

# Contributions to Statistical Modeling for Minimum Mean Square Error Estimation in Speech Enhancement

Von der Fakultät für Elektrotechnik, Informationstechnik, Physik  
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung der Würde  
eines Doktor-Ingenieurs (Dr.-Ing.)  
genehmigte

## Dissertation

von

**Balázs Fodor**

aus Esztergom, Ungarn

eingereicht am: 17. Juni 2014  
mündliche Prüfung am: 24. Oktober 2014

Erstgutachter: Prof. Dr.-Ing. Tim Fingscheidt  
Technische Universität Carolo-Wilhelmina zu Braunschweig  
Zweitgutachter: Prof. Dr.-Ing. Gerhard Schmidt  
Christian-Albrechts-Universität zu Kiel  
Prüfungsvorsitzender: Prof. Dr.-Ing. Thomas Kürner  
Technische Universität Carolo-Wilhelmina zu Braunschweig



Mitteilungen aus dem Institut für Nachrichtentechnik der  
Technischen Universität Braunschweig

Band 39

**Balázs Fodor**

**Contributions to Statistical Modeling for  
Minimum Mean Square Error Estimation  
in Speech Enhancement**

Shaker Verlag  
Aachen 2015

**Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Braunschweig, Techn. Univ., Diss., 2014

Editor of this volume:

Prof. Dr.-Ing. Tim Fingscheidt  
Institute for Communications Technology  
Technische Universität Braunschweig  
Schleinitzstrasse 22  
38106 Braunschweig  
Germany  
e-mail: [fingscheidt@ifn.ing.tu-bs.de](mailto:fingscheidt@ifn.ing.tu-bs.de)  
phone: +49-531-391-2485  
fax: +49-531-391-8218

Copyright Shaker Verlag 2015

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-3571-1

ISSN 1865-2484

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen

Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9

Internet: [www.shaker.de](http://www.shaker.de) • e-mail: [info@shaker.de](mailto:info@shaker.de)

# Preface

This dissertation was written during my research activity at the Institute for Communications Technology (in German: Institut für Nachrichtentechnik, IfN) of Technische Universität Carolo-Wilhelmina zu Braunschweig. In my opinion, a doctoral thesis is inconceivable without precious discussions with colleagues and other researchers for instance on conferences as well as without a stable familial background. It is my pleasure to thank all the people who contributed to the success of this work.

I would like to express my gratitude first to my adviser Prof. Dr.-Ing. Tim Fingscheidt. Thank you for your continuous support and competent supervision. Your suggestions, valuable feedback, and our productive discussions during my time at IfN have left a significant imprint on this dissertation.

Next, I would like to thank Prof. Dr.-Ing. Gerhard Schmidt for being a member of the examination board as well as for his interest in my research work and Prof. Dr.-Ing. Thomas Kürner for being the chair of the examination board.

As part of a fruitful research cooperation on the topic of Chapter 5, I worked together with Prof. Dr.-Ing. Timo Gerkmann. Timo, thank you for all your support and competent feedback. I always enjoyed our friendly and rewarding discussions.

I would like to thank my colleagues at IfN for their open-mindedness, helpfulness, and for creating a friendly work atmosphere. Particularly, I am grateful to Dipl.-Ing. Patrick Bauer, Marc-André Jung M.Sc., Dr.-Ing. Florian Pflug, Dipl.-Ing. Simon Receveur, Dipl.-Ing. David Scheler, Dr.-Ing. Suhadi, and Dr.-Ing. Huajun Yu for having a lot of valuable discussions which this dissertation has definitely benefited from.

Furthermore, I am thankful to Dipl.-Ing. Patrick Bauer, Samy Elshamy M.Sc., Marc-André Jung M.Sc., Dr.-Ing. Florian Pflug, Dr.-Ing. Suhadi, Peter Transfeld M.Sc., and Dr.-Ing. Huajun Yu for proofreading chapters of this thesis and providing constructive feedback regarding this dissertation.

I would like to thank Prof. István Kollár for being my adviser at Budapest University of Technology and Economics prior to my time at IfN. István, I liked your lectures and the research work with you very much. You definitely contributed the most to the decision that I wanted to keep on doing research and work towards a doctoral degree.

I am grateful to my parents for all their support and for laying the groundwork for my achievements such as this thesis. I would like to extend my deepest gratitude to my wife Kristina for her unconditional love, patience, and tireless support, especially during the entire developing process of this dissertation.

Braunschweig, December 2014

Balázs Fodor



# Abstract

This thesis deals with minimum mean square error (MMSE) speech enhancement schemes in the short-time Fourier transform (STFT) domain with a focus on statistical models for speech and corresponding estimators.

MMSE speech enhancement approaches taking speech presence uncertainty (SPU) into account usually consist of a common MMSE estimator for speech and an *a posteriori* speech presence probability (SPP) estimator. It is shown that both estimators should be based on the same statistical speech model, as they are in the same estimation framework and assume the same *a priori* knowledge. In order to give a synopsis of consistent MMSE estimation under SPU, typical common MMSE estimators and *a posteriori* SPP estimators are recapitulated. Furthermore, a new specific *a posteriori* SPP estimator is derived based on a novel statistical model for speech. Then, a synopsis of approaches to consistent MMSE estimation under SPU is given.

In the context of statistical modeling, we enhance a modern *a posteriori* SPP estimation approach based on fixed parameters. More precisely, the conservative speech model of this reference approach is replaced by an improved one. Then, a new *a posteriori* SPP estimator is derived and its fixed parameters are trained. The resulting proposed approach unifies the advantages of fixed parameters and a novel statistical speech model.

Although both speech enhancement and error concealment deal with distorted (speech) signals, there has not yet been an attempt to relate both fields to each other. However, since there are many commonalities between these disciplines, many interesting links between them are discussed based on recursive MMSE estimation. Furthermore, besides these commonalities, also interesting differences are analyzed and a general advantage of error concealment is identified. Based on this finding, research perspectives for the field of speech enhancement are sketched, inspired by error concealment.

This thesis provides a new statistical framework for *recursive* MMSE speech enhancement. This advantageously allows for applying the improved statistical models from classical, non-recursive speech enhancement to the recursive case. As a specific enhancement scheme, we extend recursive MMSE estimation by taking SPU into account.

Finally, a new reference-free signal-to-noise ratio (SNR) measurement approach is proposed in this thesis. This approach aims at estimating the SNR of a speech signal distorted by car noise as close as possible to reference-based measurement approach according to ITU-T Recommendation P.56, but in a reference-free fashion. The proposed approach achieves small estimation errors and shows high correlation with the ITU-T P.56 measurement within a typical SNR range. Furthermore, it provides relaxed computational complexity and can be applied to narrowband and wideband signals. Within ITU-T Study Group 12, the Focus Group on Car Communication (FG CarCOM) has decided to adopt the new reference-free SNR measurement approach for the draft of a recommendation proposal.



# Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit Störgeräuschreduktion (engl. speech enhancement) für Sprachsignale mittels frequenzbereichsbasierter MMSE-Schätzer (minimum mean square error, engl. für kleinster mittlerer quadratischer Fehler). Hierbei liegt ein besonderer Fokus auf den statistischen Sprachmodellen und den resultierenden Schätzregeln.

Spezifische MMSE-Verfahren, die die Unsicherheit der Sprachpräsenz (engl. speech presence uncertainty, SPU) berücksichtigen, bestehen aus einem allgemeinen MMSE-Sprachschätzer und einem Schätzer der Wahrscheinlichkeit von Sprachanwesenheit (engl. speech presence probability, SPP). Es wird gezeigt, dass beide Schätzer auf demselben statistischen Sprachmodell basieren, zudem werden üblicherweise verwendete allgemeine MMSE- und SPP-Schätzer rekapituliert. Darüber hinaus wird ein neuer SPP-Schätzer hergeleitet, der auf einem verbesserten statistischen Sprachmodell basiert. Danach wird ein Überblick über konsistente MMSE-Schätzverfahren mit SPU gegeben.

Im Kontext der statistischen Sprachmodellierung wird auch ein spezifisches, auf festen Parametern basierendes SPU-Verfahren weiterentwickelt. Das konservative Sprachmodell dieses SPU-Verfahrens wird durch ein verbessertes ersetzt und ein weiterer neuer SPP-Schätzer wird hergeleitet. Anschließend werden die festen Parameter der resultierenden Schätzregel trainiert. Dieses weiterentwickelte Verfahren vereint die Vorteile der festen Parameter und des verbesserten Sprachmodells.

Obwohl sich Störgeräuschreduktion und Fehlerverdeckung (engl. error concealment) mit der Aufgabe beschäftigen, gestörte (Sprach-)Signale zu verbessern, werden diese Verfahren typischerweise nicht in Beziehung zueinander gesehen. Da es jedoch viele Gemeinsamkeiten zwischen beiden Disziplinen gibt, werden bisher unbekannte Bezüge diskutiert. Darüber hinaus werden auch Unterschiede behandelt und ein grundsätzlicher Vorteil von Fehlerverdeckungsverfahren identifiziert. Motiviert durch diese Erkenntnis werden Forschungsperspektiven für das Themenfeld der Störgeräuschreduktion aufgezeigt.

Ferner wird eine neue, statistische Darstellung von rekursiver MMSE-Schätzung der Sprache präsentiert. Diese ermöglicht es, die modernen statistischen Modelle der klassischen, nicht-rekursiven Verfahren auf den rekursiven Fall anzuwenden. In diesem Kontext wird die rekursive MMSE-Schätzung mit einem SPU-Verfahren erweitert.

Schließlich wird ein neues, referenzfreies Messverfahren für das Signal-Rausch-Verhältnis (SNR) vorgestellt. Das Ziel des Verfahrens ist, das SNR eines von Fahrzeuggeräuschen gestörten Sprachsignals referenzfrei zu schätzen. Das Schätzergebnis soll so nah wie möglich am referenzbasierten Messverfahren nach ITU-T Recommendation P.56 liegen. Das neue Verfahren zeichnet sich durch kleine Messfehler und eine hohe Korrelation der Messwerte zum Referenzverfahren aus und kann mit Schmalband- sowie Breitbandsignalen verwendet werden. Die *Focus Group on Car Communication* (FG CarCOM) der ITU-T Study Group 12 hat beschlossen, das Verfahren in den Entwurf eines zukünftigen Standards aufzunehmen.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Outline of the Thesis . . . . .	4
<b>2</b>	<b>Synopsis of MMSE Speech Enhancement Approaches</b>	<b>7</b>
2.1	Introduction . . . . .	8
2.2	Error Criteria . . . . .	11
2.2.1	Minimum Mean Square Error (MMSE) Criterion . . . . .	11
2.2.2	Other Error Criteria . . . . .	12
2.3	Signal PDF Assumptions . . . . .	13
2.3.1	Speech Spectral PDF Assumptions . . . . .	13
2.3.2	Noise Spectral PDF Assumptions . . . . .	20
2.4	Estimation Domains . . . . .	20
2.4.1	Short-Time Spectral (STS) Estimation Domain . . . . .	21
2.4.2	Short-Time Spectral Amplitude (STSA) Estimation Domain . . . . .	21
2.4.3	Short-Time Log-Spectral Amplitude (LSA) Estimation Domain . . . . .	22
2.5	Synopsis of MMSE Estimation . . . . .	23
2.5.1	MMSE STS Estimation . . . . .	24
2.5.2	MMSE STSA Estimation . . . . .	28
2.5.3	MMSE LSA Estimation . . . . .	31
2.6	MMSE Estimation Under Speech Presence Uncertainty (SPU) . . . . .	36
2.6.1	<i>A Posteriori</i> SPP Estimation with Adapted Parameters . . . . .	37
2.6.2	<i>A Posteriori</i> SPP Estimation with Averaging and Fixed Parameters . . . . .	38
2.6.3	Estimation Domains . . . . .	42
2.7	Estimation of Noise Power, <i>A Priori</i> SNR, and <i>A Posteriori</i> SNR . . . . .	44
2.7.1	Noise Power Estimation . . . . .	45
2.7.2	<i>A Priori</i> and <i>A Posteriori</i> SNR Estimation . . . . .	47
2.8	Summary . . . . .	48
<b>3</b>	<b>Simulation Setup and Instrumental Measures</b>	<b>51</b>
3.1	Simulation Setup . . . . .	52

3.1.1	Databases . . . . .	52
3.1.2	Preprocessing . . . . .	53
3.1.3	White Box Test Setup . . . . .	53
3.2	Speech Enhancement Performance Measurement . . . . .	55
3.2.1	Speech Component . . . . .	56
3.2.2	Noise Component . . . . .	56
3.3	SNR Measurement . . . . .	57
3.3.1	Reference-Based SNR Measurement . . . . .	58
3.3.2	New Reference-Free SNR Measurement . . . . .	58
3.4	Summary . . . . .	67
<b>4</b>	<b>Consistent MMSE Estimation Under SPU</b>	<b>69</b>
4.1	Synopsis of <i>A Posteriori</i> SPP Estimation . . . . .	70
4.2	Synopsis of Consistent MMSE Estimation Under SPU . . . . .	74
4.2.1	MMSE STS Estimation Under SPU . . . . .	74
4.2.2	MMSE STSA Estimation Under SPU . . . . .	76
4.2.3	MMSE LSA Estimation Under SPU . . . . .	77
4.3	Performance Evaluation . . . . .	78
4.4	Summary . . . . .	84
<b>5</b>	<b>Consistent MMSE Estimation Under SPU with Averaging</b>	<b>85</b>
5.1	Introduction . . . . .	85
5.2	Algorithmic Approach . . . . .	86
5.2.1	<i>A Posteriori</i> SNR Averaging . . . . .	87
5.2.2	Training of the Fixed <i>A Priori</i> SNR . . . . .	89
5.3	Performance Evaluation . . . . .	92
5.4	Summary . . . . .	94
<b>6</b>	<b>Recursive MMSE Estimation and Links to Error Concealment</b>	<b>95</b>
6.1	Introduction . . . . .	96
6.2	Recursive MMSE Estimation in Speech Enhancement . . . . .	97
6.2.1	The Likelihood . . . . .	99
6.2.2	The Estimator . . . . .	99
6.2.3	The Prior . . . . .	100
6.2.4	The Kalman Filter . . . . .	102
6.3	Recursive MMSE Estimation in Error Concealment . . . . .	103
6.3.1	The Likelihood . . . . .	104
6.3.2	The Estimator . . . . .	106
6.3.3	The Prior . . . . .	106

6.4	Linking Speech Enhancement and Error Concealment . . . . .	108
6.4.1	The Likelihood . . . . .	108
6.4.2	The Estimator . . . . .	110
6.4.3	The Prior . . . . .	110
6.5	Outlook . . . . .	111
6.6	Summary . . . . .	113
<b>7</b>	<b>Recursive MMSE Estimation Under SPU</b>	<b>115</b>
7.1	Introduction . . . . .	116
7.2	Algorithmic Approach . . . . .	117
7.2.1	<i>A Posteriori</i> SPP Estimation . . . . .	118
7.2.2	<i>A Priori</i> SPP Estimation . . . . .	120
7.3	Performance Evaluation . . . . .	123
7.4	Summary . . . . .	128
<b>8</b>	<b>Conclusions</b>	<b>129</b>
<b>A</b>	<b>Bivariate and Polar Description of the Speech Prior</b>	<b>131</b>
<b>B</b>	<b>Approaches to PDF Parameter Identification</b>	<b>133</b>
<b>C</b>	<b>Bivariate and Polar Description of the Likelihood</b>	<b>137</b>
<b>D</b>	<b>Derivation of the PDF of Averaged <i>a Posteriori</i> SNRs</b>	<b>139</b>
<b>E</b>	<b>Synopsis of Recursive MMSE Estimation</b>	<b>143</b>
<b>F</b>	<b>Synopsis of Recursive <i>A Posteriori</i> SPP Estimation</b>	<b>147</b>
<b>G</b>	<b>Derivations for Recursive <i>A Priori</i> SPP Estimation</b>	<b>151</b>
	List of Symbols	155
	List of Abbreviations	159
	Bibliography	171
	Own Publications	174



# Chapter 1

## Introduction

Speech communication nowadays is widely supported by electronic devices which aim at, e.g., enhancing speech intelligibility (hearing aids), enabling speech conversations between distant users (telephones), or even reducing the distraction of communicating drivers in vehicles (hands-free systems). These devices are becoming increasingly popular, for example mobile phones are usually considered as personal assistants by their users, and are often meant to allow for robust speech communication also under harsh acoustic conditions. Speech enhancement deals with speech signals distorted by acoustic noise and aims to reduce the level of the distorting noise as much as possible, while keeping speech as unaffected as possible. Since these are conflicting goals, speech enhancement is a challenging task.

It is typically assumed in speech enhancement that speech is only observable through an acoustic channel adding distortion by superimposed acoustic noise. Commonly, both speech and acoustic noise are modeled as random processes, therefore, speech enhancement is typically a *probabilistic approach*. Generally, the goal is to *estimate* speech using the observed noisy speech and some *a priori* knowledge about the statistics of speech and the acoustic channel. This is the basic idea of *Bayes estimation*.

Being a specific Bayesian approach, minimum mean square error (MMSE) estimation is maybe the most popular estimation approach in speech enhancement and it is typically carried out in the short-time Fourier transform (STFT) domain. MMSE speech enhancement utilizes a statistical model of speech called speech prior and a statistical model of the acoustic channel called likelihood as *a priori* knowledge. Speech is then estimated by means of the observations and the underlying *a priori* knowledge. While first publications about MMSE speech enhancement were based on a Gaussian assumption for speech, e.g., [McAulay and Malpass, 1980], [Ephraim and Malah, 1984], it was later shown in [Porter and Boll, 1984], [Martin, 2002], and [Lotter and Vary, 2005] that the speech prior is actually a more heavy-tailed probability density function (PDF) than a Gaussian. Such a

PDF is also called super-Gaussian. Furthermore, speech enhancement approaches based on a super-Gaussian speech prior are shown to outperform approaches with a Gaussian assumption, e. g., [Martin, 2002], [Lotter and Vary, 2005], [Andrianakis and White, 2006], [Erkelens et al., 2007]. In [Dat et al., 2005] and [Erkelens et al., 2007] a generalized distribution is introduced which covers often employed speech spectral amplitude models as special case. Using typical assumptions, we will extend this so-called generalized gamma PDF for *complex-valued* speech STFT coefficients, resulting in a *bivariate* generalized gamma speech prior being a parametric statistical model. Usually employed (Gaussian or super-Gaussian) speech priors turn out as special case, each with an individual parameter set. Utilizing the new bivariate generalized gamma speech prior, a synopsis of typically employed speech statistical models and corresponding MMSE estimators will be given. Moreover, redefining the MMSE estimation formula by a further generalization step, the synopsis will be extended by the so-called estimation domains, namely the short-time spectral (STS), short-time spectral amplitude (STSA), and the log-spectral amplitude (LSA) domain. As a consequence, the synopsis will cover typically employed speech priors and estimation domains.

Common MMSE speech enhancement assumes that the speech signal is always present. This is clearly not fulfilled for natural speech as uttered by human beings, e. g., during natural speech pauses. Therefore, it is shown, e. g., in [McAulay and Malpass, 1980] and [Ephraim and Malah, 1984], that MMSE speech enhancement can be improved by taking speech presence uncertainty (SPU) into account. MMSE speech enhancement under SPU turns out to be a combination of two estimators: A common MMSE estimator for speech and an *a posteriori* speech presence probability (SPP) estimator. Similar to the former, the latter estimator also utilizes *a priori* knowledge about speech and the acoustic channel which, according to estimation theory, is the same as that of the common MMSE estimator [Middleton and Esposito, 1968]. Accordingly, we argue that *both estimators have to be based on the same signal spectral assumptions*, henceforth called consistent assumptions. Therefore, employing MMSE estimation under SPU, the same PDF assumptions should be employed for both the common MMSE estimator and the *a posteriori* SPP estimator. In order to see which *a posteriori* SPP estimator known from literature can be combined in a PDF-consistent way with respective common MMSE estimators, we will give a clear synopsis of SPP estimators. Furthermore, based on the new bivariate generalized gamma speech prior we will derive a new generalized *a posteriori* SPP estimator covering specific *a posteriori* SPP estimators from literature based on either Gaussian or super-Gaussian speech priors [Fodor and Fingscheidt, 2012a]. Moreover, we will derive a new specific *a posteriori* SPP estimator based on a super-Gaussian speech assumption. Then, a synopsis of PDF-consistent MMSE speech enhancement approaches under SPU will be given in different estimation domains.

Common *a posteriori* SPP estimators typically suffer from the random fluctuations of

the observations, often resulting in estimation outliers which may be perceived as annoying musical noise. In order to enhance estimation robustness and reduce estimation outliers, it is proposed in, e.g., [Suhadi and Fingscheidt, 2007] and [Gerkmann et al., 2008], to base *a posteriori* SPP estimation on averaged observations. This approach, however, does not take into account the super-Gaussian nature of speech STFT coefficients. The derivation of *a posteriori* SPP estimators using averaged observations and assuming super-Gaussian speech models turns out to be mathematically complex, therefore, only approximate solutions have been proposed yet [Fodor and Gerkmann, 2014a]. In this thesis, however, we will provide a closed-form solution instead of the approximate one [Fodor and Gerkmann, 2014b]. Furthermore, while common *a posteriori* SPP estimators are able to robustly achieve values close to one in speech presence, they typically output the *a priori* SPP during speech absence. Since values of 0.5 or larger are generally chosen as *a priori* SPP [McAulay and Malpass, 1980], the resulting *a posteriori* SPP estimates are less accurate in speech absence. To overcome this issue, fixed prior parameters (a fixed *a priori* signal-to-noise ratio (SNR) and a fixed *a priori* SPP) will be employed for SPU estimation as in [Gerkmann et al., 2008], resulting in more precise *a posteriori* SPP estimates.

MMSE estimation is a widely employed approach to, e.g., estimate signals which can only be observed through an error-prone channel. Therefore, it is often used in speech enhancement, but also in error concealment dealing with speech (or audio) signals distorted while passing through a transmission channel. While both disciplines seem to have similar tasks, there has rarely been an attempt to relate these disciplines to each other. In this thesis, we will show interesting links between speech enhancement and error concealment based on recursive MMSE estimation [Fodor et al., 2015]. In particular, utilizing the same PDF-based description as for the synopses, we will recapitulate recursive MMSE estimation in speech enhancement [Fodor et al., 2015] and recursive MMSE estimation in error concealment [Fingscheidt, 1998], [Pflug and Fingscheidt, 2013b], [Pflug, 2013] in strong analogy. Then, we will show interesting commonalities and differences between these disciplines and identify a specific strength of error concealment. Motivated by this finding, we will sketch possible research perspectives for speech enhancement. Please note that, to the best of our knowledge, the PDF-based description of recursive MMSE speech enhancement presented in this thesis is new; in speech enhancement rather a state-space representation is usual based on matrix notation. The new PDF-based framework allows for applying enhancement schemes of classical non-recursive speech enhancement to the recursive case. Such a PDF-related enhancement of recursive MMSE estimation is proposed in [Esch and Vary, 2008a], where the update step of the Kalman filter is intuitively recognized as a classical MMSE estimation step. Accordingly, the Kalman gain was replaced in a heuristic way by more modern spectral weighting rules based on a super-Gaussian assumption for the speech. We will show that this replacement is indeed optimal in an MMSE sense, if the employed

spectral weighting rule belongs to an MMSE STS estimator.

As a further PDF-related enhancement scheme, the recursive MMSE speech enhancement framework will be extended by SPU estimation in this thesis for the first time. Both a general recursive MMSE estimation formula under SPU and a general *a posteriori* SPP formula taking signal history into account will be derived. Moreover, specific *a posteriori* SPP estimators will be provided based on either Gaussian or super-Gaussian speech PDF assumptions. Different from the non-recursive case, the *a priori* SPP will turn out to make use of the signal history, therefore, a tracking algorithm will be proposed.

The knowledge of the SNR of speech signals plays an important role in speech signal processing. A possible example is the measurement of SNR improvement of speech enhancement approaches or, as an automotive example, the optimization of hands-free microphone positions with respect to (w. r. t.) the acoustic SNR. A typical reference-based SNR measurement approach is provided by Recommendation P.56 of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T). Here, the term *reference-based* means that the SNR of a noisy speech signal is measured signal-component-wise, i. e., the SNR is computed by measuring the signal level of the speech and noise components separately. In this thesis, we will propose an SNR measurement approach which aims at estimating the SNR of a speech signal distorted by car noise as close as possible to the ITU-T P.56 reference, but without using a reference. The proposed approach, therefore, is able to measure the SNR by using the noisy speech signal only and can be applied to both narrowband and wideband signals. Within ITU-T Study Group 12, the Focus Group on Car Communication (FG CarCOM) has decided to adopt the new reference-free SNR measurement approach within the draft of a current recommendation proposal.

## 1.1 Outline of the Thesis

Chapter 2 will give a short introduction of MMSE speech enhancement, focusing on usual statistical models for speech STFT coefficients in the polar domain. Furthermore, using some assumptions, a new bivariate generalized gamma speech prior will be derived which covers typically employed Gaussian or super-Gaussian speech priors as special case. A further generalization will cover typical estimation domains, such as the STS, the STSA, and the LSA domain as special case. Utilizing the new bivariate generalized gamma prior, a synopsis of common MMSE estimators for speech enhancement will be given in each estimation domain.

Chapter 3 will introduce instrumental measures for speech enhancement and the simulation setup employed in this thesis. In particular, first employed databases, simulation setup, and a white box test setup will be described which are employed in this thesis for performance

evaluation of speech enhancement approaches. This will be followed by the introduction of performance measures used for evaluation. Finally, a reference-based SNR measurement approach according to ITU-T Recommendation P.56 and a proposed reference-free SNR measurement approach will be introduced.

Chapter 4 will provide a synopsis of *a posteriori* SPP estimators w.r.t. statistical assumptions for speech in analogy to the synopsis of common MMSE estimators in Chapter 2. This synopsis will be based on a new generalized *a posteriori* SPP estimator which covers *a posteriori* SPP estimators from literature as special case. Furthermore, a new specific *a posteriori* SPP estimator will be proposed based on a super-Gaussian speech prior. Finally, both the new SPU-based approach as well as non-SPU and SPU-based approaches from Chapter 2 will be assessed by performance measures.

Chapter 5 will deal with an enhancement scheme of *a posteriori* SPP estimation. More precisely, we will replace the Gaussian speech prior of a state-of-the-art *a posteriori* SPP estimator based on averaging and fixed prior parameters by a modern, super-Gaussian speech prior. First, we will derive the new estimator which will be followed by a training of its parameters. Finally, the performance of the proposed approach along with some SPU-based approaches from Chapter 4 will be evaluated.

Chapter 6 will aim at revealing links between speech enhancement and error concealment based on recursive MMSE estimation. First, we will recapitulate recursive MMSE estimation in speech enhancement and the Kalman filter in a PDF-based manner, similar to the notation in the previous chapters. Then, we will introduce recursive MMSE estimation in error concealment, presented in analogy to speech enhancement. Next, some interesting links between those two disciplines will be shown and a general strength of error concealment will be identified. Based on this finding, possible research directions for speech enhancement will be sketched.

Chapter 7 will deal with the extension of recursive MMSE speech enhancement by SPU estimation. Using the new PDF-based recursive MMSE framework from Chapter 6, a general recursive MMSE estimation formula under SPU and a general *a posteriori* SPP estimation formula will be derived. Furthermore, the *a priori* SPP turns out to take signal history into account, thus, a tracking algorithm will be provided. Moreover, specific *a posteriori* SPP estimators will be derived utilizing either Gaussian or super-Gaussian speech PDF assumptions. Finally, the performance of recursive MMSE estimation under SPU with a Gaussian signal assumption will be evaluated.

Finally, Chapter 8 will shortly summarize this thesis.



# Chapter 2

## Synopsis of

## Minimum Mean Square Error

## Speech Enhancement Approaches

In this chapter a synopsis of speech enhancement approaches will be given based on MMSE estimation and carried out in the STFT domain. Accordingly, Section 2.1 will give an introduction to the problem of speech enhancement and will shortly recapitulate the idea behind Bayesian estimation which is a typically employed solution to this problem. Section 2.2 will give a short overview of MMSE estimation which is widely employed in speech enhancement. As a specific Bayesian approach, MMSE estimation is carried out by means of the observed noisy speech and the underlying *a priori* knowledge about the speech signal to be estimated and the acoustic channel. Generally used *a priori* knowledge will be introduced in Section 2.3, where a generalized, parametric statistical model will be employed for speech, nicely covering all commonly employed statistical speech models as special case, each with an individual parameter set. This *new bivariate generalized gamma speech prior*, being a real-valued PDF with a complex-valued argument, is an extension of the univariate generalized gamma probability density, often employed for modeling the speech spectral amplitudes. MMSE speech enhancement is typically carried out in the STS, the STSA, or the LSA estimation domain which will be introduced in Section 2.4. Based on the generalization w. r. t. the speech model, we will derive general (parametric) MMSE estimators for each estimation domain. This allows for a hierarchical overview of MMSE speech enhancement approaches in Section 2.5: Applying the parameter set of a specific speech spectral model to the generalized (parametric) estimator leads to a specific MMSE estimator from literature. By this means, relations among the resulting estimators can easily be shown. In Section 2.6 we will extend the parametric MMSE estimation formula from Section 2.2 by taking SPU into account. This allows for an overview of SPU estimation in the STS, STSA, and LSA

estimation domain. Furthermore, state-of-the-art *a posteriori* SPP estimators will shortly be reviewed. Finally, Section 2.7 will briefly recapitulate state-of-the-art SNR and noise power spectral density (PSD) estimators.

## 2.1 Introduction

Reduction of annoying background noise is a non-trivial task and implies a big challenge for the design of speech enhancement systems. In speech enhancement the following model is widely employed: We have an acoustic speech signal (e. g., that of the user of a hands-free system) and acoustic background noise signals (e. g., engine noise, wheel noise, wind noise, etc. in a vehicle which the user of the hands-free system is located in), both propagating in air and both captured by the microphone(s) of the, e. g., hands-free device. The (analog) electrical signal at the output of the microphone is subsequently digitized and the resulting digital signal is commonly modeled on signal level as follows: The samples of a time-domain random speech process  $s(n)$  with  $n$  being the discrete time index are transmitted through an acoustic channel which distorts these samples by superimposing statistically independent samples  $d(n)$  being outcomes of a random noise process. While the speech and noise samples are inaccessible in practice, the resulting noisy speech samples  $y(n) = s(n) + d(n)$  can be observed. The aim of a speech enhancement system is to reduce the background noise component  $d(n)$  of the observed noisy speech signal  $y(n)$  in such a way that the desired speech component  $s(n)$  remains as unaffected as possible.

The speech may be distorted in a different manner along the frequencies, thus, speech signals are usually processed in the frequency domain, in order to allow for a corresponding fine, frequency-dependent processing. More specifically, most often the STFT domain is employed [Vary and Martin, 2006]. This means, that the noisy speech samples  $y(n)$  are processed block-wise and sequentially, within so-called frames. Frames are built by applying an analysis window function  $w_a$  of the length  $L$  to a portion of speech samples  $s$  with the same length  $L$ . The frame length  $L$  is normally chosen from the range of 5–30 ms. This range is justified by the fact that although the speech signal  $s(n)$  is naturally non-stationary [Brehm and Stammer, 1987], the resulting short-time speech portions can be assumed to be quasi-stationary [Vary and Martin, 2006]. As a next step, the noisy speech samples within the frame can be transformed into the STFT domain by, e. g., a discrete Fourier transform (DFT) of the size  $L$ . Then, the analysis window is shifted forwards  $\Delta L$  samples and the process starts over. These steps in summary are called signal analysis and can be expressed by the STFT as [Vary and Martin, 2006]

$$Y_\ell(k) = \sum_{n'=0}^{L-1} w_a(n') \cdot y(\ell \cdot \Delta L + n') \cdot e^{-j2\pi \frac{n'k}{L}} \quad (2.1)$$

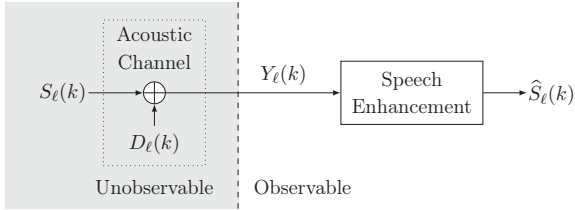


Figure 2.1: Block diagram of speech enhancement with an underlying signal and channel model

with  $\ell = 0, 1, 2, \dots$  being the frame index,  $k = 0, 1, \dots, L - 1$  denoting the frequency bin index,  $n'$  indexing samples within the frame, and  $\Delta L$  standing for the frame shift in samples. Please note that a possible criterion for choosing the frame overlap  $(L - \Delta L)/L$  and the employed analysis window  $w_a$  is that the STFT coefficients  $Y_\ell(k)$  yield exactly the original signal  $y(n)$  after signal synthesis which is the inverse operation to signal analysis. Signal synthesis is commonly divided into an inverse discrete Fourier transform (IDFT) operation

$$v_\ell(n') = \frac{1}{L} \sum_{k=0}^{L-1} Y_\ell(k) \cdot e^{j2\pi \frac{n'k}{L}}, \quad n' = 0, 1, \dots, L - 1, \quad (2.2)$$

and, e. g., an overlap-add (OLA) step

$$y(\ell \cdot \Delta L + n') = y(\ell \cdot \Delta L + n') + w_s(n') \cdot v_\ell(n') \quad (2.3)$$

with  $n' = 0, 1, \dots, L - 1$  being the sample index within a frame and  $w_s$  denoting the synthesis window. Please note that (2.3) only holds for  $(L - \Delta L)/L \leq 50\%$ .

Due to the DFT properties [Proakis and Manolakis, 2007] the acoustic noise remains additive in the STFT domain resulting in the signal and channel model as depicted in Figure 2.1: The unobservable random, complex-valued speech STFT coefficient  $S_\ell(k)$  passes through an acoustic channel which distorts it by superimposing a random, complex-valued acoustic noise STFT coefficient  $D_\ell(k)$ , i. e.,

$$Y_\ell(k) = S_\ell(k) + D_\ell(k). \quad (2.4)$$

Employing polar notation, (2.4) can equivalently be written as

$$R_\ell(k)e^{j\vartheta_\ell(k)} = A_\ell(k)e^{j\alpha_\ell(k)} + D_\ell(k) \quad (2.5)$$

with  $R_\ell(k) = |Y_\ell(k)|$ ,  $\vartheta_\ell(k) = \arg\{Y_\ell(k)\}$ ,  $A_\ell(k) = |S_\ell(k)|$ , and  $\alpha_\ell(k) = \arg\{S_\ell(k)\}$  being the noisy speech spectral amplitude, the noisy speech spectral phase, the speech spectral amplitude, and the speech spectral phase, respectively.

It is widely assumed (and will be assumed in the following chapters of this thesis) that the speech STFT coefficients  $S_\ell(k)$  are statistically independent of each other both along the

frequency bins  $k$  and frames  $\ell$ . The same assumption holds for the noise and the noisy speech STFT coefficients [Ephraim and Malah, 1984]. However, employing a dynamic speech model in the form of an autoregressive (AR) process, temporal correlation of STFT coefficients can be exploited as we will see in Chapters 6 and 7. Furthermore, it is widely assumed in speech enhancement that the speech STFT coefficient  $S_\ell(k)$  and the noise STFT coefficient  $D_\ell(k)$  are statistically independent.

While the speech STFT coefficient  $S_\ell(k)$  and the noise STFT coefficient  $D_\ell(k)$  are unobservable, the resulting noisy speech STFT coefficient  $Y_\ell(k)$  can be observed, thus, it is also called *observation*. The aim of a speech enhancement system is to recover the speech component  $S_\ell(k)$  when only its distorted replica  $Y_\ell(k)$  is observed. For this purpose, generally the well-known Bayesian estimation theory is employed.

While there are approaches utilizing merely the observations for the estimation process, Bayesian estimation further exploits *a priori* knowledge to enhance the estimation performance [Kay, 1993]. More specifically, besides the observations  $Y_\ell(k)$ , statistical models of both the desired speech STFT coefficients  $S_\ell(k)$  and the channel including the noise STFT coefficients  $D_\ell(k)$  are employed *prior to the observation* for the estimation process. The underlying *a priori* knowledge, i. e., the signal and the channel models will be discussed in Section 2.3.

As a next step, we wish to carry out an estimation of the clean speech STFT coefficient  $S_\ell(k)$  using some *a priori* knowledge and the observation  $Y_\ell(k)$ . According to Bayesian estimation theory, this can be achieved by minimizing the so-called Bayes risk which is defined as<sup>1</sup> [Eykhoff, 1974], [Lotter and Vary, 2005]

$$R(\hat{S}) = E_Y\{E_S\{C(S, \hat{S} = f(Y))\}\} \quad (2.6)$$

with  $E_Y\{\cdot\}$  and  $E_S\{\cdot\}$  being the expectation operator w. r. t. the noisy STFT coefficient  $Y$  and the speech STFT coefficient  $S$ , respectively, and with  $C(S, \hat{S} = f(Y))$  being a (mostly real-valued) cost function, e. g.,  $C(S, \hat{S}) = |S - \hat{S}|^2$ . The function  $f(\cdot)$  models the estimation process which maps each observation  $Y$  to a speech estimate  $\hat{S}$ . The goal is to find a speech estimate  $\hat{S}$  which minimizes the risk  $R(\hat{S})$  as [Eykhoff, 1974]

$$\hat{S} = \arg \min_{\hat{S}} R(\hat{S}) = \arg \min_{\hat{S}} \int_{\mathbb{C}} \int_{\mathbb{C}} C(S, \hat{S}) \cdot p_{S,Y}(S, Y) dS dY \quad (2.7)$$

with  $p_{S,Y}(S, Y)$  being the joint PDF of the clean speech STFT coefficient  $S$  and the noisy speech STFT coefficient  $Y$  and  $\mathbb{C}$  being the set of complex-valued numbers. Employing Bayes' rule  $p_{S,Y}(S, Y) = p_Y(Y)p_{S|Y}(S|Y)$ , (2.7) can be rewritten as [Vary and Martin, 2006]

<sup>1</sup>Please note that for ease of simplicity the indices  $\ell$  and  $k$  will be omitted in the remainder of this chapter.

$$\hat{S} = \arg \min_{\hat{S}} \int_{\mathbb{C}} p_Y(Y) \int_{\mathbb{C}} C(S, \hat{S}) \cdot p_{S|Y}(S|Y) dS dY \quad (2.8)$$

with  $p_Y(Y)$  and  $p_{S|Y}(S|Y)$  being the so-called *evidence* and the so-called *posterior*, respectively. The posterior  $p_{S|Y}(S|Y)$  is the PDF of the speech STFT coefficient  $S$  given a specific observation  $Y$  (being an outcome of the random speech process after passing through the acoustic channel), thus, it contains *a priori* knowledge about the speech and the channel. The inner integral in (2.8), namely  $E\{C(S, \hat{S})|Y\}$ , is called the conditional mean error, the outer integral is called the total mean error. Substituting a specific cost function  $C(S, \hat{S})$  into (2.8), the minimization problem could be solved. Considering, however, that we want to minimize the total mean error w. r. t.  $\hat{S}$ , the evidence  $p_Y(y)$  is non-negative by definition, and in this case an integral is minimal, if its integrand is minimal, it is sufficient to minimize the conditional mean error as [Van Trees, 1968], [Vary and Martin, 2006]

$$\hat{S} = \arg \min_{\hat{S}} \int_{\mathbb{C}} C(S, \hat{S}) \cdot p_{S|Y}(S|Y) dS. \quad (2.9)$$

This is the basic concept of Bayesian estimation and (2.9) is its general formula. Now in order to solve the minimization problem (2.9), a specific cost function  $C(S, \hat{S})$  is needed. In this thesis we employ the cost function according to the MMSE criterion. This cost function and a few more will briefly be introduced in the next section.

## 2.2 Error Criteria

### 2.2.1 Minimum Mean Square Error (MMSE) Criterion

The cost function of the MMSE criterion is generally defined as [Van Trees, 1968], [Hendriks et al., 2009a]

$$C_{\text{MMSE}}(g(S), \widehat{g(S)}) = \left| g(S) - \widehat{g(S)} \right|^2 \quad (2.10)$$

with  $g(\cdot)$  being an arbitrary function<sup>2</sup>. This means, that a linearly increasing absolute estimation error produces quadratically growing costs. Applying (2.10) to the Bayes estimator (2.9) and using that  $E\{g(X)\} = \int_{-\infty}^{\infty} g(X) \cdot p(X) dX$  [Papoulis and Pillai, 2002], we obtain

$$\widehat{g(S)} = \arg \min_{g(S)} \int_{\mathbb{C}} \left| g(S) - \widehat{g(S)} \right|^2 \cdot p_{S|Y}(S|Y) dS. \quad (2.11)$$

---

<sup>2</sup>Please note, that this generalization allows for introducing later in Section 2.4 the so-called estimation domains, i. e., the STS, the STSA, and the LSA estimation domain. Applying, e. g., the STS estimation domain  $g(X) = X$  with  $X \in \mathbb{C}$  results in the cost function  $C_{\text{MMSE STS}}(S, \hat{S}) = |S - \hat{S}|^2$  which is widely used in literature, e. g., in [Vary and Martin, 2006].

This means, that the cost function  $|g(S) - \widehat{g(S)}|^2$  is calculated for all possible  $\widehat{g(S)}$  values, while the posterior  $p_{S|Y}(S|Y)$  is fixed. The optimal  $\widehat{g(S)}$  results by minimizing the integral of the product  $|g(S) - \widehat{g(S)}|^2 \cdot p_{S|Y}(S|Y)$ . Please note the special case  $g(X) = X$  leading to the parabolic cost function  $|S - \widehat{S}|^2$ . Within the optimization process, this paraboloid is shifted through the whole complex plain by changing the value of  $\widehat{S}$ , while the posterior  $p_{S|Y}(S|Y)$  is fixed. Still assuming  $g(X) = X$ , the integral (2.11) is minimal, if  $\widehat{S}$  is equal to the argument of the maximum of the posterior, assuming that the latter is a rotationally symmetric PDF. Accordingly, the optimal speech estimate turns out to be  $\widehat{S} = \arg \max_S p_{S|Y}(S|Y)$  which coincides with the maximum *a posteriori* (MAP) solution, as we will see in Section 2.2.2 (cf. (2.14)).

Assuming that  $C_{\text{MMSE}}(g(S), \widehat{g(S)})$  is a convex function (which is fulfilled for usually employed functions  $g(\cdot)$ ), the solution of the optimization problem in (2.11) turns out to be [Porter and Boll, 1984], [Hendriks et al., 2009a]

$$\widehat{g(S)} = E \{g(S)|Y\} = \int_{\mathbb{C}} g(S) \cdot p_{S|Y}(S|Y) dS = \frac{\int_{\mathbb{C}} g(S) \cdot p_{Y|S}(Y|S) \cdot p_S(S) dS}{\int_{\mathbb{C}} p_{Y|S}(Y|S) \cdot p_S(S) dS} \quad (2.12)$$

with  $p_{Y|S}(Y|S)$  being the *likelihood* which can be associated with the underlying channel model for observation  $Y$ ,  $p_S(S)$  being the *prior* which can be associated with the underlying (speech) signal model, as well as with  $\int_{\mathbb{C}} p_{Y|S}(Y|S) \cdot p_S(S) dS = p_Y(Y)$  being the evidence which is usually calculated by marginalizing the PDF product in the numerator of (2.12). Please note, that the well-known Wiener filter is optimal in the MMSE sense [Kay, 1993].

Common priors and likelihoods are recapitulated in Section 2.3, then, typical functions  $g(\cdot)$  in (2.12) are discussed in Section 2.4.

### 2.2.2 Other Error Criteria

There are other error criteria which are utilized in speech enhancement. Assuming linearly growing costs as a consequence of linearly increasing absolute estimation errors, i.e., employing [Van Trees, 1968]

$$C_{\text{MAP}}(S, \widehat{S}) = \begin{cases} 0, & \text{if } |S - \widehat{S}| \equiv 0, \\ 1, & \text{otherwise} \end{cases} \quad (2.13)$$

for the Bayesian estimation formula (2.9), leads to the MAP estimation formula [Van Trees, 1968]

$$\widehat{S} = \arg \max_S p_{S|Y}(S|Y) = \arg \max_S p_{Y|S}(Y|S) p_S(S). \quad (2.14)$$

Please note that the Wiener filter is optimal also in a MAP sense [Vary and Martin, 2006]. However, in speech enhancement the MAP estimator is employed in a slightly modified form either merely for the spectral amplitude  $|S| = A$  [Lotter and Vary, 2005]

$$\hat{A} = \arg \max_A p_{R|A}(R|A) \cdot p_A(A) \quad (2.15)$$

or separately for the speech spectral amplitude  $A = |S|$  and phase  $\alpha = \arg\{S\}$  [Lotter and Vary, 2005]

$$\hat{\alpha} = \arg \max_{\alpha} p_{Y|A,\alpha}(Y|A, \alpha) p_{A,\alpha}(A, \alpha), \quad (2.16)$$

$$\hat{A} = \arg \max_A p_{Y|A,\alpha}(Y|A, \hat{\alpha}) p_{A,\alpha}(A, \hat{\alpha}), \quad (2.17)$$

respectively, leading to estimators such as those in [Wolfe and Godsill, 2003b], [Lotter and Vary, 2005], [Dat et al., 2005].

There are perceptually motivated error criteria taking psychoacoustics into account, e. g., [Wolfe and Godsill, 2000], [Gustafsson et al., 2002], [Wolfe and Godsill, 2003a], and [Loizou, 2005].

In [Fingscheidt et al., 2008, Eq. (17)] the cost function  $\||S| - \widehat{|S}||^2/|Y|^2$  is proposed in the context of data-driven speech enhancement. It is interesting to note that employing this cost function for (2.9) leads to the MMSE solution (2.12) with  $g(X) = |X|$ , because the denominator  $|Y|^2$  in the cost function has only a scaling effect on the integrand in (2.9) (being optimized w. r. t.  $S$ ) which does not influence the position of its extrema.

## 2.3 Signal PDF Assumptions

As can be seen in (2.12), the estimation is based on statistical models for the clean speech STFT component  $S$  and the acoustic channel in the form of the prior  $p_S(S)$  and the likelihood  $p_{Y|S}(Y|S)$ , respectively. In this section, typically employed priors and likelihoods are recapitulated.

### 2.3.1 Speech Spectral PDF Assumptions

The prior  $p_S(S)$  in (2.12) provides a statistical model of the (unobservable) speech STFT coefficient  $S$  and is a bivariate PDF, i. e., a real-valued function with a complex-valued argument  $S$ . Usually, however, the prior is described as a function of either the real part  $S_{\text{re}} = \text{Re}\{S\}$  and the imaginary part  $S_{\text{im}} = \text{Im}\{S\}$  of the speech STFT coefficient  $S$  or the spectral amplitude  $A = |S|$  and phase  $\alpha = \arg\{S\}$ , i. e.,  $S = S_{\text{re}} + jS_{\text{im}} = Ae^{j\alpha}$ . While there are speech enhancement schemes based on models for the real and imaginary parts, e. g., [Martin, 2002],

[Martin, 2005], [Erkelens et al., 2007], in this thesis we consider approaches based on the polar representation using  $A$  and  $\alpha$ , just as in, e.g. [McAulay and Malpass, 1980], [Ephraim and Malah, 1984], [Wolfe and Godsill, 2003b], [Lotter and Vary, 2005], [Erkelens et al., 2007].

Histogram measurements and corresponding contour plots in [Lotter and Vary, 2005], [Erkelens et al., 2007] showed that the PDF of the speech STFT coefficients  $p_S(S = Ae^{j\alpha})$  is approximately circular. Thus, it can be assumed that the speech prior is rotationally symmetric which implies for its polar description that [Erkelens et al., 2007]

- the speech spectral amplitude  $A$  and the speech spectral phase  $\alpha$  are statistically independent and
- the speech spectral phase  $\alpha$  is uniformly distributed on  $[0, 2\pi)$ , i.e.,  $p_\alpha(\alpha) = \frac{1}{2\pi}$  for  $0 \leq \alpha < 2\pi$ .

Note that these are very commonly employed assumptions in speech enhancement. Consequently, the following relationship turns out between the bivariate speech spectral PDF  $p_S(S)$  and the polar representation with  $A$  and  $\alpha$  (the derivation is given in Appendix A)

$$p_S(Ae^{j\alpha}) = \frac{1}{A} \cdot p_{A,\alpha}(A, \alpha) = \frac{1}{A} \cdot p_A(A) \cdot p_\alpha(\alpha). \quad (2.18)$$

This means that under the aforementioned assumptions the prior  $p_S(S = Ae^{j\alpha})$  can fully be described by the speech spectral phase PDF  $p_\alpha(\alpha)$  and the speech spectral amplitude PDF  $p_A(A)$ . Furthermore, since  $p_\alpha(\alpha)$  is commonly modeled by a constant, it has a rather marginal role in speech spectral modeling. Meanwhile, the speech spectral amplitude PDF  $p_A(A)$  has a very significant role in the description of the prior  $p_S(S = Ae^{j\alpha})$ . Moreover, it is rather unusual to describe the prior by a bivariate PDF. Instead, mostly a (univariate) speech spectral amplitude PDF  $p_A(A)$  is employed in conjunction with a statistically independent and uniformly distributed speech spectral phase assumption which allows for a full description of the speech prior  $p_S(S)$  according to (2.18). Thus, whenever the prior is modeled by merely  $p_A(A)$  in literature, the assumptions leading to (2.18) are implicitly made.

In speech enhancement the following priors are generally utilized which also have great influence on the speech spectral amplitude model:

- A Gaussian prior [McAulay and Malpass, 1980], [Ephraim and Malah, 1984], [Ephraim and Malah, 1985], [Wolfe and Godsill, 2003b],
- a super-Gaussian prior [Martin, 2002], [Lotter and Vary, 2005], [Andrianakis and White, 2006], [Erkelens et al., 2007], [Hendriks et al., 2009b], or
- a generalized prior which covers Gaussian and super-Gaussian priors as special case [Dat et al., 2005], [Erkelens et al., 2007].

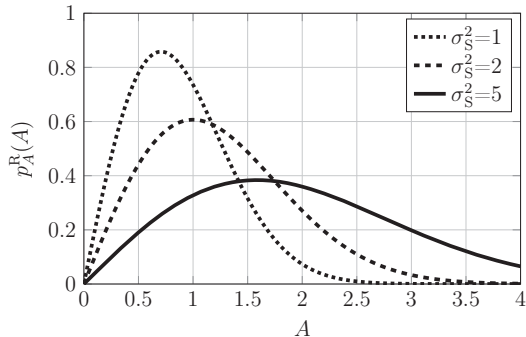


Figure 2.2: PDF of Rayleigh-distributed speech spectral amplitudes

### Gaussian Prior

The assumptions that the speech spectral amplitude and phase are statistically independent, with the latter being uniformly distributed, are fulfilled, if we assume a bivariate Gaussian distribution for the speech STFT coefficient  $S$  as [Martin, 2002]

$$p_S^G(S) = \frac{1}{\pi\sigma_S^2} \cdot e^{-\frac{|S|^2}{\sigma_S^2}}, \quad S \in \mathbb{C} \quad (2.19)$$

with superscript G denoting the Gaussian assumption and  $\sigma_S^2 = E\{|S|^2\}$  being the PSD of the speech STFT coefficient  $S$ . This assumption is widely used in literature, e.g., in [McAulay and Malpass, 1980], [Ephraim and Malah, 1984], [Ephraim and Malah, 1985], [Wolfe and Godsill, 2003b] and is usually justified by the asymptotic properties of STFT coefficients [Brillinger, 2001]: Assuming that the DFT length is sufficiently large and the span of correlation of the random samples to be transformed is sufficiently short, the central limit theorem can be applied. Thus, the STFT coefficient  $S$  can be modeled by a zero-mean bivariate Gaussian with statistically independent and identically distributed (i. i. d.) real and imaginary parts, each being a univariate Gaussian with variance  $\sigma_S^2/2$ . The corresponding speech spectral amplitude  $A = |S|$  follows the *Rayleigh* distribution [Papoulis and Pillai, 2002] and its PDF is defined as [Ephraim and Malah, 1984]

$$p_A^R(A) = 2\beta_R \cdot A \cdot e^{-\beta_R A^2}, \quad \beta_R > 0, A \geq 0 \quad (2.20)$$

with index R denoting the Rayleigh distribution and  $\beta_R = 1/\sigma_S^2$ . Please note that besides  $A$  the omitted indices  $(\ell, k)$  also apply to  $\beta_R$  and  $\sigma_S^2$ . As can be seen, this PDF only has the parameter  $\beta_R$  which is basically related to the width of the PDF (cf. Figure 2.2). Please note that a Rayleigh-distributed speech spectral amplitude  $A$  always implies a bivariate Gaussian-distributed speech STFT coefficient under the assumptions that the speech spectral

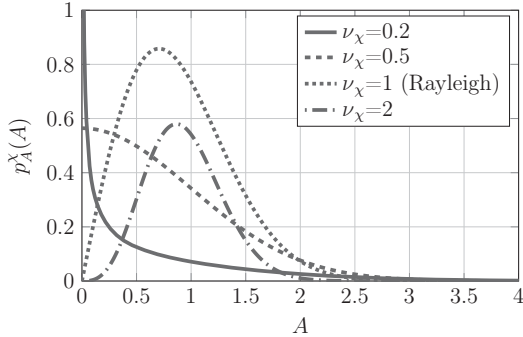


Figure 2.3: PDF of chi-distributed speech spectral amplitudes with  $\sigma_S^2 = 1$

amplitude and the phase are statistically independent and that the speech spectral phase is uniformly distributed [Papoulis and Pillai, 2002].

### Super-Gaussian Priors

The asymptotic assumptions for speech STFT coefficients are not fulfilled in practice, e.g., due to a limited frame length (cf. Section 2.1). Accordingly, it is shown in [Martin, 2002] that the real and imaginary parts of a speech STFT coefficient  $S$  rather follow a distribution with a sharper peak and heavier tails compared to a Gaussian. Furthermore, histogram measurements in [Porter and Boll, 1984] and [Lotter and Vary, 2005] showed that the speech spectral amplitudes can better be modeled by a more heavy-tailed distribution than the Rayleigh one. Accordingly, it can be assumed that speech STFT coefficients rather follow a super-Gaussian distribution instead of a Gaussian. An arbitrary random variable  $X$  (e.g., a speech STFT coefficient) is super-Gaussian distributed (or leptokurtic or heavy-tailed), if its excess [Abramowitz and Stegun, 1972]

$$\psi(X) = \frac{E\{(X - E\{X\})^4\}}{E^2\{(X - E\{X\})^2\}} - 3 \quad (2.21)$$

is positive [Vaseghi, 2008]. If  $X$  is Gaussian distributed, its excess yields zero.

Please note that the adjectives Gaussian and super-Gaussian will be related to the bivariate speech prior  $p_S(S)$  in the rest of this thesis. The Rayleigh distribution and the speech spectral amplitude distributions which will be introduced in the following all belong to the class of super-Gaussian distributions, i. e., have a positive excess.

In the context of super-Gaussian priors, most often the *chi distribution* [Andrianakis and White, 2006], [Hendriks et al., 2009b] or the *gamma distribution* [Martin, 2002], [Lotter and Vary, 2005], [Andrianakis and White, 2006], [Erkelens et al., 2007] is employed as speech

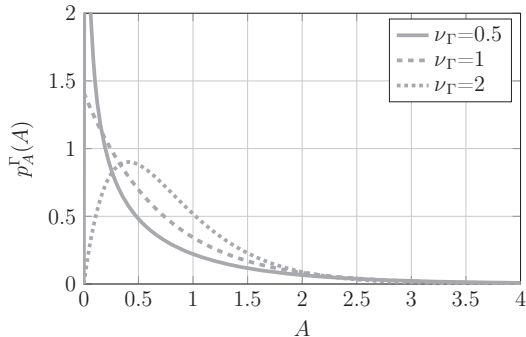


Figure 2.4: PDF of gamma-distributed speech spectral amplitudes with  $\sigma_S^2 = 1$

spectral amplitude model together with a statistically independent and uniformly distributed speech spectral phase assumption.

The PDF of *chi-distributed* speech spectral amplitudes is defined as [Andrianakis and White, 2006], [Hendriks et al., 2009b]

$$p_A^\chi(A) = \frac{2\beta_\chi^{\nu_\chi}}{\Gamma(\nu_\chi)} \cdot A^{2\nu_\chi-1} \cdot e^{-\beta_\chi A^2}, \quad \beta_\chi > 0, \nu_\chi > 0, A \geq 0 \quad (2.22)$$

with index  $\chi$  denoting the chi distribution,  $\Gamma(\cdot)$  being the gamma function [Abramowitz and Stegun, 1972, Chapter 6], and  $\beta_\chi = \nu_\chi/\sigma_S^2$ . Please note that besides  $A$  the omitted indices  $(\ell, k)$  also apply to  $\sigma_S^2$  and  $\beta_\chi$ . Compared to the Rayleigh PDF (2.20), the chi distribution has an additional (frequency-independent and time-invariant) parameter  $\nu_\chi$ . The influence of parameter  $\nu_\chi$  on the shape of the chi PDF is illustrated in Figure 2.3. Please note the following special case: Applying  $\nu_\chi = 1$  to (2.22) results in the Rayleigh PDF (2.20) (cf. Figures 2.2 and 2.3).

Based on histogram measurements, in [Lotter and Vary, 2005] a new parametric PDF is proposed for modeling the speech spectral amplitudes. However, it was later shown in [Erkelens et al., 2007] that Lotter et al. implicitly modeled the speech spectral amplitudes as *gamma distributed* by the new PDF with the proposed parameters. The PDF of the gamma-distributed speech spectral amplitudes is defined as [Andrianakis and White, 2006], [Erkelens et al., 2007]

$$p_A^\Gamma(A) = \frac{\beta_\Gamma^{\nu_\Gamma}}{\Gamma(\nu_\Gamma)} \cdot A^{\nu_\Gamma-1} \cdot e^{-\beta_\Gamma A}, \quad \beta_\Gamma > 0, \nu_\Gamma > 0, A \geq 0 \quad (2.23)$$

with index  $\Gamma$  denoting the gamma distribution and  $\beta_\Gamma = \sqrt{\nu_\Gamma(\nu_\Gamma + 1)}/\sigma_S$ . Please note that besides  $A$ , the omitted indices  $(\ell, k)$  also apply to  $\sigma_S$  and  $\beta_\Gamma$ . Similar to the PDF of the chi distribution (2.22), (2.23) has a (frequency-independent and time-invariant) parameter  $\nu_\Gamma$

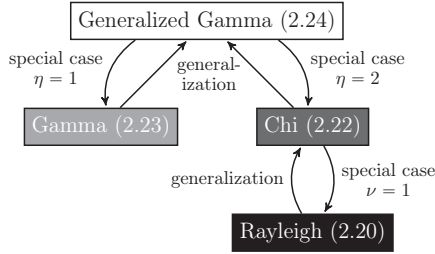


Figure 2.5: Hierarchy of typically employed speech spectral amplitude PDFs  $p_A(A)$

which influences the shape of the gamma PDF (cf. Figure 2.4), offering a nice flexibility in modeling the distribution of speech spectral amplitudes.

As mentioned earlier, Rayleigh-distributed speech spectral amplitudes with statistically independent and uniformly distributed speech spectral phases can be associated with bivariate Gaussian-distributed speech STFT coefficients with Gaussian i. i. d. real and imaginary parts. However, this does not hold for super-Gaussian speech STFT coefficients: Assuming chi- or gamma-distributed speech spectral amplitudes with statistically independent and uniformly distributed speech spectral phases, the real and imaginary parts of the corresponding super-Gaussian speech STFT coefficient will not be statistically independent [Erkelens et al., 2007].

An overview of approaches to PDF parameter identification is given in Appendix B.

### Generalized Prior

In [Dat et al., 2005] and [Erkelens et al., 2007] a generalized probability density, the so-called *generalized gamma* PDF is introduced to classify commonly employed distributions for speech spectral amplitudes. This is a parametric PDF which nicely covers the previously mentioned distributions as special cases, each with an individual parameter set. This generalized gamma PDF is defined as [Erkelens et al., 2007]

$$p_A^{g\Gamma}(A) = \frac{\eta\beta^\nu}{\Gamma(\nu)} \cdot A^{\eta\nu-1} \cdot e^{-\beta A^\eta}, \quad \eta > 0, \beta > 0, \nu > 0, A \geq 0 \quad (2.24)$$

with superscript  $g\Gamma$  denoting the term 'generalized gamma' as well as with  $\eta$ ,  $\beta$ , and  $\nu$  being the parameters of the generalized gamma PDF. Please note that parameter  $\beta$  is a function of the speech spectral variance: From the second moment of (2.24) being  $E\{A^2\} = \int_0^\infty A^2 \cdot p_A^{g\Gamma}(A) dA = E\{|S|^2\} = \sigma_S^2$  results

$$\beta = \left( \frac{\Gamma(2/\eta + \nu)}{\Gamma(\nu)} \cdot \frac{1}{\sigma_S^2} \right)^{\frac{2}{\eta}}. \quad (2.25)$$

Speech Spectral Amplitude PDF (Prior)	Parameters of the Generalized Gamma PDF (2.24)		
	$\eta$	$\beta$	$\nu$
Rayleigh (2.20) (Gaussian Prior)	2	$\beta_R = \frac{1}{\sigma_S^2}$	1
Chi (2.22) (Super-Gaussian Prior)	2	$\beta_\chi = \frac{\nu_\chi}{\sigma_S^2}$	$\nu_\chi$
Gamma (2.23) (Super-Gaussian Prior)	1	$\beta_\Gamma = \frac{\sqrt{\nu_\Gamma(\nu_\Gamma + 1)}}{\sigma_S}$	$\nu_\Gamma$

Table 2.1: Parameter sets of the generalized gamma PDF leading to a Rayleigh, chi, or gamma speech spectral amplitude model

Please note that while  $\eta$  and  $\nu$  are modeled to be frequency-independent and time-invariant, the omitted indices  $(\ell, k)$  apply to  $A$ ,  $\beta$ , and  $\sigma_S^2$ .

A summary of PDF dependencies and the respective parameters can be seen in Figure 2.5 and Table 2.1. The (univariate) generalized gamma distribution covers all previously recapitulated speech spectral amplitude models as special case, i.e., the Rayleigh, the chi, and the gamma distribution. By choosing  $\eta = 1$ , the gamma distribution results with  $\beta = \beta_\Gamma = \sqrt{\nu_\Gamma(\nu_\Gamma + 1)}/\sigma_S$  and the shape parameter  $\nu = \nu_\Gamma$ . The chi or the Rayleigh distribution is obtained by setting  $\eta = 2$ : While the chi distribution results with the shape parameter  $\nu = \nu_\chi$ , the Rayleigh PDF is obtained if  $\nu = \nu_\chi = 1$  is employed. Thus, the chi distribution can be considered as a generalization of the Rayleigh distribution.

Still assuming statistically independent speech spectral amplitudes and phases with the latter being uniformly distributed, the prior  $p_S(S)$  can be modeled by a *bivariate generalized gamma distribution* by applying (2.18) to (2.24), leading to [Fodor and Fingscheidt, 2012a]

$$p_S^{\text{g}\Gamma}(S) = p_\alpha(\alpha) \cdot \frac{1}{|S|} p_A^{\text{g}\Gamma}(|S|) = \frac{1}{2\pi} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} |S|^{\eta\nu-2} e^{-\beta|S|^\eta} \quad (2.26)$$

with  $\eta$ ,  $\beta$ , and  $\nu$  referring to the same parameters as those in the univariate case (2.24). A specific (bivariate) prior with a Rayleigh-, chi-, or gamma-distributed speech spectral amplitude and statistically independent, uniformly distributed speech spectral phase turns out if corresponding parameter sets from Table 2.1 are applied to (2.26). Thus, we will employ the generalized prior in this thesis which covers the speech spectral amplitude models from Table 2.1 and maps them into a (bivariate) prior  $p_S(S = Ae^\alpha)$  which is needed for MMSE estimation in (2.12).

### 2.3.2 Noise Spectral PDF Assumptions

It is a generally employed model in speech enhancement that the speech STFT coefficient  $S$  is superimposed with a statistically independent noise STFT coefficient  $D$  while passing through the acoustic channel, resulting in the noisy speech STFT coefficient  $Y = S + D$  (cf. Figure 2.1). Usually, the noise STFT coefficient  $D$  is modeled by a bivariate Gaussian distribution and its PDF is defined as (cf. (2.19))

$$p_D(D) = \frac{1}{\pi\sigma_D^2} \cdot e^{-\frac{|D|^2}{\sigma_D^2}}, \quad D \in \mathbb{C} \quad (2.27)$$

with  $\sigma_D^2 = E\{|D|^2\}$  being the PSD of the noise STFT coefficient  $D$ . Similar to the assumptions for speech STFT coefficients, the Gaussian assumption is commonly justified by the central limit theorem [Brillinger, 2001]. Consequently, the real and imaginary parts of the noise STFT coefficient are assumed to be univariate Gaussian i.i.d. each with the variance  $\sigma_D^2/2$ . Accordingly, (2.27) is a rotationally symmetric PDF.

The *a priori* knowledge about the acoustic channel is contained in the likelihood  $p_{Y|S}(Y|S)$  which is a bivariate PDF describing the probability density of the complex-valued noisy speech STFT coefficient  $Y$  given a specific speech STFT coefficient  $S$ . The specific speech STFT coefficient  $S$  can be interpreted as a deterministic point in the complex plain and the noisy speech STFT coefficients  $Y$  scatter around this point according to the distribution of the noise STFT coefficient  $D = Y - S$ . Thus, the likelihood actually describes the distribution of the noise and by applying (2.27), the likelihood can be written as [McAulay and Malpass, 1980]

$$p_{Y|S}(Y|S) = \frac{1}{\pi\sigma_D^2} \cdot e^{-\frac{|Y-S|^2}{\sigma_D^2}} \quad (2.28)$$

which is still a rotationally symmetric PDF centered at  $S$ . Please note that besides  $Y$ ,  $S$ , and  $D$ , the omitted indices  $(\ell, k)$  also apply to  $\sigma_D^2$ . Due to the property of rotational symmetry, the Gaussian likelihood (2.28) can be rewritten as (the proof can be found in Appendix C)

$$p_{Y|S}(Y|S = Ae^{j\alpha}) = p_{Y|A,\alpha}(Y = Re^{j\theta}|A, \alpha) = \frac{1}{\pi\sigma_D^2} \cdot e^{-\frac{R^2 + A^2 - 2AR\cos(\theta-\alpha)}{\sigma_D^2}}. \quad (2.29)$$

It is common to use (2.29) for the derivation of specific estimators. Nevertheless, in the following we will employ the likelihood (2.28) instead.

## 2.4 Estimation Domains

In Section 2.2 we introduced a general MMSE estimation formula (2.12) with an arbitrary function  $g(\cdot)$ . This function corresponds to estimation domains of speech enhancement, such

as the STS estimation domain with  $g(X) = X$ , the STSA estimation domain  $g(X) = |X|$ , and the LSA estimation domain  $g(X) = \ln |X|$  with  $\ln(\cdot)$  being the natural logarithm, which will be introduced in the following.

### 2.4.1 Short-Time Spectral (STS) Estimation Domain

Employing the simple function  $g(X) = X$ , the general MMSE estimation formula (2.12) turns out to be the general MMSE STS estimation formula (cf. [Erkelens et al., 2008])

$$\hat{S}_{\text{STS}} = E\{S|Y\} = \frac{\int_{\mathbb{C}} S \cdot p_{Y|S}(Y|S) \cdot p_S(S) dS}{\int_{\mathbb{C}} p_{Y|S}(Y|S) \cdot p_S(S) dS} \quad (2.30)$$

which estimates the complex-valued speech STFT coefficient  $\hat{S}_{\text{STS}}$ . Please note that the well-known Wiener filter [Scalart and Filho, 1996] is a typical MMSE STS estimator. Further MMSE STS estimators can be found in, e. g., [Erkelens et al., 2008].

Please note that employing a polar representation using (2.18), (2.29), and polar integration with  $S = Ae^{j\alpha}$  and  $dS = A d\alpha dA$  [Papoulis and Pillai, 2002], (2.30) can be rewritten as [Erkelens et al., 2008]

$$\hat{S}_{\text{STS}} = \frac{\int_0^{\infty} \int_0^{2\pi} A \cdot e^{j\alpha} \cdot p_{Y|A,\alpha}(Y|A, \alpha) \cdot p_A(A) \cdot p_{\alpha}(\alpha) dA d\alpha}{\int_0^{\infty} \int_0^{2\pi} p_{Y|A,\alpha}(Y|A, \alpha) \cdot p_A(A) \cdot p_{\alpha}(\alpha) dA d\alpha}. \quad (2.31)$$

However, in order to have a consistent notation with Section 2.6, we will use (2.30) in the following.

### 2.4.2 Short-Time Spectral Amplitude (STSA) Estimation Domain

Motivated by the fact that human perception is rather insensitive to the speech phase [Lim and Oppenheim, 1979], many speech enhancement proposals focus on estimating the speech spectral amplitude  $A = |S|$  only, i. e., employ  $g(X) = |X|$  in the general MMSE estimation formula (2.12). The resulting MMSE STSA estimation formula is as follows (cf. [Ephraim and Malah, 1984])

$$\hat{A}_{\text{STSA}} = E\{|S| | Y\} = \frac{\int_{\mathbb{C}} |S| \cdot p_{Y|S}(Y|S) \cdot p_S(S) dS}{\int_{\mathbb{C}} p_{Y|S}(Y|S) \cdot p_S(S) dS}. \quad (2.32)$$

MMSE STSA estimators are very successful which is reflected by their variety and the considerable number of publications on these estimators, e. g., [Ephraim and Malah, 1984],

[Andrianakis and White, 2006], [Erkelens et al., 2007], [Chen and Loizou, 2007], [Fodor and Fingscheidt, 2012a].

Again, employing polar representation using (2.18), (2.29), and polar integration as for (2.31), (2.30) yields [Ephraim and Malah, 1984]

$$\hat{A}_{\text{STSA}} = \frac{\int_0^\infty \int_0^{2\pi} A \cdot p_{Y|A,\alpha}(Y|A, \alpha) \cdot p_A(A) \cdot p_\alpha(\alpha) dA d\alpha}{\int_0^\infty \int_0^{2\pi} p_{Y|A,\alpha}(Y|A, \alpha) \cdot p_A(A) \cdot p_\alpha(\alpha) dA d\alpha} \quad (2.33)$$

which is the well-known MMSE STSA estimation formula. However, we will use (2.32) in the following in order to have a consistent notation with Chapter 2.6.

Employing MMSE STSA estimation, the speech spectral phase  $\alpha$  is out of focus. However, in order to obtain complex-valued STFT coefficients for the signal synthesis step, the estimated speech spectral amplitudes are typically combined with the noisy spectral phase. It is shown in [Ephraim and Malah, 1984] by minimizing the error criterion  $E\{1 - \cos(\alpha - \hat{\alpha})\}$  w. r. t.  $\hat{\alpha}$  that the noisy spectral phase is indeed an optimal speech spectral phase estimate

$$\boxed{\hat{\alpha}_{\text{STSA}} = \vartheta}. \quad (2.34)$$

Please note that this optimization result is independent of (2.32). Interestingly, it is shown in [Erkelens et al., 2008] that (2.34) is also optimal in the MMSE sense by minimizing  $\int_0^{2\pi} (\alpha - \hat{\alpha})^2 \cdot p_{\alpha|Y}(\alpha|Y) d\alpha$  w. r. t.  $\hat{\alpha}$ , under the assumption that the speech and noise STFT coefficients are statistically independent, the speech spectral amplitudes and phases are also statistically independent with the latter being uniformly distributed. Moreover, Gaussian i. i. d. real and imaginary parts of the noise STFT coefficient  $D$  are assumed. Please note that (2.34) holds independently of any assumptions on the speech spectral amplitude PDF  $p_A(A)$ . Finally, the estimated speech STFT coefficient is obtained by  $\hat{S}_{\text{STSA}} = \hat{A}_{\text{STSA}} \cdot e^{j\vartheta}$ .

It is worthwhile to mention that speech spectral amplitude estimation can further be improved by exploiting information carried by the spectral phase for the estimation process, as shown in [Gerkmann and Krawczyk, 2013], [Mowlae and Saeidi, 2013].

### 2.4.3 Short-Time Log-Spectral Amplitude (LSA) Estimation Domain

Motivated by the logarithmic sensitivity of human perception of sound, in [Ephraim and Malah, 1985] the LSA estimation domain is proposed. Applying  $g(X) = \ln|X|$  to the general MMSE estimation formula (2.12), the MMSE LSA estimation formula is obtained

by (cf. [Ephraim and Malah, 1985])

$$\widehat{\ln(|S|)} = \widehat{\ln(A)} = E\{\ln(|S|) | Y\} = \frac{\int_{\mathbb{C}} \ln(|S|) \cdot p_{Y|S}(Y|S) \cdot p_S(S) dS}{\int_{\mathbb{C}} p_{Y|S}(Y|S) \cdot p_S(S) dS}. \quad (2.35)$$

Further publications dealing with MMSE LSA estimation are, e.g., [Hendriks et al., 2009b], [Borgström and Alwan, 2011], [Fodor and Fingscheidt, 2012b], [Fodor and Fingscheidt, 2012d]. Please note that employing polar representation using (2.18), (2.29), and polar integration as for (2.31), (2.35) can be rewritten as [Ephraim and Malah, 1985]

$$\widehat{\ln|S|} = \widehat{\ln(A)} = \frac{\int_0^{\infty} \int_0^{2\pi} \ln(A) \cdot p_{Y|A,\alpha}(Y|A, \alpha) \cdot p_A(A) \cdot p_{\alpha}(\alpha) d\alpha dA}{\int_0^{\infty} \int_0^{2\pi} p_{Y|A,\alpha}(Y|A, \alpha) \cdot p_A(A) \cdot p_{\alpha}(\alpha) d\alpha dA} \quad (2.36)$$

which is widely used in literature. However, we will use (2.35) in the following in order to have a consistent notation with Section 2.6.

Please note that different from the MMSE STSA estimate from (2.32) or (2.33), the MMSE LSA estimate from (2.35) or (2.36) is commonly transformed back from the logarithmic domain into the linear domain. Accordingly, the speech STFT coefficient is then generally computed as

$$\hat{S}_{\text{LSA}} = \hat{A}_{\text{LSA}} \cdot e^{j\hat{\theta}} \approx \exp\left(\widehat{\ln(A)}\right) \cdot e^{j\hat{\theta}} \quad (2.37)$$

where the speech spectral amplitude is approximated by  $\hat{A}_{\text{LSA}} \approx \exp\left(\widehat{\ln(A)}\right)$  and the term  $e^{j\hat{\theta}}$  is the optimal spectral phase estimate according to (2.34).

## 2.5 Synopsis of MMSE Estimation

As outlined before, for MMSE estimation an underlying signal and channel model (cf. Section 2.3) as well as an estimation domain (cf. Section 2.4) are required. The combination of a specific speech prior and a specific likelihood leads then to a specific MMSE estimator in each estimation domain. Assuming a Gaussian likelihood (2.28), the three speech priors from Section 2.3 and the three estimation domains from Section 2.4 lead to nine specific MMSE estimators which will be introduced in this section. First, we will give a generalized estimation formula for each estimation domain employing the bivariate generalized gamma PDF (2.26) as prior, and then, each specific estimator turns out as special case. By this means, the relation among the resulting estimators can easily be shown.

The resulting speech spectral (amplitude) estimates usually feature a multiplicative relationship of the noisy speech spectrum (spectral amplitude) and a spectral gain which is also

called *spectral weighting rule*, typically denoted by  $G$ . Weighting rules are always real-valued and non-negative, if we assume that the speech spectral amplitude and phase are statistically independent, the speech spectral phase is uniformly distributed, and the noise STFT coefficients are bivariate Gaussian distributed (2.28) (circular symmetry of the likelihood) [Erkelen et al., 2008]. Furthermore, spectral weights are time-varying and frequency-dependent, i. e., are a function of the omitted indices  $(\ell, k)$ .

### 2.5.1 MMSE STS Estimation

#### Generalized Prior

Employing the bivariate *generalized gamma* distribution (2.26) for the speech STFT coefficients and a Gaussian likelihood (2.28), the general MMSE STS estimation formula (2.30) turns out to be

$$\hat{S}_{\text{g}\Gamma\text{-STS}} = \frac{\int_{\mathbb{C}} S \cdot \exp\left(\frac{|Y-S|^2}{\sigma_D^2}\right) \cdot |S|^{\eta\nu-2} \cdot \exp(-\beta|S|^\eta) dS}{\int_{\mathbb{C}} \exp\left(\frac{|Y-S|^2}{\sigma_D^2}\right) \cdot |S|^{\eta\nu-2} \cdot \exp(-\beta|S|^\eta) dS} \quad (2.38)$$

with subscript g $\Gamma$ -STS denoting the generalized gamma prior and the MMSE STS estimation domain. After employing polar integration as for (2.31) and integrating w. r. t. the spectral phase  $\alpha$  using [Gradshteyn and Ryzhik, 1965, Eq. (8.431.5)], the MMSE STS estimation formula (2.38) yields

$$\hat{S}_{\text{g}\Gamma\text{-STS}} = \frac{\int_0^\infty A^{\eta\nu} \cdot \exp\left(-\frac{A^2}{\sigma_D^2} - \beta A^\eta\right) \cdot I_1\left(2\frac{AR}{\sigma_D^2}\right) dA}{\underbrace{\int_0^\infty A^{\eta\nu-1} \cdot \exp\left(-\frac{A^2}{\sigma_D^2} - \beta A^\eta\right) \cdot I_0\left(2\frac{AR}{\sigma_D^2}\right) dA}_{\text{Spectral amplitude estimate}}} \cdot \underbrace{e^{j\vartheta}}_{\text{Spectral phase estimate}} \quad (2.39)$$

with  $I_0(\cdot)$  and  $I_1(\cdot)$  being the modified Bessel function of the zeroth and first order [Abramowitz and Stegun, 1972, Chapter 10.2], respectively. Since the integrands of (2.39) are real-valued ( $I_0(\cdot)$  and  $I_1(\cdot)$  are real-valued if their arguments are real and positive [Abramowitz and Stegun, 1972]), both integrals turn out to be real-valued. Therefore, their ratio can be associated with the spectral amplitude estimate  $\hat{A}_{\text{g}\Gamma\text{-STS}}$ .

It is interesting to note that integrating the numerator of (2.38) w. r. t. the spectral phase  $\alpha$  results in the product of the numerator in (2.39) and a spectral phase term  $e^{j\vartheta}$ . Thus, the noisy speech spectral phase  $\vartheta$  turns out to be the optimal speech spectral phase estimate  $\hat{\alpha}_{\text{STS}}$  for STS estimation in general, as the denominator of (2.38) does not influence the spectral phase estimate. Interestingly, this result coincides with the speech phase estimate in STSA estimation, cf. (2.34).

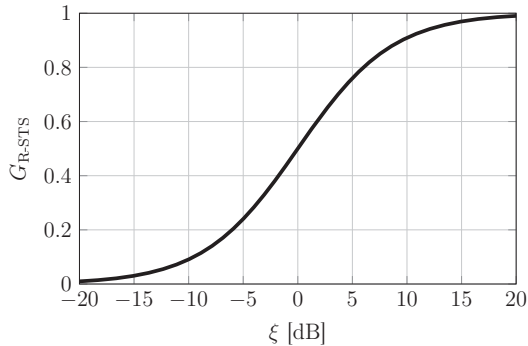


Figure 2.6: MMSE STS weighting rule with an underlying Rayleigh speech spectral amplitude model (Wiener filter) (2.41)

In the following, each of the speech spectral models from Section 2.3.1, i. e., the Rayleigh, chi, and gamma distributions, will be employed to derive specific MMSE STS estimators.

### Gaussian Prior

Assuming the *Rayleigh* distribution (2.20) for the speech spectral amplitudes (Gaussian speech prior), i. e., using the parameters  $\eta = 2$ ,  $\beta = 1/\sigma_S^2$ , and  $\nu = 1$  (cf. Table 2.1), the general MMSE STS formula (2.39) results in

$$\hat{S}_{\text{R-STS}} = \frac{\int_0^\infty A^2 \cdot \exp\left(-A^2 \left[\frac{1}{\sigma_b^2} + \frac{1}{\sigma_s^2}\right]\right) \cdot I_1\left(2\frac{AR}{\sigma_b^2}\right) dA}{\int_0^\infty A \cdot \exp\left(-A^2 \left[\frac{1}{\sigma_b^2} + \frac{1}{\sigma_s^2}\right]\right) \cdot I_0\left(2\frac{AR}{\sigma_b^2}\right) dA} \cdot e^{j\theta} \quad (2.40)$$

with subscript R-STS denoting a Rayleigh-distributed speech spectral amplitude model (Gaussian assumption for the complex speech spectral value) and the MMSE STS estimation domain. By using [Gradshteyn and Ryzhik, 1965, Eqs. (6.631.1), (8.406.3)], (2.40) yields the Wiener filter [Scalart and Filho, 1996]

$$\boxed{\hat{S}_{\text{R-STS}} = G_{\text{R-STS}} \cdot Re^{j\theta} = \frac{\xi}{1 + \xi} \cdot Y} \quad (2.41)$$

with the so-called *a priori* SNR which is defined as [McAulay and Malpass, 1980]

$$\xi = \frac{\sigma_S^2}{\sigma_D^2}. \quad (2.42)$$

Please note that besides of  $\sigma_S^2$  and  $\sigma_D^2$  the omitted indices  $(\ell, k)$  also apply to  $\xi$ . The resulting MMSE speech spectral estimator turns out to be a product of the weighting rule  $G_{\text{R-STS}}$  and

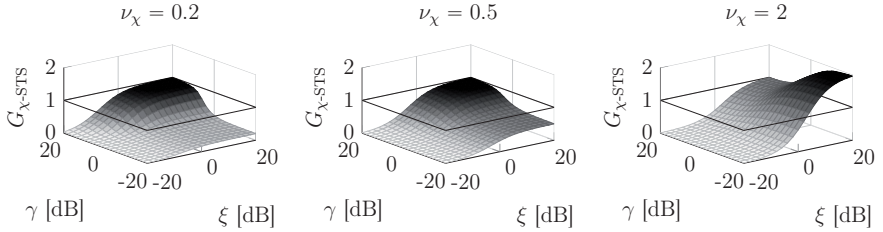


Figure 2.7: MMSE STS weighting rule with an underlying chi speech spectral amplitude model (2.43)

the noisy speech STFT coefficient  $Y$ . The spectral weights of the Wiener filter (2.41) are depicted in Figure 2.6 as a function of the *a priori* SNR  $\xi$ . As can be seen, noisy speech STFT coefficients are penalized by low spectral gains if the *a priori* SNR is low. In contrary, they are less attenuated in case of high SNRs.

### Super-Gaussian Priors

Assuming the *chi* distribution (2.22) for the speech spectral amplitudes, i. e., utilizing the parameters  $\eta = 2$ ,  $\nu = \nu_{\chi}$ , and  $\beta = \nu_{\chi}/\sigma_S^2$  (cf. Table 2.1), the general MMSE STS estimation formula (2.39) can be rewritten by means of [Gradshteyn and Ryzhik, 1965, Eqs. (6.631.1), (8.406.3)] as [Erkelens et al., 2008]

$$\hat{S}_{\chi\text{-STS}} = G_{\chi\text{-STS}} \cdot Y = \frac{\nu_{\chi} \cdot \xi}{\nu_{\chi} + \xi} \cdot \frac{{}_1F_1\left(\nu_{\chi} + 1; 2; \frac{\gamma\xi}{\nu_{\chi} + \xi}\right)}{{}_1F_1\left(\nu_{\chi}; 1; \frac{\gamma\xi}{\nu_{\chi} + \xi}\right)} \cdot Y \quad (2.43)$$

where subscript  $\chi$ -STS denotes the chi-distributed speech spectral amplitude model and the MMSE STS estimation domain and  ${}_1F_1(\cdot)$  is the confluent hypergeometric function [Gradshteyn and Ryzhik, 1965, Chapter 9.21]. In addition to the *a priori* SNR  $\xi$ , the resulting estimator (2.43) is a function of the so-called *a posteriori* SNR  $\gamma = |Y|^2/\sigma_D^2$  [McAulay and Malpass, 1980]. Please note that besides  $Y$  and  $\sigma_D^2$ , the omitted indices  $(\ell, k)$  also apply to the *a posteriori* SNR  $\gamma$ .

Again, the MMSE STS estimator (2.43) is a product of the weighting rule  $G_{\chi\text{-STS}}$  and the noisy speech STFT coefficient  $Y$ . The resulting spectral weights are plotted in Figure 2.7. Please note that for  $\nu_{\chi} = 2$ , spectral weights larger than one occur in case of small *a posteriori* SNRs and large *a priori* SNRs. We will observe this phenomenon for several estimators which can be explained by the phase relationship of the speech and noise STFT coefficients: If the speech PSD is large enough (reflected by a large *a priori* SNR), but a destructive phase relationship produces a small observed noisy spectral amplitude (indicated

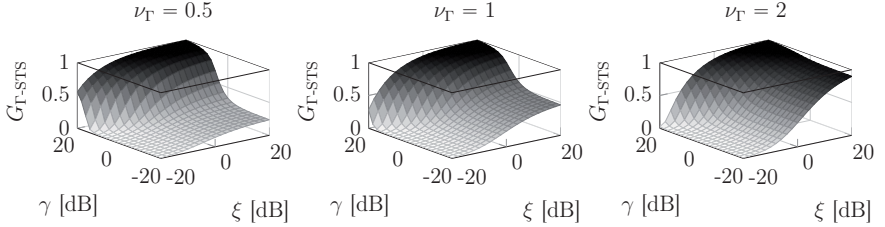


Figure 2.8: MMSE STS weighting rule with an underlying gamma speech spectral amplitude model (2.45)

by a small *a posteriori* SNR), large spectral weights amplify the attenuated speech spectral amplitudes. Please note the following special case: Applying  $\nu_\chi = 1$  to (2.43), the Wiener filter (2.41) results (cf. [Gradshteyn and Ryzhik, 1965, Eq. (9.215.1)]).

Assuming the *gamma* distribution for the speech spectral amplitudes (2.23), i. e., employing the parameters  $\eta = 1$ ,  $\nu = \nu_\chi$ , and  $\beta = \sqrt{\nu_\Gamma(\nu_\Gamma + 1)}/\sigma_S$  (cf. Table 2.1), the general MMSE STS estimation formula (2.39) turns out to be (cf. [Erkelens et al., 2008, Eq. (18)])

$$\hat{S}_{\Gamma\text{-STS}} = \frac{\int_0^\infty A^{\nu_\Gamma} \cdot \exp\left(-\frac{1}{\sigma_D^2} A^2 - \frac{\sqrt{\nu_\Gamma(\nu_\Gamma+1)}}{\sigma_S} A\right) \cdot I_1\left(2\frac{R}{\sigma_D} A\right) dA}{\int_0^\infty A^{\nu_\Gamma-1} \cdot \exp\left(-\frac{1}{\sigma_D^2} A^2 - \frac{\sqrt{\nu_\Gamma(\nu_\Gamma+1)}}{\sigma_S} A\right) \cdot I_0\left(2\frac{R}{\sigma_D} A\right) dA} \cdot e^{j\theta} \quad (2.44)$$

with subscript  $\Gamma$ -STS denoting the gamma-distributed speech spectral amplitude model and the MMSE STS estimation domain. Unfortunately, the integrals in (2.44) do not have closed-form solutions. Therefore, in [Erkelens et al., 2008] some approximations are given and two different closed-form solutions are derived for low and high SNRs. Different from this we will employ numerical methods to solve the integrals. Assuming that—similar to (2.41) and (2.43)—the estimator (2.44) is a product of a (real-valued) weighting rule and the noisy speech STFT coefficients, (2.44) can be rewritten using variable substitution with  $A = g \cdot R$  and  $dA = dg \cdot R$  as

$$\hat{S}_{\Gamma\text{-STS}} = G_{\Gamma\text{-STS}} \cdot Y = \frac{\int_0^\infty g^{\nu_\Gamma} \cdot \exp\left(-\gamma g^2 - \sqrt{\nu_\Gamma(\nu_\Gamma+1)}\sqrt{\frac{\gamma}{\xi}}g\right) \cdot I_1(2\gamma g) dg}{\int_0^\infty g^{\nu_\Gamma-1} \cdot \exp\left(-\gamma g^2 - \sqrt{\nu_\Gamma(\nu_\Gamma+1)}\sqrt{\frac{\gamma}{\xi}}g\right) \cdot I_0(2\gamma g) dg} \cdot R e^{j\theta}. \quad (2.45)$$

By this means, the integrands can be obtained as a function of the *a priori* SNR  $\xi$  and the *a posteriori* SNR  $\gamma$ . These integrals can be calculated numerically, e. g., by using the Gauss-Kronrod quadrature [Brass and Petras, 2011]. The resulting approximated weighting rule is depicted in Figure 2.8. Please note that although the exemplary spectral weights in Figure 2.8 are not larger than one, values  $G_{\Gamma\text{-STS}} > 1$  are generally possible.

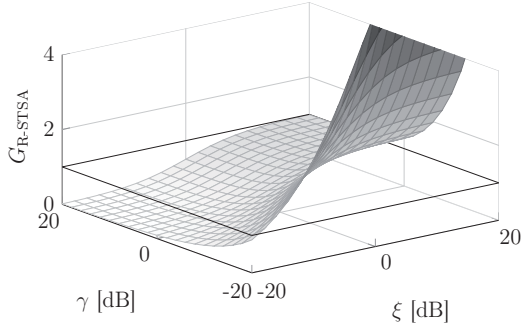


Figure 2.9: MMSE STSA weighting rule with an underlying Rayleigh speech spectral amplitude model (2.48)

## 2.5.2 MMSE STSA Estimation

### Generalized Prior

Employing the bivariate *generalized gamma* distribution (2.26) as speech prior and a Gaussian likelihood (2.28) for the general MMSE STSA estimation formula (2.32) leads to

$$\widehat{|S|}_{\text{g}\Gamma\text{-STSA}} = \widehat{A}_{\text{g}\Gamma\text{-STSA}} = \frac{\int_{\mathbb{C}} |S| \cdot \exp\left(\frac{|Y-S|^2}{\sigma_b^2}\right) \cdot |S|^{\eta\nu-2} \cdot \exp(-\beta|S|^\eta) dS}{\int_{\mathbb{C}} \exp\left(\frac{|Y-S|^2}{\sigma_b^2}\right) \cdot |S|^{\eta\nu-2} \cdot \exp(-\beta|S|^\eta) dS} \quad (2.46)$$

where subscript g $\Gamma$ -STSA denotes the generalized gamma prior and the STSA estimation domain. Utilizing polar integration as for (2.31) and integrating w. r. t. the spectral phase  $\alpha$  using [Gradshteyn and Ryzhik, 1965, Eq. (8.431.5)], (2.46) turns out to be [Fodor and Fingscheidt, 2012a]

$$\widehat{A}_{\text{g}\Gamma\text{-STSA}} = \frac{\int_0^\infty A^{\eta\nu} \cdot \exp\left(-\frac{A^2}{\sigma_b^2} - \beta A^\eta\right) \cdot I_0\left(2\frac{AR}{\sigma_b^2}\right) dA}{\int_0^\infty A^{\eta\nu-1} \cdot \exp\left(-\frac{A^2}{\sigma_b^2} - \beta A^\eta\right) \cdot I_0\left(2\frac{AR}{\sigma_b^2}\right) dA}. \quad (2.47)$$

Note that different from the MMSE STS estimation formula (2.39), (2.47) uses a Bessel function of zeroth order in the nominator and naturally does not have a spectral phase term (due to the different estimation domain). As a similarity, its integrals are real-valued, just as those in (2.39).

In the following, each of the speech spectral models from Section 2.3.1 (Rayleigh, chi, and gamma) will again be reemployed to derive specific MMSE STSA estimators.

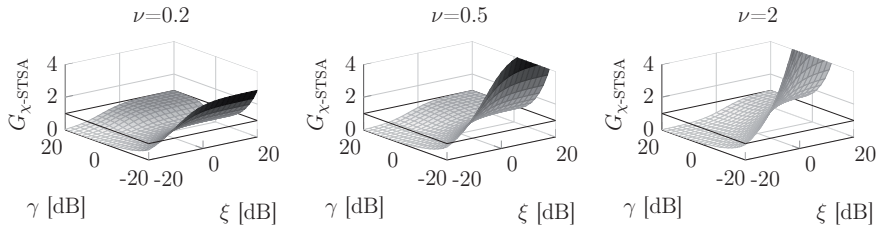


Figure 2.10: MMSE STSA weighting rule with an underlying chi speech spectral amplitude model (2.49)

### Gaussian Prior

Assuming the *Rayleigh* distribution for the speech spectral amplitudes (2.20) (Gaussian assumption for the speech), i.e., using the parameters  $\eta = 2$ ,  $\beta = 1/\sigma_S^2$ , and  $\nu = 1$  (cf. Table 2.1), the general MMSE LSA estimation formula (2.47) can be rewritten by means of [Gradshteyn and Ryzhik, 1965, Eqs. (6.631.1), (8.406.3)] as [Ephraim and Malah, 1984]

$$\hat{A}_{\text{R-STSA}} = G_{\text{R-STSA}} \cdot R = \Gamma(1.5) \sqrt{\frac{\xi}{\gamma(1+\xi)}} {}_1F_1\left(-0.5; 1; -\frac{\gamma\xi}{1+\xi}\right) \cdot R \quad (2.48)$$

with subscript R-STSA denoting the Rayleigh speech spectral amplitude model and the STSA estimation domain. The resulting weighting rule  $G_{\text{R-STSA}}$  is depicted in Figure 2.9.

### Super-Gaussian Priors

Assuming a *chi*-distributed speech spectral amplitude model (2.22), i.e., utilizing the parameters  $\eta = 2$ ,  $\nu = \nu_\chi$ , and  $\beta = \nu_\chi/\sigma_S^2$  (cf. Table 2.1) as well as using [Gradshteyn and Ryzhik, 1965, Eqs. (6.631.1), (8.406.3)], the general MMSE LSA estimation formula (2.47) yields [Andrianakis and White, 2006], [Erkelens et al., 2007]

$$\hat{A}_{\chi\text{-STSA}} = G_{\chi\text{-STSA}} \cdot R = \frac{\Gamma(\nu_\chi + 0.5)}{\Gamma(\nu_\chi)} \sqrt{\frac{\xi}{\gamma(\nu_\chi + \xi)}} \frac{{}_1F_1\left(\nu_\chi + 0.5; 1; \frac{\gamma\xi}{\nu_\chi + \xi}\right)}{{}_1F_1\left(\nu_\chi; 1; \frac{\gamma\xi}{\nu_\chi + \xi}\right)} \cdot R \quad (2.49)$$

where subscript  $\chi$ -STSA denotes the chi-distributed speech spectral amplitude model and the STSA estimation domain. The resulting weighting rule  $G_{\chi\text{-STSA}}$  is illustrated in Figure 2.10. Please note that employing  $\nu_\chi = 1$ , (2.49) yields the MMSE STSA estimator (2.48) with an underlying Gaussian speech prior as special case.

Assuming a *gamma*-distributed speech spectral amplitude model (2.23), i.e., incorporating the parameters  $\eta = 1$ ,  $\nu = \nu_\Gamma$ , and  $\beta = \sqrt{\nu_\Gamma(\nu_\Gamma + 1)}/\sigma_S$  (cf. Table 2.1), the general

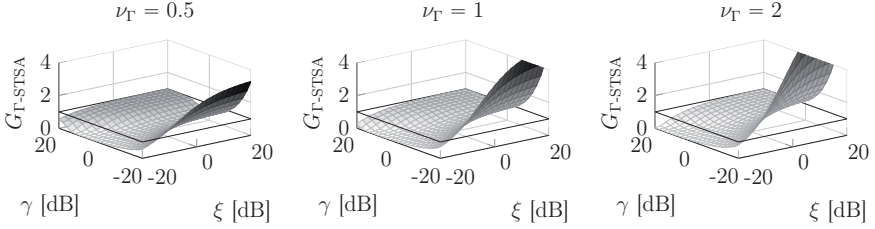


Figure 2.11: MMSE STSA weighting rule with an underlying gamma speech spectral amplitude model (2.51)

MMSE LSA estimation formula (2.47) can be reformulated as (cf. [Andrianakis and White, 2006, Eqs. (7) and (8)], [Erkelens et al., 2007, Eq. (13)])

$$\hat{A}_{\Gamma\text{-STSA}} = \frac{\int_0^{\infty} A^{\nu_{\Gamma}} \cdot \exp\left(-\frac{A^2}{\sigma_D^2} - \frac{\sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)}}{\sigma_S} A\right) \cdot I_0\left(2\frac{AR}{\sigma_D^2}\right) dA}{\int_0^{\infty} A^{\nu_{\Gamma}-1} \cdot \exp\left(-\frac{A^2}{\sigma_D^2} - \frac{\sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)}}{\sigma_S} A\right) \cdot I_0\left(2\frac{AR}{\sigma_D^2}\right) dA} \quad (2.50)$$

with subscript  $\Gamma$ -STSA denoting the employed gamma speech spectral amplitude model and the STSA estimation domain. Similar to (2.44), the integrals in (2.50) do not have closed-form solutions, therefore, in [Erkelens et al., 2007] some approximations are applied in order to yield an analytical solution. Due to the corresponding approximations, two separate speech spectral amplitude estimators are obtained one for low and one for high SNRs. However, we will employ numerical methods to solve these integrals just as in the STS estimation case. First, we reformulate (2.50) by variable substitution as for (2.45) leading to [Fodor and Fingscheidt, 2012a]

$$\hat{A}_{\Gamma\text{-STSA}} = G_{\Gamma\text{-STSA}} \cdot R = \frac{\int_0^{\infty} g^{\nu_{\Gamma}} \cdot \exp\left(-\gamma g^2 - \sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)}\sqrt{\frac{\gamma}{\xi}}g\right) \cdot I_0(2\gamma g) dg}{\int_0^{\infty} g^{\nu_{\Gamma}-1} \cdot \exp\left(-\gamma g^2 - \sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)}\sqrt{\frac{\gamma}{\xi}}g\right) \cdot I_0(2\gamma g) dg} \cdot R. \quad (2.51)$$

By this means, the integrands can be obtained as a function of the *a priori* SNR  $\xi$  and the *a posteriori* SNR  $\gamma$ . Again, solving these integrals by the Gauss-Kronrod quadrature [Brass and Petras, 2011] results in the approximated weighting rule in Figure 2.11.

## 2.5.3 MMSE LSA Estimation

### Generalized Prior

Employing the bivariate *generalized gamma* PDF (2.26) as speech prior and a Gaussian likelihood (2.28), the general MMSE LSA estimation formula (2.35) yields

$$\widehat{\ln(|S|)}_{\text{g}\Gamma\text{-LSA}} = \widehat{\ln(A)}_{\text{g}\Gamma\text{-LSA}} = \frac{\int_{\mathbb{C}} \ln(|S|) \cdot \exp\left(\frac{|Y-S|^2}{\sigma_D^2}\right) \cdot |S|^{\eta\nu-2} \cdot \exp(-\beta|S|^\eta) dS}{\int_{\mathbb{C}} \exp\left(\frac{|Y-S|^2}{\sigma_D^2}\right) \cdot |S|^{\eta\nu-2} \cdot \exp(-\beta|S|^\eta) dS} \quad (2.52)$$

where subscript g $\Gamma$ -LSA denotes the generalized gamma prior and the LSA estimation domain. Utilizing polar integration as for (2.31) and integrating w. r. t. the spectral phase  $\alpha$  using [Gradshteyn and Ryzhik, 1965, Eq. (8.431.5)], (2.52) turns out to be

$$\widehat{\ln(A)}_{\text{g}\Gamma\text{-LSA}} = \frac{\int_0^\infty \ln(A) \cdot A^{\eta\nu-1} \cdot \exp\left(-\frac{A^2}{\sigma_D^2} - \beta A^\eta\right) \cdot I_0\left(2\frac{AR}{\sigma_D^2}\right) dA}{\int_0^\infty A^{\eta\nu-1} \cdot \exp\left(-\frac{A^2}{\sigma_D^2} - \beta A^\eta\right) \cdot I_0\left(2\frac{AR}{\sigma_D^2}\right) dA} \quad (2.53)$$

Similar to the STS and STSA estimation domains (cf. (2.39) and (2.47)), its integrals are real-valued, thus, (2.53) is also real-valued. Using variable substitution as for (2.45), (2.53) can be rewritten as [Fodor and Fingscheidt, 2012d, Eqs. (6) and (7)]

$$\widehat{\ln(A)}_{\text{g}\Gamma\text{-LSA}} = \ln(R) + \frac{\int_0^\infty \ln(g) \cdot g^{\eta\nu-1} \cdot \exp\left(-\frac{g^2 R^2}{\sigma_D^2} - \beta g^\eta R^\eta\right) \cdot I_0\left(2\frac{gR^2}{\sigma_D^2}\right) dg}{\int_0^\infty g^{\eta\nu-1} \cdot \exp\left(-\frac{g^2 R^2}{\sigma_D^2} - \beta g^\eta R^\eta\right) \cdot I_0\left(2\frac{gR^2}{\sigma_D^2}\right) dg} \quad (2.54)$$

Furthermore, the introduced MMSE LSA estimators will turn out to the general form

$$\widehat{\ln(A)}_{\text{LSA}} = \ln(G_{\text{LSA}}) + \ln(R) = \ln(G_{\text{LSA}} \cdot R) \quad (2.55)$$

with  $G_{\text{LSA}}$  being an LSA spectral weighting rule. Therefore, instead of the approximation  $\widehat{A}_{\text{LSA}} \approx \exp(\widehat{\ln(A)})$  in (2.37) we can write

$$\widehat{A}_{\text{LSA}} = G_{\text{LSA}} \cdot R \quad (2.56)$$

which generally holds for MMSE LSA estimation. Accordingly, (2.54) can be rewritten as

$$\widehat{A}_{\text{g}\Gamma\text{-LSA}} = \exp\left(\frac{\int_0^\infty \ln(g) \cdot g^{\eta\nu-1} \cdot \exp\left(-g^2 \frac{R^2}{\sigma_D^2} - g^\eta \beta R^\eta\right) \cdot I_0\left(g \frac{2R^2}{\sigma_D^2}\right) dg}{\int_0^\infty g^{\eta\nu-1} \cdot \exp\left(-g^2 \frac{R^2}{\sigma_D^2} - g^\eta \beta R^\eta\right) \cdot I_0\left(g \frac{2R^2}{\sigma_D^2}\right) dg}\right) \cdot R. \quad (2.57)$$

In the following, each of the speech spectral models from Section 2.3.1 (Rayleigh, chi, gamma) will again be employed to derive specific MMSE STSA estimators.

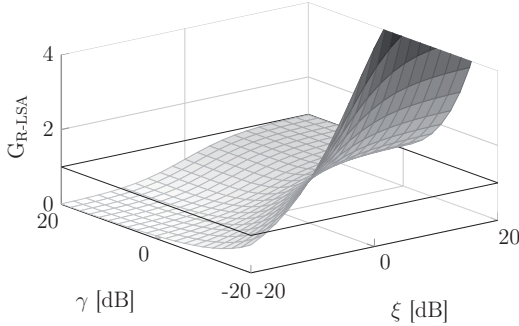


Figure 2.12: MMSE LSA weighting rule with an underlying Rayleigh speech spectral amplitude model (2.58)

### Gaussian Prior

Assuming a *Rayleigh*-distributed speech spectral amplitude model (2.20) (Gaussian assumption for the speech), i. e., using the parameters  $\eta = 2$ ,  $\beta = 1/\sigma_S^2$ , and  $\nu = 1$  (cf. Table 2.1), (2.53) can be solved by means of the moment generating function [Papoulis and Pillai, 2002, Chapter 5-5] and the result turns out to be [Ephraim and Malah, 1985]

$$\hat{A}_{\text{R-LSA}} = G_{\text{R-LSA}} \cdot R = \frac{\xi}{1 + \xi} \exp\left(\frac{1}{2} \int_v^\infty \frac{e^{-t}}{t} dt\right) \cdot R \quad (2.58)$$

with subscript R-LSA denoting the Rayleigh speech spectral amplitude model and the LSA estimation domain as well as with  $v = \gamma \cdot \xi / (1 + \xi)$ . The resulting weighting rule  $G_{\text{R-LSA}}$  is depicted in Figure 2.12.

### Super-Gaussian Priors

Assuming a *chi*-distributed speech spectral amplitude model (2.22), i. e., utilizing the parameters  $\eta = 2$ ,  $\nu = \nu_\chi$ , and  $\beta = \nu_\chi / \sigma_S^2$  (cf. Table 2.1), (2.53) can also be obtained in closed form by means of the moment generating function [Hendriks et al., 2009b]

$$\hat{A}_{\text{X-LSA}} = G_{\text{X-LSA}} \cdot R = \sqrt{\frac{\xi}{(\xi + \nu_\chi)\gamma}} \exp\left(\frac{\psi(\nu_\chi)}{2} + \frac{T(\nu_\chi, \gamma, \xi)}{{}_1F_1(\nu_\chi; 1; \frac{\gamma\xi}{\nu_\chi + \xi})}\right) \cdot R \quad (2.59)$$

with

$$T(\nu_\chi, \gamma, \xi) = \sum_{t=0}^{\infty} \frac{\Gamma(\nu_\chi + t) [\psi(\nu_\chi + t) - \psi(\nu_\chi)]}{2\Gamma(\nu_\chi)} \left(\frac{\gamma\xi}{\nu_\chi + \xi}\right)^t \frac{1}{(t!)^2}, \quad (2.60)$$

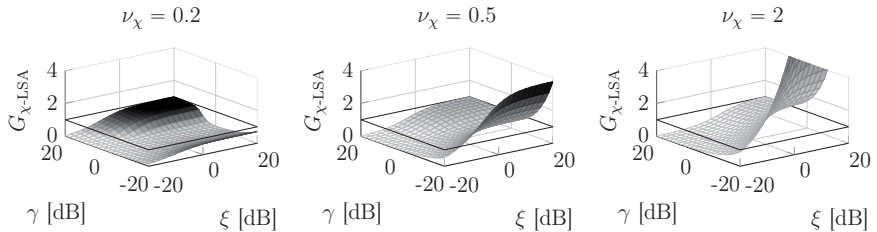


Figure 2.13: MMSE LSA weighting rule with an underlying chi speech spectral amplitude model (2.59)

$\psi(\cdot)$  being the psi function [Abramowitz and Stegun, 1972, Chapter 6], and subscript  $\chi$ -LSA denoting the chi-distributed speech spectral amplitude model and the LSA estimation domain. Please note that employing  $\nu_{\chi} = 1$ , (2.59) yields the MMSE STSA estimator with an underlying Rayleigh speech spectral amplitude model (2.58) as special case. The function  $T(\nu_{\chi}, \gamma, \xi)$  is calculated by means of an infinite series which is commonly truncated for practical implementations to an upper limit. In [Hendriks et al., 2009b] an upper limit of  $1.5 \cdot 10^4$  iterations is found to be sufficient.

The spectral weights  $G_{\chi\text{-LSA}}$  can also be solved by quadrature: Using (2.57) and the PDF parameters of the chi distribution from Table 2.1, the estimator (2.59) can be also be written as

$$\widehat{A}_{\chi\text{-LSA}} = G_{\chi\text{-LSA}} \cdot R = \exp \left( \frac{\int_0^{\infty} \ln(g) \cdot g^{2\nu_{\chi}-1} \cdot \exp \left( -g^2 \left[ \gamma + \nu_{\chi} \frac{\gamma}{\xi} \right] \right) \cdot I_0(2\gamma g) dg}{\int_0^{\infty} g^{2\nu_{\chi}-1} \cdot \exp \left( -g^2 \left[ \gamma + \nu_{\chi} \frac{\gamma}{\xi} \right] \right) \cdot I_0(2\gamma g) dg} \right) \cdot R. \quad (2.61)$$

Solving these integrals by the Gauss-Kronrod quadrature [Brass and Petras, 2011] results in the approximated weighting rule  $G_{\chi\text{-LSA}}$  in Figure 2.13.

Assuming a *gamma*-distributed speech spectral amplitude model (2.23), i. e., employing the parameters  $\eta = 1$ ,  $\nu = \nu_{\Gamma}$ , and  $\beta = \sqrt{\nu_{\Gamma}(\nu_{\Gamma} + 1)}/\sigma_S$  (cf. Table 2.1), (2.53) turns out to be

$$\widehat{\ln(A)}_{\Gamma\text{-LSA}} = \frac{\int_0^{\infty} \ln(A) \cdot A^{\nu-1} \cdot \exp \left( -\frac{A^2}{\sigma_B^2} - \frac{\sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)}}{\sigma_S} A \right) \cdot I_0 \left( 2\frac{AR}{\sigma_B^2} \right) dA}{\int_0^{\infty} A^{\nu-1} \cdot \exp \left( -\frac{A^2}{\sigma_B^2} - \frac{\sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)}}{\sigma_S} A \right) \cdot I_0 \left( 2\frac{AR}{\sigma_B^2} \right) dA} \quad (2.62)$$

with subscript  $\Gamma$ -LSA denoting the employed gamma speech spectral amplitude model and the LSA estimation domain. Similar to the MMSE STS and MMSE STSA equivalents (cf. (2.44) and (2.50)), the integrals in (2.62) do not have closed-form solutions. Therefore, in [Borgström and Alwan, 2011] some approximations are applied in order to obtain an (approximated) analytical solution. However, just as in the STS and STSA estimation domains,

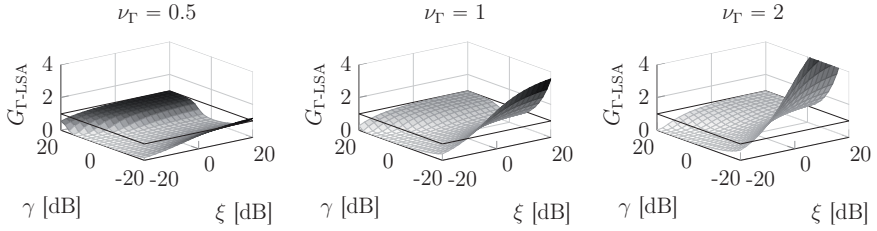


Figure 2.14: MMSE LSA weighting rule with an underlying gamma speech spectral amplitude model (2.63)

we utilize numerical methods to solve these integrals. Using the generalized MMSE LSA estimation formula (2.57) and the PDF parameters of the gamma distribution from Table 2.1, the MMSE LSA estimator assuming gamma-distributed speech spectral amplitudes can be obtained by (cf. [Fodor and Fingscheidt, 2012d, Eq. (7)])

$$\begin{aligned}
 \hat{A}_{\Gamma\text{-LSA}} &= G_{\Gamma\text{-LSA}} \cdot R \\
 &= \exp \left( \frac{\int_0^{\infty} \ln(g) \cdot g^{\nu_{\Gamma}-1} \cdot \exp \left( -\gamma g^2 - \sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)} \sqrt{\frac{\gamma}{\xi}} g \right) \cdot I_0(2\gamma g) \, dg}{\int_0^{\infty} g^{\nu_{\Gamma}-1} \cdot \exp \left( -\gamma g^2 - \sqrt{\nu_{\Gamma}(\nu_{\Gamma}+1)} \sqrt{\frac{\gamma}{\xi}} g \right) \cdot I_0(2\gamma g) \, dg} \right) \cdot R.
 \end{aligned} \tag{2.63}$$

Solving these integrals by means of the Gauss-Kronrod quadrature [Brass and Petras, 2011] results in the approximated weighting rule  $G_{\Gamma\text{-LSA}}$  depicted in Figure 2.14.

An overview of MMSE estimators recapitulated in this section is given in Table 2.2. It can be concluded that the hierarchy of speech spectral amplitude PDFs in Figure 2.5 is also true for corresponding MMSE estimators. This is due to the fact that MMSE estimators based on a specific speech spectral amplitude PDF inherit its hierarchy properties.

Speech Spectral Amplitude PDF (Prior)	Estimation Domain	
	STS (2.30)	STSA (2.32)
Generalized Gamma (2.24) (Generalized Prior) [Erkelens et al., 2007]	g $\Gamma$ -STS (2.39) [new]	g $\Gamma$ -STSA (2.47) [new] [Fodor and Fingscheidt, 2012a] [Fodor and Fingscheidt, 2012d]
Rayleigh (2.20) (Gaussian Prior) [Ephraim and Malah, 1984]	R-STS (Wiener Filter) (2.41) [Scalart and Filho, 1996]	R-STSA (2.48) [Ephraim and Malah, 1984] [Ephraim and Malah, 1985]
Chi (2.22) (Super-Gaussian Prior) [Andrianakis and White, 2006]	$\chi$ -STS (2.43) [Erkelens et al., 2008]	$\chi$ -STSA (2.49) [Andrianakis and White, 2006] [Hendriks et al., 2009b]
Gamma (2.23) (Super-Gaussian Prior) [Andrianakis and White, 2006]	$\Gamma$ -STS (2.45) [Erkelens et al., 2008]	$\Gamma$ -STSA (2.51) [Erkelens et al., 2007] [Borgström and Alwan, 2011]

Table 2.2: Overview of MMSE estimators resulting as special case of (2.12) employing the generalized or a specific prior from Section 2.3 and a specific estimation domain from Section 2.4

## 2.6 MMSE Estimation

### Under Speech Presence Uncertainty (SPU)

The general MMSE estimation formula (2.12) and the resulting specific estimators in Section 2.5 (summarized in Table 2.2) assume that *speech is always present*. However, this is not fulfilled for natural speech as uttered by human beings, e. g., during speech pauses or in frequency bins located between spectral harmonics of voiced speech sounds which generally have extreme low energy. Therefore, it was shown in several publications, e. g., [McAulay and Malpass, 1980], [Ephraim and Malah, 1984], [Azirani et al., 1996], [Cohen and Berdugo, 2001], [Gerkmann et al., 2008], that speech enhancement approaches can successfully be enhanced by taking *speech presence uncertainty* (SPU) into account.

Accordingly, extending the Bayesian estimation formula (2.9) in terms of total probability by the hypotheses  $H_0$  and  $H_1$  denoting *speech absence* and *speech presence*, respectively, results in the general Bayesian estimation formula under SPU (cf. [Middleton and Esposito, 1968])

$$\begin{aligned} \hat{S} &= \arg \min_{\hat{S}} \int_{\mathcal{C}} C(S, \hat{S}) \cdot p_{S|Y}(S|Y) \, dS \\ &= \arg \min_{\hat{S}} \int_{\mathcal{C}} C(S, \hat{S}) \cdot \overbrace{\sum_{H \in \{H_0, H_1\}} P(H|Y) \cdot p_{S|Y,H}(S|Y, H)} \, dS \quad (2.64) \\ &= \arg \min_{\hat{S}} \int_{\mathcal{C}} C(S, \hat{S}) \cdot \left[ P(H_0|Y) \cdot p_{S|Y, H_0}(S|Y, H_0) + P(H_1|Y) \cdot p_{S|Y, H_1}(S|Y, H_1) \right] dS \end{aligned}$$

with  $P(H_1|Y)$ ,  $P(H_0|Y) = 1 - P(H_1|Y)$ ,  $p_{S|Y, H_1}(S|Y, H_1)$ , and  $p_{S|Y, H_0}(S|Y, H_0)$  being the *a posteriori* speech presence probability (SPP), the *a posteriori* speech absence probability (SAP), the posterior under speech presence  $H_1$  and absence  $H_0$ , respectively. Applying the MMSE cost function (2.10) to (2.64), leads to the general MMSE estimation formula under SPU

$$\widehat{g(S)} = P(H_0|Y) \cdot E\{g(S)|Y, H_0\} + P(H_1|Y) \cdot E\{g(S)|Y, H_1\} \quad (2.65)$$

with  $E\{g(S)|Y, H_0\}$  and  $E\{g(S)|Y, H_1\}$  being the MMSE estimate for speech  $\widehat{g(S)}$  under the hypothesis of speech absence  $H_0$  and speech presence  $H_1$ , respectively. As can be seen, the MMSE estimation formula under SPU is a sum of two products: Each product consists of an MMSE estimate  $E\{g(S)|Y, H\}$  and an *a posteriori* SPP or SAP  $P(H|Y)$  with  $H \in \{H_0, H_1\}$ . The latter is also called *soft weight*. Please note that since the MMSE estimation formula (2.12) assumes permanent speech presence, it corresponds to  $E\{g(S)|Y, H_1\}$  in (2.65).

In the following, we will focus first on the soft weights, then we will take a look at the MMSE estimates  $E\{g(S)|Y, H_0\}$  and  $E\{g(S)|Y, H_1\}$  as well as the final MMSE estimators under SPU. In Section 2.6.1, a very common description of the soft weights is presented, while

in Section 2.6.2 a more sophisticated soft weight computation with averaging and fixed prior parameters is introduced. Finally, the application of the soft weights into the final MMSE estimation under SPU in different estimation domains will be presented in Section 2.6.3.

### 2.6.1 *A Posteriori* SPP Estimation with Adapted Prior Parameters

In terms of total probability, the relationship between the *a posteriori* SPP and the *a posteriori* SAP is  $P(H_1|Y) + P(H_0|Y) = 1$ . Furthermore, the *a posteriori* SPP can be written as [Middleton and Esposito, 1968], [Ephraim and Malah, 1984]

$$P(H_1|Y) = \frac{\Lambda}{1 + \Lambda} \quad (2.66)$$

with the so-called generalized likelihood ratio (GLR)

$$\Lambda = \frac{P(H_1)}{P(H_0)} \cdot \frac{p_{Y|H_1}(Y|H_1)}{p_{Y|H_0}(Y|H_0)} \quad (2.67)$$

where  $P(H_1)$ ,  $P(H_0) = 1 - P(H_1)$ ,  $p_{Y|H_1}(Y|H_1)$ , and  $p_{Y|H_0}(Y|H_0)$  are the *a priori* SPP, the *a priori* SAP, the PDF of the noisy speech assuming speech presence and absence, respectively. The PDFs  $p_{Y|H_0}(Y|H_0)$  and  $p_{Y|H_1}(Y|H_1)$  are also called the likelihood of speech absence and speech presence, respectively. The *a priori* SPP and SAP are usually modeled as constant quantities with, e. g.,  $P(H_1) = 0.5$  [McAulay and Malpass, 1980], or  $P(H_1) = 0.8$  [Ephraim and Malah, 1984], but there are also approaches based on adaptive tracking of the *a priori* SAP [Malah et al., 1999], [Cohen and Berdugo, 2001]. In this thesis, we will employ  $P(H_0) = P(H_1) = 0.5$ .

The conditional PDF  $p_{Y|H_0}(Y|H_0)$  models the probability density of the noisy speech STFT coefficient in absence of speech. That means that only the noise is present, i. e.,  $Y = D$ , therefore,

$$p_{Y|H_0}(Y|H_0) \equiv p_D(Y) \quad (2.68)$$

with  $p_D(\cdot)$  being, e. g., (2.27). Under hypothesis  $H_1$ , however, the speech STFT coefficient  $S$  is non-zero, thus, the likelihood of speech presence is defined as

$$p_{Y|H_1}(Y|H_1) \equiv p_Y(Y = S + D) = \int_{\mathcal{C}} p_{Y|S}(Y|S) \cdot p_S(S) dS. \quad (2.69)$$

It turns out that the desired PDF  $p_{Y|H_1}(Y|H_1)$  is identical to the evidence in the general MMSE estimation formula (2.12) and can be calculated by marginalizing the product of the likelihood  $p_{Y|S}(Y|S)$  and the prior  $p_S(S)$ . In the following, we will recapitulate common *a posteriori* SPP estimators  $P(H_1|Y)$ .

Analog to the PDF assumptions of the speech spectral amplitude estimator (2.48), in [Ephraim and Malah, 1984] a Gaussian speech spectral model (2.19) (Rayleigh-distributed

speech spectral amplitude assumption) and a Gaussian likelihood (2.28) are proposed for calculating the likelihood of speech absence (2.68) and speech presence (2.69), respectively. Utilizing [Gradshteyn and Ryzhik, 1965, Eq. (6.631.1)] for determining the former, the likelihoods finally yield

$$\begin{aligned} p_{Y|H_1}(Y|H_1)|_R &= \frac{1}{\pi\sigma_D^2} \cdot \exp\left(\frac{|Y|^2}{\sigma_D^2}\right) \cdot \frac{1}{1+\xi} \cdot \exp\left(\frac{|Y|^2}{\sigma_D^2} \frac{\xi}{1+\xi}\right), \\ p_{Y|H_0}(Y|H_0)|_R &= \frac{1}{\pi\sigma_D^2} \cdot \exp\left(\frac{|Y|^2}{\sigma_D^2}\right) \end{aligned} \quad (2.70)$$

with subscript R denoting the Rayleigh-distributed speech spectral amplitude model. Substituting the likelihoods (2.70) into the GLR (2.67) results in [Ephraim and Malah, 1984]

$$\Lambda_R = \frac{P(H_1)}{P(H_0)} \cdot \frac{1}{1+\xi} \cdot \exp\left(\frac{\gamma\xi}{1+\xi}\right). \quad (2.71)$$

Please note that (2.71) and the corresponding *a posteriori* SPP  $P(H_1|Y)|_R = \Lambda_R/(1+\Lambda_R)$  are a function of the *a priori* and the *a posteriori* SNR  $\xi$  and  $\gamma$ , respectively, just as the spectral weighting rules in Section 2.5.

Arguing that the speech STFT coefficients rather follow a super-Gaussian distribution than a Gaussian, in [Breithaupt and Martin, 2011] the *chi* distribution (2.22) is utilized for modeling the speech spectral amplitudes. The corresponding GLR is calculated as [Breithaupt and Martin, 2011]

$$\Lambda_\chi = \frac{P(H_1)}{P(H_0)} \cdot \left(\frac{\nu_\chi}{\nu_\chi + \xi}\right)^{\nu_\chi} \cdot {}_1F_1\left(\nu_\chi; 1; \frac{\gamma\xi}{\nu_\chi + \xi}\right) \quad (2.72)$$

with subscript  $\chi$  denoting the chi-distributed speech spectral amplitude assumption. The corresponding *a posteriori* SPP is computed straightforwardly by (2.66).

A gamma-distributed speech spectral amplitude model in conjunction with SPU estimation will be proposed in Section 4.2.

## 2.6.2 *A Posteriori* SPP Estimation

### with Averaging and Fixed Prior Parameters

The estimators in the previous section yield *a posteriori* SPP values close to one relatively robust during speech presence. In absence of speech, however, they typically output the *a priori* SPP which is usually not close to zero (cf.  $P(H_1|Y)$  in (2.66) tends to the *a priori* SPP  $P(H_1)$  instead of zero, if the *a priori* SNR in (2.71) or (2.72) approaches zero). To overcome this issue, an adaptive tracking of the *a priori* SPP is proposed in [Cohen and Berdugo, 2001]. Different from this adaptive solution, an approach with fixed *a priori* SPP

and fixed *a priori* SNR for SPU estimation is introduced in [Gerkmann et al., 2008] arguing that these quantities should be independent of the observations and, thus, reflect true *a priori* knowledge. Furthermore, it is shown that the use of a fixed *a priori* SPP and a fixed *a priori* SNR can enhance SPU estimation and can outperform approaches with adapted prior parameters by achieving *a posteriori* SPP estimates close to zero when speech is absent.

Since the *a posteriori* SPP estimator (2.66) is a function of the noisy observations (cf. (2.71) or (2.72)), the random fluctuations of the observations may result in estimation outliers which may be perceived as musical noise [Gerkmann et al., 2008]. In order to improve estimation robustness and reduce estimation outliers, in [Gerkmann et al., 2008] averaged observations are employed for SPU estimation. As can be seen in (2.71) or (2.72), the *a posteriori* SPP estimator (2.66) is not a function of the spectral phase of the noisy speech STFT coefficient  $Y$ . Therefore, assuming that hypothesis  $H_1$  is independent of the spectral phase of the noisy speech STFT coefficient  $Y$ , the *a posteriori* SPP can also be written as a function of  $\gamma$ , i. e.,  $P(H_1|Y) = P(H_1|\gamma)$ . Employing averaged *a posteriori* SNR values  $\bar{\gamma}$  for SPU estimation, the corresponding GLR (2.67) and the corresponding likelihoods (2.68) and (2.69) have to be determined. The averaged *a posteriori* SNR  $\bar{\gamma}$  is simply obtained by calculating the moving average of  $\gamma$ . However, the choice of the averaging window size is crucial: The larger the window, the more random fluctuations of  $\gamma$  can be reduced, but simultaneously the more speech distortion occurs since averaging strongly affects the fine spectral structure of speech. As a tradeoff, two averaging processes of different window sizes, i. e., a small local window and a large global window, can be applied and combined [Cohen and Berdugo, 2001]. Accordingly, the averaging is carried out as follows [Gerkmann et al., 2008]

$$\bar{\gamma}_{\Theta, \ell}(k) = \frac{1}{|\mathbb{L}_{\Theta}| \cdot |\mathbb{K}_{\Theta}|} \cdot \sum_{\lambda_{\Theta} \in \mathbb{L}_{\Theta}} \sum_{\kappa_{\Theta} \in \mathbb{K}_{\Theta}} \gamma_{\ell=\lambda_{\Theta}}(k = \kappa_{\Theta}) \quad (2.73)$$

with  $\Theta$  standing for either local or global,  $\mathbb{L}_{\Theta} = \{\ell - \Delta\ell_{\Theta}, \ell - \Delta\ell_{\Theta} + 1, \dots, \ell - 1, \ell\}$  and  $\mathbb{K}_{\Theta} = \{k - \Delta k_{\Theta}, k - \Delta k_{\Theta} + 1, \dots, k - 1, k, k + 1, \dots, k + \Delta k_{\Theta} - 1, k + \Delta k_{\Theta}\}$  being a set of frames and a set of frequency bins within the corresponding averaging window, respectively, as well as with  $|\mathbb{L}_{\Theta}|$  and  $|\mathbb{K}_{\Theta}|$  being the number of these frames and frequency bins, respectively. Besides the current frame  $\ell$ , the previous  $\Delta\ell_{\Theta}$  frames are also considered for averaging, thus, each averaging window covers  $|\mathbb{L}_{\Theta}| = \Delta\ell_{\Theta} + 1$  frames. Furthermore, each window utilizes  $|\mathbb{K}_{\Theta}| = 2 \cdot \Delta k_{\Theta} + 1$  frequency bins for averaging, comprising the current frequency bin  $k$ ,  $\Delta k_{\Theta}$  lower frequency bins, and  $\Delta k_{\Theta}$  upper frequency bins. By this means, each averaging window has the size  $|\mathbb{K}_{\Theta}| \cdot |\mathbb{L}_{\Theta}|$ . Please note that at low and high frequency bins, where the shape of the averaging windows would be reduced due to side effects, the missing observations are replaced by mirroring available ones using the smallest and largest frequency bins, respectively, as symmetry axes. Moreover, at  $k = 0$  and  $k = L/2 + 1$  no averaging is applied.

The corresponding local and global *a posteriori* SPPs driven by the averaged *a posteriori* SNRs  $\bar{\gamma}_\Theta$  are multiplicatively combined to a final *a posteriori* SPP estimate [Gerkmann et al., 2008] (cf. also [Cohen and Berdugo, 2001] and [Sørensen and Andersen, 2005])<sup>3</sup>

$$P(H_1|\gamma) = P(H_1|\bar{\gamma}_{\text{local}}) \cdot P(H_1|\bar{\gamma}_{\text{global}}). \quad (2.74)$$

Please note that (2.74) estimates speech presence  $H_1$  in a robust way: The *a posteriori* SPP  $P(H_1|\gamma)$  only achieves values close to one, if both the local and the global *a posteriori* SPPs are close to one. The local and global *a posteriori* SPPs are calculated as (cf. (2.66))

$$P(H_1|\bar{\gamma}_\Theta) = \frac{\bar{\Lambda}_\Theta}{1 + \bar{\Lambda}_\Theta} \quad (2.75)$$

with the GLR for the averaged observations (cf. (2.67))

$$\bar{\Lambda}_\Theta = \frac{P(H_1)}{P(H_0)} \cdot \frac{p_{\bar{\gamma}_\Theta|H_1}(\bar{\gamma}_\Theta|H_1)}{p_{\bar{\gamma}_\Theta|H_0}(\bar{\gamma}_\Theta|H_0)} \quad (2.76)$$

where  $p_{\bar{\gamma}_\Theta|H_1}(\bar{\gamma}_\Theta|H_1)$  and  $p_{\bar{\gamma}_\Theta|H_0}(\bar{\gamma}_\Theta|H_0)$  are the likelihood of speech presence and the likelihood of speech absence, respectively, both for averaged observations. These PDFs can be determined similar to those in (2.67) and the derivation steps will briefly be described in the following.

Starting with hypothesis  $H_1$ , the noisy speech STFT coefficient  $Y$  turns out to be Gaussian distributed, if statistically independent Gaussian-distributed speech and noise STFT coefficients are assumed. Accordingly,  $|Y|$  is Rayleigh distributed (cf. (2.20)), and  $|Y|^2$  is exponential distributed [Papoulis and Pillai, 2002]. The *a posteriori* SNR (being only a scaled version of  $|Y|^2$ ) is also exponential distributed. The sum of exponential-distributed random variables (averaged  $\gamma$  values) can be modeled by the gamma distribution [Papoulis and Pillai, 2002]. Thus, after derivation, the PDF of averaged *a posteriori* SNR values under hypothesis  $H_1$  is obtained as (cf. (2.23) and [Gerkmann et al., 2008, Eq. (6)])

$$p_{\bar{\gamma}_\Theta|H_1}^{\text{R}}(\bar{\gamma}_\Theta|H_1) = \left( \frac{\mu_\Theta}{1 + \xi} \right)^{\mu_\Theta} \cdot \frac{\bar{\gamma}_\Theta^{\mu_\Theta - 1}}{\Gamma(\mu_\Theta)} \cdot \exp\left( -\frac{\mu_\Theta \bar{\gamma}_\Theta}{1 + \xi} \right) \quad (2.77)$$

with superscript R denoting the Rayleigh assumption for the speech spectral amplitudes (Gaussian speech prior) and with the shape parameter  $\mu_\Theta$ . Please note that if the *a posteriori* SNRs are uncorrelated the shape parameter yields  $\mu_\Theta = |\mathbb{L}_\Theta| \cdot |\mathbb{K}_\Theta|$ . Furthermore, if no averaging is applied  $\mu_\Theta = 1$  results.

<sup>3</sup>Please note that in a strict mathematical sense, (2.74) should be written as  $P(H_1|\gamma) = \frac{1}{C} \cdot P(H_1|\bar{\gamma}_{\text{local}}) \cdot P(H_1|\bar{\gamma}_{\text{global}})$  with  $C$  such that  $P(H_1|\gamma) + P(H_0|\gamma) = 1$ . Using (2.74), we implicitly assume that  $P(H_0|\gamma) = 1 - P(H_1|\bar{\gamma}_{\text{local}}) \cdot P(H_1|\bar{\gamma}_{\text{global}})$  resulting in  $C = 1$ . Please note that we obtain then  $P(H_0|\gamma) = 1 - [1 - P(H_0|\bar{\gamma}_{\text{local}})] \cdot [1 - P(H_0|\bar{\gamma}_{\text{global}})] = P(H_0|\bar{\gamma}_{\text{local}}) + P(H_0|\bar{\gamma}_{\text{global}}) - P(H_0|\bar{\gamma}_{\text{local}}) \cdot P(H_0|\bar{\gamma}_{\text{global}}) \neq P(H_0|\bar{\gamma}_{\text{local}}) \cdot P(H_0|\bar{\gamma}_{\text{global}})$ .

$\Theta$	$\Delta k_\Theta$	$\Delta \ell_\Theta$	$ \mathbb{K}_\Theta  \cdot  \mathbb{L}_\Theta $	$\mu_\Theta$	$\Xi_\Theta^R$	$P(H_1)$
local	1	2	9	5.4	8 dB	0.5
global	8	2	51	25.7	3 dB	0.5

Table 2.3: Parameters of the *a posteriori* SNR averaging framework with a Gaussian speech prior (cf. [Gerkmann et al., 2008]) assuming a frame length of 32 ms, 50% frame overlap, and a Hann window

In speech absence  $\xi = 0$  holds and (2.77) reduces to (cf. [Gerkmann et al., 2008, Eq. (4)])

$$p_{\tilde{\gamma}_\Theta|H_0}^R(\tilde{\gamma}_\Theta|H_0) = \mu_\Theta^{\mu_\Theta} \cdot \frac{\tilde{\gamma}_\Theta^{\mu_\Theta-1}}{\Gamma(\mu_\Theta)} \cdot \exp(-\mu_\Theta \tilde{\gamma}_\Theta) \quad (2.78)$$

with  $\mu_\Theta$  being the same parameter as in (2.77).

Finally, using the likelihood of speech presence (2.77) and absence (2.78), the corresponding GLR can be calculated by (2.76) resulting in (cf. [Gerkmann et al., 2008, Eq. (7)])

$$\bar{\Lambda}_\Theta^R = \frac{P(H_1)}{P(H_0)} \cdot \left( \frac{1}{1+\xi} \right)^{\mu_\Theta} \cdot \exp\left( \frac{\mu_\Theta \xi}{1+\xi} \cdot \tilde{\gamma}_\Theta \right) \quad (2.79)$$

with superscript R denoting the Rayleigh assumption for the speech spectral amplitudes (Gaussian speech prior). Please note that if no averaging is applied, i.e.,  $\mu_\Theta = 1$ , (2.79) reduces to the GLR (2.71) assuming also Rayleigh-distributed speech spectral amplitudes.

Although the averaging approach can nicely reduce estimation outliers, the *a posteriori* SPP still does not approach zero at low *a priori* SNRs [Gerkmann et al., 2008]: The GLR (2.79) still tends to the value of  $P(H_1)/P(H_0)$  and, thus, the *a posteriori* SPP delivers a value close to  $P(H_1)$  if  $\xi \rightarrow 0$ . To overcome this issue, it is proposed in [Gerkmann et al., 2008] to use a fixed *a priori* SPP and a fixed *a priori* SNR for SPU estimation as summarized in Table 2.3 together with the parameters for a *a posteriori* SNR averaging. Please note that the fixed *a priori* SNR  $\Xi_\Theta^R$  is used now in (2.79) instead of the adapted one  $\xi$ .

The derivation of likelihood functions for averaged observations assuming a super-Gaussian speech model is mathematically challenging. Nevertheless, a new *a posteriori* SPP estimator using averaged observations and a chi-distributed speech spectral amplitude model in conjunction with fixed prior parameters will be proposed in Section 5.

So far, we have introduced the soft weights assuming different spectral assumptions for the speech prior as well as adapted and fixed prior parameters for MMSE speech enhancement under SPU. In the following, we will take the estimation domains STS, STSA, and LSA (cf. Chapter 2.4) into account, employing  $g(X) = X$ ,  $g(X) = |X|$ , and  $g(X) = \ln |X|$  for (2.65), respectively.

### 2.6.3 Estimation Domains

#### MMSE STS Estimation Under SPU

Employing  $g(X) = X$  for the general MMSE STS estimation formula under SPU (2.65) and considering that the expectation  $E\{S|Y, H_0\}$  is zero under hypothesis  $H_0$  (in absence of speech), the general MMSE STS estimator under SPU yields

$$\boxed{\widehat{S}_{\text{STS-SPU}} = P(H_1|Y) \cdot E\{S|Y, H_1\} = P(H_1|Y) \cdot G_{\text{STS}} \cdot Y} \quad (2.80)$$

with subscript STS-SPU denoting MMSE STS estimation under SPU. In (2.80) it is eligible to compute  $E\{S|Y, H_1\}$  by (2.30), because a permanent speech presence has already been assumed for (2.30). It turns out that a specific MMSE STS estimator from Section 2.5.1 with the weighting rule  $G_{\text{STS}}$  and an *a posteriori* SPP estimator  $P(H_1|Y)$  are required for MMSE STS estimation under SPU, both with the same underlying signal PDF assumptions.

An MMSE STS estimator under SPU with an underlying Gaussian speech prior, i. e., a combination of the common MMSE STS estimator (2.41) and the *a posteriori* SPP estimator (2.66) with the GLR (2.71) is proposed in [Azirani et al., 1996]. To the best of our knowledge, approaches based on super-Gaussian speech priors have not been published yet in literature and will be proposed in Section 4.2.1.

#### MMSE STSA Estimation Under SPU

Employing  $g(X) = |X|$  for the general MMSE STSA estimation formula under SPU (2.65) and considering that  $E\{|S||Y, H_0\}$  is zero under hypothesis  $H_0$  (in absence of speech), the MMSE STSA estimator under SPU is computed as [McAulay and Malpass, 1980, Eq. (22)]

$$\boxed{\widehat{S}_{\text{STSA-SPU}} = \widehat{A}_{\text{STSA-SPU}} = P(H_1|Y) \cdot E\{A|Y, H_1\} = P(H_1|Y) \cdot G_{\text{STSA}} \cdot R} \quad (2.81)$$

with subscript STSA-SPU denoting MMSE STSA estimation under SPU and with  $E\{A|Y, H_1\}$  being (2.32), due to the assumption of permanent speech presence.

An MMSE STSA estimator under SPU with an underlying Rayleigh-distributed speech spectral amplitude assumption, i. e., a combination of the common MMSE STSA estimator (2.48) and the *a posteriori* SPP estimator (2.66) with the GLR (2.71), is proposed in [Ephraim and Malah, 1984]. A further PDF-consistent proposal with an underlying chi-distributed speech spectral amplitude model, i. e., a combination of the common MMSE STSA estimator (2.49) and the *a posteriori* SPP estimator (2.66) with the GLR (2.72), is introduced in [Breithaupt and Martin, 2011]. To complete the picture, a new proposal for MMSE STSA estimation under SPU with an underlying gamma-distributed speech spectral amplitude model will be introduced in Chapter 4 [Fodor and Fingscheidt, 2012a].

### MMSE LSA Estimation Under SPU

Employing  $g(X) = \ln|X|$ , the general MMSE estimation formula (2.65) turns out to be

$$\widehat{\ln|S|} = \widehat{\ln(A)} = P(H_0|Y) \cdot E\{\ln(A)|Y, H_0\} + P(H_1|Y) \cdot E\{\ln(A)|Y, H_1\}. \quad (2.82)$$

Using the approximation in (2.37), (2.82) can be rewritten as

$$\begin{aligned} \widehat{S}|_{\text{LSA-SPU}} &= \hat{A}_{\text{LSA-SPU}} \\ &\approx \exp(\widehat{\ln(A)}) = \exp\left(P(H_0|Y) \cdot E\{\ln(A)|Y, H_0\} + P(H_1|Y) \cdot E\{\ln(A)|Y, H_1\}\right) \end{aligned} \quad (2.83)$$

with subscript LSA-SPU denoting MMSE LSA estimation under SPU. Unfortunately, the MMSE estimate for speech absence  $E\{\ln(A)|Y, H_0\}$  tends to minus infinity, since the speech spectral amplitude  $A$  approaches zero under hypothesis  $H_0$ . Consequently, if  $P(H_0|Y)$  is larger than zero, the speech spectral amplitude estimate  $\hat{A}_{\text{LSA}}$  in (2.83) tends to zero. Several approaches are proposed to overcome this issue which will be recapitulated in the following.

**Nonlinear MMSE LSA Estimation** In this approach proposed in [Ephraim and Malah, 1985] the term  $E\{\ln(A)|Y, H_0\}$  is implicitly assumed to be zero. Therefore, (2.83) reduces to  $\exp(P(H_1|Y) \cdot E\{\ln(A)|Y, H_1\})$  which can be rewritten as  $\exp(P(H_1|Y) \cdot \ln(G_{\text{LSA}} \cdot R))$  according to (2.55). After rearranging this expression, the MMSE LSA estimation formula under SPU results in [Ephraim and Malah, 1985]

$$\hat{A}_{\text{NL-LSA}} = \exp\left(P(H_1|Y) \cdot E\{\ln(A)|Y, H_1\}\right) = (G_{\text{LSA}} \cdot R)^{P(H_1|Y)} \quad (2.84)$$

with subscript NL-LSA denoting the nonlinear MMSE LSA estimator under SPU. As can be seen, the speech spectral amplitude estimate  $\hat{A}_{\text{NL-LSA}}$  turns out to be a nonlinear function of the noisy speech spectral amplitude  $R$ .

It is worthwhile to mention that in [Ephraim and Malah, 1985] the nonlinear MMSE LSA estimator under SPU was not found to achieve convincing results due to a low-pass effect on the enhanced signal. In [Fodor and Fingscheidt, 2012b], however, this low-pass effect could not be observed, but a strong noise attenuation in conjunction with a slight degradation of the speech quality is reported. This different observation may be produced due to the use of a different, modern noise PSD estimator in [Fodor and Fingscheidt, 2012b].

**Multiplicatively Modified MMSE LSA Estimation** Since the nonlinear MMSE LSA estimator under SPU was reported to be not worth to use in [Ephraim and Malah, 1985], the multiplicatively modified MMSE LSA estimator under SPU was proposed in [Malah et al., 1999]. Following the concept of the other estimation domains (cf. (2.80) and (2.81)), this estimator is forced to be a linear function of the noisy speech spectral amplitude  $R$  as

$$\hat{A}_{\text{MM-LSA}} = P(H_1|Y) \cdot \exp(E\{\ln(A)|H_1\}) = P(H_1|Y) \cdot G_{\text{LSA}} \cdot R \quad (2.85)$$

with subscript MM-LSA denoting multiplicatively modified MMSE LSA estimation under SPU. Please note that according to (2.64), this modification is not optimal in the MMSE sense.

**Optimally Modified MMSE LSA Estimation** To overcome the issue that the multiplicatively modified MMSE LSA estimator is not optimal in the MMSE sense, optimally modified (OM) MMSE LSA estimation was introduced in [Cohen and Berdugo, 2001]. It provides an optimal MMSE estimate and, simultaneously, a nice solution for the conceptual issue in speech absence. Contrary to the previous approaches, optimally modified MMSE LSA estimation assumes that  $E\{\ln(A)|Y, H_0\} = \ln(G_0 \cdot R)$  (cf. (2.55)) is not zero. Allowing for a small constant spectral weight  $G_0$  in absence of speech  $H_0$  results in a low, naturally sounding residual noise level. Accordingly, applying  $E\{\ln(A)|Y, H_1\} = \ln(G_{\text{LSA}} \cdot R)$  and  $E\{\ln(A)|Y, H_0\} = \ln(G_0 \cdot R)$  to the general MMSE estimation formula under SPU (2.65) and using (2.55) leads to [Cohen and Berdugo, 2001]

$$\hat{A}_{\text{OM-LSA}} = \left( G_{\text{LSA}}^{P(H_1|Y)} \cdot G_0^{P(H_0|Y)} \right) \cdot R \quad (2.86)$$

with subscript OM-LSA denoting optimally modified MMSE LSA estimation under SPU. It turns out that the speech spectral amplitude estimate  $\hat{A}_{\text{OM-LSA}}$  is a linear function of the noisy speech spectral amplitude  $R$ . Similar to the previous MMSE STS estimator under SPU (2.80) and MMSE STSA estimator under SPU (2.81), the noisy speech spectral amplitude  $R$  is weighted by a total spectral gain consisting of a weighting rule and a soft weight. Different from the previous estimators, the relationship between  $G_{\text{LSA}}$  and  $P(H_1|Y)$  is not multiplicative but nonlinear. Moreover, the total spectral gain also contains a new multiplicative term  $G_0^{P(H_0|Y)}$  where the exponent is generally calculated as  $P(H_0|Y) = 1 - P(H_1|Y)$ .

Optimally modified MMSE LSA estimators under SPU are employed in, e.g., [Cohen and Berdugo, 2001] using a Gaussian speech prior. New, super-Gaussian variants with a chi-distributed and a gamma-distributed [Fodor and Fingscheidt, 2012d] speech spectral amplitude model will be introduced in Chapter 4 (cf. Table 4.2).

## 2.7 Estimation of Noise Power, *A Priori* SNR, and *A Posteriori* SNR

So far, we have introduced various MMSE estimators for speech enhancement both with and without taking SPU into account. All estimators have in common that they are driven by the *a priori* SNR  $\xi_\ell(k)$  and/or by the *a posteriori* SNR  $\gamma_\ell(k)$ , both being a function of the noise PSD  $\sigma_{D_\ell}^2(k) = E\{|D_\ell(k)|^2\}$ . Although the noise process  $D_\ell(k)$  is assumed to be quasi-stationary, the noise PSD is naturally time-varying in practice. Therefore, it is necessary to

track the noise PSD during processing time. Typical tracking approaches are introduced in Section 2.7.1.

The *a posteriori* SNR is a function of the instantaneous value of the observations  $Y_\ell(k)$  and the noise PSD which is usually estimated by a noise PSD estimator. While the calculation of the *a posteriori* SNR is rather simple, the computation of the *a priori* SNR is not straightforward: Being a function of the speech PSD  $\sigma_{s,\ell}^2(k) = E\{|S_\ell(k)|^2\}$ , it is a strongly time-varying quantity and, therefore, its estimation is challenging. Typical *a priori* and *a posteriori* SNR estimators are recapitulated in Section 2.7.2.

## 2.7.1 Noise Power Estimation

As mentioned above, the noise PSD is essential in speech enhancement. Therefore, the accuracy of the noise PSD estimate is crucial: Underestimation may result in annoying residual noise, while overestimation may lead to a degradation of speech intelligibility due to speech distortions. The estimation of the noise PSD is a nontrivial task, since it can be estimated by merely the observations  $Y_\ell(k)$  and some *a priori* knowledge, but without any useful reference. In the following, typical noise PSD estimation approaches will briefly be summarized.

### Noise PSD Estimation Based on Voice Activity Detection (VAD)

A very simple noise PSD estimation approach is based on the fact that the observations  $Y_\ell(k)$  are dominated by the noise spectrum  $D_\ell(k)$  in speech pauses. This approach measures the PSD of  $Y_\ell(k)$  first, e. g., by recursive averaging of the noisy periodogram. Then, a voice activity detection (VAD) is used to control whenever the noise power estimate should be updated by the measured power of the observations  $Y_\ell(k)$ . Such an approach will fully be described in Section 3.3.2. Further VAD-based noise PSD estimators are proposed in, e. g., [McAulay and Malpass, 1980], [Compernelle, 1989]. A disadvantage of VAD-based approaches is that they are only able to track the noise power in speech pauses.

### Noise PSD Estimation Based on Minimum Statistics (MS)

Noise PSD estimation based on minimum statistics (MS) is proposed in [Martin, 1994] and [Martin, 2001]. Based on a statistical independence assumption between the speech and the noise STFT coefficients, this approach makes use of the fact that even during speech activity, the power of the noisy speech STFT coefficients  $Y_\ell(k)$  often decreases to the level of the noise power. Therefore, the noise PSD can be estimated by tracking the minimum of the noisy speech PSD which can be measured by calculating the recursive average of the periodogram

of the observations  $|Y_\ell(k)|^2$ . Therefore, no VAD is required and—as a clear advantage over VAD-based approaches—the noise PSD estimate can even be updated during speech activity.

Since there are different requirements regarding the smoothing of the noisy periodogram  $|Y_\ell(k)|^2$  in speech absence and speech presence, the MS approach employs an adaptive, frequency-dependent smoothing parameter. Furthermore, since the minimum of smoothed  $|Y_\ell(k)|^2$  values is generally smaller than (or equal to) their desired mean  $E\{|Y_\ell(k)|^2|H_0\}$ , tracking the minimum of power spectral values leads to an underestimation of the noise PSD. Therefore, a bias compensation is utilized to obtain an unbiased estimate  $\widehat{\sigma}_{D,\ell}^2(k)$ .

### Noise PSD Estimation

#### Based on (Improved) Minima Controlled Recursive Averaging ((I)MCRA)

Minima controlled recursive averaging (MCRA) is proposed in [Cohen and Berdugo, 2001]. Similar to the MS approach, MCRA noise PSD estimation is based on recursive averaging of the periodogram  $|Y_\ell(k)|^2$  with an adaptive smoothing factor. Different from the MS approach, this smoothing factor is driven by the SPP which is controlled by the smoothed periodogram  $|Y_\ell(k)|^2$ . Analog to MS, the MCRA approach is also able to update its noise PSD estimate during speech activity.

The improved minima controlled recursive averaging (IMCRA) is proposed in [Cohen, 2003]. In this paper the minimum tracking during speech activity and the SPP estimation of the MCRA approach is further enhanced and a new bias compensation factor is derived. According to [Cohen, 2003], the strengths of the IMCRA approach arise in presence of non-stationary noise or in low input SNR conditions.

#### MMSE Noise PSD Estimation Based on Speech Presence Probability

This approach is published in [Gerkmann and Hendriks, 2012] and it is based on an MMSE noise PSD estimator. Similar to the (I)MCRA approach, the MMSE noise PSD estimation employs also a recursive average with a smoothing factor controlled by SPPs. Different from the (I)MCRA solution, the SPP is calculated based on the *a posteriori* SPP using fixed prior parameters as in Section 2.6.2, but without any averaging of the *a posteriori* SNR. A considerable advantage of this approach over the previous ones are low estimation delay and low computational complexity.

## 2.7.2 *A Priori* and *A Posteriori* SNR Estimation

Once the noise PSD has been estimated, the *a priori* SNR and the *a posteriori* SNR can be computed.

### *A Posteriori* SNR Estimation

The *a posteriori* SNR is defined as a function of the noisy speech STFT coefficient  $Y_\ell(k)$  and the noise PSD  $\sigma_{D,\ell}^2(k)$ , and it is usually estimated by means of the estimated noise PSD as

$$\hat{\gamma}_\ell(k) = \frac{|Y_\ell(k)|^2}{\widehat{\sigma}_{D,\ell}^2(k)}. \quad (2.87)$$

### Decision Directed *A Priori* SNR Estimation

For sure the most prevalent *a priori* SNR estimator is the decision directed (DD) approach [Ephraim and Malah, 1984]

$$\hat{\xi}_\ell(k) = (1 - \beta_{DD}) \cdot \max\{\hat{\gamma}_\ell(k) - 1, 0\} + \beta_{DD} \cdot \frac{|\widehat{S}_{\ell-1}(k)|^2}{\widehat{\sigma}_{D,\ell-1}^2(k)} \quad (2.88)$$

with  $\beta_{DD} \in [0, 1]$  being the smoothing parameter. Please note that the *a priori* SNR estimate (2.88) is a weighted sum of the maximum likelihood (ML) *a priori* SNR estimate  $\hat{\gamma}_\ell(k) - 1$  and an instantaneous SNR of the last frame based on estimated values  $|\widehat{S}_{\ell-1}(k)|^2 / \widehat{\sigma}_{D,\ell-1}^2(k)$ . The ML *a priori* SNR estimation assumes a Gaussian-distributed noisy speech STFT coefficient (with statistically independent, Gaussian-distributed speech and noise spectral components) and it maximizes the resulting exponentially-distributed likelihood [Ephraim and Malah, 1984, Eq. (20)]

$$p(\gamma|\xi) = \frac{1}{1 + \xi} \cdot \exp\left(-\frac{\gamma}{1 + \xi}\right) \quad (2.89)$$

w. r. t.  $\xi$ , resulting in  $\hat{\xi}_{ML,\ell}(k) = \gamma_\ell(k) - 1$ . The lower bound of the ML estimate enforced by  $\max\{\cdot, 0\}$  ensures a plausible, non-negative SNR estimate. Since the *a posteriori* SNR is of a random nature, the ML *a priori* SNR estimate  $\gamma_\ell(k) - 1$  suffers from its fluctuations which can result in unnatural sounding musical noise [Cappe, 1994]. Therefore, the ML estimate is combined with the fraction  $|\widehat{S}_{\ell-1}(k)|^2 / \widehat{\sigma}_{D,\ell-1}^2(k)$  which can be interpreted as a very conservative *a priori* SNR estimate with significantly less estimation variance. On the other hand, this conservative *a priori* SNR estimate introduces a delay, thus, speech may be distorted, especially during speech onset or offset. Therefore, the smoothing factor  $\beta_{DD}$  in (2.88) controls the tradeoff between musical noise and speech distortion during speech transition. A typical choice for  $\beta_{DD}$  is 0.98 [Ephraim and Malah, 1984], however, it is shown

in [Yu and Fingscheidt, 2011] and [Yu, 2013] that  $\beta_{\text{DD}} = 0.98$  is not a universally best choice for all speech enhancement approaches in terms of the amount of noise reduction, the amount of musical noise, and the quality of the speech component.

Please note that under SPU the *a priori* SNR is defined slightly different from (2.42), i. e.,  $E\{|S|^2|H_1\}/\sigma_D^2$ . However, *a priori* SNR estimators, such as the DD approach (2.88), measure  $E\{|S|^2\}/\sigma_D^2$ . Nevertheless, the *a priori* SNR under SPU is easily obtained by the unconditioned *a priori* SNR (estimated by, e. g., the DD approach) as [Ephraim and Malah, 1984]

$$\frac{E\{|S|^2|H_1\}}{\sigma_D^2} = \frac{1}{P(H_1)} \cdot \frac{E\{|S|^2\}}{\sigma_D^2}. \quad (2.90)$$

### A *Priori* SNR Estimation in the Cepstral Domain

An *a priori* SNR estimator using temporal smoothing in the cepstral domain is proposed in [Breithaupt et al., 2008]. As mentioned above, the ML *a priori* SNR estimate should be smoothed in order to reduce its fluctuations which may lead to musical noise. However, a smoothing in the STFT domain is rather adversely, because it may distort the fine spectral structure of speech, leading to speech distortion.

A cepstral domain representation of the noisy speech, i. e., the IDFT of the logarithm of a noisy speech spectrum, however, has the advantage over the spectral domain that the speech and the noise can be related with specific cepstral coefficients [Breithaupt et al., 2008]. Therefore, a selective smoothing of the ML *a priori* SNR estimate in the cepstral domain allows for an adequate smoothing of the noise and a moderate smoothing of the speech, resulting in more naturally sounding residual noise and less speech distortion compared to a smoothing in the spectral domain.

## 2.8 Summary

In this chapter we gave an overview of MMSE speech enhancement. We used a general MMSE estimation formula which can be employed in different estimation domains, i. e., the STS, STSA, and LSA domain. Furthermore, typically employed *a priori* knowledge incorporated for MMSE speech enhancement was recapitulated with a focus on speech spectral modeling. We derived a new bivariate generalized speech prior which nicely covers usually employed speech priors as special case. Using the generalized speech prior, a generalized MMSE estimator was derived for each estimation domain. Then, a synopsis of MMSE speech enhancement approaches was given by means of the generalized estimators: Utilizing the parameter set of a specific speech spectral model leads to any of the specific MMSE estimators

known from literature. As a further refinement to MMSE speech enhancement, SPU was taken into account. Common SPU approaches from literature both with adapted and fixed prior parameters were recapitulated. Finally, typical noise PSD and SNR estimators were introduced.

We identified new research topics: The *a posteriori* SPP estimation formula (2.69) can be generalized by employing a generalized speech prior, such as the new bivariate generalized gamma PDF from Section 2.3. Based on the generalized PDF, new SPP estimators can potentially be identified. Accordingly, an overview of *a posteriori* SPP estimators will be given in Chapter 4. Furthermore, an *a posteriori* SPP estimator with an underlying gamma speech spectral amplitude assumption will be derived. Then, a new *a posteriori* SPP estimator with fixed prior parameters and an underlying chi-distributed speech spectral amplitude model will be proposed in Chapters 5. MMSE estimation exploiting temporal correlation of speech STFT coefficients will be investigated and some interesting links to error concealment will be shown in Chapter 6. Finally, the approach from Chapter 6 will be enhanced by SPU estimation in Chapter 7.



# Chapter 3

## Simulation Setup and Instrumental Measures

This chapter will deal with instrumental evaluation of speech enhancement systems and diverse instrumental measures. In Section 3.1, the simulation setup which is employed in this thesis for instrumental evaluation of speech enhancement systems will briefly be introduced. Besides the speech and noise data utilized for simulations, the data preprocessing steps will be recapitulated. Furthermore, a signal-component-wise evaluation procedure of the speech enhancement system under test will be described, in a so-called white box test scenario. This approach allows for analyzing the effects of signal enhancement on the speech and the noise component of the noisy signal separately. In Section 3.2 the instrumental measures will be summarized which are used in this thesis for performance evaluation. Section 3.3 will deal with the instrumental measure SNR: First, a reference-based SNR measurement approach will be recapitulated and then a new reference-free SNR measurement approach will be introduced. The reference-based SNR measurement approach will allow for measuring the SNR of noisy speech signals, e. g., those being fed into the speech enhancement system, by means of separately accessible speech and noise components as reference. The new SNR measurement approach will aim at estimating the SNR of speech signals distorted by car noise as close as possible to the measured SNR of the reference-based approach, however, in a reference-free fashion by means of the observed noisy speech only [Fodor and Fingscheidt, 2012c]. Furthermore, the new approach will be applicable to both narrowband and wideband signals. Such a reference-free SNR measurement approach allows for, e. g., finding advantageous microphone positions in the context of automotive hands-free system development. Within the ITU-T Study Group 12, the Focus Group on Car Communication (FG CarCOM) has decided to adopt the new reference-free SNR measurement approach within the draft of the recommendation proposal “Subsystem Requirements for Automotive Speech Services”. Finally, Section 3.4 will give a short summary of this chapter.

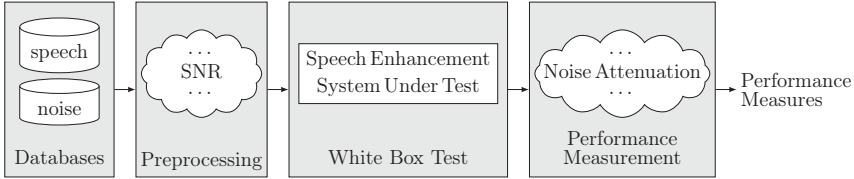


Figure 3.1: Overview of the simulation setup

## 3.1 Simulation Setup

In this section, we recapitulate the simulation setup employed in this thesis which can be divided into four main parts as depicted in Figure 3.1: First, the employed speech and noise recordings are provided by separate databases. The data is then preprocessed in the next step consisting of, e. g., artificial noisy signal generation and SNR adjustment. Then, the preprocessed data is fed into the speech enhancement system under test being analyzed in a white box test scenario. The signals gained in a white box test allow then for evaluating the speech enhancement system under test by means of performance measures. In the following, these steps will be described in detail.

### 3.1.1 Databases

As speech data, speech signals from the NTT Multi-Lingual Speech Database for Telephonometry [NTT, 1994] are employed in this thesis. This database provides 21 languages: Four female and four male speakers are assigned to each language and there are 12 recordings per speaker with 16 bit resolution, a sampling rate of 16 kHz, and a duration of 8 s each. For performance evaluation in this thesis, the English subset of the database is employed with a total of 96 recordings.

For evaluating speech enhancement approaches, car, factory, and babble noise signals from the NTT Ambient Noise Database [NTT, 1996] are employed in this thesis as noise data. This database provides recordings with 16 bit resolution, a sampling rate of 8 kHz, and a duration of 3 min each. The car noise subset consists of 84 car noise recordings made in 13 different cars. Please note that four recordings in the car noise data set made in buses were blacklisted due to the strong non-stationary nature of the recordings caused by announcements, speaking passengers, etc. The factory noise subset consists of 3 factory noise recordings. As babble noise data, a total of 14 noise files are used for simulations being recorded in environments with many different talkers, such as offices, cafes, amusement parks, exhibitions, etc.

For evaluating the proposed reference-free SNR measurement approach in Section 3.3.2

at 8 kHz and 16 kHz sampling rate, car noise signals from the ETSI database [ETSI, 2008] are employed as noise data. This database provides 9 recordings with 16 bit resolution, a sampling rate of 48 kHz, and a duration of 30 s each.

The databases provide clean speech recordings and noise only recordings. The noisy speech signals employed for evaluation in this thesis are generated artificially by superimposing clean speech and noise recordings as will be explained in the next section.

### 3.1.2 Preprocessing

The input signals are taken from the databases described above and decimated (if necessary) to 8 kHz sampling rate using software from the ITU-T Software Tool Library [ITU-T G.191, 2010]. During processing, the speech recordings from the database are processed sequentially. To each speech recording a pseudo-randomly chosen noise recording is assigned. Since the noise recordings are longer than the speech recordings (cf. Section 3.1.1), a randomly chosen noise sequence of the same length as the speech is taken from the assigned noise recording. All approaches are assessed in different SNR conditions, namely, input SNR values between -5 dB and 20 dB in 5 dB steps are used for evaluation. The pseudo-randomly chosen noise sequence remains exactly the same for different input SNR conditions apart from different scaling of course. The SNR is adjusted by scaling the speech and noise components separately, so that the resulting noisy speech being the sum of both components yields the desired SNR. Signal scaling is based on signal level measurements according to the ITU-T Recommendation P.56 [ITU-T P.56, 1993]. This recommendation defines the so-called active speech level for the speech recordings and the root mean square (RMS) signal level for the noise recordings. The *active speech level* is defined as a signal level measured *only during speech activity*, more precisely, it is an RMS signal level which is calculated by means of speech samples assigned by a time-domain VAD [ITU-T P.56, 1993]. This procedure ensures that the desired SNR is predominant during speech activity. In this thesis, an active speech level of  $-26 \text{ dB}_{\text{ov}}$  with index ov denoting the clipping level (the highest signal amplitude corresponds to  $0 \text{ dB}_{\text{ov}}$ ) and RMS noise levels of  $-26 \text{ dB}_{\text{ov}} - \text{SNR} [\text{dB}]$  are employed. Accordingly, after superimposing the adjusted speech and the noise signal, the resulting noisy speech signal exhibits the desired SNR.

### 3.1.3 White Box Test Setup

Maybe the most common test scenario in speech enhancement is depicted in Figure 3.2. This is a so-called white box test scenario [Gustafsson et al., 1996] and it assumes that the structure and the parameters of the speech enhancement system under test are known

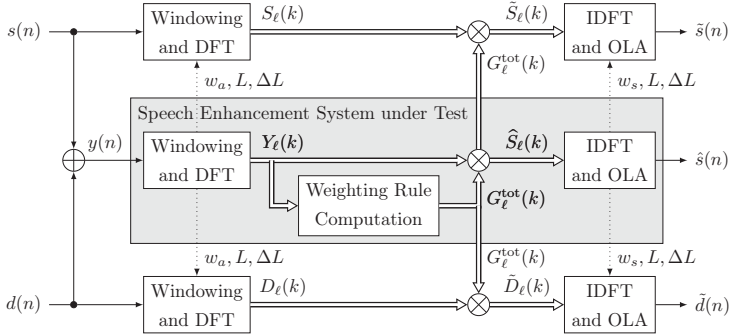


Figure 3.2: White box test setup for STFT-domain speech enhancement

and the internal variables are accessible from the outside. In case of STFT-domain speech enhancement, a white box test scenario means that the signal analysis and synthesis parameters (analysis window  $w_a$ , synthesis window  $w_s$ , frame length  $L$ , frame shift  $\Delta L$ , OLA signal reconstruction) are perfectly known to the test environment. Moreover, it is assumed that the output signal is a linear function of the input signal, more precisely, the estimated speech STFT coefficients  $\hat{S}_\ell(k)$  are obtained by multiplying the noisy speech STFT coefficients  $Y_\ell(k)$  by (real-valued) total spectral weights  $G_\ell^{\text{tot}}(k)$  (being either a plain spectral weighting rule or a combination of a spectral weighting rule and an *a posteriori* SPP) with the latter being accessible from the outside. Furthermore, it is also assumed in a white box test scenario that the signal components of the noisy speech  $y(n)$ , namely the speech component  $s(n)$  and the noise component  $d(n)$ , are accessible separately which is easy to achieve under lab conditions. Accordingly, since spectral weighting is a linear operation, the weighted noisy speech equals the sum of the weighted speech and the weighted noise, i. e.,

$$\hat{S}_\ell(k) = G_\ell^{\text{tot}}(k) \cdot Y_\ell(k) = G_\ell^{\text{tot}}(k) \cdot S_\ell(k) + G_\ell^{\text{tot}}(k) \cdot D_\ell(k) = \tilde{S}_\ell(k) + \tilde{D}_\ell(k) \quad (3.1)$$

with  $\tilde{S}_\ell(k) = G_\ell^{\text{tot}}(k) \cdot S_\ell(k)$  and  $\tilde{D}_\ell(k) = G_\ell^{\text{tot}}(k) \cdot D_\ell(k)$  being the processed speech component and the processed noise component, respectively. Furthermore, since the analysis and synthesis steps are also linear transformations, the time-domain equivalent of (3.1) exactly yields

$$\hat{s}(n) = \tilde{s}(n) + \tilde{d}(n) \quad (3.2)$$

with  $\tilde{s}(n)$  and  $\tilde{d}(n)$  being the time-domain equivalents of  $\tilde{S}_\ell(k)$  and  $\tilde{D}_\ell(k)$ , respectively. A considerable advantage of a white box test scenario is that it allows for a signal-component-based assessment of the speech enhancement system under test. That means that the effects of signal enhancement on the speech and noise components can be analyzed *separately*.

After the preprocessing step (cf. Section 3.1.2) the noisy speech signal  $y(n)$  with the desired SNR is fed into the speech enhancement system under test (cf. Figure 3.2). Here, the noisy speech is first transformed in the STFT domain employing (2.1) resulting in the noisy speech STFT coefficients  $Y_\ell(k)$ . For the frame-based processing at 8kHz sampling frequency a frame length of  $L = 256$  samples, a frame shift of  $\Delta L = 128$  samples, and a square root Hann window as analysis and synthesis window [Vary and Martin, 2006]

$$w_a(n') = w_s(n') = w(n') = \sqrt{0.5 - 0.5 \cdot \cos\left(n' \cdot \frac{2\pi}{L-1}\right)}, \quad n' = 0, 1, \dots, L-1 \quad (3.3)$$

are employed.

Then, the total spectral gain  $G_\ell^{\text{tot}}(k)$  in Figure 3.2 is computed by the following steps: A noise PSD estimate  $\widehat{\sigma}_{D_\ell}^2(k)$  is calculated by means of the noisy speech STFT coefficients  $Y_\ell(k)$  utilizing the widely employed MS approach (cf. Section 2.7.1). The *a priori* SNR is estimated by the well-known DD estimator (2.88) with parameter  $\beta_{\text{DD}} = 0.98$ , the *a posteriori* SNR is determined by (2.87). Please note that both approaches with and without SPU will be evaluated in this thesis. As mentioned in Section 2.7.2, the *a priori* SNR is defined slightly different depending on whether SPU is taken into account or not (cf. (2.90)). Instead, to allow for a fair comparison, *all* approaches incorporating the *a priori* SNR were driven by the DD estimate (2.88). Accordingly, the desired spectral weights (with or without taking SPU into account) are driven by the resulting *a priori* SNR  $\widehat{\xi}_\ell(k)$  and/or the *a posteriori* SNR  $\widehat{\gamma}_\ell(k)$ . Subsequently, the noisy speech STFT coefficients  $Y_\ell(k)$  are enhanced by the resulting spectral weights  $G_\ell^{\text{tot}}(k)$ , resulting in the estimated speech spectrum  $\widehat{S}_\ell(k)$ . The weighting rules and the *a posteriori* SPPs were implemented as lookup tables (using either numerical quadrature or, if exists, a closed-form formula) as a function of the *a priori* SNR  $\xi$  and the *a posteriori* SNR  $\gamma$  within the range [-20 dB, 20 dB] in 0.2 dB steps.

After signal enhancement, the estimated speech STFT coefficients  $\widehat{S}_\ell(k)$  are transformed back into the time domain, employing an IDFT (2.2), applying the synthesis window (3.3), and performing an OLA (2.3) step, resulting in the enhanced speech signal  $\widehat{s}(n)$ .

The performance of the speech enhancement approaches is evaluated by instrumental measures which will be introduced in the next section.

## 3.2 Speech Enhancement Performance Measurement

The signal-component-wise performance evaluation approach employed in this thesis assesses the speech and noise components separately. Therefore, the effects of signal enhancement on the speech are investigated by means of the speech component  $s(n)$  and the processed speech component  $\widehat{s}(n)$ . Meanwhile, the effects of signal enhancement on the noise are analyzed by

means of the noise component  $d(n)$  and the processed noise component  $\tilde{d}(n)$ . The processed signal components  $\tilde{s}(n)$  and  $\tilde{d}(n)$  are obtained from a white box test scenario (cf. Figure 3.2).

### 3.2.1 Speech Component

In this thesis, we measure the quality degradation of the speech component with the segmental speech-to-speech distortion ratio  $\text{SSDR}_{\text{seg}}$  [Fingscheidt et al., 2008]. This measure compares the power of the speech component  $s(n)$  to the power of the difference between the speech component and the processed speech component  $s(n) - \tilde{s}(n)$  within frames. More specifically, based on frames with the length of  $T = 256$  samples and a frame shift of also 256 samples (at a sampling rate of 8 kHz),  $\text{SSDR}_{\text{seg}}$  is defined as [Fingscheidt et al., 2008]

$$\begin{aligned} \text{SSDR}_{\text{seg}} &= \frac{1}{|\Phi|} \sum_{\lambda \in \Phi} \text{SSDR}(\lambda), \\ \text{SSDR}(\lambda) &= \mathfrak{L} \left\{ 10 \cdot \log_{10} \frac{\sum_{\tau=0}^{T-1} s^2(\tau + \lambda T)}{\sum_{\tau=0}^{T-1} [s(\tau + \lambda T) - \tilde{s}(\tau + \lambda T)]^2} \right\} \end{aligned} \quad (3.4)$$

with  $\Phi$  and  $|\Phi|$  being the frames with speech activity and the number of those frames. Please note that in a white box test setup, the speech component is accessible, thus, frames with speech activity can be identified with high reliability. Operator  $\mathfrak{L}\{\cdot\}$  limits  $\text{SSDR}(\lambda)$  to the range  $[-10, 30]$  dB. In order to identify speech active frames  $\Phi$ , a fixed threshold can be used. A frame is detected as speech active and becomes part of the set  $\Phi$ , if the mean power of the speech frame  $1/T \cdot \sum_{\tau=0}^{T-1} s^2(\tau + \lambda T)$  exceeds this threshold. In this thesis, the threshold is chosen to be 10% of the mean power of the whole speech signal.

Large  $\text{SSDR}_{\text{seg}}$  values mean a small amount of speech distortion, since in this case the power of the error  $s(n) - \tilde{s}(n)$  is significantly smaller than the speech power. A small  $\text{SSDR}_{\text{seg}}$  value means that the speech component may have significantly been distorted by the speech enhancement system.

### 3.2.2 Noise Component

Using the same frame structure as  $\text{SSDR}_{\text{seg}}$ , the amount of attenuated noise is measured by the segmental noise attenuation measure [Fingscheidt et al., 2008]

$$\text{NA}_{\text{seg}} = 10 \cdot \log_{10} \frac{1}{N_\lambda} \sum_{\lambda=0}^{N_\lambda-1} \frac{\sum_{\tau=0}^{T-1} d^2(\tau + \lambda T)}{\sum_{\tau=0}^{T-1} \tilde{d}^2(\tau + \lambda T)} \quad (3.5)$$

with  $N_\lambda$  being the number of all frames. Due to noise reduction, the residual noise power  $\tilde{d}^2(n)$  decreases compared to the input noise power  $d^2(n)$ , thus, the larger the  $\text{NA}_{\text{seg}}$  value the larger the amount of attenuated noise.

Since the noisy speech  $Y_\ell(k)$  has a random nature, the *a posteriori* SNR  $\gamma_\ell(k) = |Y_\ell(k)|^2/\sigma_{D,\ell}^2(k)$  being a function of the periodogram of the noisy speech  $|Y_\ell(k)|^2$  is also random with a large variance [Papoulis and Pillai, 2002, Chapter 12]. Therefore, driven by a randomly fluctuating *a posteriori* SNR, speech enhancement approaches (common MMSE estimators with or without *a posteriori* SPP estimators) may produce estimation outliers which may be perceived as annoying *musical noise* [Gerkmann, 2010], typically during speech absence.

The amount of musical noise can be measured by the weighted log-kurtosis ratio (LKR) [Yu and Fingscheidt, 2012], [Yu, 2013] in the STFT domain employing the same signal analysis structure and parameters as for speech enhancement (cf. Section 3.1.2). This measure is motivated by the fact that estimation outliers influence the statistical parameters of the noise. The idea is to compare the kurtosis of the *noise component* of the noisy speech signal before and after processing in absence of speech, resulting in the kurtosis ratio. The weighted kurtosis is calculated for the noise component of the noisy speech signal (denoted by subscript  $D$ ) as [Yu and Fingscheidt, 2012]

$$\Psi_D^w = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{\frac{1}{N_K} \sum_{k=0}^{N_K-1} [\alpha_n(k) \cdot |D_\ell(k)|^2 - \mu_{\text{wn},\ell}]^4}{\left( \frac{1}{N_K} \sum_{k=0}^{N_K-1} [\alpha_n(k) \cdot |D_\ell(k)|^2 - \mu_{\text{wn},\ell}] \right)^2} \quad (3.6)$$

with  $\mu_{\text{wn},\ell} = \frac{1}{N_K} \sum_{\kappa=0}^{N_K-1} \alpha_n(\kappa) \cdot |D_\ell(\kappa)|^2$ , as well as with  $\mathcal{L}$ ,  $|\mathcal{L}|$ , and  $N_K$  being the set of frames with speech absence, the total number of speech absent frames, and the number of frequency bins, respectively. The weights are computed as  $\alpha_n(k) = \left( \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} |D_\ell(k)|^2 \right)^{-1}$ . For the processed noise component  $\tilde{D}$ , the weighted kurtosis  $\Psi_{\tilde{D}}^w$  can be calculated in the same way. The weighted log-kurtosis ratio is finally obtained by  $\Delta\Psi_{\text{ln}}^w = \ln(\Psi_{\tilde{D}}^w/\Psi_D^w)$ .

The smaller the value of  $\Delta\Psi_{\text{ln}}^w$  the less estimation outliers affected the noise component and the less the amount of musical noise.

### 3.3 SNR Measurement

The knowledge of the SNR of noisy speech signals is substantial in speech enhancement. A typical example from the lab environment is performance evaluation: In order to prove the robustness of speech enhancement approaches, they are typically assessed under different SNR conditions (cf. Section 3.1.2). For this, the SNR of the noisy speech signal has to be measured. In a lab environment the SNR can be measured by using the separately accessible speech and noise components as references (cf. Section 3.1.2). However, outside a lab in a real environment these signal components are typically inaccessible, thus, SNR measurements

have to be carried out in a reference-free fashion just on the basis of the noisy speech signal. In this section, first a reference-based SNR measurement approach is recapitulated which is followed by a proposal of a new, reference-free SNR measurement approach, inspired by the reference-based one.

### 3.3.1 Reference-Based SNR Measurement

A widely employed SNR measurement approach is provided by the ITU-T Recommendation P.56 [ITU-T P.56, 1993] (cf. Section 3.1.2): The SNR of a noisy speech signal being a sum of the speech component and the noise component is determined by measuring the signal level of these components separately, as they are typically accessible under lab conditions. Thus, this approach is considered as *reference-based*. Thereby, the level of the speech and noise components are defined by the active speech level and the RMS signal level, respectively. After superposition of both signal components, the SNR of the resulting noisy speech signal is the ratio of the measured active speech level and the measured RMS noise level. PC software to measure speech and noise signal levels according to the ITU-T Recommendation P.56 is provided by the ITU-T Software Tool Library [ITU-T G.191, 2010].

### 3.3.2 New Reference-Free SNR Measurement

While the reference-based approach in Section 3.3.1 yields exact SNR values, it always requires its reference signals which are typically not accessible in a real environment. A possible example is the measurement of the SNR improvement of speech enhancement systems being fed by real recordings. In this case, the signal components are not accessible and the SNR at the input and the output of the system under test have to be measured in order to determine the SNR improvement. A further, automotive example is the investigation of advantageous hands-free microphone positions in a car with the lowest SNR levels measured in a noisy environment. In this context, often the signal components cannot be measured separately or such measurements are cost-intensive, e.g., due to the need of a free-field room. Since the new reference-free SNR measurement approach does not assume the signal components separately, the investigation of advantageous hands-free microphone positions or similar experiments become possible or can be carried out faster and more cost-effective.

We propose an STFT-domain reference-free SNR measurement approach which measures the SNR of a speech signal in the presence of car noise as close as possible to the *reference* SNR according to the ITU-T Recommendation P.56 [ITU-T P.56, 1993], but *without a reference*, i. e., without having access to the speech and noise signal components [Fodor and Fingscheidt, 2012c]. While we intended to develop an approach which measures the SNR

accurately, we aimed at keeping the computational complexity as low as possible (this means that we tried not to use approaches such as minimum statistics from Section 2.7.1). A further motivation to an as-simple-as-possible approach was that we targeted widely stationary noise conditions (automotive applications). Furthermore, the new STFT-domain approach should be able to work with narrowband ( $f_s = 8$  kHz) and wideband ( $f_s = 16$  kHz) signals, within a typical SNR range.

### Algorithmic Approach

As introduced in Section 3.3.1, the reference SNR of a noisy speech signal is defined by the active speech level and the RMS noise level, according to the ITU-T Recommendation P.56. Accordingly, in order to determine the SNR, our STFT-domain approach estimates the active speech level and the RMS noise level by observing the noisy speech only, i. e., without any reference [Fodor and Fingscheidt, 2012c]. The RMS noise level is estimated by means of a noise power estimate and, to improve the estimation robustness, a speech pause detection (SPD). Assuming that speech and noise are statistically independent, the speech power can be approximated by the difference between the noisy speech power and the previously estimated noise power. Furthermore, based on the fact that the active speech level is measured only during voice activity, the final speech power estimate is obtained by applying a voice activity detection (VAD) to the previous approximation. Since the robustness of the employed algorithms (noise power estimation, VAD, SPD) decreases significantly below 0 dB input SNR, systematic estimation errors occur in this region. In order to achieve unbiased SNR values, the raw SNR estimates are corrected by a mapping curve.

**SNR Measurement** Using the signal model from Chapter 2, we define the SNR in the STFT domain similar to the time-domain definition in Section 3.3.1 as [Fodor and Fingscheidt, 2012c]

$$\text{SNR} = \frac{\frac{1}{|\mathcal{L}_1| \cdot N_k} \sum_{\ell \in \mathcal{L}_1} \sum_k |S_\ell(k)|^2}{\frac{1}{N_\ell \cdot N_k} \sum_\ell \sum_k |D_\ell(k)|^2} = \frac{P_S}{P_D} \quad (3.7)$$

with  $\mathcal{L}_1$ ,  $|\mathcal{L}_1|$ ,  $N_\ell$ ,  $N_k$ , and  $P_S$ ,  $P_D$  being a set of speech active frames, the number of speech active frames, the number of all frames and frequency bins, as well as the speech and the noise power, respectively.

As the signal components  $S_\ell(k)$  and  $D_\ell(k)$  are not accessible in practice,  $P_S$  and  $P_D$  have to be estimated. Since speech and noise are statistically independent, the instantaneous speech power can be calculated as  $\widehat{P}_{S,\ell}(k) = \max\{|Y_\ell(k)|^2 - \widehat{\sigma}_{D,\ell}^2(k), 0\}$  with  $|Y_\ell(k)|^2$ ,  $\widehat{\sigma}_{D,\ell}^2(k)$ , and  $\max\{\cdot, 0\}$  being the periodogram of the noisy speech signal, the estimated noise PSD, and the maximum operator which avoids negative, implausible speech power estimates, respectively.

Taking only speech active frames into account, just as in the case of active speech level measurement, the speech power  $P_S$  in (3.7) can be estimated by means of the instantaneous speech power as [Fodor and Fingscheidt, 2012c]

$$\widehat{P}_S = \frac{1}{|\mathcal{L}_1| \cdot N_k} \sum_{\ell \in \mathcal{L}_1} \sum_k \max \left\{ |Y_\ell(k)|^2 - \widehat{\sigma}_{D,\ell}^2(k), 0 \right\} \quad (3.8)$$

with  $\mathcal{L}_1$  and  $|\mathcal{L}_1|$  being the set and the number of speech active frames detected by a VAD, respectively.

As can be seen above, a significant contribution to the estimation of the speech power is the noise PSD estimate and it will also be employed for the estimation of the noise power  $P_D$ , as we will see. Therefore, an accurate noise PSD estimate is crucial for SNR estimation. In order to ensure a robust noise power estimate even in low SNR regions, where the noise PSD update in the presence of speech is significantly more challenging,  $P_D$  in (3.7) is estimated in frames with speech pause [Fodor and Fingscheidt, 2012c]

$$\widehat{P}_D = \frac{1}{|\mathcal{L}_0| \cdot N_k} \sum_{\ell \in \mathcal{L}_0} \sum_k \widehat{\sigma}_{D,\ell}^2(k) \quad (3.9)$$

with  $\mathcal{L}_0$  and  $|\mathcal{L}_0|$  being the set and the number of speech pause frames detected by a SPD, respectively. Please note that for robustness reasons, a *separate conservative* SPD is used.

As can be seen, the SNR estimation is based on a noise PSD estimation, a VAD, and an SPD which will be introduced in the following.

**Noise PSD Tracking** The noise PSD estimate  $\widehat{\sigma}_{D,\ell}^2(k)$  in (3.8) and (3.9) is based on a 3-state classifier for each time-frequency unit  $(\ell, k)$ . Controlled by the smoothed periodogram of the noisy speech signal [Fodor and Fingscheidt, 2012c]

$$P_{Y,\ell}(k) = 0.5 \cdot P_{Y,\ell-1}(k) + 0.5 \cdot |Y_\ell(k)|^2 \quad (3.10)$$

and an adaptive threshold  $\Theta_\ell(k)$ , one state  $H_\ell(k)$  out of the following three is chosen [Fodor and Fingscheidt, 2012c]:

$H_{\text{sa}}$  : Speech *activity* is assumed if  $P_{Y,\ell}(k) > 2 \cdot \Theta_{\ell-1}(k)$ ,

$H_{\text{sp}}$  : Speech *pause* is assumed if  $P_{Y,\ell}(k) \leq 2 \cdot \Theta_{\ell-1}(k)$  and  $P_{Y,\ell}(k) < \widehat{\sigma}_{D,\ell-1}^2(k)$ ,

$H_{\text{st}}$  : Speech *transition* is assumed if  $P_{Y,\ell}(k) \leq 2 \cdot \Theta_{\ell-1}(k)$  and  $P_{Y,\ell}(k) \geq \widehat{\sigma}_{D,\ell-1}^2(k)$ .

The noise PSD estimate  $\widehat{\sigma}_{D,\ell}^2(k)$  is computed by recursive averaging by means of the smoothed noisy speech periodogram  $P_{Y,\ell}(k)$  as [Fodor and Fingscheidt, 2012c]

$$\widehat{\sigma}_{D,\ell}^2(k) = \epsilon_\ell(k) \cdot \widehat{\sigma}_{D,\ell-1}^2(k) + [1 - \epsilon_\ell(k)] \cdot P_{Y,\ell}(k) \quad (3.11)$$

with the initial value  $\widehat{\sigma}_{D,\ell=0}^2(k) = 0$  and  $\epsilon_\ell(k)$  being the time-varying smoothing factor which is controlled by the hypothesis of the current frame  $H_\ell(k)$  as [Fodor and Fingscheidt, 2012c]

$$\epsilon_\ell(k) = \begin{cases} 1 & \text{if } H_\ell(k) = H_{\text{sa}}, \\ 0.5 & \text{if } H_\ell(k) = H_{\text{sp}}, \\ 0.875 & \text{if } H_\ell(k) = H_{\text{st}}. \end{cases} \quad (3.12)$$

As can be seen, the noise PSD estimate  $\widehat{\sigma}_{D,\ell}^2(k)$  is not updated during speech activity hypothesis  $H_{\text{sa}}$ , updated moderately by the smoothed noisy periodogram in the state speech pause  $H_{\text{sp}}$ , and updated slightly by the smoothed noisy periodogram during speech transitions under hypothesis  $H_{\text{st}}$ . The adaptive threshold  $\Theta_\ell(k)$  is calculated by [Fodor and Fingscheidt, 2012c]

$$\Theta_\ell(k) = \begin{cases} \Delta \cdot \Theta_{\ell-1}(k) & \text{if } P_{Y,\ell}(k) > 2 \cdot \Theta_{\ell-1}(k), \\ P_{Y,\ell-1}(k) & \text{if } P_{Y,\ell}(k) < \Theta_{\ell-1}(k), \\ \Theta_{\ell-1}(k) & \text{else} \end{cases} \quad (3.13)$$

with a very large initial value  $\Theta_{\ell=0}(k) \rightarrow \infty$  and the parameter  $\Delta = 1.07$  for narrowband signals ( $\Delta = 1.2$  for wideband signals). During speech activity ( $P_{Y,\ell}(k) > 2 \cdot \Theta_{\ell-1}(k)$ ), the threshold  $\Theta_\ell(k)$  is increased slightly, in order to steadily decrease the probability of choosing state  $H_{\text{sa}}$ . In the case  $P_{Y,\ell}(k) < \Theta_{\ell-1}(k)$ , which can be interpreted as a more conservative speech absence detector than just state  $H_{\text{sp}}$ , the adaptive threshold can be adapted quickly by the last smoothed periodogram value  $P_{Y,\ell-1}(k)$ . In all other cases, the threshold is not updated.

**Voice Activity Detection (VAD)** In order to identify speech active frames  $\mathcal{L}_1$  in (3.8), a VAD is needed. We employ a frame-based VAD which is connected to the three hypotheses of the noise PSD estimator introduced above. Accordingly, frame  $\ell$  is detected as *speech active* and becomes an element of the set  $\mathcal{L}_1$ , if at least 90% of its frequency bins from the range 500...2500 Hz containing relevant speech information are classified as speech active ( $H_{\text{sa}}$ ) or transient ( $H_{\text{st}}$ ) hypothesis.

**Speech Pause Detection (SPD)** In order to identify frames with speech pause  $\mathcal{L}_0$  in (3.9), an SPD is needed. Different from the VAD, we employ a separate SPD algorithm which is independent of the noise PSD estimation. The SPD works frame-wise and is based on the noisy frame power [Fodor and Fingscheidt, 2012c]

$$\overline{P}_{Y,\ell} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} P_{Y,\ell}(k) \quad (3.14)$$

with  $P_{Y,\ell}(k)$  from (3.10) and with  $\mathcal{K}$  and  $|\mathcal{K}|$  being a set of frequency bins between 500 Hz and 2500 Hz and the number of elements in this set, respectively. This frequency range contains relevant speech information and excludes frequencies below 500 Hz where usually car noise is dominant. Similar to the noise PSD estimation, the SPD is also based on three hypotheses determined by an adaptive threshold  $\Xi_\ell$  and the noisy frame power  $\overline{P_{Y,\ell}}$ . Accordingly, one state  $H_\ell^{\text{SPD}}$  out of the following three is chosen [Fodor and Fingscheidt, 2012c]:

$H_{\text{SA}}$  : Speech *activity* is assumed if  $\overline{P_{Y,\ell}} > \Xi_\ell$ ,

$H_{\text{ST}}$  : Speech *transition* state follows every  $H_{\text{SA}}$  decision with a duration of  $L_{\text{trans}}$  frames,

$H_{\text{SP}}$  : Speech *pause* is assumed in all other situations.

We assign  $H_\ell^{\text{SPD}} = H_{\text{SP}}$ , if frame  $\ell$  is detected as part of a *speech pause* and, thus, becomes element of  $\mathcal{L}_0$ . The adaptive threshold  $\Xi_\ell$  is calculated independently of the classifier as [Fodor and Fingscheidt, 2012c]

$$\Xi_\ell = \zeta \cdot \Phi_{\ell-1} + \phi \quad (3.15)$$

with  $\zeta$  and  $\phi$  being constant terms and  $\Phi_\ell$  being the SPD floor signal which is calculated by recursive averaging (cf. (3.11)) [Fodor and Fingscheidt, 2012c]

$$\Phi_\ell = \beta_{\text{SPD},\ell} \cdot \Phi_{\ell-1} + (1 - \beta_{\text{SPD},\ell}) \cdot \overline{P_{Y,\ell}} \quad (3.16)$$

where the initial value is  $\Phi_{\ell=0} = 0$ . The smoothing of the SPD floor signal  $\Phi_\ell$  is time-varying and controlled by the smoothing factor (cf. (3.12)) [Fodor and Fingscheidt, 2012c]

$$\beta_{\text{SPD},\ell} = \begin{cases} 0.875, & \text{if } H_\ell^{\text{SPD}} \neq H_{\text{ST}} \text{ and } \overline{P_{Y,\ell}} \leq 2 \cdot \Upsilon_{\ell-1} \text{ and } \overline{P_{Y,\ell}} > \Phi_{\ell-1}, \\ 0.5, & \text{if } H_\ell^{\text{SPD}} \neq H_{\text{ST}} \text{ and } \overline{P_{Y,\ell}} \leq 2 \cdot \Upsilon_{\ell-1} \text{ and } \overline{P_{Y,\ell}} \leq \Phi_{\ell-1}, \\ 1, & \text{else.} \end{cases} \quad (3.17)$$

Accordingly, the SPD floor signal  $\Phi_\ell$  is adapted only under hypotheses  $H_{\text{SA}}$  and  $H_{\text{SP}}$ . It is updated slightly by the noisy frame power  $\overline{P_{Y,\ell}}$  during speech transitions and updated moderately by the noisy frame power in speech pauses. The above conditions, however, ensure a more conservative SPD floor signal update than just conditioning on the three hypotheses. The SPD control parameter  $\Upsilon_\ell$  is defined as (cf. (3.13)) [Fodor and Fingscheidt, 2012c]

$$\Upsilon_\ell = \begin{cases} \delta \cdot \Upsilon_{\ell-1}, & \text{if } \overline{P_{Y,\ell}} > 2 \cdot \Upsilon_{\ell-1}, \\ \overline{P_{Y,\ell}}, & \text{if } \overline{P_{Y,\ell}} < \Upsilon_{\ell-1}, \\ \Upsilon_{\ell-1}, & \text{else} \end{cases} \quad (3.18)$$

with the control update constant  $\delta$  and the initial value  $\Upsilon_{\ell=0} \rightarrow \infty$ . During speech activity ( $\overline{P_{Y,\ell}} > 2 \cdot \Upsilon_{\ell-1}$ ), the SPD control parameter  $\Upsilon_\ell$  is increased slightly, in order to steadily

Parameter	$L_{\text{trans}}$	$\zeta$	$\phi$	$\delta$
$f_s = 8$ kHz	6	5	$10^7$	1.085
$f_s = 16$ kHz	7	5	$10^8$	1.055

Table 3.1: Parameters of the SPD [Fodor and Fingscheidt, 2012c]

decrease the probability of choosing state  $H_{\text{SA}}$ . In the case  $\overline{P_{Y,\ell}} < \Upsilon_{\ell-1}$  speech is absent and, therefore, the SPD control parameter  $\Upsilon_\ell$  can quickly be adapted by the last noisy frame power  $\overline{P_{Y,\ell}}$ . In all other cases, the control parameter is not updated.

The SPD parameters differ for the narrowband and the wideband implementation and are summarized in Table 3.1 (please note that  $\phi$  is a signal-level-dependent quantity: The large parameter values occur due to the fact that the parameter training was carried out by 16 bit pulse-code modulation (PCM) signals).

**Training of the Mapping Curve** In order to prove the accuracy of the proposed reference-free SNR measurement approach, we performed the following simulations on our training speech and noise data set: 720 speech files (48 speech files in 15 different languages) recorded with a sampling rate of 16 kHz were taken from the NTT Multi-Lingual Speech Database for Telephonometry [NTT, 1994], each with a length of 8 s. Car noise signals were randomly taken from the ETSI database [ETSI, 2008] recorded with a sampling rate of 48 kHz. Both narrowband speech and narrowband noise signals were generated by downsampling the database signals to 8 kHz and filtering according to the ITU-T modified IRS filter mask [ITU-T P.830, 1996] separately in order to simulate the sending frequency characteristics of narrowband telephony terminals (including, amongst others, the frequency characteristics of the microphone). Both wideband speech and wideband noise signals were simulated by downsampling the database signals to 16 kHz and filtering according to the ITU-T P.341 filter mask [ITU-T P.830, 1996] separately in order to simulate the sending frequency characteristics of wideband telephony terminals (including, amongst others, the frequency characteristics of the microphone). Next, both narrowband and wideband speech signals was adjusted to an active speech level of  $-26$  dB<sub>ov</sub>, according to the ITU-T Recommendation P.56 [ITU-T P.56, 1993]. The RMS level of both narrowband and wideband noise signals were adjusted in such a way, that after superimposing the corresponding speech and noise signals, the resulting noisy speech signals revealed the desired reference SNR value  $\text{SNR}_{\text{ref}} = \{-15, -14, \dots, 35\}$  dB. The noisy speech signals were transformed into the STFT domain by signal analysis consisting of a segmentation step with a frame length of 256 (512) samples, a frame shift of 128 (256) samples, and a Hann window function, as well as a DFT step for 8 kHz (16 kHz) signals.

The raw SNR was estimated in the STFT domain using (3.8) and (3.9) as [Fodor and

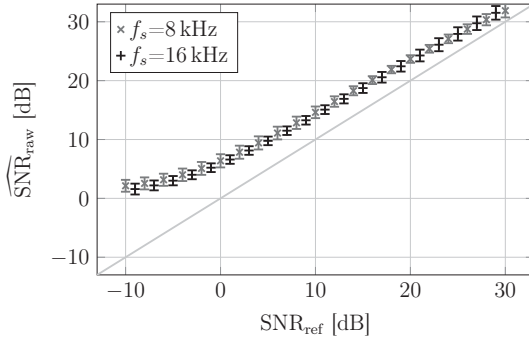


Figure 3.3: Mean (x- and +-marks) and standard deviation (error bars) of the raw measurement. The ideal linear relationship is represented by the diagonal gray line. Measurement duration was 8 s.

Fingscheidt, 2012c]

$$\widehat{\text{SNR}}_{\text{raw}}(i) = 10 \log_{10} \frac{\widehat{P}_S(i)}{\widehat{P}_D(i)} \quad (3.19)$$

for the  $i$ th database file. For (3.8) we utilized the noise PSD tracking algorithm introduced above for estimating the noise PSD  $\widehat{\sigma}_{D,\ell}^2(k)$  and the VAD algorithm introduced above for identifying speech active frames  $\mathcal{L}_1$ . For (3.9) we utilized the noise PSD tracking algorithm introduced above for estimating the noise PSD  $\widehat{\sigma}_{D,\ell}^2(k)$  and the SPD algorithm introduced above for recognizing frames with speech pause  $\mathcal{L}_0$ . Since reliable SNR estimates can only be expected in the steady state of the algorithms, the first 15 frames were not considered both in set  $\mathcal{L}_0$  and  $\mathcal{L}_1$ . The SNR estimation (3.19) was repeated for all speech files from the training data set and all reference SNR values, resulting in raw SNR estimates depicted in Figure 3.3. As can be seen, the SNR estimate is biased and at low SNRs, where the accuracy of the employed algorithms is affected, a non-linear relationship between the reference and the raw SNR can be observed. In order to correct these systematic measurement errors, we employ mapping curves which map the measured raw SNR values to the correct ones. The mapping curves can be approximated by the inverses of the mean curves in Figure 3.3. In practice, however, there is an issue to deal with: Due to the very low gradient of the curves in Figure 3.3 in the reference SNR region below 0 dB, the resulting mapping curves (being the inverses of the raw measurement curves) show very large gradients in this SNR region which leads to an unacceptable attenuation of the standard deviation (the height of the error bars) of the corrected estimates. Thus, we decided to saturate the mapping curves at -11 dB (cf. Figure 3.4). In the residual part we applied a polynomial fit [Mathews and Fink, 2004] to the inverses raw measurement curves, resulting in the mapping function [Fodor and Fingscheidt,

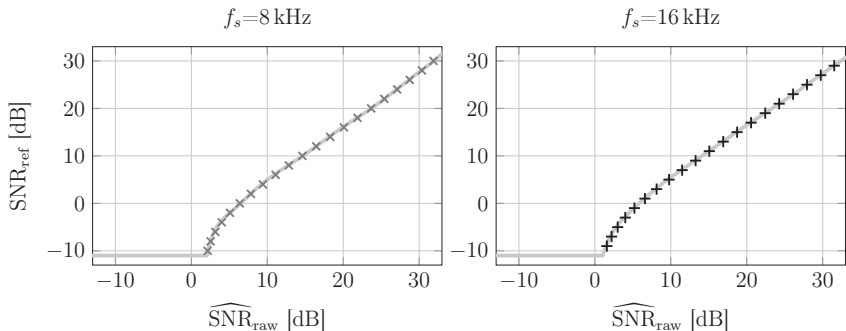


Figure 3.4: Measured mapping points (x- and +-marks) and the fitted mapping function  $\widehat{\text{SNR}} = f(\widehat{\text{SNR}}_{\text{raw}})$  (gray curves) according to (3.20)

$f_s$	$c$	$a$	$p_\mu$						
			$p_0$	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	
8 kHz	16.043	11.252	-2.1312	6.4129	2.0957	-19.199	5.0992	19.709	
16 kHz	15.461	11.798	-0.80823	2.8537	-0.3609	-7.5337	4.3304	7.2828	
$f_s$			$p_6$	$p_7$	$p_8$	$p_9$	$p_{10}$		
8 kHz			-7.7268	-6.6348	2.0857	13.066	11.555		
16 kHz			-5.0623	-1.8734	0.88424	13.06	11.48		

Table 3.2: Scaling parameters  $a$ ,  $c$  and polynomial coefficients  $p_\mu$  of the mapping curve [Fodor and Fingscheidt, 2012c]

2012c]

$$\widehat{\text{SNR}} = \max \left\{ \sum_{\mu=0}^{10} p_\mu \cdot \left( \frac{\widehat{\text{SNR}}_{\text{raw}} - c}{a} \right)^\mu, -11 \right\} \quad (3.20)$$

with  $\widehat{\text{SNR}}_{\text{raw}}$  being calculated by (3.19) and with the scaling parameters  $a$  and  $c$  as well as the polynomial coefficients  $p_\mu$  as shown in Table 3.2. The result of the fit is depicted in Figure 3.4.

### Performance Evaluation

We evaluated the proposed method including mapping (3.20) by the following simulations: Both the test speech set and the noise data set arise from the same databases and are of the same size as the training data set, however, the training and test data sets are disjoint. The test database signals were preprocessed in the same way as the training database signals and the reference SNRs were chosen to be  $\text{SNR}_{\text{ref}} = \{-15, -14, \dots, 35\}$  dB. The SNR of each

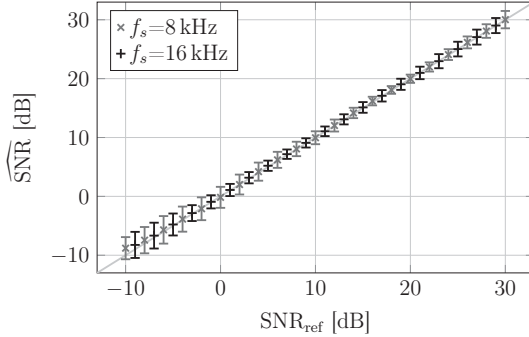


Figure 3.5: Mean (x- and +-marks) and standard deviation (error bars) of the corrected measurement. The ideal linear relationship is represented by the diagonal gray line. Measurement duration was 8 s.

noisy speech signal was estimated by (3.19) and corrected by (3.20) with the parameters from Table 3.2. After simulating all test database files, the mean and the standard deviation of the corrected SNR values  $\widehat{\text{SNR}}$  was measured. The result is depicted in Figure 3.5. Compared to Figure 3.3, the estimation bias and the nonlinearity could significantly be reduced by the mapping function.

The proposed approach was assessed by means of the absolute measurement error and the correlation coefficient between the reference-based SNR measurement according to Section 3.3.1 employing ITU-T P.56 reference levels and the proposed, reference-free approach.

Based on a measurement duration of  $T_m = 8$  s according to an underlying database signal length of 8 s, the proposed SNR measurement approach achieved a maximum of 1 dB absolute estimation error in at least 64.5% (67.8%) and a maximum of 2 dB absolute estimation error in at least 88.4% (91.6%) of the measurements for narrowband (wideband) signals, as can be seen in Table 3.3. However, applying longer speech sequences, e. g., by averaging estimation results in groups, the absolute error can substantially be decreased. This can be observed in Table 3.3 for groups of 10 and 60 files of 8 s each, reflecting a measurement duration of  $T_m = 80$  s and  $T_m = 480$  s, respectively. As a conclusion, we recommend averaging SNR measurements based on *groups* of 8 s signals, if the absolute measurement error needs to be reduced.

The correlation coefficient was calculated for all test speech signals and reference SNR

$f_s$	Absolute Estimation Error	Relative frequency		
		$T_m = 8\text{ s}$	$T_m = 80\text{ s}$	$T_m = 480\text{ s}$
8 kHz	$\leq 2\text{ dB}$	88.4 %	99.7 %	100 %
	$\leq 1\text{ dB}$	64.5 %	93.8 %	97.6 %
16 kHz	$\leq 2\text{ dB}$	91.6 %	99.7 %	100 %
	$\leq 1\text{ dB}$	67.8 %	94.9 %	97.4 %

Table 3.3: Relative frequency of absolute estimation error for different measurement durations  $T_m$  [Fodor and Fingscheidt, 2012c]

$f_s$	Correlation coefficient $\rho$		
	$T_m = 8\text{ s}$	$T_m = 80\text{ s}$	$T_m = 480\text{ s}$
8 kHz	0.9923	0.9981	0.9993
16 kHz	0.9940	0.9989	0.9996

Table 3.4: Correlation coefficient (3.21) between the corrected SNR measurements and the reference SNR values for different measurement durations  $T_m$  [Fodor and Fingscheidt, 2012c]

values by means of Pearson's formula [Burington and May, 1970]

$$\rho = \frac{\sum_i (\widehat{\text{SNR}}(i) - \overline{\widehat{\text{SNR}}}) \cdot (\text{SNR}_{\text{ref}}(i) - \overline{\text{SNR}_{\text{ref}}})}{\sqrt{\sum_i (\widehat{\text{SNR}}(i) - \overline{\widehat{\text{SNR}}})^2 \cdot \sum_i (\text{SNR}_{\text{ref}}(i) - \overline{\text{SNR}_{\text{ref}}})^2}} \quad (3.21)$$

with  $i$ ,  $\text{SNR}_{\text{ref}}$  and  $\overline{(\cdot)}$  being the test database file index, the reference SNR value of a reference-based measurement according to ITU-T Recommendation P.56, and the mean operator over  $i$ , respectively. The resulting correlation coefficients at different measurement durations  $T_m$  (simulated by building groups of the 8 s database signals) are depicted in Table 3.4. As can be seen, the proposed method achieves a correlation coefficient being larger than 0.99 in all cases.

### 3.4 Summary

In this chapter, the simulation setup was recapitulated which is used in this thesis for performance evaluation of speech enhancement approaches. Then, a brief description of the employed databases, data preprocessing steps, and a white box test setup were provided. This was followed by a summary of speech enhancement performance measures which are utilized in this thesis for a signal-component-based quality assessment in a white box test scenario. As further instrumental measurement tools first a typical, reference-based SNR

measurement method according to the ITU-T Recommendation P.56 was recapitulated and then a new reference-free SNR measurement method was proposed. The latter aims at estimating the P.56 reference SNR without using any reference as well as provides low computational complexity, small measurement errors, and a high correlation of the measurement results with the reference SNR. Moreover, it can be applied to both narrowband and wideband speech signals. Furthermore, within ITU-T Study Group 12, the Focus Group on Car Communication (FG CarCOM) has decided to adopt the new reference-free SNR measurement approach within the draft of the recommendation proposal “Subsystem Requirements for Automotive Speech Services”.

# Chapter 4

## Consistent MMSE Estimation Under Speech Presence Uncertainty

As shown in Section 2.6, MMSE estimation under SPU turns out to consist of a common MMSE estimator for speech and an *a posteriori* SPP estimator. Furthermore, since both estimators are in the same Bayesian estimation framework, they assume the same *a priori* knowledge, reflected by the same statistical models for speech and the acoustic channel. Thus, we argue that in order to obtain specific MMSE speech enhancement approaches under SPU, common MMSE estimators and *a posteriori* SPP estimators have to be combined *consistently* w. r. t. the PDF assumptions. Accordingly, this chapter will deal with PDF-consistent MMSE estimation under SPU.

In Section 4.1, a synopsis of *a posteriori* SPP estimators will be provided in analogy to Section 2.5. This synopsis will be based on the generalized gamma speech model from Section 2.3.1, resulting in a new generalized *a posteriori* SPP estimator [Fodor and Fingscheidt, 2012a]. The new generalized estimator covers existing *a posteriori* SPP estimators from literature, based on either a Rayleigh-distributed or a chi-distributed speech spectral amplitude model (cf. Section 2.6) as special case. Applying the parameter set of the gamma distribution to the generalized estimator, we will propose the respective *a posteriori* SPP estimator based on a gamma-distributed speech spectral amplitude model [Fodor and Fingscheidt, 2012a]. Then, in Section 4.2 an overview of consistent MMSE speech enhancement approaches under SPU will be given. A specific MMSE estimator with SPU is obtained, if a common MMSE estimator from Section 2.5 is coupled with a corresponding PDF-consistent *a posteriori* SPP estimator. Furthermore, it was shown in Section 2.6.3 that the *a posteriori* SPP estimator is always the same in all estimation domains. Therefore, the same *a posteriori* SPP estimators can be combined with respective common MMSE STS, MMSE STSA, or MMSE LSA estimators from Section 2.5. In Section 4.3, a performance evaluation of PDF-consistent MMSE estimation approaches under SPU will be given. To investigate the

effect of SPU estimation, respective approaches without SPU will also be taken into account. Finally, Section 4.4 will give a short summary of this chapter.

## 4.1 Synopsis of *A Posteriori* SPP Estimation

As can be seen in Section 2.6, the *a posteriori* SPP estimator is a function of both the prior  $p_S(S)$  and the likelihood  $p_{Y|S}(Y|S)$  from Chapter 2 (cf. (2.66)-(2.69)). Furthermore, since the *a posteriori* SPP (2.66) is in the same estimation framework with the common MMSE estimator  $E\{g(S)|H_1\}$  according to (2.65), they are based on the same *a priori* knowledge. Thus, it is obvious that *the same PDF assumptions should be taken both for SPU estimation and for common MMSE estimation*. Nevertheless, there are various approaches inconsistently combining different spectral PDF assumptions for the common MMSE estimator  $E\{g(S)|Y, H_1\}$  and the soft weights  $P(H_1|Y)$ , e.g., [Lotter, 2004], [Chang, 2007], [Boubakir and Berkani, 2010], and [Saha and Shimamura, 2011]. Motivated to obtain such a consistency throughout this chapter we will propose MMSE speech enhancement approaches under SPU employing consistent spectral PDF assumptions. Moreover, just as in Chapter 2, the generalization of the speech prior  $p_S(S)$  can be applied here: Employing the generalized gamma speech model, a more generalized definition of GLR can be derived. Specific GLRs and respective specific *a posteriori* SPP estimators with an underlying Rayleigh, chi, or gamma speech spectral amplitude model can be obtained by applying their parameter set to the generalized GLR.

Utilizing the bivariate generalized gamma speech prior (2.26) and a Gaussian likelihood (2.28), the likelihood of speech presence (2.69) can be reformulated as

$$p_{Y|H_1}(Y|H_1)|_{\text{g}\Gamma} = \int_{\mathbb{C}} \frac{1}{\pi\sigma_D^2} \cdot e^{-\frac{|Y-S|^2}{\sigma_D^2}} \cdot \frac{1}{2\pi} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} \cdot |S|^{\eta\nu-2} e^{-\beta|S|^\eta} dS \quad (4.1)$$

with subscript  $\text{g}\Gamma$  denoting the generalized gamma speech prior (2.26) together with its parameters  $\eta, \nu, \beta$  (cf. Table 2.1). Employing polar integration as for (2.31), as well as using (2.29) and [Gradshteyn and Ryzhik, 1965, Eq. (8.431.5)], (4.1) yields

$$p_{Y|H_1}(Y = Re^{j\theta}|H_1)|_{\text{g}\Gamma} = \underbrace{\frac{1}{\pi\sigma_D^2} \cdot e^{-\frac{|Y|^2}{\sigma_D^2}}}_{p_{Y|H_0}(Y|H_0)} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} \cdot \int_0^\infty A^{\eta\nu-1} \cdot e^{-\beta A^\eta} \cdot e^{-\frac{A^2}{\sigma_D^2}} \cdot I_0\left(\frac{2AR}{\sigma_D^2}\right) dA. \quad (4.2)$$

Please note that  $p_{Y|H_0}(Y|H_0)$  (cf. (2.68) and (2.27)) turns out to be a prefactor in (4.2) and it is simultaneously also the denominator of the GLR (2.67). Thus, after inserting (4.2) into (2.67),  $p_{Y|H_0}(Y|H_0)$  cancels out and the new GLR based on a generalized speech prior

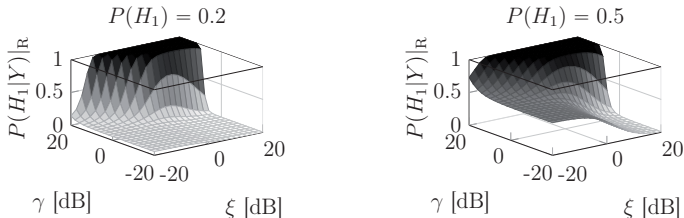


Figure 4.1: *A posteriori* SPP estimator  $P(H_1|Y)|_R$  with an underlying Rayleigh speech spectral amplitude model (Gaussian speech prior) for two different *a priori* SPPs

yields (cf. [Fodor and Fingscheidt, 2012a, Eq. (14)])

$$\Lambda_{g\Gamma} = \frac{P(H_1)}{P(H_0)} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} \cdot \int_0^\infty A^{\nu-1} \cdot e^{-\beta A^\eta} \cdot e^{-\frac{A^2}{\sigma_D^2}} \cdot I_0\left(\frac{2AR}{\sigma_D^2}\right) dA \quad (4.3)$$

where subscript  $g\Gamma$  denotes the generalized gamma speech prior with the parameters  $\eta$ ,  $\beta$ , and  $\nu$  (cf. Table 2.1). The corresponding generalized *a posteriori* SPP estimator is obtained by (2.66), i. e.,  $P(H_1|Y)|_{g\Gamma} = \Lambda_{g\Gamma}/(1 + \Lambda_{g\Gamma})$ . In the following we will derive specific *a posteriori* SPP estimators by employing speech spectral PDF models from Section 2.3 by inserting the parameter set of a specific distribution from Table 2.1 into (4.3) and then utilizing (2.66).

### Gaussian Speech Prior

Employing a *Rayleigh*-distributed speech spectral amplitude model (2.20) (Gaussian speech prior), i. e., utilizing the parameters  $\eta = 2$ ,  $\beta = 1/\sigma_S^2$ , and  $\nu = 1$  (cf. Table 2.1), as well as using [Gradshteyn and Ryzhik, 1965, Eq. (6.631.1)] results in the GLR (2.71) [Ephraim and Malah, 1984]. Employing the GLR (2.71) for (2.66) results in the *a posteriori* SPP estimate  $P(H_1|Y)|_R = \Lambda_R/(1 + \Lambda_R)$  which is depicted in Figure 4.1.

### Super-Gaussian Speech Priors

Utilizing a *chi*-distributed speech spectral amplitude model (2.22) (super-Gaussian speech prior), i. e., incorporating the parameters  $\eta = 2$ ,  $\nu = \nu_\chi$ , and  $\beta = \nu_\chi/\sigma_S^2$  (cf. Table 2.1), as well as using [Gradshteyn and Ryzhik, 1965, Eq. (6.631.1)] results in the GLR (2.72) [Breithaupt and Martin, 2011]. The corresponding *a posteriori* SPP  $P(H_1|Y)|_\chi = \Lambda_\chi/(1 + \Lambda_\chi)$  is shown in Figure 4.2. Please note that applying  $\nu_\chi = 1$  to (2.72) results in (2.71), since the chi distribution is a generalization of the Rayleigh distribution (cf. Figure 2.5).

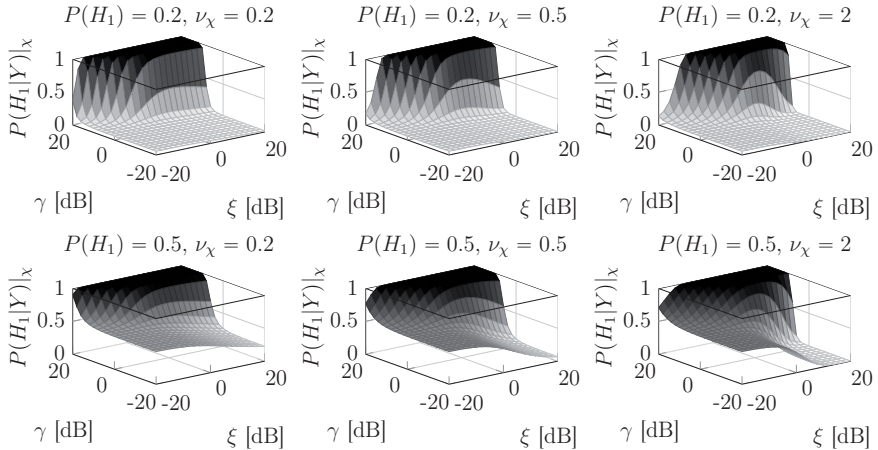


Figure 4.2: *A posteriori* SPP estimator  $P(H_1|Y)|_x$  with an underlying chi speech spectral amplitude model (super-Gaussian speech prior) for two different *a priori* SPPs (top, bottom) and three different parameters  $\nu_\chi$  (left, center, right)

Employing a *gamma*-distributed speech spectral amplitude model (2.23) (super-Gaussian speech prior), i. e., inserting the parameters  $\eta = 1$ ,  $\nu = \nu_\Gamma$ , and  $\beta = \sqrt{\nu_\Gamma(1 + \nu_\Gamma)}/\sigma_S$  (cf. Table 2.1) into the generalized GLR (4.3) results in a *new* specific GLR (cf. [Fodor and Fingscheidt, 2012a, Eq. (14)])

$$\Lambda_\Gamma = \frac{P(H_1)}{P(H_0)} \cdot \frac{[\nu_\Gamma \cdot (1 + \nu_\Gamma)]^{\frac{\nu_\Gamma}{2}}}{\Gamma(\nu_\Gamma) \cdot \sigma_S^{\nu_\Gamma}} \cdot \int_0^\infty A^{\nu_\Gamma-1} \cdot e^{-\frac{\sqrt{\nu_\Gamma(1+\nu_\Gamma)}}{\sigma_S} A} \cdot e^{-\frac{1}{\sigma_D^2} A^2} \cdot I_0\left(\frac{2AR}{\sigma_D^2}\right) dA \quad (4.4)$$

with subscript  $\Gamma$  denoting the gamma-distributed speech spectral amplitude assumption. Similar to (2.44), (2.50), and (2.62), (4.4) cannot be obtained in closed form [Fodor and Fingscheidt, 2012a]. Therefore, we employ numerical methods to calculate  $\Lambda_\Gamma$ . For convenience, however, we first reformulate (4.4) by utilizing variable substitution as for (2.45), resulting in a function of the *a priori* SNR  $\xi$  and the *a posteriori* SNR  $\gamma$  as (cf. [Fodor and Fingscheidt, 2012a, Eq. (14)])

$$\Lambda_\Gamma = \frac{P(H_1)}{P(H_0)} \cdot \frac{[\nu_\Gamma \cdot (1 + \nu_\Gamma)]^{\frac{\nu_\Gamma}{2}}}{\Gamma(\nu_\Gamma)} \cdot \left(\frac{\gamma}{\xi}\right)^{\frac{\nu_\Gamma}{2}} \cdot \int_0^\infty g^{\nu_\Gamma-1} \cdot e^{-\sqrt{\nu_\Gamma(\nu_\Gamma+1)}\sqrt{\frac{\gamma}{\xi}}g} \cdot e^{-\gamma g^2} \cdot I_0(2\gamma g) dg. \quad (4.5)$$

Employing the Gauss-Kronrod quadrature [Brass and Petras, 2011] to solve the integral in (4.5), the new *a posteriori* SPP is obtained by  $P(H_1|Y)|_\Gamma = \Lambda_\Gamma/(1 + \Lambda_\Gamma)$  which is depicted in Figure 4.3.

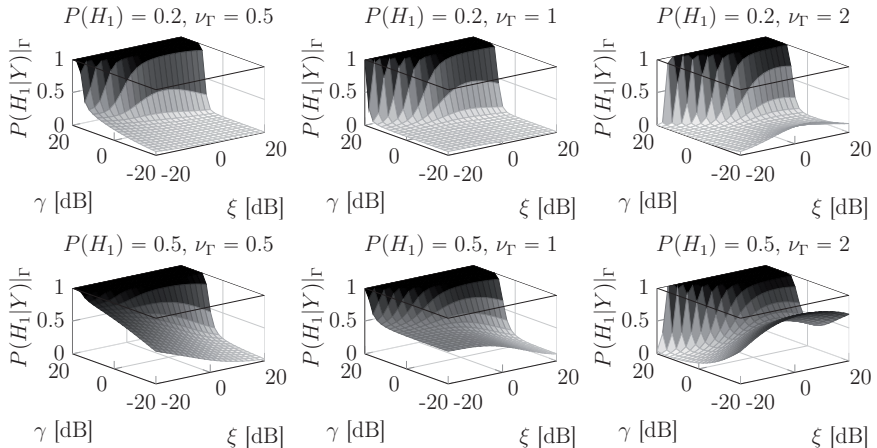


Figure 4.3: *A posteriori* SPP estimator  $P(H_1|Y)|_r$  with an underlying gamma speech spectral amplitude model (super-Gaussian speech prior) for two different *a priori* SPPs (top, bottom) and three different parameters  $\nu_T$  (left, center, right)

An overview of *a posteriori* SPP estimators w. r. t. the speech spectral amplitude assumptions is given in Table 4.1 (cf. Table 2.2). Just as in the case of MMSE speech estimation, the hierarchy of speech spectral amplitude models in 2.5 also holds for respective *a posteriori* SPP estimators.

Speech Spectral Amplitude PDF (Speech Prior)	<i>A Posteriori</i> SPP Estimator
Generalized Gamma (2.24) (Generalized Speech Prior)	(2.66) + (4.3) [new], [Fodor and Fingscheidt, 2012a]
Rayleigh (2.20) (Gaussian Speech Prior)	(2.66) + (2.71) [Ephraim and Malah, 1984]
Chi (2.22) (Super-Gaussian Speech Prior)	(2.66) + (2.72) [Breithaupt and Martin, 2011]
Gamma (2.23) (Super-Gaussian Speech Prior)	(2.66) + (4.5) [new], [Fodor and Fingscheidt, 2012a]

Table 4.1: Overview of *a posteriori* SPP estimators w. r. t. speech spectral amplitude assumptions

## 4.2 Synopsis of Consistent MMSE Estimation Under SPU

A PDF-consistent MMSE estimation approach under SPU can be obtained by combining a common MMSE estimator from Table 2.2 with a PDF-consistent *a posteriori* SPP estimator from Table 4.1, according to (2.80), (2.81), and (2.86) in the STS, STSA, and (OM) LSA estimation domain, respectively (cf. Section 2.6.3). An overview of MMSE estimation under SPU based on a consistent PDF assumption for both the common MMSE estimator  $E\{g(S)|Y, H_1\}$  and the *a posteriori* SPP estimator  $P(H_1|Y)$  is given in Table 4.2. In the following, we will revisit each table entry, separated by estimation domains.

### 4.2.1 MMSE STS Estimation Under SPU

In this section, specific MMSE STS estimators under SPU of the form (2.80) are employed.

#### Generalized Speech Prior

Our new generalized MMSE STS estimator under SPU uses the generalized gamma speech model (2.26) (generalized speech prior) consistently for the common MMSE STS estimator (2.39) and the *a posteriori* SPP estimator (2.66) with the GLR (4.3) and will be denoted as  $g\Gamma$ -STS-SPU (cf. Table 4.2). Its equivalent non-SPU approach in Table 2.2 is  $g\Gamma$ -STS.

#### Gaussian Speech Prior

The first special case of the  $g\Gamma$ -STS-SPU estimator employs a *Rayleigh*-distributed speech spectral amplitude model (Gaussian speech prior) consistently for the common MMSE STS estimator being the Wiener filter (2.41) [Scalart and Filho, 1996] and the *a posteriori* SPP estimator (2.66) with the GLR (2.71) [Ephraim and Malah, 1984] and will be denoted as R-STS-SPU (cf. Table 4.2) [Azirani et al., 1996]. Its equivalent non-SPU approach in Table 2.2 is R-STS.

#### Super-Gaussian Speech Priors

The second special case of the  $g\Gamma$ -STS-SPU estimator utilizes a *chi*-distributed speech spectral amplitude model (super-Gaussian speech prior) consistently for the common MMSE STS estimator (2.43) [Erkelens et al., 2008] and the *a posteriori* SPP estimator (2.66) with the GLR (2.72) [Breithaupt and Martin, 2011] and will be denoted as  $\chi$ -STS-SPU (cf. Table 4.2). Its equivalent non-SPU approach in Table 2.2 is  $\chi$ -STS.

Speech Spectral Amplitude PDF (Speech Prior)	Estimation Domain		
	STS-SPU (2.80)	STSA-SPU (2.81)	(OM) LSA (2.86)
Generalized Gamma (2.24) (Generalized Speech Prior)	$g\Gamma$ -STS-SPU $(2.39) + (2.66) + (4.3)$ [new]	$g\Gamma$ -STSA-SPU $(2.47) + (2.66) + (4.3)$ [new], [Fodor and Fingscheidt, 2012a]	$g\Gamma$ -LSA-SPU $(2.57) + (2.66) + (4.3)$ [new], [Fodor and Fingscheidt, 2012d]
Rayleigh (2.20) (Gaussian Speech Prior)	R-STS-SPU $(2.41) + (2.66) + (2.71)$ [Azirani et al., 1996]	R-STSA-SPU $(2.48) + (2.66) + (2.71)$ [Ephraim and Malah, 1984]	R-LSA-SPU $(2.58) + (2.66) + (2.71)$ [Cohen and Berlugo, 2001]
Chi (2.22) (Super-Gaussian Speech Prior)	$\chi$ -STS-SPU $(2.43) + (2.66) + (2.72)$ [new]	$\chi$ -STSA-SPU $(2.49) + (2.66) + (2.72)$ [Breithaupt and Martin, 2011]	$\chi$ -LSA-SPU $(2.59) + (2.66) + (2.72)$ [new]
Gamma (2.23) (Super-Gaussian Speech Prior)	$\Gamma$ -STS-SPU $(2.45) + (2.66) + (4.5)$ [new]	$\Gamma$ -STSA-SPU $(2.51) + (2.66) + (4.5)$ [new], [Fodor and Fingscheidt, 2012a]	$\Gamma$ -LSA-SPU $(2.63) + (2.66) + (4.5)$ [new], [Fodor and Fingscheidt, 2012d]

Table 4.2: Overview of MMSE speech enhancement approaches under SPU based on a consistent speech model for both the common MMSE estimator and the *a posteriori* SPP estimator

The third special case of the  $g\Gamma$ -STS-SPU estimator is based on a consistent *gamma*-distributed speech spectral amplitude model (super-Gaussian speech prior) for both the common MMSE STS estimator (2.45) [Erkelens et al., 2008] and the *a posteriori* SPP estimator (2.66) with the GLR (4.5) [Fodor and Fingscheidt, 2012a] and will be denoted as  $\Gamma$ -STS-SPU (cf. Table 4.2). Its equivalent non-SPU approach in Table 2.2 is  $\Gamma$ -STS.

To the best of our knowledge, neither the generalized approach  $g\Gamma$ -STS-SPU, nor the super-Gaussian approaches  $\chi$ -STS-SPU and  $\Gamma$ -STS-SPU have been proposed yet.

### 4.2.2 MMSE STSA Estimation Under SPU

In this section, specific MMSE STSA estimators under SPU of the form (2.81) are employed.

#### Generalized Speech Prior

Our new generalized MMSE STSA estimator under SPU employs the generalized speech model (2.26) (generalized speech prior) consistently for the common MMSE STSA estimator (2.47) [Fodor and Fingscheidt, 2012a] and the *a posteriori* SPP estimator (2.66) with the GLR (4.3) [Fodor and Fingscheidt, 2012a] and will be denoted as  $g\Gamma$ -STSA-SPU (cf. Table 4.2) [Fodor and Fingscheidt, 2012a]. Its equivalent non-SPU approach in Table 2.2 is  $g\Gamma$ -STSA.

#### Gaussian Speech Prior

The first special case of the  $g\Gamma$ -STSA-SPU estimator assumes a *Rayleigh*-distributed speech spectral amplitude model (Gaussian speech prior) consistently for the common MMSE STSA estimator (2.48) [Ephraim and Malah, 1984] and the *a posteriori* SPP estimator (2.66) with the GLR (2.71) [Ephraim and Malah, 1984] and will be denoted as R-STSA-SPU (cf. Table 4.2) [Ephraim and Malah, 1984]. Its equivalent non-SPU approach in Table 2.2 is R-STSA.

#### Super-Gaussian Speech Priors

The second special case of the  $g\Gamma$ -STSA-SPU estimator is based on a consistent *chi*-distributed speech spectral amplitude model (super-Gaussian speech prior) for both the common MMSE STSA estimator (2.49) [Andrianakis and White, 2006] and the *a posteriori* SPP estimator (2.66) with the GLR (2.72) [Breithaupt and Martin, 2011] and will be denoted as  $\chi$ -STSA-SPU (cf. Table 4.2) [Breithaupt and Martin, 2011]. Its equivalent non-SPU approach in Table 2.2 is  $\chi$ -STSA.

A further proposed approach is the third special case of the  $g\Gamma$ -STSA-SPU estimator. It utilizes a *gamma*-distributed speech spectral amplitude model (super-Gaussian speech prior) consistently for the common MMSE STSA estimator (2.51) [Erkelens et al., 2007] and the *a posteriori* SPP estimator (2.66) with the GLR (4.5) [Fodor and Fingscheidt, 2012a] and will be denoted as  $\Gamma$ -STSA-SPU (cf. Table 4.2) [Fodor and Fingscheidt, 2012a]. Its equivalent non-SPU approach in Table 2.2 is  $\Gamma$ -STSA.

### 4.2.3 MMSE LSA Estimation Under SPU

In this section, specific OM MMSE LSA estimators of the form (2.86) are employed.

#### Generalized Speech Prior

Our new generalized OM MMSE LSA estimator utilizes the generalized speech model (2.26) (generalized speech prior) consistently for the common MMSE LSA estimator (2.57) [Fodor and Fingscheidt, 2012d] and the *a posteriori* SPP estimator (2.66) with the GLR (4.3) [Fodor and Fingscheidt, 2012a] and will be denoted as  $g\Gamma$ -LSA-SPU (cf. Table 4.2) [Fodor and Fingscheidt, 2012d]. Its equivalent non-SPU approach in Table 2.2 is  $g\Gamma$ -LSA.

#### Gaussian Speech Prior

The first special case of the  $g\Gamma$ -LSA-SPU estimator uses a *Rayleigh*-distributed speech spectral amplitude model (Gaussian speech prior) consistently for the common MMSE LSA estimator (2.58) [Ephraim and Malah, 1985] and the *a posteriori* SPP estimator (2.66) with the GLR (2.71) [Ephraim and Malah, 1984] and will be denoted as R-LSA-SPU (cf. Table 4.2) [Cohen and Berdugo, 2001]. Its equivalent non-SPU approach in Table 2.2 is R-LSA.

#### Super-Gaussian Speech Priors

The second special case of the  $g\Gamma$ -LSA-SPU estimator is a further proposal. It is based on a consistent *chi*-distributed speech spectral amplitude model (super-Gaussian speech prior) for both the common MMSE LSA estimator (2.59) [Hendriks et al., 2009b] (implemented as (2.61)) and the *a posteriori* SPP estimator (2.66) with the GLR (2.72) [Breithaupt and Martin, 2011] and will be denoted as  $\chi$ -LSA-SPU (cf. Table 4.2). Its equivalent non-SPU approach in Table 2.2 is  $\chi$ -LSA.

A further proposed approach is the third special case of the  $g\Gamma$ -LSA-SPU estimator [Fodor and Fingscheidt, 2012d]. It employs a *gamma*-distributed speech spectral amplitude model (super-Gaussian speech prior) consistently for the common MMSE LSA es-

imator [Borgström and Alwan, 2011] (implemented as (2.63)) and the *a posteriori* SPP estimator (2.66) with the GLR (4.5) [Fodor and Fingscheidt, 2012a] and will be denoted as  $\Gamma$ -LSA-SPU (cf. Table 4.2) [Fodor and Fingscheidt, 2012d]. Its equivalent non-SPU approach in Table 2.2 is  $\Gamma$ -LSA.

### 4.3 Performance Evaluation

In the following, we will evaluate the performance of PDF-consistent MMSE estimation approaches under SPU from Table 4.2 using the simulation setup and the instrumental measures from Chapter 3. Furthermore, to see the effect of SPU estimation, we will also compare the estimators under SPU to respective approaches from Table 2.2 without taking SPU into account.

It can be seen in [Andrianakis and White, 2006] and [Erkelens et al., 2007] that the best choice for the shape parameter  $\nu_\chi$  of the chi distribution as a speech spectral amplitude model are values below 1. Therefore, all evaluated approaches based on a chi-distributed speech spectral amplitude model employ the shape parameter  $\nu_\chi = 0.5$ , which is shown in [Breithaupt and Martin, 2011] to achieve a good tradeoff between speech distortion and noise attenuation. All evaluated approaches with an underlying gamma-distributed speech spectral amplitude model utilize the shape parameter  $\nu_\Gamma = 1.126$ , as proposed<sup>1</sup> in [Lotter and Vary, 2005]. Furthermore, for OM MMSE LSA approaches (2.86)  $G_0 = -25$  dB was used as proposed in [Cohen and Berdugo, 2001].

#### MMSE STS Estimation

The evaluation results of MMSE STS estimation approaches are depicted in Figure 4.4. In the upper, middle, and lower figures the resulting  $\text{NA}_{\text{seg}}$ ,  $\text{SSDR}_{\text{seg}}$ , and LKR values are shown, respectively. The noise conditions are separated column-wise corresponding to car noise, factory noise, and babble noise.

In terms of residual noise levels, approaches with SPU (R-STs-SPU,  $\chi$ -STs-SPU,  $\Gamma$ -STs-SPU) show a significant larger amount of noise attenuation reflected by larger  $\text{NA}_{\text{seg}}$  values, compared to approaches without SPU (R-STs,  $\chi$ -STs,  $\Gamma$ -STs) at all input SNR levels and in all noise conditions. In the case of car noise, approaches without SPU achieve approximately the same  $\text{NA}_{\text{seg}}$  values in the range 26..28 dB. Considering approaches with SPU,  $\Gamma$ -STs-SPU achieves slightly larger  $\text{NA}_{\text{seg}}$  values than  $\chi$ -STs-SPU, while R-STs-SPU performs best in this test. In presence of factory noise, approaches without SPU show the same ranking order as in presence of car noise. Considering approaches with SPU, however, the approach  $\Gamma$ -STs-

<sup>1</sup>Please note that the shape parameter ' $\nu$ ' in [Lotter and Vary, 2005] equals  $\nu_\Gamma - 1$  [Erkelens et al., 2007].

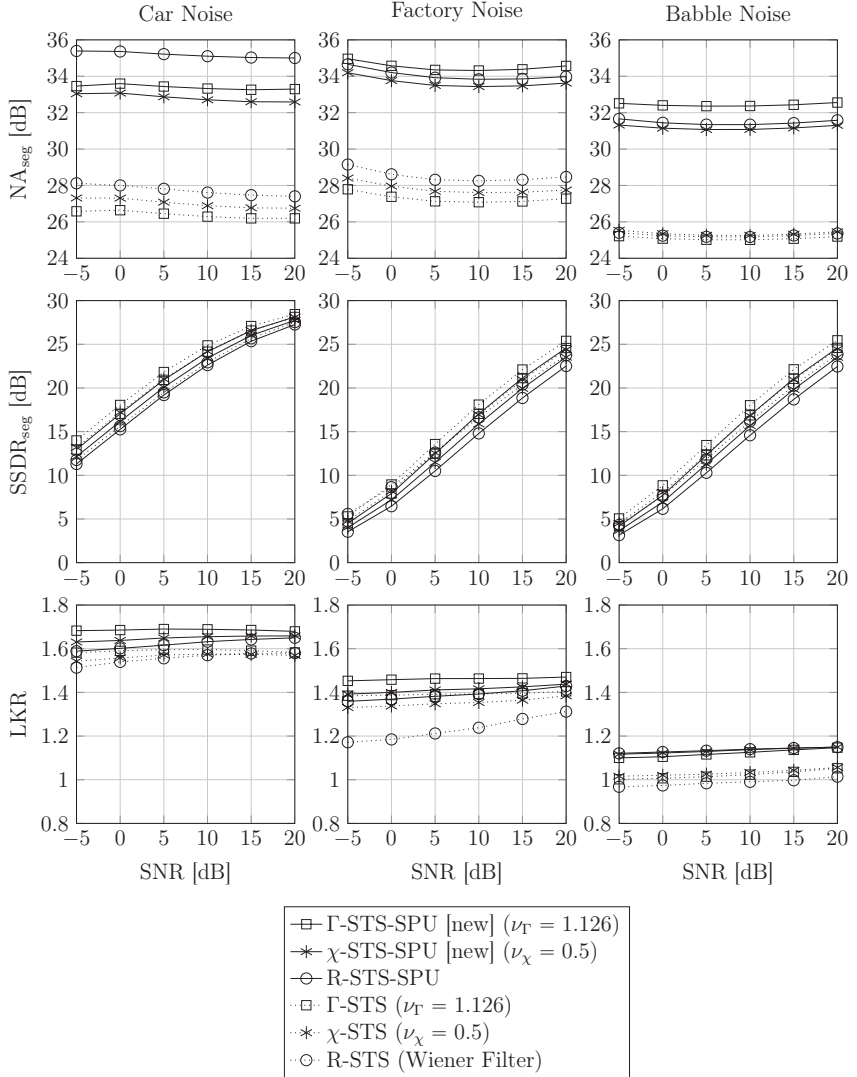


Figure 4.4: Performance evaluation results of MMSE STS approaches (R-STIS,  $\chi$ -STIS,  $\Gamma$ -STIS) and PDF-consistent MMSE STS approaches under SPU (R-STIS-SPU,  $\chi$ -STIS-SPU,  $\Gamma$ -STIS-SPU) in terms of segmental noise attenuation ( $NA_{\text{seg}}$ , the larger the better), segmental speech-to-speech distortion ratio ( $SSDR_{\text{seg}}$ , the larger the better), and weighted log-kurtosis ratio (LKR, the smaller the better) (cf. Table 4.2)

SPU performs best, followed by the approaches  $\chi$ -STS-SPU and R-STS-SPU. This ranking order is also true for approaches with SPU in presence of babble noise, while all approaches without SPU (R-STS,  $\chi$ -STS,  $\Gamma$ -STS) achieve approximately the same  $\text{NA}_{\text{seg}}$  values.

Considering the speech component, all approaches achieve approximately the same speech component quality reflected by close  $\text{SSDR}_{\text{seg}}$  values with an expected slight advantage for approaches without SPU. Moreover, approaches with SPU (without SPU) based on super-Gaussian speech priors perform slightly better than approaches with SPU (without SPU) based on a Gaussian speech prior. Furthermore, approaches with a gamma speech spectral amplitude assumption show slightly larger  $\text{SSDR}_{\text{seg}}$  values compared to approaches with a chi speech spectral amplitude assumption which have a slight advantage over approaches with a Gaussian speech prior for all input SNR levels and in all noise conditions.

It can be observed in the evaluation w. r. t. the musical noise level that approaches with SPU generate slightly more musical noise than approaches without SPU reflected by larger LKR values. Furthermore, in case of car noise and factory noise, approaches based on a gamma speech spectral amplitude model produce the largest musical noise levels reflected by the largest LKR values, while approaches based on a Gaussian speech prior achieve the lowest musical noise levels reflected by the smallest LKR values. In the presence of babble noise, all approaches with SPU achieve approximately the same LKR levels, while among the approaches without SPU  $\Gamma$ -STS and  $\chi$ -STS perform approximately equally and R-STS produces the lowest musical noise level (lowest LKR value).

### MMSE STSA Estimation

The evaluation results of MMSE STSA estimation approaches are plotted in Figure 4.5. As in the case of MMSE STS estimation, the resulting  $\text{NA}_{\text{seg}}$ ,  $\text{SSDR}_{\text{seg}}$ , and LKR values are shown in the upper, middle, and lower figures, respectively. Furthermore, car noise, factory noise, and babble noise results can be found in the first, second, and third column of plots, respectively.

Regarding the amount of residual noise, approaches with SPU (R-STSA-SPU,  $\chi$ -STSA-SPU,  $\Gamma$ -STSA-SPU) achieve a significantly larger amount of noise attenuation reflected by larger  $\text{NA}_{\text{seg}}$  values, compared to approaches without SPU (R-STSA,  $\chi$ -STSA,  $\Gamma$ -STSA). Furthermore, the  $\Gamma$ -STSA-SPU ( $\Gamma$ -STSA) estimator outperforms the approaches  $\chi$ -STSA-SPU ( $\chi$ -STSA) and R-STSA-SPU (R-STSA) at all input SNRs and in all noise conditions. Furthermore, while the approach  $\chi$ -STSA clearly outperforms the approach R-STSA, the approaches R-STSA-SPU and  $\chi$ -STSA-SPU perform approximately equally.

Concerning the amount of speech component quality, approaches without SPU perform again slightly better than approaches with SPU reflected by slightly larger  $\text{SSDR}_{\text{seg}}$  values

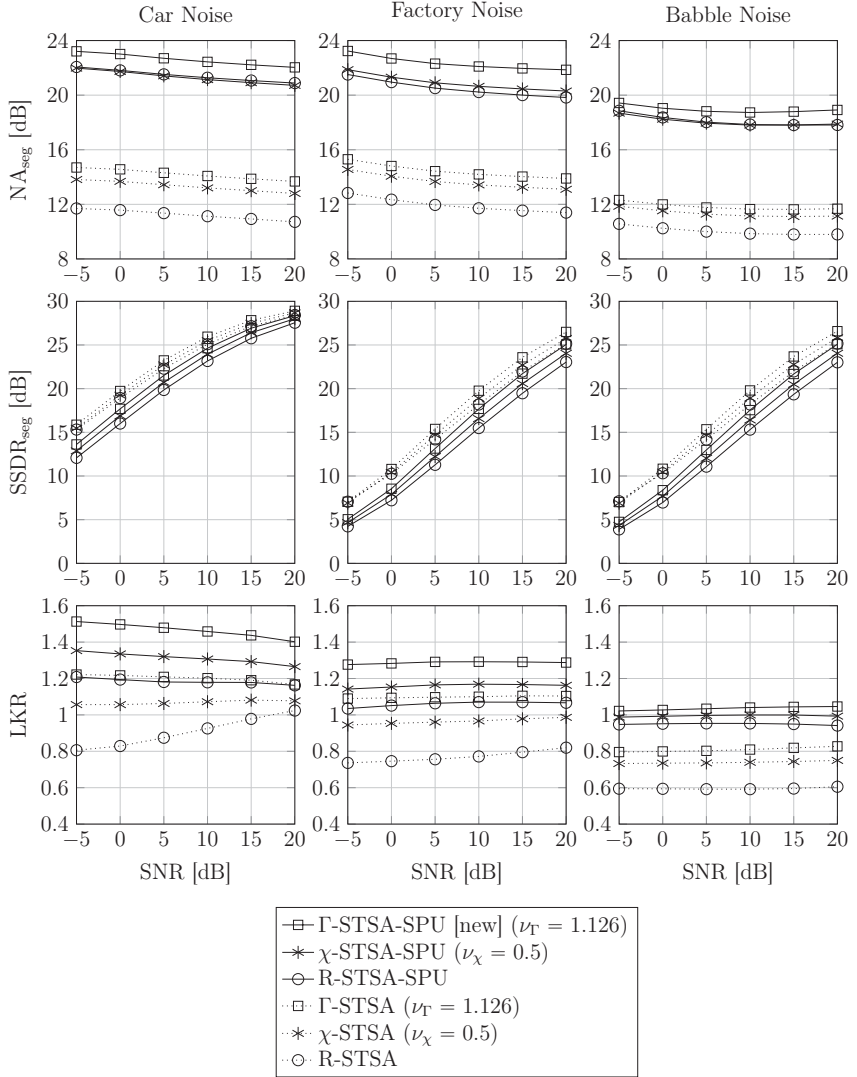


Figure 4.5: Performance evaluation results of MMSE STSA approaches (R-STSA,  $\chi$ -STSA,  $\Gamma$ -STSA) and PDF-consistent MMSE STSA approaches under SPU (R-STSA-SPU,  $\chi$ -STSA-SPU,  $\Gamma$ -STSA-SPU) in terms of segmental noise attenuation ( $NA_{\text{seg}}$ , the larger the better), segmental speech-to-speech distortion ratio ( $SSDR_{\text{seg}}$ , the larger the better), and weighted log-kurtosis ratio (LKR, the smaller the better) (cf. Table 4.2)

at all input SNRs and in all noise conditions. Just as in MMSE STS estimation, super-Gaussian speech models achieve a better speech component quality than Gaussian ones; a gamma speech spectral amplitude assumption is slightly superior to a chi speech spectral amplitude assumption reflected by slightly larger  $\text{SSDR}_{\text{seg}}$  values.

With respect to the amount of musical noise, approaches with SPU reveal larger LKR levels than respective approaches without SPU. Furthermore, approaches with a gamma speech spectral amplitude model produce higher musical noise levels (larger LKR values) than approaches with a chi speech spectral amplitude model. The lowest musical noise level (smallest LKR value) is yielded by approaches with a Gaussian speech prior at all SNR levels and in all noise conditions.

### MMSE LSA Estimation

The evaluation results of MMSE LSA estimation approaches are shown in Figure 4.6. Just as in the previous figures, the resulting  $\text{NA}_{\text{seg}}$ ,  $\text{SSDR}_{\text{seg}}$ , and LKR values are shown in the upper, middle, and lower figures, respectively. Meanwhile, car noise, factory noise, and babble noise results can be found in the first, second, and third column of plots, respectively.

Just as in the previous two estimation domains, approaches with SPU (R-LSA-SPU,  $\chi$ -LSA-SPU,  $\Gamma$ -LSA-SPU) achieve a significantly larger amount of noise attenuation reflected by larger  $\text{NA}_{\text{seg}}$  values than approaches without SPU (R-LSA,  $\chi$ -LSA,  $\Gamma$ -LSA), as shown in the upper figures. Furthermore, similar to MMSE STSA estimation, approaches based on a gamma speech spectral amplitude model perform better than respective approaches with a chi speech spectral amplitude model, while the latter achieve larger  $\text{NA}_{\text{seg}}$  values than respective approaches with a Gaussian speech prior.

As can be seen in the middle figures, approaches without SPU achieve slightly better speech component quality than approaches with SPU reflected by slightly larger  $\text{SSDR}_{\text{seg}}$  values at all input SNRs and in all noise conditions. Furthermore, a gamma speech spectral amplitude model gives the best speech preservation, followed by the chi and Rayleigh speech spectral amplitude models.

The lower figures show that approaches with SPU produce larger musical noise levels than respective approaches without SPU reflected by larger LKR values. Furthermore, approaches with a gamma speech spectral amplitude model show higher musical noise levels (larger LKR values) than approaches with a chi speech spectral amplitude model. The lowest musical noise level (smallest LKR value) yield approaches with a Gaussian speech prior at all SNR levels and in all noise conditions.

The performance evaluation results were also confirmed by informal listening tests.

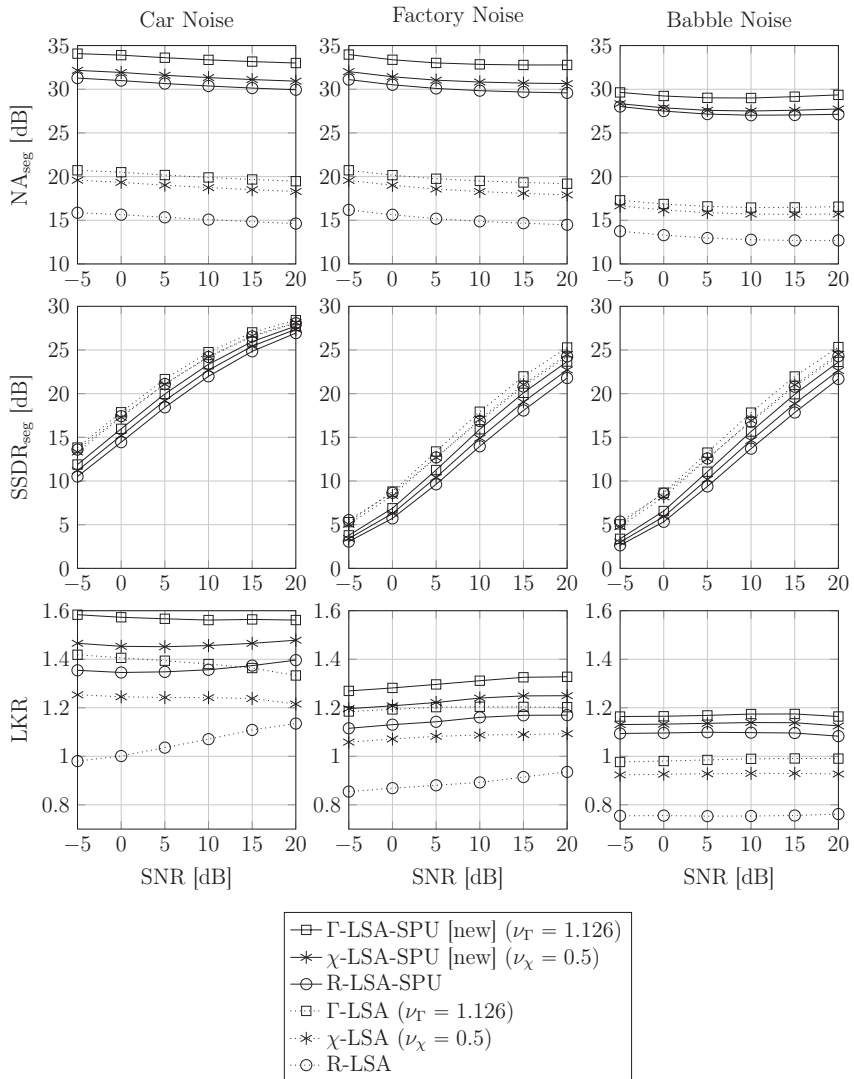


Figure 4.6: Performance evaluation results of MMSE LSA approaches (R-LSA,  $\chi$ -LSA,  $\Gamma$ -LSA) and PDF-consistent MMSE LSA approaches under SPU (R-LSA-SPU,  $\chi$ -LSA-SPU,  $\Gamma$ -LSA-SPU) in terms of segmental noise attenuation ( $NA_{\text{seg}}$ , the larger the better), segmental speech-to-speech distortion ratio ( $SSDR_{\text{seg}}$ , the larger the better), and weighted log-kurtosis ratio (LKR, the smaller the better) (cf. Table 4.2)

## 4.4 Summary

MMSE estimation under SPU can clearly be divided into common MMSE estimation for speech and a *posteriori* SPP estimation. It was shown in this chapter that both are based on the same *a priori* knowledge, therefore, the common MMSE estimator and the *a posteriori* SPP estimator should assume the same statistical models. In order to identify PDF-consistent MMSE speech enhancement approaches under SPU, a speech PDF-based overview of *a posteriori* SPP estimators was given in this chapter, similar to the synopsis of common MMSE estimators in Chapter 2. The overview was based on a new parametric *a posteriori* SPP estimator which is a generalization of the existing *a posteriori* SPP estimators from literature, based on either a Rayleigh-distributed or a chi-distributed speech spectral amplitude model. Moreover, a new specific *a posteriori* SPP estimator was derived, based on a gamma-distributed speech spectral amplitude model.

Furthermore, a synopsis of PDF-consistent MMSE approaches under SPU was given in this chapter, in analogy to Section 2.5. By sketching possible MMSE estimators under SPU, new PDF-consistent approaches could be identified. Then, both MMSE speech enhancement approaches with SPU and respective approaches without SPU from Section 2.5 were assessed using instrumental measures. It turned out that for all estimation domains and for almost all noise types that compared to non-SPU, approaches with SPU achieve much higher segmental noise attenuation at the cost of a slightly lower segmental speech-to-speech distortion ratio and higher musical noise levels reflected by larger LKR values.

Approaches assuming a gamma distribution for the speech spectral amplitudes showed best performance w. r. t. segmental speech-to-speech distortion ratio and segmental noise attenuation except for car noise and STS estimation domain. However, these approaches performed worst in terms of musical noise levels, reflected by the highest LKR values. In the next chapter, we will deal with PDF-consistent MMSE estimation approaches under SPU which can produce less musical noise and provide more accurate *a posteriori* SPP estimates.

# Chapter 5

## Consistent MMSE Estimation Under Speech Presence Uncertainty with Averaging and Fixed Parameters

Amongst others, recent improvements in SPU estimation include *a posteriori* SPP estimators based on either averaged observations or on a super-Gaussian speech model instead of a Gaussian. In this chapter, these two aspects will be combined [Fodor and Gerkmann, 2014a], [Fodor and Gerkmann, 2014b]. Furthermore, the resulting *enhanced a posteriori* SPP estimator will make use of the advantage of fixed prior parameters, i. e., a fixed *a priori* SPP and a fixed *a priori* SNR. This provides a more accurate *a posteriori* SPP estimation compared to usual techniques based on a fixed *a priori* SPP and an adapted *a priori* SNR. Moreover, similar to the previous chapter, we will propose to combine the enhanced *a posteriori* SPP estimator with a PDF-consistent common MMSE estimator in order to obtain an MMSE speech enhancement approach under SPU.

In Section 5.1, we will shortly recapitulate the motivation of enhanced *a posteriori* SPP estimation using averaged observations, a super-Gaussian assumption for speech, and fixed prior parameters. Then, Section 5.2 will give a detailed description of the algorithmic approach. In Section 5.3, the performance of the combination of the enhanced *a posteriori* SPP estimator and a PDF-consistent common MMSE estimator will be evaluated. Finally, Section 5.4 will give a short summary of this chapter.

### 5.1 Introduction

This chapter deals with an enhancement scheme of *a posteriori* SPP estimators based on the idea introduced in [Gerkmann et al., 2008], [Gerkmann, 2010, Ch. 6] and recapitulated in

Section 2.6.2. In particular, we will focus on a typical disadvantage of usual SPU estimation: *A posteriori* SPP estimators being a function of the noisy observations (cf., e. g., (2.66) and (2.71)) usually suffer from random fluctuations of the noisy speech. This often results in estimation outliers which may be perceived as annoying musical noise. Estimation outliers can successfully be reduced by averaging the observations as proposed in [Gerkmann et al., 2008].

However, the proposal in [Gerkmann et al., 2008] does not take into account the super-Gaussian nature of speech STFT coefficients (cf. Chapter 2). Therefore, in this chapter we will present a new *a posteriori* SPP estimator for averaged observations, assuming a super-Gaussian speech model. The derivation of the corresponding GLR turns out to be mathematically complex, therefore, only approximate solutions have been proposed yet [Fodor and Gerkmann, 2014a]. In this chapter, however, we will give a closed-form solution for the GLR and, therefore, the corresponding *a posteriori* SPP estimator [Fodor and Gerkmann, 2014b].

While usual *a posteriori* SPP estimators are able to robustly achieve values close to one during speech presence, they typically output the *a priori* SPP in absence of speech, which is generally not close to zero (cf. Section 2.6.2). To overcome this issue, a tracking algorithm is proposed in [Cohen and Berdugo, 2001] which achieves small *a priori* SPPs and, thus, small (close to zero) *a posteriori* SPP values during speech absence. As a different solution, in [Gerkmann et al., 2008] a fixed *a priori* SPP and a fixed *a priori* SNR are proposed for *a posteriori* SPP estimation arguing that these quantities should reflect true *a priori* knowledge and should be independent of the observations. As a result, the *a posteriori* SPP estimates robustly achieve values close to zero in absence of speech, providing potentially more accurate estimation results. Our enhanced *a posteriori* SPP estimator will also make use of the advantage of fixed prior parameters.

## 5.2 Algorithmic Approach

The proposed approach, just as the reference approach recapitulated in Section 2.6.2 [Gerkmann et al., 2008], consists of two modifications of usual SPU estimation approaches (e. g., those from Section 2.6.1 or from Chapter 4), namely, an *a posteriori* SNR averaging and a fixed *a priori* SNR instead of an adapted one. These two modification steps will be described in the following.

First, we will derive the new GLR assuming a super-Gaussian speech model and averaged observations. It will be assumed for the derivation that the speech STFT coefficients are super-Gaussian distributed with chi-distributed speech spectral amplitudes (2.22) and a statistically independent uniformly distributed phase. Furthermore, just as in [Gerkmann

et al., 2008], we will assume that the statistically independent noise STFT coefficients are Gaussian distributed and that the hypothesis  $H_1$  is statistically independent of the spectral phase of the noisy speech  $Y$  (cf., e. g., (2.71) or (2.72)). In this case the *a posteriori* SPP can be rewritten as  $P(H_1|Y) = P(H_1|\gamma)$  (cf. Section 2.6.2). Then, the averaging procedure (2.73) will be taken into account and the corresponding new *a posteriori* SPP estimator  $P(H_1|\bar{\gamma})$  will be derived under a super-Gaussian speech assumption for observed, averaged *a posteriori* SNRs  $\bar{\gamma}$ . Finally, the fixed *a priori* SNR is optimized using the resulting new *a posteriori* SPP  $P(H_1|\bar{\gamma})$ .

### 5.2.1 A Posteriori SNR Averaging

In this chapter we employ the same averaging framework as in [Gerkmann et al., 2008] to reduce estimation outliers: The *a posteriori* SNRs are smoothed by calculating the moving average (2.73) simultaneously within both a local and a global averaging window of the sizes  $\mu_\Theta = |\mathbb{K}_\Theta| \cdot |\mathbb{L}_\Theta|$ . Here,  $\Theta$  stands for either local or global. Please note that we utilize the proposed window sizes from [Gerkmann et al., 2008]. The *a posteriori* SPP is obtained by multiplying the local and the global *a posteriori* SPP estimates (2.74) which are driven by the averaged *a posteriori* SNRs (cf. (2.75) and (2.76)). The local and the global *a posteriori* SPP can be calculated by the GLR (cf. (2.75))

$$P(H_1|\bar{\gamma}_\Theta)|_\chi = \frac{\bar{\Lambda}_\Theta^\chi}{1 + \bar{\Lambda}_\Theta^\chi} \quad (5.1)$$

where index  $\chi$  denotes our chi-distributed speech spectral amplitude assumption. The GLR in (5.1) is obtained by (cf. (2.76))

$$\bar{\Lambda}_\Theta^\chi = \frac{P(H_1)}{P(H_0)} \cdot \frac{p_{\bar{\gamma}_\Theta|H_1}^\chi(\bar{\gamma}_\Theta|H_1)}{p_{\bar{\gamma}_\Theta|H_0}^\chi(\bar{\gamma}_\Theta|H_0)} \quad (5.2)$$

with superscript  $\chi$  denoting the assumption of chi-distributed speech spectral amplitudes. Furthermore,  $p_{\bar{\gamma}_\Theta|H_1}^\chi(\bar{\gamma}_\Theta|H_1)$  and  $p_{\bar{\gamma}_\Theta|H_0}^\chi(\bar{\gamma}_\Theta|H_0)$  are the PDFs of averaged *a posteriori* SNRs  $\bar{\gamma}_\Theta$  assuming speech presence and absence, respectively, which will be derived in the following.

Considering the likelihood of speech presence  $p_{\bar{\gamma}_\Theta|H_1}^\chi(\bar{\gamma}_\Theta|H_1)$  for averaged observations, we start its derivation with obtaining the PDF of noisy speech STFT coefficients  $p_{Y|H_1}^\chi(Y|H_1)$ . For this, we assume a chi-distributed speech spectral amplitude model and a statistically independent uniformly distributed speech spectral phase (super-Gaussian speech prior) as well as a statistically independent Gaussian acoustic channel noise. Next, the PDF  $p_{Y|H_1}^\chi(Y|H_1)$  is transformed to the PDF of the *a posteriori* SNR  $p_{\gamma|H_1}^\chi(\gamma|H_1)$  according to the variable transformation  $\gamma = |Y|^2/\sigma_D^2$ . Then, the PDF of averaged (i. e., summed and normalized) *a*

*a posteriori* SNRs  $p_{\bar{\gamma}|H_1}^x(\bar{\gamma}|H_1)$  can be derived by means of convolution and mathematical induction: The PDF of two statistically independent *a posteriori* SNRs is obtained by convolving the PDF  $p_{\gamma|H_1}^x(\gamma|H_1)$  by itself. Convolution of the resulting PDF again by  $p_{\gamma|H_1}^x(\gamma|H_1)$  results in the PDF of the sum of three *a posteriori* SNRs and so forth. Using mathematical induction, the likelihood of speech presence for averaged observations under a chi-distributed speech spectral amplitude assumption turns out to be (the derivation can be found in Appendix D) [Fodor and Gerkmann, 2014b]

$$p_{\bar{\gamma}_\Theta|H_1}^x(\bar{\gamma}_\Theta|H_1) = \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\mu_\Theta \nu_\chi} \cdot \frac{\mu_\Theta^{\mu_\Theta}}{\Gamma(\mu_\Theta)} \cdot \bar{\gamma}_\Theta^{\mu_\Theta - 1} \cdot e^{-\mu_\Theta \bar{\gamma}_\Theta} \cdot {}_1F_1 \left( \mu_\Theta \nu_\chi, \mu_\Theta, \bar{\gamma}_\Theta \frac{\mu_\Theta \xi}{\nu_\chi + \xi} \right) \quad (5.3)$$

with  $\nu_\chi$  and  $\mu_\Theta$  being the shape parameter of the speech model (cf. Table 2.1) and the number of averaged *a posteriori* SNR values ( $\mu_\Theta = |\mathbb{K}_\Theta| \cdot |\mathbb{L}_\Theta|$ , cf. (2.73)), respectively. Please note that in the special case  $\nu_\chi = 1$  the new likelihood (5.3) based on a chi-distributed speech spectral amplitude model reduces to (2.77) being based on a Gaussian speech model. Generally, the more *a posteriori* SNR values are averaged, the larger the parameter  $\mu_\Theta$ , and the smaller the width of the resulting likelihood (5.3). Please note that different from the reference approach [Gerkmann et al., 2008], we assume uncorrelated adjacent *a posteriori* SNR values within the averaging windows, i. e., we employ  $\mu_\Theta = |\mathbb{K}_\Theta| \cdot |\mathbb{L}_\Theta|$ .

In speech absence  $\xi = 0$  holds and (5.3) reduces to the likelihood of speech absence for averaged observations

$$p_{\bar{\gamma}_\Theta|H_0}^x(\bar{\gamma}_\Theta|H_0) = \frac{\mu_\Theta^{\mu_\Theta}}{\Gamma(\mu_\Theta)} \cdot \bar{\gamma}_\Theta^{\mu_\Theta - 1} \cdot e^{-\mu_\Theta \bar{\gamma}_\Theta} \quad (5.4)$$

with  $\mu_\Theta$  being the same parameter as in (5.3). Please note that the likelihood of speech absence for averaged observations (5.4) can also be obtained as follows: In speech absence, the noisy speech STFT coefficients are Gaussian-distributed, the *a posteriori* SNRs are exponentially distributed, and the averaged *a posteriori* SNRs turn out to be gamma-distributed with the PDF (5.4) being a function of parameter  $\mu_\Theta$  (cf. (2.23), (2.78), and (5.4)).

Substituting the likelihoods (5.3) and (5.4) into (5.2), the proposed GLR for averaged observations assuming a chi-distributed speech spectral amplitude model (super-Gaussian prior) can be expressed as [Fodor and Gerkmann, 2014b]

$$\bar{\Lambda}_\Theta^x = \frac{P(H_1)}{P(H_0)} \cdot \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\mu_\Theta \nu_\chi} \cdot {}_1F_1 \left( \mu_\Theta \nu_\chi, \mu_\Theta, \bar{\gamma}_\Theta \frac{\mu_\Theta \xi}{\nu_\chi + \xi} \right) \quad (5.5)$$

with  $\nu_\chi$  being the shape parameter of the speech model (which should be chosen consistently with the same parameter of the common MMSE estimator) and  $\mu_\Theta$  being the number of averaged *a posteriori* SNRs. Please note that (5.5) generalizes both a GLR for averaged observations under a Gaussian speech model and a GLR for non-averaged observations but a super-Gaussian speech model: In case of  $\nu_\chi = 1$  we obtain the GLR (2.79) for averaged observations assuming a Gaussian speech model, while in case of  $\mu_\Theta = 1$  we obtain the GLR (2.72) assuming chi-distributed speech spectral amplitudes and no averaging.

## 5.2.2 Training of the Fixed *A Priori* SNR

While the averaging approach can nicely reduce estimation outliers, it still does not allow for achieving *a posteriori* SPPs close to zero at low *a priori* SNRs: If variable  $\xi$  approaches zero, the new GLR (5.5) still tends to  $P(H_1)/P(H_0)$  (note that  ${}_1F_1(a, b, 0) = 1$ ) and, thus, (5.1) tends to  $P(H_1)$ . To overcome this issue, in [Gerkmann et al., 2008] the use of fixed prior parameters, i. e., fixed *a priori* SPPs and fixed *a priori* SNRs, is proposed for SPU estimation (for controlling the weighting rule, still adapted  $\xi$  values can and should be used). Similar to  $\xi = 0$  under hypothesis  $H_0$ , a fixed *a priori* SNR  $\xi = \Xi_\Theta^x$  can be defined for the likelihood of speech presence (5.3) which can be interpreted as a typical *a priori* SNR value under hypothesis  $H_1$ . In this case,  $\xi = 0$  and  $\xi = \Xi_\Theta^x$  can be understood as model parameters for speech absence and speech presence, respectively. Then, the likelihoods  $p_{\bar{\gamma}_\Theta|H_0}^x(\bar{\gamma}_\Theta|H_0)$  and  $p_{\bar{\gamma}_\Theta|H_1}^x(\bar{\gamma}_\Theta|H_1)$  can be interpreted as measures for how well the observations  $\bar{\gamma}_\Theta$  fit the underlying model parameters. As a consequence, GLR values and *a posteriori* SPP values close to zero can be achieved in absence of speech.

The fixed *a priori* SNR  $\Xi_\Theta^x$  can be optimized as follows: Interpreting the *a posteriori* SPP as a detector for speech presence/absence, the probability of misdetection can be defined. This probability turns out to be a function of the *a priori* SNR  $\xi$ . The optimal fixed *a priori* SNR  $\xi = \Xi_\Theta^x$  can be obtained by minimizing the average cost of misdetection, i. e., a combination of the miss probability  $P_M$  and the false alarm probability  $P_F$  [Gerkmann et al., 2008]. Miss occurs, if speech is present, but the *a posteriori* SPP estimate is less than 0.5, while false alarm is defined by estimated *a posteriori* SPPs larger than 0.5 in absence of speech. The joint decision threshold for miss and false alarm is at observation values  $\bar{\gamma}_\Theta$  at which the *a posteriori* SPP yields 0.5. In this case the weighted likelihoods are equal  $P(H_0) \cdot p_{\bar{\gamma}_\Theta|H_0}^x(\bar{\gamma}_\Theta|H_0) = P(H_1) \cdot p_{\bar{\gamma}_\Theta|H_1}^x(\bar{\gamma}_\Theta|H_1)$  and the GLR (5.5) is one (cf. (5.1))

$$\frac{P(H_1)}{P(H_0)} \cdot \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\mu_\Theta \nu_\chi} \cdot {}_1F_1 \left( \mu_\Theta \nu_\chi, \mu_\Theta, \bar{\gamma}_\Theta \frac{\mu_\Theta \xi}{\nu_\chi + \xi} \right) = 1. \quad (5.6)$$

Solving this equation w. r. t.  $\bar{\gamma}_\Theta$ , the decision threshold

$$\Phi_\Theta = f(\xi, \nu_\chi, \mu_\Theta) \quad (5.7)$$

for testing  $\bar{\gamma}_\Theta \underset{H_0}{\overset{H_1}{\gtrless}} \Phi_\Theta$  turns out to be a function of the *a priori* SNR. We were not able to find a closed-form solution for (5.7), therefore, we employed numerical methods and the result is depicted in Figure 5.1. Please note that given a specific *a priori* SNR  $\xi$ , a miss occurs, if  $\bar{\gamma}_\Theta < \Phi_\Theta$  assuming  $H_1$  and a false alarm occurs, if  $\bar{\gamma}_\Theta > \Phi_\Theta$  assuming  $H_0$ , as shown in Figure 5.2.

As can be seen the probability of miss and the probability of false alarm are dependent on the same threshold  $\Phi_\Theta$  which is a function of the *a priori* SNR (5.7). Therefore, the

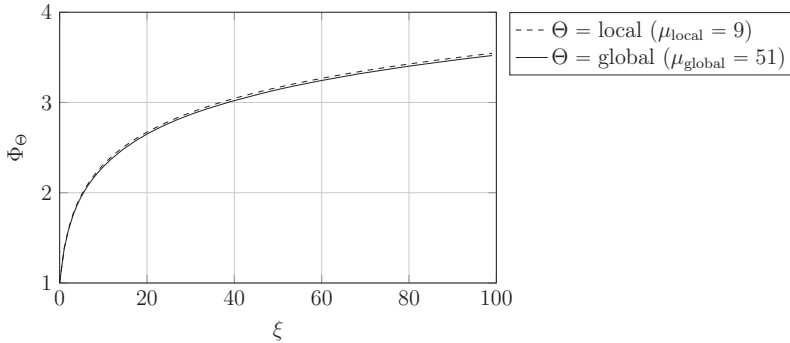


Figure 5.1: Result of the equation (5.7) (decision threshold) for the local and the global averaging window assuming  $\nu_\chi = 0.5$ ; the curves show specific combinations of  $\bar{\gamma}_\Theta$  and  $\xi$  values which lead to a unity GLR

fixed *a priori* SNR  $\Xi_\Theta^\chi$  can be optimized by minimizing the average cost of misdetection [Gerkmann et al., 2008]: Assuming a fixed *a priori* SNR  $\Xi_\Theta^\chi$  (and a corresponding fixed threshold  $\Phi_\Theta = f(\Xi_\Theta^\chi)$ ), the probability of miss is defined as [Gerkmann et al., 2008]

$$P_{M,\Theta}(\Xi_\Theta^\chi, \xi) = \int_0^{\Phi_\Theta = f(\Xi_\Theta^\chi)} \underbrace{p_{\bar{\gamma}_\Theta|H_1}^\chi(x|H_1)}_{=f(\xi)} dx \quad (5.8)$$

and the probability of false alarm can be calculated as [Gerkmann et al., 2008]

$$P_{F,\Theta}(\Xi_\Theta^\chi) = \int_{\Phi_\Theta = f(\Xi_\Theta^\chi)}^{\infty} p_{\bar{\gamma}_\Theta|H_0}^\chi(x|H_0) dx. \quad (5.9)$$

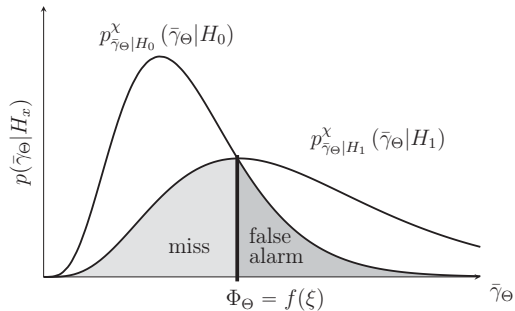
The average cost of misdetection can be determined by [Gerkmann et al., 2008]

$$R_\Theta(\Xi_\Theta^\chi, \xi) = P(H_0) \cdot P_{F,\Theta}(\Xi_\Theta^\chi) + P(H_1) \cdot P_{M,\Theta}(\Xi_\Theta^\chi, \xi). \quad (5.10)$$

The desired fixed *a priori* SNR can be obtained by minimizing the average cost  $R_\Theta(\Xi_\Theta^\chi, \xi)$  over a typical *a priori* SNR range  $[\xi_l, \xi_u]$  as [Gerkmann et al., 2008]

$$\Xi_\Theta^\chi = \arg \min_{\Xi_\Theta^\chi} \int_{\xi_l}^{\xi_u} R_\Theta(\Xi_\Theta^\chi, \xi) d\xi. \quad (5.11)$$

Assuming  $\nu_\chi = 0.5$  as in the previous chapter (cf. Section 4.3) as well as employing (5.3) and (5.4), respectively, for (5.8) and (5.9) with  $\mu_{\text{local}} = 9$  and  $\mu_{\text{global}} = 51$ , calculating (5.11) with the parameters  $\xi_l = -20$  dB,  $\xi_u = 20$  dB results in the optimal fixed *a priori* SNR values summarized in Table 5.1. The integrals in (5.8), (5.9), and (5.11) were calculated by the Gauss-Kronrod quadrature [Brass and Petras, 2011].

Figure 5.2: Probabilities of miss and false-alarm for  $P(H_1) = P(H_0) = 0.5$ 

$\Theta$	$\Delta k_\Theta$	$\Delta \ell_\Theta$	$\mu_\Theta =  \mathbb{K}_\Theta  \cdot  \mathbb{L}_\Theta $	$\Xi_\Theta^\chi$	$P(H_1)$
local	1	2	9	12.56 dB	0.5
global	8	2	51	10.42 dB	0.5

Table 5.1: Parameter set of the *a posteriori* SNR averaging framework and the proposed GLR (5.5) assuming  $\nu_\chi = 0.5$ 

These optimization steps can be interpreted as follows: Assuming a fixed threshold  $\Phi_\Theta = f(\Xi_\Theta^\chi)$ , the average cost of misdetection is measured for a typical *a priori* SNR range  $[\xi_l, \xi_u]$ . As can be seen in Figure 5.2, the probability of miss and the probability of false alarm (and, thus, the cost of misdetection) are dependent on the position of this threshold: Lower thresholds increase the probability of false alarm (residual noise level) and higher thresholds increase the probability of miss (speech distortion). Thus, repeating the average cost measurement for different fixed thresholds  $\Phi_\Theta = f(\Xi_\Theta^\chi)$ , the fixed *a priori* SNR  $\Xi_\Theta^\chi$  leading to the lowest average cost can be considered as optimal.

The proposed *a posteriori* SPP estimation steps using the fixed prior parameters can be summarized as follows:

- Average the *a posteriori* SNRs by means of (2.73) within a local and a global window using the parameters from Table 5.1.
- Calculate the local and global GLR (5.5) using the averaged *a posteriori* SNRs as well as the parameters from Table 5.1, namely, the fixed *a priori* SNR  $\Xi_\Theta^\chi$  (instead of the adapted one  $\xi$ ),  $P(H_1)$ , and  $\mu_\Theta$ . The shape parameter  $\nu_\chi$  should be chosen consistently with the shape parameter of the common MMSE estimator (the fixed *a priori* SNRs in Table 5.1 are trained assuming  $\nu_\chi = 0.5$ ).
- Calculate the local and global *a posteriori* SPPs (5.1) using the resulting GLRs.
- Combine the resulting local and global *a posteriori* SPPs to a final *a posteriori* SPP estimate using (2.74).

### 5.3 Performance Evaluation

In order to see the effect of enhanced *a posteriori* SPP estimation, we will assess the proposed approach and some reference approaches from the previous chapter without averaging and without fixed prior parameters in the following. For this, the simulation setup and the performance measures from Chapter 3 will be utilized. Since the proposal enhances SPU estimation only and SPU estimation is independent of the estimation domains (cf. Section 2.6), the following simulations will be based on MMSE STSA estimation under SPU (2.81). Nevertheless, the proposed approach can also be utilized in the other estimation domains.

We will use two reference approaches based on PDF-consistent MMSE STSA estimation under SPU from the previous chapter, both with an adapted *a priori* SNR. The first one assumes Rayleigh-distributed speech spectral amplitudes (denoted as R-STSA-SPU, cf. Table 4.2), while the second one assumes chi-distributed speech spectral amplitudes with the parameter  $\nu_\chi = 0.5$  (denoted as  $\chi$ -STSA-SPU, cf. Table 4.2).

A further reference approach is from [Gerkmann et al., 2008]: It is based on MMSE STSA estimation under SPU (2.81) assuming Rayleigh-distributed speech spectral amplitudes consistently. It utilizes R-STSA (2.48) as common MMSE estimator and employs averaged *a posteriori* SNR values and a fixed *a priori* SNR for the calculation of the *a posteriori* SPP. This approach will be denoted as R-STSA-SPU-enh. The *a posteriori* SPP is estimated as follows: The GLR is controlled by smoothed *a posteriori* SNRs (2.73) averaged within a local and a global averaging window and calculated by (2.79) using the parameters from Table 2.3, i. e., a shape parameter and a fixed *a priori* SNR. The local and global *a posteriori* SPPs are calculated by (2.75) and then multiplicatively unified to a final *a posteriori* SPP estimate by (2.74).

The proposed approach is an MMSE STSA estimation approach under SPU (2.81) assuming chi-distributed speech spectral amplitudes with  $\nu_\chi = 0.5$  (cf. Section 4.3) consistently for common MMSE estimation (2.49) and *a posteriori* SPP estimation (5.1) with the new GLR (5.5) from the last section. The GLR is driven by averaged observations and has fixed prior parameters shown in Table 5.1. The proposed approach will be denoted as  $\chi$ -STSA-SPU-enh.

The evaluation results are depicted in Figure 5.3. As in the previous chapter, the resulting  $\text{NA}_{\text{seg}}$ ,  $\text{SSDR}_{\text{seg}}$ , and LKR values are shown in the upper, middle, and lower plots, respectively. Furthermore, car noise, factory noise, and babble noise results can be found in the left, middle, and right column of figures, respectively.

Considering the amount of residual noise, approaches based on fixed prior parameters for SPU estimation ( $\chi$ -STSA-SPU-enh and R-STSA-SPU-enh) achieve significantly larger

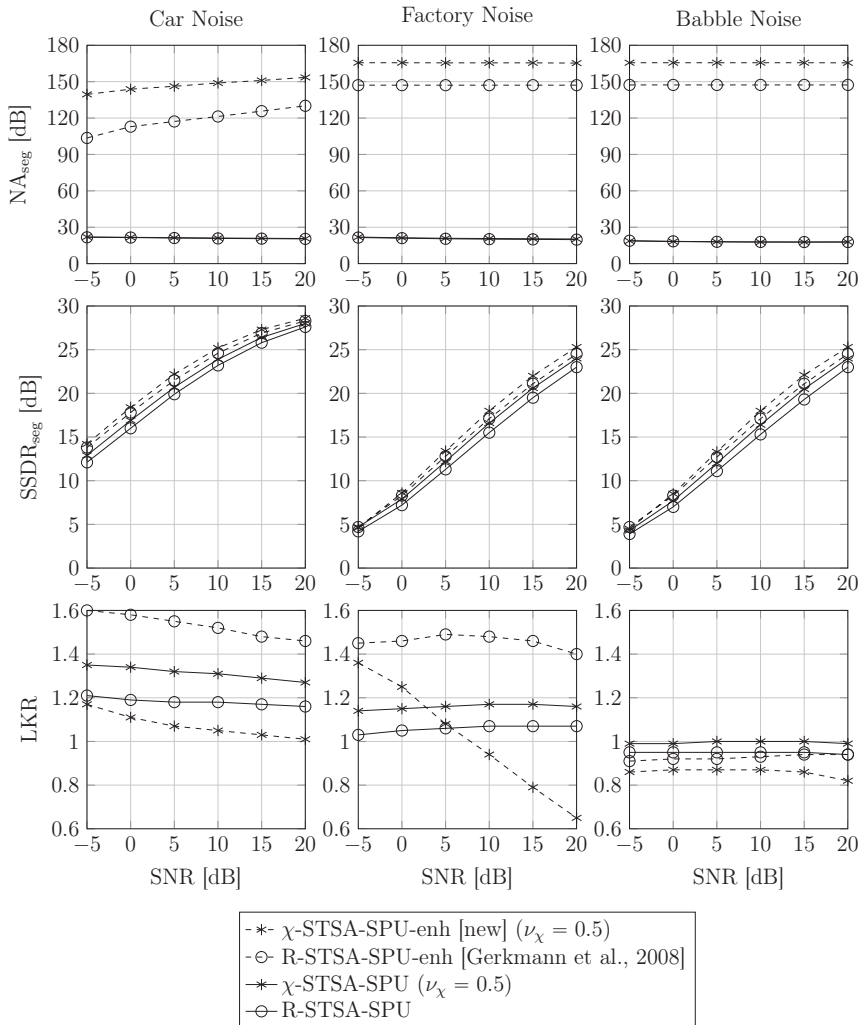


Figure 5.3: Performance evaluation results of MMSE STSA estimators with SPU based on an adapted *a priori* SNR for a *posteriori* SPP estimation (R-STSA-SPU,  $\chi$ -STSA-SPU) and MMSE STSA estimators with SPU based on averaging and a fixed *a priori* SNR for a *posteriori* SPP estimation (R-STSA-SPU-enh,  $\chi$ -STSA-SPU-enh) in terms of segmental noise attenuation ( $NA_{\text{seg}}$ , the larger the better), segmental speech-to-speech distortion ratio ( $SSDR_{\text{seg}}$ , the larger the better), and weighted log-kurtosis ratio (LKR, the smaller the better)

noise attenuation reflected by larger  $\text{NA}_{\text{seg}}$  values, compared to approaches with an adapted *a priori* SNR ( $\chi$ -STSA-SPU and R-STSA-SPU). This is due to the fact that approaches with fixed prior parameters for SPU estimation output significantly smaller (close to zero) *a posteriori* SPP values in speech absence. Furthermore, considering approaches with fixed prior parameters and averaging, the proposed approach with a super-Gaussian speech model  $\chi$ -STSA-SPU-enh outperforms the reference approach with a Gaussian speech model R-STSA-SPU-enh w. r. t.  $\text{NA}_{\text{seg}}$  at all input SNR levels and in all noise conditions.

In terms of speech component quality, all approaches perform approximately the same, reflected by close  $\text{SSDR}_{\text{seg}}$  levels. Interestingly, approaches with fixed prior parameters (R-STSA-SPU-enh,  $\chi$ -STSA-SPU-enh) show slightly larger  $\text{SSDR}_{\text{seg}}$  values than approaches without fixed prior parameters (R-STSA-SPU,  $\chi$ -STSA-SPU). Furthermore, approaches with a chi speech spectral amplitude model (super-Gaussian prior) perform slightly better than approaches with a Rayleigh speech spectral amplitude model (Gaussian prior), regardless whether an adapted or a fixed *a priori* SNR is used.

Regarding musical noise levels, the proposed approach nicely achieves the smallest LKR values in presence of car and babble noise at all input SNR levels. In case of factory noise, the proposed approach reaches the lowest LKR level only above 5 dB input SNR.

## 5.4 Summary

This chapter presented an improved MMSE estimation approach under SPU based on an enhanced *a posteriori* SPP estimator which unifies the advantages of a super-Gaussian speech model, averaged observations, and a fixed *a priori* SNR. Accordingly, an enhanced GLR for averaged *a posteriori* SNRs was derived by taking chi-distributed speech spectral amplitudes<sup>1</sup> (super-Gaussian speech model) into account. In order to overcome a typical disadvantage of common SPU estimation and to obtain *a posteriori* SPP estimates close to zero in speech absence, fixed SNR and SPP prior parameters were employed. The proposed approach was shown to outperform reference approaches that consider averaged observations but a Gaussian speech model, a super-Gaussian speech model but no averaging, and also approaches that are based on a Gaussian model without averaging.

---

<sup>1</sup>Please note that the presented approach can theoretically be combined with a gamma assumption for the speech spectral amplitudes. However, the corresponding PDFs (such as the likelihoods for averaged observations) and, thus, the final *a posteriori* SPP estimator cannot be obtained in closed-form. Accordingly, the optimization of the fixed *a priori* SNR has to be carried out by using numerically calculated PDFs.

# Chapter 6

## Recursive MMSE Estimation in Speech Enhancement and Some Links to Error Concealment

Contrary to the previous chapters dealing with non-recursive approaches, this chapter will focus on MMSE estimation carried out in a recursive manner. Recursive MMSE estimation is widely employed, there are applications in, e. g., speech enhancement, error concealment, etc. Although the aim of both speech enhancement and error concealment is to enhance disturbed (speech) signals, there has not yet been much research focus on relating these disciplines to each other. In this chapter, for the first time, we will show interesting commonalities and also differences between respective fields [Fodor et al., 2015]. Moreover, by analyzing differences, a general strength of error concealment over speech enhancement will be identified. Motivated by this finding, possible research directions for speech enhancement will briefly be sketched.

After a short introduction in Section 6.1, recursive MMSE estimation in speech enhancement and the Kalman filter as special case will be recapitulated in Section 6.2. This will be done in analogy to the non-recursive case in Chapter 2 based on PDFs. To the best of our knowledge, recursive MMSE estimation has not been derived for speech enhancement in this form before. This PDF-based description provides a basis for applying enhancement techniques of non-recursive speech enhancement (e. g., super-Gaussian speech priors, speech presence uncertainty estimation, etc.) to the recursive case. In Section 6.3, a recursive MMSE application for error concealment will be introduced in analogy to Section 6.2. In Section 6.4, interesting links between the applications for speech enhancement and error concealment will be shown. In Section 6.5, possible research directions for speech enhancement motivated by error concealment will briefly be sketched. Finally, Section 6.6 will provide a short summary.

## 6.1 Introduction

In Chapter 2 it was assumed that the speech process is memoryless, i.e., its outcomes  $S_\ell(k), S_{\ell-1}(k), \dots$  are independent of each other. Assuming further that the noise process  $D$  is also memoryless, the resulting observations  $Y_\ell(k), Y_{\ell-1}(k), \dots$  turn out to also be independent of each other. Thus, the speech is estimated sequentially in a way that each speech estimate  $\hat{S}_\ell(k)$  is obtained by means of one corresponding observation  $Y_\ell(k)$  only, signal history is merely exploited for smoothing purposes of the *a priori* SNR estimation (cf. the decision directed SNR estimator (2.88)).

However, utilizing a dynamic signal model, such as an AR speech process motivated by, e.g., the source-filter model of speech production [Rabiner and Schafer, 1978], [Vary and Martin, 2006], the resulting optimal MMSE estimator exploits signal redundancy using the current and the previous observations [Kalman, 1960]. Furthermore, under certain assumptions, the estimation can be performed *recursively* consisting of two steps, nicely relaxing computational complexity and memory usage [Haykin, 2002]. The first estimation step is called propagation (or prediction or *a priori* estimation) step which exploits signal redundancy by utilizing the previous observations and provides an *a priori* speech estimate. The second step, called the update step, corrects the *a priori* speech estimate using the current observation, resulting in an *a posteriori* speech estimate.

In speech enhancement, a typical recursive MMSE estimator is the well-known Kalman filter [Kalman, 1960]. A Kalman filter is a specific recursive MMSE estimator assuming a Gaussian distribution for both the speech prior and the likelihood. A Kalman filter is proposed in [Paliwal and Basu, 1987] for time-domain speech enhancement: Assuming an AR speech process based on a predictor and a memoryless acoustic channel, the speech estimation process turns out to be recursive, consisting of two subsequent steps. The first step utilizes the previous observations resulting in an *a priori* speech estimate which is corrected in the second step employing the current observation, resulting in an *a posteriori* speech estimate. A Kalman filter-based speech enhancement is proposed in [Wu and Chen, 1998] and [Puder, 2002] operating in subbands. The STFT-domain Kalman filter for speech enhancement in [Zavarehei and Vaseghi, 2005] assumes an AR process for the complex-valued speech STFT coefficients and a memoryless noise process. In [Esch and Vary, 2008a], different approaches to calculate the AR coefficients for determining the *a priori* speech estimate as well as different estimators for obtaining the *a posteriori* speech estimate are investigated. Furthermore, in [Esch and Vary, 2008b], the assumption of a memoryless noise process is replaced by an AR noise model. An STFT-domain Kalman filter for estimating the speech spectral amplitudes is introduced in [So et al., 2010].

While speech enhancement deals with speech signals distorted by acoustic noise, error con-

concealment is related to speech or audio signals distorted by transmission channel noise. More specifically, the aim of error concealment is to conceal residual bit errors after demodulation or channel decoding. Please note that different from speech enhancement, error concealment often operates in the time domain. As mentioned in Chapter 2, in speech enhancement the acoustic channel is assumed to distort the speech signal by superimposing statistically independent noise. In error concealment, however, the *a priori* knowledge about the channel is even richer: It is assumed that on transmitter side, speech samples or source-coded parameters are quantized, mapped to corresponding bit combinations, and transmitted over digital error-prone transmission channels. Therefore, since the number of possible channel inputs is finite and the transmitted symbols have fixed positions in the constellation diagram, channel reliability information can be exploited by the decoder on receiver side. Then, the speech samples or source-coded parameters can be estimated by MMSE estimation using a likelihood enriched by channel reliability information and a prior containing *a priori* knowledge about the speech. Here, similar to speech enhancement signal redundancy can be exploited assuming a dynamic speech model, resulting in a *recursive* MMSE estimator. This approach can be employed for, e. g., robust source decoding of speech signals [Görtz, 1998], [Fingscheidt and Vary, 2001], [Lahouti and Khandani, 2007], [Pourmir and Lahouti, 2008], [Jameel et al., 2009], [Han et al., 2013], source-coded audio signals [Adrat et al., 2000], [Pflug and Fingscheidt, 2013a], or uncompressed audio [Pflug and Fingscheidt, 2013b], [Pflug, 2013] that exploit signal redundancy in sample values or various source codec parameters (such as scaling factors, line spectral frequencies (LSFs) vectors, vector-quantized gains, adaptive codebook indices). Thereby, the signal redundancy is exploited by a time-variant modeling of the prior either using Markov chains [Görtz, 1998], [Fingscheidt and Vary, 2001], [Lahouti and Khandani, 2007], [Pourmir and Lahouti, 2008], [Jameel et al., 2009], [Adrat et al., 2000], [Han et al., 2013] or employing approaches based on linear prediction in [Fingscheidt, 1998], [Pflug and Fingscheidt, 2013a], [Pflug and Fingscheidt, 2013b], [Pflug, 2013]. Typical applications are speech and audio transmission systems such as mobile phones or digital wireless microphones.

## 6.2 Recursive MMSE Estimation in Speech Enhancement

The aim of recursive MMSE estimation is to estimate the speech STFT coefficient  $S_\ell(k)$  using the observations  $\mathbf{Y}_0^\ell(k) = [Y_0(k), Y_1(k), \dots, Y_\ell(k)]^T$  and some *a priori* knowledge, resulting in the speech estimate  $\hat{S}_\ell(k)$  (cf. Figure 6.1). The *a priori* knowledge incorporates models about the speech process  $S$  and the (acoustic) channel. Accordingly, the speech is modeled by the following AR process: The speech STFT coefficient is a sum of the predicted

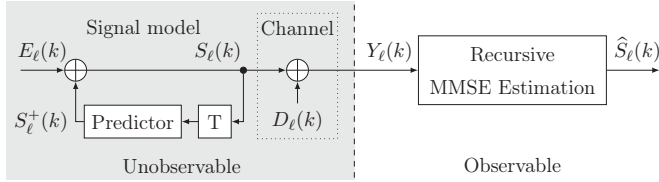


Figure 6.1: Signal model, channel, and recursive MMSE estimation

speech  $S_\ell^+(k)$  and the prediction error  $E_\ell(k)$ , i. e., [Fodor et al., 2015]

$$S_\ell(k) = S_\ell^+(k) + E_\ell(k). \quad (6.1)$$

The prediction error  $E_\ell(k)$  can be associated with the input signal of the human vocal tract and is assumed to be a zero-mean (random) signal which is statistically independent of  $S_\ell^+(k)$ . The predicted speech  $S_\ell^+(k)$  can be interpreted as a contribution of the vocal tract and it is assumed to be the output of a predictor using the previous  $L_p$  speech STFT coefficients  $\mathbf{S}_{\ell-L_p}^{\ell-1}(k) = [S_{\ell-L_p}(k), S_{\ell-L_p+1}(k), \dots, S_{\ell-1}(k)]^T$ , i. e., [Fodor et al., 2015]

$$S_\ell^+(k) = \mathbf{A}^H(k) \cdot \mathbf{S}_{\ell-L_p}^{\ell-1}(k) \quad (6.2)$$

with  $\mathbf{A}(k) = [A_{L_p}(k), A_{L_p-1}(k), \dots, A_1(k)]^T$  and  $(\cdot)^H$  being the prediction coefficients and the conjugate transpose operator, respectively. The fact that the predictor delays input coefficients  $S_\ell(k)$  by at least one frame during computing any contribution to  $S_\ell^+(k)$  is expressed in Figure 6.1 by an explicit delay unit 'T'. Assuming that the prediction coefficients are slowly time-varying, they are treated as constants for the moment. The output of the vocal tract  $S_\ell(k) = S_\ell^+(k) + E_\ell(k)$  can be observed through an acoustic channel which distorts it by superimposing the outcomes of a statistically independent random noise process  $D_\ell(k)$ , resulting in the observed noisy speech STFT coefficients  $Y_\ell(k) = S_\ell(k) + D_\ell(k)$ .

Our objective is to estimate the speech STFT coefficient  $S_\ell(k)$  utilizing the observations  $\mathbf{Y}_0^\ell(k)$  we have made so far as well as the *a priori* knowledge about the speech process and the acoustic channel introduced above. Since the current speech STFT coefficient  $S_\ell(k)$  is dependent of the previous ones (cf. (6.1) and (6.2)), it is also dependent on the previous observations, therefore, the MMSE estimator (2.12) with the underlying signal and channel model from Figure 6.1 turns out to be [Fodor et al., 2015]

$$\hat{S}_\ell(k) = E \{ S_\ell(k) | \mathbf{Y}_0^\ell(k) \} = E \{ S_\ell(k) | Y_\ell(k), \mathbf{Y}_0^{\ell-1}(k) \} = \int_{\mathcal{C}} S \cdot p(S | Y_\ell(k), \mathbf{Y}_0^{\ell-1}(k)) dS \quad (6.3)$$

with  $\mathbf{Y}_0^{\ell-1}(k) = [Y_0(k), Y_1(k), \dots, Y_{\ell-1}(k)]^T$  and with  $p(S_\ell(k) | Y_\ell(k), \mathbf{Y}_0^{\ell-1}(k))$  being the posterior. Please note that in this chapter, we omit the subscript of PDFs for ease of readability. Analog to the MMSE estimator in Chapter 2.2, the posterior is calculated by Bayes' rule as

$$p(S_\ell(k) | Y_\ell(k), \mathbf{Y}_0^{\ell-1}(k)) = \frac{p(Y_\ell(k) | S_\ell(k), \mathbf{Y}_0^{\ell-1}(k)) \cdot p(S_\ell(k) | \mathbf{Y}_0^{\ell-1}(k))}{p(Y_\ell(k) | \mathbf{Y}_0^{\ell-1}(k))} \quad (6.4)$$

where  $p(Y_\ell(k)|S_\ell(k), \mathbf{Y}_0^{\ell-1}(k))$  is the likelihood,  $p(S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k))$  is the speech prior, and  $p(Y_\ell(k)|\mathbf{Y}_0^{\ell-1}(k))$  is the evidence. Please note that the evidence is typically calculated by marginalizing the product of the prior and the likelihood (cf. Chapter 2.2).

A block diagram of recursive MMSE estimation is depicted in Figure 6.2. Please note that the lower signal path can be associated with the likelihood computation, the upper signal path is related to the speech prior computation, and the center represents the recursive MMSE estimator itself. Starting with the lower left-hand corner of Figure 6.2, the noisy speech signal  $y(n)$  being a sum of the speech  $s(n)$  and the noise signal  $d(n)$  is transformed into the STFT domain by segmentation, windowing, and a DFT (cf. Section 2.1). As a next step, the likelihood is computed by means of the resulting noisy speech STFT coefficients  $Y_\ell(k)$ .

## 6.2.1 The Likelihood

The *a priori* knowledge about the acoustic channel is contained in the likelihood (cf. Chapter 2.3). According to Figure 6.1, we assume a memoryless acoustic channel, thus, the likelihood reduces to the likelihood in (2.12) [Martin et al., 2008]

$$p(Y_\ell(k)|S_\ell(k), \mathbf{Y}_0^{\ell-1}(k)) = p(Y_\ell(k)|S_\ell(k)) = p_D(D_\ell(k) = Y_\ell(k) - S_\ell(k)) \quad (6.5)$$

meaning that the speech  $S_\ell(k)$  is a sufficient statistic for the noisy speech  $Y_\ell(k)$ . Just as in Section 2.3.2, the likelihood (6.5) can be associated with the PDF of the noise process  $D$ , thus, it is a function of the noise PSD  $\sigma_{D,\ell}^2(k)$  which is typically estimated in practice (cf. (2.28)). As introduced in Section 2.7, the noise PSD is estimated by the noisy speech STFT coefficients. Thus, in Figure 6.2 the noisy speech  $Y_\ell(k)$  is fed into a noise power estimator and the likelihood is subsequently calculated by the resulting noise PDF  $\widehat{\sigma}_{D,\ell}^2(k)$  and the noisy speech  $Y_\ell(k)$ . The resulting likelihood (6.5) turns out to be a function of the speech  $h(S_\ell(k))$  which will be the integration variable of the estimator (cf. lower signal path in Figure 6.2).

## 6.2.2 The Estimator

Inserting the likelihood (6.5) into the posterior (6.4), the MMSE estimator (6.3) turns out to be [Fodor et al., 2015]

$$\widehat{S}_\ell(k) = \frac{\int_{\mathbb{C}} S \cdot p(S|\mathbf{Y}_0^{\ell-1}(k)) \cdot p(Y_\ell(k)|S) dS}{p(Y_\ell(k)|\mathbf{Y}_0^{\ell-1}(k))} \quad (6.6)$$

with the evidence

$$p(Y_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)) = \int_{\mathbb{C}} p(S|\mathbf{Y}_0^{\ell-1}(k)) \cdot p(Y_\ell(k)|S) dS \quad (6.7)$$

and with  $p(S|\mathbf{Y}_0^{\ell-1}(k))$  being the speech prior. As can be seen, both the prior and the likelihood are a function of the integration variable  $S = S_\ell(k)$ , namely  $b(S_\ell(k))$  and  $h(S_\ell(k))$  in Figure 6.2, respectively. The result of the integration is the *a posteriori* speech estimate  $\hat{S}_\ell(k)$  which is transformed back into the time domain by an IDFT of (6.6) and a subsequent OLA step, resulting in the speech estimate  $\hat{s}(n)$  (cf. Figure 6.2).

### 6.2.3 The Prior

As mentioned in Section 6.1, within the first step of recursive MMSE estimation an *a priori* estimate for the speech  $S_\ell(k)$  is calculated by means of the previous observations  $\mathbf{Y}_0^{\ell-1}(k)$ . The *a priori* speech estimate is then corrected within a second estimation step employing the current observation  $Y_\ell(k)$ . By this means, the information carried by the previous observations becomes successively part of the *a priori* knowledge. This is reflected by the fact that the speech prior is a function of the previous observations  $\mathbf{Y}_0^{\ell-1}(k)$ , as can be seen in (6.4).

Since the objective of the propagation step is to estimate the current speech STFT coefficient  $S_\ell(k)$  by incorporating the previous observations  $\mathbf{Y}_0^{\ell-1}(k)$  only, it can be defined in an MMSE sense as  $E\{S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)\}$ . This expression is actually the statistical expectation of the speech prior  $p(S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k))$  from (6.4) and is the so-called *a priori* speech estimate (or propagation step) [Esch, 2012]

$$\begin{aligned} E\{S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)\} &= E\{E_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)\} + E\{S_\ell^+(k)|\mathbf{Y}_0^{\ell-1}(k)\} \\ &= 0 + \mathbf{A}^H(k) \cdot \begin{pmatrix} E\{S_{\ell-L_p}(k)|\mathbf{Y}_0^{\ell-L_p}(k)\} \\ E\{S_{\ell-L_p+1}(k)|\mathbf{Y}_0^{\ell-L_p+1}(k)\} \\ \vdots \\ E\{S_{\ell-2}(k)|\mathbf{Y}_0^{\ell-2}(k)\} \\ E\{S_{\ell-1}(k)|\mathbf{Y}_0^{\ell-1}(k)\} \end{pmatrix} \\ &= \mathbf{A}^H(k) \cdot \hat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k). \end{aligned} \quad (6.8)$$

Here<sup>1</sup>, we employed the speech signal model (6.1) and (6.2). Furthermore, we made use of the fact that the prediction error  $E_\ell(k)$  (or innovation, e.g., [Haykin, 2002]) is zero-mean as well as orthogonal to the previous speech STFT coefficients and, therefore, also to the previous observations  $\mathbf{Y}_0^{\ell-1}(k)$  (cf., e.g., [Haykin, 2002], [Papoulis and Pillai, 2002]). As can be seen from (6.8), the *a priori* speech estimate is a predicted quantity calculated by the

<sup>1</sup>Strictly speaking, the expectation of the predicted speech in (6.8) should be defined as  $E\{S_\ell^+(k)|\mathbf{Y}_0^{\ell-1}(k)\} = \mathbf{A}^H(k) \cdot [E\{S_{\ell-L_p}(k)|\mathbf{Y}_0^{\ell-1}(k)\}, E\{S_{\ell-L_p+1}(k)|\mathbf{Y}_0^{\ell-1}(k)\}, \dots, E\{S_{\ell-1}(k)|\mathbf{Y}_0^{\ell-1}(k)\}]^T$ . However, we do not employ so-called Kalman smoothing which means that we do not recalculate the  $\ell$ th *a posteriori* speech estimate  $E\{S_\ell(k)|\mathbf{Y}_0^\ell(k)\}$  after the  $\ell$ th observation  $Y_\ell(k)$  has been made, i.e., we assume  $E\{S_{\ell-L_p}(k)|\mathbf{Y}_0^{\ell-1}(k)\} = E\{S_{\ell-L_p}(k)|\mathbf{Y}_0^{\ell-L_p}(k)\}$ ,  $E\{S_{\ell-L_p+1}(k)|\mathbf{Y}_0^{\ell-1}(k)\} = E\{S_{\ell-L_p+1}(k)|\mathbf{Y}_0^{\ell-L_p+1}(k)\}$ , etc.

same predictor as in our speech production model (cf. Figure 6.1 and Equation (6.2)) and the last  $L_p$  speech estimates as [Esch, 2012]

$$\widehat{S}_\ell^+(k) = E\{S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)\} = \mathbf{A}^H(k) \cdot \widehat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k) \quad (6.9)$$

which is the result of the propagation step. Assuming that the *a priori* speech estimate  $\widehat{S}_\ell^+(k) = f(\mathbf{Y}_0^{\ell-1}(k))$  is a sufficient statistic for  $S_\ell(k)$ , the speech prior yields [Fodor et al., 2015]

$$p(S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)) = p(S_\ell(k)|\widehat{S}_\ell^+(k)). \quad (6.10)$$

This new speech prior is the PDF of the speech STFT coefficient  $S_\ell(k)$  given a fixed *a priori* speech estimate  $\widehat{S}_\ell^+(k)$ . The specific *a priori* speech estimate can be interpreted as a deterministic point in the complex plain and the speech coefficient  $S_\ell(k)$  scatter around this point according to the distribution of  $\bar{E}_\ell(k) = S_\ell(k) - \widehat{S}_\ell^+(k)$  which is called the *propagation error*. Thus, the speech prior actually describes the distribution of the propagation error [Fodor et al., 2015]

$$p(S_\ell(k)|\widehat{S}_\ell^+(k)) = p_{\bar{E}}(\bar{E}_\ell(k) = S_\ell(k) - \widehat{S}_\ell^+(k)) \quad (6.11)$$

with the variance [Esch, 2012]

$$\begin{aligned} E\{|S_\ell(k) - \widehat{S}_\ell^+(k)|^2\} &= E\{|S_\ell^+(k) + E_\ell(k) - \widehat{S}_\ell^+(k)|^2\} = E\{|E_\ell(k)|^2\} + E\{|S_\ell^+(k) - \widehat{S}_\ell^+(k)|^2\} \\ &= \sigma_{E,\ell}^2(k) + \sigma_{E^+,\ell}^2(k) = \sigma_{\bar{E},\ell}^2(k) \end{aligned} \quad (6.12)$$

which is not accessible directly in practice, therefore, it is typically estimated during processing [Zavarehei and Vaseghi, 2005], [Esch, 2012]. For (6.12), we employed (6.1) and made use of the fact that the prediction error  $E_\ell(k)$  is orthogonal to the previous speech STFT coefficients and, therefore, to the predicted speech  $S_\ell^+(k)$  and the *a priori* speech estimate  $\widehat{S}_\ell^+(k) = f(\mathbf{Y}_0^{\ell-1}(k))$ . Please note that (6.12) holds independently of the type of the PDF of the propagation error.

It turned out that the speech prior is a PDF being a function of the speech  $b(S_\ell(k))$  which is the integration variable in the estimator (cf. (6.6) and Figure 6.2). Furthermore, the expectation of the speech prior is the *a priori* speech estimate (6.9) being calculated by a predictor of the order  $L_p$  which employs the prediction coefficients  $\mathbf{A}(k)$  and the last  $L_p$  *a posteriori* speech estimates (cf. upper signal path in Figure 6.2).

Since the prediction coefficients  $\mathbf{A}(k)$  in (6.9) are not accessible directly in practice, they have to be estimated. Introducing again time variability, the prediction coefficients can be calculated by, e.g., the Levinson-Durbin algorithm [Markel and Gray, 1976] as being employed in [Esch, 2012] or the least-mean-squares (LMS) algorithm [Haykin, 2002] utilized in [Wu and Chen, 1998]. Alternatively, the widely-used normalized least-mean-squares (NLMS)

algorithm is also able to calculate the prediction coefficients recursively as [Haykin, 2002]

$$\widehat{\mathbf{A}}_{\ell+1}(k) = \widehat{\mathbf{A}}_{\ell}(k) + \mu \cdot \frac{\widehat{E}_{\ell}^*(k)}{\|\widehat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k)\|^2 + \Delta} \cdot \widehat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k) \quad (6.13)$$

with  $\widehat{E}_{\ell}(k) = \widehat{S}_{\ell}(k) - \widehat{S}_{\ell}^+(k)$  as well as with  $\mu$ ,  $(\cdot)^*$ ,  $\Delta$ , and  $\|\cdot\|$  denoting the step size constant, the complex conjugate, the regularization parameter, and the Euclidean norm, respectively. So far, a recursive MMSE estimation framework has been described, in the following we will introduce a specific recursive MMSE estimator assuming a Gaussian distribution for both the speech prior (6.11) and the likelihood (6.5).

### 6.2.4 The Kalman Filter

Assuming a zero-mean bivariate Gaussian-distributed propagation error  $\bar{E}_{\ell}(k)$ , the speech prior (6.11) turns out to be (cf. (2.19)) [Fodor et al., 2015]

$$p(S_{\ell}(k)|\widehat{S}_{\ell}^+(k)) = p_{\bar{E}}(\bar{E}_{\ell}(k) = S_{\ell}(k) - \widehat{S}_{\ell}^+(k)) = \frac{1}{\pi\sigma_{\bar{E},\ell}^2(k)} \cdot \exp\left(-\frac{|S_{\ell}(k) - \widehat{S}_{\ell}^+(k)|^2}{\sigma_{\bar{E},\ell}^2(k)}\right) \quad (6.14)$$

with  $\sigma_{\bar{E},\ell}^2(k)$  being the propagation error variance (6.12). Furthermore, assuming Gaussian-distributed acoustic noise  $D_{\ell}(k)$ , the likelihood turns out to be (2.28). In this case, the estimator from (6.6) can be obtained in closed form (the derivation can be found in Appendix E) and the result of the integration turns out to be the Kalman filter equations (cf. [Zavarehei and Vaseghi, 2005, Eq. (12)])

$$\widehat{S}_{\ell}(k) = \widehat{S}_{\ell}^+(k) + \widehat{E}_{\ell}(k) \quad (6.15)$$

with  $\widehat{S}_{\ell}^+(k)$  being the *a priori* speech estimate from (6.9) which merely utilizes signal history and with  $\widehat{E}_{\ell}(k)$  being an update which corrects the *a priori* estimate employing the current observation  $Y_{\ell}(k)$ . Accordingly, the update is defined as [Esch and Vary, 2008a]

$$\widehat{E}_{\ell}(k) = K_{\ell}(k) \cdot M_{\ell}(k) \quad (6.16)$$

with the Kalman gain  $K_{\ell}(k)$  and with [Esch and Vary, 2008a]

$$M_{\ell}(k) = Y_{\ell}(k) - \widehat{S}_{\ell}^+(k). \quad (6.17)$$

The Kalman gain is calculated as (cf. (2.41)) [Esch, 2012]

$$K_{\ell}(k) = \frac{\zeta_{\ell}(k)}{1 + \zeta_{\ell}(k)} \quad (6.18)$$

with (cf. (2.42))

$$\zeta_{\ell}(k) = \frac{\sigma_{\bar{E},\ell}^2(k)}{\sigma_{D,\ell}^2(k)} \quad (6.19)$$

being the *a priori* SNR as defined in recursive MMSE estimation. Please note that if no AR speech process is assumed, the prediction coefficients in (6.2) are zero and the *a priori* speech estimate yields also zero in (6.9). Furthermore, the propagation error variance (6.12) results in the speech spectral variance  $\sigma_{S,\ell}^2(k)$  in the numerator of (6.19). Accordingly, (6.19) reduces then to  $\sigma_{S,\ell}^2(k)/\sigma_{D,\ell}^2(k)$  being the classical *a priori* SNR (2.42).

The *a priori* SNR  $\zeta_\ell(k)$  can be estimated by [Esch and Vary, 2008a]

$$\hat{\zeta}_\ell(k) = (1-\beta) \max \left\{ \frac{|M_\ell(k)|^2}{\widehat{\sigma_{D,\ell}^2}(k)} - 1, 0 \right\} + \beta \frac{|\widehat{E}_{\ell-1}(k)|^2}{\widehat{\sigma_{D,\ell-1}^2}(k)} \quad (6.20)$$

with  $\beta \in [0, 1]$  being a smoothing factor. Please note the similarity between (6.20) and the decision-directed *a priori* SNR estimator (2.88). Therefore, concerning the calculation of the update  $\widehat{E}_\ell(k)$ ,  $M_\ell(k) = Y_\ell(k) - \widehat{S}_\ell^+(k)$  can be associated with the observation (cf. (2.41) and (6.17)), the fraction  $|M_\ell(k)|^2/\widehat{\sigma_{D,\ell}^2}(k)$  with the *a posteriori* SNR  $\gamma_\ell(k)$  (cf. (2.88) and (6.20)), and  $\widehat{E}_\ell(k)$  with the MMSE STS estimate (cf. (2.41)).

Please note that applying  $Y_\ell(k) = S_\ell(k) + D_\ell(k)$  to (6.17), (6.16) turns out to be  $\widehat{E}_\ell(k) = K_\ell(k) \cdot [S_\ell(k) + D_\ell(k) - \widehat{S}_\ell^+(k)] = K_\ell(k) \cdot [D_\ell(k) + \bar{E}_\ell(k)]$ . Assuming that in an ideal case the Kalman gain was able to eliminate the noise  $D_\ell(k)$  completely, the last equation would reduce to  $\widehat{E}_\ell(k) = \bar{E}_\ell(k) = S_\ell(k) - \widehat{S}_\ell^+(k)$  and (6.15) would reduce to  $\widehat{S}_\ell(k) = S_\ell(k)$ .

## 6.3 Recursive MMSE Estimation in Error Concealment

Contrary to the previous practice, this section deals with speech or audio signals which are distorted by passing through a transmission channel. The aim of error concealment is to conceal residual bit errors after demodulation or channel decoding. Theoretically, residual bit errors could be concealed on sample level using hard-decoded, transmitter-sided samples, e.g., by employing the time domain equivalent of the equations in the previous section<sup>2</sup>. However, in error concealment we typically have more *a priori* knowledge about the (transmission) channel than in speech enhancement and reliability information on a bit level can be exploited, typically improving the estimation performance [Fingscheidt and Vary, 2001], [Pflug and Fingscheidt, 2013b], [Pflug, 2013]. As we will see, this is a particular strength of bit error concealment and is a clear advantage over the speech enhancement task.

Similar to the speech enhancement approach in the previous section, in conjunction with an AR process for speech (or audio), signal redundancy can be exploited in error concealment, resulting in a recursive MMSE estimation formula. More specifically, we assume the same

<sup>2</sup>Due to the properties of the DFT, the equations in the previous section are also valid in the time domain.

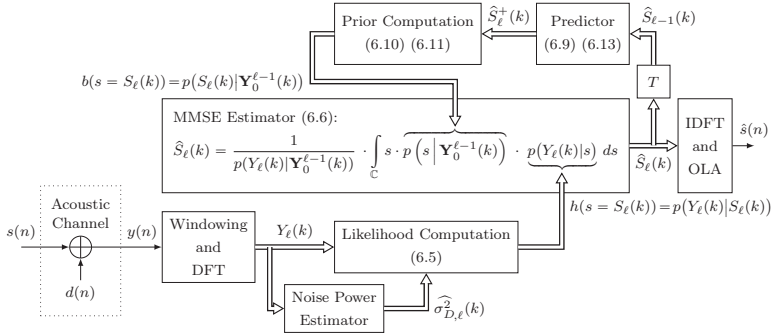


Figure 6.2: STFT domain recursive MMSE estimation for speech enhancement assuming a memoryless acoustic channel

AR process for speech as in Figure 6.1: The speech sample is modeled as a sum

$$s(n) = s^+(n) + e(n) \quad (6.21)$$

with  $s^+(n)$  and  $e(n)$  being the predicted speech and the prediction error, respectively. It is assumed that the prediction error is a zero-mean (random) signal and independent of  $s^+(n)$ . The predicted speech is assumed to be the output signal of a predictor of the order  $N_p$  [Pflug and Fingscheidt, 2013b]

$$s^+(n) = \mathbf{a}^T \cdot \mathbf{s}_{n-N_p}^{n-1} \quad (6.22)$$

with  $\mathbf{a} = [a_{N_p}, a_{N_p-1}, \dots, a_1]^T$  being the prediction coefficients and with  $\mathbf{s}_{n-N_p}^{n-1} = [s(n-N_p), s(n-N_p+1), \dots, s(n-1)]^T$ . Assuming that the prediction coefficients are slowly time-varying, they are treated as constants for the moment.

A block diagram of bit error concealment is given in Figure 6.3. As in the previous section, the lower signal path can be associated with the likelihood computation, the upper signal path can be related to the speech prior computation, and the center represents the recursive MMSE estimator itself. The input is in the lower left-hand corner of Figure 6.3.

### 6.3.1 The Likelihood

Here, the speech (or audio) samples  $s(n)$  are quantized with  $M$  bit, mapped to natural-binary bit combinations  $\mathbf{x}(n) = [x_0(n), x_1(n), \dots, x_m(n), \dots, x_{M-1}(n)]$  (cf. 'Quantization and Bit Mapping' in Figure 6.3), and transmitted through the transmission channel. Assuming a binary phase-shift keying (BPSK) modulation, the transmitted bits (BPSK symbols)  $x_m(n) \in \{-1, 1\}$  are modeled to be distorted in the transmission channel by a statistically independent, real-valued additive noise  $d_m(n)$ . Therefore, the resulting real-valued

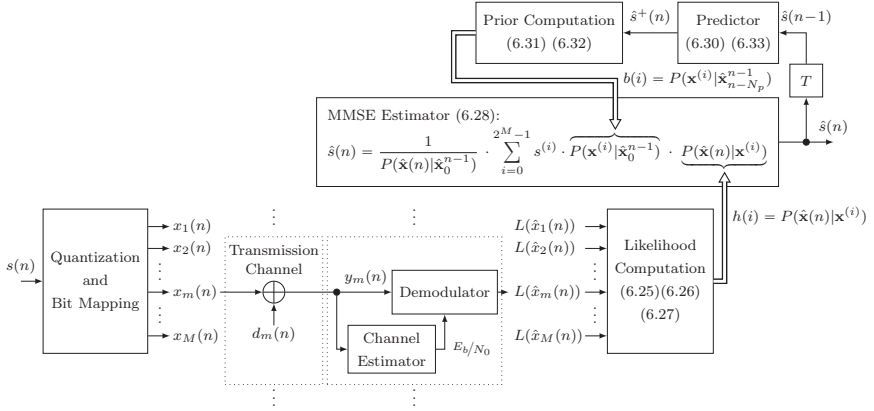


Figure 6.3: Time domain recursive MMSE estimation for error concealment assuming a memoryless transmission channel

noisy symbols  $y_m(n)$  can be observed at the output of the channel (cf. 'Transmission Channel' in Figure 6.3). Then, the demodulator calculates the log-likelihood ratio (LLR) reflecting the likelihood of a possibly transmitted bit  $x_m(n)$  (cf. 'Demodulator' in Figure 6.3) [Hagenauer, 1995]

$$L(\hat{x}_m(n)) = \ln \frac{P(\hat{x}_m(n)|x_m(n) = +1)}{P(\hat{x}_m(n)|x_m(n) = -1)} \quad (6.23)$$

where  $\hat{x}_m(n) = \text{sign}(y_m(n))$  is the observation being the receiver-sided hard-decided bit. In case of BPSK modulation, however, the LLR is calculated as [Hagenauer, 1995]

$$L(\hat{x}_m(n)) = 4 \cdot E_b/N_0 \cdot y_m(n) \quad (6.24)$$

with  $E_b/N_0$  being the so-called energy per bit to noise PSD ratio which is calculated by a channel estimator (cf. 'Channel Estimator' below 'Demodulator' in Figure 6.3).

Using the LLR, the bit-error probability [Hagenauer, 1995]

$$\text{BER}_m(n) = \frac{1}{1 + e^{|L(\hat{x}_m(n))|}} \quad (6.25)$$

can be calculated which reflects the probability that the transmitted bit  $x_m(n)$  does not match the receiver-sided hard-decided one  $\hat{x}_m(n) = \text{sign}(y_m(n))$  due to channel distortions.

Since the transmitter-sided speech sample  $s(n)$  is quantized, it assumes a discrete value from a finite set of elements  $s^{(i)}$  with  $i \in \{0, 1, \dots, 2^M-1\}$ . Moreover, each possible discrete value  $s^{(i)}$  can bijectively be mapped to a bit combination  $\mathbf{x}^{(i)} = [x_0^{(i)}, x_1^{(i)}, \dots, x_m^{(i)}, \dots, x_{M-1}^{(i)}]$ .

The *bit* likelihood reflects the probability of the observed received bit given possible channel inputs and can be computed by means of the bit-error probability as [Fingscheidt

and Vary, 2001]

$$P(\hat{x}_m(n)|x_m^{(i)}) = \begin{cases} \text{BER}_m(n), & \text{if } \hat{x}_m(n) \neq x_m^{(i)}, \\ 1 - \text{BER}_m(n), & \text{else.} \end{cases} \quad (6.26)$$

Assuming a memoryless transmission channel and statistical independence of the bit distortions  $d_m(n)$  along the bit indices  $m$ , the *sample* likelihood is obtained by multiplying the bit likelihoods (cf., e. g., [Fingscheidt and Vary, 2001])

$$P(\hat{\mathbf{x}}(n)|\mathbf{x}^{(i)}) = \prod_{m=0}^{M-1} P(\hat{x}_m(n)|x_m^{(i)}) \quad (6.27)$$

with  $\hat{\mathbf{x}}(n) = [\hat{x}_1(n), \hat{x}_2(n), \dots, \hat{x}_m(n), \dots, \hat{x}_{M-1}(n)]$ . Please note that while the likelihood computation starts on bit level, the resulting likelihood (6.27) is a sample likelihood (cf. the input and output of 'Likelihood Computation' in Figure 6.3). Moreover, further processing will be carried out on sample level, as we will see.

### 6.3.2 The Estimator

Due to the fact that the transmitter-sided samples are quantized and, therefore, their possible values are from a finite set of elements  $s^{(i)}$ , the MMSE estimator turns out to be a sum instead of an integral (cf. (6.6)) [Fingscheidt, 1998]

$$\hat{s}(n) = \frac{\sum_{i=0}^{2^M-1} s^{(i)} \cdot P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1}) \cdot P(\hat{\mathbf{x}}(n)|\mathbf{x}^{(i)})}{P(\hat{\mathbf{x}}(n)|\hat{\mathbf{x}}_0^{n-1})} \quad (6.28)$$

with the evidence

$$P(\hat{\mathbf{x}}(n)|\hat{\mathbf{x}}_0^{n-1}) = \sum_{i=0}^{2^M-1} P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1}) \cdot P(\hat{\mathbf{x}}(n)|\mathbf{x}^{(i)}) \quad (6.29)$$

and with  $P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1})$  and  $P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1})$  being the sample likelihood (6.27) and the speech prior, respectively. Please note that both the prior and the sample likelihood are functions of the summation index  $i$  of the estimator, namely  $b(i)$  (cf. upper signal path in Figure 6.3) and  $h(i)$  (cf. lower signal path in Figure 6.3), respectively. The result of (6.28) is the *a posteriori* speech estimate  $\hat{s}(n)$  (cf. right-hand side of the MMSE estimator in Figure 6.3).

### 6.3.3 The Prior

Since each possible bit combination  $\mathbf{x}^{(i)}$  can bijectively be mapped to a discrete value  $s^{(i)}$ , the speech prior can be rewritten as  $P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1}) = P(s^{(i)}|\hat{\mathbf{x}}_0^{n-1})$ . Similar to the previous section,

the *a priori* speech estimate is calculated by means of the previous  $N_p$  speech estimates (cf. 'Predictor' in Figure 6.3) [Fodor et al., 2015]

$$\hat{s}^+(n) = \mathbf{a}^T \cdot \hat{\mathbf{s}}_{n-N_p}^{n-1} = f(\hat{\mathbf{x}}_0^{n-1}) \quad (6.30)$$

with  $\hat{\mathbf{s}}_{n-N_p}^{n-1} = [\hat{s}(n-N_p), \hat{s}(n-N_p+1), \dots, \hat{s}(n-1)]^T$  and with  $\mathbf{a}$  being the prediction coefficients from (6.22). Assuming that the *a priori* speech estimate is a sufficient statistic for  $s^{(i)}$ , the speech prior can be rewritten as [Fodor et al., 2015]

$$P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1}) = P(s^{(i)}|\hat{s}^+(n)). \quad (6.31)$$

It turns out that the speech prior is driven by the *a priori* speech estimate (cf. the connection between 'Predictor' and 'Prior Computation' in Figure 6.3). Please note that  $s^{(i)}$  is a discrete real-valued quantity, while  $\hat{s}^+(n) \in \mathbb{R}$ . Nevertheless, the speech prior (6.31) can be calculated by integration [Fingscheidt and Vary, 1997]

$$P(s^{(i)}|\hat{s}^+(n)) = \int_{I_i} p_{\bar{e}}(s - \hat{s}^+(n)) ds \quad (6.32)$$

with  $I_i$  being the PCM quantization intervals ( $i \in \{0, 1, \dots, 2^M-1\}$ ) and  $p_{\bar{e}}(s(n)-\hat{s}^+(n))$  being the propagation error PDF (cf. (6.11)). Please note that in real implementations, online integration is not necessary: The discretized prior (6.32) can be implemented as a table lookup within a training process and driven by a quantized *a priori* speech estimate  $\hat{s}^{+(i)}(n)$  [Pflug and Fingscheidt, 2013b]. The training process consists of the following steps: Assuming a stationary propagation error  $\bar{e}(n)$ , the propagation error PDF  $p_{\bar{e}}(\bar{e}(n))$  in (6.32) can be obtained by histogram measurements. Then, quantizing  $\hat{s}^+(n)$  with  $M$  bits and integrating the propagation error PDF over the PCM quantization intervals, the discrete probabilities  $P(s^{(i)}|\hat{s}^+(n))$  result. Given  $\hat{s}^+(n)$  the new speech prior  $P(s^{(i)}|\hat{s}^+(n))$  is a function of the summation index of the estimator  $i$  only and can directly be fed into the estimator (cf. the connection between 'Prior Computation' and 'Estimator' in Figure 6.3). By this means, the speech prior can be obtained in a computationally efficient way during processing.

Since the prediction coefficients  $\mathbf{a}$  in (6.30) are not accessible directly in practice, they have to be estimated. Introducing again time variability, the prediction coefficients can be calculated by, e.g., the NLMS algorithm [Haykin, 2002]

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) + \mu \cdot \frac{\hat{e}(n)}{\|\hat{\mathbf{s}}_{n-N_p}^{n-1}\|^2 + \Delta} \cdot \hat{\mathbf{s}}_{n-N_p}^{n-1} \quad (6.33)$$

with  $\hat{e}(n) = \hat{s}(n) - \hat{s}^+(n)$  as well as with  $\mu$  and  $\Delta$  being the step size constant and the regularization parameter, respectively. Alternatively, the prediction coefficients can also be obtained by a slightly modified NLMS algorithm [Schuller et al., 2002], [Pflug and Fingscheidt, 2013b].

## 6.4 Linking Speech Enhancement and Error Concealment

In the previous sections, we have introduced an example application of recursive MMSE estimation for both speech enhancement and error concealment. In this section, we will show some interesting links between these two disciplines by means of these applications. Just as in the previous sections, we will first focus on the likelihood, then on the estimator, and finally on the speech prior.

### 6.4.1 The Likelihood

Starting with the likelihood, it can be seen that while in speech enhancement the acoustic channel is modeled by a continuous PDF (6.5), the likelihood in error concealment is discrete (6.27). The latter is due to the fact that the transmitted-sided samples are quantized by  $M$  bit. Thus, they assume a value from a finite set of elements  $s^{(i)}$  with  $i \in \{0, 1, \dots, 2^M - 1\}$  which can bijectively be mapped to a bit combination  $\mathbf{x}^{(i)}$  being the input of the channel. The receiver-sided observations  $\hat{\mathbf{x}}(n)$  are also discrete and from a finite set of elements, therefore, the sample likelihood (6.27) is also discrete.

Comparing the lower signal paths in Figures 6.2 and 6.3, the introduced approaches for speech enhancement and error concealment have in common that the likelihood is computed by means of the channel output ( $Y_\ell(k)$  or  $y_m(n)$ ) and the channel noise PSD ( $\sigma_{D,\ell}^2(k)$  or  $N_0$ ). However, there is a clear difference between the disciplines regarding the estimation of the noise PSD: While in speech enhancement the noise power is estimated without a reference by using the noisy speech STFT coefficients and some *a priori* knowledge only (cf. Section 2.7.1), the amount of noise in error concealment is dependent on the distance between the received symbol and all possibly transmitted symbols in the constellation diagram, the latter having fixed positions depending on the modulation scheme. Therefore, in error concealment we have more information about possible channel inputs which is a clear advantage over speech enhancement.

It is also interesting to note that while in speech enhancement usually a common PDF is employed for the likelihood (6.5) (mostly the Gaussian probability density (2.28), typically justified by the central limit theorem [Ephraim and Malah, 1984]), in error concealment the likelihood depends on the employed bit mapping. In order to compare the likelihoods from both disciplines, in the following we define the noise in error concealment as the difference between the received-sided hard-decided speech samples and the quantized transmitted ones. For 16 bit uniform PCM quantization, natural binary bit mapping, and BPSK modulation, the histogram of the transmission channel noise turns out to be spiky as can be seen on the

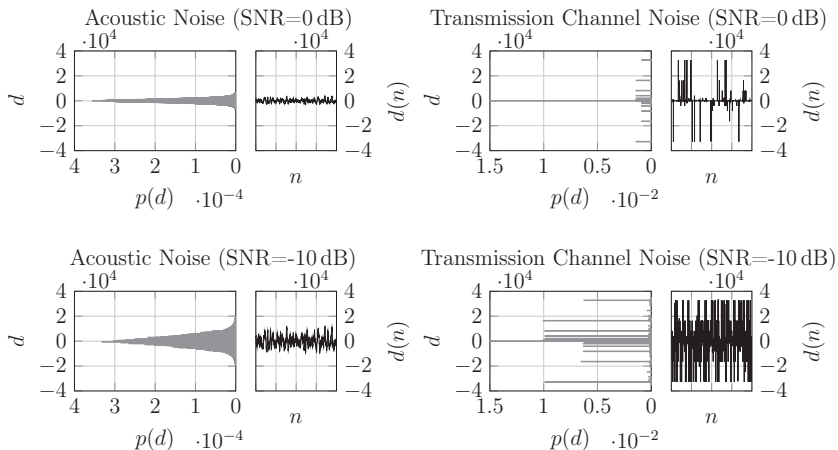


Figure 6.4: Left: Normalized time domain histograms and waveforms of car noise; right: Normalized time domain histograms and waveforms of transmission channel noise applied to 16 bit speech samples with underlying natural-binary bit mapping. The SNR values were calculated by means of a fixed speech signal level of  $-26 \text{ dB}_{\text{ov}}$  and respective noise signal levels, both measured according to ITU-T P.56 [ITU-T P.56, 1993].

right-hand side of Figure 6.4 for different signal SNRs. The bottom and top spikes in the histograms and the higher amplitudes in the waveforms can be related to bit errors at bit positions close to the most significant bit (MSB).

Considering acoustic noise, a decrease of the noise power (an increase of the SNR) is associated with a decrease of the noise signal levels. Furthermore, as a result the time-domain noise histogram becomes narrower and its height increases (cf. car noise example on the left hand-side of Figure 6.4). Considering transmission channel noise, a decrease of the noise power (an increase of the SNR) can be associated with a decreasing number of bit errors, resulting in lower peaks belonging to  $d \neq 0$  in the histogram on the right-hand side of Figure 6.4. However, it can be observed that the height of all high spikes (referring to single bit errors) belonging to  $d \neq 0$  are scaled approximately equally due to the decreased noise power. Thus, while in case of acoustic noise the noise power can be linked with the *width* of the noise PDF, in error concealment (more specifically, in case of uniform PCM quantization, natural-binary bit mapping, and BPSK modulation) the noise power can be related to the *height* of spikes in the noise histogram belonging to  $d \neq 0$ .

### 6.4.2 The Estimator

Maybe the most obvious difference between the estimators in speech enhancement and in error concealment is that while the first one (6.6) is an integral, the latter one (6.28) is a finite sum.

In case of speech enhancement, the *a posteriori* speech estimate (6.6) is obtained by calculating an integral over the whole complex plane. Unfortunately, since two-dimensional PDFs are part of its integrands, online numerical computation of (6.6) is typically hard to manage in practice. However, employing a common distribution for the speech prior and the likelihood potentially allows for a closed form solution. As we have seen in Section 6.2.4 for the example of a Gaussian speech prior and a Gaussian likelihood, the well-known Kalman filter equations (6.15)-(6.18) result as an analytical solution of (6.6).

In case of error concealment, the *a posteriori* speech estimator (6.28) cannot be obtained in closed form, since the likelihood (6.27) is usually not a common PDF. However, both the speech prior (6.32) and the likelihood (6.5) are one-dimensional PDFs and the sums in (6.28) are real-time computable (cf. [Pflug and Fingscheidt, 2013b]).

### 6.4.3 The Prior

As can be seen, assuming an AR speech process, the speech prior is the PDF of the propagation error in both speech enhancement and error concealment (cf. (6.11) and (6.32)).

In Section 6.2, we introduced the Kalman filter as a recursive MMSE application in speech enhancement. In Kalman filtering, the propagation error is assumed to be Gaussian distributed. In [Zavarehei and Vaseghi, 2005], this is justified by a tradeoff between mathematical tractability of the estimator and PDF model mismatch. Based on histogram measurements, it is shown in [Esch and Vary, 2008a] that the propagation error follows a super-Gaussian density rather than a Gaussian density (cf. left-hand side of Figure 6.5). Accordingly, MMSE estimators with an underlying super-Gaussian speech prior are proposed for obtaining the estimation update  $\hat{E}_t(k)$  in (6.15). Furthermore, since the *a priori* speech estimate depends on the channel, the propagation error also depends on the channel. Subsequently, an SNR-dependency of the propagation error is observed in [Esch and Vary, 2011] and an SNR-dependent propagation error PDF is proposed.

The non-Gaussianity of the propagation error is also observed in [Pflug and Fingscheidt, 2013b] for a recursive MMSE application to error concealment. Here, the propagation error PDF is obtained by histogram measurements assuming no channel distortions, i. e.,  $E\{(s^+ - \hat{s}^+)^2\} = 0$ . The resulting histogram turns also out to be rather super-Gaussian than a Gaussian (cf. right-hand side of Figure 6.5). Furthermore, in [Pflug and Fingscheidt, 2013b]

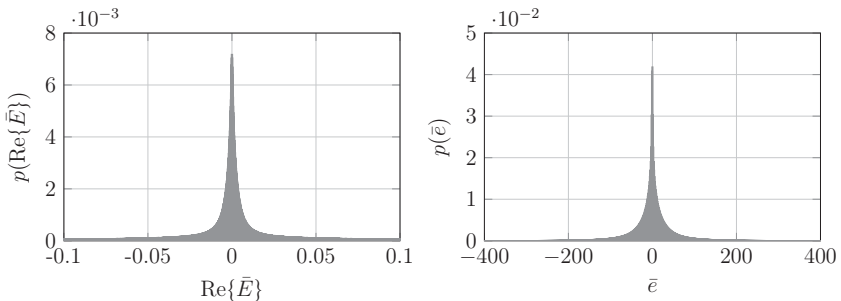


Figure 6.5: Left: Normalized frequency domain histogram of the propagation error  $\bar{E}_\ell(k) = S_\ell(k) - \hat{S}_\ell^+(k)$  measured in a speech enhancement system from Section 6.2; right: Normalized time domain histogram of the propagation error  $\bar{e}(n) = s(n) - \hat{s}^+(n)$  measured in an error concealment system from Section 6.3 assuming that  $E\{(s^+ - \hat{s}^+)^2\}$  is zero [Pflug and Fingscheidt, 2013b].

the dependency of  $\hat{s}^+$  on the propagation error PDF  $p_{\bar{e}}(\bar{e})$  is investigated and propagation error PDFs with different shapes and variances for different  $\hat{s}^+$  values are proposed.

In Sections 6.2 and 6.3, we employed the NLMS algorithm to obtain the prediction coefficients due to its robustness and low computational complexity requirements. However, there are other approaches to estimate these coefficients, such as the Levinson-Durbin algorithm [Markel and Gray, 1976]. Please note that since the propagation error depends on the *a priori* speech estimate ((6.8) or (6.30)), it also depends on the estimated prediction coefficients. Therefore, a change of the estimation algorithm for obtaining the prediction coefficients requires a new training of the propagation error PDF.

## 6.5 Outlook

In this section, we briefly introduce possible research directions for speech enhancement inspired by error concealment. As we have seen in Section 6.3, a crucial advantage of error concealment is the ability of exploiting bit reliability information. Thus, it is possible that the speech enhancement approach from Section 6.2 can benefit from using bit likelihoods. This means that the STFT coefficient likelihood (6.5) could be calculated by means of bit likelihoods as in Section 6.3 instead of (2.28). As can be seen in (6.26), the bit likelihoods can be obtained by means of the bit error probabilities (6.25). Accordingly, we experimentally measured  $\text{BER}_m(k, \text{SNR})$  values by comparing the real parts of 16 bit quantized noisy speech STFT coefficients  $Q\{\text{Re}\{Y_\ell(k)\}\}$  to real parts of 16 bit quantized clean speech STFT coefficients  $Q\{\text{Re}\{S_\ell(k)\}\}$  at bit position  $m$ , frequency bin  $k$  and SNRs of 0 dB and 20 dB,

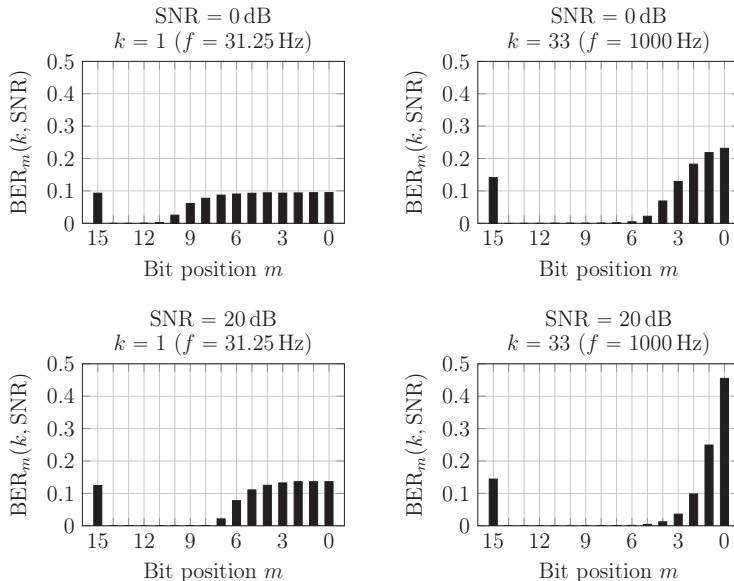


Figure 6.6: Bit error probabilities  $\text{BER}_m(k, \text{SNR})$  measured by comparing the real parts of 16 bit quantized noisy speech STFT coefficients  $Q\{\text{Re}\{Y_\ell(k)\}\}$  to real parts of quantized clean speech STFT coefficients  $Q\{\text{Re}\{S_\ell(k)\}\}$ , for different SNRs and frequencies.  $Q\{\cdot\}$  denotes the quantization operation. Natural binary bit representation with the MSB (bit position  $M-1 = 15$ ) being the sign bit is used.

respectively. The speech was distorted by car noise. It turned out that the bit likelihoods depend on the SNR and the frequency bin index  $k$ , as can be seen in Figure 6.6 for  $k = 1$  and  $k = 33$  (at 8 kHz sampling rate and a DFT length of 256 samples), reflecting  $f=31.25$  Hz and  $f=1000$  Hz, respectively. The SNR dependency is reflected by the fact that at SNR=20 dB less bits are in error than at SNR=0 dB. Furthermore, at higher frequencies less bits are erroneous compared to lower frequencies which is typical for car noise. Therefore, bit likelihoods seem to provide a useful, SNR-dependent channel model, individually for each frequency bin.

As a practical implementation, a speech enhancement approach as in Section 6.2 could be modified to include bit likelihoods. The bit likelihoods could be controlled by the bit error probabilities being trained as a function of the *a priori* SNR  $\xi_\ell(k) = E\{|S_\ell(k)|^2\}/\sigma_{D,\ell}^2(k)$  individually for each frequency bin  $k$  and bit index  $m$ . The resulting  $\text{BER}_m(k, \xi_\ell(k))$  values could be stored in a lookup table and indexed by, e.g., a decision-directed *a priori* SNR estimate (2.88). The bit likelihood could be calculated by (6.26) with  $\hat{x}_m(n)$  and  $x_m^{(i)}$  being the  $m$ th bit of the quantized real or imaginary part of an observed noisy speech STFT

coefficient and the  $m$ th bit of the real or imaginary part of a possible quantized speech STFT coefficient, respectively. Then, the coefficient likelihood (cf. (6.5)) could be obtained by (6.27) with  $\hat{\mathbf{x}}$  being the mapped bit combination of the quantized real or imaginary part of the observed noisy speech STFT coefficient and with  $\mathbf{x}^{(i)}$  being a possible bit combination for a clean speech STFT coefficient.

By this means, the simple Gaussian assumption for the noise STFT coefficients (2.28) could be replaced by such a new approach which allows for an environment-specific processing (cf. [Fingscheidt et al., 2008]). Further work towards such an approach would be to formulate the estimator and the predictor as in Section 6.3 in the STFT domain.

## 6.6 Summary

In this chapter a synopsis of recursive MMSE estimation was given for the disciplines speech enhancement and error concealment. Then, some interesting links between these research fields were shown. It turned out that recent approaches to bit error concealment based on an AR speech model are well comparable to iterative methods in speech enhancement, such as the Kalman filter approach. However, the channel models are quite different: In error concealment, powerful bit reliability information can be exploited in order to obtain robust estimation results, while in speech enhancement the channel is modeled on an STFT coefficient level without reliable reference. Motivated by this finding, some new research directions for speech enhancement were identified, inspired by error concealment.

Furthermore, the synopsis of respective approaches was given in a PDF-based manner. To the best of our knowledge, there has not been an attempt to review recursive MMSE estimation in speech enhancement based on PDFs as in Chapter 2. It advantageously provides a basis for applying enhancements of non-recursive speech enhancement to the recursive case.



# Chapter 7

## Recursive MMSE Estimation Under Speech Presence Uncertainty

With a specific focus, recursive MMSE estimation for speech enhancement was recapitulated in the previous chapter. Different from the usual practice, the description of the recursive MMSE speech estimator was based on PDFs with a similar notation as in Chapter 2 instead of the typical matrix notation. The new PDF-based description allows for applying (PDF-based) improvements of state-of-the-art non-recursive speech enhancement to the recursive case. Accordingly, as a typical enhancement scheme in non-recursive speech enhancement, SPU estimation will be applied to recursive MMSE speech enhancement in this chapter. First, a general recursive MMSE estimation formula under SPU will be given which will turn out to be a product of a common recursive MMSE speech estimator and an *a posteriori* SPP estimator as in the non-recursive case. Then, a general *a posteriori* SPP estimation formula will be provided taking signal history into account. Furthermore, specific *a posteriori* SPP estimators will be derived based on either a Gaussian or a super-Gaussian speech PDF assumption. The *a priori* SPP will also be dependent on signal history, therefore, an adaptive tracking of the *a priori* SPP will be proposed inspired by an approach known from literature employed in non-recursive speech enhancement. Finally, the performance of the a recursive MMSE approach under SPU assuming a Gaussian speech model will be evaluated.

This chapter is organized as follows: Section 7.1 will give a short introduction to recursive MMSE estimation under SPU. Then, Section 7.2 will describe the algorithmic approach in detail by introducing a new general recursive MMSE estimation formula under SPU and corresponding *a priori* and *a posteriori* SPP estimators. Furthermore, specific *a posteriori* SPP estimators will be derived assuming either a Gaussian or a super-Gaussian distribution for the propagation error. Section 7.3 will provide the instrumental evaluation of a recursive MMSE estimation approach under SPU assuming a Gaussian propagation error. Finally, Section 7.4 will summarize this chapter.

## 7.1 Introduction

As can be seen in Chapter 2, there are a lot of improvement perspectives for classical non-recursive MMSE speech enhancement, such as MMSE estimation in diverse estimation domains, MMSE estimation with a super-Gaussian speech model, e. g., [Martin, 2002], [Lotter and Vary, 2005], or MMSE estimation under SPU, e. g., [McAulay and Malpass, 1980], [Ephraim and Malah, 1984]. While classical, non-recursive MMSE speech enhancement has been a vital research field over the past decades, the number of STFT-domain *recursive* MMSE speech enhancement proposals is comparatively small. Furthermore, most often the Kalman filter is employed as recursive MMSE STS estimator which is based on a Gaussian assumption for the *a priori* estimation error (i. e., for the speech prior, cf. Chapter 6) and the acoustic noise (i. e., for the likelihood) [Zavarehei and Vaseghi, 2005]. Due to the underlying AR speech model and the corresponding temporal dependencies (cf. (6.1), (6.2), and Figure 6.1), the Kalman filter is commonly described using a state space representation based on a matrix notation. More specifically, due to the Gaussian assumptions the Kalman filter is typically characterized by mean and variance matrices (cf., e. g., [Kalman, 1960], [Zavarehei and Vaseghi, 2005]). However, a description using these two matrices does not allow for assuming a super-Gaussian speech prior or extending the Kalman filter by SPU estimation. The description of recursive MMSE estimation from the previous chapter employs a PDF-based notation similar to Chapter 2 and allows for applying enhancement schemes of non-recursive speech enhancement to the recursive case.

In [Esch and Vary, 2008a] it was argued that the update step in recursive MMSE estimation can be associated with classical non-recursive MMSE estimation and, therefore, the Kalman gain  $K_\ell(k)$  in (6.16) can be related to the common spectral weighting rule. Accordingly, the Kalman gain in (6.18) was intuitively recognized as a Wiener spectral weighting rule (2.41) and replaced in a heuristical way by more modern spectral weighting rules with underlying super-Gaussian speech priors. It is shown in Appendix E by means of the new PDF-based recursive MMSE framework from Section 6.2 that this replacement is indeed optimal in an MMSE sense, if the employed spectral weighting rule belongs to an MMSE STS estimator (cf. Section 2.5.1). Furthermore, the chosen MMSE STS weighting rule can directly be employed in the update step (6.16) by substituting the *a priori* SNR  $\xi_\ell(k)$  and the *a posteriori* SNR  $\gamma_\ell(k)$  by their recursive MMSE estimation equivalents (cf. (6.19))

$$\zeta_\ell(k) = \frac{\sigma_{E,\ell}^2(k)}{\sigma_{D,\ell}^2(k)}$$

and

$$\varsigma_\ell(k) = \frac{|Y_\ell(k) - \widehat{S}_\ell^+(k)|^2}{\sigma_{D,\ell}^2(k)}, \quad (7.1)$$

respectively. Please note that if no AR speech process is assumed, the prediction coefficients

in (6.2) are zero and the *a priori* speech estimate  $\hat{S}_\ell^+(k)$  yields also zero in (6.9). Accordingly, (7.1) reduces to  $|Y_\ell(k)|^2/\sigma_{D,\ell}^2(k)$  being the *a posteriori* SNR  $\gamma_\ell(k)$  as used in non-recursive speech enhancement.

As a further PDF-related improvement scheme, the new recursive MMSE speech enhancement framework will be extended by SPU estimation in this chapter. First, a general recursive MMSE estimation formula under SPU will be derived, followed by a general *a posteriori* SPP formula taking signal history into account. Moreover, specific *a posteriori* SPP estimators will be provided assuming either a Gaussian or a super-Gaussian model for the propagation error. Furthermore, the *a priori* SPP will turn out to make use of signal history, therefore, a tracking algorithm will be proposed. Interestingly, the resulting estimators will turn out to be very similar to those in the non-recursive case in Chapter 2.

## 7.2 Algorithmic Approach

Please note that in the remainder of this chapter, we will omit the subscript of PDFs and the frequency bin index  $k$  due to readability. Furthermore, we will assume the same signal and channel models as in the previous chapter (cf. Figure 6.1), therefore, due to temporal dependencies the frame indices will still be used. Just as in Section 2.6 and in Chapters 4 and 5 we argue that the assumption of permanent speech presence made in recursive MMSE estimation (cf. Chapter 6) is not fulfilled in practice. Therefore, we extend the general recursive MMSE estimation formula (6.3) in terms of total probability by the hypothesis of speech absence  $H_0$  and speech presence  $H_1$  (cf. Section 2.6) resulting in the general formula of recursive MMSE estimation under SPU

$$\hat{S}_\ell = P(H_0|\mathbf{Y}_0^\ell) \cdot E\{S_\ell|\mathbf{Y}_0^\ell, H_0\} + P(H_1|\mathbf{Y}_0^\ell) \cdot E\{S_\ell|\mathbf{Y}_0^\ell, H_1\} \quad (7.2)$$

with  $P(H_0|\mathbf{Y}_0^\ell)$ ,  $P(H_1|\mathbf{Y}_0^\ell) = 1 - P(H_0|\mathbf{Y}_0^\ell)$  being the *a posteriori* SAP, the *a posteriori* SPP as well as with  $E\{S_\ell|\mathbf{Y}_0^\ell, H_0\}$ ,  $E\{S_\ell|\mathbf{Y}_0^\ell, H_1\}$  being a recursive MMSE clean speech estimator for speech absence and for speech presence, respectively. Please note that this is also a generalization of the general formula of non-recursive MMSE estimation under SPU (2.65), extended by an AR speech model (cf. (6.1), (6.2), and Figure 6.1). Under hypothesis  $H_0$ , the speech is assumed to be absent, thus,  $E\{S_\ell|\mathbf{Y}_0^\ell, H_0\} \equiv 0$  results. Furthermore, the expectation  $E\{S_\ell|\mathbf{Y}_0^\ell, H_1\}$  is equal to (6.3), since it is assumed for (6.3) that speech is always present. Therefore, only the *a posteriori* SPP  $P(H_1|\mathbf{Y}_0^\ell)$  is unknown in (7.2) which will be derived in the following.

### 7.2.1 *A Posteriori* SPP Estimation

Using Bayes' rule, the *a posteriori* SPP can be rewritten as

$$P(H_1|\mathbf{Y}_0^\ell) = P(H_1|Y_\ell, \mathbf{Y}_0^{\ell-1}) = P(H_1|\mathbf{Y}_0^{\ell-1}) \cdot \frac{p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_1)}{p(Y_\ell|\mathbf{Y}_0^{\ell-1})} \quad (7.3)$$

with  $P(H_1|\mathbf{Y}_0^{\ell-1})$  and  $p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_1)$  being the *a priori* SPP and the likelihood of speech presence, respectively, both based on the previous observations  $\mathbf{Y}_0^{\ell-1}$ . The denominator in (7.3), namely  $p(Y_\ell|\mathbf{Y}_0^{\ell-1})$ , can be obtained in terms of total probability as

$$\begin{aligned} p(Y_\ell|\mathbf{Y}_0^{\ell-1}) &= p(Y_\ell, H_0|\mathbf{Y}_0^{\ell-1}) && + p(Y_\ell, H_1|\mathbf{Y}_0^{\ell-1}) \\ &= P(H_0|\mathbf{Y}_0^{\ell-1}) \cdot p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_0) && + P(H_1|\mathbf{Y}_0^{\ell-1}) \cdot p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_1) \end{aligned} \quad (7.4)$$

with  $p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_0)$  and  $P(H_0|\mathbf{Y}_0^{\ell-1}) = 1 - P(H_1|\mathbf{Y}_0^{\ell-1})$  being the likelihood of speech absence and the *a priori* SAP, respectively, both based on the previous observations  $\mathbf{Y}_0^{\ell-1}$ .

Applying (7.4) to the *a posteriori* SPP formula (7.3) allows for introducing the GLR

$$\Lambda_0^\ell = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_1)}{p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_0)}. \quad (7.5)$$

Please note the similarity to the GLR (2.67): The only difference to the non-recursive case is that (7.5) takes signal history into account in the form of the previous observations  $\mathbf{Y}_0^{\ell-1}$ . Using the GLR (7.5), the *a posteriori* SPP (7.3) can alternatively be defined as (cf. (2.66))

$$P(H_1|\mathbf{Y}_0^\ell) = \frac{\Lambda_0^\ell}{1 + \Lambda_0^\ell}. \quad (7.6)$$

Let us investigate the likelihoods  $p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_1)$  and  $p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_0)$  in (7.5). Similar to (2.69), the likelihood of speech presence can be calculated by marginalization as

$$p(Y_\ell|\mathbf{Y}_0^{\ell-1}, H_1) = \int_{\mathcal{C}} p(Y_\ell|S, \mathbf{Y}_0^{\ell-1}, H_1) \cdot p(S|\mathbf{Y}_0^{\ell-1}, H_1) dS \quad (7.7)$$

with  $p(Y_\ell|S_\ell, \mathbf{Y}_0^{\ell-1}, H_1)$  and  $p(S_\ell|\mathbf{Y}_0^{\ell-1}, H_1)$  being the likelihood (6.5) and the speech prior (6.10), respectively, both assuming permanent speech presence. As in Section 4.1, we argue that the *a posteriori* SPP estimator (7.6) and thus the GLR (7.5) should be based on the same signal PDF assumptions as the common recursive MMSE speech estimator (6.3). Therefore, the speech prior (6.10) is employed here assuming that the *a priori* speech estimate  $\widehat{S}_\ell^+$  is a sufficient statistic for the speech  $S_\ell$  (cf. Section 6.2.3). Accordingly, the speech prior turns out to be  $p(S_\ell|\mathbf{Y}_0^{\ell-1}, H_1) = p(S_\ell|\widehat{S}_\ell^+, H_1)$ . Furthermore, the likelihood (6.5) from Section 6.2.1 is utilized here assuming a memoryless acoustic channel, i. e.,  $p(Y_\ell|S_\ell, \mathbf{Y}_0^{\ell-1}, H_1) = p(Y_\ell|S_\ell, H_1)$ .

For the likelihood of speech absence in the denominator of (7.5) we can write

$$\begin{aligned}
 p(Y_\ell | \mathbf{Y}_0^{\ell-1}, H_0) &= \int_{\mathbb{C}} p(Y_\ell | S, \mathbf{Y}_0^{\ell-1}, H_0) \cdot p(S | \mathbf{Y}_0^{\ell-1}, H_0) dS \\
 &= \int_{\mathbb{C}} p(Y_\ell | S, H_0) \cdot p(S | \hat{S}_\ell^+, H_0) dS \\
 &= \int_{\mathbb{C}} p(Y_\ell | S, H_0) \cdot \delta(S - \hat{S}_\ell^+) dS \\
 &= p(Y_\ell | \hat{S}_\ell^+, H_0).
 \end{aligned} \tag{7.8}$$

Here, we assumed a memoryless acoustic channel, i. e.,  $p(Y_\ell | S, \mathbf{Y}_0^{\ell-1}, H_0) = p(Y_\ell | S, H_0)$ , and that the *a priori* speech estimate  $\hat{S}_\ell^+$  is a sufficient statistic for the speech  $S_\ell$ , i. e.,  $p(S | H_0, \mathbf{Y}_0^{\ell-1}) = p(S | \hat{S}_\ell^+, H_0)$ . Furthermore, since speech absence is assumed the speech prior  $p(S | \hat{S}_\ell^+, H_0)$  is a dirac delta function  $\delta(\cdot)$  at  $\hat{S}_\ell^+$ , i. e.,  $p(S | \hat{S}_\ell^+, H_0) = \delta(S - \hat{S}_\ell^+)$ . The rationale behind this assumption is that even if speech is absent  $H_0$ , the *a priori* speech estimate  $\hat{S}_\ell^+$  may be non-zero, e. g., because of erroneous previous *a posteriori* speech estimates  $\hat{S}_{\ell-L_p}^{\ell-1}$  or erroneous prediction coefficients  $\mathbf{A}$  (cf. (6.8)). Then, making use of the sampling property of the dirac delta function, the likelihood of speech presence turns out to be the PDF of the noise STFT coefficients shifted by the *a priori* speech estimate (cf. (6.5))

$$p(Y_\ell | \mathbf{Y}_0^{\ell-1}, H_0) = p(Y_\ell | \hat{S}_\ell^+, H_0) = p_D(Y_\ell - \hat{S}_\ell^+) \tag{7.9}$$

with  $p_D(\cdot)$  being the noise PDF, e. g., (2.27).

Using the likelihoods (7.8) and (7.7) the GLR (7.5) turns out to be

$$\Lambda_0^\ell = \frac{P(H_1 | \mathbf{Y}_0^{\ell-1})}{P(H_0 | \mathbf{Y}_0^{\ell-1})} \cdot \frac{\int_{\mathbb{C}} p(Y_\ell | S, H_1) \cdot p(S | \hat{S}_\ell^+, H_1) dS}{p(Y_\ell | \hat{S}_\ell^+, H_0)}. \tag{7.10}$$

Assuming a specific speech prior  $p(S | \hat{S}_\ell^+, H_1)$  and a likelihood  $p(Y_\ell | S, H_1)$  for (7.10), a specific GLR can be calculated. Accordingly, employing Rayleigh-distributed propagation error amplitudes and a statistically independent and uniformly distributed propagation error phase leading to a bivariate Gaussian-distributed speech prior (6.14) and utilizing a Gaussian likelihood (2.28), the GLR (7.10) turns out to be (the derivation can be found in Appendix F)

$$\Lambda_0^\ell \Big|_{\text{R}} = \frac{P(H_1 | \mathbf{Y}_0^{\ell-1})}{P(H_0 | \mathbf{Y}_0^{\ell-1})} \cdot \frac{p(Y_\ell | \mathbf{Y}_0^{\ell-1}, H_1) \Big|_{\text{R}}}{p(Y_\ell | \mathbf{Y}_0^{\ell-1}, H_0) \Big|_{\text{R}}} = \frac{P(H_1 | \mathbf{Y}_0^{\ell-1})}{P(H_0 | \mathbf{Y}_0^{\ell-1})} \cdot \frac{1}{1 + \zeta_\ell} \exp\left(\frac{c_\ell \zeta_\ell}{1 + \zeta_\ell}\right) \tag{7.11}$$

with index R denoting the Rayleigh assumption for the propagation error amplitudes and with  $\zeta_\ell$  and  $c_\ell$  being the *a priori* and *a posteriori* SNR, respectively, as defined in recursive MMSE speech enhancement. Please note the similarity between (7.11) and (2.71): The

resulting GLR (7.11) taking signal history into account has the same form as the GLR in the non-recursive case (2.71).

Further GLRs assuming super-Gaussian speech priors instead of a Gaussian as for (7.11) as well as a Gaussian likelihood are derived in Appendix F.

### 7.2.2 *A Priori* SPP Estimation

Due to the underlying AR speech model (cf. Figure 6.1), the *a priori* SPP  $P(H_1|\mathbf{Y}_0^{\ell-1})$  and the *a priori* SAP  $P(H_0|\mathbf{Y}_0^{\ell-1})$  in (7.11) take signal history into account in the form of the previous observations  $\mathbf{Y}_0^{\ell-1}$ . Please note the difference to the non-recursive case in Section 2.6.1, where the *a priori* SPP is typically a constant (although there are also proposals to track this quantity adaptively). However, in the case of an underlying AR speech model and the corresponding temporal dependencies, the *a priori* SPP turns out to take signal history into account. In order to estimate the *a priori* SPP *recursively*, we assume that the *a priori* speech estimate  $\hat{S}_\ell^+$  is a sufficient statistic for the hypothesis of speech presence  $H_1$  and speech absence  $H_0$ . Accordingly, the *a priori* SPP and the *a priori* SAP can be rewritten as

$$P(H_1|\mathbf{Y}_0^{\ell-1}) = P(H_1|\hat{S}_\ell^+) \quad (7.12)$$

and

$$P(H_0|\mathbf{Y}_0^{\ell-1}) = P(H_0|\hat{S}_\ell^+) = 1 - P(H_1|\hat{S}_\ell^+), \quad (7.13)$$

respectively. Here, we make use of the fact that the *a priori* speech estimate  $\hat{S}_\ell^+(k)$  is able to reliably signalize whether speech is present or absent in the current time-frequency unit  $(\ell, k)$  using signal history. This allows for an efficient adaptation of the *a priori* SPP (and SAP), as we will see in the following.

An *a priori* SAP tracking is proposed, e. g., in [Malah et al., 1999], however, in the context of non-recursive MMSE speech enhancement. Applying the idea to recursive MMSE speech enhancement, the *a priori* SPP (7.12) can be tracked by (cf. [Malah et al., 1999, Eq. (13)])

$$P(H_1|\hat{S}_\ell^+(k)) = \beta_{\text{SPP}} \cdot P(H_1|\hat{S}_{\ell-1}^+(k)) + (1 - \beta_{\text{SPP}}) \cdot I_\ell(k) \quad (7.14)$$

with  $\beta_{\text{SPP}} \in [0, 1]$  being a smoothing factor and with  $I_\ell(k) \in \{0, 1\}$  being the result of a decision rule. The idea here is to adaptively increase (decrease) the *a priori* SPP if speech is present (absent) using the decision  $I_\ell(k)$ .

Assuming that the hypotheses  $H_0$  and  $H_1$  are statistically independent of the spectral phase of the *a priori* speech estimate  $\hat{S}_\ell^+(k)$ , it would be meaningful to base the decision  $I_\ell(k)$  on the spectral amplitudes of the *a priori* speech estimate  $|\hat{S}_\ell^+(k)|$ : Using a small fixed threshold, hypothesis  $H_1$  could be detected if the spectral amplitude of the *a priori* speech

estimate is larger than this threshold. Otherwise, hypothesis  $H_0$  would be chosen. However, this approach would strongly be dependent on the actual signal level of  $\hat{S}_\ell^+(k)$ . Therefore, we argue that a normalized quantity should be utilized for the decision  $I_\ell(k)$ .

In [Malah et al., 1999] the *a posteriori* SNR  $\gamma_\ell(k)$  is employed for the decision process, arguing that attempts to base the decision process on the *a priori* SNR, was not fruitful. Applying the idea to recursive MMSE estimation, we propose to carry out the decision  $I_\ell(k)$  by means of the *a posteriori* SNR  $\varsigma_\ell(k)$  instead of merely the *a priori* speech estimate  $\hat{S}_\ell^+(k)$ . An advantage of  $\varsigma_\ell(k) = |Y_\ell(k) - \hat{S}_\ell^+(k)|^2 / \sigma_{D,\ell}^2(k)$  over  $\hat{S}_\ell^+(k)$  regarding *a priori* SPP estimation is that  $\varsigma_\ell(k)$  is a normalized quantity being independent of actual signal levels. Furthermore, it is still a function of the *a priori* speech estimate  $\hat{S}_\ell^+(k)$ , i. e.,  $\varsigma_\ell(k)$  implicitly contains the *a priori* knowledge carried by  $\hat{S}_\ell^+(k)$  which we aim at exploiting for the decision

$$I_\ell(k) = \begin{cases} 1, & \text{if } \varsigma_\ell(k) > \Phi_\ell(k), \\ 0, & \text{else} \end{cases} \quad (7.15)$$

with the decision threshold  $\Phi_\ell(k)$ .

The decision threshold  $\Phi_\ell(k)$  can be optimized by means of a likelihood ratio test using  $\varsigma_\ell(k)$  [Kay, 1998]: Starting with the *a posteriori* SPP  $P(H_1|\varsigma_\ell(k))$  and the *a posteriori* SAP  $P(H_0|\varsigma_\ell(k))$ , the decision rule for a 2-class problem is as follows

$$P(H_1|\varsigma_\ell(k)) \underset{H_0}{\overset{H_1}{\gtrless}} P(H_0|\varsigma_\ell(k)) \quad (7.16)$$

or, alternatively, after applying Bayes' rule and neglecting  $p(\varsigma_\ell(k))$  which does not affect the decision rule

$$p(\varsigma_\ell(k)|H_1) \cdot P(H_1) \underset{H_0}{\overset{H_1}{\gtrless}} p(\varsigma_\ell(k)|H_0) \cdot P(H_0) \quad (7.17)$$

with  $p(\varsigma_\ell(k)|H_1)$  and  $p(\varsigma_\ell(k)|H_0)$  being the likelihood of speech presence and absence, respectively. According to (7.17), hypothesis  $H_1$  is chosen if the weighted likelihood of speech presence  $p(\varsigma_\ell(k)|H_1) \cdot P(H_1)$  is larger than the weighted likelihood of speech absence  $p(\varsigma_\ell(k)|H_0) \cdot P(H_0)$ . Meanwhile, hypothesis  $H_0$  is preferred if this relation is inverted. Rearranging this decision rule by introducing the likelihood ratio  $p(\varsigma_\ell(k)|H_1)/p(\varsigma_\ell(k)|H_0)$ , the likelihood ratio test turns out to be

$$\frac{p(\varsigma_\ell(k)|H_1)}{p(\varsigma_\ell(k)|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)}. \quad (7.18)$$

The derivation of the likelihoods  $p(\varsigma_\ell(k)|H_1)$  and  $p(\varsigma_\ell(k)|H_0)$  is given in Appendix G. Employing their ratio (G.10) for the likelihood ratio test (7.18) results in

$$\frac{1}{1 + \zeta_\ell(k)} \cdot \exp\left(\frac{\varsigma_\ell(k) \cdot \zeta_\ell(k)}{1 + \zeta_\ell(k)}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)}. \quad (7.19)$$

The decision threshold is at *a posteriori* SNRs at which both sides of (7.19) are equal, i. e.,

$$\Phi_\ell(k) = \arg \min_{\zeta_\ell(k)} \left\{ \left| \frac{p(\zeta_\ell(k)|H_1)}{p(\zeta_\ell(k)|H_0)} - \frac{P(H_0)}{P(H_1)} \right| \right\} = \frac{1 + \zeta_\ell(k)}{\zeta_\ell(k)} \cdot \ln \left( \frac{P(H_0)}{P(H_1)} (1 + \zeta_\ell(k)) \right). \quad (7.20)$$

It was shown in [Gerkmann et al., 2008] and [Gerkmann, 2010] (cf. Section 2.6.2) that an adaptive decision threshold being a function of the *a priori* SNR  $\zeta_\ell(k)$  is rather disadvantageous. Assuming  $P(H_1) = P(H_0) = 0.5$ , the right side of (7.19) becomes one and the left side yields also one, if speech is absent, i. e.,  $\zeta_\ell(k) = 0$ . Thus, in speech absence  $H_0$  no clear decision between  $H_0$  and  $H_1$  is possible. This can be observed in Figure 7.1: The adapted threshold  $\Phi_\ell(k)$  follows the *a posteriori* SNR  $\zeta_\ell(k)$  while speech is absent (the employed recordings consist of two utterances, the first one between frame indices 20 and 170, the second one between frame indices 280 and 430). As a consequence, the *a priori* SPP  $P(H_1|\hat{S}_\ell^+(k))$  is not decreased during speech absence, resulting in values larger than 0.5, although speech is not present. Furthermore, the *a posteriori* SPP estimator outputs the *a priori* SPP estimate in speech absence (just as in the case of non-recursive MMSE estimation, cf. Section 2.6.2 and Chapter 5), instead of values close to zero.

To overcome this issue, in [Gerkmann and Hendriks, 2012] a fixed *a priori* SNR is proposed for the decision threshold (7.20) in the context of non-recursive MMSE estimation. As a consequence, the likelihood ratio in (7.19) successfully yields values larger than one during speech presence and values smaller than one in speech absence, resulting in more robust detection results. The fixed *a priori* SNR is optimized in [Gerkmann and Hendriks, 2012] w. r. t. misdetections (cf. Chapter 5) and the optimal fixed *a priori* SNR yields 15 dB. Please note that the optimization process in [Gerkmann and Hendriks, 2012] is based on Gaussian assumptions leading to the same likelihood ratio as in (7.19), however, with *a priori* and *a posteriori* SNRs as defined in non-recursive MMSE estimation. Therefore, a fixed *a priori* SNR optimization using the likelihood ratio in (7.19) leads to the same optimal fixed *a priori* SNR  $\zeta_{\text{fix}} = 15$  dB. Applying this fixed *a priori* SNR to (7.20) results in a fixed decision threshold  $\Phi_{\text{fix}} = 5.5$  dB.

We repeated the simulations whose results are depicted in Figure 7.1 with the fixed threshold  $\Phi_{\text{fix}}$  for decision  $I_\ell(k)$ . The result is shown in Figure 7.2. As can be seen in the third figure from the top, the initial value of the *a priori* SPP  $P(H_1|\hat{S}_\ell^+(k))$  was set to 0.5. Then, starting from this value the *a priori* SPP steadily decreases during the speech absence prior to the first sentence. During the first utterance,  $P(H_1|\hat{S}_\ell^+(k))$  increases and achieves values larger than 0.5. After the first utterance, the *a priori* SPP decreases again and yields values below 0.5. After a certain period of time, even close to zero values are possible. The same behavior can be observed for the second utterance. As can be seen, the proposed algorithm is able to adaptively track the *a priori* SPP.

## 7.3 Performance Evaluation

In this section we will evaluate the performance of recursive MMSE estimation under SPU. Moreover, in order to see the effect of SPU estimation we will also evaluate the recursive MMSE estimation approach without SPU from Section 6.2. The reference approach is, therefore, the Kalman filter (6.15)-(6.20) employing an *a priori* estimation step consisting of the prediction step (6.9) and the prediction coefficients update (6.13). We will denote this reference approach in the following as rec-R-STS standing for a recursive estimation approach working in the MMSE STS estimation domain and based on a Rayleigh distribution assumed for the propagation error amplitudes (Gaussian speech prior). The proposed approach under SPU, denoted by rec-R-STS-SPU in the following, is of the form (7.2) and consists of a common speech estimator being the reference approach rec-R-STS and an *a posteriori* SPP estimator (7.6) with the GLR (7.11) and the adaptive *a priori* SPP tracking (7.14). For the decision  $I_\ell(k)$  in (7.15) a fixed threshold  $\Phi_{\text{fix}} = 5.5$  dB is used for all frequencies.

Both evaluated approaches are based on the Kalman filter (6.15)-(6.20) employing *a priori* speech estimation (6.9) consisting of a prediction step and a prediction coefficients update. Parameter  $L_p$  of the prediction step (6.9) and the parameters  $\mu$  and  $\Delta$  of the prediction coefficients update (6.13) were trained as follows: The step size was set to  $\mu = 1$  being an ideal parameter value if there is no interference, i. e., the input of the NLMS algorithm (here: prediction error  $\widehat{E}_\ell(k) = \widehat{S}_\ell(k) - \widehat{S}_\ell^+(k)$ ) is not distorted [Hänsler and Schmidt, 2005]. The prediction order  $L_p$  and the regularization parameter  $\Delta$  were optimized by means of the prediction gain (cf. [Vary and Martin, 2006], [Esch, 2012])

$$G_p = \frac{\sum_{\ell=0}^{N_L-1} \sum_{k=0}^{N_K-1} |\widehat{S}_\ell(k)|^2}{\sum_{\ell=0}^{N_L-1} \sum_{k=0}^{N_K-1} |\widehat{S}_\ell(k) - \widehat{S}_\ell^+(k)|^2} \quad (7.21)$$

with  $N_L$  and  $N_K$  being the total number of frames and frequency bins, respectively. Within an STFT-domain Kalman filter structure (cf. (6.15)-(6.20)), speech signals distorted by car noise with an input SNR of 20 dB were processed and the prediction gain (7.21) was measured. These simulations were repeated for a large number of different parameter values  $L_p$  and  $\Delta$ . The parameter set producing the largest prediction gain was considered as optimal and yielded the values  $L_p = 1$  and  $\Delta = 10^{10}$  (please note that  $\Delta$  is a signal-level-dependent quantity: The parameter training was carried out by 16 bit PCM signals). The small prediction order  $L_p = 1$  may result from the fact that the employed frame overlap was 50% at a frame length of 256 samples and a sampling frequency of 8 kHz (cf. Section 3.1.2) which may be close to the span of correlation for speech signals.

In order to assess both the reference and the proposed approach by instrumental measures and to compare the respective performance to each other, we utilized a white-box test (cf.

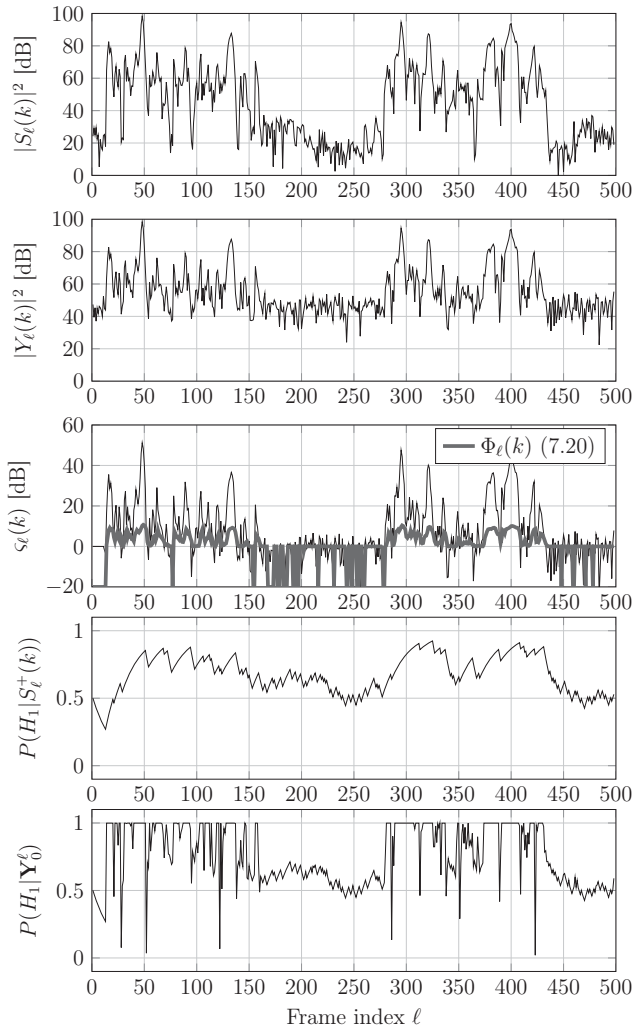


Figure 7.1: Example curves for *a posteriori* SPP estimation using adaptive *a priori* SPP estimation (7.14) with an adaptive decision threshold (7.20) for (7.15) (SNR=20 dB,  $k=32$  ( $f=1000$  Hz))

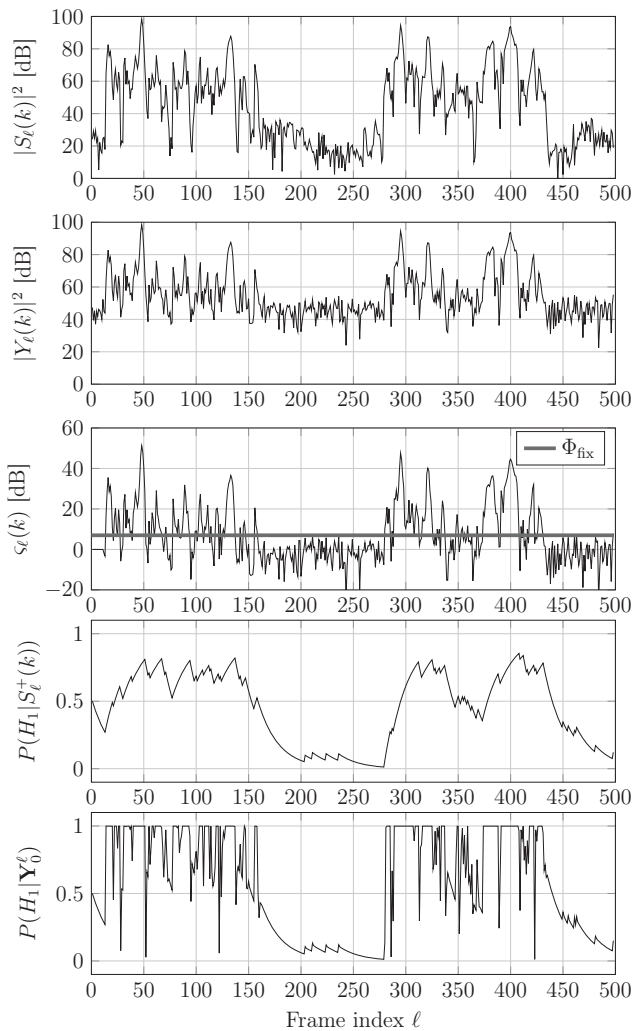


Figure 7.2: Example curves for *a posteriori* SPP estimation using adaptive *a priori* SPP estimation (7.14) with  $\beta_{\text{SPP}} = 0.95$  and a fixed decision threshold  $\Phi_{\text{fix}} = 5.5$  dB for (7.15) (SNR=20 dB,  $k=32$  ( $f=1000$  Hz))

Section 3.1.3). Please note, that both the recursive MMSE estimation formula (7.2) and the Kalman filter equations (6.15)-(6.20) including the *a priori* speech estimation step (6.9) are linear, i.e., they represent  $\hat{S} = f(Y)$  with  $f(\cdot)$  being a linear function, therefore, a signal-component-wise evaluation is possible. Accordingly, we applied the recursive MMSE estimation formula under SPU (7.2) to the speech component  $S_\ell(k)$  and the noise component  $D_\ell(k)$ , resulting in the processed speech component  $\tilde{S}_\ell(k) = P(H_1|\mathbf{Y}_0^\ell) \cdot (\tilde{S}_\ell^+(k) + K_\ell(k) \cdot [S_\ell(k) - \tilde{S}_\ell^+(k)])$  and the processed noise component  $\tilde{D}_\ell(k) = P(H_1|\mathbf{Y}_0^\ell) \cdot (\tilde{D}_\ell^+(k) + K_\ell(k) \cdot [D_\ell(k) - \tilde{D}_\ell^+(k)])$ , respectively (in case of the reference approach rec-R-STS,  $P(H_1|\mathbf{Y}_0^\ell) \equiv 1$  holds). The speech and noise components of the *a priori* speech estimate  $\tilde{S}_\ell^+(k)$  and  $\tilde{D}_\ell^+(k)$  were calculated by applying the prediction coefficients from (6.9) to the  $L_p$  last processed speech and noise components, respectively. The instrumental measures from Section 3.2 were calculated by means of the the processed speech component  $\tilde{S}_\ell(k)$  and the processed noise component  $\tilde{D}_\ell(k)$ .

The noise PSD estimate  $\widehat{\sigma}_{\tilde{D},\ell}^2(k)$  was calculated by the MS approach (cf. Section 2.7.1). The smoothing parameter in (6.20) was  $\beta = 0.99$  [Esch, 2012], the *a priori* SPP was initialized with  $P(H_1|\hat{S}_\ell^+) = 0.5$ , and the parameter to smooth the *a priori* SPP in (7.14) was set to  $\beta_{\text{SPP}} = 0.95$  [Malah et al., 1999].

The performance evaluation results are depicted in Figure 7.3. Just as in the previous figures, the resulting  $\text{NA}_{\text{seg}}$ ,  $\text{SSDR}_{\text{seg}}$ , and LKR values are shown in the upper, middle, and lower figures, respectively. Furthermore, car noise, factory noise, and babble noise results can be found in the left, middle, and right column of figures, respectively.

Analog to the case of non-recursive MMSE speech enhancement (cf. Figures 4.4, 4.5, 4.6, and 5.3) the proposed approach with SPU achieves a larger amount of noise reduction at all input SNR levels and for all employed noise types, reflected by larger  $\text{NA}_{\text{seg}}$  values in the upper figures in Figure 7.3. Furthermore, the approach rec-R-STS-SPU shows a slightly lower speech component quality as the rec-R-STS approach without SPU, reflected by slightly smaller  $\text{SSDR}_{\text{seg}}$  values in the middle figures. Finally, the proposed approach with SPU rec-R-STS-SPU yields a slightly larger amount of musical noise than rec-R-STS, reflected by slightly larger LKR values as shown in the lower plots. These are very similar results to those of the non-recursive MMSE approaches in Chapters 4 and 5: Utilizing SPU estimation allows for achieving significantly smaller residual noise levels at the cost of slight speech component quality degradation and slightly increased musical noise levels. Please note that the difference between the  $\text{NA}_{\text{seg}}$  values achieved by non-SPU and SPU approaches shown in the upper plots are not as large as in the case of non-recursive MMSE estimation (cf. Figures 4.4, 4.5, 4.6, and 5.3).

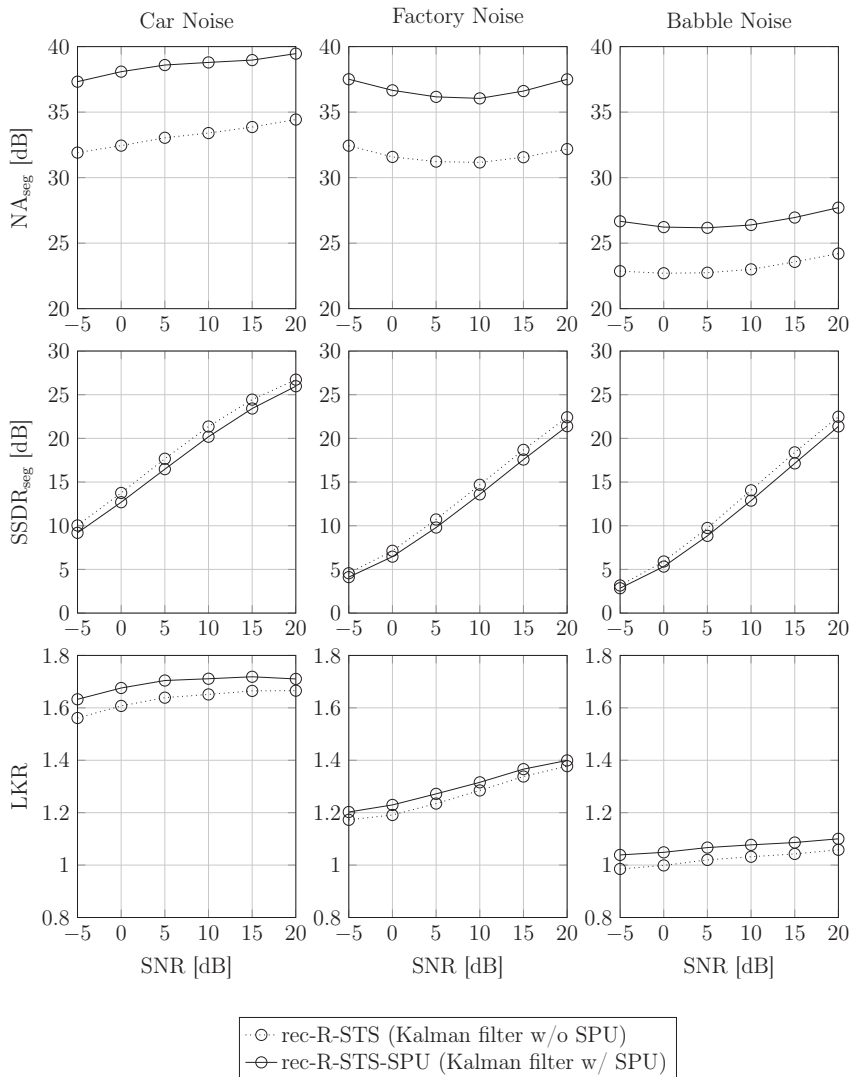


Figure 7.3: Performance evaluation results of a recursive MMSE STS estimator without SPU (rec-R-STs) and a recursive MMSE STS estimator with SPU (rec-R-STs-SPU) both based on a Rayleigh distribution assumed for the propagation error amplitudes (Gaussian prior) in terms of segmental noise attenuation ( $NA_{\text{seg}}$ , the larger the better), segmental speech-to-speech distortion ratio ( $SSDR_{\text{seg}}$ , the larger the better), and weighted log-kurtosis ratio (LKR, the smaller the better)

## 7.4 Summary

In this chapter recursive MMSE estimation under SPU was introduced, to the best of our knowledge, for the first time. It was shown that, just as in the non-recursive case, the recursive MMSE estimator under SPU is a product of a common recursive MMSE estimator for speech, e.g., the Kalman filter, and an *a posteriori* SPP estimator. Furthermore, a general formula for *a posteriori* SPP estimation was derived which, similar to the common recursive MMSE estimator for speech, turned out to be a function of the current and the previous observations. Different from non-recursive MMSE estimation, the *a priori* SPP (and SAP) being a function of the *a posteriori* SPP turned out to be dependent on signal history, therefore, an adaptive algorithm to track the *a priori* SPP was proposed. This algorithm was inspired by an approach known from literature employed in non-recursive speech enhancement.

By assuming either a Gaussian or a super-Gaussian distribution for the propagation error, specific *a posteriori* SPP estimators were derived. Interestingly, the resulting estimators have a very similar form as their non-recursive equivalents from Chapter 2. Then, the *a posteriori* SPP estimator with a Gaussian propagation error assumption was evaluated in conjunction with a Kalman filter as common recursive MMSE estimator. It was shown by performance evaluation that SPU estimation can successfully enhance recursive MMSE estimation by achieving a significantly larger amount of noise attenuation. At the same time, slight speech component quality degradation and slightly larger amount of musical noise have to be faced. It can be concluded that SPU estimation in recursive MMSE speech enhancement provides a similar contribution as SPU estimation in classical non-recursive MMSE estimation.

# Chapter 8

## Conclusions

This thesis deals with MMSE speech enhancement schemes. More specifically, there is a strong focus on statistical modeling of speech STFT coefficients and resulting estimators. Extending a generalized speech spectral amplitude PDF from the literature to complex arguments, a new bivariate generalized gamma speech prior was proposed. Based on the new speech prior, which covers typically employed Gaussian and super-Gaussian speech models, a synopsis of MMSE speech enhancement approaches was given. This also took the estimation domains, such as STS, STSA, and LSA estimation domain, into account.

As an improvement of MMSE speech enhancement we considered SPU estimation. Based on the new bivariate generalized gamma speech model, we derived a new generalized *a posteriori* SPP estimator. Furthermore, as a special case of the generalized estimator, a new super-Gaussian *a posteriori* SPP estimator was proposed. Moreover, we argued that according to estimation theory, the same PDF assumptions have to be applied to both the common MMSE estimator and the *a posteriori* SPP estimator. Therefore, based on the new bivariate generalized speech prior, a synopsis of PDF-consistent MMSE speech enhancement approaches under SPU was given in analogy to the synopsis of common MMSE estimation approaches in different estimation domains. These approaches were evaluated and it turned out that the new super-Gaussian *a posteriori* SPP estimator in conjunction with a PDF-consistent common MMSE estimator outperforms the reference approaches w. r. t. residual noise levels in almost all estimation domains and noise conditions at the cost of a decreased amount of musical noise.

As an improvement to SPU approaches, we also took a look at *a posteriori* SPP estimation with averaging and fixed prior parameters. More precisely, we replaced the Gaussian speech model of the reference approach from literature by a super-Gaussian assumption for speech. Accordingly, the enhanced *a posteriori* SPP estimator makes use of the advantages of a super-Gaussian speech prior, averaged observations, and fixed prior parameters. The proposed

approach was shown to perform better than the reference and, as a further advantage, it achieves low musical noise levels comparable with non-SPU approaches.

Amongst non-recursive MMSE approaches, this thesis also deals with recursive MMSE estimation. Particularly, we investigated some links between speech enhancement and error concealment based on recursive MMSE estimation. Since both disciplines deal with noisy (speech) signals, there are many interesting commonalities between those disciplines. Analyzing the differences, however, a general strength of error concealment over speech enhancement could be identified. Based on this finding, possible research perspectives were sketched for speech enhancement inspired by error concealment. Furthermore, a synopsis of recursive MMSE estimation was given in both disciplines based on consistent PDF-based notation similar to the description of non-recursive approaches.

To the best of our knowledge, this PDF-based description of recursive MMSE estimation for speech enhancement (leading to the classical Kalman filter equations as special case) is new and provides a basis for applying PDF-based enhancement techniques from non-recursive speech enhancement to the recursive case. Accordingly, as a PDF-related improvement, we extended recursive MMSE estimation taking SPU into account. We defined a general recursive MMSE estimation formula under SPU which turned out to be a product of a recursive MMSE estimator for speech and an *a posteriori* SPP estimator. Just as in the non-recursive case, the *a posteriori* SPP is a function of *a priori* SPP which, different from the non-recursive case, is dependent on signal history. Therefore, we proposed an *a priori* SPP tracking algorithm inspired by an approach known from literature employed in non-recursive speech enhancement. Furthermore, we derived specific common recursive MMSE speech estimators and specific *a posteriori* SPP estimators based on either Gaussian or super-Gaussian speech models. The results of a performance evaluation showed that SPU estimation in the context of recursive MMSE estimation has similar effects on the signal components as in the non-recursive case in Chapters 4 and 5.

We also proposed a new reference-free SNR measurement approach which aims at estimating the SNR of a speech signal distorted by car noise as close as possible to the ITU-T P.56 reference. The proposed approach provides small estimation errors and the measurement results show a high correlation with the P.56 reference. Furthermore, it provides relaxed computational complexity and can be applied to narrowband or wideband signals. Within ITU-T Study Group 12, the Focus Group on Car Communication (FG CarCOM) has decided to adopt the new reference-free SNR measurement approach within the draft of an ITU-T recommendation.

# Appendix A

## Bivariate and Polar Description of the Speech Prior

This appendix aims at showing that  $p_S(S) = \frac{1}{A} \cdot p_{A,\alpha}(A, \alpha) = \frac{1}{A} \cdot p_A(A) \cdot p_\alpha(\alpha)$  holds if

- the bivariate PDF of speech STFT coefficients  $p_S(S)$  is rotationally symmetric,
- the PDFs of the speech spectral amplitude  $p_A(A)$  and phase  $p_\alpha(\alpha)$  are statistically independent, and
- the speech spectral phase  $\alpha$  is uniformly distributed on  $[0, 2\pi)$ .

The (cumulative) distribution function of the speech amplitude  $A = |S|$  can be calculated by integrating the joint PDF  $p_S(S = Ae^{j\alpha})$  over a circle with the radius  $A$  [Papoulis and Pillai, 2002]. Furthermore, the PDF of the speech amplitude can be calculated by the first derivative of this (cumulative) distribution function w. r. t.  $A$ . Therefore,  $p_A(A)$  can be written by employing polar integration with  $S = Ae^{j\alpha}$  and  $dS = A d\alpha dA$  as

$$p_A(A) = \frac{\partial}{\partial A} \int_0^A \int_0^{2\pi} p_S(\tilde{A}e^{j\alpha}) \tilde{A} d\alpha d\tilde{A}. \quad (\text{A.1})$$

Using that  $\frac{\partial}{\partial \psi} \int_0^\psi g(x) dx = g(\psi)$ , (A.1) reduces to

$$p_A(A) = A \int_0^{2\pi} p_S(Ae^{j\alpha}) d\alpha. \quad (\text{A.2})$$

Considering the property of rotational symmetry of the speech prior  $p_S(S)$ , i. e.,  $p_S(Ae^{j\alpha}) \neq f(\alpha)$  and  $\int_0^{2\pi} p_S(Ae^{j\alpha}) d\alpha = 2\pi \cdot p_S(Ae^{j\alpha})$ , (A.2) can be rewritten as

$$p_A(A) = A \cdot 2\pi \cdot p_S(Ae^{j\alpha}). \quad (\text{A.3})$$

Extending both sides of the equation (A.3) by  $p_\alpha(\alpha) = \frac{1}{2\pi}$  results in

$$p_A(A) \cdot p_\alpha(\alpha) = A \cdot 2\pi \cdot p_S(Ae^{j\alpha}) \cdot \frac{1}{2\pi} = A \cdot p_S(S). \quad (\text{A.4})$$

Hence, under the assumptions listed above, the PDF of speech STFT coefficients  $p_S(S)$  can be expressed using the speech spectral amplitude PDF  $p_A(A)$  and the speech spectral phase PDF  $p_\alpha(\alpha)$  as

$$p_S(S = Ae^{j\alpha}) = \frac{1}{A} \cdot p_A(A) \cdot p_\alpha(\alpha). \quad (\text{A.5})$$

# Appendix B

## Approaches to PDF Parameter Identification

This appendix briefly summarizes approaches to identify speech spectral amplitude PDF parameters (cf. Table 2.1). A typical approach is to measure a histogram of a large amount of speech data and, subsequently, perform curve fitting using the measured histogram and the PDF whose parameters have to be identified. Since speech STFT coefficients  $S_\ell(k)$  are naturally non-stationary and not ergodic, their statistical moments, such as the variance of  $S_\ell(k)$ , are time-varying. However, an accurate histogram measurement requires a large amount of stationary data. Therefore, the crucial part of PDF parameter identification is the measurement of a speech spectral (amplitude) histogram. Relevant literature on speech spectral histogram measurement can be found, e. g., in [Martin, 2002], [Lotter, 2004], [Martin, 2005], [Lotter and Vary, 2005], and [Gerkmann and Martin, 2010]. Once a speech spectral histogram has been measured, the PDF parameters can be identified. In [Lotter, 2004] and [Lotter and Vary, 2005], the PDF parameters are obtained by minimizing the Kullback-Leibler divergence [Kullback, 1959] between the measured (and normalized) histogram and the analytical PDF.

In [Martin, 2002] and [Martin, 2005] a histogram of the real part of speech STFT coefficients  $\text{Re}\{S_\ell(k)\}$  is measured, while in [Lotter and Vary, 2005] and [Lotter, 2004] a histogram of speech spectral amplitudes  $A_\ell(k) = |S_\ell(k)|$  is determined. In [Gerkmann and Martin, 2010] both quantities are investigated. In this appendix, we will introduce the measurement steps for the spectral amplitudes only. The measurement consists of basically the same steps for the real part of speech STFT coefficients as for speech spectral amplitudes, therefore, a histogram for the former can be obtained in a similar way.

## Speech Histogram Measurement Based on Selected Data

In [Lotter, 2004] and [Lotter and Vary, 2005] a histogram of the absolute value of speech STFT coefficients is measured as follows: Taking speech data from a database, wideband noise  $d(n)$  such as white noise is superimposed artificially to the speech signal  $s(n)$  in such a way that the resulting noisy speech signal  $y(n) = s(n) + d(n)$  has a high SNR, e.g., 40 dB. Then, the noisy speech  $y(n)$  and the clean speech  $s(n)$  are transformed into the STFT domain using the same synthesis parameters (frame length, frame shift, analysis window) as for common processing, resulting in the noisy speech STFT coefficients  $Y_\ell(k)$  and the clean speech STFT coefficients  $S_\ell(k)$ , respectively. Then, the noisy speech  $Y_\ell(k)$  is processed as in a usual speech enhancement system: The noise PSD  $\widehat{\sigma}_{D,\ell}^2(k)$  is tracked by a noise PSD estimator and then the *a priori* SNR is calculated, e.g., by the decision-directed estimator (2.88) with the same smoothing parameter as for later processing. If the resulting *a priori* SNR  $\xi_\ell(k)$  is within a predefined range, e.g., 19-21 dB, the absolute value of the corresponding *clean speech* STFT coefficient  $S_\ell(k)$  becomes part of the histogram data pool. This selecting procedure ensures that the pool contains quasi-stationary data. After normalizing the variance of the pool data to one, a histogram can be measured. Please note that this approach can also be carried out in a frequency-discriminant way: In [Lotter, 2004] only speech STFT coefficients  $S_\ell(k)$  with frequency bins  $k$  belonging to frequencies within 500...2000 Hz were considered for histogram measurement, in order to collect data with essential speech information.

For the histogram measurement of speech spectral amplitudes, we first define  $M$  intervals

$$\mathcal{I}(m) = \begin{cases} [0, \mathcal{A}(m) + \frac{\Delta}{2}) & \text{for } m = 1, \\ [\mathcal{A}(m) - \frac{\Delta}{2}, \mathcal{A}(m) + \frac{\Delta}{2}) & \text{for } m = \{2, 3, \dots, M-1\}, \\ [\mathcal{A}(m) - \frac{\Delta}{2}, \infty) & \text{for } m = M \end{cases} \quad (\text{B.1})$$

with  $m \in \{1, 2, \dots, M\}$ ,  $\mathcal{A}(m) \geq 0$ , and  $\Delta$  being the histogram bin index, the center of the  $m$ th histogram bin being a specific spectral amplitude value, and the width of intermediate intervals (i. e., intervals with index  $m \in \{2, 3, \dots, M-1\}$ ), respectively. A meaningful choice for the first histogram bin center is  $\mathcal{A}(1) = \Delta/2$  allowing for interval width  $\Delta$  also for the first interval  $\mathcal{I}(1)$ . For the remaining centers holds  $\mathcal{A}(m) = \mathcal{A}(m-1) + \Delta$  with  $m = 2, 3, \dots, M$ .

During histogram measurement, each data pool entry is analyzed and a counter  $h(m)$  for the  $m$ th histogram bin is incremented by one, if the absolute value of the currently measured speech STFT coefficient is within interval  $\mathcal{I}(m)$ . After histogram measurement,  $h(m)$  represents the *total number* of speech spectral amplitudes from the data pool which are within  $\mathcal{I}(m)$ . In order to obtain the relative frequency of speech spectral amplitudes,  $h(m)$  has to be divided by the total number of pool data  $\sum_{m=1}^M h(m)$ . Then,  $h(m)/\sum_{m=1}^M h(m)$

describes the *empirical probability* that an arbitrary speech spectral amplitude from the data pool is within the interval  $\mathcal{I}(m)$ . However, instead of empirical probabilities we are interested in the *probability density* of speech spectral amplitudes at arguments  $\mathcal{A}(m)$ . Therefore, the empirical probabilities  $h(m)/\sum_{m=1}^M h(m)$  should be divided by the interval width  $\Delta$ . Accordingly, using the histogram counts  $h(m)$ , the empirical speech spectral amplitude PDF for argument  $\mathcal{A}(m)$  can be calculated as

$$p_H(\mathcal{A}(m)) = \frac{h(m)}{\Delta \cdot \sum_{m=1}^M h(m)}. \quad (\text{B.2})$$

Please note that (B.2) is an inaccurate estimator of the true speech spectral amplitude PDF for argument  $\mathcal{A}(M)$ , because the corresponding interval  $\mathcal{I}(M)$  has an infinite length instead of  $\Delta$ . Therefore,  $p_H(\mathcal{A}(M))$  should not be employed for PDF parameter identification.

## Speech Histogram Measurement Based on All Data

It is argued in [Gerkmann and Martin, 2010] that the speech spectral histograms should be measured incorporating all speech STFT coefficients belonging to speech presence, instead of just selected ones in the aforementioned approach. Furthermore, it is proposed to collect variance-normalized contributions  $|S_\ell(k)|/\sigma_{S,\ell}(k)$  in order to obtain quasi-stationary data. Accordingly, the histogram pool data can be acquired as follows: First, speech data (no noise data is required) is transformed into the STFT domain with the same signal analysis parameters as for later processing. The resulting speech STFT coefficients  $S_\ell(k)$  are employed then for speech PSD estimation employing a modified decision-directed estimator assuming no acoustic noise [Gerkmann and Martin, 2010]

$$\sigma_{S,\ell}^2(k) = \beta_{\text{DD}} \cdot |S_{\ell-1}(k)|^2 + (1 - \beta_{\text{DD}}) \cdot |S_\ell(k)|^2 \quad (\text{B.3})$$

with  $\beta_{\text{DD}}$  being the same smoothing parameter as employed for *a priori* SNR estimation in (2.88). Then, the absolute value of each speech STFT coefficient  $S_\ell(k)$  is normalized by the square root of (B.3) and  $|S_\ell(k)|/\sigma_{S,\ell}(k)$  becomes part of the histogram data pool if speech is present in the corresponding time-frequency unit  $(\ell, k)$ . Speech presence is detected for  $(\ell, k)$ , if  $|S_\ell(k)|$  is larger than or equal to a constant threshold<sup>2</sup>. Due to variance normalization, the resulting pool data exhibits unity variance and its histogram can be measured as described

<sup>1</sup>Here, we assume that  $\Delta$  is small enough so that the true (but unknown) PDF of speech spectral amplitudes can be linearized with a good accuracy within the narrow interval  $\mathcal{I}(m)$ .

<sup>2</sup>In [Gerkmann and Martin, 2010] a file-based threshold is employed: First, a whole speech database file is processed and all corresponding speech spectral amplitudes  $|S_\ell(k)|$  are calculated. Then, the threshold is defined as a value 65 dB below the largest speech spectral amplitude belonging to this particular database file.

in the previous section. The corresponding empirical probability density  $p_H$  can then be obtained by (B.2).

## PDF Parameter Identification

Once the empirical speech spectral amplitude PDF  $p_H(\mathcal{A}(m))$  has been measured, the parameters of an analytical speech spectral amplitude PDF  $p_A(A)$  can be identified. In [Lotter, 2004] and [Lotter and Vary, 2005] it is proposed to optimize the parameters of the analytical PDF  $p_A$  by means of the Kullback-Leibler divergence. Focusing on speech spectral amplitude modeling and assuming a distribution with  $\eta = 1$  or  $\eta = 2$  (cf. Table 2.1), a difference between the empirical speech spectral amplitude PDF  $p_H$  and the analytical PDF  $p_A = f(\nu)$  (e. g., (2.22) or (2.23)) can be measured by means of the symmetric Kullback-Leibler divergence [Kullback, 1959]

$$J_{H,A}(\nu) = \sum_{m=1}^{M-1} \left[ p_H(\mathcal{A}(m)) - p_A(\mathcal{A}(m), \nu) \right] \cdot \ln \left( \frac{p_H(\mathcal{A}(m))}{p_A(\mathcal{A}(m), \nu)} \right). \quad (\text{B.4})$$

Please note that the analytical PDF has to be sampled at the center of the histogram bins  $\mathcal{A}(m)$  when employed for (B.4), i. e., values  $p_A(\mathcal{A}(m))$  has to be calculated for PDF parameter identification. Furthermore, the marginal histogram bin  $m = M$  should not be employed for parameter identification due to side effects.

Then, the optimal shape parameter  $\nu_{\text{opt}}$  is obtained by minimizing (B.4) as

$$\nu_{\text{opt}} = \arg \min_{\nu} J_{H,A}(\nu). \quad (\text{B.5})$$

The corresponding function  $p_A(A, \nu_{\text{opt}})$  can be considered as an optimal speech spectral amplitude PDF estimate (of the pool data) in terms of the Kullback-Leibler divergence.

# Appendix C

## Bivariate and Polar Description of the Likelihood

This appendix aims at showing that  $p_{Y|S}(Y|S) = p_{Y|A,\alpha}(Y|A, \alpha)$ . Please note that for ease of readability the subscript of the PDFs will be omitted in this appendix.

Applying Bayes rule, the likelihood (2.28) can be rewritten as

$$p(Y|S) = p(S|Y) \cdot \frac{p(Y)}{p(S)}. \quad (\text{C.1})$$

It can be shown that the posterior  $p(S|Y)$  is rotationally symmetric if  $p(S)$  and  $p(Y|S)$  (and accordingly  $p(Y)$ , too) are rotationally symmetric. This is fulfilled if a Gaussian likelihood (2.28) and a speech prior with a uniformly distributed phase and statistically independent speech spectral amplitude and phase, e. g., (2.26) are assumed (cf. Section 2.3.1). Thus, (2.18) can be applied to the posterior leading to

$$p(S = Ae^{j\alpha}|Y) = \frac{1}{A} \cdot p(A, \alpha|Y). \quad (\text{C.2})$$

Inserting the resulting posterior into (C.1) and using (2.18) yields

$$\begin{aligned} p(Y|S) &= p(S|Y) \cdot \frac{p(Y)}{p(S)} \\ &= \frac{1}{A} \cdot p(A, \alpha|Y) \cdot \frac{p(Y)}{\frac{1}{A} \cdot p(A, \alpha)} \\ &= p(A, \alpha|Y) \cdot \frac{p(Y)}{p(A, \alpha)} \\ &= p(Y|A, \alpha). \end{aligned} \quad (\text{C.3})$$



# Appendix D

## Derivation of the PDF of Averaged *a Posteriori* SNRs Assuming Chi-Distributed Speech Spectral Amplitudes

In this appendix we will derive the PDF of averaged *a posteriori* SNR values  $\bar{\gamma}$  assuming chi-distributed speech spectral amplitudes. We will start with the PDF of the noisy speech STFT coefficients  $Y$ , then we will derive the PDF of the noisy speech amplitudes  $|Y|$ , the squared amplitudes  $|Y|^2$ , followed by the derivation of the PDF of the *a posteriori* SNR  $\gamma = |Y|^2/\sigma_D^2$ , and the PDF of averaged *a posteriori* SNRs  $\bar{\gamma}$ . Please note that we will omit the window index  $\Theta$  in this appendix for ease of readability.

We assume that the speech spectral amplitudes are chi-distributed (2.22), the statistically independent speech spectral phase is uniformly distributed (super-Gaussian speech prior (2.26) with the parameters  $\eta = 2$ ,  $\beta = \nu_\chi/\sigma_S^2$ , and  $\nu = \nu_\chi$ ), and the noise STFT coefficients  $D$  are Gaussian distributed (2.28). Assuming further that the speech STFT coefficients  $S$  and the noise STFT coefficients  $D$  are statistically independent, the distribution of the noisy speech signal  $Y = S + D$  can be calculated as (cf. [Breithaupt and Martin, 2011, Eq. (20)])

$$p_{Y|H_1}^\chi(Y|H_1) = p(Y|H_0) \cdot \left(\frac{\nu_\chi}{\nu_\chi + \xi}\right)^{\nu_\chi} \cdot {}_1F_1\left(\nu_\chi; 1; \frac{1}{\sigma_D^2} \frac{\xi}{\nu_\chi + \xi} |Y|^2\right) \quad (D.1)$$

with superscript  $\chi$  denoting the assumption of chi-distributed speech spectral amplitudes and with

$$p_{Y|H_0}(Y|H_0) = \frac{1}{\pi\sigma_D^2} e^{-\frac{|Y|^2}{\sigma_D^2}}. \quad (D.2)$$

According to [Papoulis and Pillai, 2002, Example 6-22], the distribution function of  $|Y|$  can

be calculated by integrating (D.1) w. r. t.  $Y$  over a circle with the radius  $|Y|$ . Furthermore, the first derivative of the distribution function is the PDF of  $|Y|$ . Accordingly, using polar integration with  $Y = |Y|e^{j\vartheta}$  and  $dY = |Y| \cdot d\vartheta \cdot d|Y|$  we obtain

$$p_{|Y||H_1}^x \left( |Y| |H_1 \right) = \frac{\partial}{\partial |Y|} \int_0^{|Y|} \int_0^{2\pi} |\tilde{Y}| \cdot p_{\tilde{Y}|H_1}^x (|\tilde{Y}|e^{j\vartheta} |H_1) d\vartheta d|\tilde{Y}|. \quad (\text{D.3})$$

Using  $\frac{\partial}{\partial \rho} \int_0^\rho f(x) dx = f(\rho)$ , and integrating w. r. t. the phase  $\vartheta$ , (D.3) can be rewritten as

$$p_{|Y||H_1}^x \left( |Y| |H_1 \right) = \frac{2|Y|}{\sigma_D^2} \cdot \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\nu_\chi} \cdot e^{-\frac{1}{\sigma_D^2}|Y|^2} \cdot {}_1F_1 \left( \nu_\chi; 1; \frac{1}{\sigma_D^2} \frac{\xi}{\nu_\chi + \xi} |Y|^2 \right). \quad (\text{D.4})$$

Please note that this is a generalization of the Rayleigh PDF; applying  $\nu_\chi = 1$  to (D.4) results in the Rayleigh PDF (2.20).

According to [Papoulis and Pillai, 2002, Example 5-2], the PDF of the square of the random variable  $|Y|$  can be calculated by

$$p_{|Y|^2|H_1}^x \left( |Y|^2 |H_1 \right) = \frac{1}{2|Y|} \cdot p_{|Y||H_1}^x \left( |Y| |H_1 \right), \quad (\text{D.5})$$

thus, using (D.4) the PDF of the squared noisy speech spectral amplitudes turns out to be

$$p_{|Y|^2|H_1}^x \left( |Y|^2 |H_1 \right) = \frac{1}{\sigma_D^2} \cdot \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\nu_\chi} \cdot e^{-\frac{1}{\sigma_D^2}|Y|^2} \cdot {}_1F_1 \left( \nu_\chi; 1; \frac{1}{\sigma_D^2} \frac{\xi}{\nu_\chi + \xi} |Y|^2 \right). \quad (\text{D.6})$$

Following [Papoulis and Pillai, 2002, Eq. (5-18)], the PDF of  $|Y|^2/\sigma_D^2$  with  $\sigma_D^2$  being a deterministic (slowly time-varying) quantity can be obtained from  $p_{|Y|^2|H_1}^x \left( |Y|^2 |H_1 \right)$  as

$$\begin{aligned} p_{\frac{|Y|^2}{\sigma_D^2}|H_1}^x \left( \frac{|Y|^2}{\sigma_D^2} |H_1 \right) &= \sigma_N^2 \cdot p_{|Y|^2|H_1}^x \left( \sigma_N^2 \cdot \frac{|Y|^2}{\sigma_N^2} |H_1 \right) \\ &= \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\nu_\chi} \cdot e^{-\frac{|Y|^2}{\sigma_D^2}} \cdot {}_1F_1 \left( \nu_\chi; 1; \frac{\xi}{\nu_\chi + \xi} \frac{|Y|^2}{\sigma_D^2} \right). \end{aligned} \quad (\text{D.7})$$

Next, utilizing variable substitution  $\gamma = |Y|^2/\sigma_D^2$ , (D.7) can be rewritten as

$$p_{\gamma|H_1}^x (\gamma |H_1) = \left( \frac{\nu_\chi}{\nu_\chi + \xi} \right)^{\nu_\chi} \cdot e^{-\gamma} \cdot {}_1F_1 \left( \nu_\chi; 1; \gamma \frac{\xi}{\nu_\chi + \xi} \right) \quad (\text{D.8})$$

which is the PDF of *a posteriori* SNR values in speech presence  $H_1$  assuming chi-distributed speech spectral amplitudes (denoted by superscript  $\chi$ ) with the shape parameter  $\nu_\chi$ . The *a posteriori* SNR in absence of speech  $H_0$  (special case of (D.8) with  $\xi = 0$ ) is exponentially distributed  $p_{\gamma|H_0}^x (\gamma |H_0) = e^{-\gamma}$ .

As a next step, we aim at deriving the PDF of averaged *a posteriori* SNRs. This can be achieved by mathematical induction (cf. [Nelson, 1995]): The PDF of the sum of two

i. i. d. *a posteriori* SNR values  $\bar{\gamma}_2 = \gamma + \gamma$  can be calculated by convolving their PDFs. Using [Gradshteyn and Ryzhik, 1965, Eq. (7.613.4)] and assuming the PDF (D.8)

$$\begin{aligned} p_{\bar{\gamma}_2|H_1}^{\chi}(\bar{\gamma}_2|H_1) &= \int_0^{\bar{\gamma}_2} p_{\gamma|H_1}^{\chi}(x|H_1) \cdot p_{\gamma|H_1}^{\chi}((\bar{\gamma}_2 - x)|H_1) dx \\ &= \left( \frac{\nu_{\chi}}{\nu_{\chi} + \xi} \right)^{2\nu_{\chi}} \cdot \frac{\Gamma(1)}{\Gamma(2)} \cdot \bar{\gamma}_2 \cdot e^{-\bar{\gamma}_2} \cdot {}_1F_1 \left( 2\nu_{\chi}, 2, \bar{\gamma}_2 \frac{\xi}{\nu_{\chi} + \xi} \right) \end{aligned} \quad (\text{D.9})$$

results as the PDF<sup>1</sup> of  $\gamma + \gamma$ . Then, the PDF of the sum of three i. i. d. *a posteriori* SNR values  $\bar{\gamma}_3 = \bar{\gamma}_2 + \gamma$  can be calculated by the convolution of (D.9) and (D.8). Using [Gradshteyn and Ryzhik, 1965, Eq. (7.613.4)] the PDF of the sum of three *a posteriori* SNRs yields

$$\begin{aligned} p_{\bar{\gamma}_3|H_1}^{\chi}(\bar{\gamma}_3|H_1) &= \int_0^{\bar{\gamma}_3} p_{\bar{\gamma}_2|H_1}^{\chi}(x|H_1) \cdot p_{\gamma|H_1}^{\chi}((\bar{\gamma}_3 - x)|H_1) dx \\ &= \left( \frac{\nu_{\chi}}{\nu_{\chi} + \xi} \right)^{3\nu_{\chi}} \cdot \frac{\Gamma(1)}{\Gamma(3)} \cdot (\bar{\gamma}_3)^2 \cdot e^{-\bar{\gamma}_3} \cdot {}_1F_1 \left( 3\nu_{\chi}, 3, \bar{\gamma}_3 \frac{\xi}{\nu_{\chi} + \xi} \right). \end{aligned} \quad (\text{D.10})$$

Using [Gradshteyn and Ryzhik, 1965, Eq. (7.613.4)] once again, the PDF of the sum of four i. i. d. *a posteriori* SNR values  $\bar{\gamma}_4 = \bar{\gamma}_3 + \gamma$  can be calculated by the convolution of (D.10) and (D.8) resulting in

$$\begin{aligned} p_{\bar{\gamma}_4|H_1}^{\chi}(\bar{\gamma}_4|H_1) &= \int_0^{\bar{\gamma}_4} p_{\bar{\gamma}_3|H_1}^{\chi}(x|H_1) \cdot p_{\gamma|H_1}^{\chi}((\bar{\gamma}_4 - x)|H_1) dx \\ &= \left( \frac{\nu_{\chi}}{\nu_{\chi} + \xi} \right)^{4\nu_{\chi}} \cdot \frac{\Gamma(1)}{\Gamma(4)} \cdot \bar{\gamma}_4^3 \cdot e^{-\bar{\gamma}_4} \cdot {}_1F_1 \left( 4\nu_{\chi}, 4, \bar{\gamma}_4 \frac{\xi}{\nu_{\chi} + \xi} \right). \end{aligned} \quad (\text{D.11})$$

Therefore, using mathematical induction, it can be concluded that the PDF of the sum of  $N$  i. i. d. *a posteriori* SNR values  $\bar{\gamma}_N = \sum_{i=1}^N \gamma_i$  generally is

$$p_{\bar{\gamma}_N|H_1}^{\chi}(\bar{\gamma}_N|H_1) = \left( \frac{\nu_{\chi}}{\nu_{\chi} + \xi} \right)^{N\nu_{\chi}} \cdot \frac{1}{\Gamma(N)} \cdot \bar{\gamma}_N^{N-1} \cdot e^{-\bar{\gamma}_N} \cdot {}_1F_1 \left( N\nu_{\chi}, N, \bar{\gamma}_N \frac{\xi}{\nu_{\chi} + \xi} \right). \quad (\text{D.12})$$

However, instead of the PDF of the sum of *a posteriori* SNRs  $\bar{\gamma}_N$  (D.12), we are interested in the PDF of the average of *a posteriori* SNRs  $\bar{\gamma} = 1/N \cdot \bar{\gamma}_N$ . According to [Papoulis and Pillai, 2002, Eq. (5-18)], this can be obtained by

$$p_{\bar{\gamma}|H_1}^{\chi}(\bar{\gamma}|H_1) = N \cdot p_{\bar{\gamma}_N|H_1}^{\chi}(N \cdot \bar{\gamma}|H_1) \quad (\text{D.13})$$

which finally results in (5.3) if we substitute  $N$  by the number of averaged *a posteriori* SNR values  $\mu$ .

<sup>1</sup>Please note that due to the fact that the argument of  $p_{\gamma}(\gamma)$  in (D.9) is non-negative, the upper bound of the convolution integral (D.9) reduces to  $\bar{\gamma}_2$  from  $\infty$ , cf. [Bronshtein et al., 2007, Eq. (15.93)]



# Appendix E

## Synopsis of Recursive MMSE Estimation

Please note that for ease of readability, we will omit the indices  $\ell$  and  $k$  in this appendix. The aim of this appendix is to derive the Kalman filter equations (6.15)-(6.17) using a PDF notation. Different from the commonly employed matrix notation, a PDF-based description allows for assuming super-Gaussian signal models, but of course, a Gaussian assumption is also applicable.

Employing the speech prior (6.10), (6.11), and the likelihood (6.5), the recursive MMSE STS estimator (6.6) turns out to be

$$\hat{S} = \frac{\int_{\mathbb{C}} S \cdot p_{\bar{E}}(S - \hat{S}^+) \cdot p_D(Y - S) dS}{\int_{\mathbb{C}} p_{\bar{E}}(S - \hat{S}^+) \cdot p_D(Y - S) dS} \quad (\text{E.1})$$

with  $p_{\bar{E}}(\cdot)$  and  $p_D(\cdot)$  being the propagation error PDF and the acoustic noise PDF, respectively. Introducing a new integration variable  $\bar{E} = S - \hat{S}^+$  being the propagation error (cf. (6.11)), (E.1) can be rewritten as

$$\hat{S} = \hat{S}^+ + \frac{\int_{\mathbb{C}} \bar{E} \cdot p_{\bar{E}}(\bar{E}) \cdot p_D(Y - \bar{E} - \hat{S}^+) d\bar{E}}{\int_{\mathbb{C}} p_{\bar{E}}(\bar{E}) \cdot p_D(Y - \bar{E} - \hat{S}^+) d\bar{E}} = \underbrace{\hat{S}^+}_{\text{A priori speech estimate}} + \underbrace{\hat{E}}_{\text{Update}}. \quad (\text{E.2})$$

As can be seen, the recursive MMSE estimator turns out to be a sum of the *a priori* speech estimate  $\hat{S}^+$  and a fraction being the estimation update  $\hat{E}$ . While the *a priori* speech estimate  $\hat{S}^+$  employs the previous observations only (cf. (6.9)), the second term  $\hat{E} = f(Y - \hat{S}^+)$  takes the current observation into account and performs an estimation update, resulting in the *a posteriori* speech estimate  $\hat{S}$ . Please note that  $\hat{E}$  in (E.2) is a classical (non-recursive) MMSE STS estimator, however, with an extra term ' $-\hat{S}^+$ ' in the argument of the PDF  $p_D(\cdot)$  (cf. (2.30)).

Assuming that the amplitude and the phase of the propagation error  $\bar{E}$  are statistically independent and the phase of  $\bar{E}$  is uniformly distributed, i. e., the PDF of the propagation

error  $p_{\bar{E}}(\bar{E})$  is rotationally symmetric, the generalized framework from Chapter 2 based on the bivariate generalized gamma PDF (2.26) can be employed here. Therefore, different estimators can be derived for the update step assuming either a Gaussian or a super-Gaussian propagation error. Interestingly, the resulting estimators will turn out to be the MMSE STS estimators from Section 2.5.1, however, with different variables: The *a priori* SNR  $\xi$  and the *a posteriori* SNR  $\gamma$  are defined in recursive MMSE estimation as (cf. (6.12))

$$\zeta = \frac{\sigma_{\bar{E}}^2 + \sigma_{E^+}^2}{\sigma_D^2} = \frac{\sigma_{\bar{E}}^2}{\sigma_D^2} \quad (\text{E.3})$$

and

$$\varsigma = \frac{|Y - \hat{S}^+|^2}{\sigma_D^2} = \frac{|M|^2}{\sigma_D^2}, \quad (\text{E.4})$$

respectively. Furthermore, within the update step the observation is defined as  $M = Y - \hat{S}^+$ , instead of the noisy speech STFT coefficient  $Y$  only. In the following, we will derive a general formula for the update step based on a generalized speech prior and then specific estimators will be introduced, assuming a specific statistical model for the propagation error.

## Generalized Speech Prior

Employing the new bivariate generalized gamma PDF (2.26) as speech prior  $p_{\bar{E}}(\cdot)$  and a bivariate Gaussian PDF (6.5) as likelihood  $p_D(\cdot)$ , the fraction in (E.2) yields the estimation update (cf. (2.38))

$$\hat{E} = \frac{\int_{\mathcal{C}} \bar{E} \cdot |\bar{E}|^{\eta\nu-2} \cdot e^{-\beta|\bar{E}|^\eta} \cdot e^{-\frac{|Y-\bar{E}-\hat{S}^+|^2}{\sigma_D^2}} d\bar{E}}{\int_{\mathcal{C}} |\bar{E}|^{\eta\nu-2} \cdot e^{-\beta|\bar{E}|^\eta} \cdot e^{-\frac{|Y-\bar{E}-\hat{S}^+|^2}{\sigma_D^2}} d\bar{E}}. \quad (\text{E.5})$$

Employing polar integration with  $\bar{E} = |\bar{E}|e^{j\epsilon}$  and  $d\bar{E} = |\bar{E}|d|\bar{E}|d\epsilon$ , as well as employing  $M = Y - \hat{S}^+ = |M|e^{j\theta}$ , we obtain

$$\hat{E} = \frac{\int_0^\infty \int_0^{2\pi} |\bar{E}|^{\eta\nu} \cdot e^{j\epsilon} \cdot e^{-\beta|\bar{E}|^\eta} \cdot e^{-\frac{|M|^2 + |\bar{E}|^2 - 2|\bar{E}||M|\cos(\epsilon-\theta)}{\sigma_D^2}} d\epsilon d|\bar{E}|}{\int_0^\infty \int_0^{2\pi} |\bar{E}|^{\eta\nu-1} \cdot e^{-\beta|\bar{E}|^\eta} \cdot e^{-\frac{|M|^2 + |\bar{E}|^2 - 2|\bar{E}||M|\cos(\epsilon-\theta)}{\sigma_D^2}} d\epsilon d|\bar{E}|}. \quad (\text{E.6})$$

Integrating w. r. t. the spectral phase  $\epsilon$  using [Gradshteyn and Ryzhik, 1965, (8.431.5)], (E.6) can be rewritten as

$$\hat{E} = \underbrace{\frac{\int_0^\infty |\bar{E}|^{\eta\nu} \cdot e^{-\frac{|\bar{E}|^2}{\sigma_D^2} - \beta|\bar{E}|^\eta} \cdot I_1\left(\frac{2|M|}{\sigma_D^2}|\bar{E}|^2\right) d|\bar{E}|}{\int_0^\infty |\bar{E}|^{\eta\nu-1} \cdot e^{-\frac{|\bar{E}|^2}{\sigma_D^2} - \beta|\bar{E}|^\eta} \cdot I_0\left(\frac{2|M|}{\sigma_D^2}|\bar{E}|\right) d|\bar{E}|}}_{\text{Spectral amplitude estimate}} \cdot \underbrace{e^{j\theta}}_{\text{Spectral phase estimate}}. \quad (\text{E.7})$$

The formula of the update step of recursive MMSE STS estimation based on a generalized speech prior turns out to be the same form as the classical non-recursive MMSE STS estimator (2.39) consisting of well separated amplitude and phase estimates. In order to obtain specific estimators, one of the specific speech priors from Table 2.1 has to be chosen, each having an own parameter set. Therefore, employing the speech priors from Section 2.3.1 for (E.7) results in specific estimators for the update step which will be introduced in the following.

## Gaussian Speech Prior

Assuming the *Rayleigh* distribution for the propagation error amplitudes (Gaussian propagation error), i. e., using the parameters  $\eta = 2$ ,  $\beta = 1/\sigma_E^2$ , and  $\nu = 1$  (cf. Table 2.1), the generalized update step (E.7) turns out to be

$$\hat{E}_R = \frac{\int_0^\infty |\bar{E}|^2 \cdot e^{-|\bar{E}|^2 \left[ \frac{1}{\sigma_E^2} + \frac{1}{\sigma_D^2} \right]} \cdot I_1 \left( \frac{2|M|}{\sigma_D} |\bar{E}| \right) d|\bar{E}|}{\int_0^\infty |\bar{E}| \cdot e^{-|\bar{E}|^2 \left[ \frac{1}{\sigma_E^2} + \frac{1}{\sigma_D^2} \right]} \cdot I_0 \left( \frac{2|M|}{\sigma_D} |\bar{E}| \right) d|\bar{E}|} \cdot e^{je} \quad (\text{E.8})$$

with subscript R denoting the Rayleigh assumption for the the propagation error amplitudes. Integrating w. r. t. the propagation error amplitudes  $|\bar{E}|$  using [Gradshteyn and Ryzhik, 1965, (6.631.1)], (E.8) yields

$$\hat{E}_R = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_D^2} \cdot |M| e^{je} = K_R \cdot (Y - \hat{S}^+) = \frac{\zeta}{1 + \zeta} \cdot (Y - \hat{S}^+) \quad (\text{E.9})$$

with  $\zeta$  and  $K_R = \zeta/(1 + \zeta)$  being the *a priori* SNR and the Kalman gain (6.18), respectively. Please note that the Kalman gain  $K_R$  and the update step  $\hat{E}_R$  can be associated with the Wiener spectral weighting rule  $G_{R\text{-STS}}$  in (2.41) and the MMSE STS estimator (2.41), respectively.

## Super-Gaussian Speech Priors

Assuming the *chi* distribution for the propagation error amplitudes (super-Gaussian distribution for the propagation error), i. e., using the parameters  $\eta = 2$ ,  $\nu = \nu_\chi$ , and  $\beta = \nu_\chi/\sigma_E^2$  (cf. Table 2.1) as well as [Gradshteyn and Ryzhik, 1965, Eqs. (6.631.1), (8.406.3)], the generalized update step (E.7) can be rewritten as (cf. (2.43))

$$\hat{E}_\chi = K_\chi \cdot (Y - \hat{S}^+) = \frac{\nu_\chi \cdot \zeta}{\nu_\chi + \zeta} \cdot \frac{{}_1F_1 \left( \nu_\chi + 1; 2; \frac{\zeta \zeta}{\nu_\chi + \zeta} \right)}{{}_1F_1 \left( \nu_\chi; 1; \frac{\zeta \zeta}{\nu_\chi + \zeta} \right)} \cdot (Y - \hat{S}^+) \quad (\text{E.10})$$

where subscript  $\chi$  denotes the chi assumption for the propagation error amplitudes. Different from the Gaussian case (E.9), (E.10) is also a function of the *a posteriori* SNR  $\zeta$  which is only following a slightly different definition as in the non-recursive case. Please note that, just as in the non-recursive case, (E.9) is a special case of (E.10) with  $\nu_\chi = 1$ . Furthermore, the Kalman gain  $K_\chi$  and the update step  $\hat{E}_\chi$  can be associated with the MMSE STS spectral weighting rule  $G_{\chi\text{-STS}}$  in (2.43) and the MMSE STS estimator (2.43), respectively.

Assuming the *gamma* distribution for the propagation error amplitudes (super-Gaussian distribution for the propagation error), i.e., using the parameters  $\eta = 1$ ,  $\nu = \nu_\Gamma$ , and  $\beta = \sqrt{\nu_\Gamma(\nu_\Gamma + 1)}/\sigma_E$  (cf. Table 2.1), the generalized update step (E.7) results in (cf. (2.44))

$$\hat{E}_\Gamma = \frac{\int_0^\infty |\bar{E}|^\nu \cdot \exp\left(-\frac{1}{\sigma_D^2} |\bar{E}|^2 - \frac{\sqrt{\nu_\Gamma(\nu_\Gamma+1)}}{\sigma_E} |\bar{E}|\right) \cdot I_1\left(\frac{2|M|}{\sigma_D^2} |\bar{E}|^2\right) d|\bar{E}|}{\int_0^\infty |\bar{E}|^{\nu-1} \cdot \exp\left(-\frac{1}{\sigma_D^2} |\bar{E}|^2 - \frac{\sqrt{\nu_\Gamma(\nu_\Gamma+1)}}{\sigma_E} |\bar{E}|\right) \cdot I_0\left(\frac{2|M|}{\sigma_D^2} |\bar{E}|\right) d|\bar{E}|} \cdot e^{je} \quad (\text{E.11})$$

with subscript  $\Gamma$  denoting the gamma assumption for the propagation error amplitudes. Just as in the non-recursive case, the integrals in (E.11) do not have closed form solutions, therefore, either some approximations or numerical methods are required. Just as in Chapter 2 we propose calculating the integrals in (E.11) by quadrature. For convenience, we first rewrite (E.11) using variable substitution with  $|\bar{E}| = k \cdot |M|$  and  $d|\bar{E}| = dk \cdot |M|$  in order to obtain the integrands as a function of  $\zeta$  and  $\varsigma$  as

$$\hat{E}_\Gamma = K_\Gamma \cdot (Y - \hat{S}^+) = \frac{\int_0^\infty k^{\nu_\Gamma} \cdot \exp\left(-\varsigma k^2 - \sqrt{\nu_\Gamma(\nu_\Gamma+1)} \sqrt{\frac{\varsigma}{\zeta}} k\right) \cdot I_1(2\varsigma k) dk}{\int_0^\infty k^{\nu_\Gamma-1} \cdot \exp\left(-\varsigma k^2 - \sqrt{\nu_\Gamma(\nu_\Gamma+1)} \sqrt{\frac{\varsigma}{\zeta}} k\right) \cdot I_0(2\varsigma k) dk} \cdot (Y - \hat{S}^+). \quad (\text{E.12})$$

These integrals can be calculated numerically by, e.g., the Gauss-Kronrod quadrature for a typical range of  $\zeta$  and  $\varsigma$  [Brass and Petras, 2011]. Please note that the Kalman gain  $K_\Gamma$  and the update step (E.12) can be associated with the MMSE STS spectral weighting rule  $G_{\Gamma\text{-STS}}$  in (2.45) and the MMSE STS estimator (2.45), respectively.

# Appendix F

## Synopsis of

## Recursive *A Posteriori* SPP Estimation

The aim of this appendix is to derive specific GLRs for *a posteriori* SPP estimation assuming Gaussian or super-Gaussian propagation errors. Just as in Chapter 4, we will first derive a generalized GLR based on the bivariate generalized gamma PDF as speech prior. Specific GLRs assuming either Gaussian or super-Gaussian distributions for the propagation error will turn out as special case. Please note that for ease of readability, we will omit the indices  $\ell$  and  $k$  in this appendix. However, sequences will still keep their time indices.

### Generalized Speech Prior

Employing the bivariate generalized gamma PDF (2.26) as speech prior  $p(S|\hat{S}^+, H_1)$  and a bivariate Gaussian PDF (6.5) as likelihood  $p(Y|S, H_1)$ , the common GLR formula (7.10) yields (cf. (2.38))

$$\Lambda_0^\ell \Big|_{\text{g}\Gamma} = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{\int_{\mathbb{C}} \frac{1}{\pi\sigma_D^2} \cdot \frac{\eta\beta^\nu}{2\pi\Gamma(\nu)} \cdot e^{-\frac{|Y-\hat{S}^+|^2}{\sigma_D^2}} \cdot |S - \hat{S}^+|^{\nu-2} \cdot e^{-\beta|S-\hat{S}^+|^\eta} dS}{\frac{1}{\pi\sigma_D^2} \cdot e^{-\frac{|Y-\hat{S}^+|^2}{\sigma_D^2}}} \quad (\text{F.1})$$

with subscript g $\Gamma$  denoting the generalized gamma model for the propagation error. By introducing a new integration variable  $\bar{E} = S - \hat{S}^+$  being the propagation error, (F.1) can be rewritten as

$$\Lambda_0^\ell \Big|_{\text{g}\Gamma} = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{\int_{\mathbb{C}} \frac{\eta\beta^\nu}{2\pi\Gamma(\nu)} \cdot e^{-\frac{|Y-\hat{S}^+-\bar{E}|^2}{\sigma_D^2}} \cdot |\bar{E}|^{\nu-2} \cdot e^{-\beta|\bar{E}|^\eta} d\bar{E}}{e^{-\frac{|Y-\hat{S}^+|^2}{\sigma_D^2}}} \quad (\text{F.2})$$

Employing polar integration with  $\bar{E} = |\bar{E}|e^{j\epsilon}$  and  $d\bar{E} = |\bar{E}|d|\bar{E}|d\epsilon$  as well as employing  $M = Y - \hat{S} = |M|e^{j\epsilon}$ , we obtain

$$\begin{aligned} \Lambda_0^\ell \Big|_{\text{gr}} &= \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{\eta\beta^\nu}{2\pi \cdot \Gamma(\nu)} \cdot \int_0^\infty |\bar{E}|^{\eta\nu-1} \cdot e^{-\beta|\bar{E}|^\eta} \frac{\int_0^{2\pi} e^{-\frac{|M|^2 + |\bar{E}|^2 - 2|M||\bar{E}|\cos(\theta-\epsilon)}{\sigma_D^2}} d\epsilon d|\bar{E}|}{e^{-\frac{|M|^2}{\sigma_D^2}}} d|\bar{E}| \\ &= \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{\eta\beta^\nu}{2\pi \cdot \Gamma(\nu)} \cdot \int_0^\infty |\bar{E}|^{\eta\nu-1} \cdot e^{-\beta|\bar{E}|^\eta} \cdot e^{-\frac{|E|^2}{\sigma_D^2}} \int_0^{2\pi} e^{\frac{2|M||\bar{E}|\cos(\theta-\epsilon)}{\sigma_D^2}} d\epsilon d|\bar{E}|. \end{aligned} \quad (\text{F.3})$$

Integrating w. r. t. the spectral phase  $\epsilon$  using [Gradshteyn and Ryzhik, 1965, (8.431.5)], (F.3) can be rewritten as

$$\Lambda_0^\ell \Big|_{\text{gr}} = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} \cdot \int_0^\infty |\bar{E}|^{\eta\nu-1} \cdot e^{-\beta|\bar{E}|^\eta} \cdot e^{-\frac{|E|^2}{\sigma_D^2}} \cdot I_0\left(\frac{2|M||\bar{E}|}{\sigma_D^2}\right) d|\bar{E}|. \quad (\text{F.4})$$

This is a generalized GLR formula for recursive MMSE estimation under SPU. The corresponding generalized *a posteriori* SPP estimator can be obtained by (7.6). Please note the similarity to (4.3) being the non-recursive equivalent of (F.4). In the following, we will derive specific GLRs assuming either a Gaussian or a super-Gaussian propagation error.

## Gaussian Speech Prior

Assuming the *Rayleigh* distribution for the propagation error amplitudes (Gaussian speech prior), i. e., using parameters  $\eta = 2$ ,  $\beta = 1/\sigma_E^2$ , and  $\nu = 1$  (cf. Table 2.1), the generalized GLR (F.4) can be rewritten as

$$\Lambda_0^\ell \Big|_{\text{R}} = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{2}{\sigma_E^2} \cdot \int_0^\infty |\bar{E}| \cdot \exp\left(-|\bar{E}|^2 \left[\frac{1}{\sigma_E^2} + \frac{1}{\sigma_D^2}\right]\right) \cdot I_0\left(\frac{2|M||\bar{E}|}{\sigma_D^2}\right) d|\bar{E}| \quad (\text{F.5})$$

with R denoting the Rayleigh assumption for the propagation error amplitudes. Utilizing [Gradshteyn and Ryzhik, 1965, Eq. (6.631.1)], (F.5) turns out to be (7.11). The corresponding *a posteriori* SPP estimator can be obtained by (7.6).

## Super-Gaussian Speech Priors

Assuming the *chi* distribution for the propagation error amplitudes (super-Gaussian speech prior), i. e., using parameters  $\eta = 2$ ,  $\beta = \nu_\chi/\sigma_E^2$ , and  $\nu = \nu_\chi$  (cf. Table 2.1), the generalized GLR (F.4) turns out to be

$$\Lambda_0^\ell \Big|_{\chi} = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{2\nu_\chi^{\nu_\chi}}{\Gamma(\nu_\chi) \cdot \sigma_E^{2\nu_\chi}} \cdot \int_0^\infty |\bar{E}|^{2\nu_\chi-1} \cdot e^{-\nu_\chi \frac{|E|^2}{\sigma_E^2}} \cdot e^{-\frac{|E|^2}{\sigma_D^2}} \cdot I_0\left(\frac{2|M||\bar{E}|}{\sigma_D^2}\right) d|\bar{E}| \quad (\text{F.6})$$

where subscript  $\chi$  denotes the chi assumption for the propagation error amplitudes. Using [Gradshteyn and Ryzhik, 1965, Eq. (6.631.1)] to solve the integral w. r. t.  $|\bar{E}|$ , (F.6) can be redefined as (cf. (2.72))

$$\Lambda_0^\ell \Big|_\chi = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \left( \frac{\nu_\chi}{\nu_\chi + \zeta} \right)^{\nu_\chi} \cdot {}_1F_1 \left( \nu_\chi; 1; \frac{\zeta\zeta}{\nu_\chi + \zeta} \right). \quad (\text{F.7})$$

The corresponding generalized *a posteriori* SPP estimator can be obtained by (7.6)

Assuming the *gamma* distribution for the propagation error amplitudes (super-Gaussian speech prior), i. e., employing parameters  $\eta = 1$ ,  $\nu = \nu_\Gamma$ , and  $\beta = \sqrt{\nu_\Gamma(\nu_\Gamma + 1)}/\sigma_{\bar{E}}$  for the generalized estimator (E.7) (cf. Table 2.1), the generalized GLR (F.4) yields (cf. (4.4))

$$\Lambda_0^\ell \Big|_\Gamma = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{[\nu_\Gamma(\nu_\Gamma + 1)]^{\frac{\nu_\Gamma}{2}}}{\Gamma(\nu_\Gamma) \cdot \sigma_{\bar{E}}^{\nu_\Gamma}} \cdot \int_0^\infty |\bar{E}|^{\nu_\Gamma-1} \cdot e^{-\sqrt{\nu_\Gamma(\nu_\Gamma+1)}\frac{|\bar{E}|}{\sigma_{\bar{E}}}} \cdot e^{-\frac{|\bar{E}|^2}{\sigma_{\bar{E}}^2}} \cdot I_0 \left( \frac{2|M| \cdot |\bar{E}|}{\sigma_D^2} \right) d|\bar{E}| \quad (\text{F.8})$$

with subscript  $\Gamma$  denoting the chi assumption for the propagation error amplitudes. Similar to (4.4), the integral in (F.8) cannot be obtained in closed form. Using quadrature, however, the result can be calculated numerically. For this, it is meaningful to reformulate (F.8) using variable substitution  $|\bar{E}| = k \cdot |M|$  and  $d|\bar{E}| = k \cdot d|M|$  leading to (cf. (4.5))

$$\Lambda_0^\ell \Big|_\Gamma = \frac{P(H_1|\mathbf{Y}_0^{\ell-1})}{P(H_0|\mathbf{Y}_0^{\ell-1})} \cdot \frac{[\nu_\Gamma(\nu_\Gamma + 1)]^{\frac{\nu_\Gamma}{2}}}{\Gamma(\nu_\Gamma)} \cdot \left( \frac{\zeta}{\zeta} \right)^{\frac{\nu_\Gamma}{2}} \cdot \int_0^\infty k^{\nu_\Gamma-1} \cdot e^{-\sqrt{\nu_\Gamma(\nu_\Gamma+1)}\sqrt{\zeta}k} \cdot e^{-\zeta k^2} \cdot I_0(2\zeta k) dk. \quad (\text{F.9})$$

This reformulation allows for obtaining the GLR as a function of the *a priori* SNR  $\zeta$  and the *a posteriori* SNR  $\varsigma$ . The GLR (F.9) based on a super-Gaussian propagation error assumption with gamma-distributed amplitudes can be computed by, e. g., the Gauss-Kronrod quadrature [Brass and Petras, 2011]. Finally, the corresponding generalized *a posteriori* SPP estimator can be obtained by (7.6).



# Appendix G

## Derivations for Recursive *A Priori* SPP Estimation

The aim of this appendix is to derive the likelihoods  $p(\varsigma_\ell(k)|H_1)$  and  $p(\varsigma_\ell(k)|H_0)$  for the likelihood ratio test (7.18). For ease of readability, we will omit the frequency bin index  $k$  in the following.

Starting with the likelihood of speech presence, we have to derive the PDF  $p(\varsigma_\ell|H_1)$  with  $\varsigma_\ell = |Y_\ell - \widehat{S}_\ell^+|^2 / \sigma_{D,\ell}^2$ . Please note that the argument within the absolute value operator can be rearranged using the signal and channel model from Figure 6.1 as follows  $Y_\ell - \widehat{S}_\ell^+ = (S_\ell + D_\ell) - \widehat{S}_\ell^+ = D_\ell + (S_\ell - \widehat{S}_\ell^+) = D_\ell + \bar{E}_\ell$  with  $D_\ell$  and  $\bar{E}_\ell$  being the acoustic noise and the propagation error, respectively. Since  $D_\ell$  and  $\bar{E}_\ell$  are statistically independent, we can write

$$\begin{aligned} p_{M_\ell|H_1}(M_\ell = Y_\ell - \widehat{S}_\ell^+ | H_1) &= p_{M_\ell|H_1}(M_\ell = D_\ell + \bar{E}_\ell | H_1) \\ &= \int_{\mathbb{C}} p_{D_\ell}(Y_\ell - S_\ell) \cdot p_{E_\ell}(S_\ell - \widehat{S}_\ell^+) dS_\ell. \end{aligned} \quad (\text{G.1})$$

Assuming Gaussian noise (2.27) and Gaussian propagation errors (6.14) as well as employing polar integration using [Gradshteyn and Ryzhik, 1965, Eqs. (6.631.1), (8.406.3), and (8.431.5)], the PDF (G.1) yields

$$p_{M_\ell|H_1}(M_\ell = Y_\ell - \widehat{S}_\ell^+ | H_1) = \frac{1}{\pi(\sigma_{D,\ell}^2 + \sigma_{\bar{E},\ell}^2)} \cdot e^{-\frac{|Y_\ell - \widehat{S}_\ell^+|^2}{\sigma_{D,\ell}^2 + \sigma_{\bar{E},\ell}^2}} = \frac{1}{\pi(\sigma_{D,\ell}^2 + \sigma_{\bar{E},\ell}^2)} \cdot e^{-\frac{|Y_\ell - \widehat{S}_\ell^+|^2}{\sigma_{D,\ell}^2}} \cdot \frac{1}{1 + \varsigma_\ell} \quad (\text{G.2})$$

which is still a Gaussian PDF with the variance  $\sigma_{D,\ell}^2 + \sigma_{\bar{E},\ell}^2$ . The quantity  $|M_\ell| = |Y_\ell - \widehat{S}_\ell^+|$  will accordingly follow the Rayleigh distribution. Its PDF can be obtained by polar integration

of (G.2) using  $M_\ell = |M_\ell| \cdot e^{j\hat{\rho}}$  as (cf. (D.3))

$$\begin{aligned} p_{|M_\ell|H_1} \left( |M_\ell| = |Y_\ell - \hat{S}_\ell^+ | H_1 \right) &= \frac{\partial}{\partial |M_\ell|} \int_0^{|M_\ell|} \int_0^{2\pi} p_{M_\ell H_1} \left( |\tilde{M}_\ell| \cdot e^{j\hat{\rho}} | H_1 \right) \cdot |\tilde{M}_\ell| \, d\hat{\rho} \, d|\tilde{M}_\ell| \\ &= \frac{2|Y_\ell - \hat{S}_\ell^+|}{\sigma_{D,\ell}^2 + \sigma_{E,\ell}^2} \cdot e^{-\frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2} \cdot \frac{1}{1+\zeta_\ell}}. \end{aligned} \quad (\text{G.3})$$

Then,  $|Y_\ell - \hat{S}_\ell^+|^2$  will follow the exponential distribution and its PDF is calculated according to [Papoulis and Pillai, 2002, Example 5-2] as

$$\begin{aligned} p_{|M_\ell|^2|H_1} \left( |M_\ell|^2 = |Y_\ell - \hat{S}_\ell^+|^2 | H_1 \right) &= \frac{1}{2|Y_\ell - \hat{S}_\ell^+|} \cdot p_{M_\ell H_1} \left( |Y_\ell - \hat{S}_\ell^+ | H_1 \right) \\ &= \frac{1}{\sigma_{D,\ell}^2 + \sigma_{E,\ell}^2} \cdot e^{-\frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2} \cdot \frac{1}{1+\zeta_\ell}}. \end{aligned} \quad (\text{G.4})$$

Finally,  $\zeta_\ell = |Y_\ell - \hat{S}_\ell^+|^2 / \sigma_D^2$  remains also exponential distributed and its PDF can be obtained according to [Papoulis and Pillai, 2002, Eq. (5-18)] as

$$\begin{aligned} p_{\zeta_\ell|H_1} \left( \zeta_\ell = \frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2} | H_1 \right) &= \sigma_D^2 \cdot p_{|M_\ell|^2|H_1} \left( \sigma_{D,\ell}^2 \cdot \frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2} | H_1 \right) \\ &= \frac{1}{1 + \zeta_\ell} \cdot e^{-\frac{\zeta_\ell}{1+\zeta_\ell}}. \end{aligned} \quad (\text{G.5})$$

Assuming speech absence, the likelihood  $p(\zeta_\ell|H_0)$  can be derived in a similar way. Starting with (G.1) and assuming that  $p_{\bar{E}_\ell|H_0}(\bar{E}_\ell = S_\ell - \hat{S}_\ell^+ | H_0) = \delta(\bar{E}_\ell = S_\ell - \hat{S}_\ell^+)$  as in (7.8), the likelihood of speech absence turns out to be (7.9). Assuming a bivariate Gaussian distribution for the noise  $D_\ell$ , the likelihood of speech absence yields

$$p_{M_\ell|H_0}(M_\ell = Y_\ell - \hat{S}_\ell^+ | H_0) = \frac{1}{\pi \sigma_{D,\ell}^2} \cdot e^{-\frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2}}. \quad (\text{G.6})$$

The absolute value of  $Y_\ell - \hat{S}_\ell^+$  will then follow the Rayleigh distribution with the PDF

$$\begin{aligned} p_{|M_\ell|H_0} \left( |M_\ell| = |Y_\ell - \hat{S}_\ell^+ | H_0 \right) &= \frac{\partial}{\partial |M_\ell|} \int_0^{|M_\ell|} \int_0^{2\pi} p_{M_\ell|H_0} \left( |\tilde{M}_\ell| \cdot e^{j\hat{\rho}} | H_0 \right) \cdot |\tilde{M}_\ell| \, d\hat{\rho} \, d|\tilde{M}_\ell| \\ &= \frac{2|Y_\ell - \hat{S}_\ell^+|}{\sigma_{D,\ell}^2} \cdot e^{-\frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2}}. \end{aligned} \quad (\text{G.7})$$

Then, the square of Rayleigh-distributed  $|Y_\ell - \hat{S}_\ell^+|$  values will follow the exponential distribution, i. e., the PDF of  $|Y_\ell - \hat{S}_\ell^+|^2$  in speech absence turns out to be

$$\begin{aligned} p_{|M_\ell|^2|H_0} \left( |M_\ell|^2 = |Y_\ell - \hat{S}_\ell^+|^2 | H_0 \right) &= \frac{1}{2|Y_\ell - \hat{S}_\ell^+|} \cdot p_{M_\ell|H_0} \left( |Y_\ell - \hat{S}_\ell^+ | H_0 \right) \\ &= \frac{1}{\sigma_{D,\ell}^2} \cdot e^{-\frac{|Y_\ell - \hat{S}_\ell^+|^2}{\sigma_{D,\ell}^2}}. \end{aligned} \quad (\text{G.8})$$

Finally,  $\varsigma_\ell = |Y_\ell - \widehat{S}_\ell^+|^2 / \sigma_{D,\ell}^2$  will also follow the exponential distribution and the likelihood of speech absence can be written as

$$\begin{aligned} p_{\varsigma_\ell|H_0} \left( \varsigma_\ell = \frac{|Y_\ell - \widehat{S}_\ell^+|^2}{\sigma_{D,\ell}^2} \middle| H_0 \right) &= \sigma_D^2 \cdot p_{|M_\ell|^2|H_0} \left( \sigma_{D,\ell}^2 \cdot \frac{|Y_\ell - \widehat{S}_\ell^+|^2}{\sigma_{D,\ell}^2} \middle| H_0 \right) \\ &= e^{-\varsigma_\ell}. \end{aligned} \quad (\text{G.9})$$

Accordingly, the ratio of the likelihood of speech presence and the likelihood of speech absence yields

$$\frac{p(\varsigma_\ell|H_1)}{p(\varsigma_\ell|H_0)} = \frac{1}{1 + \zeta_\ell} \cdot e^{\frac{\varsigma_\ell \zeta_\ell}{1 + \zeta_\ell}}. \quad (\text{G.10})$$



# List of Symbols

${}_1F_1(\cdot)$	confluent hypergeometric function
$A_\ell(k)$	speech spectral amplitude
$\alpha_\ell(k)$	speech spectral phase
$\mathbf{A}$	prediction coefficients
$\beta$	parameter of the generalized gamma PDF
$\beta_{\text{DD}}$	smoothing factor of the decision-directed SNR estimator
$\mathbb{C}$	set of all complex numbers
$D_\ell(k)$	noise STFT coefficient
$d(n)$	noise signal
$E\{\cdot\}$	statistical expectation operator
$\bar{E}$	propagation ( <i>a priori</i> estimation) error
$\eta$	parameter of the generalized gamma PDF
$\exp(\cdot)$	exponential function
$e^{(\cdot)}$	exponential function
$f(\cdot)$	arbitrary function
$\Gamma(\cdot)$	gamma function
$\gamma_\ell(k)$	<i>a posteriori</i> SNR
$g(\cdot)$	function to model estimation domains, such as the STS, the STSA, or the LSA domain
$H_0$	hypothesis of speech absence

---

$H_1$	hypothesis of speech presence
$K_\ell(k)$	Kalman gain
$k$	frequency bin index
$\Lambda$	generalized likelihood ratio (GLR)
$\ell$	frame index
$\ln(\cdot)$	natural logarithm
$\nu$	parameter of the generalized gamma PDF
$n$	discrete time index
$P(H_0   Y)$	<i>a posteriori</i> SAP
$P(H_0)$	<i>a priori</i> SAP
$P(H_1   Y)$	<i>a posteriori</i> SPP
$P(H_1)$	<i>a priori</i> SPP
$p_{Y S}(\cdot)$	likelihood
$p_A(\cdot)$	speech spectral amplitude PDF
$p_{S Y}(\cdot)$	posterior
$p_S(\cdot)$	prior
$p_Y(\cdot)$	evidence
$R_\ell(k)$	noisy speech spectral amplitude
$\mathbb{R}$	set of all real numbers
$S_\ell(k)$	speech STFT coefficient
$\sigma_{D,\ell}^2(k)$	noise PSD
$\sigma_{S,\ell}^2(k)$	speech PSD
$\sigma_{E,\ell}^2(k)$	propagation (or <i>a priori</i> estimation) error PSD
$\varsigma_\ell(k)$	<i>a posteriori</i> SNR as defined in recursive MMSE estimation
$\hat{S}_\ell(k)$	estimated speech STFT coefficient

---

$\widehat{S}_\ell^+(k)$	<i>a priori</i> estimate of the speech STFT coefficient
$s(n)$	clean speech signal
$\vartheta_\ell(k)$	noisy speech spectral phase
$\xi_\ell(k)$	<i>a priori</i> SNR
$Y_\ell(k)$	noisy speech STFT coefficient
$y(n)$	noisy speech signal
$\zeta_\ell(k)$	<i>a priori</i> SNR as defined in recursive MMSE estimation



# List of Abbreviations

AR	autoregressive
BPSK	binary phase-shift keying
DD	decision directed
DFT	discrete Fourier transform
GLR	generalized likelihood ratio
i. i. d.	statistically independent and identically distributed
IDFT	inverse discrete Fourier transform
IMCRA	improved minima controlled recursive averaging
ITU-T	International Telecommunication Union, Telecom. Standardization Sector
LLR	log-likelihood ratio
LMS	least-mean-squares
LSA	log-spectral amplitude
MAP	maximum <i>a posteriori</i>
MCRA	minima controlled recursive averaging
ML	maximum likelihood
MMSE	minimum mean square error
MS	minimum statistics
MSB	most significant bit
NLMS	normalized least-mean-squares
OLA	overlap-add
OM	optimally modified
PCM	pulse-code modulation
PDF	probability density function
PSD	power spectral density
RMS	root mean square
SAP	speech absence probability
SNR	signal-to-noise ratio
SPP	speech presence probability
SPU	speech presence uncertainty
STFT	short-time Fourier transform

STS	short-time spectral
STSA	short-time spectral amplitude
VAD	voice activity detection
w. r. t.	with respect to

# Bibliography

- [Abramowitz and Stegun, 1972] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York, NY, USA: Dover Publications, Inc., 1972.
- [Adrat et al., 2000] M. Adrat, J. Spittka, S. Heinen, and P. Vary, “Error Concealment by Near Optimum MMSE-Estimation of Source Codec Parameters,” in *Proc. of IEEE Workshop on Speech Coding (SCW)*, Delavan, WI, USA, Sep. 2000, pp. 84–86.
- [Andrianakis and White, 2006] I. Andrianakis and P. R. White, “MMSE Speech Spectral Amplitude Estimators With Chi and Gamma Speech Priors,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Toulouse, France, May 2006, pp. 1068–1071.
- [Azirani et al., 1996] A. A. Azirani, R. L. B. Jeanns, and G. Faucon, “Speech Enhancement Using a Wiener Filtering Under Signal Presence Uncertainty,” in *Proc. of European Signal Processing Conference (EUSIPCO)*, Trieste, Italy, Sep. 1996, pp. 1–4.
- [Borgström and Alwan, 2011] B. J. Borgström and A. Alwan, “Log-spectral Amplitude Estimation with Generalized Gamma Distributions for Speech Enhancement,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4756–4759.
- [Boubakir and Berkani, 2010] C. Boubakir and D. Berkani, “Speech Enhancement Using Minimum Mean-Square Error Amplitude Estimators Under Normal and Generalized Gamma Distribution,” *Journal of Computer Science*, vol. 6, no. 7, pp. 700–705, Jul. 2010.
- [Brass and Petras, 2011] H. Brass and K. Petras, *Quadrature Theory: The Theory of Numerical Integration on a Compact Interval*, ser. Mathematical Surveys and Monographs. Providence, RI, USA: American Mathematical Society, 2011, vol. 178.

- [Brehm and Stammer, 1987] H. Brehm and W. Stammer, "Description and Generation of Spherically Invariant Speech-Model Signals," *Elsevier Signal Processing*, vol. 12, no. 2, pp. 119–141, Mar. 1987.
- [Breithaupt and Martin, 2011] C. Breithaupt and R. Martin, "Analysis of the Decision-Directed SNR Estimator for Speech Enhancement With Respect to Low-SNR and Transient Conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 277–289, Feb. 2011.
- [Breithaupt et al., 2008] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4897–4900.
- [Brillinger, 2001] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 2001.
- [Bronshtein et al., 2007] I. Bronshtein, K. Semendyayev, G. Musiol, and H. Mühlig, *Handbook of Mathematics*. Springer, 2007.
- [Burington and May, 1970] R. Burington and D. May, *Handbook of Probability and Statistics with Tables*, 2nd ed. New York, NY, USA: McGraw-Hill, 1970.
- [Cappe, 1994] 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. 345–349, Apr. 1994.
- [Chang, 2007] J.-H. Chang, "Complex Laplacian Probability Density Function for Noisy Speech Enhancement," *IEICE Electronics Express*, vol. 4, no. 8, pp. 245–250, Feb. 2007.
- [Chen and Loizou, 2007] B. Chen and P. C. Loizou, "A Laplacian-Based MMSE Estimator for Speech Enhancement," *Elsevier Speech Communication*, vol. 49, no. 2, pp. 134–143, Feb. 2007.
- [Cohen, 2003] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
- [Cohen and Berdugo, 2001] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Elsevier Signal Processing*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

- [Compernelle, 1989] D. V. Compernelle, “Noise Adaptation in a Hidden Markov Model Speech Recognition System,” *Computer Speech and Language*, vol. 3, no. 2, pp. 151–167, 1989.
- [Dat et al., 2005] T. H. Dat, K. Takeda, and F. Itakura, “Generalized Gamma Modeling of Speech and Its Online Estimation for Speech Enhancement,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Philadelphia, PA, USA, Mar. 2005, pp. 181–184.
- [Ephraim and Malah, 1984] 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, Dec. 1984.
- [Ephraim and Malah, 1985] —, “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, Apr. 1985.
- [Erkelens et al., 2007] 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, Aug. 2007.
- [Erkelens et al., 2008] J. S. 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, Jan. 2008.
- [Esch, 2012] T. Esch, “Model-Based Speech Enhancement Exploiting Temporal and Spectral Dependencies,” Ph.D. dissertation, Aachener Beiträge zu Digitalen Nachrichtensystemen, P. Vary, Ed., vol. 32, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, 2012, Available online: <http://darwin.bth.rwth-aachen.de/opus3/volltexte/2012/4035/pdf/4035.pdf>.
- [Esch and Vary, 2008a] T. Esch and P. Vary, “Speech Enhancement Using a Modified Kalman Filter Based on Complex Linear Prediction and Supergaussian Priors,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, pp. 4877–4880.
- [Esch and Vary, 2008b] —, “Modified Kalman Filter Exploiting Interframe Correlation of Speech and Noise Magnitudes,” in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Seattle, WA, USA, Sep. 2008, pp. 1–4.

- [Esch and Vary, 2011] —, “Model-Based Speech Enhancement Using SNR Dependent MMSE Estimation,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4073–4076.
- [ETSI, 2008] ETSI, “Speech Processing, Transmission and Quality Aspects (STQ), Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation technique and Background Noise Database,” ETSI EG 202 396-1, European Telecommunications Standards Institute (ETSI), 2008.
- [Eykhoff, 1974] P. Eykhoff, *System Identification: Parameter and State Estimation*. Wiley-Interscience, 1974.
- [Fingscheidt and Vary, 1997] T. Fingscheidt and P. Vary, “Robust Speech Decoding: A Universal Approach to Bit Error Concealment,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Munich, Germany, Apr. 1997, pp. 1667–1670.
- [Fingscheidt and Vary, 2001] —, “Softbit Speech Decoding: A New Approach to Error Concealment,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 240–251, Mar. 2001.
- [Fingscheidt et al., 2008] T. Fingscheidt, S. Suhadi, and S. Stan, “Environment-Optimized Speech Enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.
- [Fingscheidt, 1998] T. Fingscheidt, “Softbit-Sprachdecodierung in digitalen Mobilfunksystemen,” Ph.D. dissertation, Aachener Beiträge zu Digitalen Nachrichtensystemen, P. Vary, Ed., vol. 9, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, 1998.
- [Fodor and Fingscheidt, 2012a] B. Fodor and T. Fingscheidt, “MMSE Speech Enhancement Under Speech Presence Uncertainty Assuming (Generalized) Gamma Speech Priors Throughout,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4033–4036.
- [Fodor and Fingscheidt, 2012b] —, “Comparison and Signal-Component-Wise Instrumental Evaluation of MMSE Log-Spectral Amplitude Estimation Under Speech Presence Uncertainty,” in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 43–46.
- [Fodor and Fingscheidt, 2012c] —, “Reference-Free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise,” in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 199–202.

- [Fodor and Fingscheidt, 2012d] —, “MMSE Log-Spectral Amplitude Estimation Under Speech Presence Uncertainty Using Generalized Gamma Speech Priors,” in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 1–4.
- [Fodor and Gerkmann, 2014a] B. Fodor and T. Gerkmann, “A Speech Presence Uncertainty Estimator Based on Fixed Priors and a Heavy-Tailed Speech Model,” accepted at the *European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, Sep. 2014, pp. 1–5.
- [Fodor and Gerkmann, 2014b] —, “A Posteriori Speech Presence Probability Estimation Based on Averaged Observations and a Super-Gaussian Speech Model,” accepted at the *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes-Juan les Pins, France, Sep. 2014, pp. 11–15.
- [Fodor et al., 2015] B. Fodor, F. Pflug, and T. Fingscheidt, “Linking Speech Enhancement and Error Concealment Based on Recursive MMSE Estimation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 13, pp. 1–13, Feb. 2015.
- [Gerkmann and Hendriks, 2012] T. Gerkmann and R. Hendriks, “Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.
- [Gerkmann et al., 2008] T. Gerkmann, C. Breithaupt, and R. Martin, “Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, Jul. 2008.
- [Gerkmann, 2010] T. Gerkmann, “Statistical Analysis of Cepstral Coefficients and Applications in Speech Enhancement,” Ph.D. dissertation, Ruhr-Universität Bochum, Bochum, Germany, 2010.
- [Gerkmann and Krawczyk, 2013] T. Gerkmann and M. Krawczyk, “MMSE-Optimal Spectral Amplitude Estimation Given the STFT-Phase,” *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129–132, Feb. 2013.
- [Gerkmann and Martin, 2010] T. Gerkmann and R. Martin, “Empirical Distributions of DFT-Domain Speech Coefficients Based on Estimated Speech Variances,” in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tel Aviv, Israel, Aug. 2010, pp. 1–4.

- [Görtz, 1998] N. Görtz, “Joint Source Channel Decoding Using Bit-Reliability Information and Source Statistics,” in *Proc. of IEEE International Symposium on Information Theory*, Cambridge, MA, USA, Aug. 1998, p. 9.
- [Gradshteyn and Ryzhik, 1965] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integral, Series, and Products*, 4th ed. Academic Press, 1965.
- [Gustafsson et al., 1996] S. Gustafsson, R. Martin, and P. Vary, “On the Optimization of Speech Enhancement Systems Using Instrumental Measures,” in *Workshop on Quality Assessment in Speech, Audio and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.
- [Gustafsson et al., 2002] 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, Jul. 2002.
- [Hagenauer, 1995] J. Hagenauer, “Source-Controlled Channel Decoding,” *IEEE Transactions on Communications*, vol. 43, no. 9, pp. 2449–2457, 1995.
- [Han et al., 2013] S. Han, F. Pflug, and T. Fingscheidt, “Improved AMR Wideband Error Concealment for Mobile Communications,” in *Proc. of European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.
- [Hänsler and Schmidt, 2005] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, ser. Adaptive and Learning Systems for Signal Processing, Communications and Control Series. Wiley, 2005.
- [Haykin, 2002] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.
- [Hendriks et al., 2009a] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, “On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 16, no. 10, pp. 885–888, Oct. 2009.
- [Hendriks et al., 2009b] R. C. Hendriks, R. Heusdens, and J. Jensen, “Log-Spectral Magnitude MMSE Estimators under Super-Gaussian Densities,” in *Proc. of Annual Conference of the International Speech Communication Association (ISCA INTERSPEECH)*, Brighton, UK, Sep. 2009, pp. 1319–1322.
- [ITU-T G.191, 2010] ITU-T G.191, “Software Tools for Speech and Audio Coding Standardization,” ITU-T Recommendation G.191, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Mar. 2010. [Online]. Available: <http://www.itu.int/rec/T-REC-G/>

- [ITU-T P.56, 1993] ITU-T P.56, "Objective Measurement of Active Speech Level," ITU-T Recommendation P.56, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Mar. 1993. [Online]. Available: <http://www.itu.int/rec/T-REC-P/>
- [ITU-T P.830, 1996] ITU-T P.830, "Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs," ITU-T Recommendation P.830, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 1996. [Online]. Available: <http://www.itu.int/rec/T-REC-P/>
- [Jameel et al., 2009] A. J. Jameel, H. Adnan, Y. Xiaohu, and A. Hussain, "Error Concealment of EVRC Speech Decoder Using Residual Redundancy," in *Proc. of Developments in eSystems Engineering (DeSE)*, Abu Dhabi, United Arab Emirates, Dec. 2009, pp. 84–88.
- [Kalman, 1960] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME-Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.
- [Kay, 1993] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice Hall, 1993, vol. 1.
- [Kay, 1998] —, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ, USA: Prentice Hall, 1998, vol. 2.
- [Kullback, 1959] S. Kullback, *Information Theory and Statistics*. John Wiley and Sons, 1959.
- [Lahouti and Khandani, 2007] F. Lahouti and A. K. Khandani, "Soft Reconstruction of Speech in the Presence of Noise and Packet Loss," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 44–56, Jan. 2007.
- [Lim and Oppenheim, 1979] 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, Dec. 1979.
- [Loizou, 2005] P. C. Loizou, "Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, Sep. 2005.
- [Lotter and Vary, 2005] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, May 2005.

- [Lotter, 2004] T. Lotter, “Single and Multimicrophone Speech Enhancement for Hearing Aids,” Ph.D. dissertation, Aachener Beiträge zu Digitalen Nachrichtensystemen, P. Vary, Ed., vol. 18, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, 2004.
- [Malah et al., 1999] D. Malah, R. Cox, and A. Accardi, “Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Phoenix, AZ, USA, 1999, pp. 789–792.
- [Markel and Gray, 1976] J. Markel and A. Gray, *Linear Prediction of Speech*, ser. Communication and Cybernetics. Springer-Verlag, 1976.
- [Martin, 2001] R. Martin, “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
- [Martin, 2002] —, “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Orlando, FL, USA, May 2002, pp. 253–256.
- [Martin, 2005] —, “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, Sep. 2005.
- [Martin et al., 2008] R. Martin, P. U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*. Wiley, 2008.
- [Martin, 1994] R. Martin, “Spectral Subtraction Based on Minimum Statistics,” in *Proc. of European Signal Processing Conference (EUSIPCO)*, Edinburgh, UK, Sep. 1994, pp. 1182–1185.
- [Mathews and Fink, 2004] J. H. Mathews and K. D. Fink, *Numerical Methods Using MATLAB*, 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2004.
- [McAulay and Malpass, 1980] 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. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
- [Middleton and Esposito, 1968] D. Middleton and R. Esposito, “Simultaneous Optimum Detection and Estimation of Signals in Noise,” *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 434–444, May 1968.

- [Mowlae and Saeidi, 2013] P. Mowlae and R. Saeidi, “On Phase Importance in Parameter Estimation in Single-Channel Speech Enhancement,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 7462–7466.
- [Nelson, 1995] R. Nelson, *Probability, Stochastic Processes, and Queueing Theory*. Springer, 1995.
- [NTT, 1994] NTT, “Multi-Lingual Speech Database for Telephonometry,” NTT Advanced Technology Corporation (NTT-AT), 1994.
- [NTT, 1996] —, “Ambient Noise Database for Telephonometry,” NTT Advanced Technology Corporation (NTT-AT), 1996.
- [Paliwal and Basu, 1987] K. K. Paliwal and A. Basu, “A Speech Enhancement Method Based on Kalman Filtering,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, Apr. 1987, pp. 177–180.
- [Papoulis and Pillai, 2002] A. Papoulis and U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.
- [Pflug and Fingscheidt, 2013a] F. Pflug and T. Fingscheidt, “Delayless Robust DPCM Audio Transmission for Digital Wireless Microphones,” in *Proc. of 134th International Audio Engineering Society (AES) Convention*, Rome, Italy, May 2013, pp. 1–8.
- [Pflug and Fingscheidt, 2013b] —, “Robust Ultra-Low Latency Soft-Decision Decoding of Linear PCM Audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2324–2336, Nov. 2013.
- [Pflug, 2013] F. Pflug, “Funkübertragung von Audiosignalen mit prädiktiver Soft-Decision-Dekodierung,” Ph.D. dissertation, Mitteilungen aus dem Institut für Nachrichtentechnik der Technischen Universität Braunschweig, T. Fingscheidt, Ed., vol. 31, Technische Universität Braunschweig, Braunschweig, Germany, 2013.
- [Porter and Boll, 1984] J. Porter and S. Boll, “Optimal Estimators for Spectral Restoration of Noisy Speech,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, San Diego, CA, USA, Mar. 1984, pp. 53–56.
- [Pourmir and Lahouti, 2008] A. M. Pourmir and F. Lahouti, “Joint Source Channel Speech Decoding Using Long-Term Residual Redundancy,” in *Proc. of 16th International Conference on Software, Telecommunications, and Computer Networks (SoftCOM)*, Split, Croatia, Sep. 2008, pp. 329–333.

- [Proakis and Manolakis, 2007] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing*, 4th ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2007.
- [Puder, 2002] 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, 2002.
- [Rabiner and Schafer, 1978] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.
- [Saha and Shimamura, 2011] A. Saha and T. Shimamura, “Generalized Gamma Distributed Bayesian Estimator Under Speech Presence Probability,” in *Proc. of the 11th WSEAS International Conference on Applied Computer Science (ACS)*, Penang, Malaysia, Oct. 2011, pp. 118–123.
- [Scalart and Filho, 1996] P. Scalart and J. Filho, “Speech Enhancement Based on a Priori Signal to Noise Estimation,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Atlanta, GA, USA, May 1996, pp. 629–632.
- [Schuller et al., 2002] G. D. T. Schuller, B. Yu, D. Huang, and B. Edler, “Perceptual Audio Coding Using Adaptive Pre- and Post-Filters and Lossless Compression,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 379–390, Sep. 2002.
- [So et al., 2010] S. So, K. K. Wójcicki, and K. K. Paliwal, “Single-Channel Speech Enhancement Using Kalman Filtering in the Modulation Domain,” in *Proc. of Annual Conference of the International Speech Communication Association (ISCA INTERSPEECH)*, Makuhari, Chiba, Japan, Sep. 2010, pp. 993–996.
- [Sørensen and Andersen, 2005] K. V. Sørensen and S. V. Andersen, “Speech Enhancement with Natural Sounding Residual Noise Based on Connected Time-Frequency Speech Presence Regions,” *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 2954–2964, 2005.
- [Suhadi and Fingscheidt, 2007] S. Suhadi and T. Fingscheidt, “Speech Enhancement with Improved A Posteriori SNR Computation,” in *Proc. of Annual Conference of the International Speech Communication Association (ISCA INTERSPEECH)*, Antwerp, Belgium, Aug. 2007, pp. 962–965.
- [Van Trees, 1968] H. L. Van Trees, *Detection, Estimation and Modulation Theory. Vol 1. Detection, Estimation and Linear Modulation Theory*. Wiley, 1968.

- [Vary and Martin, 2006] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, 2006.
- [Vaseghi, 2008] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed. John Wiley and Sons, 2008.
- [Wolfe and Godsill, 2000] P. J. Wolfe and S. J. Godsill, “Towards a Perceptually Optimal Spectral Amplitude Estimator for Audio Signal Enhancement,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Istanbul, Turkey, Jun. 2000, pp. 821–824.
- [Wolfe and Godsill, 2003a] —, “A Perceptually Balanced Loss Function for Short-Time Spectral Amplitude Estimation,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, Hong Kong, Apr. 2003, pp. 425–428.
- [Wolfe and Godsill, 2003b] —, “Efficient Alternatives to the Ephraim and Malah Suppression Rule for Audio Signal Enhancement,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, Sep. 2003.
- [Wu and Chen, 1998] 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, Aug. 1998.
- [Yu and Fingscheidt, 2011] H. Yu and T. Fingscheidt, “A Figure of Merit for Instrumental Optimization of Noise Reduction Algorithms,” in *Proc. of 5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, Sep. 2011, pp. 1–8.
- [Yu and Fingscheidt, 2012] —, “Instrumental Musical Tones Measurement of Arbitrary Noise Reduction Systems,” in *Proc. of 38th German Annual Conference on Acoustics (DAGA)*, Darmstadt, Germany, Mar. 2012, pp. 255–256.
- [Yu, 2013] H. Yu, “Post-Filter Optimization for Multichannel Automotive Speech Enhancement,” Ph.D. dissertation, Mitteilungen aus dem Institut für Nachrichtentechnik der Technischen Universität Braunschweig, T. Fingscheidt, Ed., vol. 29, Technische Universität Braunschweig, Braunschweig, Germany, 2013, Available online: <http://theses.eurasip.org/theses/509/post-filter-optimization-for-multichannel/>.
- [Zavarehei and Vaseghi, 2005] E. Zavarehei and S. Vaseghi, “Speech Enhancement in Temporal DFT Trajectories Using Kalman Filters,” in *Proc. of Annual Conference of the International Speech Communication Association (ISCA INTERSPEECH)*, Lisbon, Portugal, Sep. 2005, pp. 2077–2080.



# Own Publications

**B. Fodor** and I. Kollár, “ADC Testing with Verification,” in *In Proc. of IEEE Instrumentation and Measurement Technology Conference (IMTC)*, Warsaw, Poland, May 2007, pp. 1–6.

L. Balogh, **B. Fodor**, A. Sárhegyi, and I. Kollár, “Maximum Likelihood Estimation of ADC Parameters from Sine Wave Test Data,” in *In Proc. of 15th IMEKO TC4 Symposium on Novelities in Electrical Measurements and Instrumentation: 12th Workshop on ADC Modeling and Testing*, Iasi, Romania, Sep. 2007, pp. 1–6.

**B. Fodor** and I. Kollár, “ADC Testing With Verification,” *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 12, pp. 2762–2768, Dec. 2008.

**B. Fodor**, D. Scheler, S. Suhadi, and T. Fingscheidt, “Talk-And-Push (TAP) - Towards More Natural Speech Dialog Initiation,” in *In Proc. of AES 36th International Conference*, Dearborn, MI, USA, Jun. 2009, pp. 1–8.

**B. Fodor**, D. Scheler, and T. Fingscheidt, “A Novel Way to Start Speech Dialogs in Cars by Talk-and-Push (TAP),” in *In Proc. of 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety*, Dallas, TX, USA, Jun. 2009, pp. 1–6.

**B. Fodor** and T. Fingscheidt, “Speech Enhancement Using a Joint MAP Estimator with Gaussian Mixture Model for (Non-)Stationary Noise,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4768–4771.

**B. Fodor** and T. Fingscheidt, “MMSE Speech Spectral Amplitude Estimation Assuming Non-Gaussian Noise,” in *Proc. of European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 2011, pp. 2314–2318.

**B. Fodor**, D. Scheler, and T. Fingscheidt, “A Novel Way to Start Speech Dialogs in Cars by Talk-and-Push (TAP),” in *Digital Signal Processing for In-Vehicle Systems and Safety*, J. Hansen, P. Boyraz, K. Takeda, and H. Abut, Eds. Springer, 2012, ch. 7, pp. 123–131.

- B. Fodor** and T. Fingscheidt, “MMSE Speech Enhancement Under Speech Presence Uncertainty Assuming (Generalized) Gamma Speech Priors Throughout,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4033–4036.
- B. Fodor** and T. Fingscheidt, “Reference-free SNR Measurement for Speech in Car Noise,” in *In Proc. of 38th German Annual Conference on Acoustics (DAGA)*, Darmstadt, Germany, Mar. 2012, pp. 259–260.
- B. Fodor** and T. Fingscheidt, “MMSE Log-Spectral Amplitude Estimation Under Speech Presence Uncertainty Using Generalized Gamma Speech Priors,” in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 1–4.
- B. Fodor** and T. Fingscheidt, “Reference-free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise,” in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 199–202.
- B. Fodor** and T. Fingscheidt, “Comparison and Signal-Component-Wise Instrumental Evaluation of MMSE Log-Spectral Amplitude Estimation Under Speech Presence Uncertainty,” in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 43–46.
- B. Fodor** and T. Gerkmann, “A Speech Presence Uncertainty Estimator Based on Fixed Priors and a Heavy-Tailed Speech Model,” accepted at the *European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, Sep. 2014, pp. 1–5.
- B. Fodor** and T. Gerkmann, “A Posteriori Speech Presence Probability Estimation Based on Averaged Observations and a Super-Gaussian Speech Model,” accepted at the *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes-Juan les Pins, France, Sep. 2014, pp. 11–15.
- B. Fodor**, F. Pflug, and T. Fingscheidt, “Linking Speech Enhancement and Error Concealment Based on Recursive MMSE Estimation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 13, pp. 1–13, Feb. 2015.