes the hypothesis that speech is present in the current spectral coefficient $Y(\lambda,\mu)$. Several solutions for Eq. 2.18 are presented in [Coh05a] based on different statistical models which are assumed for speech and noise. Finally, the result of Eq. 2.18 is utilized to determine the a priori SNR according to Eq. 2.15 using an adequate noise PSD estimate. The model parameters $\sigma^2_{\min}$, ρ_{G} and δ_{G} have to be determined in advance from a training database.

2.5 Statistical Weighting Rules

As depicted in Fig. 2.1, the actual spectral weighting is performed by multiplying the noisy spectrum $Y(\lambda,\mu)$ by weighting gains $G(\lambda,\mu)$ resulting in estimates of the clean speech DFT coefficients according to Eq. 2.5. The calculation of these weighting gains is dependent on the respective noise reduction algorithm and is usually a function of the short-term noise PSD estimate $\hat{\sigma}^2_N(\lambda,\mu)$ and the SNR estimates $\hat{\gamma}(\lambda,\mu)$ and $\hat{\xi}(\lambda,\mu)$ as addressed before. In essence, the determination of the spectral weighting gains relies on mathematical criteria and specific statistical models which are assumed either for the complex-valued DFT coefficients of speech and noise or for their real-valued magnitudes and phases. Motivated by the central limit theorem [ABBN04], it is often assumed that real and imaginary parts of speech and noise spectra follow a Gaussian distribution. However, especially for speech signals, this condition is not exactly met in reality as the respective frame and DFT sizes applied in the analysis system are too small. A better approximation is obtained by using a super-Gaussian distribution as, e.g., Laplacian or Gamma for the speech DFT coefficients.

In the following, a selection of some well-known statistical noise suppression techniques relying on both Gaussian and super-Gaussian models is presented. The derivations and the resulting spectral weighting gains are given for each algorithm. More detailed information can be found in literature.

2.5.1 Wiener Filter

The Wiener filter is derived from the optimal filter theory, e.g., [Vas96, LO79]. It is a linear estimator that minimizes the mean square error between the clean speech DFT coefficients $S(\lambda,\mu)$ and the enhanced DFT coefficients $\hat{S}(\lambda,\mu)$. With $\hat{S}(\lambda,\mu) = G(\lambda,\mu)\cdot Y(\lambda,\mu)$ according to Eq. 2.5, it follows:

$$\mathbb{E}\{|S(\lambda,\mu)-\hat{S}(\lambda,\mu)|^2\} = \mathbb{E}\{|S(\lambda,\mu)-G(\lambda,\mu)\cdot Y(\lambda,\mu)|^2\} \quad \rightarrow \quad \min. \tag{2.19}$$

Assuming mutual independence of real and imaginary parts, the partial derivations of Eq. 2.19 with respect to the real and imaginary parts of $G(\lambda,\mu)$ yield [Vas96]:

$$\frac{\partial\mathbb{E}\{|S(\lambda,\mu)-\hat{S}(\lambda,\mu)|^2\}}{\partial\mathrm{Re}\{G(\lambda,\mu)\}} = 0 \quad \rightarrow \quad \mathrm{Re}\{G(\lambda,\mu)\} = \frac{\mathbb{E}\{|S(\lambda,\mu)|^2\}}{\mathbb{E}\{|S(\lambda,\mu)|^2\}+\mathbb{E}\{|N(\lambda,\mu)|^2\}}, \tag{2.20}$$

as well as

$$\frac{\partial\mathbb{E}\{|S(\lambda,\mu)-\hat{S}(\lambda,\mu)|^2\}}{\partial\mathrm{Im}\{G(\lambda,\mu)\}} = 0 \quad \rightarrow \quad \mathrm{Im}\{G(\lambda,\mu)\} = 0, \tag{2.21}$$

with $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denoting real and imaginary parts. Hence, the spectral enhanced DFT coefficients $\hat{S}(\lambda,\mu)$ can be stated as:

$$\hat{S}(\lambda,\mu) = \frac{\mathbb{E}\{|S(\lambda,\mu)|^2\}}{\mathbb{E}\{|S(\lambda,\mu)|^2\}+\mathbb{E}\{|N(\lambda,\mu)|^2\}}\cdot Y(\lambda,\mu) = \underbrace{\frac{\xi(\lambda,\mu)}{\xi(\lambda,\mu)+1}}_{G(\lambda,\mu)}\cdot Y(\lambda,\mu), \tag{2.22}$$

resulting in a weighting gain $G(\lambda,\mu)$ which is only dependent on the a priori SNR $\xi(\lambda,\mu)$:

$$G(\lambda,\mu) = \frac{\xi(\lambda,\mu)}{\xi(\lambda,\mu)+1}. \tag{2.23}$$

Using a Gaussian signal model for the real and imaginary parts of speech and noise, the Wiener filter solution equals the conditional expectation $\mathbb{E}\{S(\lambda,\mu)\,|\,Y(\lambda,\mu)\}$, e.g., [VM06].

2.5.2 MMSE Short-Time Spectral Amplitude Estimator

In speech and audio applications, a certain amount of phase distortions is mostly tolerable due to the mentioned insensitivity of the human ear with respect to phase errors [WL82, Var85, ZF99]. Moreover, the estimation of the magnitudes of the short-time Fourier speech coefficients is much easier to achieve than the estimation of the corresponding phases. The MMSE *Short-Time Spectral Amplitude* (STSA) estimator [EM84] belongs to this important class of amplitude estimators, as it only estimates the spectral magnitudes and uses the noisy input phase $\vartheta(\lambda,\mu)$ for reconstruction.

The STSA estimator minimizes the quadratic error of the spectral speech amplitudes according to:

$$\mathbb{E}\{(A(\lambda,\mu)-\hat{A}(\lambda,\mu))^2\} \quad \rightarrow \quad \min. \tag{2.24}$$

Under the assumption of a Gaussian speech and noise model and statistical independence of the respective real and imaginary parts of $S(\lambda,\mu)$, $N(\lambda,\mu)$ and $Y(\lambda,\mu)$, the final weighting rule is given by [EM84]:

$$\begin{aligned}\hat{S}(\lambda,\mu) &= \mathbb{E}\{A(\lambda,\mu)|Y(\lambda,\mu)\}\cdot\exp(\mathrm{j}\cdot\vartheta(\lambda,\mu)) \\ &= \underbrace{\frac{\sqrt{\nu(\lambda,\mu)}}{\gamma(\lambda,\mu)}\Gamma(1.5)F_1(-0.5,1,-\nu)}_{G(\lambda,\mu)}\cdot Y(\lambda,\mu),\end{aligned} \tag{2.25}$$

where $\nu(\lambda,\mu)=\frac{\xi(\lambda,\mu)}{1+\xi(\lambda,\mu)}\cdot\gamma(\lambda,\mu)$, $\Gamma(\cdot)$ denotes the Gamma function and $F_1(\cdot,\cdot,\cdot)$ the hypergeometric function [GRJZ00]. In addition, $\exp(\cdot)$ represents the exponential function and j the imaginary unit.

2.5.3 MMSE Log Spectral Amplitude Estimator

Similar to the MMSE STSA estimator the MMSE *Log Spectral Amplitude* (LSA) weighting rule estimates only the magnitudes of the short-time Fourier coefficients of the clean speech signal [EM85]. However, in order to adapt to the logarithmic response of the human ear to sound intensity changes [ZF90], this estimator minimizes the mean square error of the logarithmically weighted amplitudes as follows:

$$\mathbb{E}\{(\log_{\mathrm{e}}(A(\lambda,\mu))-\log_{\mathrm{e}}(\hat{A}(\lambda,\mu)))^2\} \quad \rightarrow \quad \min., \tag{2.26}$$

with e the Euler number. Using the same assumptions as for the MMSE STSA estimator that speech and noise DFT coefficients are complex-Gaussian distributed and that real and imaginary parts of $S(\lambda,\mu)$, $N(\lambda,\mu)$ and $Y(\lambda,\mu)$ are statistically independent, it follows:

$$\begin{aligned}\hat{S}(\lambda,\mu) &= \exp\left(\mathbb{E}\{\log_{\mathrm{e}}(A(\lambda,\mu))|Y(\lambda,\mu)\}\right)\cdot\exp(\mathrm{j}\cdot\vartheta(\lambda,\mu)) \\ &= \underbrace{\frac{\xi(\lambda,\mu)}{1+\xi(\lambda,\mu)}\cdot\exp\left(\frac{1}{2}\int\limits_{\nu(\lambda,\mu)}^{\infty}\frac{\exp(-t)}{t}\mathrm{d}t\right)}_{G(\lambda,\mu)}\cdot Y(\lambda,\mu),\end{aligned} \tag{2.27}$$

where again $\nu(\lambda,\mu)=\frac{\xi(\lambda,\mu)}{1+\xi(\lambda,\mu)}\cdot\gamma(\lambda,\mu)$. In literature, the MMSE LSA estimator is often used as reference and is probably the most cited weighting rule for noise reduction.

2.5.4 Laplacian MMSE Estimator

In [MB03], a super-Gaussian MMSE estimator is proposed which uses a Laplacian speech and a Gaussian noise model. The algorithm minimizes the quadratic error of the clean and estimated speech DFT coefficients as follows:

$$\mathbb{E}\{(S(\lambda,\mu)-\hat{S}(\lambda,\mu))^2\} \quad \rightarrow \quad \min. \tag{2.28}$$

Under the assumption that real and imaginary parts of $S(\lambda,\mu)$, $N(\lambda,\mu)$ and $Y(\lambda,\mu)$ are statistically independent, the estimator can be divided into a separate estimator for both parts according to:

$$\begin{aligned}\hat{S}(\lambda,\mu) &= \mathbb{E}\{S(\lambda,\mu)|Y(\lambda,\mu)\} \\ &= \mathbb{E}\{\mathrm{Re}\{S(\lambda,\mu)\}|\mathrm{Re}\{Y(\lambda,\mu)\}\} + \mathrm{j}\cdot\mathbb{E}\{\mathrm{Im}\{S(\lambda,\mu)\}|\mathrm{Im}\{Y(\lambda,\mu)\}\}.\end{aligned} \tag{2.29}$$

Using the abbreviated forms $Y_{\mathrm{Re}} = \mathrm{Re}\{Y(\lambda,\mu)\}$, $Y_{\mathrm{Im}} = \mathrm{Im}\{Y(\lambda,\mu)\}$, $S_{\mathrm{Re}} = \mathrm{Re}\{S(\lambda,\mu)\}$ and $S_{\mathrm{Im}} = \mathrm{Im}\{S(\lambda,\mu)\}$, real as well as imaginary parts of $\hat{S}(\lambda,\mu)$ result in:

$$\mathbb{E}\{S_{\mathrm{Re}}|Y_{\mathrm{Re}}\} = ... \\ \frac{\sigma_N(\lambda,\mu)\left(L_{\mathrm{Re}_+}\exp\left(L^2_{\mathrm{Re}_+}\right)\mathrm{erfc}(L_{\mathrm{Re}_+}) - L_{\mathrm{Re}_-}\exp\left(L^2_{\mathrm{Re}_-}\right)\mathrm{erfc}(L_{\mathrm{Re}_-})\right)}{\exp\left(L^2_{\mathrm{Re}_+}\right)\mathrm{erfc}(L_{\mathrm{Re}_+}) + \exp\left(L^2_{\mathrm{Re}_-}\right)\mathrm{erfc}(L_{\mathrm{Re}_-})}, \tag{2.30}$$

and

$$\mathbb{E}\{S_{\mathrm{Im}}|Y_{\mathrm{Im}}\} = ... \\ \frac{\sigma_N(\lambda,\mu)\left(L_{\mathrm{Im}_+}\exp\left(L^2_{\mathrm{Im}_+}\right)\mathrm{erfc}(L_{\mathrm{Im}_+}) - L_{\mathrm{Im}_-}\exp\left(L^2_{\mathrm{Im}_-}\right)\mathrm{erfc}(L_{\mathrm{Im}_-})\right)}{\exp\left(L^2_{\mathrm{Im}_+}\right)\mathrm{erfc}(L_{\mathrm{Im}_+}) + \exp\left(L^2_{\mathrm{Im}_-}\right)\mathrm{erfc}(L_{\mathrm{Im}_-})}, \tag{2.31}$$

where $L_{\mathrm{Re}_+} = \frac{1}{\sqrt{\xi(\lambda,\mu)}} + \frac{Y_{\mathrm{Re}}}{\sigma_N(\lambda,\mu)}$, $L_{\mathrm{Re}_-} = \frac{1}{\sqrt{\xi(\lambda,\mu)}} - \frac{Y_{\mathrm{Re}}}{\sigma_N(\lambda,\mu)}$, $L_{\mathrm{Im}_+} = \frac{1}{\sqrt{\xi(\lambda,\mu)}} + \frac{Y_{\mathrm{Im}}}{\sigma_N(\lambda,\mu)}$, $L_{\mathrm{Im}_-} = \frac{1}{\sqrt{\xi(\lambda,\mu)}} - \frac{Y_{\mathrm{Im}}}{\sigma_N(\lambda,\mu)}$ and $\mathrm{erfc}(\cdot)$ denotes the complementary error function [GRJZ00].

2.5.5 Super-Gaussian MAP Estimator

A more generalized super-Gaussian approach is proposed in [LV05]. Here, the following parametric function is used to approximate the PDF of the spectral speech amplitude $A(\lambda,\mu)$:

$$p(A(\lambda,\mu)) = \frac{{\delta_{\mathrm{M}}}^{\eta_{\mathrm{M}}+1}}{\Gamma(\eta_{\mathrm{M}}+1)}\frac{A(\lambda,\mu)^{\eta_{\mathrm{M}}}}{\sigma_S(\lambda,\mu)^{\eta_{\mathrm{M}}+1}}\exp\left(-\delta_{\mathrm{M}}\frac{A(\lambda,\mu)}{\sigma_S(\lambda,\mu)}\right), \tag{2.32}$$

where η_{M} and δ_{M} denote the model parameters. Thereby, η_{M} describes the PDF at small values and δ_{M} represents the slope of the decay at higher values. The model function allows the approximation, e.g., of a Laplacian PDF (for η_{M}=1 and δ_{M}=2.5) or of a Gamma PDF (for η_{M}=0.01 and δ_{M}=1.5).

In [LV05], the histogram of real speech amplitudes is measured based on a large speech database. After normalization, the measured PDF is approximated by the model PDF in Eq. 2.32. Therefore, the model parameters are adjusted according to the Kullback Leibler distance [KL51] which is an information theoretic measure for the similarity of two PDFs. The optimal approximation, i.e., the minimum Kullback Leibler distance, is given for the parameters η_{M}=0.126 and δ_{M}=1.74 [LV05] and lies between a Laplacian and a Gamma *Probability Density Function* (PDF).

The proposed MAP estimator jointly maximizes the a posteriori PDF of the amplitude $A(\lambda,\mu)$ and the phase $\phi(\lambda,\mu)$ of the clean speech signal given the noisy magnitude $R(\lambda,\mu)$ according to:

$$\hat{A}(\lambda,\mu) = \arg\max_{A(\lambda,\mu)} p\left(A(\lambda,\mu), \phi(\lambda,\mu) \,|\, R(\lambda,\mu)\right) \quad \text{and} \tag{2.33}$$

$$\hat{\phi}(\lambda,\mu) = \arg\max_{\phi(\lambda,\mu)} p\left(A(\lambda,\mu), \phi(\lambda,\mu) \,|\, R(\lambda,\mu)\right). \tag{2.34}$$

Using a Gaussian model for the noise signal and assuming that the phase distribution of $\phi(\lambda,\mu)$ is independent from the amplitude distribution given in Eq. 2.32, the resulting weighting rule yields:

$$G(\lambda,\mu) = u_{\mathrm{M}}(\lambda,\mu) + \sqrt{u_{\mathrm{M}}^2(\lambda,\mu) + \frac{\eta_{\mathrm{M}}}{2\gamma(\lambda,\mu)}}, \tag{2.35}$$

where $u_{\mathrm{M}}(\lambda,\mu) = \frac{1}{2} - \frac{\delta_{\mathrm{M}}}{4\sqrt{\gamma(\lambda,\mu)\xi(\lambda,\mu)}}$.

3

Speech Enhancement Incorporating Temporal Correlation

Most state-of-the-art noise suppression systems are based on the so-called *spectral subtraction* approach. Spectral subtraction was introduced in [Bol79] and originally denotes the subtraction of the estimated noise spectrum from the noisy input spectrum in order to estimate the speech signal. In the meantime, the term 'spectral subtraction' is also used in literature to perform noise suppression in the short-term *Discrete Fourier Transform* (DFT) domain by applying individual adaptive gains to the noisy frequency domain coefficients. There exists a huge amount of different weighting rules. Most of these rules aim to minimize a specific distortion measure of a mathematical cost function between original and estimated speech signal under certain assumptions about the statistics of speech and noise, cf. Chapter 2. As these methods only consider the probability distributions of the DFT coefficients, they can be classified as methods relying on memory-less *A Priori Knowledge* (AK) (or A Priori Knowledge of order zero, AK0). Temporal correlation of speech and noise is explicitly not taken into account except for smoothing purposes, e.g., in [EM84].

In literature, the so-called *source-filter* model is widely accepted and often used in order to model the speech production process, e.g., [RS78, Par86, Qua01, VM06]. In its most simple form, this model consists of two components: the *excitation signal generator* representing the physical effects caused by the lungs as well as the vocal cords and the time-varying *digital vocal tract filter* that approximates the influence of the human vocal tract. The spectral envelope of a speech signal and its parametric representation are often realized by using *Autoregressive* (AR) modeling in conjunction with *Linear Prediction* (LP) techniques [VM06, AS70, MGj76]. A speech signal exhibits correlation across time originating on the one hand from a pulsed excitation signal (in voiced speech segments) and on the other hand from the vocal tract filter. Nevertheless, speech signals can be considered to be correlated only for short-time periods between 10 and 100 ms as the coefficients of the vocal tract filter change quickly over time [RS78, PLW10, Hab05].

Almost every speech coding standard is explicitly taking advantage from the mentioned model of speech production by exploiting the fact that the AR filter directly

corresponds to the vocal tract filter [JN84]. Moreover, it is known from the field of error concealment and joint source channel decoding that the performance of a communication system can be improved dramatically if temporal correlation of speech signals is additionally exploited [FV01, HGV98], e.g., in terms of transition probabilities of a first order Markov model (AK1) as used in [CVA06, SV10]. Furthermore, it should be noted that speech and audio coding techniques do not only rely on purely mathematical criteria. Depending on the application, more sophisticated models for the speech production process as well as for the human auditory system are used.

For noise suppression so far, only a very limited number of proposals is known which take into account the correlation of either speech or noise samples. Generally, speech and noise are assumed to be stationary at least for a short period of time and the speech production process is not incorporated in the respective derivations. In contrast to the statistical estimators that are discussed in Chapter 2, the *Kalman filter* performs optimal estimation in a linear dynamic system in which a non-stationary target signal is disturbed by additive noise [Kal60]. The authors in [PB87] were the first to propose the use of a Kalman filter for the purpose of speech enhancement assuming the speech signal to be disturbed by *White Gaussian Noise* (WGN). Compared to the Wiener filtering method, the performance of this model-based approach was shown to be considerably better. In [GKG91], the Kalman filter was extended in order to exploit correlation of colored noise signals as well. Kalman filtering in subbands was proposed in [WC98] and [Pud02] and achieved better results than the corresponding full-band time domain approaches as well as a reduction in complexity. Most of these techniques only consider the temporal correlation within one frame and only a few publications are known which also take into account the correlation of successive speech frames, e.g., [EV11, EV08c, EV08b, EV08a, ZVY06b, ZVY06a].

In this chapter, a novel Kalman filter approach is derived which is applied in the frequency domain to the noisy short-time DFT coefficients. In a first step, a *complex-valued predictor* is used in order to exploit the temporal correlation of speech and noise in successive frames. The resulting estimates of the DFT coefficients are updated in a second step by applying adequate statistical weighting rules, amongst others *Signal-to-Noise-Ratio* (SNR)-dependent *Minimum Mean Square Error* (MMSE) estimators which are adapted to (measured) statistics of the speech prediction error signal.

The approach is mainly designed for speech communication systems. As the acoustic bandwidth of today's fixed-line and mobile networks is still limited to narrowband, i.e., the frequency range between 300 Hz and 3.4 kHz, the presented speech enhancement technique in this chapter is developed and evaluated for narrowband signals using a sampling frequency $f_s = 8\,\text{kHz}$. However, by doubling sampling rate and transform length, the same algorithm can in principle be applied to higher frequencies as well, e.g., in wideband communication systems ($f_s = 16\,\text{kHz}$). An alternative method for wideband noise suppression can be found in Chapter 4.

The remainder of this chapter is organized as follows. At first, the general concept of Kalman filtering for speech enhancement is introduced. Afterwards it is shown how correlation between adjacent frames can be exploited in the frequency domain.

Therefore, the *Probability Density Function* (PDF) of the speech prediction error signal within the Kalman filter is approximated either as Gaussian PDF or as generalized Gamma PDF. In addition, the influence of the input SNR on the statistics of the speech prediction error signal is investigated leading to an adaptive weighting rule which takes into account the mentioned SNR-dependency. The developed noise suppression technique is evaluated by means of instrumental measurements and auditory judgments. Finally, conclusions are drawn at the end of this chapter.

3.1 Kalman Filter for Speech Enhancement

The following section covers the derivation of the well-known *Kalman filter* [Kal60] for the purpose of speech enhancement. The Kalman filter is a recursive filter which estimates the state of a linear dynamic system from a series of noisy measurements. In contrast to the purely statistical estimators presented in Chapter 2, it explicitly incorporates changes in the temporal course of the input signals, i.e., it is not based on the stationarity assumption and can exploit correlation over time. The Kalman filter consists of two steps. In the first step, called *propagation step*, noisy measurements up to the previous time instance are considered and used to predict the target (speech) signal at the current time instance. In order to update this first prediction, a second step is applied in which the current noisy measurement is included in the estimation process. The second step is therefore called *update step*.

In the following, it is assumed that the speech signal $s(k)$ can be modeled as AR process given by:

$$s(k) = \sum_{\kappa=1}^{N_\mathrm{K}} a_\kappa(k) s(k-\kappa) + e_s(k), \tag{3.1}$$

where N_K is the model order, $a_\kappa(k)$ the κ-th AR coefficient at time instance k and $e_s(k)$ the excitation signal or so-called process noise. In this application, the Kalman filter addresses the general problem of trying to estimate the state vector

$$\mathbf{s}(k) = \begin{pmatrix} s(k-N_\mathrm{K}+1) \\ s(k-N_\mathrm{K}+2) \\ \vdots \\ s(k-1) \\ s(k) \end{pmatrix} \tag{3.2}$$

of the discrete-time AR process that is governed by the linear stochastic difference equation

$$\mathbf{s}(k) = \mathbf{A}(k)\mathbf{s}(k-1) + \mathbf{g}_s e_s(k) \tag{3.3}$$

from the measurement

$$y(k) = \mathbf{h}_s^T \mathbf{s}(k) + n(k), \tag{3.4}$$

where $(\cdot)^T$ denotes the transpose of a vector or matrix,

$$\mathbf{A}(k) = \begin{pmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ a_{N_\mathrm{K}}(k) & a_{N_\mathrm{K}-1}(k) & \dots & a_1(k) \end{pmatrix} \tag{3.5}$$

states the transition matrix consisting of AR coefficients, $n(k)$ the additive (so-called measurement) noise and

$$\mathbf{g}_s = \mathbf{h}_s = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{3.6}$$

The composition of the noisy signal $y(k)$ is illustrated in Fig. 3.1. The matrix $\mathbf{A}(k)$ in Eq. 3.3 relates the state at the previous time step $k-1$ to the state at the current step k in absence of the process noise. Note that in practice $\mathbf{A}(k)$ changes with each time step and has to be estimated in advance. In the following derivation, $\mathbf{A}(k)$ is assumed to be constant.

The aim of the Kalman filter approach is to estimate the current (updated) state vector $\hat{\mathbf{s}}_\mathrm{up}(k)$ based on the conditional expectation vector as follows:

$$\hat{\mathbf{s}}_\mathrm{up}(k) = \begin{pmatrix} \hat{s}_\mathrm{up}(k - N_\mathrm{K} + 1) \\ \hat{s}_\mathrm{up}(k - N_\mathrm{K} + 2) \\ \vdots \\ \hat{s}_\mathrm{up}(k-1) \\ \hat{s}_\mathrm{up}(k) \end{pmatrix} = \mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\}, \tag{3.7}$$

where $\mathbf{y}(k)$ is defined as measurement history up to time instance k:

$$\mathbf{y}(k) = \begin{pmatrix} y(0) \\ \vdots \\ y(k) \end{pmatrix}. \tag{3.8}$$

Since the state of the dynamic system changes over time, it is necessary to define an intermediate, propagated estimate $\hat{\mathbf{s}}_\mathrm{prop}(k)$:

$$\hat{\mathbf{s}}_\mathrm{prop}(k) = \begin{pmatrix} \hat{s}_\mathrm{prop}(k - N_\mathrm{K} + 1) \\ \hat{s}_\mathrm{prop}(k - N_\mathrm{K} + 2) \\ \vdots \\ \hat{s}_\mathrm{prop}(k-1) \\ \hat{s}_\mathrm{prop}(k) \end{pmatrix} = \mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k-1)\}, \tag{3.9}$$

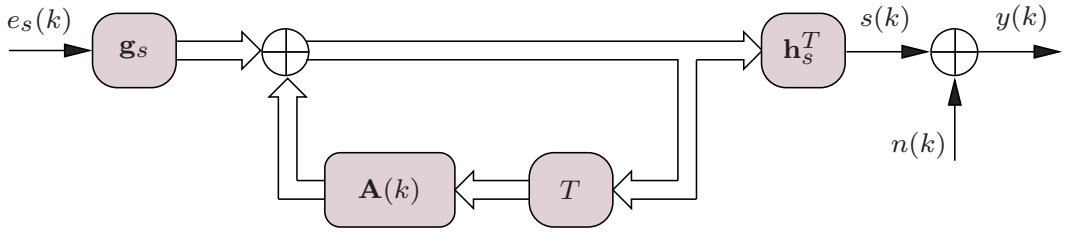

Figure 3.1: Synthesis of noisy signal $y(k)$ modeling the speech signal as AR process.

which describes how the state evolves in between measurements by incorporating information $\mathbf{y}(k-1)$ only up to time instance $k-1$. The corresponding estimation error vectors are given by:

$$\mathbf{e}^s_{\text{up}}(k) = \mathbf{s}(k) - \hat{\mathbf{s}}_{\text{up}}(k) \quad \text{and} \tag{3.10}$$

$$\mathbf{e}^s_{\text{prop}}(k) = \mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k), \tag{3.11}$$

leading to the following error covariance matrices:

$$\mathbf{P}^s_{\text{up}}(k) = \mathbb{E}\{\mathbf{e}^s_{\text{up}}(k)\left(\mathbf{e}^s_{\text{up}}(k)\right)^H\} \quad \text{and} \tag{3.12}$$

$$\mathbf{P}^s_{\text{prop}}(k) = \mathbb{E}\{\mathbf{e}^s_{\text{prop}}(k)\left(\mathbf{e}^s_{\text{prop}}(k)\right)^H\}, \tag{3.13}$$

where $(\cdot)^H$ represents the hermitian of the corresponding vector or matrix.

In the next section, the Kalman filter is derived for a discrete-time Gauss-Markov system assuming that the speech signal $s(k)$ as well as the noise signal $n(k)$ exhibit a Gaussian PDF.

3.1.1 Gaussian Model

In this section, the problem of estimating $s(k)$ given the noisy measurements $\mathbf{y}(k)$ is covered relying on Eqs. 3.3 and 3.4. It is assumed that $y(k)$, $s(k)$, $n(k)$ as well as $e_s(k)$ are Gaussian distributed with zero-mean having the powers $\sigma_y^2(k)$, $\sigma_s^2(k)$, $\sigma_n^2(k)$ and $\sigma_{e_s}^2(k)$, respectively. Furthermore, the noise signal $n(k)$ shall be independent of the speech signal $s(k)$ and the excitation signal $e_s(k)$.

The following derivation of the Kalman filter is divided into two parts. At first, only the temporal correlation of speech signals is exploited regarding the AR model in Eq. 3.1. Afterwards possible correlation of noise signals is utilized as well requiring an extension of the system.

3.1.1.1 Exploiting Correlation of Speech Signals

As mentioned before, the Kalman filter can be split into two steps, namely the propagation step and the update step. In the propagation step, the process state is estimated by projecting the previous state forward in time (see Eq. 3.9). In the subsequent update step, the system obtains feedback information from the current

(noisy) measurement (see Eq. 3.7) and updates the initial prediction. Both steps are outlined in the following.

Propagation Step

In this step, measurements $\mathbf{y}(k-1)$ up to time instance $k-1$ are considered in order to estimate the current speech state vector $\mathbf{s}(k)$. The previous estimate $\hat{\mathbf{s}}_{\text{up}}(k-1)$ is therefore propagated in time in order to account for the systematic changes of the underlying system. Using Eqs. 3.3, 3.7 and 3.9, it follows in the stationary case ($\mathbf{A}(k)$=const.):

$$\begin{aligned}\hat{\mathbf{s}}_{\text{prop}}(k) &= \mathbb{E}\{\mathbf{A}(k)\mathbf{s}(k-1) + \mathbf{g}_s e_s(k)|\mathbf{y}(k-1)\} \\ &= \mathbf{A}(k)\mathbb{E}\{\mathbf{s}(k-1)|\mathbf{y}(k-1)\} + \mathbb{E}\{\mathbf{g}_s e_s(k)|\mathbf{y}(k-1)\} \\ &= \mathbf{A}(k)\hat{\mathbf{s}}_{\text{up}}(k-1) + \underbrace{\mathbb{E}\{\mathbf{g}_s e_s(k)\}}_{=0} \\ &= \mathbf{A}(k)\hat{\mathbf{s}}_{\text{up}}(k-1). \end{aligned} \tag{3.14}$$

Similarly, an expression for the error covariance matrix $\mathbf{P}^s_{\text{prop}}(n)$ can be derived with Eqs. 3.3 and 3.10 for the propagation step:

$$\begin{aligned}\mathbf{P}^s_{\text{prop}}(n) &= \mathbb{E}\{(\mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k))\,(\mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k))^H\} \\ &= \mathbb{E}\{(\mathbf{s}(k) - \mathbf{A}(k)\hat{\mathbf{s}}_{\text{up}}(k-1))\,(\mathbf{s}(k) - \mathbf{A}(k)\hat{\mathbf{s}}_{\text{up}}(k-1))^H\} \\ &= \mathbb{E}\{\left(\mathbf{A}(k)\mathbf{e}^s_{\text{up}}(k-1) + \mathbf{g}_s e_s(k)\right)\left(\mathbf{A}(k)\mathbf{e}^s_{\text{up}}(k-1) + \mathbf{g}_s e_s(k)\right)^H\} \\ &= \mathbf{A}(k)\mathbb{E}\{\mathbf{e}^s_{\text{up}}(k-1)\left(\mathbf{e}^s_{\text{up}}(k-1)\right)^H\}\mathbf{A}(k)^H + \mathbf{g}_s\sigma^2_{e_s}(k-1)\mathbf{g}_s^H \\ &= \mathbf{A}(k)\mathbf{P}^s_{\text{up}}(k-1)\mathbf{A}(k)^H + \mathbf{g}_s\sigma^2_{e_s}(k-1)\mathbf{g}_s^H. \end{aligned} \tag{3.15}$$

In general, the predictions in the propagation step are erroneous and a non-zero estimation error vector $\mathbf{e}^s_{\text{prop}}(k)$ and a non-zero covariance matrix $\mathbf{P}^s_{\text{prop}}(k)$ occur. The following update step estimates the resulting estimation errors by incorporating the current (noisy) measurement $y(k)$.

Update Step

In contrast to the propagation step, the update step considers all information available at time instance k, cf. Eq. 3.7. The conditional expectation $\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\}$ in Eq. 3.7 is derived using the conditional PDF $p(\mathbf{s}(k)|\mathbf{y}(k))$ according to:

$$\hat{\mathbf{s}}_{\text{up}}(k) = \mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \mathbf{s}(k)\cdot p(\mathbf{s}(k)|\mathbf{y}(k))\,\mathrm{d}\mathbf{s}(k). \tag{3.16}$$

As all signals are assumed to be Gaussian distributed, the PDF $p(\mathbf{s}(k)|\mathbf{y}(k))$ follows a Gaussian distribution as well. The estimate in the update step finally results in (see Appendix A.1):

$$\hat{\mathbf{s}}_{\text{up}}(k) = \hat{\mathbf{s}}_{\text{prop}}(k) + \mathbf{k}^s(k)\left(y(k) - \mathbf{h}_s^T\hat{\mathbf{s}}_{\text{prop}}(k)\right), \tag{3.17}$$

where $\mathbf{k}^s(k)$ states the *Kalman filter gain* (see Appendix A.1):

$$\mathbf{k}^s(k) = \mathbf{P}^s_{\text{prop}}(k) \cdot \mathbf{h}_s \left(\mathbf{h}_s^T \mathbf{P}^s_{\text{prop}}(k) \mathbf{h}_s + \sigma_n^2(k)\right)^{-1}. \tag{3.18}$$

The difference $y(k) - \mathbf{h}_s^T \hat{\mathbf{s}}_{\text{prop}}(k)$ in Eq. 3.17 is called *measurement innovation* or *differential signal* $d_s(k)$.

In order to determine an expression for the error covariance matrix $\mathbf{P}^s_{\text{up}}(k)$ in the update step, the estimation error $\mathbf{e}^s_{\text{up}}(k)$ is also stated as update recursion at first:

$$\begin{aligned}\mathbf{e}^s_{\text{up}}(k) &= \mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k) - \mathbf{k}^s(k) \left(y(k) - \mathbf{h}_s^T \hat{\mathbf{s}}_{\text{prop}}(k)\right) \\ &= \mathbf{e}^s_{\text{prop}}(k) - \mathbf{k}^s(k) \left(y(k) - \mathbf{h}_s^T \hat{\mathbf{s}}_{\text{prop}}(k)\right). \end{aligned} \tag{3.19}$$

Inserting Eqs. 3.4 and 3.19 in Eq. 3.12 yields:

$$\begin{aligned}\mathbf{P}^s_{\text{up}}(k) =& \mathbb{E}\Big\{ \left(\mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k) - \mathbf{k}^s(k) \left(\mathbf{h}_s^T \mathbf{s}(k) + n(k) - \mathbf{h}_s^T \hat{\mathbf{s}}_{\text{prop}}(k)\right)\right) \\ & \cdot \left(\mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k) - \mathbf{k}^s(k) \left(\mathbf{h}_s^T \mathbf{s}(k) + n(k) - \mathbf{h}_s^T \hat{\mathbf{s}}_{\text{prop}}(k)\right)\right)^H \Big\} \\ =& \mathbb{E}\Big\{ \left(\left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right) \left(\mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k)\right) - \mathbf{k}^s(k) n(k)\right) \\ & \cdot \left(\left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right) \left(\mathbf{s}(k) - \hat{\mathbf{s}}_{\text{prop}}(k)\right) - \mathbf{k}^s(k) n(k)\right)^H \Big\} \\ =& \mathbb{E}\Big\{ \left(\left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right) \mathbf{e}^s_{\text{prop}}(k) - \mathbf{k}^s(k) n(k)\right) \\ & \cdot \left(\left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right) \mathbf{e}^s_{\text{prop}}(k) - \mathbf{k}^s(k) n(k)\right)^H \Big\} \\ =& \left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right) \mathbf{P}^s_{\text{prop}}(k) \left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right)^H + \mathbf{k}^s(k) \sigma_n^2(k) \left(\mathbf{k}^s(k)\right)^H \\ =& \mathbf{P}^s_{\text{prop}}(k) - \mathbf{k}^s(k) \mathbf{h}_s^T \mathbf{P}^s_{\text{prop}}(k) - \mathbf{P}^s_{\text{prop}}(k) \mathbf{h}_s (\mathbf{k}^s(k))^H \\ & + \mathbf{k}^s(k) (\mathbf{h}_s^T \mathbf{P}^s_{\text{prop}}(k) \mathbf{h}_s + \sigma_n^2(k)) (\mathbf{k}^s(k))^H, \end{aligned} \tag{3.20}$$

with $\mathbf{I}$ being the identity matrix. Using the Kalman filter gain of Eq. 3.18 for the last summand, the error covariance matrix $\mathbf{P}^s_{\text{up}}(k)$ is given by:

$$\begin{aligned}\mathbf{P}^s_{\text{up}}(k) =& \mathbf{P}^s_{\text{prop}}(k) - \mathbf{k}^s(k) \mathbf{h}_s^T \mathbf{P}^s_{\text{prop}}(k) - \mathbf{P}^s_{\text{prop}}(k) \mathbf{h}_s (\mathbf{k}^s(k))^H \\ & + \mathbf{P}^s_{\text{prop}}(k) \mathbf{h}_s (\mathbf{k}^s(k))^H \\ =& \left(\mathbf{I} - \mathbf{k}^s(k) \mathbf{h}_s^T\right) \mathbf{P}^s_{\text{prop}}(k). \end{aligned} \tag{3.21}$$

3.1.1.2 Interpretation

The afore derived Kalman filter can be applied in a noise reduction system and is able to exploit temporal correlation of speech signals. In principle, the estimator can be summarized into the following five steps, which have to be carried out recursively after initialization of the system:

System Initialization			
Propagation Step:	$\hat{\mathbf{s}}_{\text{prop}}(k)$	$=$	$\mathbf{A}(k)\hat{\mathbf{s}}_{\text{up}}(k-1)$
	$\mathbf{P}^s_{\text{prop}}(k)$	$=$	$\mathbf{A}(k)\mathbf{P}^s_{\text{up}}(k-1)\mathbf{A}(k)^H + \mathbf{g}_s\sigma^2_{e_s}(k-1)\mathbf{g}_s^H$
Update Step:	$\mathbf{k}^s(k)$	$=$	$\mathbf{P}^s_{\text{prop}}(k)\cdot\mathbf{h}_s\left(\mathbf{h}_s^T\mathbf{P}^s_{\text{prop}}(k)\mathbf{h}_s + \sigma^2_n(k)\right)^{-1}$
	$\hat{\mathbf{s}}_{\text{up}}(k)$	$=$	$\hat{\mathbf{s}}_{\text{prop}}(k) + \mathbf{k}^s(k)\left(y(k) - \mathbf{h}_s^T\hat{\mathbf{s}}_{\text{prop}}(k)\right)$
	$\mathbf{P}^s_{\text{up}}(k)$	$=$	$\left(\mathbf{I} - \mathbf{k}^s(k)\mathbf{h}_s^T\right)\mathbf{P}^s_{\text{prop}}(k).$

In this process, the propagation step acts as a predictor yielding a first, initial speech estimate which is then corrected/updated in the update step.

In the following, the update step is examined in more detail. The covariance matrix $\mathbf{P}^s_{\text{prop}}(k)$ can be written as:

$$\mathbf{P}^s_{\text{prop}}(k) = \begin{pmatrix} \sigma^2_{e^s_{\text{prop}}}\binom{k-N_{\text{K}}+1}{k-N_{\text{K}}+1} & \cdots & \sigma^2_{e^s_{\text{prop}}}\binom{k-N_{\text{K}}+1}{k} \\ \vdots & \ddots & \vdots \\ \sigma^2_{e^s_{\text{prop}}}\binom{k}{k-N_{\text{K}}+1} & \cdots & \sigma^2_{e^s_{\text{prop}}}\binom{k}{k} \end{pmatrix}, \tag{3.22}$$

where $\sigma^2_{e^s_{\text{prop}}}\binom{k-\kappa_1+1}{k-\kappa_2+1} = \mathbb{E}\{e^s_{\text{prop}}(k-\kappa_1+1)\left(e^s_{\text{prop}}(k-\kappa_2+1)\right)^*\}$. The operator $(\cdot)^*$ denotes the complex-conjugate and $\kappa_1, \kappa_2 \in \{1, ..., N_{\text{K}}\}$. Inserting Eq. 3.22 into Eq. 3.18, the Kalman filter gain is given by:

$$\mathbf{k}^s(k) = \frac{1}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k} + \sigma^2_n(k)} \begin{pmatrix} \sigma^2_{e^s_{\text{prop}}}\binom{k-N_{\text{K}}+1}{k} \\ \vdots \\ \sigma^2_{e^s_{\text{prop}}}\binom{k}{k} \end{pmatrix}. \tag{3.23}$$

In the update step, $\mathbf{k}^s(k)$ is applied to the differential signal $d_s(k)$ consisting of the prediction error $e^s_{\text{prop}}(k)$ disturbed by the initial noise signal $n(k)$, cf. Eq. 3.17:

$$\begin{aligned} d_s(k) &= y(k) - \mathbf{h}_s^T\hat{\mathbf{s}}_{\text{prop}}(k) \\ &= s(k) + n(k) - \hat{s}_{\text{prop}}(k) \\ &= e^s_{\text{prop}}(k) + n(k). \end{aligned} \tag{3.24}$$

Hence, the task in the update step becomes a 'classical' noise reduction problem: Decomposition of the 'noisy' input signal $d_s(k)$ into the (new) 'target' signal $e^s_{\text{prop}}(k)$

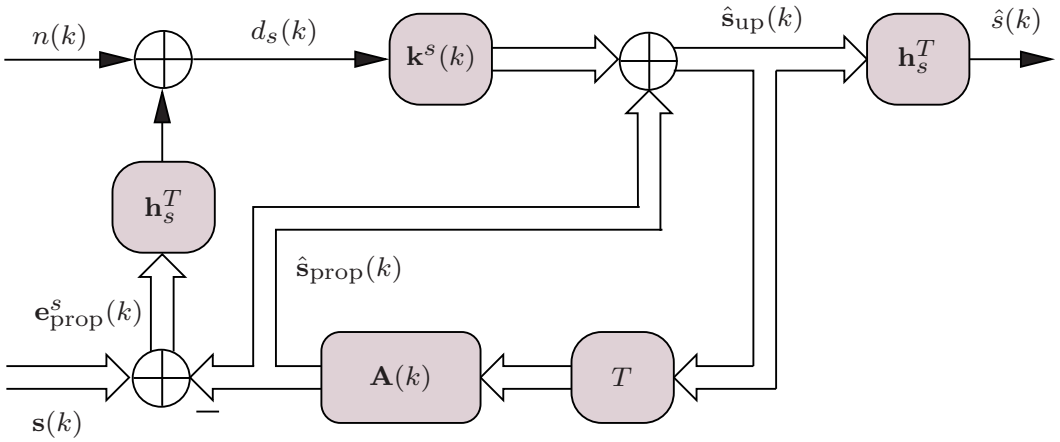

Figure 3.2: Time domain Kalman filter exploiting correlation of speech signals.

and the noise signal $n(k)$. Using Eq. 3.24, the update part of Eq. 3.17 can be stated as:

$$\mathbf{k}^s(k)\left(y(k)-\mathbf{h}_s^T\hat{\mathbf{s}}_{\text{prop}}(k)\right)=\frac{d_s(k)}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}+\sigma_n^2(k)}\begin{pmatrix}\sigma^2_{e^s_{\text{prop}}}\binom{k-N_{\text{K}}+1}{k}\\ \vdots\\ \sigma^2_{e^s_{\text{prop}}}\binom{k}{k}\end{pmatrix}. \tag{3.25}$$

It can be shown that this expression equals the conditional expectation $\mathbb{E}\{\mathbf{e}^s_{\text{prop}}(k)|d_s(k)=e^s_{\text{prop}}(k)+n(k)\}$ (see Appendix A.2) under the assumption that $e^s_{\text{prop}}(k)$ and $n(k)$ are Gaussian distributed. The signal flow of the derived Kalman filter is illustrated in Fig. 3.2. Please note that $\mathbf{s}(k)$ and $n(k)$ are depicted separately just for demonstration purposes. In a real system, only the noisy input vector $\mathbf{y}(k)$ is accessible.

3.1.1.3 Extension to Colored Noise Signals

In order to additionally exploit the temporal correlation of the noise signal $n(k)$, the Kalman filter system that is presented in the previous sections is extended in the following. Therefore, the noise signal is also modeled as AR process, cf. Eq. 3.1 and Fig. 3.3, given by:

$$n(k)=\sum_{\tau=1}^{M_{\text{K}}}b_\tau(k)n(k-\tau)+e_n(k), \tag{3.26}$$

with noise AR coefficients $b_\tau(k)$, noise model order M_{K} and noise excitation signal $e_n(k)$ exhibiting the power $\sigma^2_{e_n}(k)$. The corresponding difference equation according to the speech case in Eq. 3.3 results in:

$$\mathbf{n}(k)=\mathbf{B}(k)\mathbf{n}(k-1)+\mathbf{g}_n e_n(k), \tag{3.27}$$

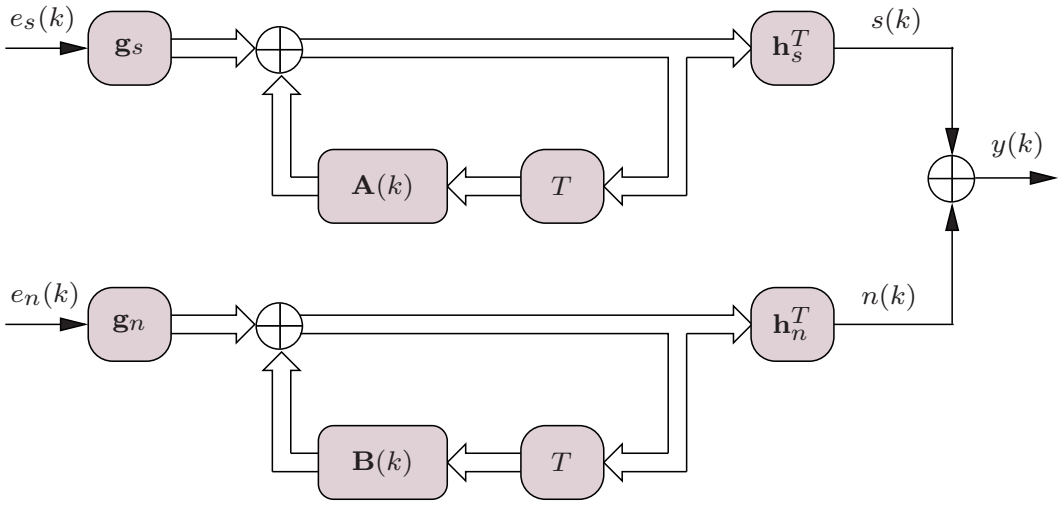

Figure 3.3: Synthesis of noisy signal $y(k)$ modeling the speech *and* noise as AR processes.

depending on the noise state vector $\mathbf{n}(k)$:

$$\mathbf{n}(k) = \begin{pmatrix} n(k - M_\mathrm{K} + 1) \\ n(k - M_\mathrm{K} + 2) \\ \vdots \\ n(k-1) \\ n(k) \end{pmatrix}, \tag{3.28}$$

the noise transition matrix $\mathbf{B}(k)$:

$$\mathbf{B}(k) = \begin{pmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ b_{M_\mathrm{K}}(k) & b_{M_\mathrm{K}-1}(k) & \dots & b_1(k) \end{pmatrix}, \tag{3.29}$$

and

$$\mathbf{g}_n = \mathbf{h}_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{3.30}$$

The extended system also provides a propagation step and an update step for the noise signal, which can be derived analogously to the speech case in Sec. 3.1.1.1. In the *propagation step*, possible correlation of the noise signal is exploited by estimating the current noise sample $n(k)$ based on the noisy information $\mathbf{y}(k-1)$ from the past:

$$\hat{\mathbf{n}}_\mathrm{prop}(k) = \begin{pmatrix} \hat{n}_\mathrm{prop}(k - M_\mathrm{K} + 1) \\ \hat{n}_\mathrm{prop}(k - M_\mathrm{K} + 2) \\ \vdots \\ \hat{n}_\mathrm{prop}(k-1) \\ \hat{n}_\mathrm{prop}(k) \end{pmatrix} = \mathbb{E}\{\mathbf{n}(k)|\mathbf{y}(k-1)\}. \tag{3.31}$$

This first prediction is updated in the *update step* by additionally considering the current (noisy) measurement $y(k)$ resulting in:

$$\hat{\mathbf{n}}_{\text{up}}(k) = \begin{pmatrix} \hat{n}_{\text{up}}(k - M_{\text{K}} + 1) \\ \hat{n}_{\text{up}}(k - M_{\text{K}} + 2) \\ \vdots \\ \hat{n}_{\text{up}}(k-1) \\ \hat{n}_{\text{up}}(k) \end{pmatrix} = \mathbb{E}\{\mathbf{n}(k)|\mathbf{y}(k)\}. \tag{3.32}$$

Applying a similar derivation as in Sec. 3.1.1.1, both steps can be summarized for the noise signal as follows:

System Initialization			
Propagation Step:	$\hat{\mathbf{n}}_{\text{prop}}(k)$	$=$	$\mathbf{B}(k)\hat{\mathbf{n}}_{\text{up}}(k-1)$
	$\mathbf{P}^n_{\text{prop}}(k)$	$=$	$\mathbf{B}(k)\mathbf{P}^n_{\text{up}}(k-1)\mathbf{B}(k)^H + \mathbf{g}_n\sigma^2_{e_n}(k-1)\mathbf{g}_n^H$
Update Step:	$\mathbf{k}^n(k)$	$=$	$\mathbf{P}^n_{\text{prop}}(k)\cdot\mathbf{h}_n\left(\mathbf{h}_n^T\mathbf{P}^n_{\text{prop}}(k)\mathbf{h}_n + \sigma^2_s(k)\right)^{-1}$
	$\hat{\mathbf{n}}_{\text{up}}(k)$	$=$	$\hat{\mathbf{n}}_{\text{prop}}(k) + \mathbf{k}^n(k)\left(y(k) - \mathbf{h}_n^T\hat{\mathbf{n}}_{\text{prop}}(k)\right)$
	$\mathbf{P}^n_{\text{up}}(k)$	$=$	$\left(\mathbf{I} - \mathbf{k}^n(k)\mathbf{h}_n^T\right)\mathbf{P}^n_{\text{prop}}(k).$

The corresponding error vectors and covariance matrices are given by:

$$\mathbf{e}^n_{\text{up}}(k) = \mathbf{n}(k) - \hat{\mathbf{n}}_{\text{up}}(k), \tag{3.33}$$

$$\mathbf{e}^n_{\text{prop}}(k) = \mathbf{n}(k) - \hat{\mathbf{n}}_{\text{prop}}(k), \tag{3.34}$$

$$\mathbf{P}^n_{\text{up}}(k) = \mathbb{E}\{\mathbf{e}^n_{\text{up}}(k)\left(\mathbf{e}^n_{\text{up}}(k)\right)^H\} \quad \text{and} \tag{3.35}$$

$$\mathbf{P}^n_{\text{prop}}(k) = \mathbb{E}\{\mathbf{e}^n_{\text{prop}}(k)\left(\mathbf{e}^n_{\text{prop}}(k)\right)^H\}. \tag{3.36}$$

The update steps of the speech and noise signal can be combined by considering the joint differential signal $d(k)$:

$$\begin{aligned} d(k) &= y(k) - \underbrace{\left(\mathbf{h}_s^T\hat{\mathbf{s}}_{\text{prop}}(k) + \mathbf{h}_n^T\hat{\mathbf{n}}_{\text{prop}}(k)\right)}_{\hat{y}_{\text{prop}}(k)} \\ &= s(k) - \hat{s}_{\text{prop}}(k) + n(k) - \hat{n}_{\text{prop}}(k) \\ &= e^s_{\text{prop}}(k) + e^n_{\text{prop}}(k). \end{aligned} \tag{3.37}$$

In this case, the update step equations for speech and noise can be reformulated as:

$$\mathbf{k}^s(k) = \mathbf{P}^s_{\text{prop}}(k)\cdot\mathbf{h}_s\left(\mathbf{h}_s^T\mathbf{P}^s_{\text{prop}}(k)\mathbf{h}_s + \sigma^2_{e^n_{\text{prop}}}\begin{pmatrix}k\\k\end{pmatrix}\right)^{-1}, \tag{3.38}$$

$$\hat{\mathbf{s}}_{\text{up}}(k) = \hat{\mathbf{s}}_{\text{prop}}(k) + \mathbf{k}^s(k)d(k), \tag{3.39}$$

where $\sigma^2_{e^n_{\text{prop}}}\binom{k-\tau_1+1}{k-\tau_2+1} = \mathbb{E}\{e^n_{\text{prop}}(k-\tau_1+1)\left(e^n_{\text{prop}}(k-\tau_2+1)\right)^*\}$ with $\tau_1, \tau_2 \in \{1, ..., M_{\text{K}}\}$ and:

$$\mathbf{k}^n(k) = \mathbf{P}^n_{\text{prop}}(k) \cdot \mathbf{h}_n \left(\mathbf{h}_n^T \mathbf{P}^n_{\text{prop}}(k)\mathbf{h}_n + \sigma^2_{e^s_{\text{prop}}}\binom{k}{k}\right)^{-1}, \tag{3.40}$$

$$\hat{\mathbf{n}}_{\text{up}}(k) = \hat{\mathbf{n}}_{\text{prop}}(k) + \mathbf{k}^n(k)d(k). \tag{3.41}$$

Using the extended system, the task of the update step is to estimate the two prediction errors $e^s_{\text{prop}}(k)$ and $e^n_{\text{prop}}(k)$ given the 'noisy' signal $d(k)$ according to Eq. 3.37. Assuming Gaussian distributions for both prediction error signals and statistical independence[1] of $e^s_{\text{prop}}(k)$ and $e^n_{\text{prop}}(k)$, this problem results again in estimating the conditional expectation vectors $\mathbb{E}\{\mathbf{e}^s_{\text{prop}}(k)|d(k)\}$ and $\mathbb{E}\{\mathbf{e}^n_{\text{prop}}(k)|d(k)\}$, cf. Sec. 3.1.1.2. A block diagram of the extended Kalman filter system which is capable of exploiting the correlation of speech *and* noise signals is shown in Fig. 3.4. Here again, the speech and noise vectors $\mathbf{s}(k)$ and $\mathbf{n}(k)$ are usually not available separately in a realistic scenario and depicted only to illustrate the composition of the two prediction error vectors.

3.2 Exploiting Inter-Frame Correlation in the Frequency Domain

In the previous section, speech and noise signals are modeled as AR processes characterized by excitation signals and AR coefficients $a_\kappa(k)$ and $b_\tau(k)$, respectively. The AR coefficients describe the spectral envelopes of either the speech or the noise signal and can be interpreted as filter parameters for the vocal tract filter in the speech case. In [Pud02], it is shown that for a sampling frequency $f_{\text{s}} = 8\,\text{kHz}$ very high model orders of $N_{\text{K}} > 80$ are required for speech signals in order to obtain good approximations of the fine structure of the actual magnitude spectrum. High model orders are necessary especially during voiced speech segments to be able to resolve the fine pitch structure. A division into *Linear Predictive Coding* (LPC) and *Long Term Prediction* (LTP) parts, as known from the field of speech coding [VM06, KP95] in order to reduce the number of AR coefficients, is a challenging task as it is very difficult to accurately estimate the pitch frequency from a noisy signal, especially in low SNR conditions [SN09]. The dimension of the required noise model order M_{K} can not be stated in general as it highly depends on the specific type of noise signal. Nevertheless, when using such high model orders for speech and/or noise, it is very difficult to obtain reliable estimates of the corresponding AR coefficients. In order to prevent this problem, subband Kalman filtering is proposed, e.g., in [Pud02]. In this approach, the input signal is decomposed into 16 subbands ($f_{\text{s}} = 8\,\text{kHz}$) and one Kalman filter is separately applied in each subband to the speech and the noise signal. Therefore, [Pud02] mainly focuses on an adequate estimation of the required

[1] Strictly speaking, the independence assumption for the prediction errors only holds if a perfect prediction in the propagation step is possible, i.e., if $e^s_{\text{prop}}(k) = e_s(k)$ and $e^n_{\text{prop}}(k) = e_n(k)$.

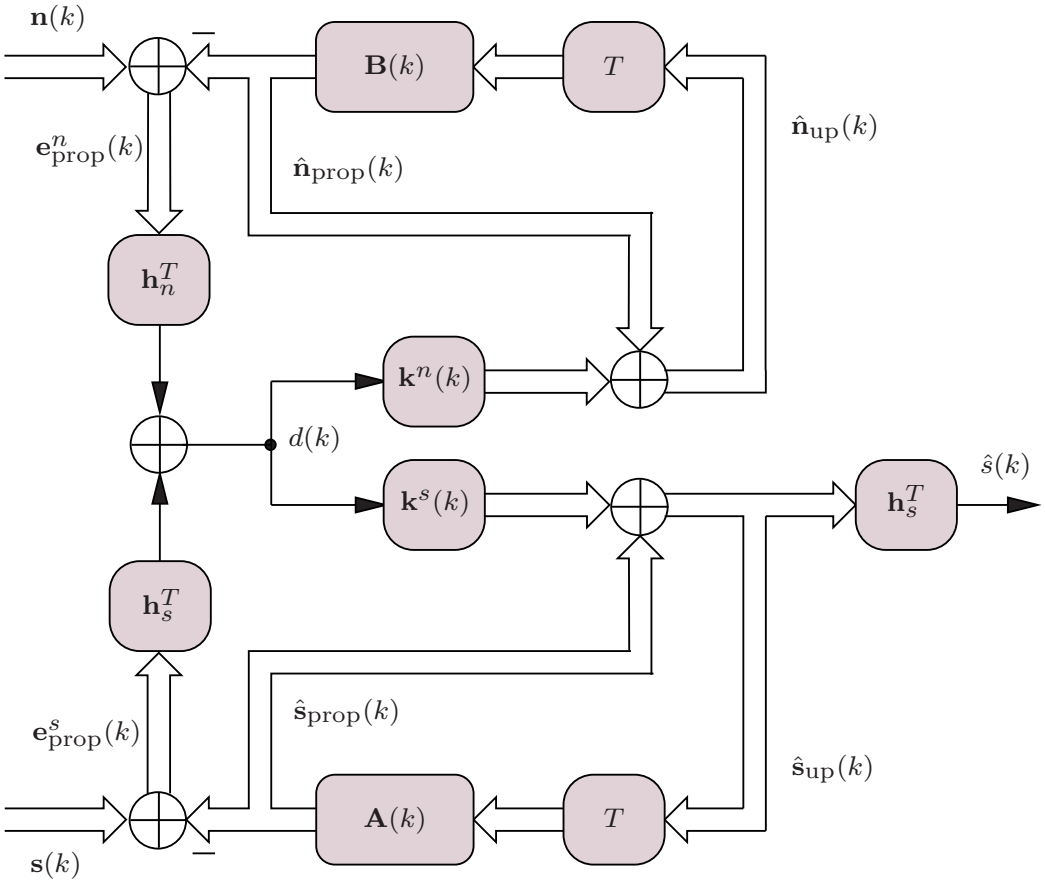

Figure 3.4: Time domain Kalman filter exploiting correlation of speech *and* noise signals.

model parameters for the subbands if the speech signal is degraded by car noise. Since the spectral envelopes of speech and noise in each subband are smoother than their corresponding full-band signals, their shape can be estimated from the noisy data more easily. Thus, lower model orders N_{K} and M_{K} are sufficient leading to low-order Kalman filters. The reduction of the model order is proportional to the number of subbands or the downsampling rate in the subbands, respectively. In addition to better model estimation properties, subband Kalman filtering can also reduce the computational complexity of the system. In [WC98], it is shown that the number of operations required for the Kalman filter structure according to Sec. 3.1.1.3 is of order $\mathcal{O}(N_{\text{K}}^2 + M_{\text{K}}^2)$. Depending on the noise model order M_{K}, the computational load can be reduced by subband processing up to 80% compared to the full-band approach [WC98].

For the derivation of the Kalman filter in Sec. 3.1, it is assumed that the transition matrices $\mathbf{A}(k)$ for the speech signal and $\mathbf{B}(k)$ for the noise signal remain constant during the whole process. In analogy to the autocorrelation method, e.g., [VM06], where the LPC coefficients are first derived for stationary signals but actually determined framewise using the short-term autocorrelation, the solution of Sec. 3.1 is also valid for time-varying AR coefficients assuming the observed process to be stationary only

over a limited period of time. In this case, the respective AR parameters of speech and noise have to be estimated in advance either in each time step or in each frame given the noisy input signal. Common methods for this purpose are, e.g.,:

- Estimation of the autocorrelation matrix based on the noisy input data or enhanced signals from the past by means of the *Yule Walker equations* [Kay88] and use of the *Levinson-Durbin algorithm* [MGj76, Hay96] to determine the AR coefficients.
- Recursive estimation using the *Burg algorithm* [Pud02] which can be combined with *Voice Activity Detection* (VAD) in order to improve the estimation results [Kap05].
- Iterative *Expectation Maximization* (EM)-based algorithms as proposed, e.g., in [GBW98], [CD08] and [Ton77].
- Recursive *Least-Square* (LS) algorithms as proposed, e.g., in [Gab05], [SA79], [Zhe99] and [JJYW02]. A comparative evaluation of different LS methods can be found in [JKYW03].
- Model-based approaches exploiting a priori information of the AR coefficients of speech and noise, e.g., [KK01, KK06, WASA07, LMS96].

In [ZVY06b], a Kalman filter system is applied in the frequency domain to the real and imaginary parts of the noisy DFT coefficients. The temporal trajectories of the real and imaginary parts of speech and noise are modeled as low-order AR processes separately for each frequency bin. Therefore, two Kalman filters are required: the first one is applied to the real-parts of the noisy DFT coefficients and the second one to the imaginary parts. Both Kalman filters include propagation and update steps for the speech and the noise signal. The speech estimates of both filters are finally combined to get an estimate of the complex-valued speech DFT coefficients. As proposed in [ZVY06b], the AR coefficients of the noise signal are obtained using a simple VAD algorithm that assumes speech inactivity in the first few frames. From these frames a noise model is derived, which is updated in noise-only frames. In contrast, the AR coefficients of the speech signal are estimated continuously from the previous enhanced spectral coefficients using the Levinson-Durbin algorithm [MGj76, Hay96].

Kalman filtering in the frequency domain possesses two advantages:

1. The *Fast Fourier Transform* (FFT) size is usually considerably higher than the number of subbands used, e.g., in [WC98] and [Pud02]. This has the effect that the temporal trajectories of speech and noise can be estimated more accurately in noisy environments using even lower model orders.
2. It is well known that successive speech DFT coefficients are correlated over time depending on frame shift and frequency [Coh05b]. This temporal correlation can be exploited in the frequency domain between adjacent frames.

In this thesis, a novel Kalman filter approach is proposed which is applied in the frequency domain *directly* to the complex-valued DFT coefficients. Compared to previous solutions, modifications and improvements are made in both the propagation *and* update step. In the propagation step, *complex-valued* prediction is used to exploit the temporal correlation of successive speech and noise DFT coefficients. As will be shown, this new concept works better than estimating real and imaginary parts or magnitudes and phases separately. The resulting prediction errors of the propagation step are estimated in the update step applying different statistical estimators. As a novelty, the SNR-dependent statistics of the differential signal are intensively studied and exploited, finally leading to the application of *SNR-dependent MMSE estimators* in the update step which are adapted to the (measured) statistics of the speech prediction error signal.

In the following, the developed model-based approach is presented. Therefore, at first an overview of the new system is given followed by a detailed description of the individual Kalman filter steps, i.e., the propagation as well as the update step.

3.2.1 System Overview

Figure 3.5 illustrates a simplified block diagram of the proposed system. For the decomposition of the speech and the noise signal, the noisy signal $y(k)$ is segmented into overlapping frames of length L_{F} with frame shift size L_{FS}. The FFT is applied to these frames after windowing and zero-padding, cf. Sec. 2.2. Hence, the spectral DFT coefficients of the noisy input signal at frame λ and frequency bin μ are given by:

$$Y(\lambda, \mu) = S(\lambda, \mu) + N(\lambda, \mu), \tag{3.42}$$

where $S(\lambda, \mu)$ and $N(\lambda, \mu)$ represent the spectral DFT coefficients of the speech and the noise signal.

In contrast to [ZVY06b], the proposed model-based system is directly applied to the *complex-valued* DFT coefficients $Y(\lambda, \mu)$ with *low-order Kalman filters* for speech and noise running in parallel for each frequency bin. In accordance to the Kalman filter for speech enhancement in the time domain (see Sec. 3.1), this approach also consists of two steps:

- In the *propagation step*, temporal correlation of successive frames is exploited. The current DFT coefficients of speech $S(\lambda, \mu)$ and noise $N(\lambda, \mu)$ are propagated in time based on information taken from previous, enhanced DFT coefficients using linear prediction techniques as shown later. From the resulting vectors $\hat{\mathbf{S}}_{\mathrm{prop}}(\lambda, \mu)$ and $\hat{\mathbf{N}}_{\mathrm{prop}}(\lambda, \mu)$ which consist of the speech and noise estimates from the previous N_{K} and M_{K} frames, the current predictions $\hat{S}_{\mathrm{prop}}(\lambda, \mu)$ and $\hat{N}_{\mathrm{prop}}(\lambda, \mu)$ are

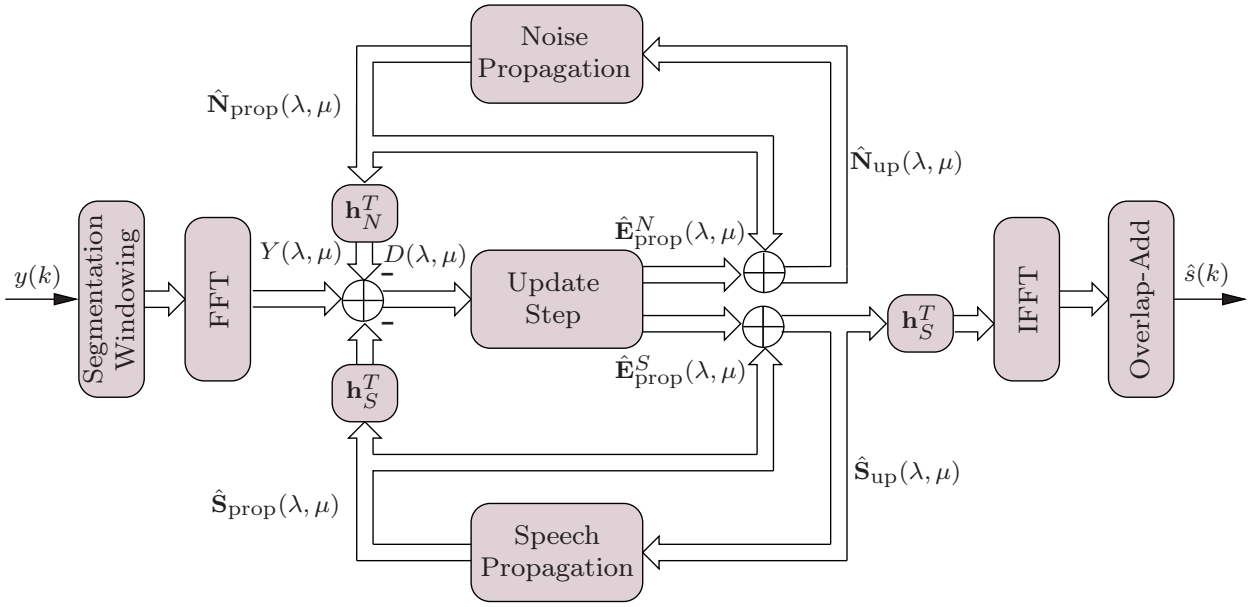

Figure 3.5: Block diagram of the proposed Kalman filter structure working in the frequency domain.

summed up to get estimates of the current noisy DFT coefficients:

$$\begin{aligned}\hat{Y}_{\text{prop}}(\lambda,\mu) &= \mathbf{h}_S^T \underbrace{\begin{pmatrix} \hat{S}_{\text{prop}}(\lambda - N_{\text{K}} + 1,\mu) \\ \hat{S}_{\text{prop}}(\lambda - N_{\text{K}} + 2,\mu) \\ \vdots \\ \hat{S}_{\text{prop}}(\lambda - 1,\mu) \\ \hat{S}_{\text{prop}}(\lambda,\mu) \end{pmatrix}}_{\hat{\mathbf{S}}_{\text{prop}}(\lambda,\mu)} + \mathbf{h}_N^T \underbrace{\begin{pmatrix} \hat{N}_{\text{prop}}(\lambda - M_{\text{K}} + 1,\mu) \\ \hat{N}_{\text{prop}}(\lambda - M_{\text{K}} + 2,\mu) \\ \vdots \\ \hat{N}_{\text{prop}}(\lambda - 1,\mu) \\ \hat{N}_{\text{prop}}(\lambda,\mu) \end{pmatrix}}_{\hat{\mathbf{N}}_{\text{prop}}(\lambda,\mu)} \\ &= \hat{S}_{\text{prop}}(\lambda,\mu) + \hat{N}_{\text{prop}}(\lambda,\mu), \end{aligned} \tag{3.43}$$

where $\mathbf{h}_S$ and $\mathbf{h}_N$ are the equivalents to $\mathbf{h}_s$ and $\mathbf{h}_n$ in the time domain.

The prediction in the propagation step generally is erroneous as well, resulting in the following two prediction error vectors:

$$\mathbf{E}^S_{\text{prop}}(\lambda,\mu) = \begin{pmatrix} S(\lambda - N_{\text{K}} + 1,\mu) \\ S(\lambda - N_{\text{K}} + 2,\mu) \\ \vdots \\ S(\lambda - 1,\mu) \\ S(\lambda,\mu) \end{pmatrix} - \hat{\mathbf{S}}_{\text{prop}}(\lambda,\mu) \quad \text{and} \tag{3.44}$$

$$\mathbf{E}^N_{\text{prop}}(\lambda,\mu) = \begin{pmatrix} N(\lambda - M_{\text{K}} + 1,\mu) \\ N(\lambda - M_{\text{K}} + 2,\mu) \\ \vdots \\ N(\lambda - 1,\mu) \\ N(\lambda,\mu) \end{pmatrix} - \hat{\mathbf{N}}_{\text{prop}}(\lambda,\mu). \tag{3.45}$$

- The objective in the *update step* is to estimate these error vectors incorporating the new measurement information $Y(\lambda,\mu)$ of the current frame:

$$\begin{aligned} D(\lambda,\mu) &= Y(\lambda,\mu) - \hat{Y}_{\text{prop}}(\lambda,\mu) \\ &= Y(\lambda,\mu) - \mathbf{h}_S^T \hat{\mathbf{S}}_{\text{prop}}(\lambda,\mu) - \mathbf{h}_N^T \hat{\mathbf{N}}_{\text{prop}}(\lambda,\mu). \end{aligned} \tag{3.46}$$

  Following the Kalman filter approach in the time domain of Sec. 3.1, the differential signal $D(\lambda,\mu)$ is multiplied by Kalman gains $\mathbf{K}$ in order to determine estimates of the prediction error vectors in the update step:

$$\hat{\mathbf{E}}_{\text{prop}}^S(\lambda,\mu) = \mathbf{K}^S(\lambda,\mu) \cdot D(\lambda,\mu) \tag{3.47}$$

$$\hat{\mathbf{E}}_{\text{prop}}^N(\lambda,\mu) = \mathbf{K}^N(\lambda,\mu) \cdot D(\lambda,\mu), \tag{3.48}$$

  where $\mathbf{K}^S(\lambda,\mu)$ and $\mathbf{K}^N(\lambda,\mu)$ state the Kalman filter gains for the speech and the noise cases in the frequency domain. To obtain the final enhanced DFT coefficient vectors $\hat{\mathbf{S}}_{\text{up}}(\lambda,\mu)$ and $\hat{\mathbf{N}}_{\text{up}}(\lambda,\mu)$, the initial predictions of the propagation step are updated:

$$\hat{\mathbf{S}}_{\text{up}}(\lambda,\mu) = \hat{\mathbf{S}}_{\text{prop}}(\lambda,\mu) + \hat{\mathbf{E}}_{\text{prop}}^S(\lambda,\mu) \tag{3.49}$$

$$\hat{\mathbf{N}}_{\text{up}}(\lambda,\mu) = \hat{\mathbf{N}}_{\text{prop}}(\lambda,\mu) + \hat{\mathbf{E}}_{\text{prop}}^N(\lambda,\mu). \tag{3.50}$$

The *Inverse Fast Fourier Transform* (IFFT) of $\mathbf{h}_S^T \hat{\mathbf{S}}_{\text{up}}(\lambda,\mu)$ and the overlap-add method yield an estimate of the enhanced speech signal $\hat{s}(k)$ in the time domain, cf. Sec. 2.2.

3.2.2 Propagation Step

This section addresses the basic principles of the aforementioned propagation steps for speech and noise as seen in Fig. 3.5. For the estimation of the current speech and noise DFT coefficients $S(\lambda,\mu)$ and $N(\lambda,\mu)$ within this step, the temporal trajectories of each frequency bin are separately considered for speech as well as for noise. An example for speech is shown in Fig. 3.6 where the temporal trajectory of *one* complex-valued speech DFT coefficient is depicted for a short time segment. It can clearly be seen that successive DFT coefficients are correlated over time. In contrast to conventional time domain Kalman filters where the correlation of consecutive samples is exploited (e.g., [PB87]), the proposed solution in this thesis utilizes the *inter-frame* correlation of speech and noise in the frequency domain.

In analogy to the AR modeling of speech and noise in the time domain, cf. Sec. 3.1, the current speech and noise DFT coefficients are divided into a *prediction part*, i.e., a function depending on DFT coefficients from the past, and an *innovation part* consisting of uncorrelated process noise as follows:

$$S(\lambda,\mu) = f\big(\, S(\lambda - N_{\text{K}},\mu),\, \ldots\, , S(\lambda-1,\mu)\,\big) + E_S(\lambda,\mu) \quad \text{and} \tag{3.51}$$

$$N(\lambda,\mu) = f\big(\, N(\lambda - M_{\text{K}},\mu),\, \ldots\, , N(\lambda-1,\mu)\,\big) + E_N(\lambda,\mu), \tag{3.52}$$

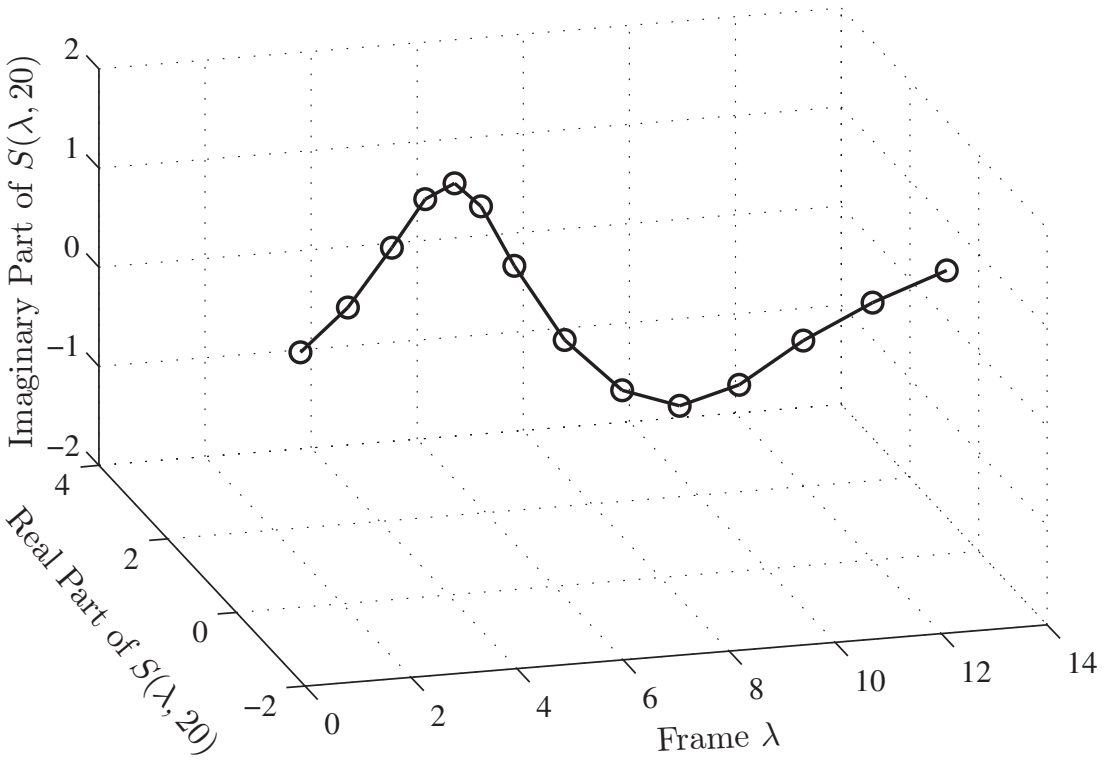

Figure 3.6: Temporal trajectory of one complex-valued speech DFT coefficient for a short speech segment. The example is taken from the NTT speech database [NC94] and corresponds to the word 'Help' (male voice) at frequency bin $\mu = 20$ using an FFT size $M_{\mathrm{F}} = 256$, a frame shift $L_{\mathrm{FS}} = 40$ and a sampling frequency $f_{\mathrm{s}} = 8\,\mathrm{kHz}$.

where $E_S(\lambda, \mu)$ and $E_N(\lambda, \mu)$ represent the 'excitation' DFT coefficients of speech and noise, respectively.

As mentioned before, the authors in [ZVY06b] propose a system that depends on two separate Kalman filters for real and imaginary parts, i.e., one Kalman filter including propagation *and* update step for speech and noise is applied to the real parts and the other one to the imaginary parts of the noisy input DFT coefficients. Eventually, both estimates are combined to obtain the desired complex-valued estimation of the speech DFT coefficients. In contrast to [ZVY06b], the proposed system already requires one *complex-valued* prediction after each propagation step. For this purpose, the following three alternative methods are investigated to perform the estimation in the propagation step based on the enhanced coefficients $\hat{S}_{\mathrm{up}}$ and $\hat{N}_{\mathrm{up}}$ from previous frames, cf. Fig. 3.5:

1. Prediction of magnitudes and phases separately,
2. Prediction of real and imaginary parts separately,
3. Prediction using complex-valued coefficients.

In the sequel, the evaluation is carried out individually for the speech and the noise signal. Therefore, the prediction gain, which can be achieved in the propagation step, is taken as quality measure.

3.2.2.1 Complex-Valued Prediction of Speech Signals

The *first method* uses two separate predictors, the first one for the current speech magnitudes $A(\lambda, \mu)$ and the second one for the current speech phase coefficients $\phi(\lambda, \mu)$ according to:

$$\hat{A}_{\text{prop}}(\lambda, \mu) = \sum_{\kappa=1}^{N_{\text{K}}} \hat{a}_{\kappa,\text{m}}(\lambda, \mu) \left| \hat{S}_{\text{up}}(\lambda - \kappa, \mu) \right| \tag{3.53}$$

$$\hat{\phi}_{\text{prop}}(\lambda, \mu) = \sum_{\kappa=1}^{N_{\text{K}}} \hat{a}_{\kappa,\text{p}}(\lambda, \mu) \angle \left\{ \hat{S}_{\text{up}}(\lambda - \kappa, \mu) \right\}, \tag{3.54}$$

where $|\cdot|$ represents the magnitude operator, $\angle\{\cdot\}$ the 'unwrapped' phase operator and a_κ the κ-th AR coefficients which have to be estimated in advance. Both predictions of this method are combined in order to obtain the complex-valued estimates of the current speech DFT coefficients:

$$\hat{S}^{(1)}_{\text{prop}}(\lambda, \mu) = \hat{A}_{\text{prop}}(\lambda, \mu) \cdot \exp\left(\text{j} \cdot \hat{\phi}_{\text{prop}}(\lambda, \mu) \right), \tag{3.55}$$

where $\exp(\cdot)$ denotes the exponential function and j the imaginary unit.

For the *second method*, a similar procedure is carried out relying on separate predictors for the real and the imaginary parts of the DFT coefficients[2]:

$$\hat{S}^{(2)}_{\text{prop}}(\lambda, \mu) = \sum_{\kappa=1}^{N_{\text{K}}} \hat{a}_{\kappa,\text{Re}}(\lambda, \mu) \text{Re}\left\{ \hat{S}_{\text{up}}(\lambda - \kappa, \mu) \right\} + \text{j} \sum_{\kappa=1}^{N_{\text{K}}} \hat{a}_{\kappa,\text{Im}}(\lambda, \mu) \text{Im}\left\{ \hat{S}_{\text{up}}(\lambda - \kappa, \mu) \right\}, \tag{3.56}$$

with $\text{Re}\{\cdot\}$ and $\text{Im}\{\cdot\}$ denoting real and imaginary parts.

In the *third method*, one complex-valued predictor estimates the temporal trajectory of the speech signal in the frequency domain. Therefore, complex-valued prediction coefficients $\hat{a}_\kappa$ are directly applied to the complex-valued estimates $\hat{S}_{\text{up}}$ from the previous frames:

$$\hat{S}^{(3)}_{\text{prop}}(\lambda, \mu) = \sum_{\kappa=1}^{N_{\text{K}}} \hat{a}_\kappa(\lambda, \mu) \hat{S}_{\text{up}}(\lambda - \kappa, \mu). \tag{3.57}$$

Determination of the prediction coefficients

As mentioned above, all prediction coefficients have to be estimated in advance in each frame for each frequency bin before the current speech DFT coefficients can be

[2] Please note that this second method differs from the propagation step which is applied in [ZVY06b]. In [ZVY06b], estimates of real and imaginary parts are combined not until the update step.

predicted. For the computation of the prediction coefficients, the minimization of the prediction error energy is used as optimization criterion for all three methods, i.e.:

$$\left|S(\lambda,\mu)-\hat{S}_{\text{prop}}^{(\zeta)}\right|^2 \rightarrow \text{ min.},$$

with $\zeta \in \{1,2,3\}$ indicating the prediction method. The real-valued prediction coefficients in Eqs. 3.53, 3.54 and 3.56 and also the complex-valued prediction coefficients in Eq. 3.57 (see Appendix A.3 for details) can be obtained by solving the corresponding Yule-Walker equations [PM96] using, e.g., the well-known Levinson-Durbin algorithm [MGj76], which can be applied to complex-valued input parameters as well, e.g., [Hay96]. The autocorrelation vectors and matrices which are required for this algorithm have to be known a priori or estimated in advance.

Evaluation

In order to investigate which one of the aforementioned methods performs best, the prediction gains $G_{\text{P},S}^{(\zeta)}$ are measured for each method as follows:

$$G_{\text{P},S}^{(\zeta)} = \frac{\mathbb{E}\left\{|S(\lambda,\mu)|^2\right\}}{\mathbb{E}\left\{\left|S(\lambda,\mu)-\hat{S}_{\text{prop}}^{(\zeta)}(\lambda,\mu)\right|^2\right\}}. \tag{3.58}$$

In the investigation, idealistic conditions are assumed. The predictions in Eqs. 3.55, 3.56 and 3.57 are based on clean speech DFT coefficients, i.e., the estimates $\hat{S}_{\text{up}}(\lambda-\kappa,\mu)$ are replaced by $S(\lambda-\kappa,\mu)$ and the prediction coefficients are determined from the previous clean speech DFT coefficients. Therefore, the Levinson-Durbin algorithm is applied as described above[3] using the most recent L_{AC} speech DFT coefficients to estimate the autocorrelation function. The results are averaged over time and frequency by incorporating only frequency bins contributing to speech activity[4].

The investigation is carried out at a sampling frequency $f_{\text{s}} = 8\,\text{kHz}$, the frame size is set to 20 ms ($L_{\text{F}} = 160$) and an FFT length $M_{\text{F}} = 256$ is used. The data is obtained from about 30 minutes of speech randomly selected from the NTT speech database [NC94]. As speech data of finite length is used, the measured prediction gain $\hat{G}_{\text{P},S}^{(\zeta)}$ states only an estimate of $G_{\text{P},S}^{(\zeta)}$ as defined in Eq. 3.58. In order to determine appropriate values for the frame shift L_{FS}, the autocorrelation function length L_{AC} as well as the model order N_{K}, different settings are investigated. The frame shift is set to the following shift sizes which are commonly used in speech enhancement applications [Ben07]:

- L_{FS}=40 ($\hat{=}$ 5 ms corresponding to 75% frame overlap),
- L_{FS}=60 ($\hat{=}$ 7.5 ms corresponding to 62.5% frame overlap) or

[3] Idealistic conditions are assumed only for the evaluation. In the final implementation, $\hat{S}_{\text{prop}}^{(\zeta)}$ is determined from entities which are available in the real system.

[4] A simple power constrained threshold is applied to the clean speech signal for VAD.

- L_{FS}=80 ($\hat{=}$ 10 ms corresponding to 50% frame overlap).

In addition, the length of the autocorrelation function is varied between $2 \leq L_{\mathrm{AC}} \leq 8$ and the model order is adapted depending on L_{AC} in the range $1 \leq N_{\mathrm{K}} \leq L_{\mathrm{AC}}$.

Figure 3.7 depicts the maximum prediction gains $\hat{G}^{(\zeta)}_{\mathrm{P},S,\max}(L_{\mathrm{AC}})$ for the three complex-valued prediction methods plotted over the autocorrelation function length L_{AC} for the different frame shift sizes. The entity $\hat{G}^{(\zeta)}_{\mathrm{P},S,\max}(L_{\mathrm{AC}})$ is defined as the maximum prediction gain which can be achieved for a given autocorrelation function length L_{AC} while varying the model order N_{K} according to:

$$\hat{G}^{(\zeta)}_{\mathrm{P},S,\max}(L_{\mathrm{AC}}) = \max_{N_{\mathrm{K}}=1,\,\ldots,\,L_{\mathrm{AC}}} \hat{G}^{(\zeta)}_{\mathrm{P},S}(L_{\mathrm{AC}}, N_{\mathrm{K}}), \tag{3.59}$$

where $\hat{G}^{(\zeta)}_{\mathrm{P},S}(L_{\mathrm{AC}}, N_{\mathrm{K}})$ illustrates the prediction gain $\hat{G}^{(\zeta)}_{\mathrm{P},S}$ using L_{AC} and N_{K}.

The results show that the highest prediction gains are obtained by using complex-valued prediction coefficients for all three frame shift sizes. This prediction method consistently outperforms the other two approaches where real and imaginary parts and magnitudes and phases are predicted separately. In the latter approach, even negative values are achieved for the prediction gain which is due to the fact that there is only little temporal correlation in successive phase coefficients. Figure 3.7 also illustrates that the prediction gain is depending on the autocorrelation length L_{AC} as well as on the frame shift size L_{FS}. With increasing autocorrelation length, the prediction gains of the Methods (2) and (3) are increasing at first and finally decreasing again at higher values of L_{AC}. For Method (1), the opposite effect can be recognized. In contrast to common LPC techniques used, e.g., for speech coding, the current clean speech DFT coefficients or estimates of them are not available in the final system and therefore not incorporated to determine the autocorrelation function. As expected, the prediction gain becomes higher for smaller frame shift sizes L_{FS}, i.e., for a larger frame overlap. If L_{FS} is chosen to be too large, the correlation between DFT coefficients in consecutive frames can not be exploited any more due to the relative short period of time in which a speech signal can be assumed to be stationary. Shorter frame shift sizes are not considered in this work in order to limit the computational cost of the system. Considering all curves in Fig. 3.7, the highest prediction gain is achieved using complex-valued prediction coefficients, a frame shift size $L_{\mathrm{FS}} = 40$ corresponding to 75% frame overlap, an autocorrelation function length $L_{\mathrm{AC}} = 6$ and a speech model order $N_{\mathrm{K}} = 3$. These parameters are used in the final system for the prediction of the current speech DFT coefficients in the propagation step according to Eq. 3.57.

3.2.2.2 Complex-Valued Prediction of Noise Signals

Based on the results of the previous section, the investigations for the noise signal are directly limited to a frame shift size $L_{\mathrm{FS}} = 40$ as well as on the application of

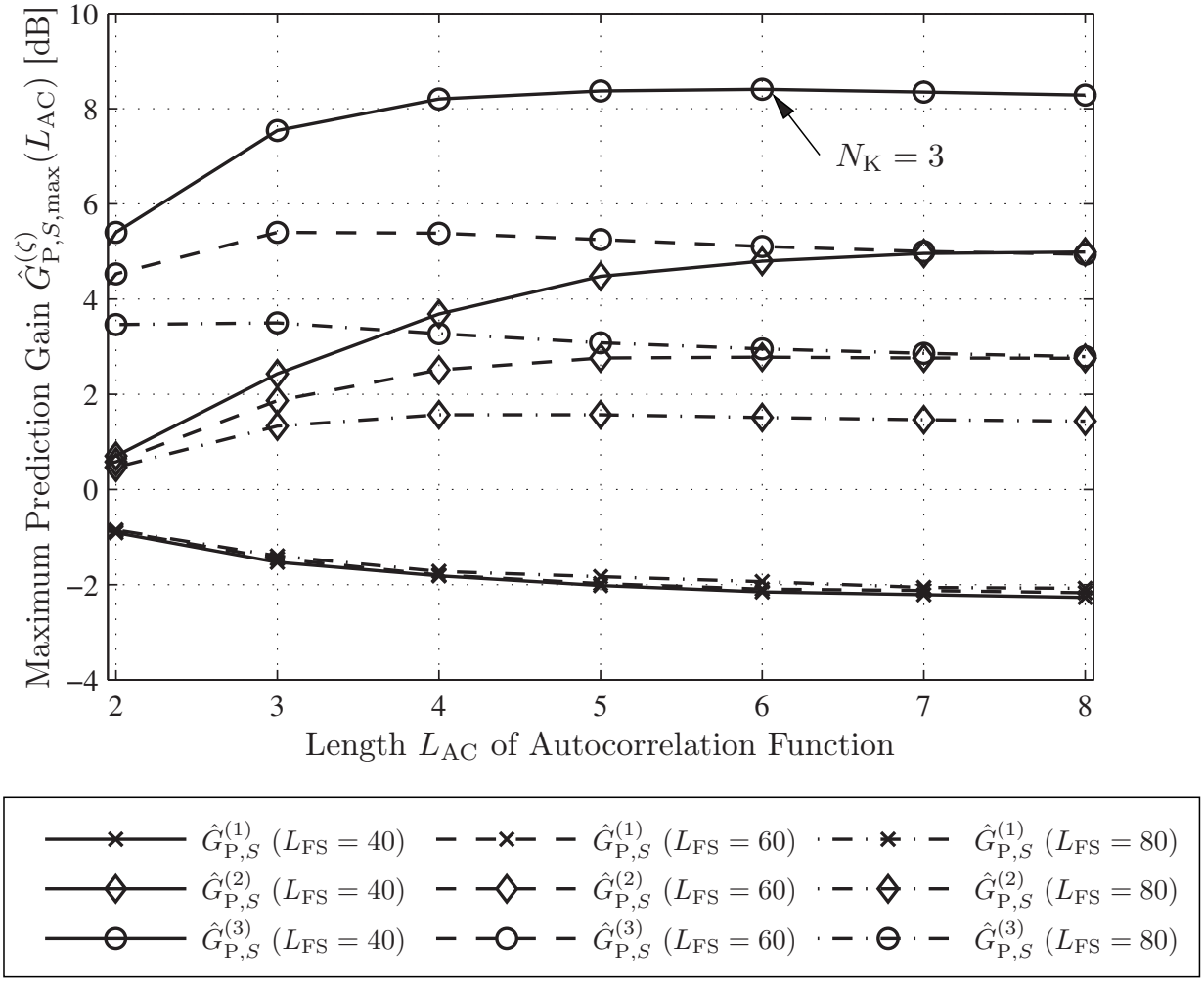

Figure 3.7: Maximum speech prediction gains for the three complex-valued prediction methods: (1) separate prediction of magnitudes and phases, (2) separate prediction of real and imaginary parts, (3) prediction using complex-valued prediction coefficients, depending on autocorrelation function length L_{AC}, model order N_{K} and frame shift L_{FS}. For a given length L_{AC}, the prediction gains for all model orders in the range $1 \leq N_{\mathrm{K}} \leq L_{\mathrm{AC}}$ are measured but only the highest prediction gain out of this set is depicted in the figure in each case. Overall, the highest prediction gain is achieved by complex-valued linear prediction using $L_{\mathrm{AC}} = 6$, $N_{\mathrm{K}} = 3$ and $L_{\mathrm{FS}} = 40$.

complex-valued prediction coefficients $b_\tau(\lambda, \mu)$ for the prediction of the current noise DFT coefficients according to:

$$\hat{N}_{\mathrm{prop}}(\lambda, \mu) = \sum_{\tau=1}^{M_{\mathrm{K}}} \hat{b}_\tau(\lambda, \mu) \hat{N}_{\mathrm{up}}(\lambda - \tau, \mu). \tag{3.60}$$

In order to find out adequate settings for the noise model order M_{K} and the autocorrelation function length L'_{AC} for the noise signal, a similar procedure as for the speech signal is applied. The investigations are carried out under the same idealistic conditions. The previous M_{K} true noise DFT coefficients $N(\lambda - M_{\mathrm{K}}, \mu)$, ... , $N(\lambda - 1, \mu)$

are applied to predict the current noise DFT coefficients $N(\lambda, \mu)$ using prediction coefficients which are determined from the previous L'_{AC} true noise DFT coefficients. Based on the prediction gain:

$$G_{\mathrm{P},N} = \frac{\mathbb{E}\left\{|N(\lambda,\mu)|^2\right\}}{\mathbb{E}\left\{\left|N(\lambda,\mu) - \hat{N}_{\mathrm{prop}}(\lambda,\mu)\right|^2\right\}}, \tag{3.61}$$

different combinations of L'_{AC} and M_{K} are evaluated. Figure 3.8 shows the maximum noise prediction gains $\hat{G}_{\mathrm{P},N,\max}(L'_{\mathrm{AC}})$ which are achieved for a given length L'_{AC} of the autocorrelation function when the model order M_{K} is altered between 1 and L'_{AC}, cf. Eq. 3.59. The figure depicts the results for four different types of noise signals taken from the NOISEX-92 database [VS93]: babble noise, f16 noise, factory noise and WGN. The highest prediction gain (equivalent to the highest temporal correlation) is achieved for 'babble noise', the lowest for WGN. Please note that the frame overlap leads to an oversampling such that temporal correlation as well as a non-zero prediction gain is also possible for WGN. The stationarity of the considered noise signals over time is greater than that of speech signals. Therefore, it would be reasonable to choose a larger length L'_{AC} for the noise autocorrelation function as evident from the continuously increasing predicting gains in Fig. 3.8 beyond $L'_{\mathrm{AC}} = 8$. However, the model orders M_{K} which yield the corresponding maximum prediction gains are lower than that for the speech signals and are all in the range $M_{\mathrm{K}} \in \{1, 2\}$ indicating a shorter correlation time compared to the speech signals. With regard to the computational complexity, the autocorrelation function lengths of speech and noise are chosen to be equivalent ($L'_{\mathrm{AC}} = L_{\mathrm{AC}} = 6$) accepting a small loss in prediction performance for the noise signal. The noise model order is set to $M_{\mathrm{K}} = 2$ in the following and the sizes L_{AC} and L'_{AC} are used interchangeably from now on.

3.2.2.3 Summary and Remarks

Summarizing the previous two subsections, the application of complex-valued prediction coefficients in the propagation step was shown to be beneficial in comparison to separately predicting magnitudes and phases or real and imaginary parts. Moreover, a frame shift of 5 ms (corresponding to a frame overlap of 75%) for the considered analysis-synthesis system results in a good compromise between good prediction properties and reasonable computational complexity. The required prediction coefficients for speech and noise are determined using the Levinson-Durbin algorithm. Therefore, the previous $L_{\mathrm{AC}} = 6$ *enhanced* DFT coefficients $\hat{S}_{\mathrm{up}}$ and $\hat{N}_{\mathrm{up}}$ will be applied in the final system to estimate autocorrelation vectors and matrices of speech and noise, respectively. Based on the investigations under idealistic conditions, the model orders for speech and noise are set to $N_{\mathrm{K}} = 3$ and $M_{\mathrm{K}} = 2$. Please note that the estimation of the prediction coefficients given a noisy signal is not the main focus of this thesis. More sophisticated techniques can be found in literature, see, e.g., the list at the beginning of Sec. 3.2.

The prediction errors made in the propagation step when predicting speech and noise DFT coefficients are defined as $\mathbf{E}^S_{\mathrm{prop}}(\lambda, \mu)$ and $\mathbf{E}^N_{\mathrm{prop}}(\lambda, \mu)$, cf. Eqs. 3.44

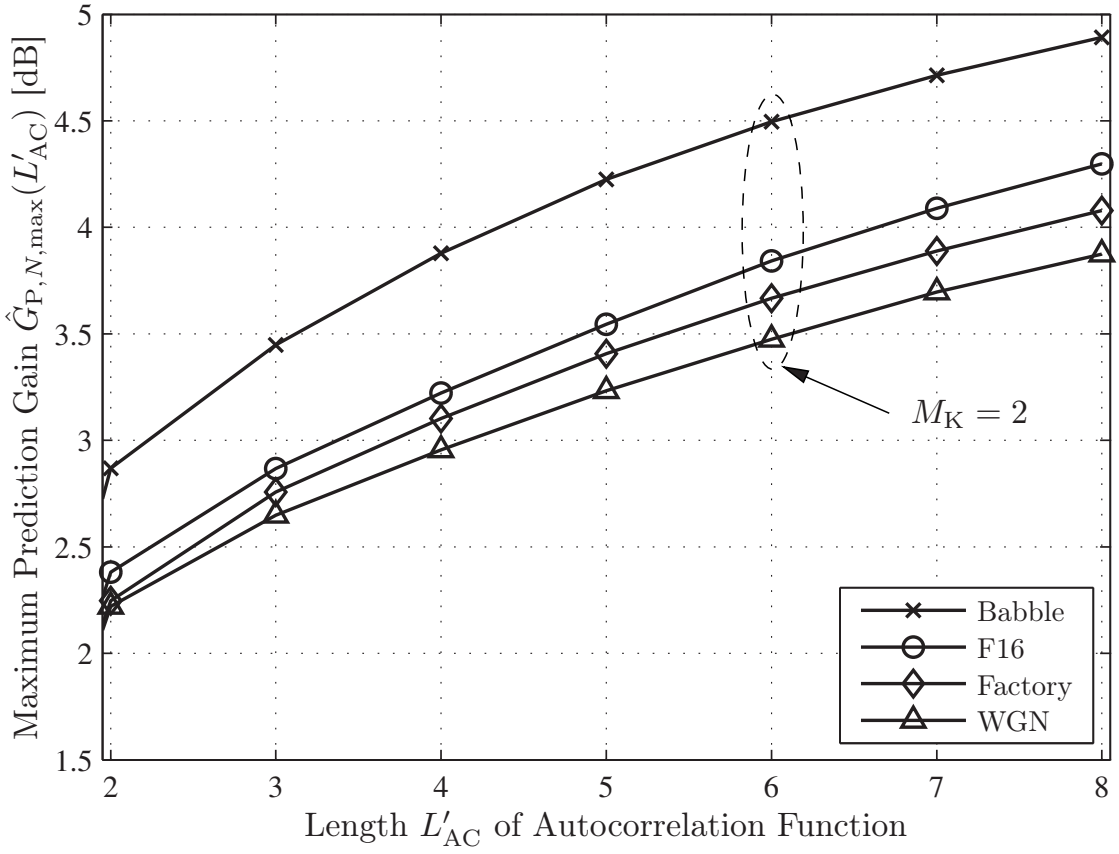

Figure 3.8: Maximum noise prediction gains depending on autocorrelation function length L'_{AC} and model order M_{K} using complex-valued prediction coefficients and a frame shift size $L_{\mathrm{FS}} = 40$. For a given length L'_{AC}, the prediction gains for all model orders in the range $1 \leq M_{\mathrm{K}} \leq L'_{\mathrm{AC}}$ are measured but only the highest prediction gain out of this set is depicted in the figure in each case. Setting $L'_{\mathrm{AC}} = 6$, the respective highest prediction gain is achieved for $M_{\mathrm{K}} = 2$ independent of the noise type.

and 3.45. Similar to Sec. 3.1.1, an expression for the error covariance matrices $\mathbf{P}^S_{\mathrm{prop}}(\lambda,\mu)$ and $\mathbf{P}^N_{\mathrm{prop}}(\lambda,\mu)$ of the propagation step can be derived from these entities in the frequency domain as well resulting in:

$$\begin{aligned}\mathbf{P}^S_{\mathrm{prop}}(\lambda,\mu) &= \mathbb{E}\{\mathbf{E}^S_{\mathrm{prop}}(\lambda,\mu)\left(\mathbf{E}^S_{\mathrm{prop}}(\lambda,\mu)\right)^H\} \\ &= \mathbf{A}(\lambda,\mu)\mathbf{P}^S_{\mathrm{up}}(\lambda-1,\mu)\mathbf{A}(\lambda,\mu)^H + \mathbf{g}_S\sigma^2_{E_S}(\lambda,\mu)\mathbf{g}_S^H \qquad (3.62)\\ \mathbf{P}^N_{\mathrm{prop}}(\lambda,\mu) &= \mathbb{E}\{\mathbf{E}^N_{\mathrm{prop}}(\lambda,\mu)\left(\mathbf{E}^N_{\mathrm{prop}}(\lambda,\mu)\right)^H\} \\ &= \mathbf{B}(\lambda,\mu)\mathbf{P}^N_{\mathrm{up}}(\lambda-1,\mu)\mathbf{B}(\lambda,\mu)^H + \mathbf{g}_N\sigma^2_{E_N}(\lambda,\mu)\mathbf{g}_N^H, \qquad (3.63)\end{aligned}$$

where $\mathbf{A}(\lambda,\mu)$ and $\mathbf{B}(\lambda,\mu)$ denote transition matrices of speech and noise including the respective prediction coefficients. $\mathbf{P}^S_{\mathrm{up}}(\lambda-1,\mu)$ and $\mathbf{P}^N_{\mathrm{up}}(\lambda-1,\mu)$ state the covariance matrices of the update step from the previous frame, $\sigma^2_{E_S}(\lambda,\mu)$ and $\sigma^2_{E_S}(\lambda,\mu)$ the powers of $E_S(\lambda,\mu)$ and $E_N(\lambda,\mu)$ and $\mathbf{g}_S$ and $\mathbf{g}_N$ the equivalents to $\mathbf{g}_s$ and $\mathbf{g}_n$ in the time domain. Estimates of the error covariance matrices $\mathbf{P}^S_{\mathrm{prop}}(\lambda,\mu)$ and $\mathbf{P}^N_{\mathrm{prop}}(\lambda,\mu)$ are required in the following update step.

3.2.3 Update Step

While temporal correlation of successive speech and noise DFT coefficients is exploited in the propagation step, the update step utilizes the statistical characteristics of both signals. The objective in this step is to estimate the two prediction error vectors $\mathbf{E}^S_{\text{prop}}(\lambda,\mu)$ and $\mathbf{E}^N_{\text{prop}}(\lambda,\mu)$ arising from the propagation step. Therefore, the differential signal $D(\lambda,\mu)$ as well as the error covariance matrices $\mathbf{P}^S_{\text{prop}}(\lambda,\mu)$ and $\mathbf{P}^N_{\text{prop}}(\lambda,\mu)$ are available in the current frame. Based on these parameters, the conditional expectation vectors $\mathbb{E}\{\mathbf{E}^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ and $\mathbb{E}\{\mathbf{E}^N_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ yield the required estimates of the prediction error vectors in the update step and finally update the initial predictions of the propagation step according to Eqs. 3.49 and 3.50.

In order to estimate the aforementioned expectation vectors, the differential signal can be decomposed into the sum of the two prediction errors $E^S_{\text{prop}}(\lambda,\mu)$ and $E^N_{\text{prop}}(\lambda,\mu)$ of the current frame, cf. Sec. 3.1.1.3:

$$\begin{aligned} D(\lambda,\mu) &= Y(\lambda,\mu) - \hat{Y}_{\text{prop}}(\lambda,\mu), \\ &= Y(\lambda,\mu) - \mathbf{h}_S^T\hat{\mathbf{S}}_{\text{prop}}(\lambda,\mu) - \mathbf{h}_N^T\hat{\mathbf{N}}_{\text{prop}}(\lambda,\mu), \\ &= S(\lambda,\mu) - \mathbf{h}_S^T\hat{\mathbf{S}}_{\text{prop}}(\lambda,\mu) + N(\lambda,\mu) - \mathbf{h}_N^T\hat{\mathbf{N}}_{\text{prop}}(\lambda,\mu), \\ &= E^S_{\text{prop}}(\lambda,\mu) + E^N_{\text{prop}}(\lambda,\mu). \end{aligned} \tag{3.64}$$

As seen before, the estimation problem in the update step therewith becomes a 'classical' noise reduction problem: The 'target' signal $E^S_{\text{prop}}(\lambda,\mu)$ is degraded by the additive 'noise' signal $E^N_{\text{prop}}(\lambda,\mu)$. Given the 'noisy' signal $D(\lambda,\mu)$, a conventional statistical estimator can be applied in order to estimate the prediction errors $E^S_{\text{prop}}(\lambda,\mu)$ and $E^N_{\text{prop}}(\lambda,\mu)$ which can afterwards be used to determine the required prediction error vectors $\mathbf{E}^S_{\text{prop}}(\lambda,\mu)$ and $\mathbf{E}^N_{\text{prop}}(\lambda,\mu)$.

In the following, at first both conditional expectation vectors are derived under the assumption that the prediction errors of speech *and* noise follow a Gaussian distribution as assumed in the original Kalman filter approach. Afterwards, it is shown that the statistics of the speech prediction error rather follow a super-Gaussian model which is influenced by the input SNR. In order to exploit this fact, an adequate SNR-dependent statistical weighting rule is proposed. Finally, the procedure for estimating the required prediction error powers is presented before a brief summary closes this section.

3.2.3.1 Gaussian Model

The conditional expectation vectors $\mathbb{E}\{\mathbf{E}^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ and $\mathbb{E}\{\mathbf{E}^N_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ are determined in this subsection assuming multivariate Gaussian distributions for both vectors. Therefore, at first an expression for $\mathbb{E}\{E^S_{\text{prop}}|D(\lambda,\mu)\}$ is derived based on the assumption that real and imaginary parts

of all DFT coefficients are statistically independent. In this case, the MMSE[5] estimate separates into two independent estimators for the real and the imaginary parts as shown in the following:

$$\begin{aligned}\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\} =&\mathbb{E}\{\text{Re}\{E^S_{\text{prop}}(\lambda,\mu)\}|\text{Re}\{D(\lambda,\mu)\}\}\\ &+\text{j}\cdot\mathbb{E}\{\text{Im}\{E^S_{\text{prop}}(\lambda,\mu)\}|\text{Im}\{D(\lambda,\mu)\}\}.\end{aligned} \tag{3.65}$$

Considering the real part first and using *Bayes' theorem* [Bay63], the conditional PDF $p(\text{Re}\{E^S_{\text{prop}}(\lambda,\mu)\}|\text{Re}\{D(\lambda,\mu)\})$ can be stated as a function dependent on the PDFs $p(\text{Re}\{D(\lambda,\mu)\})$, $p(\text{Re}\{E^S_{\text{prop}}(\lambda,\mu)\})$ and $p(\text{Re}\{D(\lambda,\mu)\}|\text{Re}\{E^S_{\text{prop}}(\lambda,\mu)\})$. Using the abbreviated forms $D_{\text{R}}(\lambda,\mu)=\text{Re}\{D(\lambda,\mu)\}$, $E^S_{\text{R,prop}}(\lambda,\mu)=\text{Re}\{E^S_{\text{prop}}(\lambda,\mu)\}$ and $E^N_{\text{R,prop}}(\lambda,\mu)=\text{Re}\{E^N_{\text{prop}}(\lambda,\mu)\}$, the respective PDFs are given as follows for the Gaussian case:

$$p(D_{\text{R}}(\lambda,\mu))=\frac{1}{\sqrt{2\pi\sigma^2_{D_{\text{R}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}}\exp\left(-\frac{(D_{\text{R}}(\lambda,\mu))^2}{2\sigma^2_{D_{\text{R}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\right), \tag{3.66}$$

$$p(E^S_{\text{R,prop}}(\lambda,\mu))=\frac{1}{\sqrt{2\pi\sigma^2_{E^S_{\text{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}}\exp\left(-\frac{\left(E^S_{\text{R,prop}}(\lambda,\mu)\right)^2}{2\sigma^2_{E^S_{\text{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\right)\quad\text{and} \tag{3.67}$$

$$p(D_{\text{R}}(\lambda,\mu)|E^S_{\text{R,prop}}(\lambda,\mu))=\frac{1}{\sqrt{2\pi\sigma^2_{E^N_{\text{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}}\exp\left(-\frac{\left(D_{\text{R}}(\lambda,\mu)-E^S_{\text{R,prop}}(\lambda,\mu)\right)^2}{2\sigma^2_{E^N_{\text{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\right), \tag{3.68}$$

where $\sigma^2_{D_{\text{R}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)=\mathbb{E}\{|D_{\text{R}}(\lambda,\mu)|^2\}$, $\sigma^2_{E^S_{\text{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)=\mathbb{E}\{|E^S_{\text{R,prop}}(\lambda,\mu)|^2\}$ and $\sigma^2_{E^N_{\text{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)=\mathbb{E}\{|E^N_{\text{R,prop}}(\lambda,\mu)|^2\}$. Based on these PDFs, the real part of Eq. 3.65

[5] In order to form an estimate $\hat{a}$ for a parameter a by using the disturbed observation b, conditional estimation can be applied. Therefore, the integral over the joint PDF of the undisturbed and disturbed value is minimized finally leading to $\mathbb{E}\{C(a,\hat{a})|b\}=\int_{-\infty}^{\infty}C(a,\hat{a})\cdot p(a|b)\text{d}a$ with the cost function $C(\cdot)$. If the cost function is chosen to be square, i.e., $C(a,\hat{a})=(a-\hat{a})^2$, the conditional mean estimator equals the MMSE solution, e.g., [Var08].

results in (cf. Appendix A.2):

$$\begin{aligned}\mathbb{E}\{E^S_{\mathrm{R,prop}}(\lambda,\mu)|D_{\mathrm{R}}(\lambda,\mu)\} &= \int_{-\infty}^{\infty} E^S_{\mathrm{R,prop}}(\lambda,\mu) \\ &\qquad \cdot p(E^S_{\mathrm{R,prop}}(\lambda,\mu)|D_{\mathrm{R}}(\lambda,\mu))\,\mathrm{d}E^S_{\mathrm{R,prop}}(\lambda,\mu) \\ &= \int_{-\infty}^{\infty} E^S_{\mathrm{R,prop}}(\lambda,\mu)\cdot\frac{p(D_{\mathrm{R}}(\lambda,\mu)|E^S_{\mathrm{R,prop}}(\lambda,\mu))}{p(D_{\mathrm{R}}(\lambda,\mu))} \\ &\qquad \cdot p(E^S_{\mathrm{R,prop}}(\lambda,\mu))\,\mathrm{d}E^S_{\mathrm{R,prop}}(\lambda,\mu) \\ &= \frac{\sigma^2_{E^S_{\mathrm{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{D_{\mathrm{R}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\cdot D_{\mathrm{R}}(\lambda,\mu) \\ &= \frac{\sigma^2_{E^S_{\mathrm{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{E^S_{\mathrm{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)+\sigma^2_{E^N_{\mathrm{R,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\cdot D_{\mathrm{R}}(\lambda,\mu).\end{aligned} \tag{3.69}$$

The expression for the imaginary part can be derived analogously and yields:

$$\mathbb{E}\{E^S_{\mathrm{I,prop}}(\lambda,\mu)|D_{\mathrm{I}}(\lambda,\mu)\} = \frac{\sigma^2_{E^S_{\mathrm{I,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{E^S_{\mathrm{I,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)+\sigma^2_{E^N_{\mathrm{I,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\cdot D_{\mathrm{I}}(\lambda,\mu), \tag{3.70}$$

where the index $(\cdot)_{\mathrm{I}}$ is used to denote the imaginary part and $\sigma^2_{E^S_{\mathrm{I,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right) = \mathbb{E}\{|E^S_{\mathrm{I,prop}}(\lambda,\mu)|^2\}$ as well as $\sigma^2_{E^N_{\mathrm{I,prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right) = \mathbb{E}\{|E^N_{\mathrm{I,prop}}(\lambda,\mu)|^2\}$. Combined with the solution for the real part, the conditional expectation of Eq. 3.65 is finally given by:

$$\begin{aligned}\mathbb{E}\{E^S_{\mathrm{prop}}(\lambda,\mu)|D(\lambda,\mu)\} &= \frac{\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{D}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\cdot D(\lambda,\mu) \\ &= \frac{\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)+\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\cdot D(\lambda,\mu),\end{aligned} \tag{3.71}$$

with $\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\kappa_1+1\\ \lambda-\kappa_2+1\end{smallmatrix},\mu\right) = \mathbb{E}\{E^S_{\mathrm{prop}}(\lambda-\kappa_1+1,\mu)\left(E^S_{\mathrm{prop}}(\lambda-\kappa_2+1,\mu)\right)^*\}$, $\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\tau_1+1\\ \lambda-\tau_2+1\end{smallmatrix},\mu\right) = \mathbb{E}\{E^N_{\mathrm{prop}}(\lambda-\tau_1+1,\mu)\left(E^N_{\mathrm{prop}}(\lambda-\tau_2+1,\mu)\right)^*\}$ and $\sigma^2_{D}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right) = \mathbb{E}\{|D(\lambda,\mu)|^2\}$.

The result is identical to the well-known Wiener filter solution [LO79], see Sec. 2.5.1. In the derivation of Eq. 3.71, it is implicitly assumed that $E^S_{\mathrm{prop}}(\lambda,\mu)$ and $E^N_{\mathrm{prop}}(\lambda,\mu)$

are statistically independent, in particular that $\mathbb{E}\{|D(\lambda,\mu)|^2\} = \mathbb{E}\{|E^S_{\text{prop}}(\lambda,\mu)|^2\} + \mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\}$. If the noisy input signal can be decomposed perfectly and the prediction in the propagation step works ideally, the two prediction errors are equal to the 'excitation' DFT coefficients $E_S(\lambda,\mu)$ and $E_N(\lambda,\mu)$ and the assumption is correct, cf. Eqs. 3.51 and 3.52. In reality, the system is not perfect and E^S_{prop} as well as E^N_{prop} are both estimates from the same observation Y leading to a statistical dependency. However, the resulting error caused by this approximation is very small and can be neglected in the sequel as shown in Appendix B.

As known from conventional statistical estimators, the weighting rules can often be expressed by the *a posteriori* SNR γ_{K} and the *a priori* SNR ξ_{K} which are defined in the update step as follows:

$$\gamma_{\text{K}}(\lambda,\mu) = \frac{|D(\lambda,\mu)|^2}{\mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\}} = \frac{|D(\lambda,\mu)|^2}{\sigma^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)} \quad \text{and} \tag{3.72}$$

$$\xi_{\text{K}}(\lambda,\mu) = \frac{\mathbb{E}\{|E^S_{\text{prop}}(\lambda,\mu)|^2\}}{\mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\}} = \frac{\sigma^2_{E^S_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}. \tag{3.73}$$

Using the definitions in Eqs. 3.72 and 3.73, the expression in Eq. 3.71 can be rewritten as:

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\} = \frac{\xi_{\text{K}}(\lambda,\mu)}{\xi_{\text{K}}(\lambda,\mu)+1}\cdot D(\lambda,\mu). \tag{3.74}$$

Equation 3.71 can be used to obtain the missing conditional expectation values $\mathbb{E}\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|D(\lambda,\mu)\}$ $(1 \le \kappa \le N_{\text{K}})$ which are still required in the update step to determine $\mathbb{E}\{\mathbf{E}^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$. If $\mathbf{E}^S_{\text{prop}}(\lambda,\mu)$ follows a multivariate complex-Gaussian distribution of order N_{K}, the conditional expectation $\mathbb{E}\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|E^S_{\text{prop}}(\lambda,\mu)\}$ is given by [KBJ00]:

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|E^S_{\text{prop}}(\lambda,\mu)\} = \frac{\sigma^2_{E^S_{\text{prop}}}\left(\begin{smallmatrix}\lambda-\kappa+1\\ \lambda\end{smallmatrix},\mu\right)}{\sigma^2_{E^S_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\cdot E^S_{\text{prop}}(\lambda,\mu). \tag{3.75}$$

Thus, $\mathbb{E}\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|D(\lambda,\mu)\}$ results in[6]:

$$\begin{aligned}
\mathbb{E}\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|D(\lambda,\mu)\} &= ... \\
&\mathbb{E}\left\{\mathbb{E}\left\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|E^S_{\text{prop}}(\lambda,\mu),D(\lambda,\mu)\right\}|D(\lambda,\mu)\right\} \\
&= \mathbb{E}\left\{\left.\frac{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda-\kappa+1}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}E^S_{\text{prop}}(\lambda,\mu)\right|D(\lambda,\mu)\right\} \\
&= \frac{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda-\kappa+1}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\} \\
&= \frac{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda-\kappa+1}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}\frac{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)+\sigma^2_{E^N_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}\cdot D(\lambda,\mu) \\
&= \frac{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda-\kappa+1}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)+\sigma^2_{E^N_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}\cdot D(\lambda,\mu), \qquad (3.76)
\end{aligned}$$

which exactly matches the Kalman filter equations that have been derived in Sec. 3.1.1, cf. Eq. 3.25.

Following a very similar derivation based on the same assumptions for the noise signal, the noise prediction error vector $\mathbf{E}^N_{\text{prop}}(\lambda,\mu)$ can be estimated based on:

$$\mathbb{E}\{E^N_{\text{prop}}(\lambda-\tau+1,\mu)|D(\lambda,\mu)\} = \frac{\sigma^2_{E^N_{\text{prop}}}\binom{\lambda-\tau+1}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)+\sigma^2_{E^N_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}\cdot D(\lambda,\mu), \quad (3.77)$$

where $1 \leq \tau \leq M_{\text{K}}$. The first factors in Eqs. 3.76 and 3.77 constitute the Kalman filter gains $\mathbf{K}^S(\lambda,\mu)$ and $\mathbf{K}^N(\lambda,\mu)$ in the frequency domain. In matrix notation, they can be written in their known forms as:

$$\mathbf{K}^S(\lambda,\mu) = \mathbf{P}^S_{\text{prop}}(\lambda,\mu)\cdot\mathbf{h}_S\left(\mathbf{h}_S^T\mathbf{P}^S_{\text{prop}}(\lambda,\mu)\mathbf{h}_S+\sigma^2_{E^N_{\text{prop}}}\binom{\lambda}{\lambda},\mu)\right)^{-1} \quad (3.78)$$

$$\mathbf{K}^N(\lambda,\mu) = \mathbf{P}^N_{\text{prop}}(\lambda,\mu)\cdot\mathbf{h}_N\left(\mathbf{h}_N^T\mathbf{P}^N_{\text{prop}}(\lambda,\mu)\mathbf{h}_N+\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)\right)^{-1}. \quad (3.79)$$

and are used to update the speech and noise predictions in Eqs. 3.47-3.50. Following an analog derivation as presented in Sec. 3.1.1, the error covariance matrices of the update step are finally determined according to:

$$\mathbf{P}^S_{\text{up}}(\lambda,\mu) = \left(\mathbf{I}-\mathbf{K}^S(\lambda,\mu)\mathbf{h}_S^T\right)\mathbf{P}^S_{\text{prop}}(\lambda,\mu) \quad (3.80)$$

$$\mathbf{P}^N_{\text{up}}(\lambda,\mu) = \left(\mathbf{I}-\mathbf{K}^N(\lambda,\mu)\mathbf{h}_N^T\right)\mathbf{P}^N_{\text{prop}}(\lambda,\mu). \quad (3.81)$$

[6] Assuming a, b and c to be random variables, a general property of conditional expectation yields $\mathbb{E}\{a|b\} = \mathbb{E}\{\mathbb{E}\{a|b,c\}|b\}$ [Dur95].

In addition to the independence assumption of E^S_{prop} and E^N_{prop}, the above derivation of the spectral Kalman filter approach rests on two more assumptions for the two prediction errors:

1. Real and imaginary parts of the speech and noise *prediction error* coefficients E^S_{prop} and E^N_{prop} are each statistically independent resulting in separate estimators for real and imaginary parts, cf. Eq. 3.65, and

2. the *prediction errors* E^S_{prop} as well as E^N_{prop} are complex-Gaussian distributed leading to the Wiener filter weighting rules in the update step.

Both assumptions are analyzed in the following.

3.2.3.2 Generalized Gamma Model

The Kalman filter solutions, which have been proposed for speech enhancement in literature so far, assume Gaussian distributions for the prediction errors of speech and noise, regardless of whether they are implemented in the time domain [PB87, GKG91], in subbands [WC98, Pud02] or in the DFT domain [ZVY06b]. In this thesis, the statistics of the speech prediction error signal are explicitly taken into account showing that the distribution of E^S_{prop} follows a generalized Gamma model rather than a Gaussian model. As a novel feature, adapted statistical weighting rules are derived for the use within the update step in the following.

The statistical properties of the noise prediction error coefficients E^N_{prop} vary with different noise types. In order to keep the algorithm general and to not become dependent on a particular noise signal, it is still assumed in this section that the DFT coefficients E^N_{prop} follow a complex-Gaussian distribution and only the statistics of the speech prediction error signal E^S_{prop} are investigated in the sequel.

The assumption that real and imaginary parts of E^S_{prop} are statistically independent is evaluated in Fig. 3.9 where contour plots of measured histograms of $\text{Re}\{E^S_{\text{prop}}\}$ and $\text{Im}\{E^S_{\text{prop}}\}$ are illustrated. For the measurement, about 1.5 hours of speech taken from the NTT database [NC94] are processed by the proposed system under the Gaussian model presented in the previous Sec. 3.2.3.1. As the statistics of E^S_{prop} might be dependent on the input SNR, the speech signals are disturbed by additive WGN at SNR values varying between -25 dB and 35 dB (step size: 5 dB) and the results are averaged. Figure 3.9(a) shows the contours of the joint distribution $p\left(\text{Re}\{E^S_{\text{prop}}\}, \text{Im}\{E^S_{\text{prop}}\}\right)$ and Fig. 3.9(b) the contours for the product of the marginal distributions $p\left(\text{Re}\{E^S_{\text{prop}}\}\right) \cdot p\left(\text{Im}\{E^S_{\text{prop}}\}\right)$. It can be seen that both PDFs are different showing that real and imaginary parts of E^S_{prop} are statistically dependent. Similar investigations are carried out in [EHH08] and [HEH08] for real and imaginary parts of speech signals.

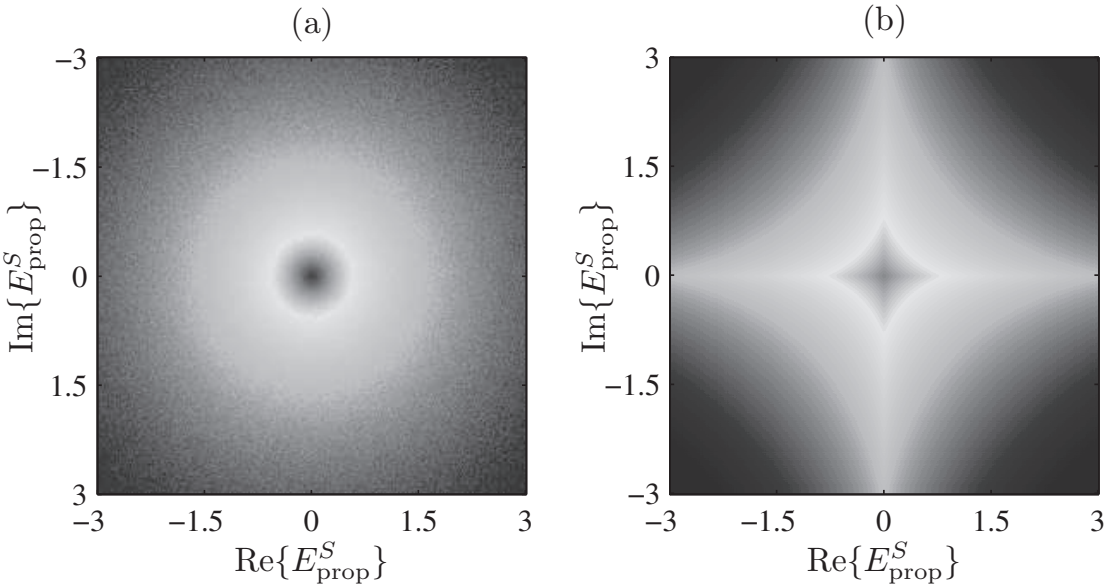

Figure 3.9: Contour lines of measured distributions of $\mathrm{Re}\{E^S_{\mathrm{prop}}\}$ and $\mathrm{Im}\{E^S_{\mathrm{prop}}\}$: (a) joint distribution, (b) product of marginal distributions.

In order to prove the assumption that the speech *prediction error* is complex-Gaussian distributed, Fig. 3.10 depicts measured histograms of the real part[7] of E^S_{prop} and of its amplitude using the same set of data as before. For comparison reasons, the figure also shows the PDFs of Gaussian, Rayleigh, Laplacian and Gamma distributions, all normalized to unit power. It can clearly be seen that the real part of E^S_{prop} is not Gaussian distributed as implicitly assumed by the original Kalman filter approach within the update step. This mismatch is equivalent with the fact that the amplitude histogram is not well fitted by a Rayleigh PDF. Instead, the statistics of the speech prediction error are better described by a super-Gaussian distribution like Laplace or Gamma. Similar investigations on the distribution of speech signals have been carried out in [LV05] and [GM10a].

In the following, a modified statistical estimator is derived for the use within the update step which is not based on the independence assumption of $\mathrm{Re}\{E^S_{\mathrm{prop}}\}$ and $\mathrm{Im}\{E^S_{\mathrm{prop}}\}$. Moreover, the estimator is adapted to the statistics measured in Fig. 3.10. As shown in Fig. 3.9, the joint PDF $p\left(\mathrm{Re}\{E^S_{\mathrm{prop}}\}, \mathrm{Im}\{E^S_{\mathrm{prop}}\}\right)$ is circularly symmetric meaning that the phase distribution of E^S_{prop} is uniform and independent from the amplitude distribution. This allows to state the joint PDF of amplitude and phase of the prediction error E^S_{prop} as:

$$p(|E^S_{\mathrm{prop}}|, \angle\{E^S_{\mathrm{prop}}\}) = \frac{p(|E^S_{\mathrm{prop}}|)}{2\pi}. \tag{3.82}$$

[7] The same distribution holds for the imaginary part.

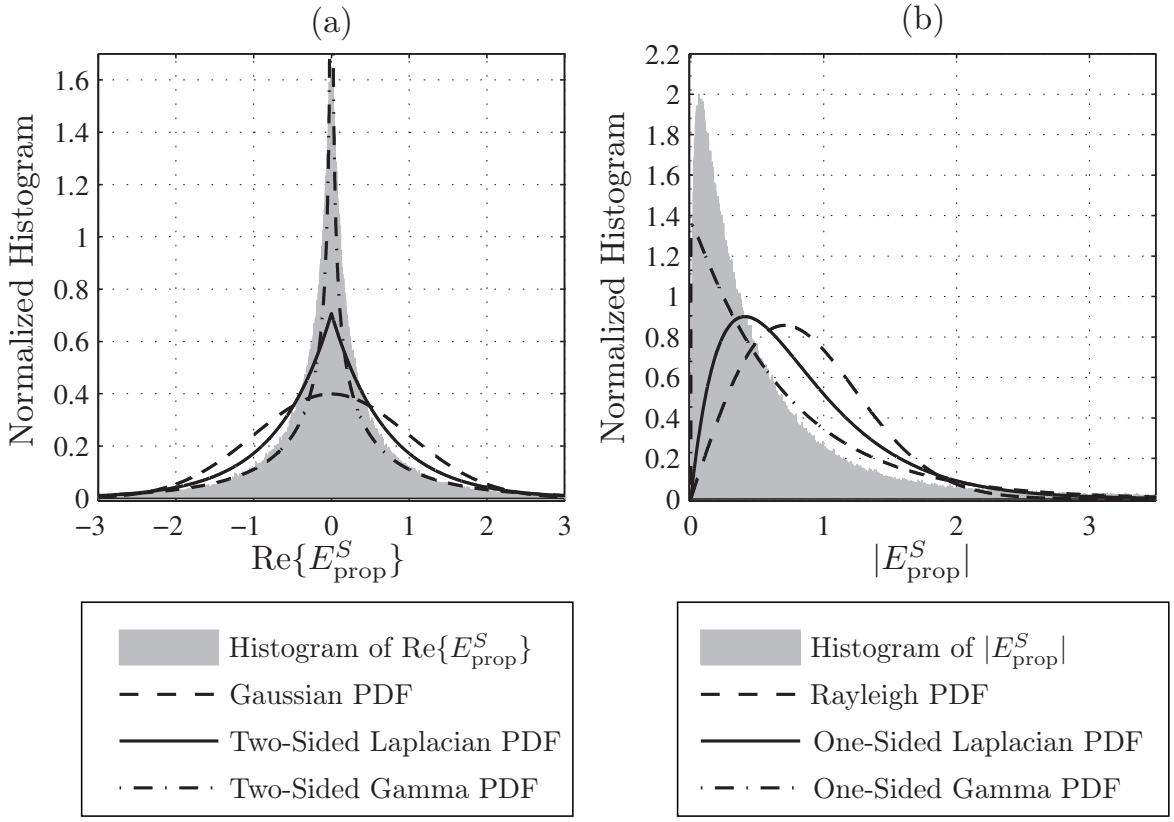

Figure 3.10: Histograms of speech prediction error E^S_{prop} normalized to unit power: (a) real part, (b) amplitude.

In the update step, the objective is to estimate the conditional expectation vectors $\mathbb{E}\{\mathbf{E}^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ and $\mathbb{E}\{\mathbf{E}^N_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$. Therefore, at first an expression for $\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ is again derived in the sequel under the new considerations. For this purpose, the conditional PDF $p(D(\lambda,\mu)|E^S_{\text{prop}}(\lambda,\mu))$ is required as well as an adequate PDF for the amplitude $|E^S_{\text{prop}}|$ which matches the measured histogram in Fig. 3.10(b). Still assuming the Gaussian model for the noise *prediction error*, the first PDF is given by:

$$p(D(\lambda,\mu)|E^S_{\text{prop}}(\lambda,\mu)) = \frac{1}{\pi\sigma^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)} \cdot \exp\left(-\frac{\left|D(\lambda,\mu) - E^S_{\text{prop}}(\lambda,\mu)\right|^2}{\sigma^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)}\right). \quad (3.83)$$

For the amplitude of the speech *prediction error* E^S_{prop}, the following single-sided generalized Gamma PDF is applied:

$$p(|E^S_{\text{prop}}(\lambda,\mu)|) = \frac{\theta\delta^\rho}{\Gamma(\rho)}|E^S_{\text{prop}}(\lambda,\mu)|^{(\theta\rho-1)} \cdot \exp\left(-\delta|E^S_{\text{prop}}(\lambda,\mu)|^\theta\right), \quad (3.84)$$

with $\delta > 0$, $\theta > 0$ and $\rho > 0$. $\Gamma(\cdot)$ represents the Gamma function and θ, δ and ρ are model parameters which can be adjusted to approximate the measured histogram. The parameter δ depends on θ, ρ and $\sigma^2_{E^S_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$. For $\theta = 1$,

$\delta = \sqrt{\rho(\rho+1)/\sigma_{E^S_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix}, \mu\right)}$ and for $\theta = 2$, $\delta = \rho/\sigma_{E^S_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix}, \mu\right)$ [EHHJ07]. Several special cases are included in Eq. 3.84, e.g., a Laplacian and a Gamma PDF.

In [EHH08], the required expectation $\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ is derived based on the aforementioned PDFs for the cases $\theta = 1$ and $\theta = 2$.

For $\theta = 1$, there is no closed form solution available for the respective integrals which include modified Bessel functions among other things. However, approximations are available for low and high SNR values which can be solved analytically [EHH08] resulting in:

- Approximation for low SNR values using Taylor series expansion of length $L_{\max}$:

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}_{\text{LSNR}} = K_{\text{LSNR}}(\lambda,\mu)\cdot D(\lambda,\mu), \tag{3.85}$$

  where

$$K_{\text{LSNR}}(\lambda,\mu) = \frac{1}{2}\frac{\sum_{l=0}^{L_{\max}-1}\frac{1}{l!(l+1)!}\left(\frac{\gamma_{\text{K}}(\lambda,\mu)}{2}\right)^l \Gamma(\rho+2l+2)\mathcal{D}_{-(\rho+2l+2)}(\mathcal{X})}{\sum_{l=0}^{L_{\max}-1}\left(\frac{1}{l!}\right)^2\left(\frac{\gamma_{\text{K}}(\lambda,\mu)}{2}\right)^l \Gamma(\rho+2l)\mathcal{D}_{-(\rho+2l)}(\mathcal{X})}, \tag{3.86}$$

  and '!' denoting the factorial operator. $\mathcal{D}_{\rho'}(\cdot)$ states the parabolic cylinder function of order ρ' [GRJZ00] and $\mathcal{X} = \sqrt{\rho(\rho+1)/(2\xi_{\text{K}}(\lambda,\mu))}$.

- Approximation for high SNR values:

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}_{\text{HSNR}} = K_{\text{HSNR}}(\lambda,\mu)\cdot D(\lambda,\mu), \tag{3.87}$$

  where

$$K_{\text{HSNR}}(\lambda,\mu) = \frac{(\rho-1/2)\left(\mathcal{X}\mathcal{D}_{-(\rho+1/2)}(\mathcal{X}') + (\rho+\frac{1}{2})\mathcal{D}_{-(\rho+3/2)}(\mathcal{X}')\right)}{2\gamma_{\text{K}}(\lambda,\mu)\mathcal{D}_{-(\rho-1/2)}(\mathcal{X}')} - \frac{\rho}{2\gamma_{\text{K}}(\lambda,\mu)}, \tag{3.88}$$

with $\mathcal{X}' = \mathcal{X} - \sqrt{2\gamma_{\text{K}}(\lambda,\mu)}$. The final estimate for $\theta = 1$ combines the two approximations using the procedure in [EHHJ07]:

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}^{\theta=1} = \max\big(K_{\text{LSNR}}(\lambda,\mu), K_{\text{HSNR}}(\lambda,\mu)\big)\cdot D(\lambda,\mu). \tag{3.89}$$

Due to the approximations of the Bessel functions within the derivation, the solution is inappropriate for $\rho < 0.5$. According to [EHHJ07], $L_{\max} = 5$ is sufficient for the purpose of speech enhancement leading to an overall approximation error between +3.7 dB and -0.2 dB where a positive error means that the approximated gain function is smaller than its actual value.

For $\theta = 2$, a closed form solution can be derived given as [EHH08]:

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}^{\theta=2} = \frac{\rho\xi_{\text{K}}(\lambda,\mu)}{\rho+\xi_{\text{K}}(\lambda,\mu)} \frac{\mathcal{M}\left(\rho+1;2;\frac{\gamma_{\text{K}}(\lambda,\mu)\xi_{\text{K}}(\lambda,\mu)}{\rho+\xi_{\text{K}}(\lambda,\mu)}\right)}{\mathcal{M}\left(\rho;1;\frac{\gamma_{\text{K}}(\lambda,\mu)\xi_{\text{K}}(\lambda,\mu)}{\rho+\xi_{\text{K}}(\lambda,\mu)}\right)} \cdot D(\lambda,\mu), \tag{3.90}$$

with $\mathcal{M}(\cdot)$ representing the confluent hypergeometric function [GRJZ00].

In order to get a good approximation to the measured histogram in Fig. 3.10(b), the minimum Kullback Leibler distance[8] between modeled and measured PDF is determined by varying the model parameters θ and ρ in Eq. 3.84 while keeping the power normalized to $\sigma^2_{E^S_{\text{prop}}} = 1$. Based on the minimum Kullback Leibler distance, the optimal approximation is given for the parameters $\theta_0 = 1$ and $\rho_0 = 0.9093$. Figure 3.11 illustrates the measured histogram for the amplitude of the speech prediction error together with the resulting fitted approximation.

The relation given in Eq. 3.75 holds for multivariate Gamma distributed variables as well, e.g., [Iza65] and [KBJ00]. Thus, the required conditional expectation vector $\mathbb{E}\{\mathbf{E}^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ based on the generalized Gamma model is given by (cf. Eq. 3.76):

$$\mathbb{E}\{E^S_{\text{prop}}(\lambda-\kappa+1,\mu)|D(\lambda,\mu)\} = \frac{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda-\kappa+1}{\lambda},\mu)}{\sigma^2_{E^S_{\text{prop}}}\binom{\lambda}{\lambda},\mu)} \mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}^{\theta=1|2}, \tag{3.91}$$

where $1 \leq \kappa \leq N_{\text{K}}$ and $\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}^{\theta=1|2}$ is obtained using the parameters θ_0 and ρ_0.

In order to determine the expectation of the noise prediction error conditioned on the differential signal $D(\lambda,\mu)$, Eq. 3.64 can be exploited. Using the afore derived expression for the speech prediction error, the conditional expectation $\mathbb{E}\{E^N_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}$ at the current frame yields:

$$\begin{aligned}\mathbb{E}\{E^N_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\} &= \mathbb{E}\{D(\lambda,\mu) - E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\} \\ &= \mathbb{E}\{D(\lambda,\mu)|D(\lambda,\mu)\} - \mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\} \\ &= D(\lambda,\mu) - \mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu)\}. \end{aligned} \tag{3.92}$$

Please note that the same expression also holds for the Gaussian model in Sec. 3.2.3.1, cf. Eqs. 3.76 and 3.77 for $\kappa = 1$ and $\tau = 1$, respectively. As the noise prediction error

[8] Information theoretic measure for the similarity of two PDFs: To differentiate between an analytical PDF $p_{\text{a}}(x)$ and a measured PDF $p_{\text{m}}(x)$, the Kullback Leibler distance can be calculated according to [KL51]: $J = \int_{x=-\infty}^{\infty} (p_{\text{a}}(x) - p_{\text{m}}(x)) \log_{10}\left(\frac{p_{\text{a}}(x)}{p_{\text{m}}(x)}\right) \text{d}x$.

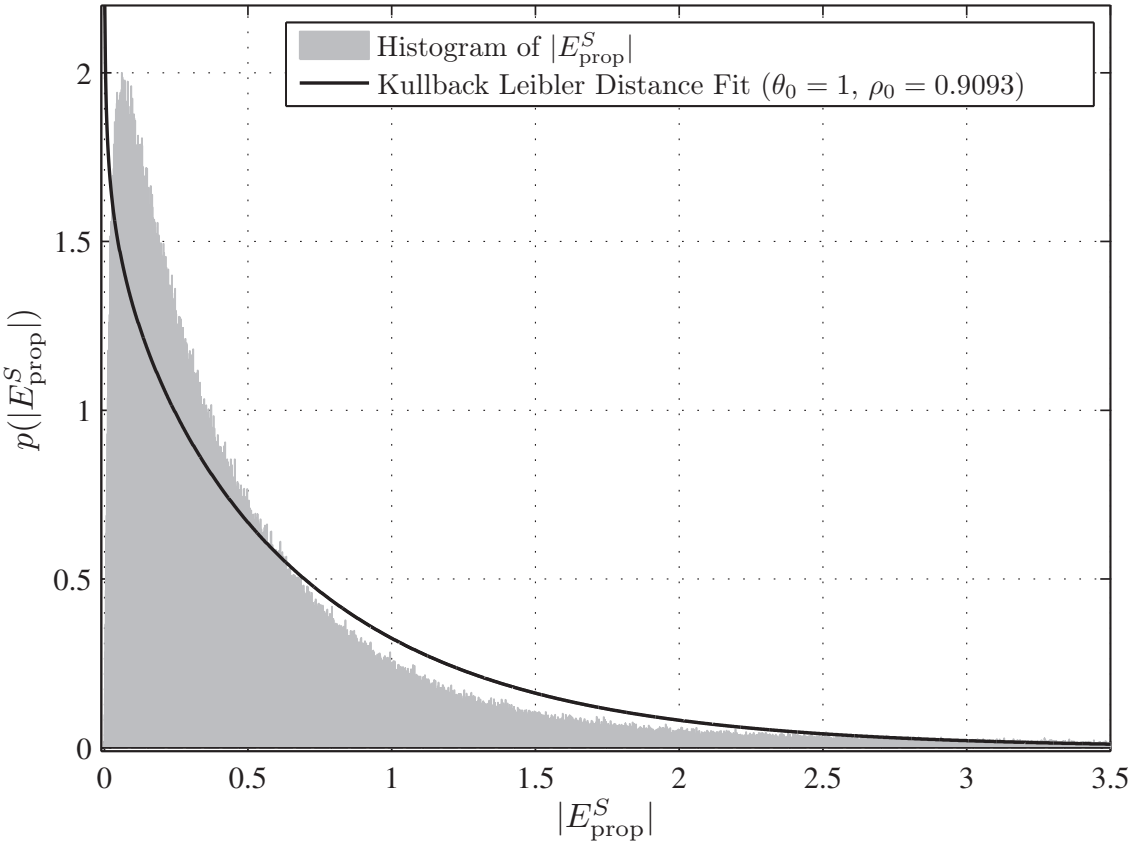

Figure 3.11: Histogram of speech prediction error amplitudes $|E^S_{\text{prop}}|$ and fitted approximation of Eq. 3.84 according to minimum Kullback Leibler distance.

is still assumed to be complex-Gaussian distributed, the missing elements of the noise prediction error vector $\mathbf{E}^N_{\text{prop}}(\lambda, \mu)$ are given as follows:

$$\mathbb{E}\{E^N_{\text{prop}}(\lambda-\tau+1,\mu)|D(\lambda,\mu)\} = \frac{\sigma^2_{E^N_{\text{prop}}}\binom{\lambda-\tau+1}{\lambda},\mu)}{\sigma^2_{E^N_{\text{prop}}}\binom{\lambda}{\lambda},\mu)}\mathbb{E}\{E^N_{\text{prop}}(\lambda,\mu)|D(\lambda,\mu\}, \quad (3.93)$$

for $1 \leq \tau \leq M_{\text{K}}$.

3.2.3.3 SNR Influence on Statistics of Prediction Error Signal

Figure 3.11 depicts the measured histogram of $|E^S_{\text{prop}}|$ averaged over a broad SNR range from -25 dB to 35 dB. In this section, it is shown that the input SNR has a relevant influence on the statistics of the speech prediction error signal in the propagation step. As a novelty, this characteristic is taken into account in the update step by using different SNR-dependent MMSE estimators which rely on the generalized Gamma priors introduced in the previous section.

For the following evaluation, the Kalman filter system of Sec. 3.2.3.1 is used again with the same amount of data as in Sec. 3.2.3.2 relying on about 1.5 hours of speech taken from the NTT speech database [NC94] and WGN. Depending on the input

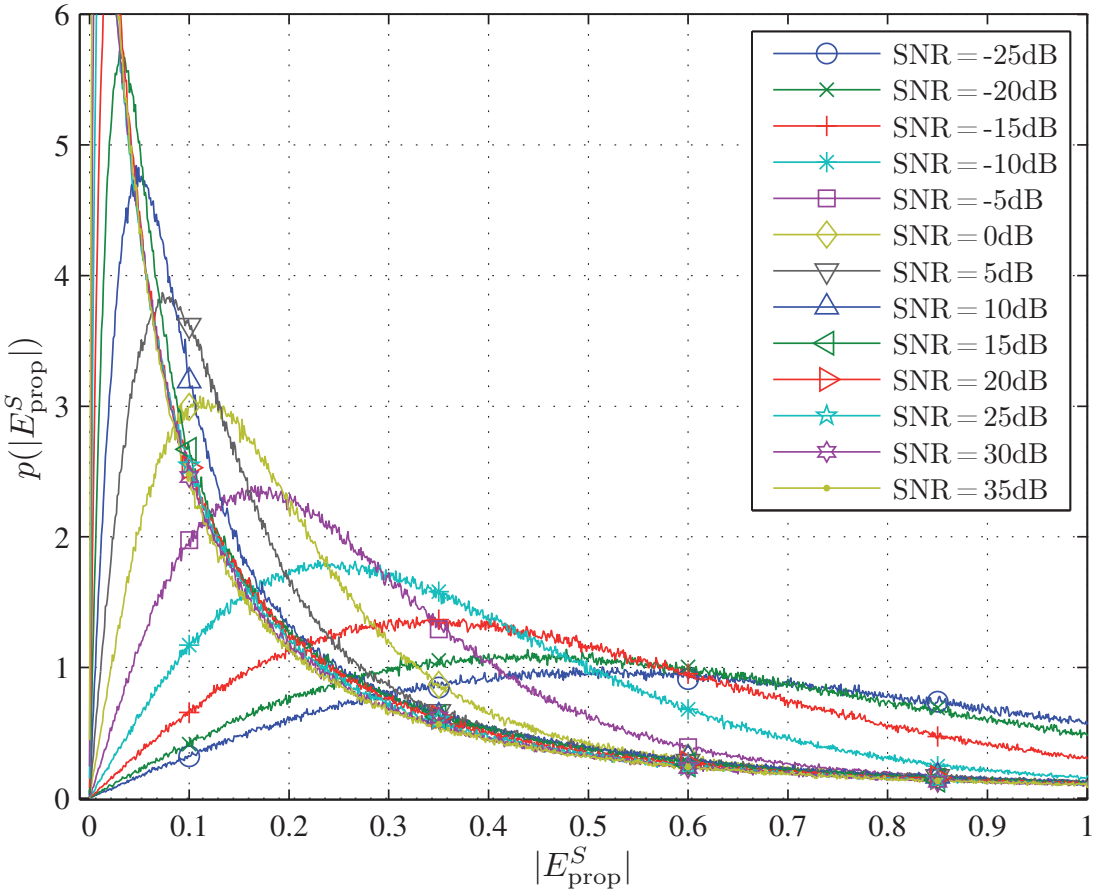

Figure 3.12: Normalized histograms of $|E^S_{\text{prop}}|$ dependent on input SNR.

SNR, which is varied in the range from -25 dB to 35 dB (step size: 5 dB), the histograms of the speech prediction error E^S_{prop} are measured, this time *separately* for each SNR value and not averaged over the entire SNR range. Figure 3.12 shows the measured histograms of the amplitude E^S_{prop} depending on different SNR values. The histograms are normalized to a power of $\sigma^2_{E^S_{\text{prop}}} = 1$ to illustrate the dependencies of the shape of the PDF on the input SNR. The SNR-dependency can clearly be seen. The steepness of the respective probability density functions around zero is getting larger for higher input SNR values showing that smaller prediction error values occur proportionally more often at higher SNR values. This behavior goes along with the fact that the prediction in the propagation step performs better the higher the input SNR is [EV08a].

In order to exploit this SNR-dependency, the complex-valued DFT estimator of the previous section is adapted to each of the measured histograms in Fig. 3.12. For the approximation, a similar procedure as in Sec. 3.2.3.2 is applied. The Kullback Leibler distances between modeled and measured PDFs are minimized by altering the parameters θ and ρ in Eq. 3.84 in the specified domains. The resulting settings which contribute to the best approximations are given in Tab. 3.1 for each SNR value. Due to the constraint that the parameter ρ has to be greater than or equal to 0.5 for the case $\theta = 1$ (see Sec. 3.2.3.2), the parameter sets for SNR values greater than 25 dB are identical. For all other SNR values, different MMSE estimators arise for the use

SNR [dB]	≤ -25	-20	-15	-10	-5	0	5
θ	1	1	1	1	1	1	1
ρ	2.2936	1.9541	1.5101	1.1629	0.9491	0.8085	0.7074

SNR [dB]	10	15	20	25	30	≥ 35
θ	1	1	1	1	1	1
ρ	0.6254	0.5584	0.5137	0.5000	0.5000	0.5000

Table 3.1: Parameter settings for complex-valued DFT MMSE estimator depending on the input SNR according to the model PDF in Eq. 3.84.

within the update step and the input SNR decides which parameter set to use in the current frame and frequency bin.

In the real system, the SNR has to estimated based on enhanced speech and noise DFT coefficients from the past. In order to determine which settings are used in the current frame, the averaged and quantized SNR estimates of the previous N_{K} frames are utilized. The decision is made individually for each frequency bin.

3.2.3.4 Estimation of Prediction Error Powers

In order to determine the speech and noise prediction error vectors $\mathbf{E}^S_{\mathrm{prop}}(\lambda,\mu)$ and $\mathbf{E}^N_{\mathrm{prop}}(\lambda,\mu)$ for the Gaussian model as well as for the generalized Gamma model in Eqs. 3.76, 3.77, 3.91 and 3.93, the prediction error powers $\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\kappa+1\\ \lambda\end{smallmatrix},\mu\right)$ and $\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\tau+1\\ \lambda\end{smallmatrix},\mu\right)$ are required. These entities can directly be obtained from the error covariance matrices $\mathbf{P}^S_{\mathrm{prop}}(\lambda,\mu)$ and $\mathbf{P}^N_{\mathrm{prop}}(\lambda,\mu)$ in Eqs. 3.62 and 3.63 for $2 \leq \kappa \leq N_{\mathrm{K}}$ and $2 \leq \tau \leq M_{\mathrm{K}}$ using the transition matrices $\mathbf{A}$ and $\mathbf{B}$. For $\kappa = 1$ and $\tau = 1$, i.e., the prediction error powers $\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ and $\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ of the current frame, additional estimates of the excitation signal powers $\sigma^2_{E_S}(\lambda,\mu)$ and $\sigma^2_{E_N}(\lambda,\mu)$ are necessary, see Eqs. 3.62 and 3.63. Therefore, the estimated autocorrelation functions and prediction coefficients of the speech and the noise signals can be used, e.g., [Pud02]. However, after long periods of speech inactivity, when the Kalman filter output $\hat{S}_{\mathrm{up}}(\lambda,\mu)$ converges to zero, these estimates are obtained from past speech DFT coefficients which are all almost zero leading to suppressed speech at speech onsets. To counteract this behavior, speech onset periods can be detected and treated separately [ZVY06b] by applying additional computational effort.

As a novel feature, the prediction error powers $\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ and $\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ of the current frame are directly estimated in this approach in order to prevent the requirement of determining speech onsets. For this, the estimates of the error covariance matrices for $\kappa = 1$ and $\tau = 1$ are skipped and the update step is considered again as stand-alone noise suppression system with the objective to decompose the 'noisy' input signal $D(\lambda,\mu)$ into 'differential target' signal $E^S_{\mathrm{prop}}(\lambda,\mu)$ and 'differential noise' signal $E^N_{\mathrm{prop}}(\lambda,\mu)$. At first, the 'noise' power $\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ is estimated by applying

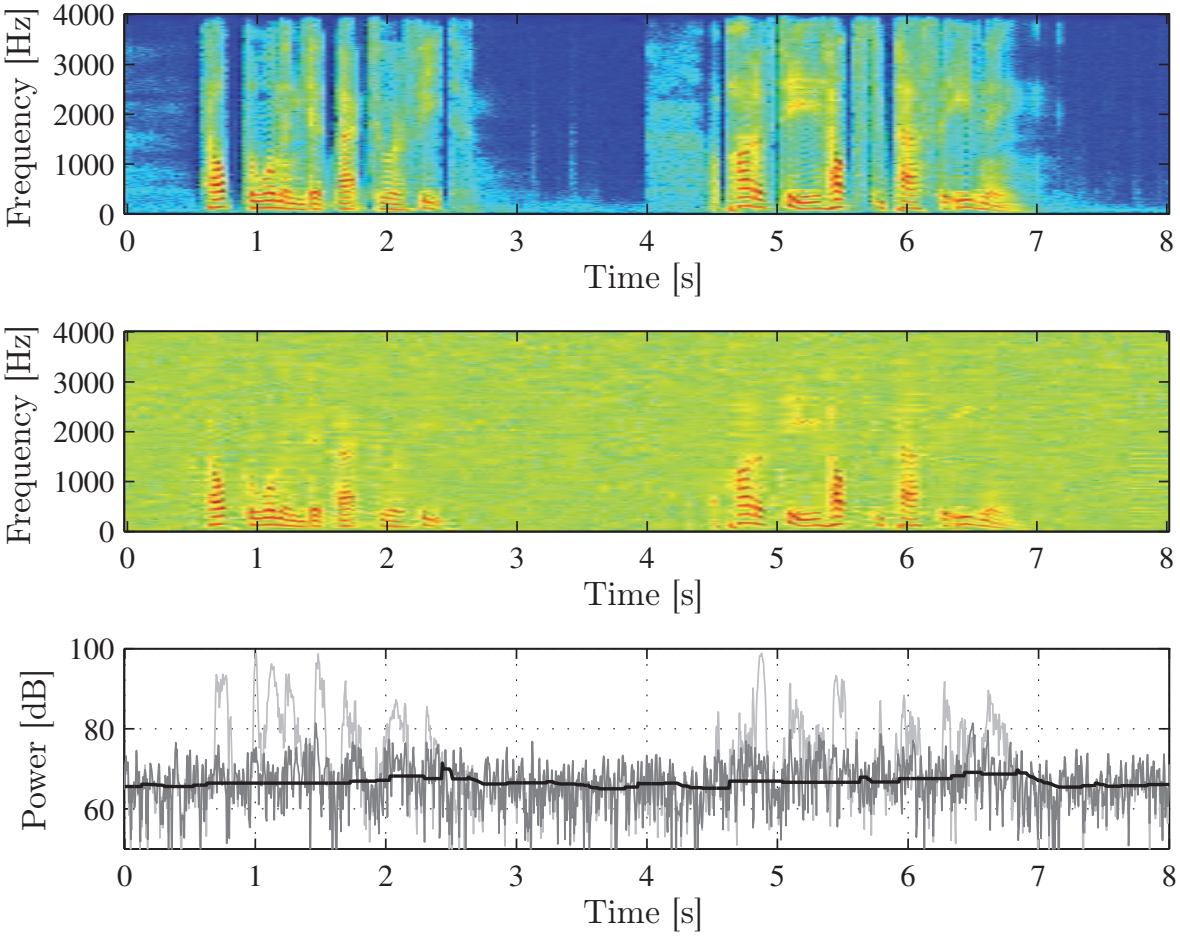

Figure 3.13: Estimation of noise prediction error power using Minimum Statistics within the update step. The clean speech signal "Help the woman get back to her feet. A pot of tea helps to pass the evening." (male voice) in the upper plot is disturbed by WGN at 10 dB input SNR. The spectrogram of the noisy signal is depicted in the middle plot. The lower plot illustrates the 'noisy' squared magnitude $|D(\lambda,\mu)|^2$ (light grey), the true 'noise' power $\sigma^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ (grey) and the estimated 'noise' power $\hat{\sigma}^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ (black) at frequency bin $\mu = 15$ using an FFT size of $M_{\text{F}} = 256$.

the well-known *Minimum Statistics* (MS) [Mar01] approach (see Sec. 2.3.2) to the differential signal $D(\lambda,\mu)$. Although Minimum Statistics was originally developed for speech signals disturbed by additive background noise, it also works well in the update step. An example for the estimation performance is depicted in Fig. 3.13 for a speech signal taken from the NTT speech database [NC94] which is disturbed by WGN at 10 dB input SNR. As can be seen, the power of the differential signal often decays to the power level of the 'noise' signal E^N_{prop}. Therefore, it is possible to track the 'noise' power $\sigma^2_{E^N_{\text{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ in the update step using the original MS approach of [Mar01].

All weighting rules which are presented in Secs. 3.2.3.1, 3.2.3.2 and 3.2.3.3 only depend on the a posteriori SNR and a priori SNR. If an estimate of the 'noise' power is

available, the a posteriori SNR $\gamma_{\mathrm{K}}(\lambda,\mu)$ can easily be measured, see Eq. 3.72. For the estimation of the a priori SNR $\xi_{\mathrm{K}}(\lambda,\mu)$, an estimate of the speech prediction error power $\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ is required in addition, cf. Eq. 3.73. For this purpose, the *decision-directed* approach [EM84] is applied as follows:

$$\hat{\xi}_{\mathrm{K}}(\lambda,\mu) = \alpha_{\mathrm{K}}^{\mathrm{DD}} \frac{|\hat{E}^S_{\mathrm{prop}}(\lambda-1,\mu)|^2}{\hat{\sigma}^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-1\\ \lambda-1\end{smallmatrix},\mu\right)} + (1-\alpha_{\mathrm{K}}^{\mathrm{DD}})\ \max\bigl(\hat{\gamma}_{\mathrm{K}}(\lambda,\mu)-1,0\bigr). \tag{3.94}$$

The smoothing factor $\alpha_{\mathrm{K}}^{\mathrm{DD}}$ here also states a tradeoff between noise attenuation and musical tones. Applied within the update step, a value of $\alpha_{\mathrm{K}}^{\mathrm{DD}} = 0.875$ achieves a good compromise in the proposed system.

3.2.3.5 Summary

In the following, the procedures in the update step for the Gaussian model as well as for the generalized Gamma model are briefly summarized.

1. Estimation of noise prediction error power $\sigma^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$ using the Minimum Statistics approach applied to the differential signal $D(\lambda,\mu)$.
2. Estimation of a posteriori SNR γ_{K} and a priori SNR ξ_{K} based on the decision-directed approach.
3. Determination of the Kalman gains $K(\lambda,\mu)$ for the current frame λ:
   - Gaussian Model

$$\begin{aligned}\text{Speech:}\quad & K^S(\lambda,\mu) = K^S_{\mathrm{G}}(\lambda,\mu) = \frac{\hat{\xi}_{\mathrm{K}}(\lambda,\mu)}{\hat{\xi}_{\mathrm{K}}(\lambda,\mu)+1}\\ \text{Noise:}\quad & K^N(\lambda,\mu) = K^N_{\mathrm{G}}(\lambda,\mu) = 1 - K^S_{\mathrm{G}}(\lambda,\mu)\end{aligned}$$

   - Generalized Gamma Model[9] using $L_{\max} = 5$ (see Eq. 3.86)

$$\begin{aligned}\text{Speech:}\quad & K^S(\lambda,\mu) = K^S_{\mathrm{gG}}(\lambda,\mu) = \max\left(K_{\mathrm{LSNR}}(\lambda,\mu), K_{\mathrm{HSNR}}(\lambda,\mu)\right)\\ \text{Noise:}\quad & K^N(\lambda,\mu) = K^N_{\mathrm{gG}}(\lambda,\mu) = 1 - K^S_{\mathrm{gG}}(\lambda,\mu)\end{aligned}$$

4. Estimation of speech prediction error power $\sigma^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)$:

$$\hat{\sigma}^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right) = \hat{\xi}_{\mathrm{K}}(\lambda,\mu)\cdot\hat{\sigma}^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right).$$

[9] The appropriate values for the model parameter ρ are set according to Secs. 3.2.3.2 and 3.2.3.3.

5. Determination of missing Kalman gain vector entries for $2 \leq \kappa \leq N_{\mathrm{K}}$ and $2 \leq \tau \leq M_{\mathrm{K}}$:

$$\text{Speech:} \quad K^S(\lambda-\kappa,\mu) = \frac{\hat{\sigma}^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\kappa+1\\ \lambda\end{smallmatrix},\mu\right)}{\hat{\sigma}^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)} K^S(\lambda,\mu)$$

$$\text{Noise:} \quad K^N(\lambda-\tau,\mu) = \frac{\hat{\sigma}^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\tau+1\\ \lambda\end{smallmatrix},\mu\right)}{\hat{\sigma}^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda\\ \lambda\end{smallmatrix},\mu\right)} K^N(\lambda,\mu),$$

where $K^S(\lambda,\mu)$ and $K^N(\lambda,\mu)$ are the respective gains for the Gaussian or generalized Gamma model according to Step 3. The values for $\hat{\sigma}^2_{E^S_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\kappa+1\\ \lambda\end{smallmatrix},\mu\right)$ and $\hat{\sigma}^2_{E^N_{\mathrm{prop}}}\left(\begin{smallmatrix}\lambda-\tau+1\\ \lambda\end{smallmatrix},\mu\right)$ result from the estimated error covariance matrices $\hat{\mathbf{P}}^S_{\mathrm{prop}}(\lambda,\mu)$ and $\hat{\mathbf{P}}^N_{\mathrm{prop}}(\lambda,\mu)$ of the propagation step in Eqs. 3.62 and 3.63.

6. Estimation of prediction error vectors:

$$\hat{\mathbf{E}}^S_{\mathrm{prop}}(\lambda,\mu) = \mathbf{K}^S(\lambda,\mu)D(\lambda,\mu)$$
$$\hat{\mathbf{E}}^N_{\mathrm{prop}}(\lambda,\mu) = \mathbf{K}^N(\lambda,\mu)D(\lambda,\mu),$$

where the vectors $\mathbf{K}^S(\lambda,\mu)$ and $\mathbf{K}^N(\lambda,\mu)$ contain the afore computed Kalman gains either for the Gaussian or the generalized Gamma model. $\hat{\mathbf{E}}^S_{\mathrm{prop}}(\lambda,\mu)$ and $\hat{\mathbf{E}}^N_{\mathrm{prop}}(\lambda,\mu)$ are used to update the initial predictions of the propagation step according to Eqs. 3.49 and 3.50.

7. Update of error covariance matrices $\hat{\mathbf{P}}^S_{\mathrm{up}}(\lambda,\mu)$ and $\hat{\mathbf{P}}^N_{\mathrm{up}}(\lambda,\mu)$:

$$\hat{\mathbf{P}}^S_{\mathrm{up}}(\lambda,\mu) = \left(\mathbf{I} - \mathbf{K}^S(\lambda,\mu)\mathbf{h}_S^T\right)\hat{\mathbf{P}}^S_{\mathrm{prop}}(\lambda,\mu)$$
$$\hat{\mathbf{P}}^N_{\mathrm{up}}(\lambda,\mu) = \left(\mathbf{I} - \mathbf{K}^N(\lambda,\mu)\mathbf{h}_N^T\right)\hat{\mathbf{P}}^N_{\mathrm{prop}}(\lambda,\mu).$$

3.3 Performance Results

The novel noise reduction system developed in this chapter exploits temporal correlation between adjacent frames in the frequency domain and can be realized with different statistical estimators in the update step. In this section, the presented model-based approaches are analyzed and compared with several state-of-the-art noise suppression techniques with regard to speech quality, speech distortions and noise attenuation. At first, purely statistical weighting rules as well as the Kalman filter proposed in [ZVY06b] serve as reference for the Kalman filter presented in Sec. 3.2.3.1 under the Gaussian model. Thereafter, it is shown that the exploitation of the prediction error statistics under the generalized Gamma model in Secs. 3.2.3.2 and 3.2.3.3 leads

to further improvements followed by investigations of the SNR-dependent prediction gain within the propagation step. The instrumental measurements are confirmed by an informal listening test and finally, a visualization example of the processed signals based on spectrograms closes this chapter. The computational complexity as well as the memory requirements of the proposed modified Kalman filter approach are analyzed in Appendix C.

3.3.1 Instrumental Measurements

For the investigation, several instrumental measurements are applied, see Appendix D for more details. The main parameter settings that are used in the simulations are listed in Tab. 3.2. Five speech signals from the NTT speech database [NC94] are each degraded by six different noise types (f16, babble, car, factory1, factory2, white), taken from the NOISEX-92 database [VS93] at an input SNR varying between -10 dB and 35 dB with an increment of 5 dB. Among the five speech signals, there are three sequences from male and two from female speakers, each with a length of 8 seconds.

For the simulations, the proposed Kalman filter system is initialized as follows: the speech transition matrix $\mathbf{A}(0,\mu)$ is set to $\mathbf{0}$ whereas the noise transition matrix $\mathbf{B}(0,\mu)$ is determined from the first M_{K} noisy coefficients $N(\check{\lambda},\mu)$ with $0 \leq \check{\lambda} < M_{\mathrm{K}}$. Moreover, the error covariance matrices of propagation and update step $\mathbf{P}^{S}_{\mathrm{prop}}(0,\mu)$, $\mathbf{P}^{S}_{\mathrm{up}}(0,\mu)$, $\mathbf{P}^{N}_{\mathrm{prop}}(0,\mu)$ and $\mathbf{P}^{N}_{\mathrm{up}}(0,\mu)$ are all initialized by zeros.

3.3.1.1 Gaussian Model

In this section, the developed Kalman filter system as discussed in Sec. 3.2.3.1 is evaluated using the Gaussian model in the update step. In the propagation step, two procedures are differentiated:

- **Proposed KF (*S* only, Gauss)** – The complex-valued prediction is applied only to the speech signal. For this approach, the noise prediction vector $\hat{\mathbf{N}}_{\mathrm{prop}}(\lambda,\mu)$ in Sec. 3.2.3.1 is set to $\mathbf{0}$, cf. Fig. 3.5.
- **Proposed KF (*S*+*N*, Gauss)** – The complex-valued prediction is applied to the speech signal *and* the noise signal as depicted in Fig. 3.5.

The proposed system is compared with the well-known Wiener filter [LO79] and the *Log Spectral Amplitude* (LSA) estimator [EM85] to illustrate the advantage of exploiting temporal correlation in addition to a priori knowledge of zeroth order. Moreover, the Kalman filter approach of [ZVY06b] is included in the investigations. This approach is applied in the frequency domain as well and, as described before, uses one Kalman filter to estimate the real parts of the speech and noise DFT coefficients and one Kalman filter for the respective imaginary parts. All investigated noise reduction techniques in this subsection assume a complex-Gaussian distribution for the spectral DFT coefficients of the speech and the noise signal.

Parameter	*Settings*
Sampling frequency	8 kHz
Frame length L_{F}	160 (20 ms)
FFT length M_{F}	256 (including zero-padding)
Frame overlap	75% (Hann window)
Input SNR	-10 dB ... 35 dB (step size: 5 dB)

Propagation Step	
AC length[10] L_{AC}	6 (see Fig. 3.7)
Model order N_{K}	3 (see Eq. 3.57)
Model order M_{K}	2 (see Eq. 3.60)

Update Step	
Noise estimation	Minimum Statistics [Mar01]
SNR estimation	Decision-directed approach [EM84]

Table 3.2: System settings.

The averaged results are depicted in Figs. 3.14 and 3.15. Figure 3.14 shows the differences between segmental noise and speech attenuation over the input SNR and Fig. 3.15 the segmental speech SNR plotted over the noise attenuation with the input SNR as control variable. Thus, a fair comparison with respect to the tradeoff between noise attenuation and speech distortions is possible. In Fig. 3.14, a higher score indicates a better performance in which a value greater than 0 dB justifies the application of noise suppression. In Fig. 3.15, high values for both the segmental speech SNR and the noise attenuation are desirable. More details about the instrumental measurements can be found in Appendix D.

The results show that the MMSE-LSA estimator yields a better performance than the Wiener filter for the entire SNR range as known from literature [Loi07]. In Fig. 3.14, the reference Kalman filter approach [ZVY06b] performs worst for input SNR values lower than 5 dB. However, its performance becomes better at higher input SNR values and it outperforms the two purely statistical weighting rules Wiener filter and MMSE-LSA estimator beyond an input SNR of 15 dB. Moreover, the reference Kalman filter approach consistently achieves a higher noise attenuation in Fig. 3.15 compared to the statistical estimators. However, this benefit comes at the expense of a lower segmental speech SNR comparing the corresponding markers in Fig. 3.15 for each input SNR value separately.

The instrumental measurements demonstrate that both novel variants of the proposed Kalman filter system *Proposed KF (S only, Gauss)* and *Proposed KF (S+N, Gauss)* consistently outperform all other noise reduction techniques including the two conventional approaches [LO79] and [EM85] as well as the reference Kalman filter system [ZVY06b]. Compared to the Wiener filter, gains of up to 2.2 dB can be achieved in

[10] The previous L_{AC} *enhanced* DFT coefficients $\hat{S}_{\mathrm{up}}$ and $\hat{N}_{\mathrm{up}}$ are applied to estimate autocorrelation vectors and matrices of speech and noise, respectively.

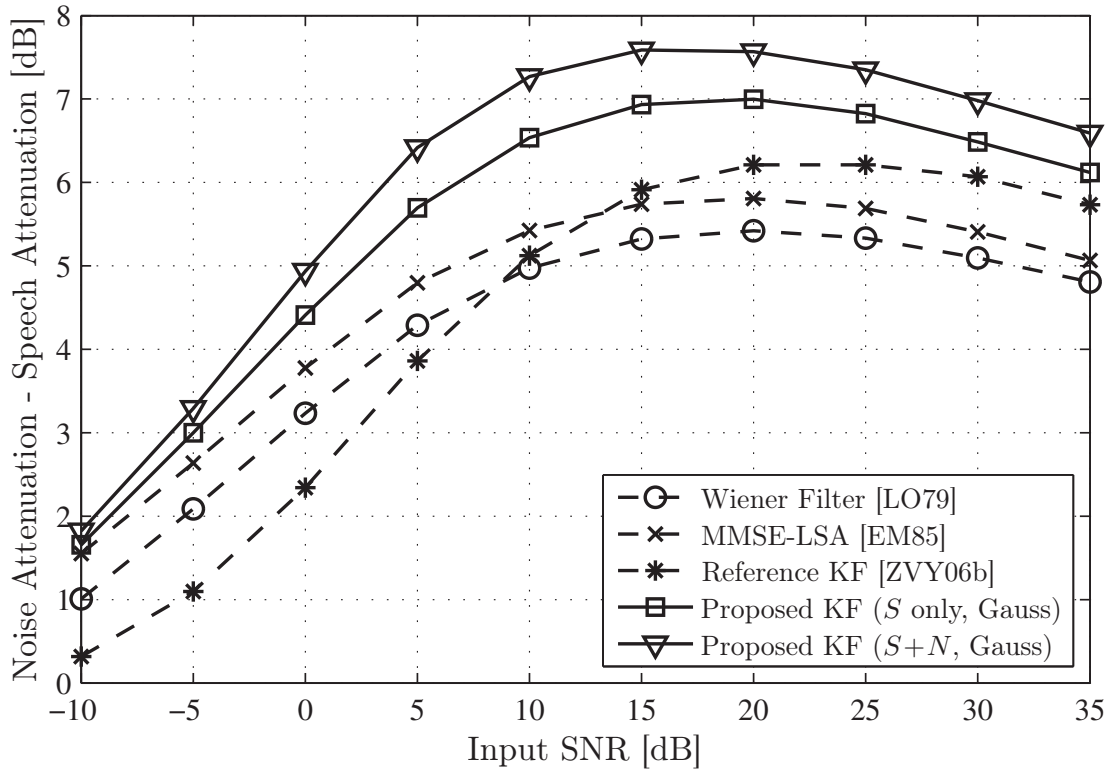

Figure 3.14: Difference between noise attenuation and speech attenuation plotted over input SNR. The different *Kalman Filter* (KF) setups are explained on page 61.

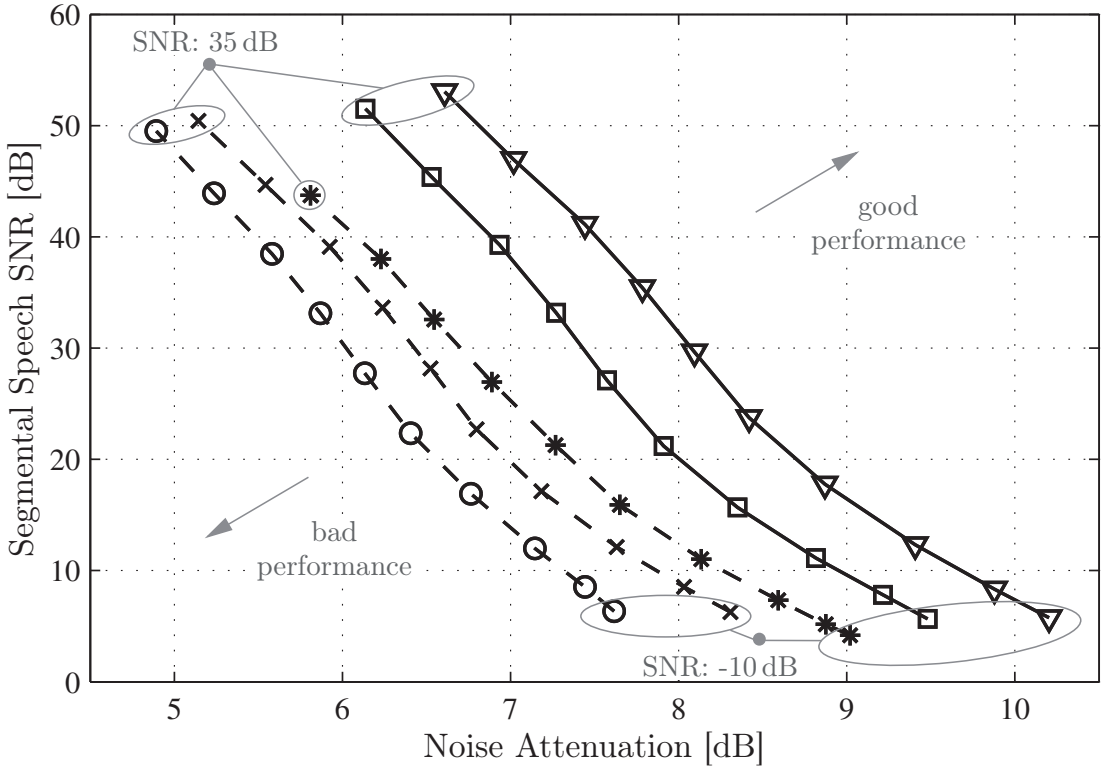

Figure 3.15: Segmental speech SNR plotted over noise attenuation.

terms of the deviation between noise and speech attenuation. If the segmental speech SNR is kept constant, an increase in noise attenuation of up to 2 dB is possible with the new model-based techniques as shown in Fig. 3.15. In addition to the exploitation of the temporal speech correlation, the results show that the temporal correlation of the applied noise signals can be exploited as well. If the proposed Kalman filter approach is applied to the speech *and* noise signal, the performance can be further increased.

3.3.1.2 Generalized Gamma Model

The investigation in this subsection analyzes the two developed Kalman filter approaches of Secs. 3.2.3.2 and 3.2.3.3 which explicitly exploit the statistics of the speech prediction error coefficients:

- **Proposed KF ($S+N$, genGam, SNR-indep.)** – The parameter set of the model PDF in Eq. 3.84 is kept fixed using θ_0 and ρ_0 for all input SNR values.
- **Proposed KF ($S+N$, genGam, SNR-dep.)** – Different parameter sets are used within the MMSE estimator depending on the input SNR, cf. Tab. 3.1.

They are compared with two purely statistical noise reduction techniques relying on super-Gaussian models for the speech signal as well: the Laplacian MMSE estimator [MB03] (see Sec. 2.5.4) and the super-Gaussian *Maximum A Posteriori* (MAP) estimator [LV05] (see Sec. 2.5.5). As reference, the results of the Wiener filter and the results of the proposed Kalman filter based on the Gaussian model *Proposed KF ($S+N$, Gauss)* are reproduced from the previous subsection.

The results are illustrated in Figs. 3.16 and 3.17. For the evaluation, the same instrumental measurements as in the previous subsection are applied. Figure 3.16 depicts the deviation between noise and speech attenuation plotted over the input SNR and in Fig. 3.17, the segmental speech SNR is plotted over the noise attenuation dependent on the input SNR.

The instrumental measurements in this section illustrate again the advantages of the proposed novel Kalman filter solutions compared to the considered statistical weighting rules, i.e., Wiener filter, Laplacian MMSE estimator and super-Gaussian MAP estimator. In addition, the results of the statistical estimators show that the more the estimator is adapted to the statistics of the target signal, the better the overall performance. This behavior can also be observed for the Kalman filter approaches: Figures 3.16 and 3.17 show consistent improvements due the application of the generalized Gamma model for the speech prediction error signal. Compared to the SNR-independent MMSE estimator applied in the update step, the SNR-dependent approach achieves better results in terms of noise attenuation and speech distortions for the entire SNR range. Thus, the additional exploitation of the prediction error SNR-dependency leads to further improvements and clearly outperforms all other considered noise reduction techniques. The proposed approach *Proposed KF ($S+N$,*

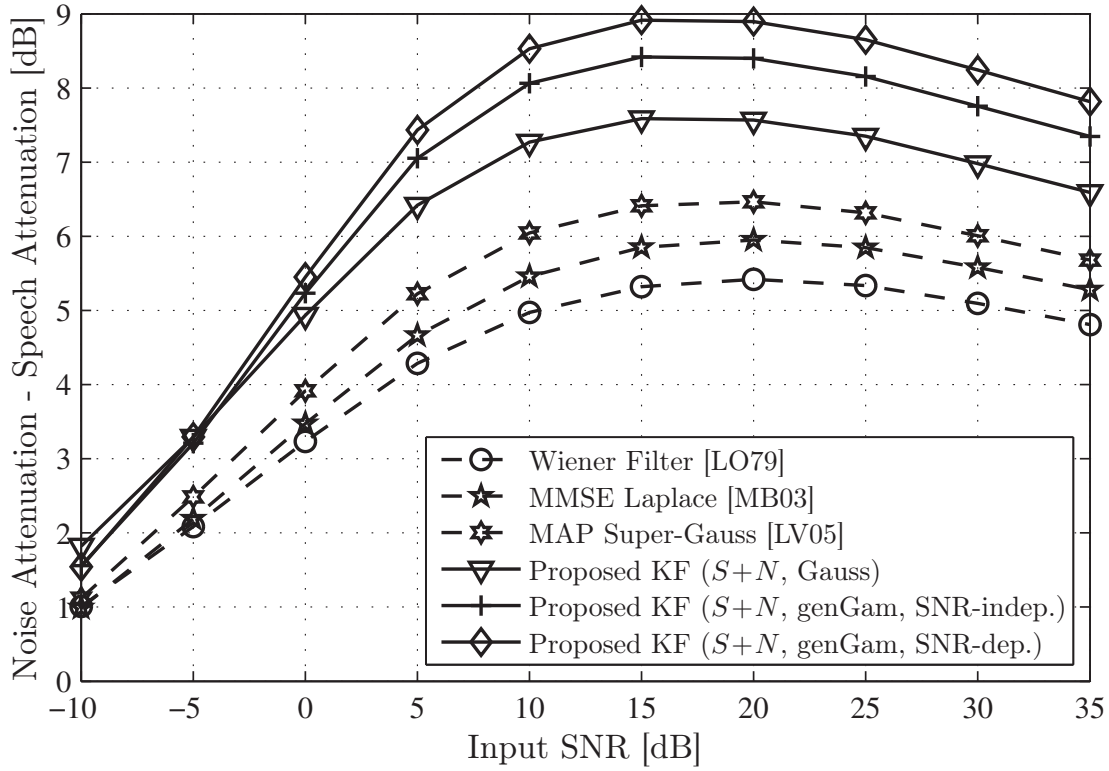

Figure 3.16: Difference between noise attenuation and speech attenuation plotted over input SNR. The different *Kalman Filter* (KF) setups are explained on pages 61 and 64.

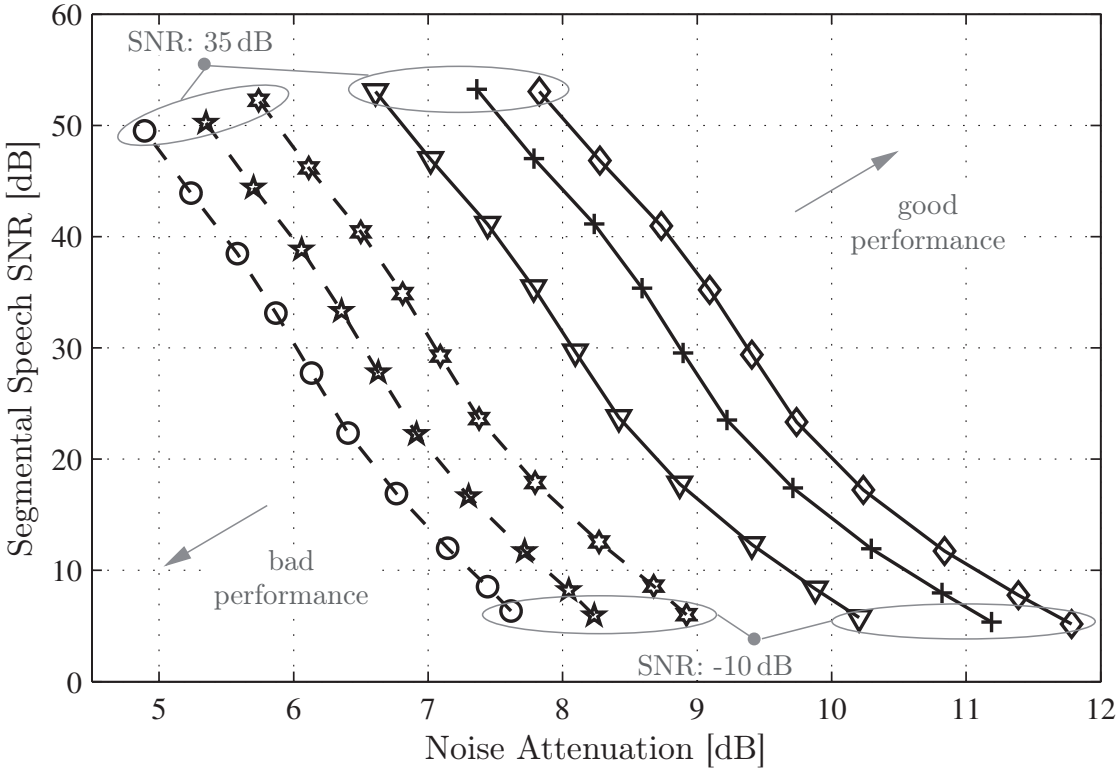

Figure 3.17: Segmental speech SNR plotted over noise attenuation.

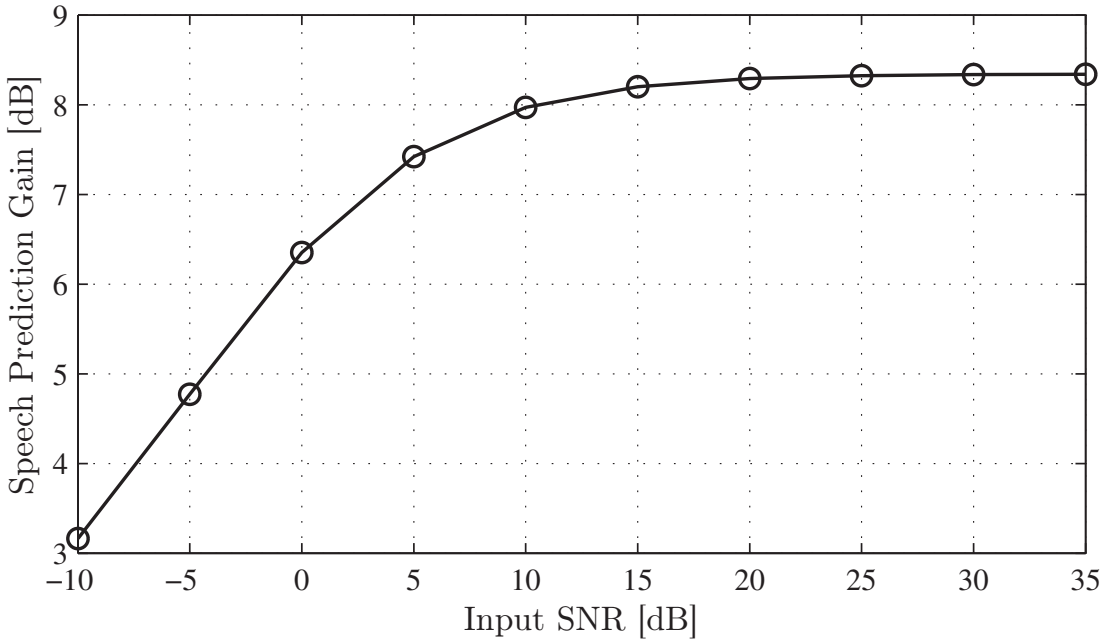

Figure 3.18: Prediction gain of speech DFT coefficients achieved in propagation step plotted over input SNR.

genGam, SNR-dep.) yields the best compromise between speech and noise attenuation as shown in Fig. 3.16 and the highest noise attenuation if the segmental speech SNR is kept constant as shown in Fig. 3.17. Nevertheless, the occurrence of *musical tones* is slightly increased the more the respective estimator is adapted to the statistics of the target signal. This applies to the super-Gaussian MAP estimator as well as to the proposed Kalman filter using the SNR-dependent MMSE estimators within the update step. However, one of the countermeasures presented in Sec. 5.1 can be used successfully to prevent the problem of musical noise.

3.3.1.3 Prediction Gain

In Sec. 3.2.3.3, it is shown that the PDF of the speech prediction error E^S_{prop} is dependent on the input SNR with smaller values occurring proportionally more often the higher the SNR is. This property results from the fact that the prediction gain within the propagation step is dependent on the SNR as well: the higher the input SNR, the better the performance of the complex-valued linear prediction. Figure 3.18 depicts the effective speech prediction gain of the proposed Kalman filter approach based on the SNR-dependent generalized Gamma model[11] plotted over the input SNR. It can be seen that the proposed system already starts to benefit from the propagation step at -10 dB input SNR and nearly reaches the level of 'ideal' prediction at 20 dB, cf. Fig. 3.7.

[11] Similar results are achieved using the proposed Kalman filter approach based on the Gaussian MMSE estimator in the update step.

3.3.2 Auditory Judgments

In addition to the instrumental measurements, an informal *Comparison Category Rating* (CCR) test according to [ITU96] was conducted which presented two processed samples per question to the participants: a processed signal from Method A and a processed signal from Method B. Two different scenarios were evaluated. On the one hand, two estimators relying on Gaussian models were compared, namely the Wiener filter [LO79] and the novel Kalman filter *Proposed KF (S+N, Gauss)* proposed in Sec. 3.2.3.1. On the other hand, the performance of two weighting rules was analyzed which can explicitly be adapted to the measured statistics of the respective target signal. Here, the results of the super-Gaussian MAP estimator [LV05] was compared with the novel Kalman filter approach *Proposed KF (S+N, genGam, SNR-dep.)* based on the SNR-dependent MMSE estimator which is presented in Sec. 3.2.3.3. For the evaluation, the labels 'Method A' and 'Method B' were randomly assigned to one of the respective noise reduction techniques in each scenario. The noisy input signals consisted of speech signals randomly taken from the NTT speech database disturbed by a noise signal from the NOISEX-92 database at an input SNR varying between 5 dB and 15 dB. 19 experienced listeners were asked to judge the overall speech quality in terms of noise attenuation, speech distortions and occurrence of musical tones. They could choose between the following rating options: Sample A sounds {much better | better | slightly better | about the same | slightly worse | worse | much worse} than Sample B. Each test person had to judge 10 signals (5 per scenario), i.e., the total results are based on $10 \cdot 19 = 190$ votes. The samples could be played ad libitum before the probands had to make their judgments.

The averaged results are separately illustrated in Fig. 3.19 for the two Scenarios (a) and (b). Both results clearly show that most listeners preferred the novel Kalman filter approaches compared to the corresponding purely statistical estimators. When combining the options {much better | better | slightly better }, approximately 75% of the test listeners favored the proposed model-based solution in Scenario (a) and about 77% in Scenario (b). The participants who preferred the purely statistical estimators in some isolated cases, explained their decisions by a slightly higher occurrence of musical tones in the processed signals of the Kalman filters which, in these cases, was perceived as more annoying than the higher noise suppression of the model-based approaches.

When comparing the quality of the novel Kalman filter techniques with the quality of the investigated reference Kalman filter [ZVY06b], the results of the instrumental measurements were confirmed as well. Informal listening tests here also showed the superiority of the Kalman filter solutions proposed in this thesis.

3.3.3 Spectrograms

Figure 3.20 allows a comparison of the proposed Kalman filter techniques by means of spectrograms. In the upper row, the spectrograms of the clean speech signal and the noisy input signal are depicted. The utterance "Help the woman get back

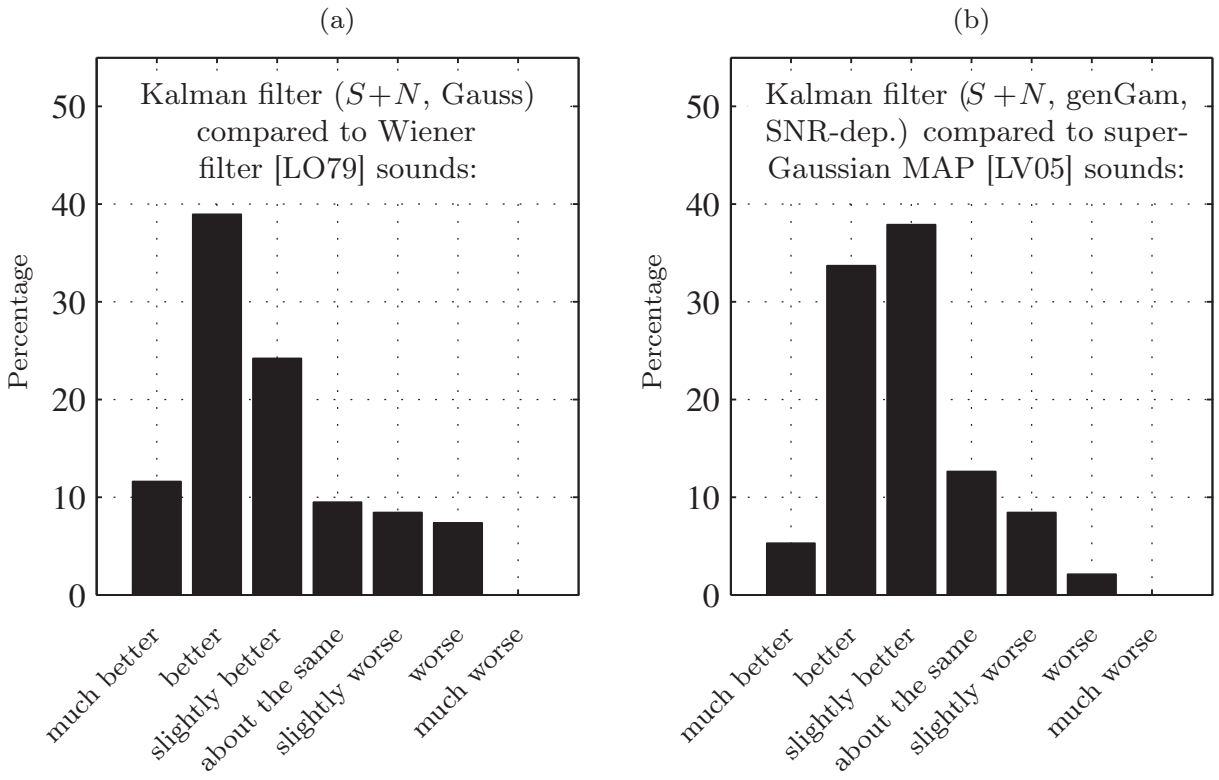

Figure 3.19: Results of the informal listening test comparing (a) the *proposed Kalman filter (S+N, Gauss)* with the Wiener filter [LO79] and (b) the *proposed Kalman filter (S+N, genGam, SNR dep.)* with the super-Gaussian MAP estimator [LV05]. The different Kalman filter setups are explained on pages 61 and 64.

to her feet. A pot of tea helps to pass the evening." originates from the NTT speech database [NC94] and is disturbed by factory noise from the NOISEX-92 database [VS93] at 5 dB input SNR. The middle and lower row show the spectrograms of the processed signals, among them the results of the two purely statistical estimators Wiener filter [LO79] and super-Gaussian MAP estimator [LV05] as well as the results of the two novel Kalman filter techniques relying on the Gaussian and the SNR-dependent generalized Gamma model for the speech prediction error signal. The instrumental measurements presented before are confirmed by the spectrograms as well. It can be seen that the model-based approaches in the lower row achieve a higher noise attenuation compared to the corresponding statistical weighting rules in the middle row without affecting the speech quality. In terms of the tradeoff noise attenuation and speech distortions, the Kalman filter techniques which uses the SNR-dependent generalized Gamma MMSE estimator within the update step yields the best results. Nevertheless, a new randomly fluctuating type of noise, referred to as musical noise, can be observed in the spectrogram at the bottom right of the figure. As mentioned before, countermeasures presented in Sec. 5.1 effectively help to avoid such effects.

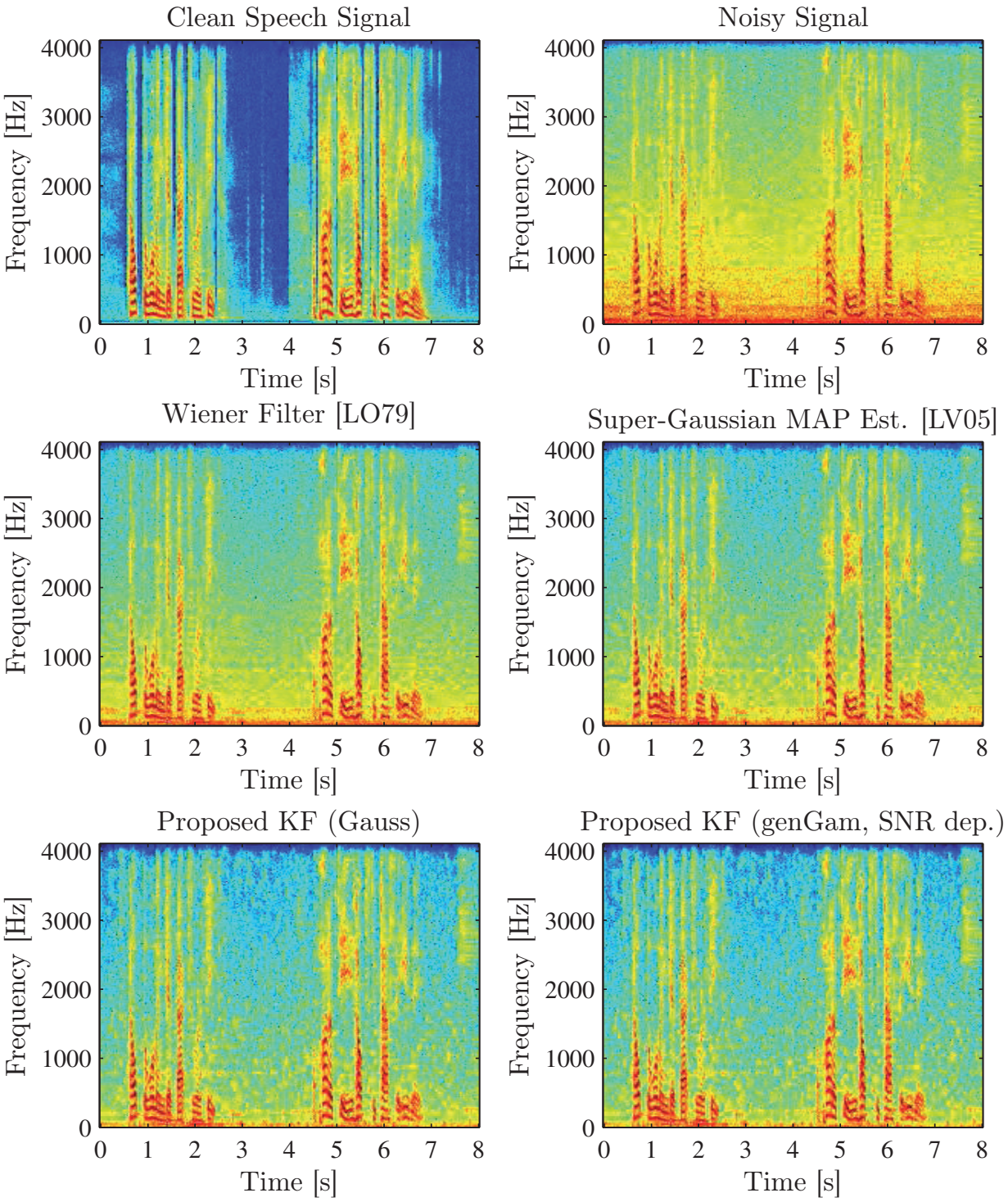

Figure 3.20: Spectrograms of clean speech signal, noisy signal (speech+factory noise, SNR: 5 dB), processed signals by Wiener filter, super-Gaussian MAP estimator and novel Kalman filter techniques using Gaussian model as well as SNR-dependent generalized Gamma model. The sentences "Help the woman get back to her feet. A pot of tea helps to pass the evening." are spoken by a male voice.

3.4 Conclusions

In this chapter, a novel Kalman filter for single-channel speech enhancement in the frequency domain is presented. The approach is based on a modified propagation and update step which are both applied directly to the complex-valued DFT coefficients of the noisy input signal. In the propagation step, temporal correlation of successive frames is exploited using low-order models to approximate the trajectories of speech and noise DFT coefficients. Investigations show that complex-valued linear prediction yields higher prediction gains than estimating real and imaginary parts or magnitudes and phases separately. The proposed system is able to exploit temporal correlation of speech already at very low input SNR values and nearly reaches the level of 'ideal' prediction at 20 dB input SNR. In the second (update) step, the first predictions are updated utilizing an appropriate statistical weighting rule in order to estimate the prediction errors caused in the propagation step. As novelty, not only the conventional Kalman filter gain (assuming Gaussian distributions for speech and noise) is taken into account for this purpose but also different SNR-dependent MMSE estimators which are explicitly adapted to measured histograms of the speech prediction error signal. Moreover, a new possibility to estimate the prediction error powers of speech and noise is presented. In the evaluation, the proposed system clearly outperforms several purely statistical estimators as well as the Kalman filter approach presented in [ZVY06b]. Especially the incorporation of the SNR-dependency on the statistics of the speech prediction error leads to significant improvements. The instrumental measurements are confirmed by an informal listening test in which about 75% of the test listeners preferred the signals processed by the novel Kalman filter solutions. Compared to state-of-the-art noise suppression systems, the overall computational load of the proposed system is increased by a factor of $5-7$. However, possible methods for an effective reduction of the complexity are presented in Appendix C.

4

Speech Enhancement Exploiting Spectral Dependencies

The quality of today's telephone speech is designed to achieve a reasonable intelligibility. The acoustic bandwidth in telephony systems is typically limited to the frequency range between 300 Hz and 3.4 kHz. However, the typical 'telephone sound' cannot satisfy the increased demands as the perceived speech quality and intelligibility is considerably reduced compared to the full audio bandwidth which usually ranges between 100 Hz and 10 kHz for human voices [VM06]. As a reasonable compromise, various wideband (50 Hz – 7 kHz) speech codecs have been developed in the past, e.g., [ITU88, ITU99, ITU06a, 3GP01, 3GP04], and are about to be introduced in current mobile networks. Nevertheless, most of these codecs are mainly designed for nearly noise-free input speech signals and do not perform well if the input signal is degraded by acoustic background noise. In order to improve the listening comfort and to keep the high quality also in noisy environments, noise suppression techniques are required for wideband communication systems.

One of the popular methods for enhancing degraded speech is based on modeling the noisy input signal in the *Short-Time Fourier Transform* (STFT) domain and to apply individual adaptive gains to the noisy STFT coefficients of each frequency bin. Most of the rules proposed in literature have been derived for low band signals (50 Hz – 4 kHz) under certain assumptions about the statistics of the speech and noise signals (cf. the weighting rules outlined in Chapter 2) or by using model-based techniques as presented in Chapter 3. When it comes to wideband noise reduction, an established method is to double sampling rate and transform length and to apply the low band algorithms also for higher frequencies. Thereby, neither the unequal spectral energy distribution of a speech and noise signal nor the properties of the human auditory system are considered. For a variety of different noise sources, it can be shown that the *Signal-to-Noise-Ratio* (SNR) significantly degrades beyond 4 kHz [EHGV10] leading to an imprecise noise reduction and fluctuating weighting gains at higher frequencies which result in the increased occurrence of *musical noise*. So far, only a very limited number of proposals are known which take into account the aforementioned aspects when enhancing wideband speech signals, e.g., [EHGV10, HEGV10, BSV06].

In contrast to wideband noise suppression, wideband speech coding has experienced a lot of progress in recent years. Depending on the infrastructure that is available at the terminals and within the telephone network, several possibilities exist how to set up a wideband connection. The use of dedicated wideband codecs, e.g., [3GP01], and embedded codecs, e.g., [ITU06a, ITU06b], achieves a high speech quality but requires a modification of the whole communication system. If the telephone network and applied protocols only support narrowband connections, additional information about the high band signal ($4-7\,\text{kHz}$) can be embedded into the bitstream of a narrowband codec by steganographic techniques, e.g., [VG07, GJV05, GV07, GV08, Esc06]. Therefore, the encoder as well as the decoder has to be modified in order to establish a wideband connection. Only the decoder has to be changed when using *Artificial Bandwidth Extension* (ABWE). This technique exploits spectral dependencies of speech signals in order to recover missing high frequency components by utilizing only the low band speech signal, i.e., ABWE aims at increasing the perceived speech quality if only the low band signal is available, e.g., [JV03b, JV06, GJV$^+$07]. Of course, the resulting speech quality is slightly worse compared to dedicated wideband codecs.

In the derivation of most noise reduction weighting rules, it is often assumed that adjacent *Discrete Fourier Transform* (DFT) coefficients within one frame are statistically independent. However, due to the harmonic structure of speech and a frame-based processing by means of a windowing operation, this assumption may not be justified in practice. For the purpose of noise reduction, only few publications are known which make use of correlation between individual frequency bins within one frame, e.g., [FBS05], [Plo09] and [HS08, Chapter 4].

In this chapter, wideband speech enhancement is combined with techniques known from ABWE. While a conventional noise suppression technique is used for the low band, a novel approach is applied for the speech enhancement in the high band. Based on a trained *Hidden Markov Model* (HMM), parameters from the processed (enhanced) low band signal are extracted and used to estimate subband energies of the high band speech signal. The resulting weighting gains determined from these energy estimates are adaptively combined with conventional gains for the high band. In addition, this chapter comprises an information theoretic view on ABWE under noisy conditions. A performance bound is formulated and the influence of noise reduction prior to ABWE is investigated by real entropy measurements.

The remainder of this chapter is organized as follows. At first, the general concept of ABWE is introduced when used for the purpose of speech coding. Afterwards, the proposed wideband noise reduction system is presented including the procedure of the combined noise suppression in the high band in detail. Thereafter, the mutual information between low and high band in noisy environments is analyzed. Finally, experimental results are shown and conclusions are drawn.

4.1 Artificial Bandwidth Extension

As mentioned at the beginning of this chapter, the application of artificial bandwidth extension is independent from the sending side of the transmission and therefore fully compatible with existing narrowband speech communication systems. This is important as the change of the current bandwidth limitation in public telephony systems, especially in the fixed-line networks, will not happen abruptly. Although ABWE does not achieve the full quality of true wideband coding, it can be used to improve the acceptance by the user while achieving a smooth transition between narrowband and wideband speech coding.

In this section, the basic concept of ABWE is summarized. The algorithm is based on the source-filter model of the human speech production system which is already briefly introduced at the beginning of Chapter 3. The estimation of the missing frequency components in the high band can be divided into two parts: extension of the narrowband *excitation signal* and estimation of the wideband *spectral envelope* using only information from the narrowband signal. After a short overview of the ABWE system, both steps are outlined in the following.

4.1.1 System Overview

In Fig. 4.1, a simplified block diagram of an ABWE system is depicted as proposed in [Jax02] and [JV03b]. At first, the sampling frequency of the low band input speech signal $s_{\mathrm{lb}}(k)$ is increased from $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ to $f_{\mathrm{s}} = 16\,\mathrm{kHz}$ by interpolation and subsequent low-pass filtering. From now on all further steps are applied to the upsampled low band signal $s_{\mathrm{lb}}(k')$ at the sampling frequency $f_{\mathrm{s}} = 16\,\mathrm{kHz}$ where k' denotes the time instance in the upsampled domain.

Based on $s_{\mathrm{lb}}(k')$ the spectral envelope of the narrowband signal is extended in the upper part of the block diagram. A feature vector $\mathbf{a}_{\mathrm{wb}}$ representing the spectral envelope of the wideband signal $s_{\mathrm{wb}}(k')$ is estimated. The vector $\mathbf{a}_{\mathrm{wb}}$ consists, e.g., of *Autoregressive* (AR) coefficients and is determined by exploiting information from an observation vector $\mathbf{x}_{\mathrm{lb}}$ as well as a priori knowledge provided by a pre-trained statistical model. Usually the vector $\mathbf{x}_{\mathrm{lb}}$ itself also contains information about the spectral envelope of the input signal $s_{\mathrm{lb}}(k')$.

The estimated vector $\hat{\mathbf{a}}_{\mathrm{wb}}$ is used to form a *Finite Impulse Response* (FIR) analysis filter which is applied to the input signal $s_{\mathrm{lb}}(k')$ in order to obtain an estimate of the bandlimited low band excitation signal $\hat{e}_{\mathrm{lb}}(k')$. In the next step, the missing high band frequencies in the excitation signal are determined. As the human ear is relatively insensitive to variations of the spectral fine structure at high frequencies, the procedure can be implemented quite efficiently. Different approaches can be found in literature for this purpose, see Sec. 4.1.2.

Finally, the estimated wideband excitation signal $\hat{e}_{\mathrm{wb}}(k')$ is combined with the envelope features of the vector $\hat{\mathbf{a}}_{\mathrm{wb}}$ using a synthesis filter which is inverse to the ap-

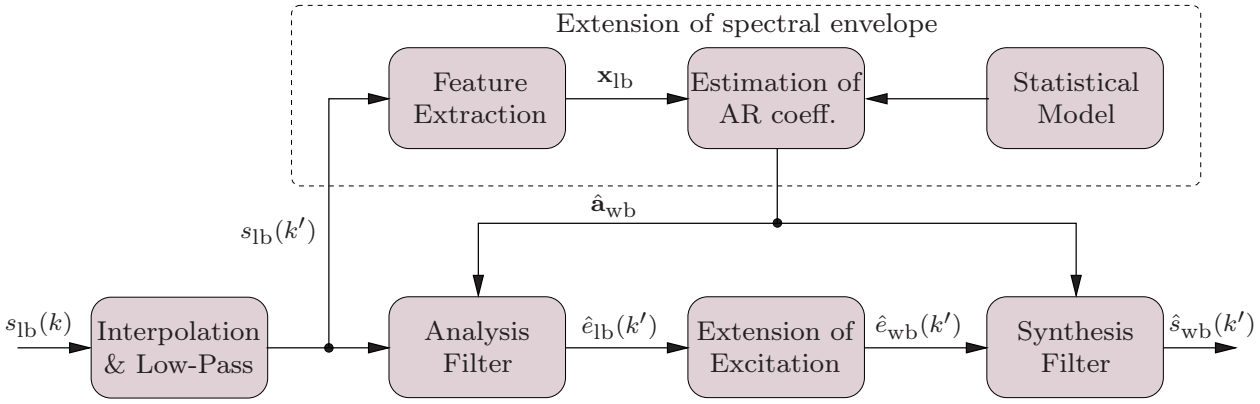

Figure 4.1: Block diagram of the *Artificial Bandwidth Extension* (ABWE) System.

plied analysis filter. The resulting signal $\hat{s}_{\mathrm{wb}}(k')$ provides an estimate of the wideband speech signal and exhibits transparency with respect to the low band input signal $s_{\mathrm{lb}}(k')$.

In the derivation and training phase of an ABWE system, clean speech signals are available and can be applied to the system. However, if the algorithm is used in practical speech communication systems, the quality of the low band input signal is often impaired due to background noise. In this case, the performance of ABWE significantly degrades and additional procedures are necessary, e.g., [SAD05].

4.1.2 Extension of the Excitation Signal

The objective of this step is to recover the spectral fine structure of the missing frequency components in the high band. Therefore, it has to be guaranteed that the estimated wideband excitation signal $\hat{e}_{\mathrm{wb}}(k')$ fully includes the low band excitation signal in order to accomplish the mentioned transparency with respect to the low band signal. A variety of different methods for the extension of the excitation signal can be found in literature, e.g., [Jax02]. Although this step is not required in the proposed wideband speech enhancement system, a selection of extension techniques is briefly outlined in the following for the sake of completeness.

Explicit Signal Generation

This method is based again on the source-filter model of the speech production system. The pitch frequency as well as a voiced/unvoiced decision is extracted from the low band excitation signal $\hat{e}_{\mathrm{lb}}(k')$. Afterwards, the missing components of the high frequencies are generated synthetically using a signal generator which consists of an impulse and noise generator. In order to achieve good results, an accurate estimate of the pitch frequency is necessary.

Non-Linear Processing

A non-linear function $g(\cdot)$ is employed in order to perform the extension of the excitation signal. Using this technique, the estimated wideband excitation signal results in:

$$\hat{e}_{\mathrm{wb}}^{\mathrm{NLP}}(k') = g\left(\hat{e}_{\mathrm{lb}}(k')\right), \tag{4.1}$$

and contains harmonic distortions which reflect the desired new components in the missing frequency band. For the choice of $g(\cdot)$, various non-linear functions are possible. It has been shown that good results can be achieved using a simple quadratic function, i.e., $\hat{e}_{\mathrm{wb}}^{\mathrm{NLP}}(k') = \left(\hat{e}_{\mathrm{lb}}(k')\right)^2$. Unfortunately, this method also produces undesired non-linear distortions making an additional postprocessing necessary.

Modulation in the Time Domain

Using this method, the wideband excitation signal arises by modulating the low band signal $\hat{e}_{\mathrm{lb}}(k')$ with a weighted cosine function as follows:

$$\hat{e}_{\mathrm{wb}}^{\mathrm{MTD}}(k') = \hat{e}_{\mathrm{lb}}(k') \cdot \zeta^{\mathrm{MTD}} \cdot \cos(\Omega_M k'). \tag{4.2}$$

A modulation in the time domain corresponds to a translation in the frequency domain, i.e., the low band signal $\hat{e}_{\mathrm{lb}}(k')$ is shifted in the frequency domain by Ω_M making a re-use of the original low band excitation signal possible. The factor ζ^{MTD} controls the power correction of the wideband excitation signal.

The frequency Ω_M can be chosen to be fixed, e.g., half of the sampling frequency or set according to the pitch frequency. The latter ensures that the harmonic structure of the excitation signal is maintained.

Pitch Scaling

Pitch scaling implies a doubling of the pitch frequency realized by a downsampling of factor two and subsequent time stretching of the low band excitation signal. During voiced periods, the generated pitch doubled signal contains tonal components at even integer multiples of the original pitch frequency. After pitch doubling, the signal is highpass filtered and added to a delayed version of the original low band excitation signal in order to obtain the wideband excitation signal.

4.1.3 Estimation of the Spectral Envelope

A very crucial part of any ABWE algorithm is the estimation of the spectral wideband envelope. Therefore, it is exploited that a typical speech vocabulary consists only of a limited number of sounds. The estimation of these sounds or of the respective feature vectors $\mathbf{a}_{\mathrm{wb}}$ of dimension b_{wb} which describe the spectral envelopes is performed using a statistical model that has to be trained in advance. The model is based on an HMM containing N_{bwe} different states. Each state $\mathcal{S}_i$ is assigned to a typical

speech sound which corresponds to a specific vector $\hat{\mathbf{a}}_{\mathrm{wb}}^{i}$ with $1 \leq i \leq N_{\mathrm{bwe}}$. In a training phase, a *Vector Quantizer* (VQ)[1] and a large training set of wideband speech are applied in order to generate N_{bwe} different vectors $\hat{\mathbf{a}}_{\mathrm{wb}}^{i}$ such that the number of states corresponds to the number of codebook entries. In the real application, the objective is to estimate the state $\mathcal{S}_i(\lambda)$ or coefficient vector $\hat{\mathbf{a}}_{\mathrm{wb}}^{i}(\lambda)$ of the current frame λ using only the available low band speech signal $s_{\mathrm{lb}}(k')$. Therefore, an observation vector $\mathbf{x}_{\mathrm{lb}}$ is extracted from the low band signal containing parameters which should deliver maximum information about the current state $\mathcal{S}_i(\lambda)$. Possible parameters for this purpose are, e.g., the autocorrelation function, the *Zero Crossing Rate* (ZCR), the frame energy or the local kurtosis [JV04].

In the estimation process, the link between observation $\mathbf{x}_{\mathrm{lb}}$ and the state sequence (envelope vector $\mathbf{a}_{\mathrm{wb}}$) is made by an HMM. The required statistical properties between both vectors have to be measured offline in terms of the state *Probability Mass Function* (PMF) $P(\mathcal{S}_i)$, the observation *Probability Density Function* (PDF) $p(\mathbf{x}_{\mathrm{lb}}|\mathcal{S}_i)$, the emission PDF $p(\mathbf{a}_{\mathrm{wb}}|\mathcal{S}_i)$ as well as the transition PMF $P(\mathcal{S}_{i_1}(\lambda+1)|\mathcal{S}_{i_2}(\lambda))$ with $i_1, i_2 \in \{1, ..., N_{\mathrm{bwe}}\}$. While the vector codebook implicitly gives information about the emission probability, the state and transition probabilities have to be measured from the wideband speech data considering the true state transitions. The observation PDF $p(\mathbf{x}_{\mathrm{lb}}|\mathcal{S}_i)$ is approximated using *Gaussian Mixture Model* (GMM) techniques [RR95].

Let $\mathbf{X}_{\mathrm{lb}} = (\mathbf{x}_{\mathrm{lb}}(0), ..., \mathbf{x}_{\mathrm{lb}}(\lambda))$ be a sequence of observation vectors from the low band of frames 0 to λ. The final estimate of the vector $\mathbf{a}_{\mathrm{wb}}$ representing the spectral wideband envelope is derived in the *Minimum Mean Square Error* (MMSE) sense by minimizing the conditional expectation $\mathbb{E}\{||\mathbf{a}_{\mathrm{wb}} - \hat{\mathbf{a}}_{\mathrm{wb}}||^2|\mathbf{X}_{\mathrm{lb}}\}$ where $\hat{\mathbf{a}}_{\mathrm{wb}}$ is the respective estimate. The solution yields [JV03a]:

$$
\begin{aligned}
\hat{\mathbf{a}}_{\mathrm{wb}} &= \mathbb{E}\{\mathbf{a}_{\mathrm{wb}}|\mathbf{X}_{\mathrm{lb}}\} \\
&= \int_{\mathbb{R}^{b_{\mathrm{wb}}}} \mathbf{a}_{\mathrm{wb}} p\left(\mathbf{a}_{\mathrm{wb}}|\mathbf{X}_{\mathrm{lb}}\right) \mathrm{d}\mathbf{a}_{\mathrm{wb}} \\
&= \int_{\mathbb{R}^{b_{\mathrm{wb}}}} \mathbf{a}_{\mathrm{wb}} \sum_{i=1}^{N_{\mathrm{bwe}}} p\left(\mathbf{a}_{\mathrm{wb}}, \mathcal{S}_i|\mathbf{X}_{\mathrm{lb}}\right) \mathrm{d}\mathbf{a}_{\mathrm{wb}} \\
&= \int_{\mathbb{R}^{b_{\mathrm{wb}}}} \mathbf{a}_{\mathrm{wb}} \sum_{i=1}^{N_{\mathrm{bwe}}} p\left(\mathbf{a}_{\mathrm{wb}}|\mathcal{S}_i, \mathbf{x}_{\mathrm{lb}}\right) P\left(\mathcal{S}_i|\mathbf{X}_{\mathrm{lb}}\right) \mathrm{d}\mathbf{a}_{\mathrm{wb}} \\
&= \sum_{i=1}^{N_{\mathrm{bwe}}} \left(P\left(\mathcal{S}_i|\mathbf{X}_{\mathrm{lb}}\right) \int_{\mathbb{R}^{b_{\mathrm{wb}}}} \mathbf{a}_{\mathrm{wb}} p\left(\mathbf{a}_{\mathrm{wb}}|\mathcal{S}_i, \mathbf{x}_{\mathrm{lb}}\right) \mathrm{d}\mathbf{a}_{\mathrm{wb}} \right) \\
&= \sum_{i=1}^{N_{\mathrm{bwe}}} P\left(\mathcal{S}_i|\mathbf{X}_{\mathrm{lb}}\right) \hat{\mathbf{a}}_{\mathrm{wb}}^{i}, \qquad (4.3)
\end{aligned}
$$

which essentially is a weighted sum over the N_{bwe} centroids of the codebook. The PMF $P\left(\mathcal{S}_i|\mathbf{X}_{\mathrm{lb}}\right)$ can be determined using Bayes' theorem [Bay63] and the assumption

[1] The codebook entries $\hat{\mathbf{a}}_{\mathrm{wb}}^{i}$ can be generated using, e.g., the well-known LBG algorithm [LBG80].

that the HMM is of first order [JV03a]. The vector $\hat{\mathbf{a}}_{\mathrm{wb}}$ is estimated in each frame and used to build the analysis and synthesis filters according to Fig. 4.1.

In the following section, techniques of the afore presented ABWE are used in order to support the noise reduction in the high band.

4.2 Wideband Noise Reduction

Only very few publications can be found in literature so far which explicitly cover wideband noise reduction. As mentioned before, almost all known approaches process the low band signal (50 Hz – 4 kHz) and the components of the high band signal (4 – 7 kHz) in the same way. In this thesis, a new wideband (50 Hz – 7 kHz) speech enhancement system is presented that uses techniques known from ABWE in order to improve the spectral estimation process. Therefore, statistical dependencies between the low band and the high band are exploited. Conventional noise suppression is used in the low band, while a novel approach is applied to the high band. Parameters from the processed (enhanced) low band signal are extracted and used to estimate subband energies of the high band. The resulting weighting gains determined from these energy estimates are adaptively combined with conventional gains obtained in addition for the high band. Thereby, the enhanced low band signal can be re-used in order to improve the results of a conventional noise suppression technique in the high band. In the following presentation of the proposed system, a sampling frequency of $f_{\mathrm{s}} = 16\,\mathrm{kHz}$ is assumed for the noisy input signal $y(k')$. However, the approach can also be adapted to other sampling frequencies.

4.2.1 System Overview

A simplified block diagram of the proposed novel wideband speech enhancement system is depicted in Fig. 4.2. In analogy to Chapters 2 and 3, it is assumed that the noisy input signal $y(k')$ consists of the clean speech signal $s(k')$ and the additive noise signal $n(k')$ according to $y(k') = s(k') + n(k')$. In order to suppress the noise signal, different processing schemes are applied in the low band and the high band. Therefore, a 2-channel FIR *Quadrature Mirror Filter* (QMF) bank with critical sampling and perfect reconstruction is used to split the wideband signal $y(k')$ into the low band and the high band signal. The frequency responses of the FIR low-pass and high-pass filters (filter length: 64) used within the QMF bank are illustrated in Fig. 4.3. Due to the decomposition, individual analysis-synthesis structures and different algorithms can be used in each band making it possible to re-use existing low band noise reduction systems. After downsampling the low-pass and high-pass filtered signals by a factor of 2, a conventional noise reduction technique is applied to the low band signal $y_{\mathrm{lb}}(k)$. In the high band, the noisy signal $y_{\mathrm{hb}}(k)$ is enhanced by using additional information from the improved low band signal. For this, parameters from the vector $\hat{\mathbf{S}}_{\mathrm{lb}}$, consisting of the DFT coefficients from the enhanced low band signal, are extracted as will be explained in the next section.

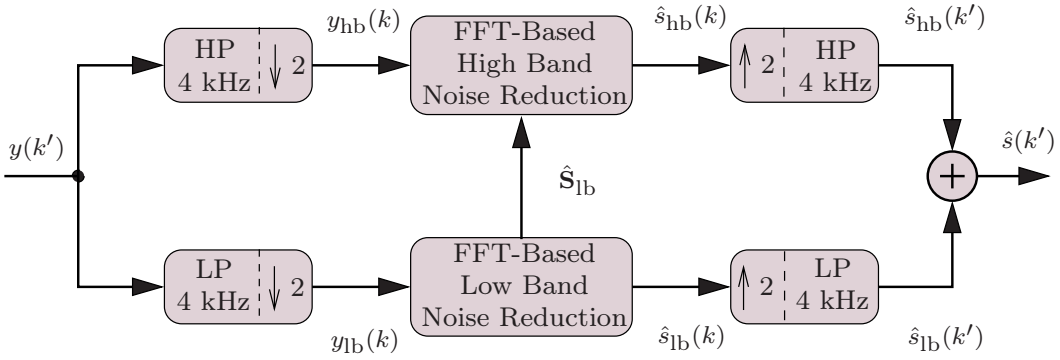

Figure 4.2: Wideband noise reduction using different techniques in low band and high band exploiting spectral dependencies.

In both bands, the noise suppression is performed in the frequency domain. Therefore, $y_{\mathrm{xx}}(k)$ is segmented into overlapping frames of length L_{F}, where the index 'xx' denotes either the low band 'lb' or the high band 'hb'. After windowing and zero-padding, the *Fast Fourier Transform* (FFT) of length M_{F} is applied to these frames. The spectral coefficients of the noisy input signal at frame λ and frequency bin μ are given by:

$$Y_{\mathrm{xx}}(\lambda, \mu) = S_{\mathrm{xx}}(\lambda, \mu) + N_{\mathrm{xx}}(\lambda, \mu), \tag{4.4}$$

where $S_{\mathrm{xx}}(\lambda, \mu)$ and $N_{\mathrm{xx}}(\lambda, \mu)$ represent the spectral DFT coefficients of the speech and the noise signal of the low band and the high band, respectively. For the sake of brevity, the frame index λ is omitted in the sequel.

The enhanced signals $\hat{s}_{\mathrm{lb}}(k)$ and $\hat{s}_{\mathrm{hb}}(k)$ are upsampled and interpolated by low-pass and high-pass filtering. Finally, both signals are added in order to obtain the enhanced wideband signal $\hat{s}(k')$.

4.2.2 Joint Noise Reduction in the High Band

The main energy of a speech signal is usually located in the frequency range between 500 Hz and 3 kHz [VM06]. Assuming that the energy of speech signals declines stronger than the energy of noise signals beyond 3 kHz, the SNR in the low band is usually significantly higher than in the high band. Table 4.1 shows some quantitative examples of how much the SNR in the low band is higher than in the high band for different speakers and different noise environments. It can be seen that in most cases the SNR significantly degrades in the high band which leads to an imprecise noise reduction and fluctuating weighting gains if solely a conventional noise suppression technique is applied at higher frequencies. To counteract this problem, a joint noise reduction method is presented in the following for the high band signal which makes use of the spectral dependencies between low band and high band.

Figure 4.4 shows the basic principle of the combined noise suppression scheme in the high band. The analysis and synthesis structures remain the same as for the low band signal. After the transformation into the frequency domain, two separate noise suppression modules are applied to the noisy high band spectrum $Y_{\mathrm{hb}}(\lambda)$ finally resulting

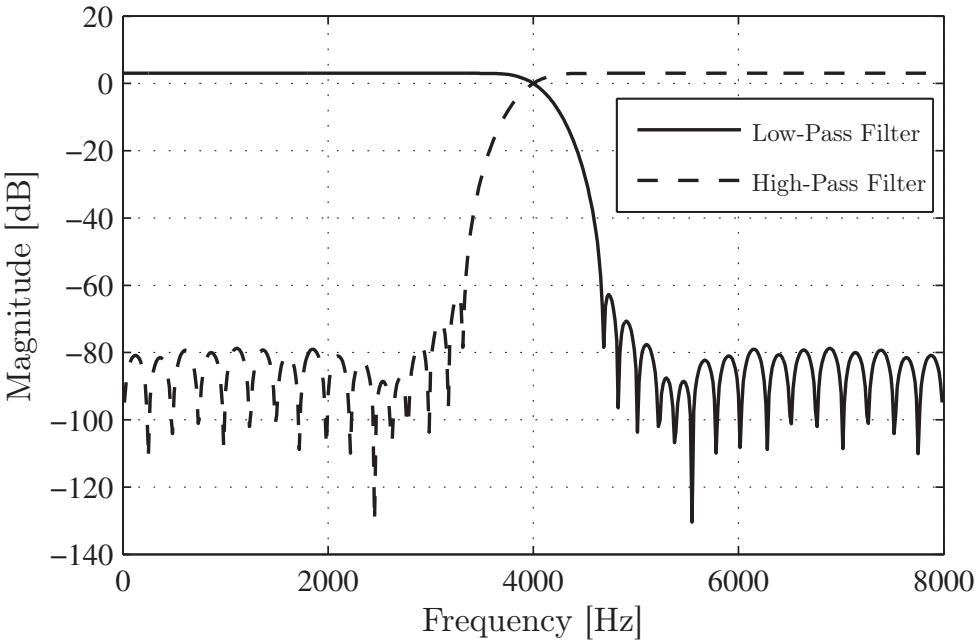

Figure 4.3: Frequency responses of applied FIR QMF bank (filter length: 64, $f_\mathrm{s} = 16\,\mathrm{kHz}$).

in the calculation of the high band weighting gains $G_\mathrm{hb}(\mu')$ where μ' represents the frequency index in a subsampled frequency domain as will be explained later.

As depicted in Fig. 4.4, a first (conventional) and a second (novel) gain calculation is performed for the high band spectrum. For the first gain calculation any 'regular' noise reduction technique can be used, e.g., the proposed Kalman filter approach of Chapter 3 or any other statistical weighting rule like [LO79], [EM84] or [LV05], including noise power estimation and SNR estimation, cf. Chapter 2. As the resulting weighting gains $G_\mathrm{hb}^\mathrm{conv}(\mu)$ exhibit a high variance over time, they are further processed in order to limit temporal fluctuations. In this post processing stage, the frequency resolution of $G_\mathrm{hb}^\mathrm{conv}(\mu)$ is decreased from M_F to M_F' by combining adjacent frequency bins to frequency bands using 50 %-overlapping Hann windows of the same length.

Noise Type	Average Deviation of the Low Band SNR from the High Band SNR for	
	Male Speakers	Female Speakers
Cockpit	+15.39 dB	+13.98 dB
Babble	+0.55 dB	-0.86 dB
Factory 1	+12.55 dB	+11.14 dB
Buccaneer	+15.64 dB	+14.23 dB
WGN	+26.81 dB	+25.39 dB

Table 4.1: Average SNR deviation of the low band from the high band for different noise types. For the measurement, six speech signals (three male and three female speakers) from the NTT database [NC94] were used. The noise signals have been taken from the NOISEX-92 database [VS93].

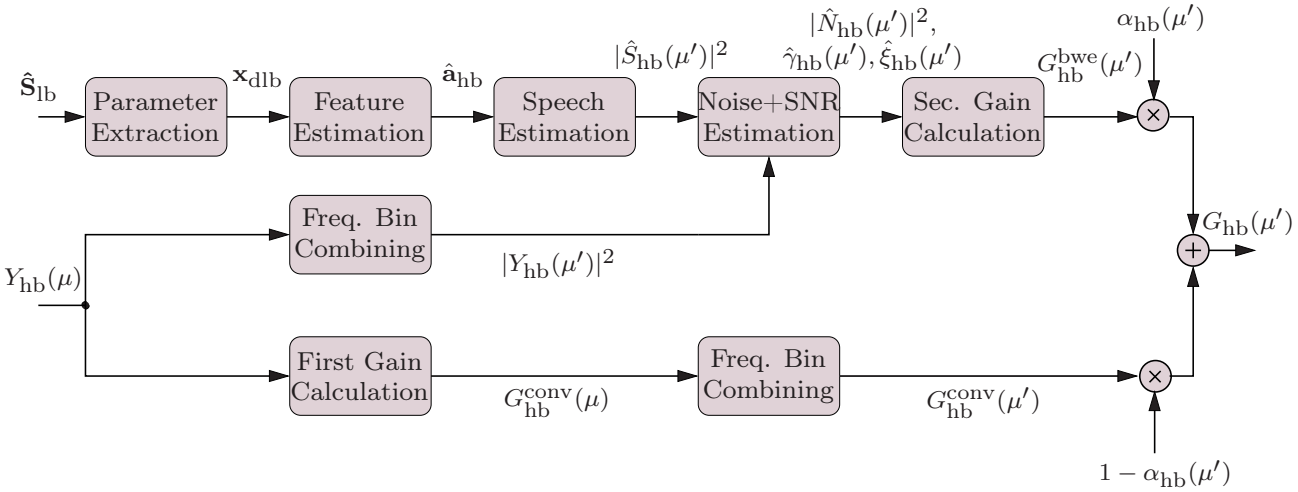

Figure 4.4: High band noise reduction exploiting spectral dependencies between low band and high band.

In the upper branch of Fig. 4.4, artificial bandwidth extension techniques are used to perform the second gain calculation (see next section for details). All required processing steps are therefore performed at the reduced frequency resolution M_{F}' as well. The resulting weighting gains $G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu')$ are adaptively combined with $G_{\mathrm{hb}}^{\mathrm{conv}}(\mu')$ according to:

$$G_{\mathrm{hb}}(\mu') = \alpha_{\mathrm{hb}}(\mu') \cdot G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu') + (1 - \alpha_{\mathrm{hb}}(\mu')) \cdot G_{\mathrm{hb}}^{\mathrm{conv}}(\mu'), \tag{4.5}$$

where $0 \leq \mu' \leq M_{\mathrm{F}}' - 1$. The cross-fading factor $\alpha_{\mathrm{hb}}(\mu') \in [0, 1]$ is frame and frequency-dependent as will be shown in Sec. 4.2.4. At the end, the frequency resolution of the high band weighting gains $G_{\mathrm{hb}}(\mu')$ is expanded back from M_{F}' to the original resolution M_{F} using overlap-add of the scaled Hann windows which have been used before. A spectral weighting of the noisy high band coefficients $Y_{\mathrm{hb}}(\mu)$ with the resulting weighting gains $G_{\mathrm{hb}}(\mu)$ yields an estimate $\hat{S}_{\mathrm{hb}}(\mu)$ of the clean high band DFT coefficients $S_{\mathrm{hb}}(\mu)$:

$$\hat{S}_{\mathrm{hb}}(\mu) = Y_{\mathrm{hb}}(\mu) \cdot G_{\mathrm{hb}}(\mu). \tag{4.6}$$

Finally, an *Inverse Fast Fourier Transform* (IFFT) and overlap-add are applied to obtain the enhanced signal $\hat{s}_{\mathrm{hb}}(k)$ in the time domain.

4.2.3 Noise Reduction Exploiting Spectral Dependencies

In order to exploit the dependencies in the frequency domain between low band and high band, techniques known from ABWE are applied for the second gain calculation in Fig. 4.4. The main principle that is used here for the ABWE is partly included in [GV07] and based on the approach which is presented in Sec. 4.1.

In this realization, the observation vector[2] $\mathbf{x}_{\mathrm{dlb}}$ from the low band consists of N_{C} *Mel-Frequency Cepstral Coefficients* (MFCCs) [RJ93] and the *Zero Crossing*

[2] In contrast to Sec. 4.1, it is assumed here that only the observation vector $\mathbf{x}_{\mathrm{dlb}}$ of the disturbed low band signal is available although noise reduction is applied. The degree of distortion is, of course, SNR-dependent.

Rate (ZCR) [RS78] of the low band signal. Both parameters provide high information on the spectral envelope of the high band signal [JV04] which is represented in Fig. 4.4 by the feature vector $\mathbf{a}_{\mathrm{hb}}$. Other parameters for the purpose of noise reduction are investigated in [HEGV10]. In contrast to the approach in Sec. 4.1, only the spectral envelope of the high band is required in the following and not the envelope of the full wideband signal as indicated by $\mathbf{a}_{\mathrm{wb}}$ in Fig. 4.1.

Given the observation vector $\mathbf{x}_{\mathrm{dlb}}$ of the low band, the estimation process described in Sec. 4.1.3 is applied in order to estimate the feature vector $\mathbf{a}_{\mathrm{hb}}$. Therefore, the HMM comprises N_{bwe} states and M_{GM} mixture components are used to approximate the observation probabilities $p(\mathbf{x}_{\mathrm{lb}}|\mathcal{S}_i)$ using GMMs.

The estimated feature vector $\hat{\mathbf{a}}_{\mathrm{hb}}$ contains the M'_{F} logarithmic subband energies of the high band signal according to:

$$\hat{\mathbf{a}}_{\mathrm{hb}} = \left(\log_{10}\left(|\hat{S}_{\mathrm{hb}}(0)|^2\right),\ \ldots,\ \log_{10}\left(|\hat{S}_{\mathrm{hb}}(M'_{\mathrm{F}}-1)|^2\right)\right), \tag{4.7}$$

where $\hat{S}_{\mathrm{hb}}(\mu')$ represents the estimated spectral speech coefficient of the μ'-th subband. Once the instantaneous energies $|\hat{S}_{\mathrm{hb}}(\mu')|^2$ of the M'_{F} subbands are determined, they are used to estimate the short-time noise energies in the high band as follows:

$$|\hat{N}_{\mathrm{hb}}(\mu')|^2 = \max\left(|Y_{\mathrm{hb}}(\mu')|^2 - |\hat{S}_{\mathrm{hb}}(\mu')|^2, 0\right). \tag{4.8}$$

Finally, the *a posteriori* SNR $\gamma_{\mathrm{hb}}(\mu')$ and *a priori* SNR $\xi_{\mathrm{hb}}(\mu')$ of the high band can be estimated according to:

$$\hat{\gamma}_{\mathrm{hb}}(\mu') = \frac{|Y_{\mathrm{hb}}(\mu')|^2}{|\hat{N}_{\mathrm{hb}}(\mu')|^2} \quad \text{and} \quad \hat{\xi}_{\mathrm{hb}}(\mu') = \frac{|\hat{S}_{\mathrm{hb}}(\mu')|^2}{|\hat{N}_{\mathrm{hb}}(\mu')|^2}. \tag{4.9}$$

Both SNR quantities are required in order to calculate the weighting gains $G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu')$ based on a conventional noise reduction algorithm.

4.2.4 Cross-Fading Factor

As mentioned before, the two weighting gains $G_{\mathrm{hb}}^{\mathrm{conv}}(\mu')$ and $G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu')$ are adaptively combined using the cross-fading factor $\alpha_{\mathrm{hb}}(\mu')$, cf. Eq. 4.5. In the following, the *reference* cross-fading factor $\alpha_{\mathrm{hb}}^{\mathrm{ref}}(\mu')$ is defined as:

$$\alpha_{\mathrm{hb}}^{\mathrm{ref}}(\mu') = \frac{\left(G_{\mathrm{hb}}^{\mathrm{opt}}(\mu') - G_{\mathrm{hb}}^{\mathrm{conv}}(\mu')\right)^2}{\left(G_{\mathrm{hb}}^{\mathrm{opt}}(\mu') - G_{\mathrm{hb}}^{\mathrm{conv}}(\mu')\right)^2 + \left(G_{\mathrm{hb}}^{\mathrm{opt}}(\mu') - G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu')\right)^2}, \tag{4.10}$$

where $G_{\mathrm{hb}}^{\mathrm{opt}}(\mu')$ represents the *optimum* weighting gain which could in theory (or by a dedicated simulation setup) be determined from the clean speech and noise signal according to the *optimum* a posteriori SNR $\gamma_{\mathrm{hb}}^{\mathrm{opt}}(\mu')$ and *optimum* a priori SNR $\xi_{\mathrm{hb}}^{\mathrm{opt}}(\mu')$:

$$\gamma_{\mathrm{hb}}^{\mathrm{opt}}(\mu') = \frac{|Y_{\mathrm{hb}}(\mu')|^2}{|N_{\mathrm{hb}}(\mu')|^2} \quad \text{and} \quad \xi_{\mathrm{hb}}^{\mathrm{opt}}(\mu') = \frac{|S_{\mathrm{hb}}(\mu')|^2}{|N_{\mathrm{hb}}(\mu')|^2}, \tag{4.11}$$

which are also determined at the reduced frequency resolution M'_{F} by combining adjacent frequency bins as described before. If the conventional (first) noise suppression technique in Fig. 4.4 performs better than the ABWE approach, i.e., if $\left(G_{\mathrm{hb}}^{\mathrm{opt}} - G_{\mathrm{hb}}^{\mathrm{conv}}\right)^2 < \left(G_{\mathrm{hb}}^{\mathrm{opt}} - G_{\mathrm{hb}}^{\mathrm{bwe}}\right)^2$, $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ in Eq. 4.10 tends to smaller values leading to a stronger weighting of $G_{\mathrm{hb}}^{\mathrm{conv}}$ in Eq. 4.5 and vice versa. Moreover, the extreme cases are correctly mapped as follows:

$$\begin{aligned} G_{\mathrm{hb}}^{\mathrm{opt}} - G_{\mathrm{hb}}^{\mathrm{conv}} = 0 \quad &\Rightarrow \quad \alpha_{\mathrm{hb}}^{\mathrm{ref}} = 0 \\ G_{\mathrm{hb}}^{\mathrm{opt}} - G_{\mathrm{hb}}^{\mathrm{bwe}} = 0 \quad &\Rightarrow \quad \alpha_{\mathrm{hb}}^{\mathrm{ref}} = 1 \\ G_{\mathrm{hb}}^{\mathrm{opt}} - G_{\mathrm{hb}}^{\mathrm{conv}} = G_{\mathrm{hb}}^{\mathrm{opt}} - G_{\mathrm{hb}}^{\mathrm{bwe}} \quad &\Rightarrow \quad \alpha_{\mathrm{hb}}^{\mathrm{ref}} = 0.5 \end{aligned}$$

In a realistic scenario $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ can not be applied as $G_{\mathrm{hb}}^{\mathrm{opt}}$ is not available. In order to estimate the reference cross-fading factor, first $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ is recorded in an offline training process for every frame λ and every subband μ' together with the respective subband SNR $\xi_{\mathrm{hb}}^{\mathrm{opt}}(\mu')$ of the high band and the averaged SNR $\bar{\xi}_{\mathrm{lb}}^{\mathrm{opt}}$ of the low band:

$$\bar{\xi}_{\mathrm{lb}}^{\mathrm{opt}} = \frac{1}{M_{\mathrm{F}}} \sum_{\mu=0}^{M_{\mathrm{F}}-1} \frac{|S_{\mathrm{lb}}(\mu)|^2}{|N_{\mathrm{lb}}(\mu)|^2}. \tag{4.12}$$

Based on representative training data, a look-up table for the estimation of $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ is generated for each subband. Therefore, $\xi_{\mathrm{hb}}^{\mathrm{opt}}(\mu')$ and $\bar{\xi}_{\mathrm{lb}}^{\mathrm{opt}}$ are quantized (e.g., 1 dB step size) and the associated values for $\alpha_{\mathrm{hb}}^{\mathrm{ref}}(\mu')$ are averaged within the quantization levels. At the end, a final look-up table provides one value $\bar{\alpha}_{\mathrm{hb}}(\mu')$ for each quantized combination of $\xi_{\mathrm{hb}}^{\mathrm{opt}}(\mu')$ and $\bar{\xi}_{\mathrm{lb}}^{\mathrm{opt}}$. A typical example of this two-dimensional look-up table can be seen in Fig. 4.5. The figure demonstrates a strong correlation between the averaged factor $\bar{\alpha}_{\mathrm{hb}}$ and the two SNR quantities showing that the ABWE approach in Eq. 4.5 is preferred with a decreasing high band SNR. Moreover, in the high band SNR range $-15\,\mathrm{dB} \leq \xi_{\mathrm{hb}} \leq 0\,\mathrm{dB}$, it can be seen that the cross-fading factor $\bar{\alpha}_{\mathrm{hb}}$ becomes larger for higher low band SNR values $\bar{\xi}_{\mathrm{lb}}$ showing that the ABWE (trained with clean speech) performs better the higher the input SNR is in the low band.

In a real application, $\xi_{\mathrm{hb}}^{\mathrm{opt}}$ and $\bar{\xi}_{\mathrm{lb}}^{\mathrm{opt}}$ are not available to estimate the reference cross-fading factor $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$. Here, the respective SNR estimates of the conventional noise suppression technique in the low band and high band are utilized to determine $\bar{\alpha}_{\mathrm{lb}}(\mu')$ using M'_{F} different pre-trained look-up tables.

The look-up tables implicitly show the existence of spectral dependencies between low band and high band in noisy environments as is apparent from the different colored areas in Fig. 4.5. However, in order to quantify the amount of mutual information and to investigate the benefit of noise suppression before determining the observation vector $\mathbf{x}_{\mathrm{dlb}}$, the subject is approached from an information theoretic point of view in the following.

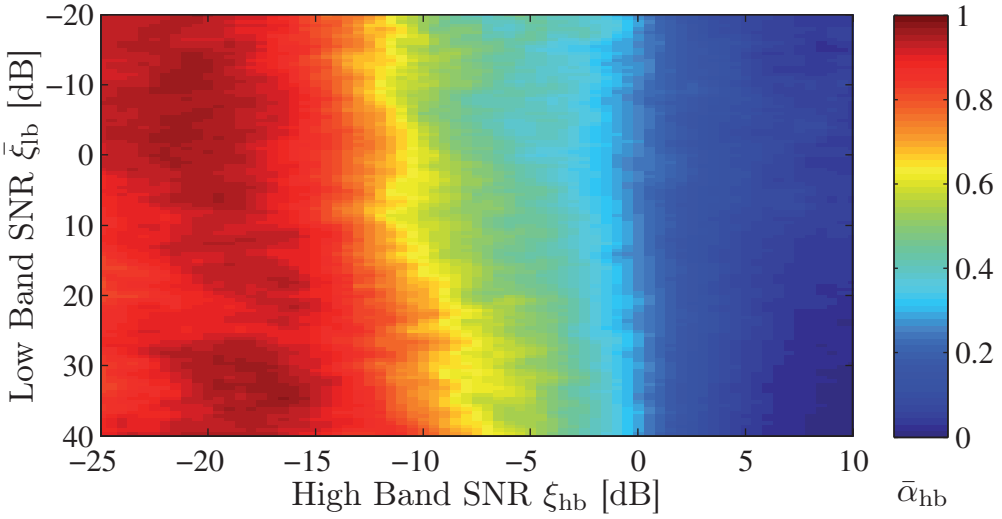

Figure 4.5: Visualization example of look-up table used to estimate $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ in subband $\mu' = 0$ using $M'_{\mathrm{F}} = 24$.

4.3 Mutual Information in Noisy Environments

In this section, the mutual information $I(\mathbf{x}_{\mathrm{lb}}; \mathbf{a}_{\mathrm{hb}})$ between the observation vector $\mathbf{x}_{\mathrm{lb}}$ from the clean low band signal and the feature vector $\mathbf{a}_{\mathrm{hb}}$ from the high band signal is analyzed and a bound for the estimation performance in noisy environments is formulated. The mutual information $I(\mathbf{x}_{\mathrm{lb}}; \mathbf{a}_{\mathrm{hb}})$ describes the linear and non-linear dependencies between $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{a}_{\mathrm{hb}}$, i.e., a high mutual information between the two vectors is desirable in order to obtain a good estimate of $\mathbf{a}_{\mathrm{hb}}$. In the sequel, real measurements show that the mutual information between the low and high bands is significantly higher if noise reduction is applied to the disturbed low band signal prior to ABWE.

The mutual information between frequency bands in speech is examined, e.g., in [NEH00], [NGAK02] and [JV02]. Following the approach of [Ber98], an upper bound on the quality of ABWE techniques is derived in [JV02] for the case that the clean low band signal is available. However, in a realistic scenario, it is likely that the speech signal is disturbed by ambient noise. Therefore, the derivation of the bound in [JV02] is extended in this thesis considering the signal-flow model shown in Fig. 4.6a). The b_{lb}-dimensional observation vector $\mathbf{x}_{\mathrm{lb}}$ is assumed to be degraded by the additive noise vector $\mathbf{n}_{\mathrm{lb}}$ of dimension b_{lb} as well. Afterwards, ABWE is performed based on the resulting noisy observation vector $\mathbf{x}_{\mathrm{dlb}}$. The ABWE process is described by the function $f_{\mathrm{bwe}}(\cdot)$, yielding the estimated feature vector $\hat{\mathbf{a}}_{\mathrm{hb}}$ of dimension b_{hb} according to:

$$\hat{\mathbf{a}}_{\mathrm{hb}} = f_{\mathrm{bwe}}\left(\mathbf{x}_{\mathrm{dlb}}\right). \tag{4.13}$$

The estimation error of the ABWE process is defined as b_{hb}-dimensional vector $\mathbf{n}_{\mathrm{hb}}$ and states the difference between $\mathbf{a}_{\mathrm{hb}}$ and $\hat{\mathbf{a}}_{\mathrm{hb}}$:

$$\mathbf{n}_{\mathrm{hb}} = \mathbf{a}_{\mathrm{hb}} - \hat{\mathbf{a}}_{\mathrm{hb}}. \tag{4.14}$$

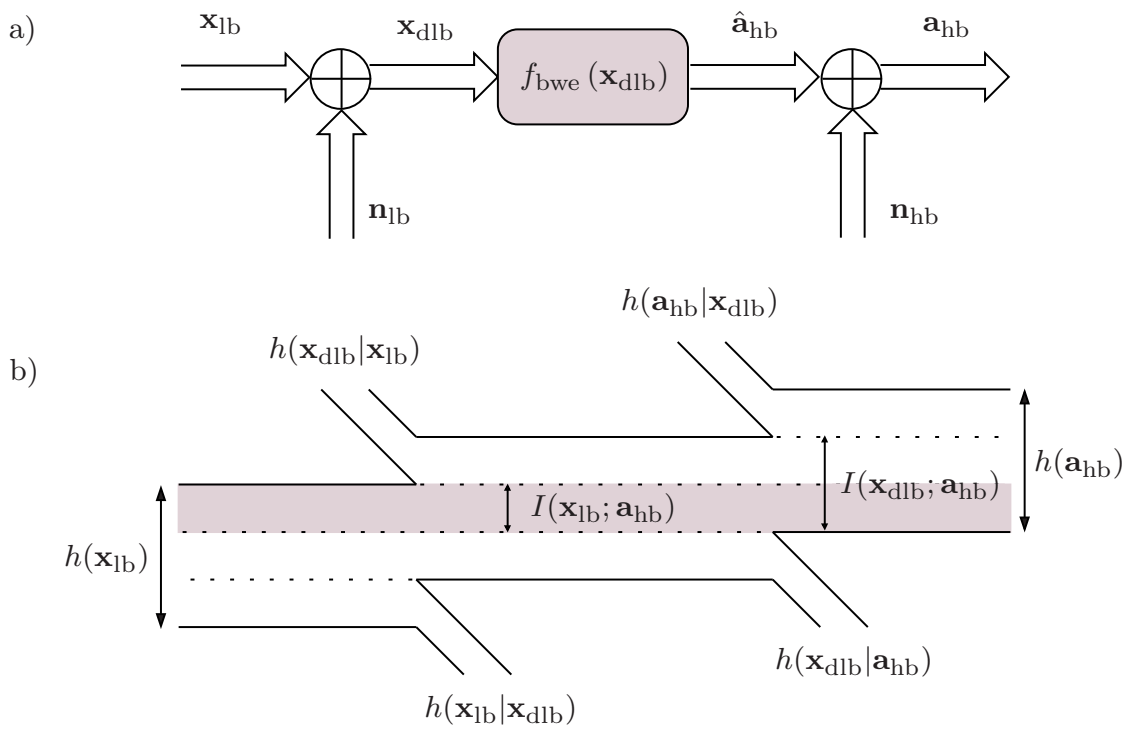

Figure 4.6: Artificial bandwidth extension in noisy environments based on two memoryless channels: a) signal-flow model and b) mutual information between $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{a}_{\mathrm{hb}}$.

In the following, a lower bound for the mutual information between $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{a}_{\mathrm{hb}}$ is derived. Following the concept in [JV02], it is assumed that the ABWE system uses a memoryless estimator $f_{\mathrm{bwe}}(\cdot)$ which is not relying on information from previous or subsequent frames.

4.3.1 Performance Bound

In order to derive a lower bound for the mutual information $I(\mathbf{x}_{\mathrm{lb}};\mathbf{a}_{\mathrm{hb}})$, the process of the disturbances by $\mathbf{n}_{\mathrm{lb}}$ and $\mathbf{n}_{\mathrm{hb}}$ including the ABWE estimation is modeled by two independent memoryless, additive noisy channels. The resulting information theoretic dependencies between $\mathbf{x}_{\mathrm{lb}}$, $\mathbf{x}_{\mathrm{dlb}}$ and $\mathbf{a}_{\mathrm{hb}}$ are depicted in Fig. 4.6b) based on (conditional) differential entropies $h(\cdot)$. The mutual information $I(\mathbf{x}_{\mathrm{lb}};\mathbf{a}_{\mathrm{hb}})$ between $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{a}_{\mathrm{hb}}$ can be expressed as:

$$\begin{aligned} I(\mathbf{x}_{\mathrm{lb}};\mathbf{a}_{\mathrm{hb}}) &= I(\mathbf{x}_{\mathrm{dlb}};\mathbf{a}_{\mathrm{hb}}) - h(\mathbf{x}_{\mathrm{dlb}}|\mathbf{x}_{\mathrm{lb}}) \\ &= h(\mathbf{a}_{\mathrm{hb}}) - h(\mathbf{a}_{\mathrm{hb}}|\mathbf{x}_{\mathrm{dlb}}) - h(\mathbf{x}_{\mathrm{dlb}}|\mathbf{x}_{\mathrm{lb}}), \end{aligned} \tag{4.15}$$

where $I(\mathbf{x}_{\mathrm{dlb}};\mathbf{a}_{\mathrm{hb}})$ represents the mutual information between $\mathbf{x}_{\mathrm{dlb}}$ and $\mathbf{a}_{\mathrm{hb}}$, $h(\mathbf{a}_{\mathrm{hb}})$ the differential entropy of $\mathbf{a}_{\mathrm{hb}}$, $h(\mathbf{a}_{\mathrm{hb}}|\mathbf{x}_{\mathrm{dlb}})$ the conditional differential entropy of $\mathbf{a}_{\mathrm{hb}}$ when $\mathbf{x}_{\mathrm{dlb}}$ is given and $h(\mathbf{x}_{\mathrm{dlb}}|\mathbf{x}_{\mathrm{lb}})$ the conditional differential entropy of $\mathbf{x}_{\mathrm{dlb}}$ when

$\mathbf{x}_{\mathrm{lb}}$ is known. The latter can further be simplified to [CT06]:

$$\begin{aligned} h(\mathbf{x}_{\mathrm{dlb}}|\mathbf{x}_{\mathrm{lb}}) &= h(\mathbf{x}_{\mathrm{lb}} + \mathbf{n}_{\mathrm{lb}}|\mathbf{x}_{\mathrm{lb}}) \\ &= h(\mathbf{n}_{\mathrm{lb}}|\mathbf{x}_{\mathrm{lb}}) \\ &= h(\mathbf{n}_{\mathrm{lb}}). \end{aligned} \tag{4.16}$$

The disturbance of the observation vector $\mathbf{x}_{\mathrm{lb}}$ by $\mathbf{n}_{\mathrm{lb}}$ can be interpreted as transmission over b_{lb} different channels in parallel. Assuming a fixed variance for the observation vector, an upper bound for the entropy $h(\mathbf{n}_{\mathrm{lb}})$ is given for the case that all channels are statistically independent[(a)] *Additive White Gaussian Noise* (AWGN)[(b)] channels with same variances[(c)] $\sigma^2_{n_{\mathrm{lb}},j} = \sigma^2_{n_{\mathrm{lb}}}$ for $0 \leq j < b_{\mathrm{lb}}$ [CT06]:

$$\begin{aligned} h(\mathbf{n}_{\mathrm{lb}}) &= h(n_{\mathrm{lb},0},\, n_{\mathrm{lb},1},\, \ldots,\, n_{\mathrm{lb},b_{\mathrm{lb}}-1}) \\ &\overset{(a)}{=} \sum_{j=0}^{b_{\mathrm{lb}}-1} h(n_{\mathrm{lb},j}) \\ &\overset{(b)}{\leq} \sum_{j=0}^{b_{\mathrm{lb}}-1} \frac{1}{2} \log_2(2\pi \mathrm{e} \sigma^2_{n_{\mathrm{lb}},j}) \\ &\overset{(c)}{=} b_{\mathrm{lb}} \log_2 \left(\sqrt{2\pi \mathrm{e} \sigma^2_{n_{\mathrm{lb}}}} \right), \end{aligned} \tag{4.17}$$

where e states the Euler number. The unit of the (differential) entropies and mutual information in this thesis is *bits/vector*. Under the assumption that the observation vector $\mathbf{x}_{\mathrm{lb}}$ consists of the first b_{lb} *cepstral coefficients*[3] $c_{\mathrm{lb},j}$ [MGj76] of the clean low band signal $s_{\mathrm{lb}}(k)$, i.e.:

$$x_{\mathrm{lb,j}} = \begin{cases} \frac{1}{\sqrt{2}} c_{\mathrm{lb},0} & \text{for } j = 0, \\ c_{\mathrm{lb},j} & \text{for } 1 \leq j < b_{\mathrm{lb}}, \end{cases} \tag{4.18}$$

the noise vector $\mathbf{n}_{\mathrm{lb}}$ is related to the *Log Spectral Distortion* (LSD) between the spectral envelopes of the clean low band signal $s_{\mathrm{lb}}(k)$ and the disturbed low band signal $y_{\mathrm{lb}}(k)$, represented by the vectors $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{x}_{\mathrm{dlb}}$, respectively. The LSD measure $d_{\mathrm{lb}}^{\mathrm{LSD}}$ correlates well with the subjective speech quality and is defined as in [JV02]:

$$d_{\mathrm{lb}}^{\mathrm{LSD}} = \frac{\sqrt{2} \cdot 10}{\log_{\mathrm{e}}(10)} \sqrt{\mathbb{E}\left\{ \frac{1}{2} \left(c_{\mathrm{lb},0} - c_{\mathrm{dlb},0} \right)^2 + \sum_{j=1}^{\infty} \left(c_{\mathrm{lb},j} - c_{\mathrm{dlb},j} \right)^2 \right\}}, \tag{4.19}$$

where $c_{\mathrm{dlb},j}$ represents the j-th cepstral coefficient of y_{lb} included in $\mathbf{x}_{\mathrm{dlb}}$. The unit of $d_{\mathrm{lb}}^{\mathrm{LSD}}$ is dB. A lower bound for $d_{\mathrm{lb}}^{\mathrm{LSD}}$ is given by:

$$d_{\mathrm{lb}}^{\mathrm{LSD}} \geq \frac{\sqrt{2} \cdot 10}{\log_{\mathrm{e}}(10)} \sqrt{\mathbb{E}\left\{ |\mathbf{n}_{\mathrm{lb}}|^2 \right\}} \geq \frac{\sqrt{2} \cdot 10}{\log_{\mathrm{e}}(10)} \sqrt{b_{\mathrm{lb}} \cdot \sigma^2_{n_{\mathrm{lb}}}}, \tag{4.20}$$

[3] It is shown in [JV04] that cepstral coefficients also provide high information on the spectral envelope of the missing frequency band.

which can be re-arranged to:

$$\sqrt{\sigma_{n_{\mathrm{lb}}}^2} \leq \frac{d_{\mathrm{lb}}^{\mathrm{LSD}} \cdot \log_{\mathrm{e}}(10)}{\sqrt{2} \cdot 10 \cdot \sqrt{b_{\mathrm{lb}}}}. \tag{4.21}$$

Using Ineqs. 4.21 and 4.17 and Eq. 4.16, the mutual information $I(\mathbf{x}_{\mathrm{lb}}; \mathbf{a}_{\mathrm{hb}})$ in Eq. 4.15 is bounded by:

$$I(\mathbf{x}_{\mathrm{lb}}; \mathbf{a}_{\mathrm{hb}}) \geq \underbrace{h(\mathbf{a}_{\mathrm{hb}}) - h(\mathbf{a}_{\mathrm{hb}}|\mathbf{x}_{\mathrm{dlb}})}_{I(\mathbf{x}_{\mathrm{dlb}};\mathbf{a}_{\mathrm{hb}})} - b_{\mathrm{lb}} \log_2 \left(\frac{\sqrt{\pi \mathrm{e}} \log_{\mathrm{e}}(10)}{10 \cdot \sqrt{b_{\mathrm{lb}}}} d_{\mathrm{lb}}^{\mathrm{LSD}} \right). \tag{4.22}$$

The mutual information $I(\mathbf{x}_{\mathrm{dlb}}; \mathbf{a}_{\mathrm{hb}})$ incorporates estimation errors of the artificial bandwidth extension which are represented in Fig. 4.6 by $\mathbf{n}_{\mathrm{hb}}$. An expression for $I(\mathbf{x}_{\mathrm{dlb}}; \mathbf{a}_{\mathrm{hb}})$ is derived in [JV02] using very similar calculus as above. Assuming that the upper frequency band is represented in $\mathbf{a}_{\mathrm{hb}}$ by cepstral coefficients of the high band as well, the conditional entropy $h(\mathbf{a}_{\mathrm{hb}}|\mathbf{x}_{\mathrm{dlb}})$ is also bounded by the LSD $d_{\mathrm{hb}}^{\mathrm{LSD}}$ of the high band according to:

$$h(\mathbf{a}_{\mathrm{hb}}|\mathbf{x}_{\mathrm{dlb}}) = h(\mathbf{n}_{\mathrm{hb}}) \leq b_{\mathrm{hb}} \log_2 \left(\frac{\sqrt{\pi \mathrm{e}} \log_{\mathrm{e}}(10)}{10 \cdot \sqrt{b_{\mathrm{hb}}}} d_{\mathrm{hb}}^{\mathrm{LSD}} \right), \tag{4.23}$$

finally leading to the following lower bound for the mutual information between $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{a}_{\mathrm{hb}}$:

$$I(\mathbf{x}_{\mathrm{lb}}; \mathbf{a}_{\mathrm{hb}}) \geq h(\mathbf{a}_{\mathrm{hb}}) - b_{\mathrm{hb}} \log_2 \left(\frac{\sqrt{\pi \mathrm{e}} \log_{\mathrm{e}}(10)}{10 \cdot \sqrt{b_{\mathrm{hb}}}} d_{\mathrm{hb}}^{\mathrm{LSD}} \right) - b_{\mathrm{lb}} \log_2 \left(\frac{\sqrt{\pi \mathrm{e}} \log_{\mathrm{e}}(10)}{10 \cdot \sqrt{b_{\mathrm{lb}}}} d_{\mathrm{lb}}^{\mathrm{LSD}} \right). \tag{4.24}$$

Estimation errors occurring in the ABWE process are considered in the first subtrahend and the information loss due to disturbances of the low band signal is expressed by the second subtrahend. Knowing the distortion caused by a specific ABWE estimator as well as the degradation of the narrowband signal, the mutual information which is at least included in $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{a}_{\mathrm{hb}}$ is given by Ineq. 4.24.

In the following, two scenarios are considered in which ABWE in noisy environments is analyzed based on real entropy measurements.

4.3.2 Measurements

In this section, the performance bound is evaluated based on real speech data. In addition, the influence of low band noise reduction prior to ABWE is investigated. Therefore, the proposed Kalman filter solution based on the SNR-dependent MMSE estimator (see Sec. 3.2.3.3) is applied to the noisy input signal $y_{\mathrm{lb}}(k)$ yielding the speech estimate $\hat{s}_{\mathrm{lb}}(k)$. In order to compare the cases before and after noise suppression, the observation vector $\mathbf{x}_{\mathrm{dlb}}$ is extracted from both, the noisy low band signal $y_{\mathrm{lb}}(k)$ and the enhanced signal $\hat{s}_{\mathrm{lb}}(k)$.

Two scenarios are considered in the following characterized by the choice of parameters used for the observation and feature vectors. In Scenario I, the theoretical bound of Ineq. 4.24 is analyzed according to the measured differential entropy $h(\mathbf{a}_{\mathrm{hb}}^{\mathrm{I}})$ and the LSD measures $d_{\mathrm{lb}}^{\mathrm{LSD}}$ and $d_{\mathrm{hb}}^{\mathrm{LSD}}$. With respect to the derivation of this lower bound, cepstral coefficients are chosen as parameters for the low band as well as for the high band according to:

Scenario I	Parameter(s)	Dimension
Observation vector $\mathbf{x}_{\mathrm{lb}}^{\mathrm{I}}$	Cepstral coefficients	$b_{\mathrm{lb}} = 10$
Observation vector $\mathbf{x}_{\mathrm{dlb}}^{\mathrm{I}}$	Cepstral coefficients	$b_{\mathrm{lb}} = 10$
Feature vector $\mathbf{a}_{\mathrm{hb}}^{\mathrm{I}}$	Cepstral coefficients	$b_{\mathrm{hb}} = 10$

In Scenario II, the effective mutual information $I(\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}}; \mathbf{a}_{\mathrm{hb}}^{\mathrm{II}})$ which can be achieved in the presented wideband noise suppression system of Sec. 4.2 by ABWE is investigated and its dependency on the input SNR is shown. In order to get results which can be transferred to the novel system, the same parameters and dimensions as used in the final implementation of the proposed system are applied for the observation and feature vectors as follows:

Scenario II	Parameter(s)	Dimension
Observation vector $\mathbf{x}_{\mathrm{lb}}^{\mathrm{II}}$	Mel frequency cepstral coefficients + zero crossing rate	$b_{\mathrm{lb}} = 14$
Observation vector $\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}}$	Mel frequency cepstral coefficients + zero crossing rate	$b_{\mathrm{lb}} = 14$
Feature vector $\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}}$	Logarithmic subband energies	$b_{\mathrm{hb}} = 12$

The measurements of $h(\mathbf{a}_{\mathrm{hb}}^{\mathrm{I}})$, $h(\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}})$ and $I(\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}}; \mathbf{a}_{\mathrm{hb}}^{\mathrm{II}})$, which are required in Scenarios I and II are carried out by using the well-known *k-nearest neighbor* algorithm [KL87] to estimate the necessary PDFs [CT06]. This algorithm is data efficient, adaptive and achieves minimal bias. The number k here decides how many neighbors influence the final classification. In the sequel, the results for the two scenarios are presented using $k = 1$.

For the evaluation, about 16 minutes of speech taken from the NTT speech database [NC94] (sampling frequency f_{s} =16 kHz) are disturbed by additive *White Gaussian Noise* (WGN)[4] at different SNR values. The wideband signals are split into the respective low band and high band part using the FIR filters of Fig. 4.3 before the signals are downsampled by a factor of 2. In the simulation setup, the speech and noise signals are both available and can therefore be filtered and downsampled

[4]Similar results can be achieved with other noise signals.

Log spectral distortion $\hat{d}_{\text{lb}}^{\text{LSD}}$	Input SNR		
	-15 dB	0 dB	15 dB
Without noise reduction	4.627 dB	2.529 dB	1.083 dB
With noise reduction	2.047 dB	1.115 dB	0.944 dB

Table 4.2: Averaged log spectral distortion measures $\hat{d}_{\text{lb}}^{\text{LSD}}$ of disturbed low band signal before and after noise suppression.

separately. Hence, clean and noisy versions of low band as well as high band signals are accessible. Based on these signals, the observation vectors $\mathbf{x}_{\text{lb}}$ and $\mathbf{x}_{\text{dlb}}$ as well as the feature vector $\mathbf{a}_{\text{hb}}$ are extracted based on 20 ms non-overlapping frames contributing to speech activity[5]. In total, the investigation in this section relies on about 21000 speech frames.

Scenario I (Narrowband Noise Reduction)

In order to apply Ineq. 4.24, this scenario considers the use of cepstral coefficients as parameters within the observation and feature vectors. The LSD $d_{\text{lb}}^{\text{LSD}}$ of the low band as well as the differential entropy $h(\mathbf{a}_{\text{hb}}^{\text{I}})$ are determined from real data. Therefore, the low band speech signal $s_{\text{lb}}(k)$ is disturbed by additive WGN at input SNR values of -15 dB, 0 dB and 15 dB. It has to be mentioned that the disturbance of $s_{\text{lb}}(k)$ by additive WGN does not necessarily mean that the elements of the vector $\mathbf{n}_{\text{lb}}$ are Gaussian distributed as well. However, in any case, Ineq. 4.24 is valid as the entropy of a scalar signal with variance $\sigma_{n_{\text{lb}}}^2$ is upper bounded by the entropy of a normally distributed variable with the same variance [CT06]. Table 4.2 shows the averaged LSD values $\hat{d}_{\text{lb}}^{\text{LSD}}$ measured before and after noise suppression is applied[6]. Moreover, the differential entropy $h(\mathbf{a}_{\text{hb}}^{\text{I}})$ is measured and yields $h(\mathbf{a}_{\text{hb}}^{\text{I}}) \approx 9.533$ bits/vector in this setup. Based on these measurements, Fig. 4.7 depicts the theoretical lower bounds for the mutual information $I(\mathbf{x}_{\text{lb}}^{\text{I}}; \mathbf{a}_{\text{hb}}^{\text{I}})$ according to Ineq. 4.24 while varying $d_{\text{hb}}^{\text{LSD}}$ from 1 dB to 6 dB[7]. The figure clearly shows the dependency of $I(\mathbf{x}_{\text{lb}}^{\text{I}}; \mathbf{a}_{\text{hb}}^{\text{I}})$ on the two distortion measures: the lower $\hat{d}_{\text{lb}}^{\text{LSD}}$ or $d_{\text{hb}}^{\text{LSD}}$ the higher the theoretical bound and vice versa. In addition, the advantages of applying noise suppression prior to ABWE can be seen. According to Tab. 4.2, noise reduction achieves a reduction of the LSD measure $\hat{d}_{\text{lb}}^{\text{LSD}}$ leading to a higher bound $I(\mathbf{x}_{\text{lb}}^{\text{I}}; \mathbf{a}_{\text{hb}}^{\text{I}})$ compared to the case when no noise reduction is applied. Figure 4.7 illustrates this behavior for the three investigated SNR values where the discrepancy is especially high for -15 dB input SNR.

[5] A simple power constrained threshold is applied to the clean speech signal for *Voice Activity Detection* (VAD).

[6] Please note that $\hat{d}_{\text{lb}}^{\text{LSD}}$ is only an estimate of $d_{\text{lb}}^{\text{LSD}}$ as the sum in Eq. 4.19 is truncated after b_{lb} elements.

[7] According to the literature, speech quality can be classified as 'transparent' for LSD values less than 1 dB [JV02]. Please note that, in general, $d_{\text{hb}}^{\text{LSD}}$ can not be adjusted for a specific ABWE realization.

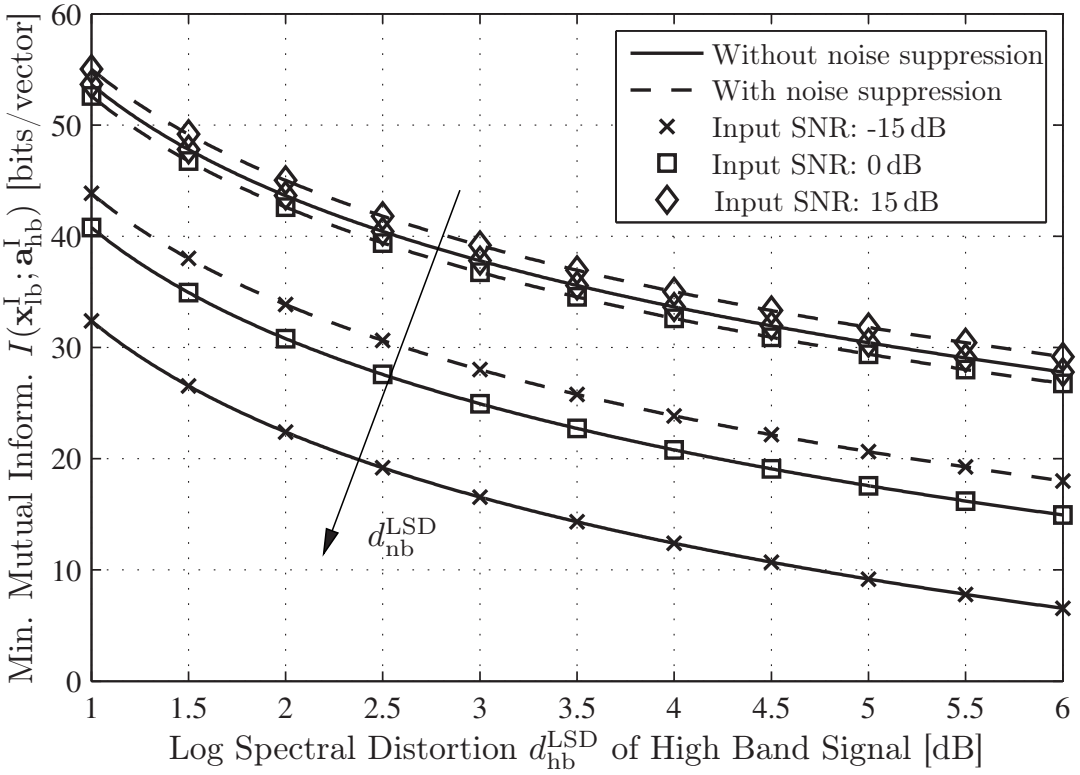

Figure 4.7: Theoretical lower bound according to Ineq. 4.24 for the mutual information $I(\mathbf{x}_{\mathrm{lb}}^{\mathrm{I}};\mathbf{a}_{\mathrm{hb}}^{\mathrm{I}})$ between high and low band vectors depending on the log spectral distortions $d_{\mathrm{lb}}^{\mathrm{LSD}}$ and $d_{\mathrm{hb}}^{\mathrm{LSD}}$.

Scenario II (Wideband Noise Reduction Supported by ABWE)

In this scenario, the reader's attention is drawn back to the proposed wideband speech enhancement system which is presented in Sec. 4.2. The system exploits mutual information between the enhanced low band signal $\hat{s}_{\mathrm{lb}}(k)$ and the high band speech signal $s_{\mathrm{hb}}(k)$ in order to support a conventional high band noise suppression by ABWE techniques. It seems obvious that the mutual information between the disturbed observation vector $\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}}$, which is available in the system, and the feature vector $\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}}$ is depending on the degree of distortion of $\hat{s}_{\mathrm{lb}}(k)$. Although this fact is implicitly already incorporated in the look-up tables (cf. Fig. 4.5), which are used to determine the cross-fading factors $\bar{\alpha}_{\mathrm{hb}}$, it is also confirmed by real mutual information measurements in this scenario. Therefore, the mutual information $I(\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}};\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}})$ is measured in dependence of the input SNR and whether noise suppression is applied to the low band signal or not. For the evaluation purpose, the feature vector $\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}}$ is extracted directly from the extension band of the original wideband signal and not from the estimated signal after ABWE. The results are shown in Fig. 4.8 for input SNR values varying from -20 dB to 35 dB. It can be seen that the mutual information is continuously increasing with the input SNR and finally converges to $I(\mathbf{x}_{\mathrm{lb}}^{\mathrm{II}};\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}})$, i.e., the case where $\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}} = \mathbf{x}_{\mathrm{lb}}^{\mathrm{II}}$ or $\mathbf{n}_{\mathrm{lb}} = \mathbf{0}$. In addition, the figure also motivates the application

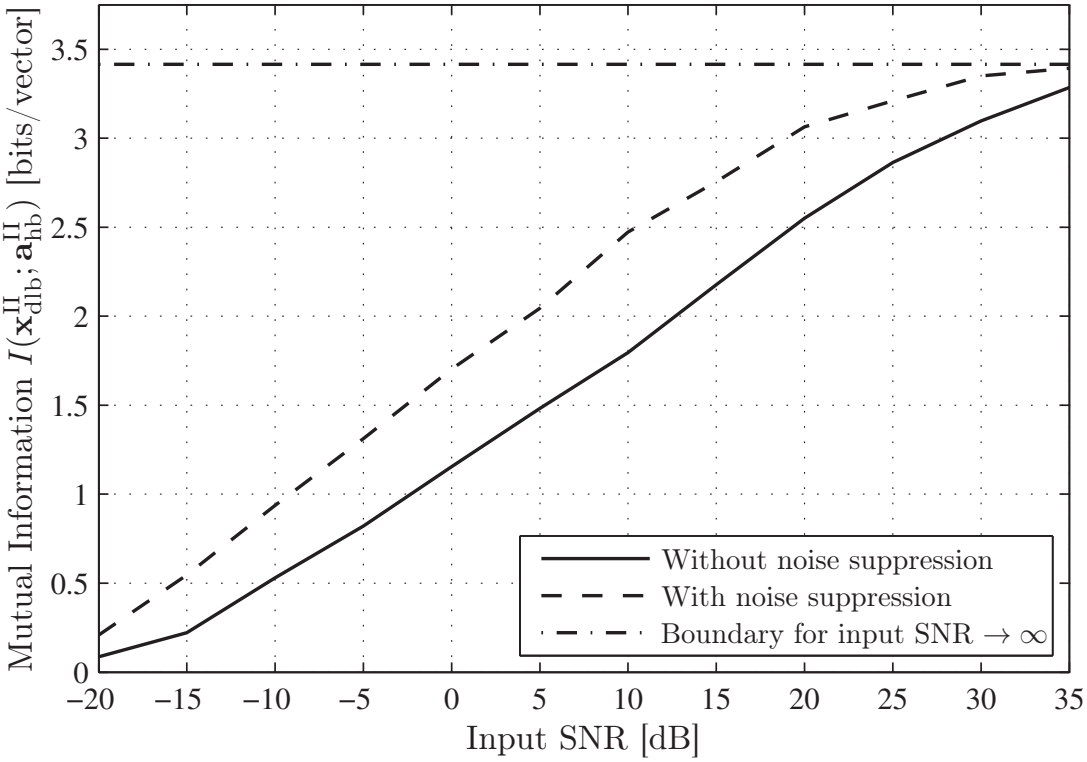

Figure 4.8: Measured amount of mutual information $I(\mathbf{x}_{\mathrm{dlb}}^{\mathrm{II}};\mathbf{a}_{\mathrm{hb}}^{\mathrm{II}})$ between high and low band as a function of the input SNR.

of noise suppression before ABWE is performed: the mutual information with prior noise reduction is significantly higher than without noise reduction.

In total, the investigations in this section quantitatively show the existence of dependencies between low and high band even if the low band signal is severely disturbed by additive noise. In order to exploit this mutual information, the application of noise suppression to the low band signal is advantageous by all means before estimating the high band parameters.

4.4 Performance Results

After demonstrating the theoretical functionality of artificial bandwidth extension also in highly disturbed noisy environments, the entire wideband noise reduction system which is presented in Sec. 4.2 is investigated in this section. In principle, any noise reduction technique can be applied within the proposed system to perform the suppression in the low band and to estimate the conventional (first) weighting gains $G_{\mathrm{hb}}^{\mathrm{conv}}$ in the high band, see Figs. 4.2 and 4.4. In the following, the Wiener filter [LO79] as well as the proposed Kalman filter (KF) approach of Sec. 3.2.3.3 are used to enhance the low band signal and to compute $G_{\mathrm{hb}}^{\mathrm{conv}}$ in the high band.

In order to determine the (second) ABWE weighting gains $G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu')$ in the upper branch of Fig. 4.4, the well-known Wiener filter approach [LO79] is applied according to:

$$G_{\mathrm{hb}}^{\mathrm{bwe}}(\mu') = \frac{\hat{\xi}_{\mathrm{hb}}(\mu')}{\hat{\xi}_{\mathrm{hb}}(\mu') + 1}, \tag{4.25}$$

using the a priori SNR estimates $\hat{\xi}_{\mathrm{hb}}(\mu')$ obtained by ABWE techniques, cf. Eq. 4.9.

In the investigation, the novel wideband noise suppression technique with the use of $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ and $\bar{\alpha}$ is compared with systems, where only the Wiener filter [LO79] or only the proposed Kalman filter (KF) approach of Sec. 3.2.3.3 are separately applied to the low band and the *entire* high band signal. Moreover, an upper bound for the performance in the high band is given by utilizing only the optimum weighting gains $G_{\mathrm{hb}}^{\mathrm{opt}}$ at high frequencies. The weighting gains $G_{\mathrm{hb}}^{\mathrm{opt}}$ are therefore also determined by using the Wiener filter based on $\xi_{\mathrm{hb}}^{\mathrm{opt}}$ according to Eq. 4.11.

In order to achieve a fair comparison, the same analysis and synthesis structure is applied to all techniques including low-pass and high-pass filtering, downsampling and FFT transformation according to Fig. 4.2. In total, the following six different setups are investigated where Setups ④ and ⑤ are based on the new wideband speech enhancement system as presented in Sec. 4.2:

Setup	Low Band Noise Reduction	High Band Noise Reduction
①	Wiener filter	Wiener filter
②	Proposed Kalman filter	Wiener filter
③	Proposed Kalman filter	Proposed Kalman filter
④	Proposed Kalman filter	$G_{\mathrm{hb}}^{\mathrm{conv}} \;\hat{=}\;$ Proposed Kalman filter $G_{\mathrm{hb}}^{\mathrm{bwe}} \;\hat{=}\;$ Wiener filter $\alpha_{\mathrm{hb}} = \alpha_{\mathrm{hb}}^{\mathrm{ref}}$
⑤	Proposed Kalman filter	$G_{\mathrm{hb}}^{\mathrm{conv}} \;\hat{=}\;$ Proposed Kalman filter $G_{\mathrm{hb}}^{\mathrm{bwe}} \;\hat{=}\;$ Wiener filter $\alpha_{\mathrm{hb}} = \bar{\alpha}_{\mathrm{hb}}$
⑥	Proposed Kalman filter	Optimum weighting gains $G_{\mathrm{hb}}^{\mathrm{opt}}$

In Setup ④, the reference cross-fading factor $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$, which is defined in Eq. 4.10, is used to perform the fading between $G_{\mathrm{hb}}^{\mathrm{conv}}$ and $G_{\mathrm{hb}}^{\mathrm{bwe}}$. This shows the maximum quality which can be achieved by Setup ⑤, where $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ is estimated using trained look-up tables.

The parameters used in the simulations are listed in Tab. 4.3. Although the sampling frequency of the input signal $y(k')$ is $f_{\mathrm{s}} = 16\,\mathrm{kHz}$, the noise reduction techniques are applied to the filtered and downsampled signals $y_{\mathrm{lb}}(k)$ and $y_{\mathrm{hb}}(k)$. Therefore, frame and FFT length refer to signals which are sampled at $f_{\mathrm{s}} = 8\,\mathrm{kHz}$. For the Wiener filter and the proposed Kalman filter solution the same settings as in Chapter 3 are applied.

Parameter	*Settings*
Sampling frequency	16 kHz
Frame length L_{F}	160 ($\hat{=}$20 ms due to downsampling)
FFT length M_{F}	256 (including zero-padding)
Frame overlap	75% (Hann window)
Input SNR	-10 dB ... 35 dB (step size: 5 dB)
QMF filter length	64
Number subbands M'_{F}	24
Number MFCCs N_{C}	13
Codebook size N_{bwe}	128 (training based on 1.5 h speech)
GMM Mixture Components M_{GM}	8

Table 4.3: System settings.

According to Sec. 4.2, the observation vector $\mathbf{x}_{\mathrm{dlb}}$ for the ABWE consists of $N_{\mathrm{C}} = 13$ MFCCs and the ZCR of the low band signal $\hat{s}_{\mathrm{lb}}$ whereas the feature vector $\mathbf{a}_{\mathrm{hb}}$ comprises $b_{\mathrm{hb}} = M'_{\mathrm{F}}/2 = 12$ logarithmic subband energies of the high band. For the training of the HMM about 1.5 hours of clean speech are taken randomly from the NTT speech database [NC94], including different male and female speakers. In the system, the clean observation vector $\mathbf{x}_{\mathrm{lb}}$ is not available and the disturbed vector $\mathbf{x}_{\mathrm{dlb}}$ extracted from the enhanced low band signal $\hat{s}_{\mathrm{lb}}(k)$ is used instead, see Fig. 4.4. The look-up tables required for the estimation of $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ are generated based on 10 minutes of clean speech from the NTT database disturbed by WGN at different input SNR values. Here again, white Gaussian noise is used in the training process in order to become as independent of a specific noise type as possible.

For the investigations, the same instrumental measurements as in Chapter 3 are applied, namely the segmental noise and speech attenuation as well as the segmental speech SNR (see Appendix D). For the instrumental evaluation of the different noise reduction schemes, five speech signals from the NTT speech database are each degraded by six different noise types (f16, babble, car, factory1, buccaneer, white), taken from the NOISEX-92 database [VS93]. Among the five speech signals, there are three sequences from male and two from female speakers, each with a length of 8 seconds. The speech signals used for the evaluation are not included in the training data for the HMM and the look-up tables.

Figure 4.9 depicts the averaged results for the difference between noise and speech attenuation plotted over the input SNR. Figure 4.10 illustrates the segmental speech SNR dependent on the noise attenuation with the input SNR as control variable. The points of best performance are placed in the upper right corner of the figure.

At first, the instrumental measurements demonstrate once again the advantage of the novel Kalman filter system presented in Chapter 3 compared to the Wiener filter solution. Especially regarding the noise reduction in the low band signal (Setup ① versus Setup ②), the results show a considerable improvement due to the model-based approach in both figures. As the energy of speech and noise signals usually decays towards higher frequency, the advantage of the Kalman filter system becomes smaller in the high band (Setup ② versus Setup ③), but is still noticeable. The instrumental

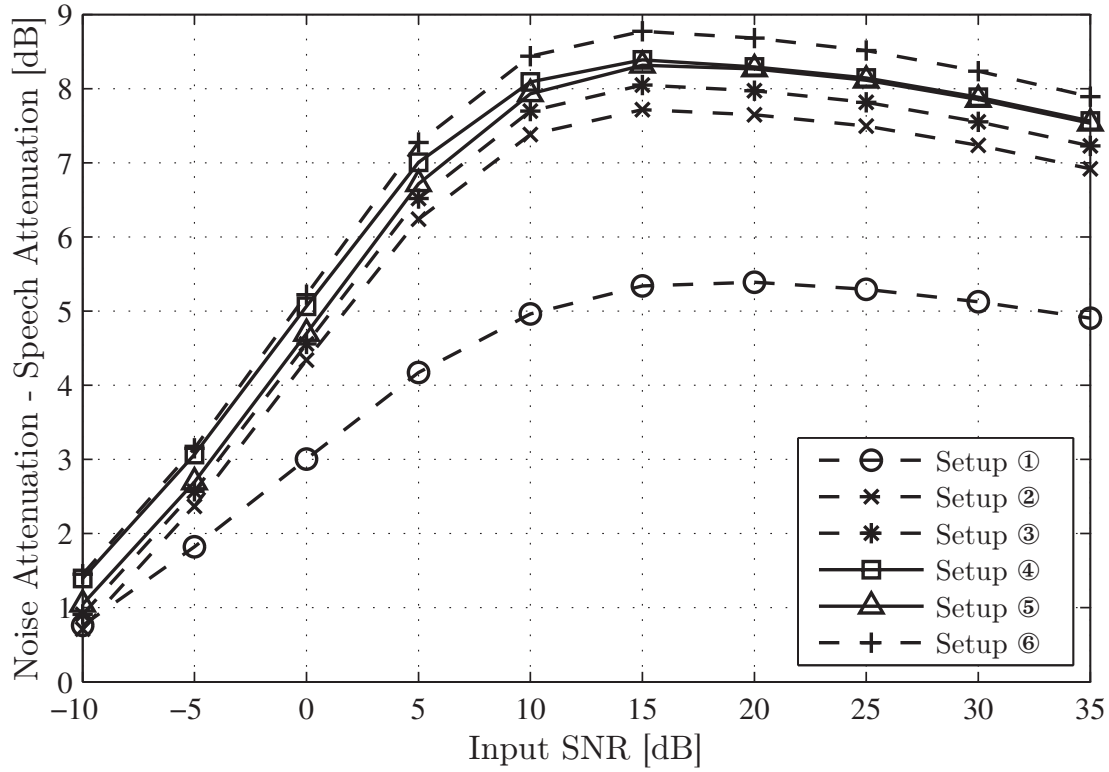

Figure 4.9: Difference between noise attenuation and speech attenuation plotted over input SNR. The different setups are explained on page 91.

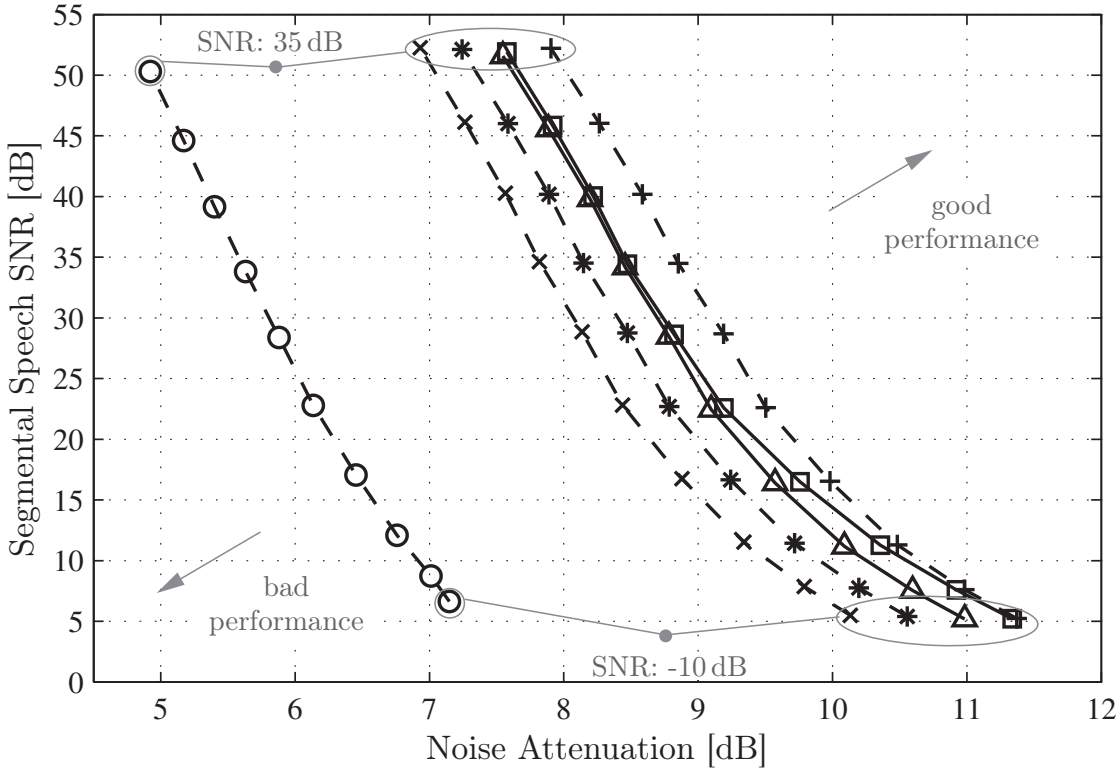

Figure 4.10: Segmental speech SNR plotted over noise attenuation.

measurements show that the performance can further be improved by using the new wideband speech enhancement system as proposed in this chapter. The additional use of the artificial bandwidth extension in the high band yields consistently better results compared to Setups ① – ③. Especially at low input SNR values, where mainly the ABWE approach is used (see Fig. 4.5), the new method using $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ outperforms the Kalman filter approach of Setup ③. It can be seen in both figures that the performance of the proposed method using $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ compared to $\bar{\alpha}_{\mathrm{hb}}$ slightly diverges in bad SNR conditions whereas the results are very similar at higher SNR values. As it is more difficult to obtain a precise SNR estimate in highly disturbed environments, this divergence can be explained by SNR estimation errors which lead to a suboptimal determination of the fading factors at low SNR values. Figures 4.9 and 4.10 also illustrate that the choice of the reference cross-fading factor $\alpha_{\mathrm{hb}}^{\mathrm{ref}}$ seems to be appropriate especially in low SNR conditions as Setup ⑤ performs only slightly worse than the 'optimal' estimator given by Setup ⑥ and the usage of $G_{\mathrm{hb}}^{\mathrm{opt}}$. In particular, at very low SNR values both curves are very close together. Informal listening tests confirmed the instrumental measurements and showed that the occurrence of musical tones is slightly reduced by the novel wideband noise suppression technique.

4.5 Conclusions

This chapter presents a new possibility to exploit spectral dependencies of speech signals for the purpose of wideband speech enhancement. By using techniques known from artificial bandwidth extension, the information of the enhanced low band signal is utilized again in order to achieve a better speech quality in the high band. Instrumental measurements demonstrate the superiority of the novel approach compared to 'conventional' wideband noise suppression techniques. The results are confirmed by information theoretic considerations which show the existence of spectral dependencies between the two bands already at very low input SNR values. The additional complexity which is required by the ABWE varies and is mainly depending on the number of HMM states and the number of GMM mixture components [Jax02].

It is demonstrated that the mutual information between low and high bands can be significantly increased if noise suppression is applied prior to ABWE. A slightly modified version of the proposed system can additionally be used if only the noisy narrowband signal is available in order to perform a joint noise reduction of the narrowband signal and artificial bandwidth extension.

Moreover, the approach is not strictly restricted to *one* low frequency band which facilitates the noise suppression in *one* high frequency band. Using a modified training process, the system can be adapted to support the speech enhancement in an arbitrary frequency band or even individual frequency bin.

5

Additional Methods for Quality Improvements

When dealing with noise reduction systems, there is always some kind of 'trilemma' between high noise suppression, low speech distortions and low occurrence of musical noise. The Kalman filter approach which is presented in Chapter 3 provides a good tradeoff between the first two properties. However, it also produces slightly more musical tones compared to less aggressive algorithms. Therefore, two different *musical noise countermeasures* are presented in the first part of this chapter which can be applied in a postprocessing stage separately or even combined.

Conventional noise estimation algorithms usually rely on the assumption that noise is stationary or quasi-stationary, see Sec. 2.3. However, realistic background noise can be rapidly time-varying and highly non-stationary. In order to cope with such strong variations when the noise is *harmonic*, the well-known Minimum Statistics algorithm [Mar01] is modified in the second part of this chapter yielding a significantly better performance compared to state-of-the-art noise estimation techniques.

5.1 Musical Noise Countermeasures

Noise suppression algorithms provide an improvement in terms of noise attenuation. Nevertheless, they often affect the actual speech signal and produce some artificial, randomly fluctuating type of noise, referred to as *musical noise*. The phenomenon of musical tones can be explained by noise or *Signal-to-Noise-Ratio* (SNR) estimation errors leading to spurious peaks in the processed spectrum. When the enhanced signal is reconstructed in the time domain, these peaks result in short sinusoidals whose frequencies vary from frame to frame. In particular, musical noise is very annoying during speech pauses and in low SNR conditions when it is not masked by the speech signal. As mentioned before, it is possible to gain control over the tradeoff between noise attenuation and speech distortions, e.g., by using the proposed Kalman filter system described in Chapter 3. However, this algorithm can not prevent the generation of musical noise.

In literature, a variety of different methods for reducing musical tones is proposed. A lower limit to the a priori SNR is applied in [Cap94] resulting in a flooring of the spectral weighting gains. The well-known decision-directed approach [EM84] prevents the musical noise phenomenon by recursive smoothing over time the a priori SNR. A time smoothed gain factor is proposed in [GNC99] in order to reduce the dynamics of the weights. In [GTT98], a postprocessing method is presented to suppress the annoying artifacts based on a speech/musical noise classification. Cepstral smoothing is applied to the spectral weighting gains in [BGM07] and [GM10b] enabling a selective smoothing of speech and musical tones.

In this chapter, two different approaches are presented that effectively suppress musical noise. The first technique performs adaptive spectral smoothing of the weighting gains relying on a low SNR detector. In contrast, the second method is based on noise suppression with adaptive frequency resolution where the resolution is lower during speech pauses in order to reduce the tonality of the residual noise. Both musical noise countermeasures can be applied to the spectral weighting gains in a postprocessing stage either separately or sequentially combined in order to achieve an even higher suppression. In the sequel, a sampling frequency of $f_\mathrm{s} = 8\,\mathrm{kHz}$ is assumed for the input signal. However, the techniques have also been implemented for signals sampled at $f_\mathrm{s} = 16\,\mathrm{kHz}$ and can simply be adapted to other sampling frequencies without difficulties.

In the following, at first a brief overview of the proposed system is given. Afterwards, the two postprocessing concepts are presented in detail. Finally, both techniques are evaluated by means of instrumental measurements and auditory judgments.

5.1.1 System Overview

A simplified block diagram of the proposed noise reduction system including postprocessing of the spectral weighting gains is depicted in Fig. 5.1. The system can be used for both proposed musical noise countermeasures in this section and uses the same analysis/synthesis structures (i.e., segmentation, windowing, *Fast Fourier Transform* (FFT), *Inverse Fast Fourier Transform* (IFFT) and overlap-add) as applied in Chapters 2–4 in order to transform the noisy input signal $y(k) = s(k) + n(k)$ into the frequency domain.

Both postprocessing techniques treat the estimation of the initial weighting gains as *black box* and can therefore be applied to an arbitrary noise reduction system working in the frequency domain. Possible estimators are, e.g., purely statistical weighting rules as outlined in Chapter 2 or the model-based approach presented in Chapter 3. While for the statistical estimators the weighting gains can often be determined directly from a specific weighting rule, the weighting gains for the proposed Kalman filter technique in Chapter 3 are given by (see Fig. 3.5):

$$G(\lambda,\mu) = G_\mathrm{K}(\lambda,\mu) = \frac{\hat{S}_\mathrm{up}(\lambda,\mu)}{Y(\lambda,\mu)}, \tag{5.1}$$

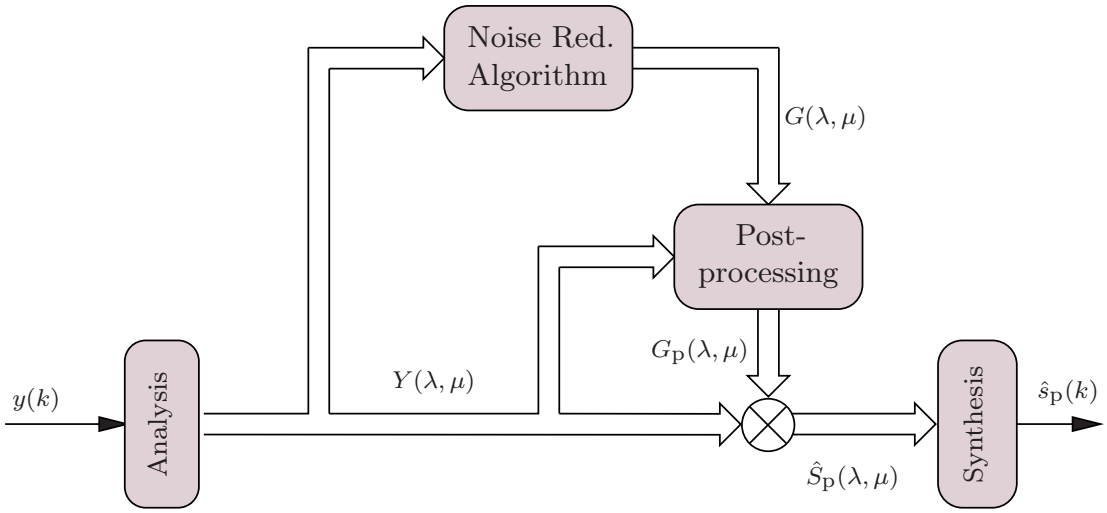

Figure 5.1: Block diagram of noise reduction system with postprocessing.

where λ and μ denote frame and frequency indices, respectively. In order to reduce the occurrence of musical tones, the actual weighting gains additionally run through a postprocessing stage. Therefore, the spectral smoothing approach as well as the adaptive bandwidth solution generate modified weighting gains $G_{\mathrm{p}}(\lambda,\mu)$ relying on the original weighting gains and the noisy input coefficients $Y(\lambda,\mu)$. The actual spectral weighting is performed by multiplying the noisy spectrum $Y(\lambda,\mu)$ with the new weighting gains $G_{\mathrm{p}}(\lambda,\mu)$:

$$\hat{S}_{\mathrm{p}}(\lambda,\mu) = G_{\mathrm{p}}(\lambda,\mu) \cdot Y(\lambda,\mu). \tag{5.2}$$

Finally, the postfiltered spectrum $\hat{S}_{\mathrm{p}}(\lambda,\mu)$ is transformed back into the time domain yielding the processed signal $\hat{s}_{\mathrm{p}}(k)$.

5.1.2 Spectral Smoothing of Weighting Gains

This section presents a novel *Postfilter* (PF) for the spectral weighting gains which is capable of reducing musical noise in a simple but efficient way. It includes a robust detector for speech pauses and low SNR conditions and adaptively smoothes the weighting gains over frequency based on soft-decisions. In addition to noise suppression [EV09], it is shown in [JSEV10] that the proposed postfilter concept can also be applied for the purpose of speech dereverberation.

5.1.2.1 Concept

As mentioned before, the nature of the musical noise phenomenon can be described as additional noise which arises in the processed signal due to estimation errors. Randomly spaced spectral peaks occur in the weighting gains and are perceived as time-varying tones in the output signal. The main idea of this PF concept is to adaptively eliminate these peaks in low SNR conditions. Therefore, at first a reliable and robust detector for the respective low SNR regions is required which is presented in

the following. Based on the results of this detector, spectral smoothing of the initial magnitudes $|G(\lambda,\mu)|$ is performed.

Low SNR Detection

It turned out that the power ratio $\psi(\lambda)$ between the processed speech *Discrete Fourier Transform* (DFT) coefficients $\hat{S}(\lambda,\mu)$ from the initial noise suppression technique and the noisy DFT coefficients $Y(\lambda,\mu)$ provides a good indicator for speech presence or absence in the current frame λ and can thus be applied as low SNR detector according to:

$$\psi(\lambda) = \frac{\sum\limits_{\mu=0}^{M_\mathrm{F}-1} |G(\lambda,\mu)\cdot Y(\lambda,\mu)|^2}{\sum\limits_{\mu=0}^{M_\mathrm{F}-1} |Y(\lambda,\mu)|^2} = \frac{\sum\limits_{\mu=0}^{M_\mathrm{F}-1} |\hat{S}(\lambda,\mu)|^2}{\sum\limits_{\mu=0}^{M_\mathrm{F}-1} |Y(\lambda,\mu)|^2}, \tag{5.3}$$

where M_F indicates the FFT length. If the current frame mainly contains speech (high SNR), the power of the processed frame is equal or only slightly lower to the power of the noisy input frame, i.e., $\psi(\lambda) \approx 1$. By contrast, the noise reduction system is supposed to strongly attenuate the input signal in low SNR conditions (or during a speech pause), resulting in a power ratio $\psi(\lambda) \approx 0$.

An example is depicted in Fig. 5.2 for a sequence of 8 seconds length. The upper plot shows the spectrogram of the clean speech signal, the lower plots the resulting values for the power ratio $\psi(\lambda)$ at the input SNR values -5 dB, 5 dB and 15 dB, respectively. The speech signal is taken from the NTT speech database [NC94] and is disturbed by 'factory noise' from the NOISEX-92 database [VS93]. The weighting gains which are required in Eq. 5.3 are calculated based on the super-Gaussian *Maximum A Posteriori* (MAP) estimator [LV05]. As can be seen, the power ratio $\psi(\lambda)$ correlates well with the speech activity even at very low input SNR values.

In order to detect only low SNR regions, a threshold ψ_thr is applied to $\psi(\lambda)$ setting the power ratio equal to 1 if $\psi(\lambda) \geq \psi_\mathrm{thr}$ as follows:

$$\psi_\mathrm{T}(\lambda) = \begin{cases} 1, & \text{if} \quad \psi(\lambda) \geq \psi_\mathrm{thr} \\ \psi(\lambda), & \text{else.} \end{cases} \tag{5.4}$$

The threshold ψ_thr controls the tradeoff between speech distortions and musical noise reduction as will be seen below.

In principle, the detection of low SNR regions could also be achieved by directly utilizing the estimated a priori SNR which is already available in most noise suppression techniques, cf. Chapter 2. However, the proposed method described above works more reliable due to the additional consideration of the weighting gains $G(\lambda,\mu)$ within the current frame when determining the power ratio ψ. Moreover, the noise reduction system can be treated as black box by this novel procedure as only the noisy input signal and the enhanced output signal or the spectral weighting gains are required for the detection.

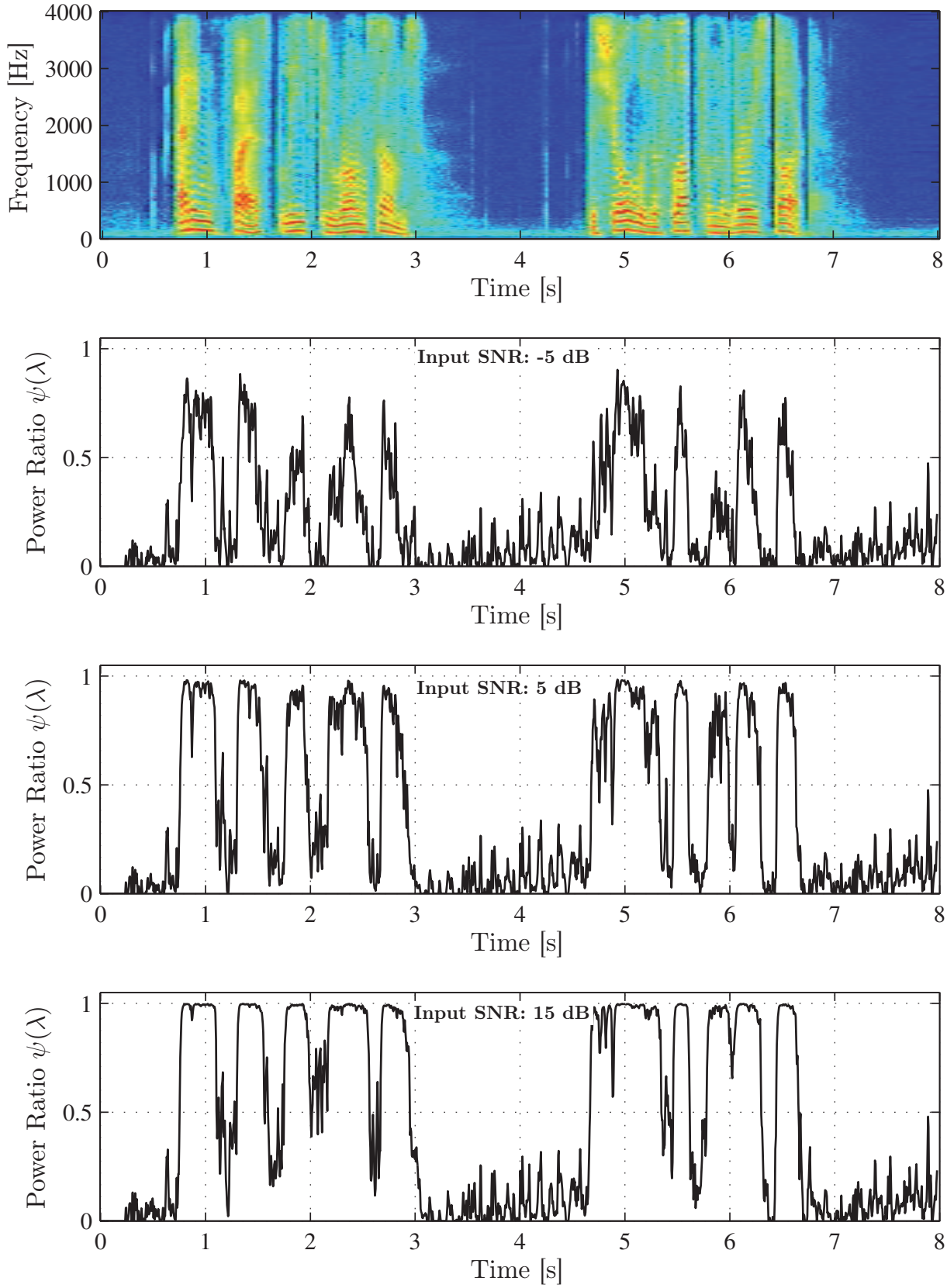

Figure 5.2: *Upper plot*: Spectrogram of the clean speech signal: "Adding fast leads to wrong sums. The show was a flop from the very start." (male voice). *Lower plots*: Results of the power ratio $\psi(\lambda)$ for different input SNR values (noise type: factory).

Adaptive Spectral Smoothing

The introduced power ratio $\psi_{\mathrm{T}}(\lambda)$ yields a reliable detection of low SNR regions. In order to prevent the annoying musical tones within these regions, the magnitudes of the initial weighting gains $G(\lambda,\mu)$ are adaptively smoothed over frequency using a *Moving Average* (MA) window. Therefore, the odd length L_{MA} of the MA window is adjusted framewise based on $\psi_{\mathrm{T}}(\lambda)$ and ψ_{thr} according to:

$$L_{\mathrm{MA}}(\lambda) = \begin{cases} 1, & \text{if} \quad \psi_{\mathrm{T}}(\lambda) = 1 \\ 2 \cdot \mathrm{round}\left(\left(1 - \frac{\psi_{\mathrm{T}}(\lambda)}{\psi_{\mathrm{thr}}}\right) \cdot \chi\right) + 1, & \text{else.} \end{cases} \tag{5.5}$$

The term $1 - \frac{\psi_{\mathrm{T}}(\lambda)}{\psi_{\mathrm{thr}}}$ provides a soft-decision which states the reliability of the low SNR detection. The function round$(\cdot)$ rounds the element to the nearest integer and χ is a scaling factor that determines the maximum degree of smoothing. Equation 5.5 ensures that the more reliable a low-SNR-frame is detected, the longer the window length and the stronger the corresponding smoothing of the weighting gains.

Applying a moving average window of length $L_{\mathrm{MA}}(\lambda)$ to the initial weighting gains $G(\lambda,\mu)$ is equivalent to a linear filtering over frequency using the impulse respond $H_{\mathrm{MA}}(\lambda,\mu)$ given by:

$$H_{\mathrm{MA}}(\lambda,\mu) = \begin{cases} \frac{1}{L_{\mathrm{MA}}(\lambda)}, & \text{if} \quad \mu < L_{\mathrm{MA}}(\lambda) \\ 0, & \text{else} \end{cases}, \quad \text{where} \quad 0 \leq \mu < M_{\mathrm{F}}. \tag{5.6}$$

Figure 5.3 depicts the Fourier transform of $H_{\mathrm{MA}}(\lambda,\mu)$ for different values of L_{MA}. Please note that the term 'frequency' in this context is somewhat misleading as $H_{\mathrm{MA}}(\lambda,\mu)$ is already applied in the frequency domain. However, Fig. 5.3 shows the low-pass characteristic of the filter $H_{\mathrm{MA}}(\lambda,\mu)$ whose cut-off 'frequency' is decreasing with an increasing window length L_{MA} leading to a stronger smoothing of the resulting weighting gains.

In order to obtain the smoothed weighting gains $G_{\mathrm{p}}(\lambda,\mu)$ within the postfilter, the weighting gain magnitudes of the initial noise reduction system are convolved by the respective filter $H_{\mathrm{MA}}(\lambda,\mu)$ (convolution over frequency index μ) and the phase information of the original weighting gains is re-used:

$$G_{\mathrm{p}}(\lambda,\mu) = G_{\mathrm{p}}^{\mathrm{PF}}(\lambda,\mu) = |G(\lambda,\mu)| * H_{\mathrm{MA}}(\lambda,\mu) \cdot \exp\left(\mathrm{j} \cdot \angle\{G(\lambda,\mu)\}\right), \tag{5.7}$$

where '$*$' states the convolution operator. According to Fig. 5.1, the weighting gains $G_{\mathrm{p}}^{\mathrm{PF}}(\lambda,\mu)$ are finally applied to the noisy DFT coefficients $Y(\lambda,\mu)$ and the enhanced speech DFT coefficients $\hat{S}_{\mathrm{p}}(\lambda,\mu)$ are transformed back into the time domain.

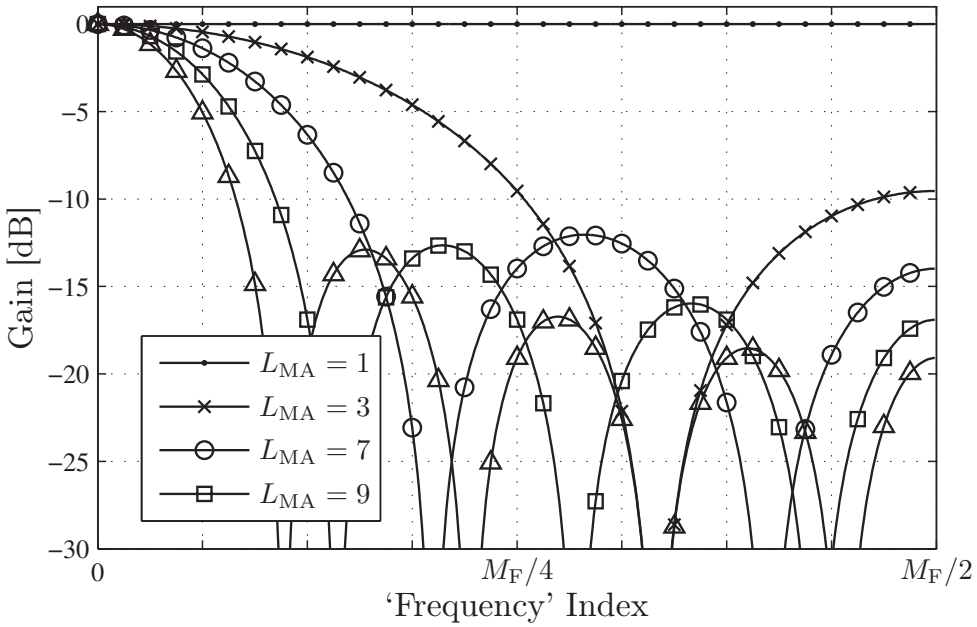

Figure 5.3: Fourier transform of $H_{\mathrm{MA}}(\lambda, \mu)$ for different values of L_{MA}. In this example, the FFT size is set to $M_{\mathrm{F}} = 256$.

5.1.3 Adaptive Bandwidth Resolution

The purpose of spectral transformations in speech processing is to exploit special properties of the input signal which are better accessible in the transform domain. Most of the noise reduction techniques separate speech and noise in the frequency domain using the DFT or an analysis/synthesis filter-bank with uniform resolution in the spectral domain. The resulting frequency bands hold the same bandwidth and are equidistant distributed on the frequency scale in contrast to the perception abilities of the human auditory system [ZF90]. So far, there are considerably less approaches published for noise reduction which use a non-uniform frequency resolution achieved, e.g., by wavelet-based transforms [GEH98, LGO$^+$96, SB97] or allpass transformed DFT filter-banks [HS06, Chapter 2]. The advantage of these methods is the possibility to adjust the spectral resolution with respect to psychoacoustical criteria, e.g, according to the well-known Bark scale [ZF90] showing a high resolution at low frequencies and a low resolution at high frequencies. Using Bark bands for speech enhancement reduces the tonality of the residual noise (musical tones) especially at higher frequencies as the variance of the noise power and SNR estimates is decreased due to the higher bandwidths. However, it also leads to disturbances during speech activity preventing the enhancement of voiced speech by making the speech signal sound muffled to some extent. In order to compensate the tradeoff between musical tones and muffled speech, a time-varying frequency resolution concept is proposed in this thesis as second postprocessing procedure and is presented in the following.

5.1.3.1 Concept

The main idea of this second musical noise countermeasure is to perform a variant spectral analysis of the weighting gains with framewise adaptive frequency resolution such that the spectral resolution is high during speech activity (low bandwidths) and low during speech pauses (high bandwidths). A similar approach is presented in [GLH03]. In the following, a detailed description of the proposed postprocessing technique is given outlining all differences compared to [GLH03].

According to Fig. 5.1, at first the weighting gains $G(\lambda, \mu)$ are determined based on any noise reduction method using the full uniform frequency resolution M_{F}. In the following postprocessing step, the spectral resolution of the weighting gains is adjusted recursively. Therefore, the signal-to-noise-ratio SNR_b in each Bark band b is initially estimated according to:

$$\mathrm{SNR}_b = \frac{\sum\limits_{\tilde{\mu}=B_{\mathrm{l}}(b)}^{B_{\mathrm{u}}(b)} \left|G(\lambda,\tilde{\mu}) \cdot Y(\lambda,\tilde{\mu})\right|^2}{\sum\limits_{\tilde{\mu}=B_{\mathrm{l}}(b)}^{B_{\mathrm{u}}(b)} \max\left(\left|Y(\lambda,\tilde{\mu})\right|^2 - \left|G(\lambda,\tilde{\mu}) \cdot Y(\lambda,\tilde{\mu})\right|^2, \epsilon_0\right)}, \tag{5.8}$$

where $1 \leq b \leq M_{\mathrm{B}}/2$. Thereby, $M_{\mathrm{B}}/2$ represents the total number of Bark bands and $B_{\mathrm{l}}(b)$ and $B_{\mathrm{u}}(b)$ are the lower and upper frequency bin limits of the respective b-th Bark band. The constant ϵ_0 states a very small number. The relation between Bark scale and the lower and upper frequency bounds in Hz is given in Tab. 5.1 up to a frequency of 7.7 kHz. Hence, working with narrowband signals (sampling frequency f_{s} = 8 kHz) leads to $M_{\mathrm{B}}/2$ = 17 Bark bands whereas $M_{\mathrm{B}}/2$ = 21 Bark bands are used when sampling at f_{s}=16 kHz. Afterwards the SNR estimates of all $M_{\mathrm{B}}/2$ Bark bands are compared with a threshold ϵ_{B} in order to check the speech presence status of each band. If the estimated SNR of Bark band b is lower than ϵ_{B}, i.e., $\mathrm{SNR}_b < \epsilon_{\mathrm{B}}$, it is assumed that this band contains no speech or only weak speech components and the respective weighting gains within this band are merged. Therefore, all weighting gains $G_{\mathrm{p}}^{\mathrm{aB}}$ of this band are set to the median value of the corresponding initial weighting gains G according to:

$$G_{\mathrm{p}}^{\mathrm{aB}}(\lambda,\mu) = \mathrm{median}\Big(G\big(\lambda, B_{\mathrm{l}}(b)\big),\ ..., G\big(\lambda, B_{\mathrm{u}}(b)\big)\Big), \quad \text{where} \quad B_{\mathrm{l}}(b) \leq \mu \leq B_{\mathrm{u}}(b). \tag{5.9}$$

The respective band remains unconsidered in the following procedure. However, if the SNR of Bark band b is higher or equal to ϵ_{B}, i.e., $\mathrm{SNR}_b \geq \epsilon_{\mathrm{B}}$, the spectral resolution in this band is increased by a factor of 2 assuming speech is active. To achieve this, the specific b-th band is split into two new subbands of same bandwidth $(B_{\mathrm{u}}(b) - B_{\mathrm{l}}(b))/2$, which is half of the bandwidth of the original band. In the next step, the SNR of the new bands is estimated following Eq. 5.8 by using the new lower and upper frequency bin limits. Afterwards, the new estimated SNR values are again compared with the threshold ϵ_{B} leading to the same consequences as above: if the SNR is

Bark band b	1	2	3	4	5	6	7
Lower bound [Hz]	0	100	200	300	400	510	630
Upper bound [Hz]	100	200	300	400	510	630	770

Bark band b	8	9	10	11	12	13	14
Lower bound [Hz]	770	920	1080	1270	1480	1720	2000
Upper bound [Hz]	920	1080	1270	1480	1720	2000	2380

Bark band b	15	16	17	18	19	20	21
Lower bound [Hz]	2380	2700	3150	3700	4400	5300	6400
Upper bound [Hz]	2700	3150	3700	4400	5300	6400	7700

Table 5.1: Bark scale and corresponding frequency bands [ZF90].

lower than ϵ_{B}, the median value of all spectral weighting gains within the specific band is calculated, otherwise the frequency band is once again split into two new subbands. This procedure is repeated in every frame as long as the estimated SNR in each subband is lower than the threshold ϵ_{B} or as long as the resulting bandwidth resolution corresponds to the original resolution of the FFT, i.e., $M_{\mathrm{F}}/2$. The final number of frequency bands used in frame λ is denoted by $M'_{\mathrm{F}}(\lambda)/2$ and varies between $M_{\mathrm{B}}/2$ and $M_{\mathrm{F}}/2$. The adaptation scheme is exemplary illustrated in Fig. 5.4 by using a subband-tree structure. The resulting weighting gains $G_{\mathrm{p}}(\lambda,\mu) = G_{\mathrm{p}}^{\mathrm{aB}}(\lambda,\mu)$ are finally used for the spectral weighting as depicted in Fig. 5.1.

An example for the effective number $M'_{\mathrm{F}}(\lambda)/2$ of subbands which are used after applying the proposed postprocessing method can be seen in Fig. 5.5. The clean speech signal depicted in the upper figure is taken from the NTT database [NC94] and disturbed by *White Gaussian Noise* (WGN) at 10 dB input SNR. After processing the noisy signal by using the super-Gaussian MAP estimator [LV05], the spectral resolution of the weighting gains is adjusted adaptively as shown above. The resulting number of subbands $M'_{\mathrm{F}}/2$ is depicted in the lower figure over time. It can be seen that during speech pauses, a low frequency resolution is used leading to a better suppression of musical tones. In contrast, the spectral resolution is considerably higher when speech is active. Thus, the speech signal is almost not affected by the postprocessing method. In the example, a sampling frequency of $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ and an FFT size of $M_{\mathrm{F}} = 256$ are used limiting the minimum number of subbands to $M'_{\mathrm{F,min}}/2 = M_{\mathrm{B}}/2 = 17$ (see Tab. 5.1) and the maximum number to $M'_{\mathrm{F,max}}/2 = M_{\mathrm{F}}/2 = 128$. The threshold ϵ_{B} is determined empirically and set to $\epsilon_{\mathrm{B}} = 0.5$.

In comparison to the proposed postprocessing technique in this work, the approach in [GLH03] uses a uniform partitioning of the frequency scale in the first step instead of utilizing Bark bands. In addition, the mean value is taken in [GLH03] to combine the weighting gains within the subbands and not the median value. The benefit of initializing the frequency scale by Bark bands can be recognized especially in low SNR

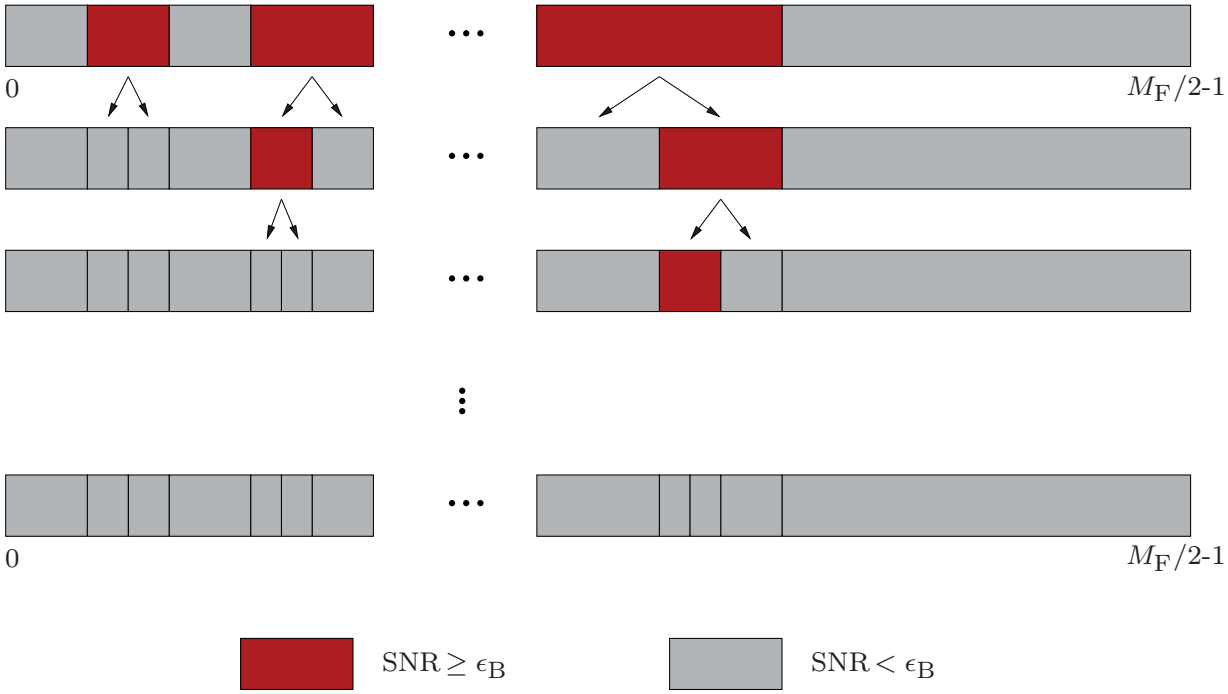

Figure 5.4: Adaptation scheme of bandwidths.

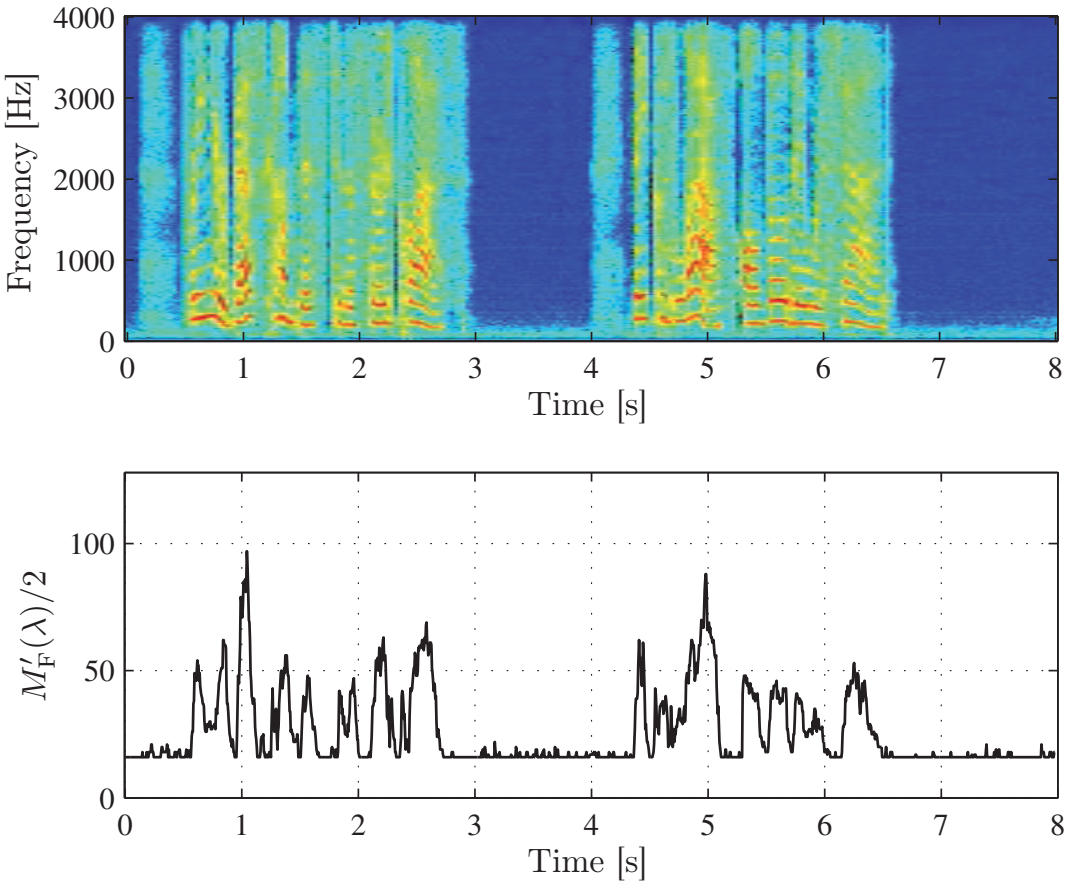

Figure 5.5: *Upper plot*: Spectrogram of the clean speech signal: "Bring your best compass to the third class. They could laugh although they were sad." (female voice). *Lower plot*: Resulting time-varying subband number $M'_F/2$ after gain calculation [LV05] and postprocessing of disturbed speech signal (noise type: WGN, input SNR: 10 dB, $\epsilon_B = 0.5$, $M_F = 256$).

conditions. Here, the non-uniform frequency resolution leads to less fluctuations in the resulting weighting gains at higher frequencies where musical tones typically arise. Using the median instead of the mean value avoids that outliers do affect the resulting weighting gain estimation in each subband.

5.1.4 Performance Results

The two musical noise countermeasures which are presented in the previous subsections can in principle be applied to the weighting gains of any noise reduction system. In the following, the postprocessing techniques are used in combination with the super-Gaussian MAP estimator [LV05] (see Sec. 2.5.5) as well as with the proposed Kalman filter approach based on the SNR-dependent *Minimum Mean Square Error* (MMSE) estimator which is presented in Chapter 3. As seen in Chapter 3, both noise suppression methods possess the slight tendency to produce musical tones and are therefore qualified for the evaluation here. The investigation is based on both instrumental measurements and auditory judgments. Instrumental measurements are used to analyze the influence of the postprocessing methods with respect to noise and speech attenuation as well as speech distortions. However, these entities allow no real statement about the quality of the processed background noise, i.e., the occurrence of musical tones in the output signal. Therefore, an informal subjective listening test is conducted in addition.

For the super-Gaussian MAP estimator and the Kalman filter the same system parameters as in Chapter 3 are applied. Further simulation settings are listed in Tab. 5.2. The values for the thresholds ψ_{thr} and ϵ_{B} as well as the scaling factor χ are determined empirically and provide a good compromise between speech distortions and musical noise suppression.

Parameter	*Settings*
Sampling frequency	8 kHz
Frame length L_{F}	160 (20 ms)
FFT length M_{F}	256 (including zero-padding)
Frame overlap	75% (Hann window)
Input SNR	-10 dB ... 35 dB (step size: 5 dB)

Spectral Smoothing of Weighting Gains	
Threshold ψ_{thr}	0.4
Scaling factor χ	10

Adaptive Bandwidth Resolution	
Threshold ϵ_{B}	0.5
Constant ϵ_0	2^{-52}

Table 5.2: System settings.

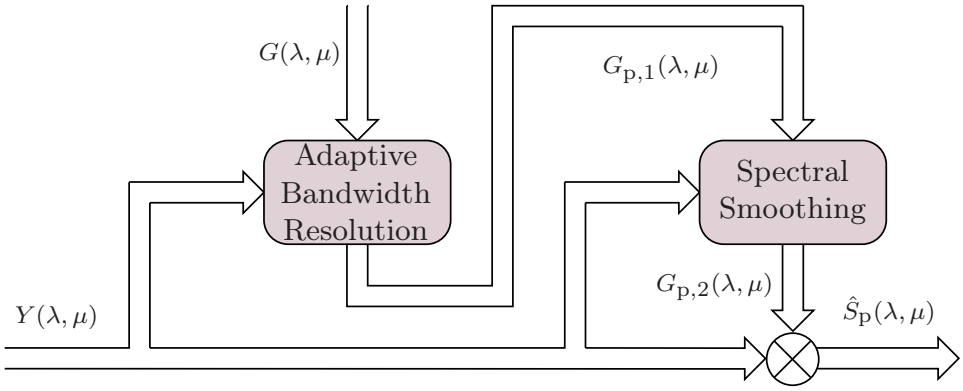

Figure 5.6: Serial connection of the two postprocessing techniques.

5.1.4.1 Instrumental Measurements

In the instrumental evaluation, four different cases are investigated for each of the two noise suppression techniques (see Fig. 5.1):

1. Noise Reduction without any postprocessing, i.e., $G_{\mathrm{p}}(\lambda,\mu) = G(\lambda,\mu)$.
2. Noise Reduction using the spectral smoothing approach of the weighting gains (SSWG) as proposed in Sec. 5.1.2, i.e., $G_{\mathrm{p}}(\lambda,\mu) = G_{\mathrm{p}}^{\mathrm{PF}}(\lambda,\mu)$.
3. Noise Reduction using the adaptive bandwidth resolution (ABR) method as proposed in Sec. 5.1.3, i.e., $G_{\mathrm{p}}(\lambda,\mu) = G_{\mathrm{p}}^{\mathrm{aB}}(\lambda,\mu)$.
4. Noise Reduction using both musical noise countermeasures in combination. In order to increase the musical noise suppression, at first the adaptive bandwidth resolution (ABR) method is applied to the weighting gains $G(\lambda,\mu)$. On top of this first method, the resulting gains $G_{\mathrm{p},1}(\lambda,\mu)$ are further processed using the spectral smoothing (SSWG) approach according to Fig. 5.6. It could be shown that the order 'ABR + SSWG' yields a slightly better performance compared to the other way round, i.e., 'SSWG + ABR'.

For the evaluation, the same instrumental measurements as already used in Chapters 3 and 4 are applied, i.e., the segmental noise and speech attenuation as well as the segmental speech SNR (see Appendix D). Furthermore, the same speech and noise signals as in Chapter 3 are utilized: five speech signals from the NTT speech database [NC94]) are each degraded by six different noise types (f16, babble, car, factory1, factory2, white) taken from the NOISEX-92 database [VS93]. Among the five speech signals, there are three sequences from male and two from female speakers, each with a length of 8 seconds.

While in Fig. 5.7 the averaged deviation between segmental noise and speech attenuation is depicted, Fig. 5.8 illustrates the results for the segmental speech SNR plotted over the noise attenuation. Therefore, the input SNR serves as control variable. In Fig. 5.8, the points of best performance are placed in the upper right corner.

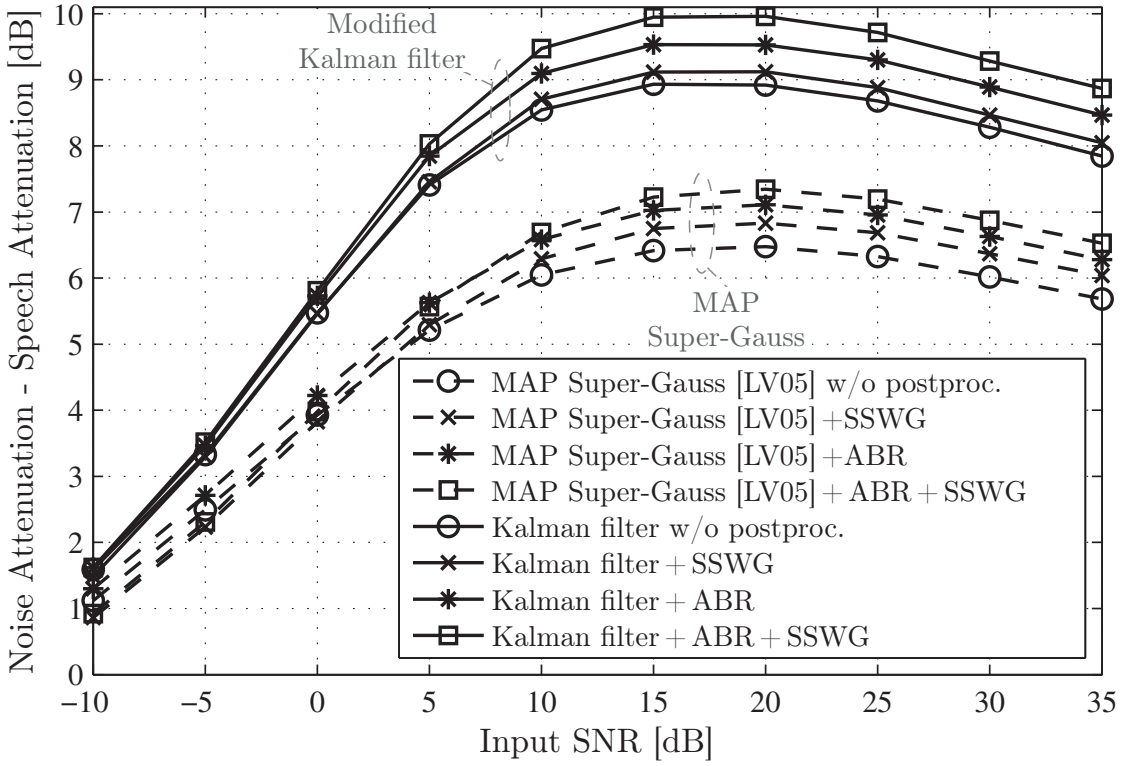

Figure 5.7: Difference between noise attenuation and speech attenuation plotted over input SNR.

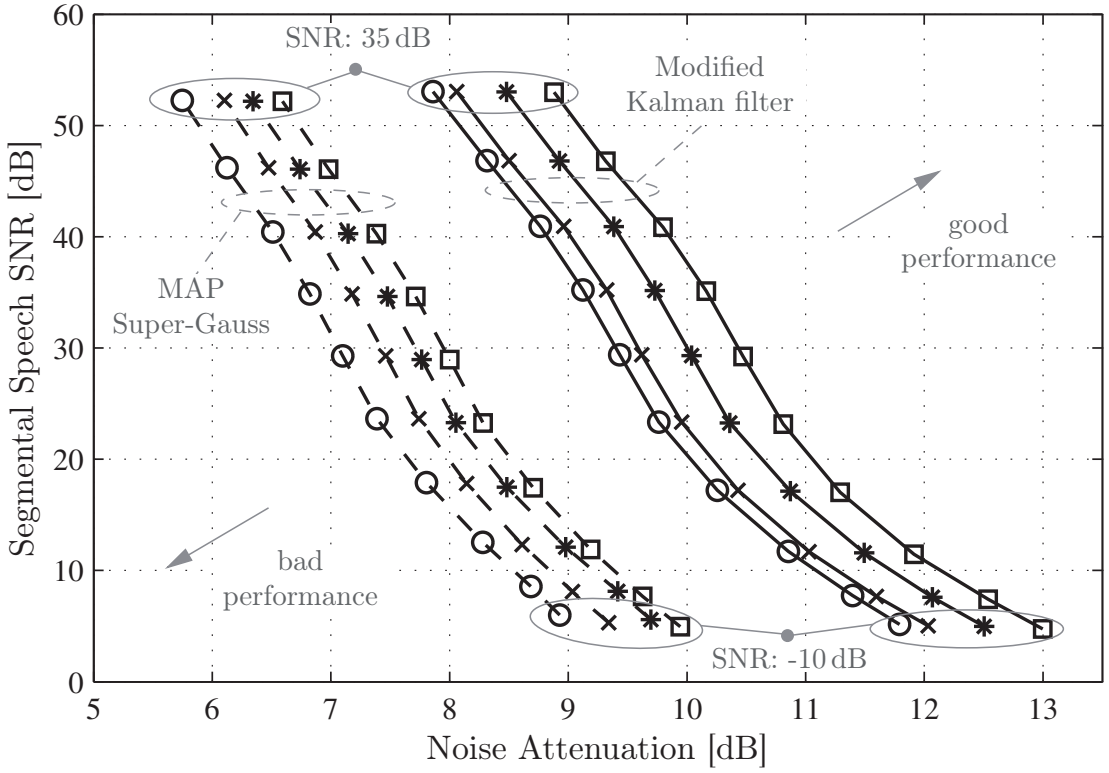

Figure 5.8: Segmental speech SNR plotted over noise attenuation.

The instrumental measurements show that the proposed postprocessing schemes are able to improve the results of both investigated estimators. Applied to the super-Gaussian MAP estimator, the spectral smoothing of the weighting gains and the adaptive bandwidth resolution technique achieve a better tradeoff between noise and speech attenuation, especially at higher input SNR values. They gain in noise attenuation when keeping the segmental speech SNR constant. Comparing the two musical noise countermeasures among each other, the adaptive decomposition of the frequency scale performs slightly better as can be seen in both figures. When concatenating both postfilter approaches at low input SNR values, the results are worse than the individual applications but still better than the MAP estimator without postprocessing in Fig. 5.7. In good SNR conditions however, the output of the combined system 'ABR+SSWG' gets better and eventually outperforms the other combinations beyond 5 dB input SNR as illustrated in Figs. 5.7 and 5.8.

A very similar behavior can be seen for the Kalman filter approach. Here, the use of the two postprocessing techniques consistently improves the results of the Kalman filter. The adaptive bandwidth resolution approach achieves again better results than the adaptive smoothing procedure of the weighting gains. In contrast to the MAP estimator, the concatenation of both techniques leads to a higher noise attenuation without affecting the speech quality over the entire SNR range.

Overall, the instrumental measurements show that the proposed musical noise countermeasures contribute to an additional increase in noise attenuation without producing more speech distortions except for the combined method 'ABR+SSWG' at very low SNR values when applied subsequent to the MAP estimator.

The improvements in terms of noise attenuation and speech quality are of course desirable but as mentioned before give no evidence about the suppression of musical noise in the processed output signals. Therefore, an informal listening test was also carried out, the results of which are presented in the following.

5.1.4.2 Auditory Judgments

In addition to the instrumental measurements, an informal *Comparison Category Rating* (CCR) test was conducted according to [ITU96] which presents two samples per question to the participants: a processed signal from Method A and a processed signal from Method B. One of the two noise suppression techniques (super-Gaussian MAP estimator [LV05] or the proposed Kalman filter approach of Sec. 3.2.3.3) was randomly assigned to Method A or B without any postprocessing. The respective other sample was processed by the same noise reduction method with subsequent musical noise suppression using the concatenation of both proposed techniques, i.e., 'ABR+SSWG'. The noisy input signals consisted of a speech signal randomly taken from the NTT speech database disturbed by a noise signal from the NOISEX-92 database at an input SNR varying between 5 dB and 15 dB. 19 experienced listeners were asked to judge the overall speech quality by choosing between the follow-

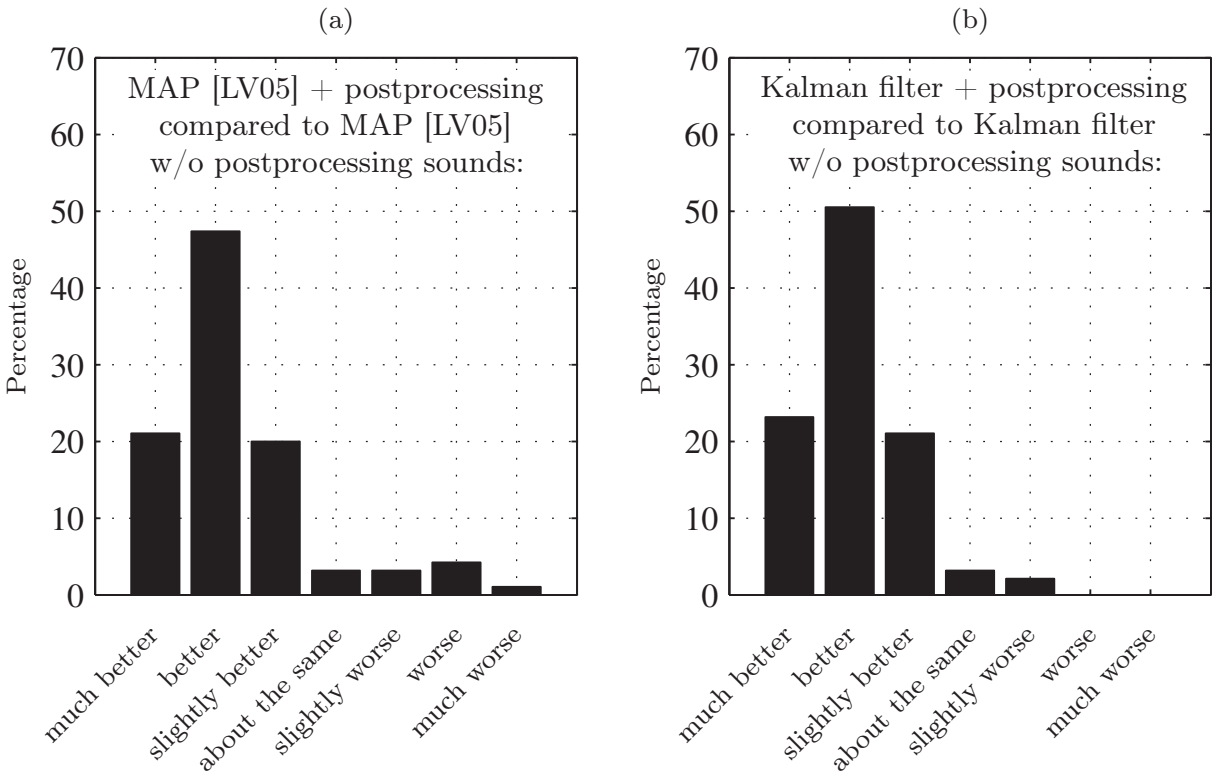

Figure 5.9: Results of the informal listening test comparing (a) the super-Gaussian MAP estimator [LV05] and (b) the Kalman filter approach with and without the proposed postprocessing techniques.

ing rating options: Sample A sounds {much better | better | slightly better | about the same | slightly worse | worse | much worse} than Sample B. Each test person had to judge 10 signals (5 per noise reduction method), i.e., the total results are based on $10 \cdot 19 = 190$ votes. The samples could be played ad libitum before the probands had to make their judgments.

The averaged results are separately illustrated in Fig. 5.9 for the MAP estimator as well as for the Kalman filter. It can clearly be seen that most listeners preferred the processed samples with subsequent postprocessing in both cases. As reason, they stated the reduction of musical noise while preserving the speech quality and noise attenuation. In some isolated cases, the participants favored the samples without postprocessing and explained their choices with a slightly 'muffled' sounding of the postprocessed signal. This indicates that the postprocessing techniques have sometimes been adjusted a little bit too aggressive. However, this problem can be solved by changing the respective system parameters in Tab. 5.2 at the expense of a lower musical noise suppression. In total, i.e., when averaging over the results of both estimators and when combining the options {much better | better | slightly better }, approximately 92% of the test listeners preferred the samples that were generated with the new postprocessing techniques.

5.1.5 Conclusions

In this section, two postprocessing methods are presented which effectively suppress musical noise. They can be applied to the spectral weighting gains of an arbitrary noise reduction technique. The first technique adaptively smoothes the spectral weighting gains over frequency based on soft-decisions of a low SNR detector. The second approach uses a framewise adaptive frequency resolution such that the spectral resolution is high during speech activity and low during speech pauses. In order to increase the suppression of musical tones, both techniques can be combined. Instrumental measurements in terms of noise and speech attenuation as well as segmental speech SNR show improvements of the new approaches when applied subsequent to two noise reduction methods. The instrumental results have been confirmed by an informal listening test.

5.2 Noise Estimation in Rapidly Varying Harmonic Noise Environments

A crucial component of any practical speech enhancement system is the estimation of the noise power spectrum. For this purpose, several approaches can be found in literature, e.g., the application of a *Voice Activity Detection* (VAD) [SKS99], the Minimum Statistics approach [Mar01] or the MMSE based noise *Power Spectral Density* (PSD) tracking algorithm [HHJ10] (see Sec. 2.3). All of these techniques assume stationary or slowly time-varying noise and have severe problems in tracking sudden noise variations leading to an underestimation of the noise power.

In this section, speech enhancement in noisy environments with rapidly time-varying harmonic noise and stationary random noise is investigated. Possible application areas can be found in intercom systems for motorcycles or in the interior of other motor vehicles, e.g., in order to communicate via a hands-free device inside a car where engine, wind and tires are the main noise sources. An example is depicted in Fig. 5.10 showing the noisy spectrogram of a speech signal which is disturbed by a real noise signal recorded inside a car. The strong spectral components of the harmonic noise signal are present at multiples of a fundamental frequency f_0 and contribute to the main noise power.

In the following, a novel noise suppression system consisting of two stages is presented which effectively reduces the considered harmonic and stationary noise [ERHV10b]. In the first stage, harmonic noise components are suppressed using a modified Minimum Statistics approach [ERHV10a]. Therefore, it is assumed that the instantaneous fundamental frequency f_0 for each frame is available to the noise reduction system, e.g., received from the vehicle's onboard computer which is possible in modern vehicles or estimated from a second reference microphone placed, e.g., near the engine. In the second stage, the remaining residual stationary background noise is reduced.

In principle, the proposed method can be applied to an arbitrary noise reduction system relying on an estimate of the noise PSD. Hence, it could also support the noise

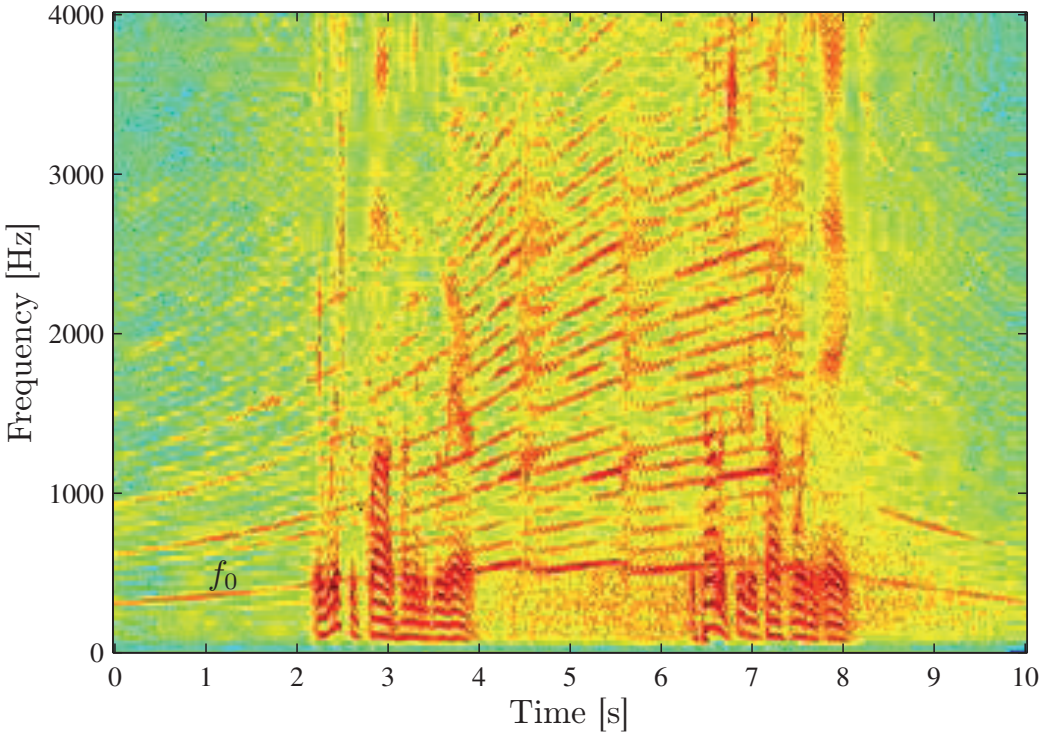

Figure 5.10: Spectrogram of noisy input signal showing strong spectral components of harmonic noise at multiples of the fundamental frequency f_0.

estimation process within the update step of the modified Kalman filter approach presented in Chapter 3. However, the use in this section is limited only to purely statistical estimation rules in order to allow a fair comparison with conventional noise estimation techniques which have been developed exactly for these methods.

The remainder of this section is organized as follows: at first a brief overview of the proposed noise reduction system is given. Afterwards, the two stages of the noise estimation technique are comprised in detail including the estimation of the harmonic and stationary noise power. Experimental results finally close this section.

5.2.1 System Overview

A simplified block diagram of the proposed system is shown in Fig. 5.11. The speech signal $s(k)$ is assumed to be degraded by an additive harmonic noise signal $n_{\mathrm{h}}(k)$ and a stationary noise signal $n_{\mathrm{s}}(k)$ according to:

$$y(k) = s(k) + n_{\mathrm{h}}(k) + n_{\mathrm{s}}(k). \tag{5.10}$$

Analog to the analysis procedure in the previous chapters, the signal $y(k)$ is transformed into the frequency domain using segmentation, windowing and the application of the FFT (transform length M_{F}). The spectrum of the noisy input signal is given by:

$$Y(\lambda, \mu) = S(\lambda, \mu) + N_{\mathrm{h}}(\lambda, \mu) + N_{\mathrm{s}}(\lambda, \mu), \tag{5.11}$$

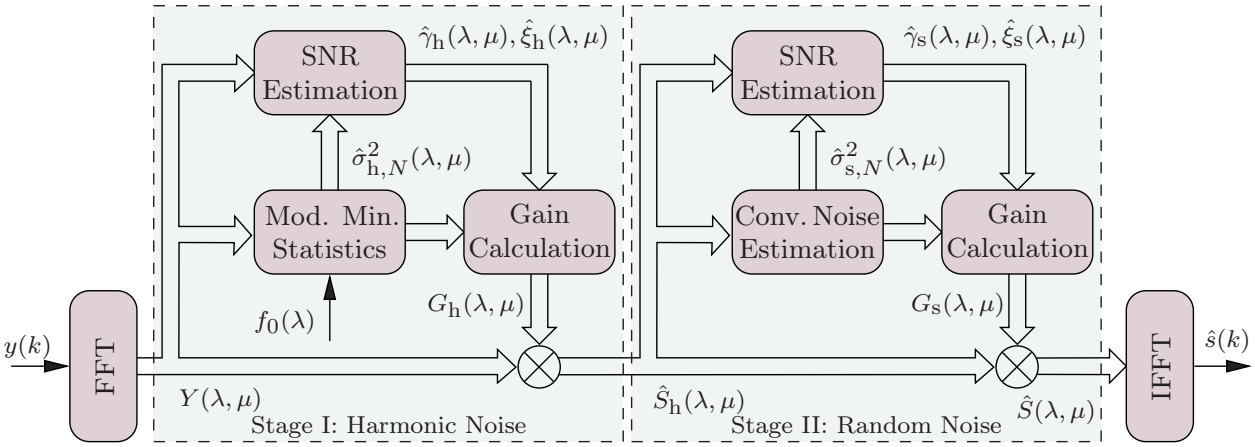

Figure 5.11: Proposed two stage noise reduction system for the elimination of rapidly time-varying harmonic and stationary noise.

where $S(\lambda,\mu)$, $N_\mathrm{h}(\lambda,\mu)$ and $N_\mathrm{s}(\lambda,\mu)$ represent the spectral DFT coefficients of the speech and the noise signals, respectively.

The proposed solution consists of the concatenation of two noise suppression stages relying on different noise PSD estimators. In the first stage, the *harmonic* noise powers $\sigma^2_{\mathrm{h},N}(\lambda,\mu)$ are estimated using a modified Minimum Statistics approach which exploits the knowledge of the instantaneous fundamental frequency f_0, see Sec. 5.2.2 for more details. Based on the estimates $\hat{\sigma}^2_{\mathrm{h},N}(\lambda,\mu)$, the *a posteriori SNR* $\gamma_\mathrm{h}(\lambda,\mu)$ and the *a priori SNR* $\xi_\mathrm{h}(\lambda,\mu)$ of the first stage are calculated according to:

$$\gamma_\mathrm{h}(\lambda,\mu) = \frac{|Y(\lambda,\mu)|^2}{\sigma^2_{\mathrm{h},N}(\lambda,\mu)} \quad \text{and} \quad \xi_\mathrm{h}(\lambda,\mu) = \frac{\mathbb{E}\{|S(\lambda,\mu)|^2\}}{\sigma^2_{\mathrm{h},N}(\lambda,\mu)}, \tag{5.12}$$

where the a priori SNR can again be determined, e.g., by using the decision-directed approach [EM84]. The actual spectral weighting in this stage is performed by multiplying the noisy spectrum $Y(\lambda,\mu)$ by weighting gains $G_\mathrm{h}(\lambda,\mu)$ resulting in an estimate of the harmonic noise-reduced spectrum:

$$\hat{S}_\mathrm{h}(\lambda,\mu) = G_\mathrm{h}(\lambda,\mu) \cdot Y(\lambda,\mu). \tag{5.13}$$

In the second stage, the enhanced spectrum $\hat{S}_\mathrm{h}(\lambda,\mu)$ is further improved with regard to the residual *stationary random* noise. While for the noise estimation, a conventional technique is used, e.g., the original Minimum Statistics algorithm [Mar01], the same methods as in the first stage are applied for the subsequent a posteriori SNR $\gamma_\mathrm{s}(\lambda,\mu)$ and a priori SNR $\xi_\mathrm{s}(\lambda,\mu)$ estimation as well as for the gain calculation $G_\mathrm{s}(\lambda,\mu)$ in this second stage.

The overall spectral weighting results in the estimate:

$$\hat{S}(\lambda,\mu) = G_\mathrm{h}(\lambda,\mu) \cdot G_\mathrm{s}(\lambda,\mu) \cdot Y(\lambda,\mu), \tag{5.14}$$

and IFFT with overlap-add is applied in order to obtain the enhanced time domain signal $\hat{s}(k)$.

5.2.2 Harmonic Noise PSD Estimation

According to Eq. 5.10, speech signals are assumed to be disturbed by stationary *and* harmonic noise characterized by (strong) spectral components at multiples of the (time-varying) fundamental frequency f_0. As the fundamental frequency might change over time very fast (e.g., when the engine accelerates or when a gear change occurs), conventional noise estimation techniques usually fail in tracking the spectral harmonics as they originally have been developed to estimate short-term *stationary* noise.

In the following, a novel modified Minimum Statistics algorithm is presented which is adapted to the specific noise environment in order to achieve a significantly better noise estimation performance. Therefore, at first the original Minimum Statistics procedure is briefly revised followed by a description of the necessary modifications.

Original Minimum Statistics Approach

The original Minimum Statistics approach [Mar01] performs well in stationary and slowly changing noise conditions as the minimum at each frequency bin within a search time window provides a good estimate of the actual noise power. However, when it comes to a sudden rise in the noise power in one specific frequency bin, Minimum Statistics is not able to track this rise due to the large window length D_{MS}, which should correspond to a duration of approximately 1.5 seconds [Mar01], see Sec. 2.3.2.

In the original Minimum Statistics approach, the noisy periodogram $|Y(\lambda,\mu)|^2$ is recursively smoothed over time. The smoothed signal power $\hat{\sigma}_Y^2(\lambda,\mu)$ is given by:

$$\hat{\sigma}_Y^2(\lambda,\mu) = \alpha_{\mathrm{MS}}(\lambda,\mu)\cdot\hat{\sigma}_Y^2(\lambda-1,\mu) + (1-\alpha_{\mathrm{MS}}(\lambda,\mu))\cdot|Y(\lambda,\mu)|^2, \tag{5.15}$$

where $\alpha_{\mathrm{MS}}(\lambda,\mu)\in[0,1]$ is the adaptive smoothing factor. The signal powers $\hat{\sigma}_Y^2(\lambda,\mu)$ of the previous D_{MS} frames are buffered in the matrix:

$$\hat{\boldsymbol{\Sigma}}_Y^2(\lambda) = \begin{pmatrix} \hat{\sigma}_Y^2(\lambda-D_{\mathrm{MS}}+1,0) & \dots & \hat{\sigma}_Y^2(\lambda,0) \\ \hat{\sigma}_Y^2(\lambda-D_{\mathrm{MS}}+1,1) & \dots & \hat{\sigma}_Y^2(\lambda,1) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_Y^2(\lambda-D_{\mathrm{MS}}+1,\mu) & \dots & \hat{\sigma}_Y^2(\lambda,\mu) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_Y^2(\lambda-D_{\mathrm{MS}}+1,M_{\mathrm{F}}/2-1) & \dots & \hat{\sigma}_Y^2(\lambda,M_{\mathrm{F}}/2-1) \end{pmatrix}. \tag{5.16}$$

Afterwards, the minimum is tracked in $\hat{\boldsymbol{\Sigma}}_Y^2(\lambda)$ in each row *separately* for each frequency bin according to:

$$\hat{\sigma}_{Y,\min}^2(\lambda,\mu) = \min\left(\hat{\boldsymbol{\sigma}}_Y^2(\lambda,\mu)\right), \tag{5.17}$$

where $\hat{\boldsymbol{\sigma}}_Y^2(\lambda,\mu) = \begin{pmatrix}\hat{\sigma}_Y^2(\lambda-D_{\mathrm{MS}}+1,\mu) & \dots & \hat{\sigma}_Y^2(\lambda,\mu)\end{pmatrix}$ represents the μ-th row of $\hat{\boldsymbol{\Sigma}}_Y^2(\lambda)$. The duration of the time window D_{MS} for the minimum search states

a tradeoff between fast noise tracking and speech distortions after spectral weighting. The minimum value is multiplied by a bias correction factor $B(\lambda, \mu)$ [Mar01], which is mainly dependent on the variance of the noisy signal. The final noise PSD estimation is given by:

$$\hat{\sigma}_N^2(\lambda, \mu) = B(\lambda, \mu) \cdot \hat{\sigma}_{Y,\min}^2(\lambda, \mu). \tag{5.18}$$

Modified Minimum Statistics Approach

In Stage I of the proposed system, a novel modified Minimum Statistics procedure is used to estimate the harmonic noise powers $\sigma_{\mathrm{h},N}^2(\lambda, \mu)$. The new concept is illustrated and compared with the original one in Fig. 5.12. The figure shows the temporal course of four harmonic oscillations over frequency. To determine the noise PSD in frame λ_0 at one particular frequency bin, the original Minimum Statistics algorithm tracks the minimum within the search window by considering entities *only* at this specific frequency bin, see Method (a). In contrast, the modified method adaptively 'looks back' inclined according to the evolution of the harmonics in the time-frequency domain, see Method (b). Following one specific harmonic oscillation over time, the short-term PSDs of the harmonic components are no longer fluctuating that much but relatively stationary. Thus, tracking the minimum along the courses of the harmonics will lead to much better noise estimation results.

In order to achieve this tilted 'viewing direction', the entries of the matrix $\hat{\mathbf{\Sigma}}_Y^2(\lambda)$ need to be modified according to the fundamental frequency f_0. The harmonic oscillation of the current frame λ_0 at frequency $f_0(\lambda_0)$ can be found in the frame $\lambda_0 - D_{\mathrm{MS}} + 1 + j$ at frequency $f_0(\lambda_0 - D_{\mathrm{MS}} + 1 + j)$ with $0 \leq j < D_{\mathrm{MS}}$. In order to estimate the noise power at frame λ_0, the j-th column of the matrix $\hat{\mathbf{\Sigma}}_Y^2(\lambda_0)$ is therefore compressed/expanded according to the ratio:

$$r_{\mathrm{MS}}(\lambda_0, j) = \frac{f_0(\lambda_0)}{f_0(\lambda_0 - D_{\mathrm{MS}} + 1 + j)}. \tag{5.19}$$

After transformation, the j-th column of the modified matrix $\hat{\mathbf{\Sigma}}_{Y,\mathrm{mod}}^2(\lambda_0)$ comprises the noisy signal power at the new positions $\breve{\mu} = \frac{\mu}{r_{\mathrm{MS}}(\lambda_0, j)}$. For this curve fitting problem, conventional linear interpolation [Mei02] is used. The proposed frequency warping technique is schematically illustrated in Fig. 5.13. Visually speaking, the course of the harmonics is brought into a horizontal position within the time-frequency domain. Afterwards, the original minimum tracking concept of the Minimum Statistics algorithm can be applied to the 'warped' spectrogram.

In the case of compression ($r_{\mathrm{MS}} < 1$), the elements of the j-th column are missing for $\breve{\mu} > M_{\mathrm{F}}/2 - 1$ and replaced by $\hat{\sigma}_Y^2(\lambda_0 - D_{\mathrm{MS}} + 1 + j, M_{\mathrm{F}}/2 - 1)$.

Based on the modified matrix $\hat{\mathbf{\Sigma}}_{Y,\mathrm{mod}}^2(\lambda)$, the minimum is again tracked for each row as in the original approach:

$$\hat{\sigma}_{Y,\mathrm{mod},\min}^2(\lambda, \mu) = \min\left(\hat{\boldsymbol{\sigma}}_{Y,\mathrm{mod}}^2(\lambda, \mu)\right), \tag{5.20}$$

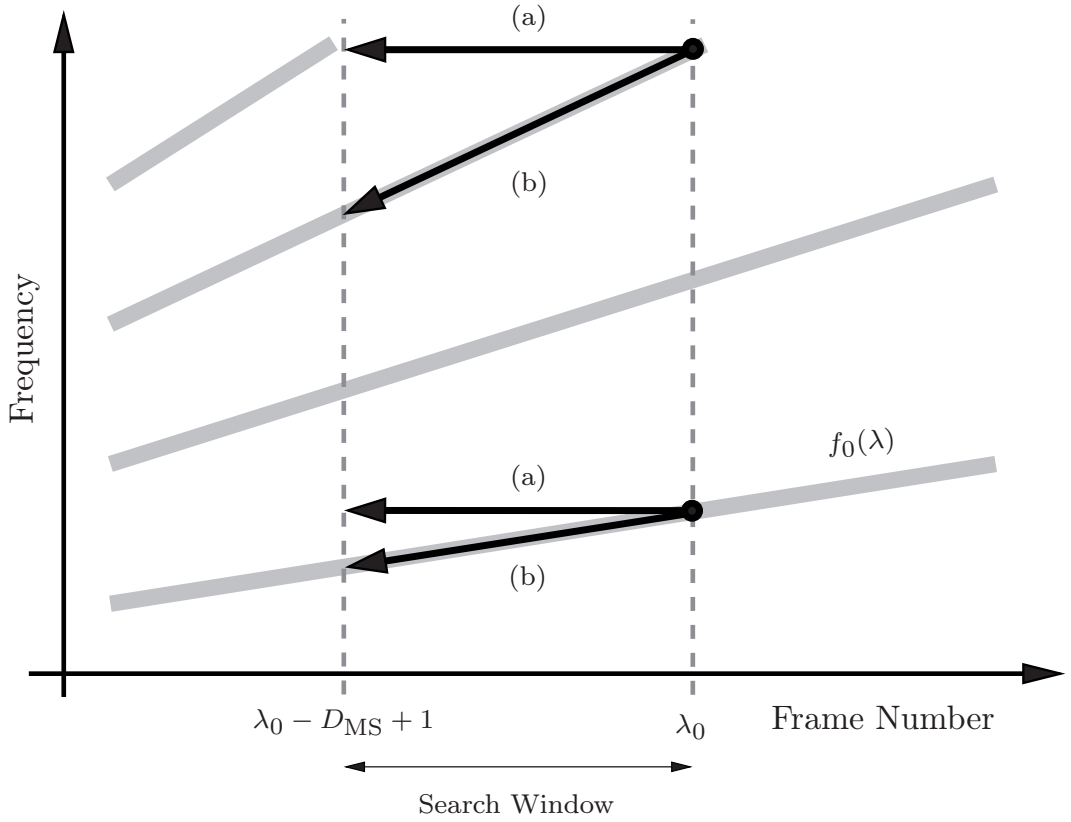

Figure 5.12: 'Direction of view' of (a) original Minimum Statistics and (b) modified Minimum Statistics algorithm.

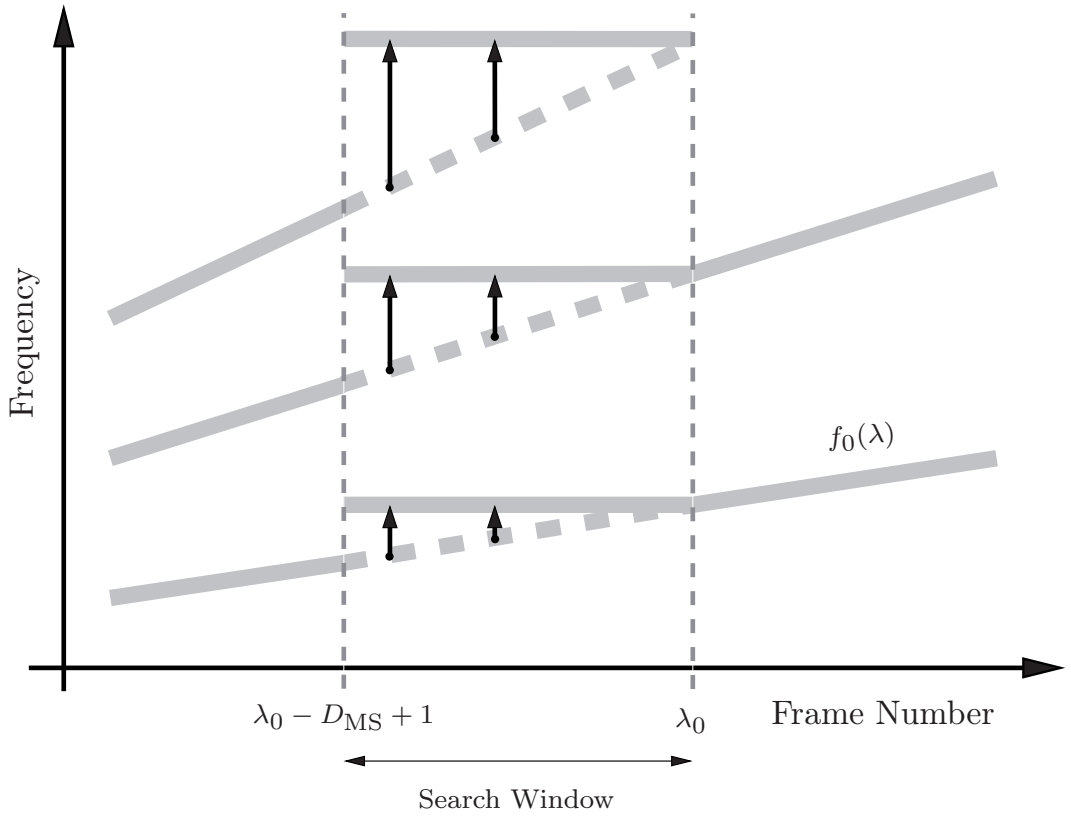

Figure 5.13: Frequency warping of all frames within search window according to the ratio $r_{\mathrm{MS}}(\lambda_0, j)$ using linear interpolation.

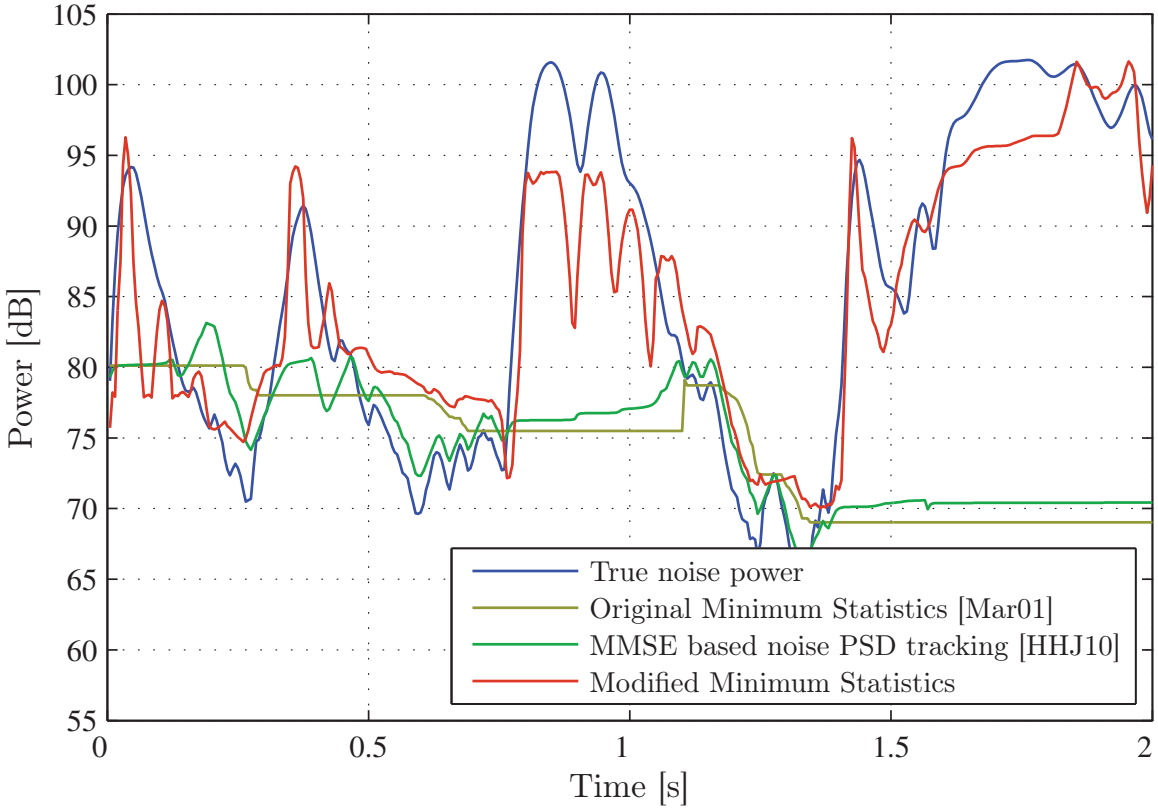

Figure 5.14: Noise PSD estimation results for a noisy speech segment at frequency bin $\mu = 25$ using an FFT length $M_{\mathrm{F}} = 256$. The speech signal was taken from the NTT speech database [NC94] and disturbed by a real noise signal with time-varying harmonic components at 0 dB input SNR.

where $\hat{\boldsymbol{\sigma}}^2_{Y,\mathrm{mod}}(\lambda, \mu)$ represents the μ-th row of $\hat{\boldsymbol{\Sigma}}^2_{Y,\mathrm{mod}}(\lambda)$. From the Minimum Statistics' point of view, the harmonics in the time-frequency domain of $\hat{\boldsymbol{\Sigma}}^2_{Y,\mathrm{mod}}(\lambda)$ appear more stationary over time. Finally, the bias is calculated according to the original Minimum Statistics approach and Eq. 5.18 is applied.

For the determination of the adaptive smoothing factors α_{MS} (see Eq. 5.15), estimates from the previous frame are required as well [Mar01]. Before applying these estimates, they are 'warped' in the same way by using the ratio $r_{\mathrm{MS}}(\lambda_0, D_{\mathrm{MS}} - 2)$.

Figure 5.14 shows an example of noise PSD estimation comparing the novel approach with the original Minimum Statistics approach and the MMSE based noise PSD tracking algorithm proposed in [HHJ10]. The figure shows the estimation results of the three techniques as well as the true noise PSD for one specific frequency bin over time. Therefore, a speech signal was disturbed by a time-varying harmonic noise signal at 0 dB input SNR and the fundamental frequency f_0 was available directly from the engine control. It can clearly be seen that the two conventional estimation techniques have problems in tracking sudden noise variations occurring in acceleration or deceleration phases. In contrast, the modified Minimum Statistics algorithm is able to incorporate adjacent frequency bins from the past leading to a significantly better noise PSD tracking result in the considered environment.

5.2.3 Random Noise PSD Estimation

The first stage of the proposed noise suppression system effectively reduces (rapidly) time-varying harmonic noise, originating, e.g., from a vehicle's engine. However, as the modified noise estimation technique is adapted to the fundamental frequency, the random stationary noise components (e.g., wind or tire noise) would be suppressed by this first stage only satisfactorily in the case of white noise. Therefore, a second stage is applied that reduces the random parts of the noise. As depicted in Fig. 5.11, conventional noise estimation techniques can be used for this purpose, e.g., the original Minimum Statistics approach [Mar01] or the MMSE based noise PSD tracking algorithm [HHJ10] as investigated in the following evaluation.

5.2.4 Performance Results

The performance of the proposed noise estimation technique for the application in harmonic and random noise environments is compared with the results of the original Minimum Statistics approach [Mar01] and the MMSE based noise PSD tracking algorithm [HHJ10]. Therefore, the speech enhancement system depicted in Fig. 5.11 is used incorporating f_0 which is directly provided by the engine control unit of a vehicle. The a priori SNR is estimated according to the decision-directed approach [EM84] and the well-known Wiener filter rule [LO79] is used to calculate the spectral weighting gains. Referring to Fig. 5.11, the following noise estimation techniques are applied in Stages I and II:

Method	Stage I	Stage II
A	disabled (G_h=1)	Minimum Statistics [Mar01]
B	disabled (G_h=1)	MMSE based noise PSD tracking [HHJ10]
C	modified Minimum Statistics (see Sec. 5.2.2)	disabled (G_s=1)
D	modified Minimum Statistics (see Sec. 5.2.2)	Minimum Statistics [Mar01]
E	modified Minimum Statistics (see Sec. 5.2.2)	MMSE based noise PSD tracking [HHJ10]

For the evaluation, four different (real) noise signals are used which were recorded inside a car during acceleration and deceleration phases, i.e., the signals contain a relatively large portion of rapidly time-varying harmonic engine noise. The recordings are each added to three male and two female speech sequences (each with a length of 8 seconds taken randomly from the NTT speech database [NC94]) at input SNR values varying between -10 dB and 35 dB with an increment of 5 dB. The parameters which are used in the simulations are listed in Tab. 5.3.

For the instrumental evaluation, on the one hand the same instrumental measurements as in Chapters 3 and 4 are applied using the segmental noise and speech

Parameter	*Settings*
Sampling frequency	8 kHz
Frame length L_{F}	160 (20 ms)
FFT length M_{F}	256 (including zero-padding)
Frame overlap	50% (Hann window)

Table 5.3: System settings.

attenuation as well as the segmental speech SNR (see Appendix D). On the other hand, the noise tracking performances of the different techniques are explicitly analyzed. Therefore, the log-error distortion measure LogERR is used which is defined as follows [HJH08]:

$$\mathrm{LogERR} = \frac{1}{M_{\mathrm{F}} N_{\mathrm{F}}} \sum_{\mu=0}^{M_{\mathrm{F}}-1} \sum_{\lambda=0}^{N_{\mathrm{F}}-1} \left| 10 \log_{10} \frac{\tilde{\sigma}_N^2(\lambda,\mu)}{\hat{\sigma}_N^2(\lambda,\mu)} \right|, \tag{5.21}$$

where N_{F} represents the total number of evaluated frames and $\hat{\sigma}_N^2(\lambda,\mu)$ states the noise power estimate of the respective investigated technique, i.e., the results of the original or the modified Minimum Statistics or the MMSE based noise PSD tracking algorithm. As it is almost impossible to obtain one overall noise estimate for the proposed 2-stage system, i.e., Methods D and E, only the noise tracking capabilities of the modified Minimum Statistics algorithm (Method C) are investigated. The reference value $\tilde{\sigma}_N^2(\lambda,\mu)$ in Eq. 5.21 is given as a smoothed version of the original noise periodogram according to:

$$\tilde{\sigma}_N^2(\lambda,\mu) = 0.9 \cdot \tilde{\sigma}_N^2(\lambda-1,\mu) + 0.1 \cdot |N(\lambda,\mu)|^2 . \tag{5.22}$$

The smoothing factor 0.9 provides good noise tracking capabilities of $\tilde{\sigma}_N^2(\lambda,\mu)$ while obtaining a reduced temporal variance compared to the true noise PSD, see [HHJ10] and [TTM+11]. The lower the value of LogERR, the better the performance of the noise estimation technique. The measure LogERR is given in dB.

The averaged results for the segmental noise and speech attenuation as well as the speech SNR are depicted in Figs. 5.15 and 5.16. Figure 5.15 shows the difference between noise and speech attenuation where higher scores indicate a better performance of the respective approach. It can be seen that all methods using the modified Minimum Statistics approach, i.e., Methods C, D and E consistently improve the results of the conventional noise estimation techniques (Methods A and B). Moreover, the proposed 2-stage system (Methods D and E) achieves a better tradeoff between noise and speech attenuation when compared to the modified Minimum Statistics algorithm in Method C. Overall, the combined system consisting of the modified Minimum Statistics and the MMSE based noise PSD tracking algorithm yields the best performance in this measurement and outperforms the other approaches.

In Fig. 5.16, the segmental speech SNR is plotted over the noise attenuation. The aim in this graph is to achieve a high segmental speech SNR and a high noise attenuation. Hence, the more the respective curve is placed in the upper right corner, the better

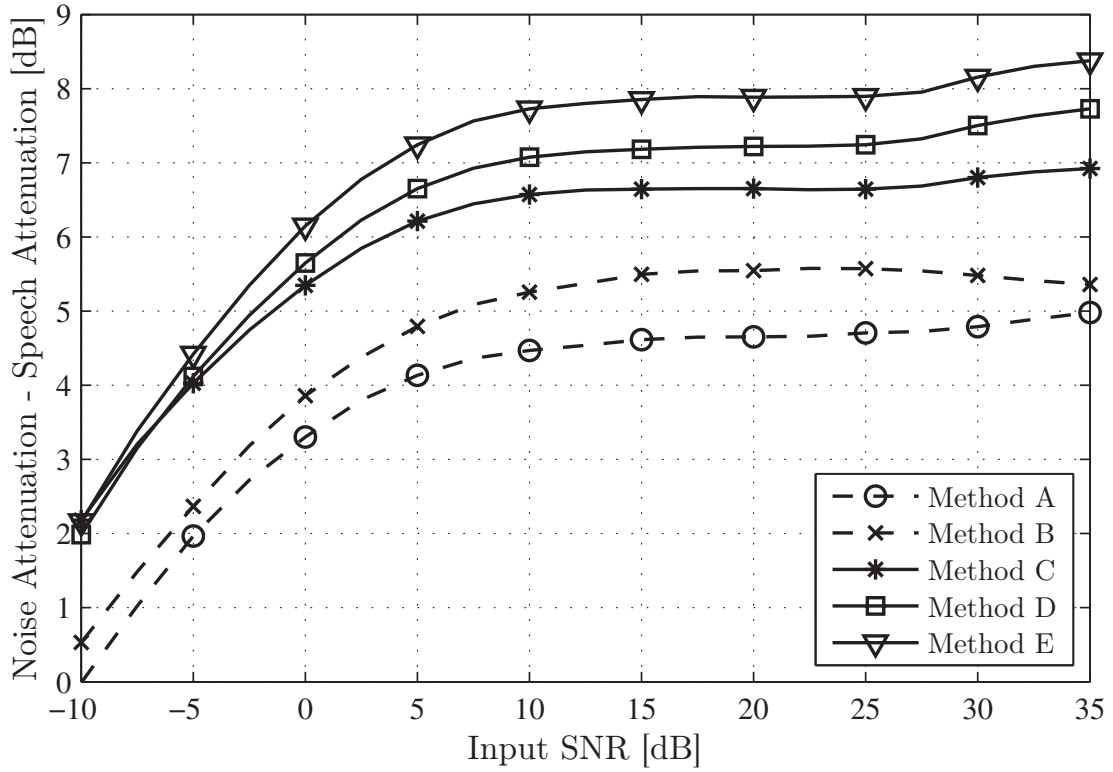

Figure 5.15: Difference between noise attenuation and speech attenuation plotted over input SNR. The different methods are explained on page 117.

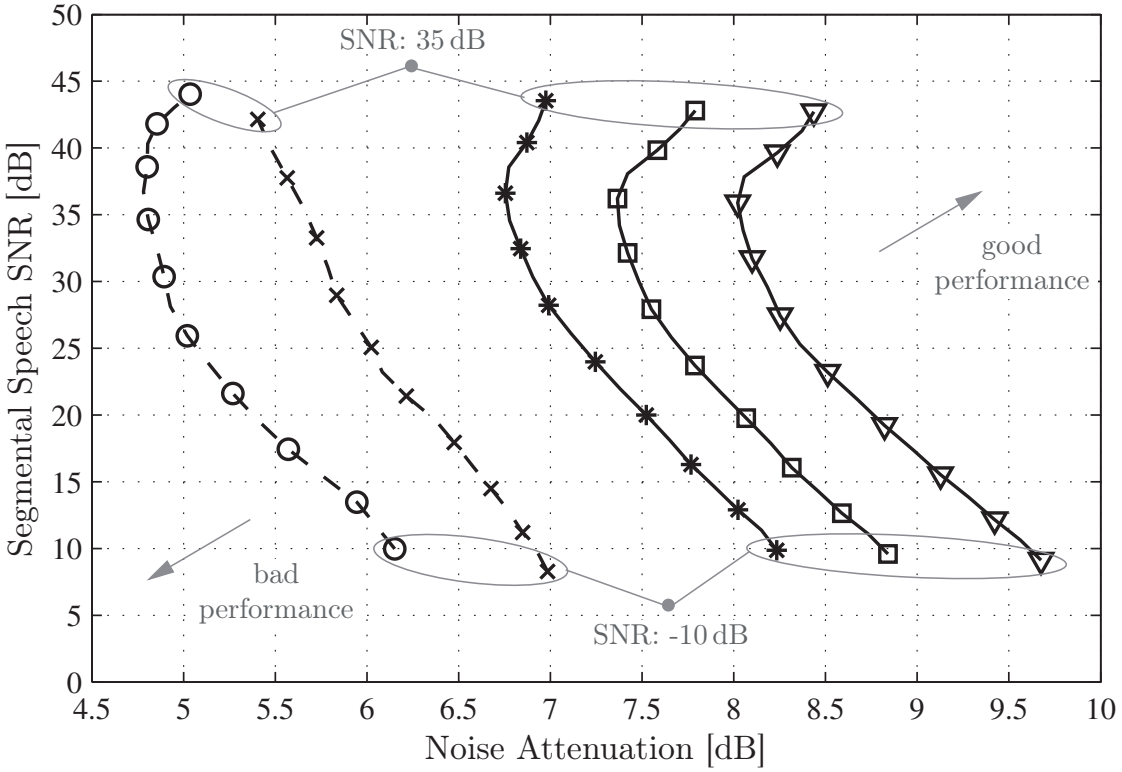

Figure 5.16: Segmental speech SNR plotted over noise attenuation.

the performance. The results show again the improvements of the new Methods C, D and E compared to the original Minimum Statistics approach (Method A) and the MMSE based noise PSD tracking algorithm (Method B). Method B is able to achieve a higher noise attenuation than Method A. However, this benefit comes at the expense of a lower segmental speech SNR comparing the particular input SNR markers. Regarding the proposed techniques, the 2-stage system again provides the best compromise in terms of noise attenuation and segmental speech SNR. In contrast to the measurements in Fig. 5.15, the choice for the conventional noise estimation technique in the second stage of Fig. 5.11 can not be conclusively attributed to the MMSE based noise PSD tracking algorithm and depends on the application. Whereas Method E outperforms all other approaches in terms of noise attenuation, it affects the speech quality slightly more than Method D especially at low input SNR values. Interestingly, when using Methods A, C, D and E, the amount of noise attenuation is not decreasing monotonically in this environment for higher input SNR values. Instead, there is a gain in noise suppression beyond 25 dB input SNR as can be seen in Fig. 5.16. A possible explanation for this behavior can be the fact that the original and modified Minimum Statistics approaches are able to track the power of the stationary noise components proportionally better in good SNR conditions where the influence of the time-varying harmonics decreases.

Figure 5.17 depicts the averaged results for the LogERR measure plotted over the input SNR for Methods A, B and C. It can clearly be seen that the modified Minimum Statistics approach (Method C) achieves the lowest distortion measures for almost the entire SNR range and thus provides the best noise estimation performance in this specific noise environment. Especially at low input SNR values, a clear preference for the modified Minimum Statistics algorithm can be observed. In good SNR conditions however, the distortion measure of Method C increases stronger than that of the other two noise estimation techniques. As mentioned before, the influence of the time-varying harmonic noise components decreases at high input SNR values. Hence, the advantage of the proposed frequency warping in order to reduce the harmonics diminishes beyond 30 dB input SNR and the modified Minimum Statistics approach achieves a less precise noise estimation performance compared to Methods A and B. To counteract this problem, a subsequent (conventional) noise reduction system is added in the proposed system as depicted in Fig. 5.11.

Spectrograms of the processed signals are shown in Fig. 5.18. In the upper row, the spectrogram of the clean and the noisy input signals are depicted. The speech sentences "A wisp of cloud hung in the blue air. A pound of sugar costs more than eggs." are taken from the NTT database [NC94] and disturbed at 5 dB input SNR by a real noise signal recorded inside a car. In addition to stationary background noise, it can be seen that the engine mainly contributes to the noise signal. The speech signal is highly disturbed by the spectral harmonics. The spectrograms of the processed signals are shown in the middle and lower rows for the different approaches A, B, D and E. While the conventional noise estimation (Methods A and B) fails in this noise environment (stationary background noise slightly reduced but spectral harmonics remain almost unchanged), the new approaches (Methods D and E) perform significantly better and are able to suppress a larger amount of the engine and stationary

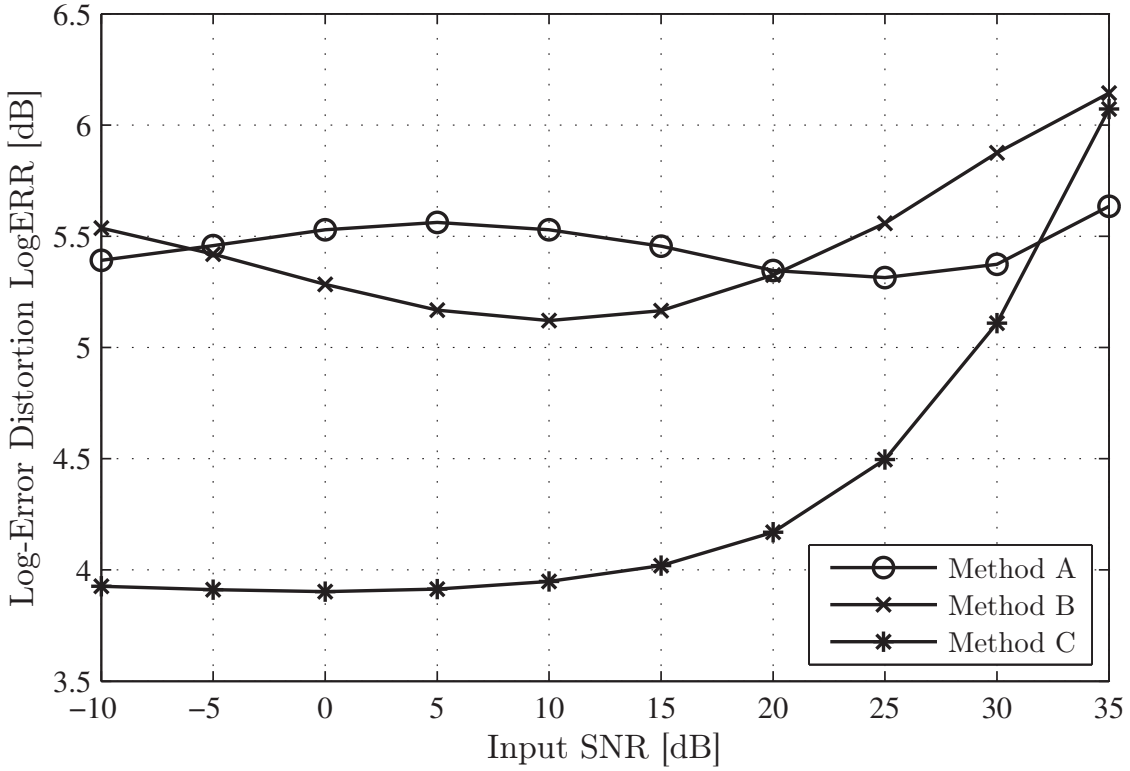

Figure 5.17: Log-error distortion LogERR plotted over input SNR. An explanation of the different methods can be found on page 117.

background noise. As can be seen, the harmonics and the stationary background noise are effectively reduced leading to a more comfortable listening condition.

In principle, the application of the proposed 'warping' technique is not restricted to the Minimum Statistics algorithm and can be applied to any noise estimation method which takes into account noisy or enhanced DFT coefficients from previous frames.

5.2.5 Conclusions

In this section, a novel noise PSD estimation algorithm is proposed for the application in noisy environments consisting of time-varying harmonic and stationary random noise. Conventional noise estimation techniques usually fail in this specific environment as the tracking of rapidly time-varying noise often leads to an underestimation of the noise power. Thus, the harmonic noise components are reduced in a first stage by using a modified Minimum Statistics approach which performs frequency warping according to the harmonic's fundamental frequency in order to track and suppress the harmonic noise quite effectively. The remaining random noise components in the signal are estimated and reduced in a second stage using conventional noise estimation techniques. Instrumental measurements show a consistent improvement in terms of noise/speech attenuation and segmental speech SNR compared to the original Minimum Statistics approach [Mar01] and the MMSE based noise PSD tracking algorithm [HHJ10].

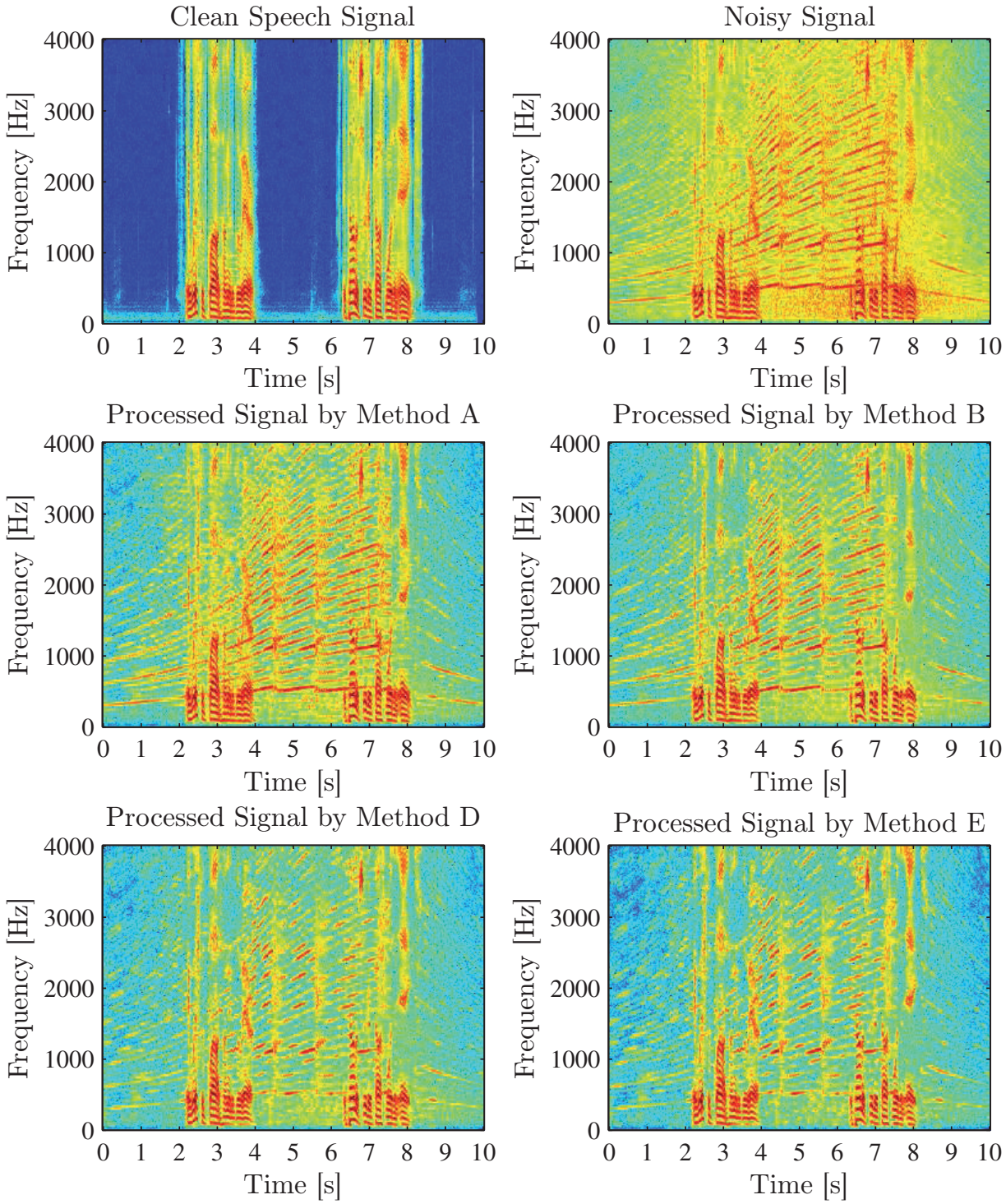

Figure 5.18: Spectrograms of clean speech signal, noisy signal (speech+car noise, SNR: 5 dB), and processed signals with Method A: enhanced signal using original Minimum Statistics approach [Mar01], Method B: enhanced signal using original MMSE based noise PSD tracking algorithm [HHJ10], Method D: enhanced signal using new approach by combining modified and original Minimum Statistics approach and Method E: enhanced signal using new approach by combining modified Minimum Statistics and MMSE based noise PSD tracking algorithm. The sentences "A wisp of cloud hung in the blue air. A pound of sugar costs more than eggs." are spoken by a male voice. An explanation of the different methods can be found on page 117.

6

Summary

This thesis addresses the problem of single-channel speech enhancement for the application in mobile phones, conferencing systems, digital hearing aids or hands-free devices. The developed algorithms focus on exploiting *temporal and spectral dependencies* of speech as well as noise signals. In addition, they provide novel solutions for suppressing *musical noise* and *rapidly time-varying harmonic noise*. All speech enhancement techniques within this work have been thoroughly evaluated by means of instrumental measurements and auditory judgments. It turned out that the proposed noise reduction techniques achieve distinctly better results compared to state-of-the-art approaches with respect to noise attenuation and speech distortions.

In the *first part* of this thesis, a novel Kalman filter approach for noise suppression in the frequency domain has been presented. The solution is based on a modified propagation and update step which are both applied directly to the complex-valued DFT coefficients of the noisy input signal. In the propagation step, *temporal correlation* of successive frames is exploited using low-order models to approximate the trajectories of the speech and the noise DFT coefficients. It has been shown that complex-valued linear prediction yields higher prediction gains than estimating real and imaginary parts or magnitudes and phases separately. The proposed system is able to exploit temporal correlation of speech already at very low input SNR values and nearly reaches the level of 'ideal' prediction at 20 dB input SNR. In the second (update) step, the first predictions are updated utilizing an appropriate statistical weighting rule in order to estimate the prediction errors caused in the propagation step. As novelty, not only the conventional Kalman filter gain (assuming Gaussian distributions for speech and noise) has been taken into account for this purpose but also different SNR-dependent MMSE estimators which have been explicitly adapted to the measured histograms of the speech prediction error signal. Moreover, a new possibility to estimate the prediction error powers of speech and noise has been presented. In the evaluation, the proposed system has clearly outperformed several purely statistical estimators as well as the Kalman filter approach presented in [ZVY06b]. Especially the incorporation of the SNR-dependency on the statistics of the speech prediction error has led to significant improvements. The instrumental measurements have been confirmed by an informal listening test in which about 75% of the test listeners preferred the signals processed by the novel Kalman filter solution. Compared to

state-of-the-art noise suppression techniques, the overall computational load of the proposed system has been increased by a factor of $5-7$. However, possible methods for an effective reduction of the complexity have been presented.

In literature, only very few publications can be found so far which explicitly cover wideband ($50\,\text{Hz}-7\,\text{kHz}$) noise reduction. Almost all known approaches process the low band ($50\,\text{Hz}-4\,\text{kHz}$) and the high band ($4\,\text{kHz}-7\,\text{kHz}$) components in the same way. In the *second part* of this work, a new possibility to exploit *spectral dependencies* of speech signals by means of wideband speech enhancement has been presented. The developed method uses techniques known from *Artificial Bandwidth Extension* (ABWE) to improve the results of a conventional noise suppression technique in the high band. Therefore, ABWE is applied to the processed (enhanced) low band signal and utilized in order to estimate subband energies of the high band. The resulting weighting gains determined from these energy estimates are combined with gains that are additionally obtained by a conventional noise reduction technique in the high band. For this purpose, cross-fading with an adaptive factor is used which is depending on input SNR estimates. The performance of this novel wideband speech enhancement system has been shown to be consistently better than state-of-the-art noise reduction approaches with respect to noise attenuation and speech distortions. Moreover, informal listening tests have revealed that the occurrence of musical tones can be slightly reduced by the proposed method. The results have been verified by information theoretic considerations which quantify the existence of spectral dependencies between low and high frequencies already at very low input SNR values. In addition, the mutual information between the low and high band could be significantly increased if noise suppression was applied prior to ABWE. Overall, the proposed technique is not strictly restricted to two broad bands. Using a modified training process, the system can be adapted to support the speech enhancement in an arbitrary frequency band or even individual frequency bin.

The *third part* of this thesis has covered postprocessing methods to improve the overall speech quality of a communication system. Noise suppression techniques often suffer from generating annoying musical tones, especially when they yield a good compromise between noise attenuation and speech distortions in other respects. Moreover, most speech enhancement algorithms fail as soon as the environmental noise becomes non-stationary. Both problems have been successfully tackled in this third part.
In order to suppress *musical tones*, two effective postprocessing methods have been presented which can be applied to the spectral weighting gains of an arbitrary noise reduction technique. The first technique utilizes soft-decisions of a low input SNR detector in order to adaptively smooth the spectral gains over frequency. The second approach uses a framewise adaptive frequency resolution such that the spectral resolution is higher during speech activity and lower during speech pauses. Both techniques can be beneficially concatenated. Instrumental measurements have shown that both musical noise countermeasures yield improvements when they are applied subsequent to a noise reduction system. Furthermore, the results of an informal listening tests clearly demonstrated the advantages of the developed approaches with respect to musical noise suppression. In total, approximately 92% of the test listeners preferred the samples generated by the proposed postprocessing techniques.

In order to cope with *rapidly time-varying harmonic* noise and *stationary random* noise, a novel noise PSD estimation algorithm has been presented. The harmonic noise components are effectively reduced in a first stage by using a modified Minimum Statistics approach which performs frequency warping according to the harmonic's fundamental frequency. The remaining random noise components in the signal are estimated and suppressed in a second stage using conventional noise estimation techniques. While the investigated state-of-the-art noise estimation techniques have failed in the considered noise environment, instrumental measurements have shown that the new approach performs significantly better and is able to suppress a larger amount of the harmonic and stationary background noises leading to a more comfortable listening condition.

A

Derivations

A.1 Kalman Filter Equations

In the following, the Kalman filter equations used in the update step are determined based on the derivations which can be found in [Mey00] and [SC08]. In order to derive the conditional expectation $\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\}$, the conditional *Probability Density Function* (PDF) $p(\mathbf{s}(k)|\mathbf{y}(k))$ is required. Using Bayes' theorem [Bay63], $p(\mathbf{s}(k)|\mathbf{y}(k))$ is given by:

$$p(\mathbf{s}(k)|\mathbf{y}(k)) = \frac{p(\mathbf{y}(k), \mathbf{s}(k))}{p(\mathbf{y}(k))}. \tag{A.1}$$

At first, the joint PDF $p(\mathbf{y}(k), \mathbf{s}(k))$ is determined based on a new vector $\mathbf{w}$ which combines $\mathbf{y}(k)$ and $\mathbf{s}(k)$ according to:

$$\mathbf{w} = \left(\, \mathbf{y}^H(k),\, \mathbf{s}^H(k) \,\right)^H = (\, y(0),\, \ldots,\, y(k),\, s(k - N_\mathrm{K} + 1),\, \ldots,\, s(k) \,)^H . \tag{A.2}$$

Assuming that $\mathbf{y}(k)$, $\mathbf{s}(k)$ as well as the additive noise signal $n(k)$ are Gaussian distributed, the PDF $p(\mathbf{w}) = p(\mathbf{y}(k), \mathbf{s}(k))$ follows a Gaussian distribution as well. Its mean $\bar{\mathbf{w}}$ and covariance $\mathbf{M}$ are given by:

$$\bar{\mathbf{w}} = \mathbb{E}\{\mathbf{w}\} = \begin{pmatrix} \mathbb{E}\{\mathbf{y}(k)\} \\ \mathbb{E}\{\mathbf{s}(k)\} \end{pmatrix}, \tag{A.3}$$

and

$$\begin{aligned} \mathbf{M} &= \mathbb{E}\{(\mathbf{w} - \mathbb{E}\{\mathbf{w}\})\,(\mathbf{w} - \mathbb{E}\{\mathbf{w}\})^H\} \\ &= \begin{pmatrix} \mathbf{M}_{yy} & \mathbf{M}_{ys} \\ \mathbf{M}_{sy} & \mathbf{M}_{ss} \end{pmatrix}, \end{aligned} \tag{A.4}$$

where $\mathbf{M}_{yy} = \mathbb{E}\{(\mathbf{y}(k) - \mathbb{E}\{\mathbf{y}(k)\}) \cdot (\mathbf{y}(k) - \mathbb{E}\{\mathbf{y}(k)\})^H\}$, $\mathbf{M}_{ys} = \mathbb{E}\{(\mathbf{y}(k) - \mathbb{E}\{\mathbf{y}(k)\}) \cdot (\mathbf{s}(k) - \mathbb{E}\{\mathbf{s}(k)\})^H\}$, $\mathbf{M}_{sy} = \mathbb{E}\{(\mathbf{s}(k) - \mathbb{E}\{\mathbf{s}(k)\}) \cdot (\mathbf{y}(k) - \mathbb{E}\{\mathbf{y}(k)\})^H\}$ and $\mathbf{M}_{ss} =$

$\mathbb{E}\{(\mathbf{s}(k) - \mathbb{E}\{\mathbf{s}(k)\}) \cdot (\mathbf{s}(k) - \mathbb{E}\{\mathbf{s}(k)\})^H\}$. The PDF $p(\mathbf{w}) = p(\mathbf{y}(k), \mathbf{s}(k))$ results in [SC08]:

$$p(\mathbf{w}) = p(\mathbf{y}(k), \mathbf{s}(k)) = \frac{1}{(2\pi)^{(k+N_\mathrm{K})/2}|\mathbf{M}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^H \mathbf{M}^{-1}(\mathbf{w} - \bar{\mathbf{w}})\right). \tag{A.5}$$

Since the PDF $p(\mathbf{y}(k))$ is a Gaussian distribution as well with mean $\mathbb{E}\{\mathbf{y}(k)\}$ and covariance $\mathbf{M}_{yy}$, the conditional PDF $p(\mathbf{s}(k)|\mathbf{y}(k))$ can be determined using Eq. A.1. Finally, the required conditional expectation $\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\}$ is derived according to:

$$\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{s}(k) \cdot p(\mathbf{s}(k)|\mathbf{y}(k)) \, \mathrm{d}\mathbf{s}(k). \tag{A.6}$$

In the update step, the conditional mean $\hat{\mathbf{s}}_\mathrm{prop}(k) = \mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k-1)\}$ and covariance matrix $\mathbf{P}^s_\mathrm{prop}(k)$ from the propagation step are already available and only the current measurement $y(k)$ needs to be incorporated to obtain $\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\}$. Using Eqs. 3.14 and 3.15, Eq. A.6 is therefore given by [Mey00, SC08]:

$$\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\} = \mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k-1)\} + \mathbf{M}_{sy}\mathbf{M}_{yy}^{-1}(y(k) - \mathbb{E}\{y(k)|\mathbf{y}(k-1)\}), \tag{A.7}$$

which can further be derived to:

$$\begin{aligned}
\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\} &= \hat{\mathbf{s}}_\mathrm{prop}(k) + \mathbf{M}_{sy}\mathbf{M}_{yy}^{-1}(y(k) - \mathbb{E}\{\mathbf{h}_s^T\mathbf{s}(k) + n(k)|\mathbf{y}(k-1)\}) \\
&= \hat{\mathbf{s}}_\mathrm{prop}(k) + \mathbf{M}_{sy}\mathbf{M}_{yy}^{-1}(y(k) - \mathbb{E}\{\mathbf{h}_s^T\mathbf{s}(k)|\mathbf{y}(k-1)\}) \\
&= \hat{\mathbf{s}}_\mathrm{prop}(k) + \mathbf{M}_{sy}\mathbf{M}_{yy}^{-1}(y(k) - \mathbf{h}_s^T\hat{\mathbf{s}}_\mathrm{prop}(k)),
\end{aligned} \tag{A.8}$$

with [Mey00]:

$$\begin{aligned}
\mathbf{M}_{sy} &= \mathbb{E}\{(\mathbf{s}(k) - \mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k-1)\}) \cdot (y(k) - \mathbb{E}\{y(k)|\mathbf{y}(k-1)\})^H\} \\
&= \mathbb{E}\{(\mathbf{s}(k) - \hat{\mathbf{s}}_\mathrm{prop}(k)) \cdot \left(\mathbf{h}_s^T\mathbf{s}(k) + n(k) - \mathbf{h}_s^T\hat{\mathbf{s}}_\mathrm{prop}(k)\right)^H\} \\
&= \mathbb{E}\{(\mathbf{s}(k) - \hat{\mathbf{s}}_\mathrm{prop}(k)) \cdot \left(\mathbf{h}_s^T(\mathbf{s}(k) - \hat{\mathbf{s}}_\mathrm{prop}(k)) + n(k)\right)^H\} \\
&= \mathbf{P}^s_\mathrm{prop}(k) \cdot \mathbf{h}_s,
\end{aligned} \tag{A.9}$$

and

$$\begin{aligned}
\mathbf{M}_{yy} &= \mathbb{E}\{(y(k) - \mathbb{E}\{y(k)|\mathbf{y}(k-1)\}) \cdot (y(k) - \mathbb{E}\{y(k)|\mathbf{y}(k-1)\})^H\} \\
&= \mathbb{E}\{\left(\mathbf{h}_s^T(\mathbf{s}(k) - \hat{\mathbf{s}}_\mathrm{prop}(k)) + n(k)\right) \cdot \left(\mathbf{h}_s^T(\mathbf{s}(k) - \hat{\mathbf{s}}_\mathrm{prop}(k)) + n(k)\right)^H\} \\
&= \mathbf{h}_s^T \cdot \mathbf{P}^s_\mathrm{prop}(k) \cdot \mathbf{h}_s + \sigma_n^2(k).
\end{aligned} \tag{A.10}$$

Combining Eqs. A.8, A.9 and A.10 results in:

$$\mathbb{E}\{\mathbf{s}(k)|\mathbf{y}(k)\} = \hat{\mathbf{s}}_\mathrm{up}(k) = \hat{\mathbf{s}}_\mathrm{prop}(k) + \mathbf{k}^s(k)\left(y(k) - \mathbf{h}_s^T\hat{\mathbf{s}}_\mathrm{prop}(k)\right), \tag{A.11}$$

with

$$\mathbf{k}^s(k) = \mathbf{P}^s_\mathrm{prop}(k) \cdot \mathbf{h}_s \left(\mathbf{h}_s^T\mathbf{P}^s_\mathrm{prop}(k)\mathbf{h}_s + \sigma_n^2(k)\right)^{-1}. \tag{A.12}$$

A.2 MMSE Estimation in Update Step under Gaussian Assumption

In the update step of the Kalman filter, the objective is to estimate the speech prediction error vector $\mathbf{e}^s_{\text{prop}}(k)$ having access only to the differential signal $d_s(k)$ given by:

$$d_s(k) = e^s_{\text{prop}}(k) + n(k), \tag{A.13}$$

where $e^s_{\text{prop}}(k)$ and $n(k)$ are statistically independent. Performing a *Minimum Mean Square Error* (MMSE) estimation of $\mathbf{e}^s_{\text{prop}}(k)$, i.e., minimizing $\mathbb{E}\{(\mathbf{e}^s_{\text{prop}}(k) - \hat{\mathbf{e}}^s_{\text{prop}}(k))^2\}$, the solution equals the conditional expectation $\mathbb{E}\{\mathbf{e}^s_{\text{prop}}(k)|d_s(k)\}$ assuming Gaussian models for $\mathbf{e}^s_{\text{prop}}(k)$ and $n(k)$ as shown in the following.

The derivation of $\mathbb{E}\{\mathbf{e}^s_{\text{prop}}(k)|d_s(k)\}$ is decomposed into two parts. In the first part, the conditional expectation $\mathbb{E}\{e^s_{\text{prop}}(k)|d_s(k)\}$ is determined based on the following Gaussian *Probability Density Functions* (PDFs):

$$p(d_s(k)) = \frac{1}{\sqrt{2\pi\sigma^2_{d_s}(k)}} \cdot \exp\left(-\frac{d_s^2(k)}{2\sigma^2_{d_s}(k)}\right), \tag{A.14}$$

$$p(e^s_{\text{prop}}(k)) = \frac{1}{\sqrt{2\pi\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}}} \cdot \exp\left(-\frac{\left(e^s_{\text{prop}}(k)\right)^2}{2\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}}\right) \quad \text{and} \tag{A.15}$$

$$p(d_s(k)|e^s_{\text{prop}}(k)) = \frac{1}{\sqrt{2\pi\sigma^2_n(k)}} \cdot \exp\left(-\frac{\left(d_s(k) - e^s_{\text{prop}}(k)\right)^2}{2\sigma^2_n(k)}\right), \tag{A.16}$$

where $\sigma^2_{d_s}(k) = \mathbb{E}\{d_s^2(k)\}$. In the second part, an expression for the remaining components $\mathbb{E}\left\{\left(e^s_{\text{prop}}(k - N_{\text{K}} + 1), ..., e^s_{\text{prop}}(k-1)\right)|d_s(k)\right\}$ is derived.

Using the assumed PDFs given in Eqs. A.14-A.16, the expectation $\mathbb{E}\{e^s_{\text{prop}}(k)|d_s(k)\}$ results in the well-known Wiener filter solution (see, e.g, [LO79], [VM06, Chapter 11])

given by:

$$
\begin{aligned}
&\mathbb{E}\{e^s_{\text{prop}}(k)|d_s(k)\} = ... \\
&\quad= \int_{-\infty}^{\infty} e^s_{\text{prop}}(k) \cdot p(e^s_{\text{prop}}(k)|d_s(k))\,\mathrm{d}e^s_{\text{prop}}(k) \\
&\quad= \int_{-\infty}^{\infty} e^s_{\text{prop}}(k) \cdot \frac{p(d_s(k)|e^s_{\text{prop}}(k))}{p(d_s(k))} \cdot p(e^s_{\text{prop}}(k))\,\mathrm{d}e^s_{\text{prop}}(k) \\
&\quad= \frac{1}{p(d_s(k))} \int_{-\infty}^{\infty} \frac{e^s_{\text{prop}}(k)}{2\pi\sqrt{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k} \cdot \sigma^2_n(k)}} \exp\left(- \frac{\left(d_s(k) - e^s_{\text{prop}}(k)\right)^2}{2\sigma^2_n(k)} \right. \\
&\qquad \left. - \frac{\left(e^s_{\text{prop}}(k)\right)^2}{2\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}} \right) \mathrm{d}e^s_{\text{prop}}(k) \\
&\quad= \frac{\exp\left(-\frac{d^2_s(k)}{2\sigma^2_n(k)}\right)}{p(d_s(k))2\pi\sqrt{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k} \cdot \sigma^2_n(k)}} \int_{-\infty}^{\infty} e^s_{\text{prop}}(k) \exp\left(\frac{d_s(k)e^s_{\text{prop}}(k)}{\sigma^2_n(k)} \right. \\
&\qquad \left. - \frac{\left(\sigma^2_{e^s_{\text{prop}}}\binom{k}{k} + \sigma^2_n(k)\right)(e^s_{\text{prop}}(k))^2}{2\sigma^2_{e^s_{\text{prop}}}\binom{k}{k} \cdot \sigma^2_n(k)} \right) \mathrm{d}e^s_{\text{prop}}(k) \\
&\quad= \frac{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k} + \sigma^2_n(k)} \cdot d_s(k). \qquad \text{(A.17)}
\end{aligned}
$$

If $\mathbf{e}^s_{\text{prop}}(k)$ follows a multivariate Gaussian distribution, the following condition holds for adjacent samples of the prediction error [KBJ00]:

$$
\mathbb{E}\{e^s_{\text{prop}}(k-\kappa+1)|e^s_{\text{prop}}(k)\} = \frac{\sigma^2_{e^s_{\text{prop}}}\binom{k-\kappa+1}{k}}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}} \cdot e^s_{\text{prop}}(k), \qquad \text{(A.18)}
$$

with $1 \leq \kappa \leq N_{\text{K}}$.

The remaining estimates $\left(\hat{e}^s_{\text{prop}}(k-N_{\text{K}}+1), ..., \hat{e}^s_{\text{prop}}(k-1)\right)$ can be derived using Eqs. A.17 and A.18 as well as general properties of conditional expectation values

[Dur95] according to:

$$\begin{aligned}\mathbb{E}\{e^s_{\text{prop}}(k-\kappa+1)|d_s(k)\} &= \mathbb{E}\left\{\mathbb{E}\left\{e^s_{\text{prop}}(k-\kappa+1)|e^s_{\text{prop}}(k), d_s(k)\right\}|d_s(k)\right\} \\ &= \mathbb{E}\left\{\left.\frac{\sigma^2_{e^s_{\text{prop}}}\binom{k-\kappa+1}{k}}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}} e^s_{\text{prop}}(k)\right| d_s(k)\right\} \\ &= \frac{\sigma^2_{e^s_{\text{prop}}}\binom{k-\kappa+1}{k}}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}}\mathbb{E}\{e^s_{\text{prop}}(k)|d_s(k)\} \\ &= \frac{\sigma^2_{e^s_{\text{prop}}}\binom{k-\kappa+1}{k}}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}+\sigma^2_n(k)}\cdot d_s(k). \end{aligned} \tag{A.19}$$

In vector notation, Eq. A.19 can be written as:

$$\mathbb{E}\{\mathbf{e}^s_{\text{prop}}(k)|d_s(k)\} = \frac{d_s(k)}{\sigma^2_{e^s_{\text{prop}}}\binom{k}{k}+\sigma^2_n(k)}\begin{pmatrix}\sigma^2_{e^s_{\text{prop}}}\binom{k-N_{\text{K}}+1}{k}\\ \vdots \\ \sigma^2_{e^s_{\text{prop}}}\binom{k}{k}\end{pmatrix}, \tag{A.20}$$

and matches Eq. 3.25 in the update step of the Kalman filter.

A.3 Complex-Valued Autoregressive Coefficients

In this section, the optimal *Autoregressive* (AR) coefficients are derived for the purpose of *Linear Prediction* (LP). In contrast to conventional LP applications, e.g., speech coding, the input sequence x is assumed to be complex-valued in the following. Based on previous samples $x(k-i) \in \mathbb{C}$ with $1 \leq i \leq N_{\text{AR}}$, the current sample $x(k)$ is predicted according to:

$$\hat{x}(k) = \sum_{i=1}^{N_{\text{AR}}} a_i \cdot x(k-i), \tag{A.21}$$

by using complex-valued AR coefficients a_i and a model order N_{AR}. The optimal AR coefficients are derived by minimizing the mean square prediction error $e_x(k)$ defined as:

$$e_x(k) = x(k) - \hat{x}(k). \tag{A.22}$$

Therefore, the partial derivation of $\mathbb{E}\{|e_x(k)|^2\}$ with respect to the real part of a_j $(1 \leq j \leq N_{\mathrm{AR}})$ yields:

$$\begin{aligned}
\frac{\partial \mathbb{E}\{|e_x(k)|^2\}}{\partial \mathrm{Re}\{a_j\}} &= \frac{\partial \mathbb{E}\{(x(k)-\hat{x}(k))\,(x(k)-\hat{x}(k))^*\}}{\partial \mathrm{Re}\{a_j\}} \\
&= \frac{\partial \mathbb{E}\left\{\left(x(k)-\sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot x(k-i)\right)\left(x(k)-\sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot x(k-i)\right)^*\right\}}{\partial \mathrm{Re}\{a_j\}} \\
&= \mathbb{E}\Bigg\{-x(k-j)\left(x(k)-\sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot x(k-i)\right)^* \\
&\quad - x^*(k-j)\left(x(k)-\sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot x(k-i)\right)\Bigg\} \\
&= -\varphi_{xx}^*(j) + \sum_{i=1}^{N_{\mathrm{AR}}} a_i^* \cdot \varphi_{xx}^*(j-i) - \varphi_{xx}(j) + \sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot \varphi_{xx}(j-i) \\
&= -2 \cdot \mathrm{Re}\left\{\varphi_{xx}(j) - \sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot \varphi_{xx}(j-i)\right\} \stackrel{!}{=} 0,
\end{aligned} \tag{A.23}$$

with complex-valued autocorrelation function $\varphi_{xx}(j) = \mathbb{E}\{x(k) \cdot x^*(k-j)\}$ having the property that $\varphi_{xx}(-j) = \varphi_{xx}^*(j)$. The expression for the imaginary part can be derived in a similar way and results in:

$$\frac{\partial \mathbb{E}\{|e_x(k)|^2\}}{\partial \mathrm{Im}\{a_j\}} = -2\mathrm{j} \cdot \mathrm{Im}\left\{\varphi_{xx}(j) - \sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot \varphi_{xx}(j-i)\right\} \stackrel{!}{=} 0, \tag{A.24}$$

Combining both derivations, the optimal AR coefficients have to fulfill the following condition:

$$\varphi_{xx}(j) - \sum_{i=1}^{N_{\mathrm{AR}}} a_i \cdot \varphi_{xx}(j-i) = 0. \tag{A.25}$$

For $j = 1, ..., N_{\mathrm{AR}}$, the Yule-Walker equations [PM96] for complex-valued signals arise, given by:

$$\begin{pmatrix} \varphi_{xx}(1) \\ \varphi_{xx}(2) \\ \vdots \\ \varphi_{xx}(N_{\mathrm{AR}}) \end{pmatrix} = \underbrace{\begin{pmatrix} \varphi_{xx}(0) & \varphi_{xx}^*(1) & \dots & \varphi_{xx}^*(1-N_{\mathrm{AR}}) \\ \varphi_{xx}(1) & \varphi_{xx}(0) & \dots & \varphi_{xx}^*(2-N_{\mathrm{AR}}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{xx}(N_{\mathrm{AR}}-1) & \varphi_{xx}(N_{\mathrm{AR}}-2) & \dots & \varphi_{xx}(0) \end{pmatrix}}_{\mathbf{R}_{xx}} \cdot \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{N_{\mathrm{AR}}} \end{pmatrix}, \tag{A.26}$$

which have to be solved with respect to the AR coefficients a_j. Since $\mathbf{R}_{xx}$ is hermitian, Eq.A.26 can be solved efficiently, e.g., by using the Levinson-Durbin algorithm [Hay96].

B

Independence Assumption of Prediction Errors

In the derivation of the spectral Kalman filter approach in Sec. 3.2, the task in the update step becomes a 'classical' noise reduction problem: the 'noisy' (disturbed) DFT coefficients $D(\lambda, \mu)$ shall be decomposed into 'target' DFT coefficients $E^S_{\text{prop}}(\lambda, \mu)$ and 'noise' DFT coefficients $E^N_{\text{prop}}(\lambda, \mu)$. As the differential signal $D(\lambda, \mu)$ consists of the sum of $E^S_{\text{prop}}(\lambda, \mu)$ and $E^N_{\text{prop}}(\lambda, \mu)$, i.e., $D(\lambda, \mu) = E^S_{\text{prop}}(\lambda, \mu) + E^N_{\text{prop}}(\lambda, \mu)$, a conventional statistical estimator can be applied in the update step which is adapted to the statistics of the two prediction errors $E^S_{\text{prop}}(\lambda, \mu)$ and $E^N_{\text{prop}}(\lambda, \mu)$. For this purpose, MMSE estimators based on Gaussian or generalized Gamma priors are proposed in Sec. 3.2. In analogy to the original noise reduction approaches which are directly applied to $Y(\lambda, \mu) = S(\lambda, \mu) + N(\lambda, \mu)$, it is assumed here that target and noise signal, i.e., the two *prediction errors* of speech and noise, are *statistically independent*. However, only in the case of perfect prediction, the prediction errors equal the 'excitation' DFT coefficients $E_S(\lambda, \mu)$ and $E_N(\lambda, \mu)$ as defined in Eqs. 3.51 and 3.52, and the assumption is fulfilled. Otherwise the independence assumption introduces a small error which is analyzed in the following.

Considering the two prediction errors as complex-valued random variables, $E^S_{\text{prop}}(\lambda, \mu)$ and $E^N_{\text{prop}}(\lambda, \mu)$ are statistical independent only if the joint PDF $p(E^S_{\text{prop}}(\lambda, \mu), E^N_{\text{prop}}(\lambda, \mu))$ equals the product of the marginal PDFs $p(E^S_{\text{prop}}(\lambda, \mu))$ and $p(E^N_{\text{prop}}(\lambda, \mu))$ [Rou97], i.e.:

$$p(E^S_{\text{prop}}(\lambda, \mu), E^N_{\text{prop}}(\lambda, \mu)) = p(E^S_{\text{prop}}(\lambda, \mu)) \cdot p(E^N_{\text{prop}}(\lambda, \mu)). \tag{B.1}$$

As the PDF of a complex-valued random variable is fully characterized by the joint PDF of its real and imaginary parts [Say03], statistical independence between $E^S_{\text{prop}}(\lambda, \mu)$ and $E^N_{\text{prop}}(\lambda, \mu)$ can also be expressed by relating the statistics of the respective real and imaginary parts to each other. Therefore, the following conditions

have to be fulfilled all together for statistical independence:

$$p(\mathrm{Re}\{E^S_{\mathrm{prop}}(\lambda,\mu)\},\mathrm{Re}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) = p(\mathrm{Re}\{E^S_{\mathrm{prop}}(\lambda,\mu)\}) \cdot p(\mathrm{Re}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) \tag{B.2}$$

$$p(\mathrm{Im}\{E^S_{\mathrm{prop}}(\lambda,\mu)\},\mathrm{Im}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) = p(\mathrm{Im}\{E^S_{\mathrm{prop}}(\lambda,\mu)\}) \cdot p(\mathrm{Im}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) \tag{B.3}$$

$$p(\mathrm{Re}\{E^S_{\mathrm{prop}}(\lambda,\mu)\},\mathrm{Im}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) = p(\mathrm{Re}\{E^S_{\mathrm{prop}}(\lambda,\mu)\}) \cdot p(\mathrm{Im}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) \tag{B.4}$$

$$p(\mathrm{Im}\{E^S_{\mathrm{prop}}(\lambda,\mu)\},\mathrm{Re}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}) = p(\mathrm{Im}\{E^S_{\mathrm{prop}}(\lambda,\mu)\}) \cdot p(\mathrm{Re}\{E^N_{\mathrm{prop}}(\lambda,\mu)\}). \tag{B.5}$$

In order to prove the statistical independence of $E^S_{\mathrm{prop}}(\lambda,\mu)$ and $E^N_{\mathrm{prop}}(\lambda,\mu)$ based on Eqs. B.2 – B.5, the required joint PDFs as well as the marginal distributions are measured within the proposed Kalman filter system. Therefore, about 1.5 hours of speech is randomly selected from the NTT database [NC94] and disturbed by *White Gaussian Noise* (WGN) at different input *Signal-to-Noise-Ratio* (SNR) values varying between -10 dB and 35 dB with 5 dB step size. As proposed in Sec. 3.2.3.1, the Gaussian model is applied within the update step of the Kalman filter for this investigation.

Figure B.1 shows the measured histograms for the different constellations averaged over all input SNR values. In each row, the plot on the left side depicts the contours of the joint distribution and the plot on the right side the contours for the product of the respective marginal distributions. Overall, it can be seen that the particular histograms look very similar. The differences at the corners of the figures (the surfaces on the left side look a little bit more grainy than the plots on the right side) are due to the fact that the same amount of data is used to determine both distributions resulting in a lower graphical resolution for the joint distributions. However, marginal deviations can be observed comparing Figs. B.1(a) and B.1(b) as well as Figs. B.1(c) and B.1(d). Here, the declines of the PDFs slightly vary in the first and third quadrants leading to slight unbalances for the joint distributions which indicate possible statistical dependencies between the respective real and imaginary parts. In order to quantify these dependencies, further investigations are applied in the following.

In the derivation process of Sec. 3.2, not the full statistical independence between the two prediction errors is required. It is only assumed that $\mathbb{E}\{|D(\lambda,\mu)|^2\} = \mathbb{E}\{|E^S_{\mathrm{prop}}(\lambda,\mu)|^2\} + \mathbb{E}\{|E^N_{\mathrm{prop}}(\lambda,\mu)|^2\}$. Therefore, only this assumption is evaluated in the sequel. An exact calculation of $\mathbb{E}\{|D(\lambda,\mu)|^2\}$ for the proposed system results in:

$$\begin{aligned}\mathbb{E}\{|D(\lambda,\mu)|^2\} =& \mathbb{E}\{(E^S_{\mathrm{prop}}(\lambda,\mu) + E^N_{\mathrm{prop}}(\lambda,\mu))(E^S_{\mathrm{prop}}(\lambda,\mu) + E^N_{\mathrm{prop}}(\lambda,\mu))^*\} \\ =& \mathbb{E}\{|E^S_{\mathrm{prop}}(\lambda,\mu)|^2\} + \mathbb{E}\{|E^N_{\mathrm{prop}}(\lambda,\mu)|^2\} \\ &+ \mathbb{E}\{E^S_{\mathrm{prop}}(\lambda,\mu) \cdot (E^N_{\mathrm{prop}}(\lambda,\mu))^* + \mathbb{E}\{(E^S_{\mathrm{prop}}(\lambda,\mu))^* \cdot E^N_{\mathrm{prop}}(\lambda,\mu)\},\end{aligned} \tag{B.6}$$

where in general $\mathbb{E}\{E^S_{\mathrm{prop}}(\lambda,\mu) \cdot (E^N_{\mathrm{prop}}(\lambda,\mu))^*\} + \mathbb{E}\{(E^S_{\mathrm{prop}}(\lambda,\mu))^* \cdot E^N_{\mathrm{prop}}(\lambda,\mu)\} \neq 0$.

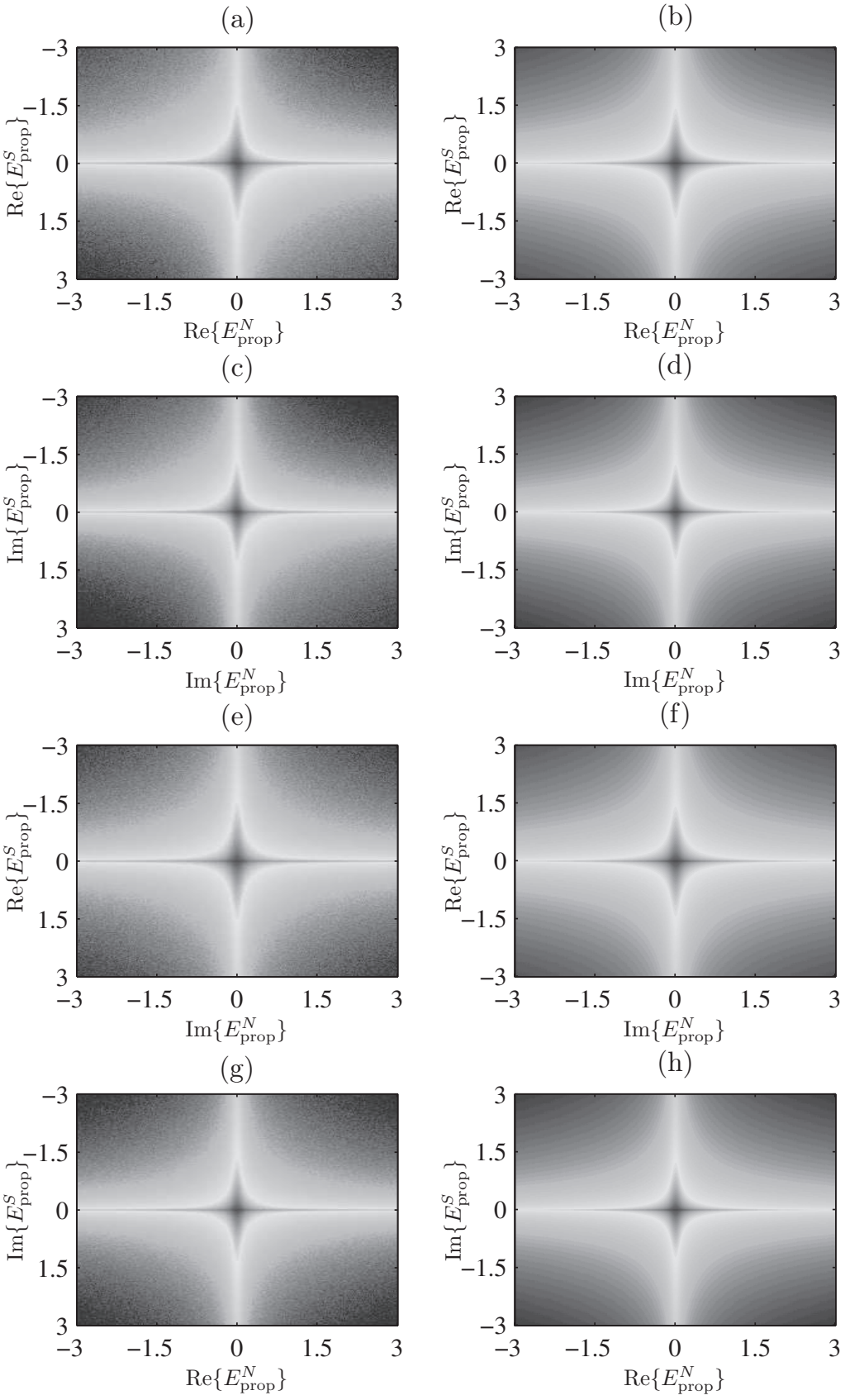

Figure B.1: Contour lines of measured distributions for real and imaginary parts of E^S_{prop} and E^N_{prop}: joint distributions in (a), (c), (e) and (g), products of marginal distributions in (b), (d), (f) and (h).

The expression $\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)\cdot(E^N_{\text{prop}}(\lambda,\mu))^*\}$ in Equation B.6 can be simplified to:

$$\begin{aligned}\mathbb{E}\{E^S_{\text{prop}}(\lambda,\mu)\cdot(E^N_{\text{prop}}(\lambda,\mu))^*\} = \ &... \\ &\mathbb{E}\left\{(S(\lambda,\mu)-S_{\text{prop}}(\lambda,\mu))\cdot(N(\lambda,\mu)-N_{\text{prop}}(\lambda,\mu))^*\right\} \\ =&\underbrace{\mathbb{E}\left\{S(\lambda,\mu)\cdot(N(\lambda,\mu))^*\right\}}_{=0}-\underbrace{\mathbb{E}\left\{S(\lambda,\mu)\cdot(N_{\text{prop}}(\lambda,\mu))^*\right\}}_{=0} \\ &-\underbrace{\mathbb{E}\left\{S_{\text{prop}}(\lambda,\mu)\cdot(N(\lambda,\mu))^*\right\}}_{=0}+\mathbb{E}\left\{S_{\text{prop}}(\lambda,\mu)\cdot(N_{\text{prop}}(\lambda,\mu))^*\right\} \\ =&\mathbb{E}\left\{S_{\text{prop}}(\lambda,\mu)\cdot(N_{\text{prop}}(\lambda,\mu))^*\right\}. \end{aligned} \tag{B.7}$$

In a similar way, the term $\mathbb{E}\{(E^S_{\text{prop}}(\lambda,\mu))^*\cdot E^N_{\text{prop}}(\lambda,\mu)\}$ results in $\mathbb{E}\left\{(S_{\text{prop}}(\lambda,\mu))^*\cdot N_{\text{prop}}(\lambda,\mu)\right\}$. Hence, $\mathbb{E}\{|D(\lambda,\mu)|^2\}$ is given by:

$$\begin{aligned}\mathbb{E}\{|D(\lambda,\mu)|^2\}=&\mathbb{E}\{|E^S_{\text{prop}}(\lambda,\mu)|^2\}+\mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\} \\ &+\mathbb{E}\left\{S_{\text{prop}}(\lambda,\mu)\cdot(N_{\text{prop}}(\lambda,\mu))^*\right\}+\mathbb{E}\left\{(S_{\text{prop}}(\lambda,\mu))^*\cdot N_{\text{prop}}(\lambda,\mu)\right\} \\ =&\mathbb{E}\{|E^S_{\text{prop}}(\lambda,\mu)|^2\}+\mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\} \\ &+2\cdot\mathbb{E}\{\text{Re}\{S_{\text{prop}}(\lambda,\mu)\cdot N_{\text{prop}}(\lambda,\mu)\}\}. \end{aligned} \tag{B.8}$$

As mentioned before, the expression $\mathbb{E}\{\text{Re}\{S_{\text{prop}}(\lambda,\mu)\cdot N_{\text{prop}}(\lambda,\mu)\}\}$ generally is non-zero as S_{prop} as well as N_{prop} are both estimates from the same noisy observation Y of previous frames. In order to analyze the introduced error when neglecting the last summand in Eq. B.8, the logarithmic error LogERR is investigated according to:

$$\begin{aligned}\text{LogERR}&=10\cdot\log_{10}\left(\frac{\mathbb{E}\{|D(\lambda,\mu)|^2\}}{\mathbb{E}\{|E^S_{\text{prop}}(\lambda,\mu)|^2\}+\mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\}}\right) \\ &=10\cdot\log_{10}\left(1+\frac{2\cdot\mathbb{E}\{\text{Re}\{S_{\text{prop}}(\lambda,\mu)\cdot N_{\text{prop}}(\lambda,\mu)\}\}}{\mathbb{E}\{|E^S_{\text{prop}}(\lambda,\mu)|^2\}+\mathbb{E}\{|E^N_{\text{prop}}(\lambda,\mu)|^2\}}\right), \end{aligned} \tag{B.9}$$

which can be approximated by:

$$\widehat{\text{LogERR}}\approx\frac{1}{M_{\text{F}}}\sum_{\mu=0}^{M_{\text{F}}-1}\left|10\cdot\log_{10}\left(1+\frac{2\cdot\sum\limits_{\lambda=0}^{N_{\text{F}}-1}\text{Re}\{S_{\text{prop}}(\lambda,\mu)\cdot N_{\text{prop}}(\lambda,\mu)\}}{\sum\limits_{\lambda=0}^{N_{\text{F}}-1}|E^S_{\text{prop}}(\lambda,\mu)|^2+\sum\limits_{\lambda=0}^{N_{\text{F}}-1}|E^N_{\text{prop}}(\lambda,\mu)|^2}\right)\right|, \tag{B.10}$$

where N_{F} and M_{F} represent the total number of evaluated frames and frequency bins, respectively. The unit of the logarithmic error is dB and an error near 0 dB is desirable. For the investigation, the Kalman filter system presented in Sec. 3.2.3.1 is again used relying on Gaussian models for both prediction errors. Moreover, five different speech signals from the NTT speech database [NC94] are each degraded by six different noise signals (f16, babble, car, factory1, factory2, white) taken from the NOISEX-92 database [VS93]. In total, $N_{\text{F}}=40000$ frames and $M_{\text{F}}=256$ frequency bins are used.

For this scenario, the averaged, SNR-dependent results of $\widehat{\text{LogERR}}$ are as follows:

SNR	-10 dB	-5 dB	0 dB	5 dB	10 dB
$\widehat{\text{LogERR}}$	0.0097 dB	0.0112 dB	0.0120 dB	0.0108 dB	0.0097 dB

SNR	15 dB	20 dB	25 dB	30 dB	35 dB
$\widehat{\text{LogERR}}$	0.0072 dB	0.0052 dB	0.0034 dB	0.0027 dB	0.0022 dB

It can be seen that the deviation is extremely small showing that the influence of the expression $2 \cdot \mathbb{E}\{\text{Re}\{S_{\text{prop}}(\lambda, \mu) \cdot N_{\text{prop}}(\lambda, \mu)\}\}$ within the derivation of the estimators in Sec. 3.2 is irrelevant and can be neglected.

C

Computational Complexity and Memory Requirements

In this section, the theoretical complexity as well as the memory requirements for the proposed Kalman filter approach of Sec. 3.2 are analyzed.

Theoretical Complexity

In the following, the additional number of operations which are necessary for the complex-valued Kalman filter approach on top of a conventional, statistical estimator is evaluated. Therefore, the computational complexity of the analysis-synthesis structure and the statistical weighting rule including noise *Power Spectral Density* (PSD) estimation and *Signal-to-Noise-Ratio* (SNR) estimation is not considered. Moreover, possible methods for reducing the computational complexity are proposed at the end.

Using a speech model of order N_K and a noise model of order M_K, the number of operations in terms of additions, multiplications and divisions solely for the Kalman filter (propagation and update step) yields:

# Additions	# Multiplications	# Divisions
$2(M_\mathrm{K}^2 + N_\mathrm{K}^2) + M_\mathrm{K} + N_\mathrm{K} + 3$	$2(M_\mathrm{K}^2 + N_\mathrm{K}^2) + 9(M_\mathrm{K} + N_\mathrm{K}) - 2$	1

The operations are a mixture of complex-valued and real-valued instructions. In order to apply the Kalman filter equations, the prediction coefficients of speech and noise have to be estimated in advance. Therefore, the Levinson-Durbin algorithm as well as the complex-valued autocorrelation function of length L_AC are necessary for the speech and the noise signal. Using a very efficient implementation of the Levinson-Durbin algorithm based on Toeplitz matrices [Mus84], the number of operations required for this purpose is given as follows:

# Additions	# Multiplications	# Divisions
$1.5 \cdot (N_\mathrm{K}^2 + M_\mathrm{K}^2) +$	$3.5 \cdot (N_\mathrm{K}^2 + M_\mathrm{K}^2) +$	
$(L_\mathrm{AC} + 0.5)(N_\mathrm{K} + M_\mathrm{K}) +$	$(L_\mathrm{AC} - 2.5)(N_\mathrm{K} + M_\mathrm{K}) +$	$N_\mathrm{K} + M_\mathrm{K} + 4$
$2 \cdot L_\mathrm{AC} - 4$	$2 \cdot L_\mathrm{AC}$	

In order to get a rough estimate of the additional complexity for the Kalman filter structure, the following weights are applied for the different instructions:

	Addition	Multiplication	Division
real-valued	1	1	16
complex-valued	2	4	not required

If possible, multiplication and addition are combined and counted as one operation assuming that a digital signal processor with a corresponding multiply-accumulate instruction is used for the implementation. Using the parameter settings as proposed in Sec. 3.3, i.e., sampling frequency $f_\mathrm{s} = 8\,\mathrm{kHz}$, *Fast Fourier Transform* (FFT) size $M_\mathrm{F} = 256$, frame length $L_\mathrm{F} = 160$, frame shift size $L_\mathrm{FS} = 40$, model orders $N_\mathrm{K} = 3$ and $M_\mathrm{K} = 2$ and autocorrelation function length $L_\mathrm{AC} = 6$, the proposed Kalman filter approach requires about 18.74 *Weighted Million Operations per Second* (WMOPS) in addition to a conventional, statistical estimator. The computational complexity of a state-of-the-art noise reduction technique working in the frequency domain lies in the order of 3 – 5 WMOPS. In [JMV+00], a noise suppression system for the *Adaptive Multi-Rate* (AMR) codec is proposed which is based on the same system parameters as above. The approach uses a simplified Minimum Statistics algorithm [Mar94] and applies the MMSE *Log Spectral Amplitude* (LSA) weighting rule [EM85]. In total, this algorithm requires 3.39 WMOPS. Hence, the *overall* complexity of the complex-valued Kalman filter approach exceeds that of a conventional system by a factor of 5 – 7. However, there are different possibilities to reduce the computational costs as stated below.

Complexity Reduction

In order to reduce the complexity, the following methods are proposed but not further covered in this thesis:

- *Frequency-dependent model orders for speech and noise*
  As proposed in [Pud02], the model orders for the speech and the noise signal can be chosen frequency-dependent. At least for speech signals, it is known that lower frequencies exhibit a higher temporal correlation compared to higher frequencies [Coh05b]. Thus, even lower model orders can be used for higher frequencies in order to save complexity.

- *Update rate of prediction coefficients*
  In the current realization of the algorithm, the prediction coefficients are calculated in each frame, i.e., every 5 ms. At the expense of a small performance loss, the prediction coefficients can be kept fixed for more than one frame, e.g., two or three frames to reduce the computational load.

- *Calculation of prediction coefficients*
  Instead of using the Levinson-Durbin algorithm in order to determine the prediction coefficients, less complex methods can be applied as, e.g., an adaptive prediction error filter based on a normalized *Least Mean Square* (LMS) algorithm [WC98].

Memory Requirements

This section addresses the evaluation of the memory requirements for the proposed Kalman filter system including the use of the different SNR-dependent MMSE estimators within the update step as proposed in Sec. 3.2.3.3. Similar to the previous section, only the extra memory which is required in addition to a conventional, statistical estimator is analyzed. A distinction is drawn between static and dynamic memory which is stored in the *Read-Only-Memory* (ROM) and *Random-Access-Memory* (RAM), respectively. It is assumed that each real-valued variable requires one ROM or RAM word and accordingly each complex-valued variable two ROM or RAM words. The requirements are listed in the following, separately for the propagation step and the update step.

Propagation Step

Within the propagation step, the previous L_{AC} enhanced speech and noise DFT coefficients have to be stored in order to determine the autocorrelation functions which are required for the estimation of the prediction coefficients. Moreover, the transition matrices $\mathbf{A}$ and $\mathbf{B}$ as well as the error covariance matrices $\mathbf{P}^{S}_{\mathrm{prop}}$ and $\mathbf{P}^{N}_{\mathrm{prop}}$ take up memory as follows:

Algorithm Components	Memory Requirements in	
	RAM Words	ROM Words
Enhanced speech and noise DFT coefficients	$M_{\mathrm{F}} \cdot L_{\mathrm{AC}}$	
Transition matrices $\mathbf{A}$ and $\mathbf{B}$	$M_{\mathrm{F}} \cdot (N_{\mathrm{K}} + M_{\mathrm{K}})$	
Matrices $\mathbf{P}^{S}_{\mathrm{prop}}$ and $\mathbf{P}^{N}_{\mathrm{prop}}$	$M_{\mathrm{F}} \cdot (N_{\mathrm{K}}^2 + M_{\mathrm{K}}^2)$ $-\frac{M_{\mathrm{F}}}{2} \cdot (N_{\mathrm{K}} + M_{\mathrm{K}})$	

Update Step

One look-up table has to be stored for each of the N_{MMSE} MMSE estimators which are used for the SNR-dependent estimation in the update step. Each look-up table consists of a two-dimensional matrix containing the weighting gains for all quantized a priori SNR and a posteriori SNR values. The total number of rows and columns of each matrix depends on the minimum and maximum SNR values $\mathrm{SNR}_{\mathrm{min}}$ and $\mathrm{SNR}_{\mathrm{max}}$ as well as the step size Δ_{SNR}. In addition to the look-up tables, the error covariance matrices $\mathbf{P}_{\mathrm{up}}^{S}$ and $\mathbf{P}_{\mathrm{up}}^{N}$ have to be stored. All memory requirements within the update step are listed in the following:

Algorithm Components	Memory Requirements in	
	RAM Words	ROM Words
Look-up tables		$N_{\mathrm{MMSE}} \cdot \left(\frac{\mathrm{SNR}_{\mathrm{max}} - \mathrm{SNR}_{\mathrm{min}}}{\Delta_{\mathrm{SNR}}}\right)^2$
Matrices $\mathbf{P}_{\mathrm{up}}^{S}$ and $\mathbf{P}_{\mathrm{up}}^{N}$	$M_{\mathrm{F}} \cdot (N_{\mathrm{K}}^2 + M_{\mathrm{K}}^2)$ $-\frac{M_{\mathrm{F}}}{2} \cdot (N_{\mathrm{K}} + M_{\mathrm{K}})$	

Using the same parameter settings as in the proposed system of Sec. 3.3, i.e., FFT size $M_{\mathrm{F}} = 256$, model orders $N_{\mathrm{K}} = 3$ and $M_{\mathrm{K}} = 2$, autocorrelation function length $L_{\mathrm{AC}} = 6$, $N_{\mathrm{MMSE}} = 11$ different MMSE estimators, $\mathrm{SNR}_{\mathrm{min}} = -40\,\mathrm{dB}$, $\mathrm{SNR}_{\mathrm{max}} = 50\,\mathrm{dB}$ as well as a step size $\Delta_{\mathrm{SNR}} = 1\,\mathrm{dB}$, the proposed Kalman filter approach requires about 8192 RAM words and 89100 ROM words more than a conventional, statistical estimator.

D

Instrumental Measurements

In order to evaluate the performance of different algorithms in the field of speech processing, the best way is to conduct a listening test. However, such tests are cumbersome and costly as a large number of probands is required. Moreover, each test person has to judge several audio samples which additionally takes a lot of time.

For the purpose of noise reduction, researchers have been busy to find instrumental measurements that predict the subjectively perceived speech quality of the processed signals in terms of speech distortions and noise attenuation. Although a listening test will probably never be replaced completely by instrumental measures, several instrumental quantities can be found in literature correlating well with the listener's impressions.

For the evaluation of the different algorithms in this thesis, the system and measures proposed in [Gus99, GMV96, Lot04] are applied. A brief overview about the framework as well as the instrumental measurements is given in the following.

Framework

A prerequisite for the measurements stated below is that not only the enhanced signal $\hat{s}(k)$ is available in the system but also purely the filtered speech signal $\tilde{s}(k)$ and the filtered noise signal $\tilde{n}(k)$. This allows to investigate the influences of the noise reduction algorithm on the speech and noise signals separately.

The framework used for this purpose is depicted in Fig. D.1. The weighting gains $G(\lambda, \mu)$ are determined in the frequency domain based on the noisy input signal $y(k)$ which is the sum of the clean speech signal $s(k)$ and noise signal $n(k)$. Afterwards, not only the noisy signal is filtered by $G(\lambda, \mu)$ but also the clean speech and noise *Discrete Fourier Transform* (DFT) coefficients $S(\lambda, \mu)$ and $N(\lambda, \mu)$ resulting in the filtered DFT coefficients $\tilde{S}(\lambda, \mu)$ and $\tilde{N}(\lambda, \mu)$. Finally, all three signals are transformed back into the time domain.

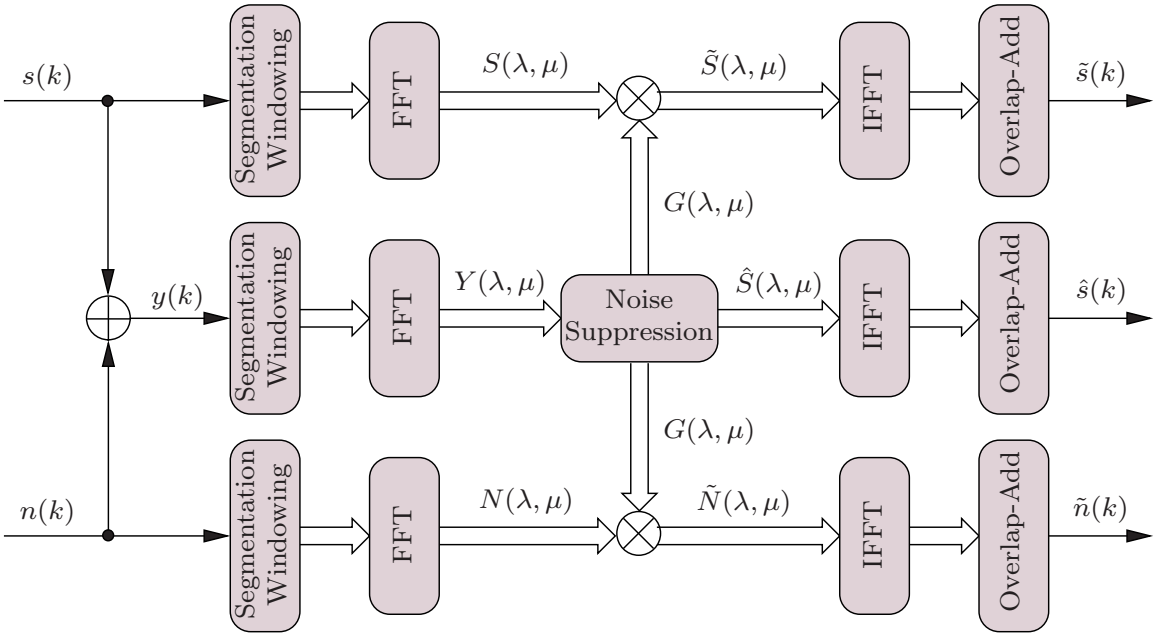

Figure D.1: System setup for the instrumental measurements.

Segmental Speech and Noise Attenuation

Segmental speech attenuation SegSA and segmental noise attenuation SegNA are defined as the segmental power ratios between original speech/noise signal and filtered speech/noise signal defined as follows:

$$\mathrm{SegSA} = \frac{1}{\mathcal{C}(K_s)} \sum_{l \in K_s} \left(10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} s^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \tilde{s}^2(k + l \cdot L_\mathrm{F})} \right) \right) \tag{D.1}$$

$$\mathrm{SegNA} = \frac{1}{\mathcal{C}(K_n)} \sum_{l \in K_n} \left(10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} n^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \tilde{n}^2(k + l \cdot L_\mathrm{F})} \right) \right) \tag{D.2}$$

where K_s represents the set of frames corresponding to speech activity, K_n the set of frames to be evaluated in total and $\mathcal{C}(\cdot)$ the number of elements in a set. Moreover, L_F denotes the length of one segment. The unit of SegNA and SegSA is dB.

Although the segmental speech attenuation does not state anything about how the speech is distorted, the difference between noise and speech attenuation SegNA-SegSA indicates the effective noise reduction and justifies the application of noise suppression for values greater than 0 dB.

Segmental Speech Signal-to-Noise Ratio

The segmental speech signal-to-noise ratio SegSSNR is defined as the geometric mean of the signal-to-noise ratios of short segments, while the squared difference between the clean speech signal $s(k)$ and the filtered speech signal $\tilde{s}$ is interpreted as noise. It is defined according to:

$$\mathrm{SegSSNR}_{s-\tilde{s}}^{s,(l)} = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} s^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \left(s(k + l \cdot L_\mathrm{F}) - \tilde{s}(k + l \cdot L_\mathrm{F})\right)^2} \right)$$

$$\mathrm{SegSSNR} = \frac{1}{\mathcal{C}(K_s)} \sum_{l \in K_s} \mathrm{SegSSNR}_{s-\tilde{s}}^{s,(l)}. \tag{D.3}$$

The SegSSNR measure is given in dB as well. It provides an indicator for the distortion of the speech signal and does not give any information about possible noise attenuation. The higher the result for SegSSNR the better the performance.

E

Deutschsprachige Kurzfassung

Die Benutzung des Mobiltelefons ist heutzutage aus dem alltäglichen Gebrauch der meisten Menschen nicht mehr wegzudenken. Mittlerweile ist der Informationsaustausch via Sprache zu jedem Zeitpunkt von fast jedem Ort der Welt aus möglich. Obwohl die Vision nach einer permanenten Erreichbarkeit und Konnektivität inzwischen fast weltweit realisiert worden ist, besteht weiterhin der Bedarf, die existierenden Kommunikationssysteme hinsichtlich Sprachqualität und Sprachverständlichkeit zu verbessern. Der Hörkomfort kann insbesondere bei der Sprachübertragung aus gestörten Umgebungen durch akustische Hintergrundstörungen, wie beispielsweise Verkehrslärm oder Bürogeräuschen erheblich beeinträchtigt werden.

In dieser Arbeit wird ein neuartiges, modellbasiertes Sprachverbesserungssystem zur einkanaligen Störgeräuschreduktion vorgestellt. Im Gegensatz zu konventionellen Verfahren steht bei den entwickelten Algorithmen die Ausnutzung *zeitlicher und spektraler Abhängigkeiten* von Sprach- und Störsignalen explizit im Fokus. Zur Berücksichtigung der zeitlichen Korrelation wird ein *modifiziertes Kalman-Filter* im Frequenzbereich abgeleitet. Wichtigste Neuerungen bilden hierbei die Verwendung einer komplexwertigen Prädiktion zur Schätzung der aktuellen DFT-Koeffizienten von Sprache und Störung sowie der Einsatz von SNR-abhängigen MMSE-Schätzregeln, welche an die gemessenen Statistiken des Eingangssignals angepasst sind. Um zusätzlich spektrale Abhängigkeiten von Sprachsignalen auszunutzen, zeigt diese Arbeit als neue Möglichkeit auf, Techniken der *künstlichen Bandbreitenerweiterung* für ein breitbandiges Störgeräuschreduktionssystem zu nutzen. Das vorgestellte Konzept verwendet dabei das bereits prozessierte und verbesserte Signal von tiefen Frequenzen erneut, um die Ergebnisse einer konventionellen Störreduktion bei höheren Frequenzen zu verbessern. Darüber hinaus beschäftigt sich diese Arbeit mit wirksamen Gegenmaßnahmen zur Reduzierung von sogenannten *Musical Tones* und bietet eine neuartige Lösung zur Unterdrückung von *zeitlich stark veränderlichen, harmonischen Störungen.*

Alle entwickelten Verfahren zur Sprachverbesserung wurden in der vorliegenden Arbeit anhand von instrumentellen Messungen und subjektiven Höreindrücken ausgiebig bewertet und evaluiert. Im Vergleich zu konventionellen Verfahren der Störgeräuschreduktion stellte sich dabei heraus, dass die vorgestellten Algorithmen in Bezug auf Störd\"ampfung und Sprachverzerrungen deutlich bessere Ergebnisse

erzielten. Das neue modellbasierte System ist dabei nicht auf die Anwendung in Mobiltelefonen beschränkt. Es kann zusätzlich verwendet werden, um die Sprachqualität von Freisprecheinrichtungen, Konferenzsystemen oder digitalen Hörgeräten zu verbessern.

Bibliography

[3GP01] 3GPP TS 26.171. "AMR Wideband Speech Codec; General Description", March 2001.

[3GP04] 3GPP TS 26.290. "Extended AMR Wideband Codec; Transcoding Functions", September 2004.

[ABBN04] S. Artstein, K. Ball, F. Barthe, and A. Naor. "Solution of Shannon's Problem on the Monotonicity of Entropy". *Journal of the American Mathematical Society*, vol. 17, pp. 975–982, May 2004.

[AS70] B. S. Atal and M. R. Schroeder. "Predictive Coding of Speech Signals". *Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, October 1970.

[Bay63] T. Bayes. "An Essay Towards Solving a Problem in the Doctrine of Chances". *Philosophical Transactions*, vol. 53, pp. 370–418, January 1763.

[BCHC09] J. Benesty, J. Chen, Y. Huang, and I. Cohen. *Noise Reduction in Speech Processing*. Springer-Verlag, Berlin, Heidelberg, New York, 2009.

[BCR01] F. Beritelli, S. Casale, and G. Ruggeri. "Performance Evaluation and Comparison of ITU-T/ETSI Voice Activity Detectors". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001.

[Ben07] J. Benesty. *Springer Handbook of Speech Processing*. Springer-Verlag, Berlin, Heidelberg, New York, 2007.

[Ber98] H.-P. Bernhard. "A Tight Upper Bound on the Gain of Linear and Nonlinear Predictors for Stationary Stochastic Processes". *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2909–2917, November 1998.

[BGM07] C. Breithaupt, T. Gerkmann, and R. Martin. "Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement without Musical Noise". *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

[BKM08] C. Breithaupt, M. Krawczyk, and R. Martin. "Parameterized MMSE Spectral Magnitude Estimation for the Enhancement of Noisy Speech". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, April 2008.

[Bol79] S. F. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[BSV06] C. Beaugeant, M. Schönle, and I. Varga. "Challenges of 16 kHz in Acoustic Pre- and Post-Processing for Terminals". *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, May 2006.

[Cap94] O. Cappe. "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor". *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–37, April 1994.

[CB02] I. Cohen and B. Berdugo. "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement". *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12 –15, January 2002.

[CD08] B. Chabane and B. Daoued. "On the Use of Kalman Filter for Enhancing Speech Corrupted by Colored Noise". *WSEAS Transactions on Signal Processing*, vol. 4, no. 12, pp. 657–666, December 2008.

[Coh03] I. Cohen. "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging". *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[Coh04] I. Cohen. "Modeling Speech Signals in the Time-Frequency Domain Using GARCH". *Signal Processing*, vol. 84, no. 12, pp. 2453–2459, December 2004.

[Coh05a] I. Cohen. "Speech Spectral Modeling and Enhancement Based on Autoregressive Conditional Heteroscedasticity Models". *Signal Processing*, vol. 86, no. 4, pp. 698–709, June 2005.

[Coh05b] I. Cohen. "Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation". *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, September 2005.

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Chichester, UK, 2006.

[CVA06] T. Clevorn, P. Vary, and M. Adrat. "Parameter SNR Optimized Index Assignments and Quantizers Based on First Order A Priori Knowledge for Iterative Source-Channel Decoding". *Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, March 2006.

[Dav02] G. M. Davis. *Noise Reduction in Speech Applications*. CRC Press Inc., Boca Raton, FL, USA, 2002.

[DBC91] M. Dendrinos, S. Bakamidis, and G. Carayannis. "Speech Enhancement from Noise: A Regenerative Approach". *Speech Communication*, vol. 10, pp. 45–67, February 1991.

[Dur95] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, USA, 1995.

[EHGV10] T. Esch, F. Heese, B. Geiser, and P. Vary. "Wideband Noise Suppression Supported by Artificial Bandwidth Extension Techniques". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, March 2010.

[EHH08] J. Erkelens, R. C. Hendriks, and R. Heusdens. "On the Estimation of Complex Speech DFT Coefficients without Assuming Independent Real and Imaginary Parts". *IEEE Signal Processing Letters*, vol. 15, pp. 213–216, January 2008.

[EHHJ07] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients with Generalized Gamma Priors". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, August 2007.

[EM84] Y. Ephraim and D. Malah. "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[EM85] Y. Ephraim and D. Malah. "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[EMJ89] Y. Ephraim, D. Malah, and B.-H. Juang. "On the Application of Hidden Markov Models for Enhancing Noisy Speech". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1846 –1856, December 1989.

[Eph92] Y. Ephraim. "Statistical-Model-Based Speech Enhancement Systems". *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526 –1555, October 1992.

[ERHV10a] T. Esch, M. Rüngeler, F. Heese, and P. Vary. "A Modified Minimum Statistics Algorithm for Reducing Time Varying Harmonic Noise". *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010.

[ERHV10b] T. Esch, M. Rüngeler, F. Heese, and P. Vary. "Combined Reduction of Time Varying Harmonic and Stationary Noise Using Frequency Warping". *Conference Record of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, Pacific Grove, CA, USA, November 2010.

[Esc06] T. Esch. "Wideband Coding of Speech and Audio Signals using Bandwidth Extension Techniques". *Proceedings of International Student Conference on Electrical Engineering (POSTER)*, Prague, Czech Republic, May 2006.

[EV08a] T. Esch and P. Vary. "Speech Enhancement Using a Modified Kalman Filter Based on Complex Linear Prediction and Supergaussian Priors". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, March 2008.

[EV08b] T. Esch and P. Vary. "Modified Kalman Filter Exploiting Interframe Correlation of Speech and Noise Magnitudes". *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, September 2008.

[EV08c] T. Esch and P. Vary. "Exploiting Temporal Correlation of Speech and Noise Magnitudes Using a Modified Kalman Filter for Speech Enhancement". *ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, October 2008.

[EV09] T. Esch and P. Vary. "Efficient Musical Noise Suppression for Speech Enhancement Systems". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipeh, Taiwan, May 2009.

[EV11] T. Esch and P. Vary. "Model-Based Speech Enhancement Using SNR Dependent MMSE Estimation". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[EVT95] Y. Ephraim and H. L. Van Trees. "A Signal Subspace Approach for Speech Enhancement". *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251 –266, July 1995.

[FBS05] T. Fingscheidt, C. Beaugeant, and S. Suhadi. "Overcoming the Statistical Independence Assumption w.r.t. Frequency in Speech Enhancement". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, March 2005.

[FV01] T. Fingscheidt and P. Vary. "Softbit Speech Decoding: A New Approach to Error Concealment". *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 240 – 251, March 2001.

[Gab05] M. Gabrea. "An Adaptive Kalman Filter for the Enhancement of Speech Signals in Colored Noise". *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2005.

[GBW98] S. Gannot, D. Burshtein, and E. Weinstein. "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms". *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[GEH98] T. Gülzow, A. Engelsberg, and U. Heute. "Comparison of a Discrete Wavelet Transformation and a Nonuniform Polyphase Filterbank Applied to Spectral-Subtraction Speech Enhancement". *Signal Processing*, vol. 64, no. 1, pp. 5 – 19, January 1998.

[GJV05] B. Geiser, P. Jax, and P. Vary. "Artificial Bandwidth Extension of Speech Supported by Watermark-Transmitted Side Information". *Proceedings of European Conference on Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, September 2005.

[GJV+07] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaumé, and S. Ragot. "Bandwidth Extension for Hierarchical Speech and Audio Coding in ITU-T Rec. G.729.1". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, November 2007.

[GKG91] J. Gibson, B. Koo, and S. Gray. "Filtering of Colored Noise for Speech Enhancement and Coding". *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, August 1991.

[GL84] D. Griffin and J. Lim. "Signal Estimation from Modified Short-Time Fourier Transform". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236 – 243, April 1984.

[GLH03] T. Gülzow, T. Ludwig, and U. Heute. "Spectral-Subtraction Speech Enhancement in Multirate Systems with and without Non-Uniform and Adaptive Bandwidths". *Signal Processing*, vol. 83, no. 8, pp. 1613–1631, August 2003.

[GM10a] T. Gerkmann and R. Martin. "Empirical Distributions of DFT-Domain Speech Coefficients Based on Estimated Speech Variances". *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, August 2010.

[GM10b] T. Gerkmann and R. Martin. "Cepstral Smoothing with Reduced Computational Complexity". *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010.

[GMJV02] S. Gustafsson, R. Martin, P. Jax, and P. Vary. "A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction". *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, July 2002.

[GMV96] S. Gustafsson, R. Martin, and P. Vary. "On the Optimization of Speech Enhancement Systems Using Instrumental Measures". *Proceedings of Workshop on Quality Assessment in Speech, Audio and Image Communication*, Darmstadt, Germany, March 1996.

[GNC99] H. Gustafsson, S. Nordholm, and I. Claesson. "Spectral Subtraction with Adaptive Averaging of the Gain Function". *Proceedings of European Conference on Speech Communication and Technology (INTERSPEECH)*, Budapest, Hungary, September 1999.

[GRJZ00] I. S. Gradshteyn, I. M. Ryzhik, A. Jeffrey, and D. Zwillinger. *Table of Integrals, Series, and Products*. Academic Press, 2000.

[GTT98] Z. Goh, K.-C. Tan, and B. T. G. Tan. "Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction". *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, May 1998.

[Gus99] S. Gustafsson. *Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction*. Phd thesis, RWTH Aachen University, Aachen, Germany, 1999.

[GV07] B. Geiser and P. Vary. "Backwards Compatible Wideband Telephony in Mobile Networks: CELP Watermarking and Bandwidth Extension". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. IV, Honolulu, HI, USA, April 2007.

[GV08] B. Geiser and P. Vary. "High Rate Data Hiding in ACELP Speech Codecs". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, March 2008.

[Hab05] E. Habets. "Multi-Channel Speech Dereverberation Based on a Statistical Model of Late Reverberation". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, 2005.

[Hay96] M. H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Chichester, UK, 1996.

[HEGV10] F. Heese, T. Esch, B. Geiser, and P. Vary. "Noise Reduction for Wideband Speech Exploiting Spectral Dependencies Based on Conditional Estimation". *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010.

[HEH08] R. C. Hendriks, J. Erkelens, and R. Heusdens. "Comparison of Complex-DFT Estimators with and without the Independence Assumption of Real and Imaginary Parts". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008.

[HEV11] F. Heese, T. Esch, and P. Vary. "Dual Channel Reduction of Rapidly Varying Harmonic and Random Noise Using a Spot Microphone". *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Aachen, Germany, September 2011.

[HGV98] S. Heinen, A. Geiler, and P. Vary. "MAP Channel Decoding by Exploiting Multilevel Source A Priori Knowledge". *ITG-Fachtagung Codierung für Quelle, Kanal und Übertragung*, Aachen, Germany, March 1998.

[HHJ10] R. C. Hendriks, R. Heusdens, and J. Jensen. "MMSE Based Noise PSD Tracking with Low Complexity". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, March 2010.

[HJH08] R. C. Hendriks, J. Jensen, and R. Heusdens. "Noise Tracking Using DFT Domain Subspace Decompositions". *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 541–553, March 2008.

[HL03] Y. Hu and P. C. Loizou. "A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise". *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334 – 341, July 2003.

[HS06] E. Hänsler and G. Schmidt, editors. *Topics in Acoustic Echo and Noise Control, Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise and Speech Processing.* Springer-Verlag, Berlin, Heidelberg, New York, 2006.

[HS08] E. Hänsler and G. Schmidt, editors. *Speech and Audio Processing in Adverse Environments.* Springer-Verlag, Berlin, Heidelberg, New York, 2008.

[ITU88] ITU-T Recommendation G.722. "7 kHz Audio Coding within 64 kBit/s", November 1988.

[ITU96] ITU-T Recommendation P.800. "Methods for Subjective Determination of Transmission Quality - Series P: Telephone Transmission Quality; Methods for Objective and Subjective Assessment of Quality", August 1996.

[ITU99] ITU-T Recommendation G.722.1. "Coding at 24 and 32 kbit/s for Hands-Free Operation in Systems with Low Frame Loss", September 1999.

[ITU06a] ITU-T Recommendation G.729.1. "An 8-32 kbit/s Scalable Wideband Coder Bitstream Interoperable with G.729", May 2006.

[ITU06b] ITU-T Recommendation G.729.1-SWB. "G.729-Based Embedded Variable Bit-Rate Coder: An 8-32 kbit/s Scalable Wideband Coder Bitstream Interoperable with G.729: New Annex E on Superwideband Scalable Extension for G.729.1", May 2006.

[Iza65] T. Izawa. "Two or Multi-Dimensional Gamma-Type Distribution and its Application to Rainfall Data". *Papers in Meteorology and Geophysics*, vol. 15, pp. 167–200, February 1965.

[Jax02] P. Jax. *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds.* Phd thesis, RWTH Aachen University, Aachen, Germany, 2002.

[JJYW02] C.-Z. Jin, L.-J. Jia, Z.-J. Yang, and K. Wada. "On Convergence of a BCLS Algorithm for Noisy Autoregressive Process Estimation". *Proceedings of IEEE Conference on Decision and Control*, Las Vegas, NV, USA, December 2002.

[JKYW03] L.-J. Jia, S. Kanae, Z.-J. Yang, and K. Wada. "On Bias Compensation Estimation for Noisy AR Process". *Proceedings of IEEE Conference on Decision and Control*, Maui, HI, USA, December 2003.

[JMV+00] P. Jax, R. Martin, P. Vary, M. Adrat, I. Varga, W. Frank, and M. Ihle. "A Noise Suppression System for the AMR Speech Codec". *ITG-Fachtagung Konvens 2000 / Sprachkommunikation*, Ilmenau, Germany, October 2000.

[JN84] N. S. Jayant and P. Noll. *Digital Coding of Waveforms.* Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

[JSEV10] M. Jeub, M. Schäfer, T. Esch, and P. Vary. "Model-Based Dereverberation Preserving Binaural Cues". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18 - Special Issue on Processing Reverberant Speech, no. 7, pp. 1732 – 1745, September 2010.

[JV02] P. Jax and P. Vary. "An Upper Bound on the Quality of Artificial Bandwidth Extension of Narrowband Speech Signals". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, May 2002.

[JV03a] P. Jax and P. Vary. "Artificial Bandwidth Extension of Speech Signals Using MMSE Estimation Based on a Hidden Markov Model". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China, April 2003.

[JV03b] P. Jax and P. Vary. "On Artificial Bandwidth Extension of Telephone Speech". *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, August 2003.

[JV04] P. Jax and P. Vary. "Feature Selection for Improved Bandwidth Extension of Speech Signals". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, May 2004.

[JV06] P. Jax and P. Vary. "Bandwidth Extension of Speech Signals: A Catalyst for the Introduction of Wideband Speech Coding?". *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106–111, May 2006.

[Kal60] R. Kalman. "A New Approach to Linear Filtering and Prediction Problems". *Transactions of the ASME – Journal of Basic Engineering*, vol. 82, pp. 35–45, March 1960.

[Kap05] A. Kaps. "Acoustic Noise Reduction Using a Multiple-Input Single-Output Kalman Filter". *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, Netherlands, 2005.

[Kay88] S. M. Kay. *Modern Spectral Estimation-Theory and Application*. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[KBJ00] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions; Models and Applications*. John Wiley & Sons, Chichester, UK, 2000.

[KC09] J.-M. Kum and J.-H. Chang. "Speech Enhancement Based on Minima Controlled Recursive Averaging Incorporating Second-Order Conditional MAP Criterion". *IEEE Signal Processing Letters*, vol. 16, no. 7, pp. 624 –627, July 2009.

[KK01] M. Kuropatwinski and W. B. Kleijn. "Estimation of the Excitation Variances of Speech and Noise AR-Models for Enhanced Speech Coding". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, May 2001.

[KK06] M. Kuropatwinski and W. B. Kleijn. "Estimation of the Short-Term Predictor Parameters of Speech Under Noisy Conditions". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1645 –1655, September 2006.

[KL51] S. Kullback and R. A. Leibler. "On Information and Sufficiency". *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, March 1951.

[KL87] L. F. Kozachenko and N. N. Leonenko. "A Statistical Estimate for the Entropy of a Random Vector". *Problems of Information Transmission*, vol. 23, no. 2, pp. 95–101, 1987.

[KP95] W. B. Kleijn and K. K. Paliwal. *Speech Coding and Synthesis* . Elsevier Science Inc., New York, NY, USA, 1995.

[KSE+09] H. Krüger, T. Schumacher, T. Esch, B. Geiser, and P. Vary. "RTPROC: Rapid Real-Time Prototyping for Audio Signal Processing". *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Dresden, Germany, September 2009.

[LBG80] Y. Linde, A. Buzo, and R. Gray. "An Algorithm for Vector Quantizer Design". *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[LGO+96] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells Jr. "Noise Reduction Using an Undecimated Discrete Wavelet Transform". *IEEE Signal Processing Letters*, vol. 3, no. 1, pp. 10 –12, January 1996.

[Lim83] J. S. Lim. *Speech Enhancement*. Prentice Hall, Upper Saddle River, New Jersey, 1983.

[LMS96] K. Y. Lee, S. McLaughlin, and K. Shirai. "Speech Enhancement Based on Extended Kalman Filter and Neural Predictive Hidden Markov Model". *Proceedings of IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, Piscataway, NJ, USA, September 1996.

[LO79] J. S. Lim and A. V. Oppenheim. "Enhancement and Bandwidth Compression of Noisy Speech". *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

[Loi07] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press Inc., Boca Raton, FL, USA, 2007.

[Lot04] T. Lotter. *Single and Multimicrophone Speech Enhancement for Hearing Aids*. Phd thesis, RWTH Aachen University, Aachen, Germany, 2004.

[LV05] T. Lotter and P. Vary. "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model". *EURASIP Journal on Applied Signal Processing*, pp. 1110–1126, January 2005.

[Mar94] R. Martin. "Spectral Subtraction Based on Minimum Statistics". *European Signal Processing Conference (EUSIPCO)*, Edinburgh, UK, September 1994.

[Mar01] R. Martin. "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics". *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 501–512, July 2001.

[Mar05] R. Martin. "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors". *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845 – 856, September 2005.

[MB03] R. Martin and C. Breithaupt. "Speech Enhancement in the DFT Domain Using Laplacian Speech Priors". *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, September 2003.

[Mei02] E. Meijering. "A Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing". *Proceedings of the IEEE*, vol. 90, no. 3, pp. 319 –342, March 2002.

[Mey00] H. Meyr. *Regelungstechnik und Systemtheorie*. Verlag Mainz in Aachen, Aachen, Germany, 2000.

[MGj76] J. D. Markel and A. H. Gray jr. *Linear Prediction of Speech*. Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[MM80] R. J. McAulay and M. L. Malpass. "Speech Enhancement Using a Soft-Decision Noise Suppression Filter". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, April 1980.

[Mus84] B. R. Musicus. "Levinson and Fast Choleski Algorithms for Toeplitz and Almost Toeplitz Matrices". *Technical report, Research Lab. of Electronics, M.I.T.*, 1984.

[NC94] NTT-Corporation. "Multi-Lingual Speech Database for Telephonometry", 1994.

[NEH00] F. Norden, T. Eriksson, and P. Hedelin. "An Information Theoretic Perspective on the Speech Spectrum Process". *Proceedings of IEEE Workshop on Speech Coding*, Delavan, WI, USA, September 2000.

[NGAK02] M. Nilsson, H. Gustafsson, S. V. Andersen, and B. W. Kleijn. "Gaussian Mixture Model Based Mutual Information Estimation Between Frequency Bands in Speech". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, April 2002.

[OS98] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1998.

[Par86] T. W. Parsons. *Voice and Speech Processing*. McGraw-Hill, New York, 1986.

[PB87] K. Paliwal and A. Basu. "A Speech Enhancement Method Based on Kalman Filtering". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, April 1987.

[PC08] Y.-S. Park and J.-H. Chang. "A Probabilistic Combination Method of Minimum Statistics and Soft Decision for Robust Noise Power Estimation in Speech Enhancement". *IEEE Signal Processing Letters*, vol. 15, pp. 95–98, January 2008.

[Plo09] E. Plourde. *Bayesian Short-Time Spectral Amplitude Estimators for Single Channel Speech Enhancement*. Phd thesis, McGill University, Montreal, Quebec, Canada, 2009.

[PLW10] K. Paliwal, J. Lyons, and K. Wójcicki. "Preference for 20–40 ms Window Duration in Speech Analysis". *Proceedings of IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, Gold Coast, QLD, Australia, December 2010.

[PM96] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications*. Prentice Hall, Upper Saddle River, New Jersey, 1996.

[Pud02] H. Puder. "Kalman-Filters in Subbands for Noise Reduction with Enhanced Pitch-Adaptive Speech Model Estimation". *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 139–148, March 2002.

[Qua01] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 2001.

[RJ93] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[Rou97] G. G. Roussas. *A Course in Mathematical Statistics*. Academic Press, 1997.

[RR95] D. Reynolds and R. Rose. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72 –83, January 1995.

[RS78] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[SA79] H. Sakai and M. Arase. "Recursive Parameter Estimation of an Autoregressive Process Disturbed by White Noise". *International Journal of Control*, vol. 30, no. 6, pp. 949–966, 1979.

[SAD05] M. L. Seltzer, A. Acero, and J. Droppo. "Robust Bandwidth Extension of Noise-corrupted Narrowband Speech". *Proceedings of European Conference on Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, September 2005.

[Say03] A. H. Sayed. *Fundamentals of Adaptive Filtering*. Wiley-IEEE Press, 2003.

[SB97] J. Seok and K. Bae. "Speech Enhancement with Reduction of Noise Components in the Wavelet Domain". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, April 1997.

[SC08] J. L. Speyer and W. H. Chung. *Stochastic Processes, Estimation, and Control*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[Sch65] M. R. Schroeder. "Apparatus for Suppressing Noise and Distortion in Communication Signals". *U.S. Patent No. 3180936*, April 1965.

[SKS99] J. Sohn, N. S. Kim, and W. Sung. "A Statistical Model-Based Voice Activity Detection". *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, January 1999.

[SN09] D. Sharma and P. A. Naylor. "Evaluation of Pitch Estimation in Noisy Speech for Application in Non-Intrusive Speech Quality Assessment". *European Signal Processing Conference (EUSIPCO)*, Glasgow, UK, August 2009.

[SV10] L. Schmalen and P. Vary. "Reconstruction of Multiple Descriptions by MMSE Estimation". *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010.

[TJ09] M. Triki and K. Janse. "Minimum Subspace Noise Tracking for Noise Power Spectral Density Estimation". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipeh, Taiwan, May 2009.

[Ton77] H. Tong. "More on Autoregressive Model Fitting with Noisy Data by Akaike's Information Criterion". *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 409 – 410, May 1977.

[TPM93] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos. "Speech Enhancement Using Psychoacoustic Criteria". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Minneapolis, MN, USA, April 1993.

[TTM+11] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin. "An Evaluation of Noise Power Spectral Density Estimation Algorithms in Adverse Acoustic Environments". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[Var85] P. Vary. "Noise Suppression by Spectral Magnitude Estimation-Mechanism and Theoretical Limits". *Signal Processing*, vol. 8, no. 4, pp. 387–400, July 1985.

[Var08] P. Vary. "Speech Enhancement by Conditional Estimation - Noise Reduction, Error Concealment & Bandwidth Extension, What Makes the Difference?". *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, September 2008.

[Vas96] S. V. Vaseghi. *Advanced Signal Processing and Digital Noise Reduction*. John Wiley & Sons, Chichester, UK, 1996.

[vC89] D. van Compernolle. "Noise Adaptation in a Hidden Markov Model Speech Recognition System". *Computer Speech and Language*, vol. 3, pp. 151–167, 1989.

[VG07] P. Vary and B. Geiser. "Steganographic Wideband Telephony Using Narrowband Speech Codecs". *Conference Record of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, Pacific Grove, CA, USA, November 2007.

[VHH98] P. Vary, U. Heute, and W. Hess. *Digitale Sprachsignalverarbeitung*. Teubner Verlag, Stuttgart, Germany, 1998.

[VM06] P. Vary and R. Martin. *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, Chichester, UK, 2006.

[VS93] A. Varga and H. J. M. Steeneken. "Assessment for Automatic Speech Recognition II: NOISEX-92: a Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems". *Speech Communication*, vol. 12, pp. 247–251, July 1993.

[WASA07] Y. Wang, J. An, V. Sethu, and E. Ambikairajah. "Perceptually Motivated Pre-Filter for Speech Enhancement Using Kalman Filtering". *Proceedings of IEEE International Conference on Information, Communications and Signal Processing*, Singapore, Singapore, December 2007.

[WC98] W.-R. Wu and P.-C. Chen. "Subband Kalman Filtering for Speech Enhancement". *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 1072–1083, August 1998.

[WL82] D. L. Wang and J. S. Lim. "The Unimportance of Phase in Speech Enhancement". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, August 1982.

[ZF90] E. Zwicker and H. Fastl. *Psychoacoustics*. Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[ZF99] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin, Heidelberg, New York, 1999.

[Zhe99] W. X. Zheng. "A Least-Squares Based Method for Autoregressive Signals in the Presence of Noise". *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 1, pp. 81 –85, January 1999.

[ZVY06a] E. Zavarehei, S. Vaseghi, and Q. Yan. "Temporal Modelling and Kalman Filtering of DFT Trajectories for Enhancement of Noisy Speech". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[ZVY06b] E. Zavarehei, S. Vaseghi, and Q. Yan. "Inter-Frame Modeling of DFT Trajectories of Speech and Noise for Speech Enhancement Using Kalman Filters". *Speech Communication*, vol. 48, no. 11, pp. 1545 – 1555, November 2006.