Graz University of Technology

Dipl.-Ing. Johannes Karl-Wilhelm Stahl, BSc.

Contributions to Single-Channel Speech Enhancement with a Focus on the Spectral Phase

Doctoral Thesis

in partial fulfillment of the requirements for the degree of

Doctor of Technical Sciences

submitted to

Graz University of Technology, Austria
Doctoral School of Information and Communications Engineering

Supervisor and First Assessor/Examiner:
Priv.-Doz. Pejman Mowlaee, PhD
Widex A/S
Nymøllevej 6, 3540 Lynge, Denmark

Supervisor and Second Assessor/Examiner:
Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

Third Assessor/Examiner:
Prof. Dr.-Ing. Tim Fingscheidt
Institute for Communications Technology
Technische Universität Braunschweig, Germany

Graz, January 2019

This work was supported by the Austrian Science Fund (FWF): P28070-N33.

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral dissertation.

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe. Die vorliegende Fassung entspricht der im TUGRAZonline eingereichten elektronischen Version.

Date

Signature

Abstract

Single-channel speech enhancement refers to the reduction of noise signal components in a single-channel signal composed of both speech and noise. Spectral speech enhancement methods are among the most popular approaches to solving this problem. Since the short-time spectral amplitude has been identified as a highly perceptually relevant quantity, most conventional approaches rely on processing the amplitude spectrum only, ignoring any information that may be contained in the spectral phase. As a consequence, the noisy short-time spectral phase is neither enhanced for the purpose of signal reconstruction nor is it used for refining short-time spectral amplitude estimates.

This thesis investigates the use of the spectral phase and its structure in algorithms for single-channel speech enhancement. This includes the analysis of the spectral phase in the context of theoretically optimal speech estimators. The resulting knowledge is exploited in formulating single-channel speech enhancement algorithms. On the one hand, the developed algorithms process the noisy spectral magnitude using spectral phase information and also modify the noisy spectral phase itself. On the other hand, the findings about the spectral phase also result in the conclusion that in certain cases, phase-aware processing should be deliberately circumvented.

Besides objective evaluation of the algorithms presented in this thesis, two subjective listening tests have been conducted in order to evaluate the perceptual relevance of the proposals. The results show that the proposed algorithms consistently improve the perceived speech quality of noisy speech signals while improving the speech intelligibility compared to their conventional counterparts.

Kurzfassung

Einkanalige Sprachverbesserung meint die Reduktion von Rauschsignalkomponenten in einem einkanaligen Signal, das sich sowohl aus Sprach- als auch Rauschanteilen zusammensetzt. Die Klasse der spektralen Sprachverbesserungsmethoden gehört zu den beliebtesten Ansätzen dieses Problem zu lösen. Da die spektrale Amplitude als perzeptiv höchst relevante Größe eingestuft wird, beruhen die meisten konventionellen Ansätze darauf, nur das Kurzzeitamplitudenspektrum zu verarbeiten, wobei zumeist jegliche in der spektralen Phase enthaltene Information ignoriert wird. Infolgedessen wird das verrauschte Kurzzeitphasenspektrum weder zum Zwecke der Signalrekonstruktion verbessert noch zur Verfeinerung der Kurzzeitamplitudenschätzung verwendet.

Die vorliegende Arbeit untersucht die Rolle der spektralen Phase in Algorithmen zur einkanaligen Sprachverbesserung. Dies beinhaltet ihre Analyse im Zusammenhang mit theoretisch optimalen Sprachschätzern. Das so gewonnene Wissen wird für die Formulierung von Sprachverbesserungsalgorithmen genutzt. Diese Vorgehensweise resultiert in Schätzern, die sowohl die verrauschte spektrale Phase als auch die verrauschte spektrale Amplitude modifizieren. Überdies führen die Erkenntnisse über die spektrale Phase auch zu der Schlussfolgerung, dass phasensensitive Verarbeitung in bestimmten Fällen bewusst umgangen werden sollte.

Neben der objektiven Bewertung der in dieser Arbeit vorgestellten Algorithmen wurden zwei subjektive Hörtests durchgeführt, um auch die perzeptive Relevanz der Beiträge zu evaluieren. Die Ergebnisse zeigen, dass die vorgeschlagenen Algorithmen die wahrgenommene Sprachqualität von verrauschten Sprachsignalen konsistent verbessern und gleichzeitig die Sprachverständlichkeit im Vergleich zu ihren konventionellen Gegenstücken verbessern.

Acknowledgments

First of all, I would like to thank my doctoral advisors Pejman Mowlaee and Gernot Kubin. Pejman, thank you for being such a dedicated and enthusiastic supervisor. I highly appreciate your guidance throughout the past years, even after you had (physically) left Graz and the SPSC. Thank you Gernot for your steady support and making meetings possible, even in stressful times. Your ability to immediately grasp and develop ideas further is not only astonishing but also made me reconsider my work, always from new and exciting perspectives.

I thank Prof. Tim Fingscheidt for agreeing to assess this thesis and coming to Graz to be a part of the examination board.

I want to express my gratitude to all members of the SPSC - I really enjoyed my time at the institute. Particularly, I want to thank Anna, Christian K., Elmar, Erik, Jamilla, Stefan, and Wolfgang for the great discussions we had, for your advice whenever I needed some, and last but not least for the fun we had throughout the last years. Those with whom I have worked most closely are Florian, Josef, and Sean - thanks for your invaluable help! Thanks to my office mates Daniel, Harald, Karl, and Katharina for creating such an inspiring and pleasant atmosphere. Finally, thanks to all participants of the listening tests I conducted over the last years.

Special thanks go to Flo, Hannes, Johannes, and Patrick for being exceptional roommates, fellow students, bandmates, and, most importantly, friends. In this context, there is no way around thanking The Snakes - Andi, Kurti, Nico, Rainer - without the evenings in our rehearsal room, the last years would have been far less enjoyable. Also thank you Maria, you are an endless source of inspiration.

I want to thank my family, in particular, Bettina, Michael, Ute, and Lilli - thanks for your support and being such an amazing bunch of crazy people!

My deepest gratitude goes to Lissi, thank you so much for being there for me! This included listening to my probably seldomly interesting stories (not only) about this thesis. Writing this thesis made me review the past years - I'm glad I spent them with you.

Contents

List of Figures

List of Tables

List of Acronyms

ACS	autocorrelation sequence
AMS	analysis, modification, synthesis
AR	autoregressive
ASR	automatic speech recognition
CLT	central limit theorem
COKA	complex-valued Kalman filter
CUP	complex spectral speech coefficients given uncertain phase information
DCT	discrete cosine transform
DD	decision-directed
DFT	discrete Fourier transform
DTFT	discrete-time Fourier transform
ESR	error-to-signal ratio
FFT	fast Fourier transform
GD	group delay
GL	Griffin-Lim
HF	hop factor
IF	instantaneous frequency
KLT	Karhunen-Loève transform
LMD-STSA	linear multidimensional short-time spectral amplitude
LS	least squares
LSA	log spectral amplitude
MAP	maximum a posteriori
MD	multidimensional
MD-WF	multidimensional Wiener filter
MD-STSA	multidimensional short-time spectral amplitude
ML	maximum likelihood
MMSE	minimum mean-square error
MSE	mean-square error
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
MVDR	minimum-variance distortionless response
NA	noise attenuation
NMSE	normalized mean-square error
PA	pitch-adaptive
PACO	pitch-adaptive complex-valued
PADDi	pitch-adaptive decision-directed
PASTFT	pitch-adaptive short-time Fourier transform
pdf	probability density function
PESQ	perceputal evaluation of perceived quality
PS	pitch-synchronous
PSD	power spectral density

PSDE	pitch-synchronous detection-estimation
RV	random variable
SD	stochastic-deterministic
SEGAN	speech enhancement generative adversarial network
SNR	signal-to-noise ratio
SPU	signal presence uncertainty
SSDR	speech to speech distortion ratio
STFT	short-time Fourier transform
STFTPI	short-time Fourier transform phase improvement
STOI	short-time objective intelligibility
STSA	short-time spectral amplitude
STSP	short-time spectral phase
SUC	symmetric uncertainty coefficient
VAD	voice activity detection
WLKR	weighted log kurtosis ratio

1 Introduction

Speech communication technology has become an indispensable part of our lives. It is hard to imagine a world without mobile telephony, smart devices, or medical devices such as hearing aids, all of them equipped with dedicated speech and/or language processing units. As we rely more and more on the functionality of these devices, we expect speech communication technology to work robustly and with high performance. By mitigating the impact of acoustic noise, reliable single-channel speech enhancement algorithms play a key role in ensuring the functionality of the respective application.

This thesis presents work that investigates the role that the spectral phase can play in a single-channel speech enhancement framework, together with practical methods that make use of the resultant insights.

The rest of this chapter is structured as follows. Section 1.1 explains the motivation behind the work that forms this thesis together with the formulation of the underlying research question and Section 1.2 presents a selection of PhD theses that are relevant in the context of the present work. Finally, Section 1.3 introduces the publications that resulted from the work included in this thesis and Section 1.4 further explains their relation to the remaining chapters of this thesis.

1.1 Motivation and Research Question

Single-channel speech enhancement algorithms aim to improve the performance of speech communication systems in noisy environments. This includes extracting a high quality speech signal from a recorded signal that contains not only the desired speech signal but also acoustic noise that deteriorates certain perceptual aspects of the original (clean) reference speech signal. It depends on the specific application what we expect the speech enhancement algorithm to improve. As an example, in a phone conversation we are interested in high intelligibility while also the perceived quality should be high. In automatic speech recognition (ASR) applications, on the other hand, the only objective of the speech enhancement algorithm is to increase the word accuracy rate of the ASR system, generally without any constraints on the perceived quality of the speech enhancement system's output signal.

The literature on single-channel speech enhancement dates back for decades. Among the pioneering works are the studies in [12–14]. These works and many of their most prominent successors, e.g. [15–18], formulate the problem of estimating speech from the noisy mixture signal in the short-time Fourier transform (STFT) domain. In general, the discrete Fourier transform (DFT) yields complex-valued expansion coefficients, rendering the problem of retrieving the clean speech signal from the noisy mixture a problem of estimating real and imaginary parts, or equivalently magnitude and phase, of the masked speech coefficients.

The spectral magnitude is typically considered to be the most perceptually relevant quantity [15]. While this has been justified by the human ears' insensitivity to relative phase changes in the sinusoidal components of a signal [19], in [20] it was shown that rapid fluctuations in the relative phases in sinusoidal speech signal components significantly degrade perceived speech quality. The study of Wang and Lim in [19] from 1982, boldly entitled "The unimportance of phase in speech enhancement", has been a key reference for works that only consider estimation of the speech spectral magnitude. More recently however, in 2011, Paliwal et al. in [21] reconsidered the role of the spectral phase in the context of speech enhancement, finding it to be perceptually important. Following this study, a plethora of phase-sensitive approaches to speech enhancement have been proposed, notable works include [22–31].[1]

With the advent of deep learning, it may arise the impression that classical statistical model-based speech enhancement approaches will become more and more outdated. However, the classical approaches circumvent lengthy training phases, are known to generalize well [34], and in principle machine learning-based approaches may be fused with the a priori knowledge about the speech signal characteristics and the perceptual constraints that typically form the basis of classical methods. Perhaps most relevant in the context of this thesis is the fact that the majority of recent neural network-based approaches to speech enhancement, e.g. [35, 36], also only estimate the magnitude of the STFT from the noisy magnitude, i.e. do not feed any phase information into the network.[2]

As already mentioned, for many speech enhancement approaches that solely rely on estimating the spectral magnitude in order to enhance the noisy speech signal, Wang and Lim's study in [19] has been a key reference for arguing why such a strategy is promising. However, it is the author's belief that the very same study actually motivates further research into the role of the spectral phase in speech enhancement as Wang and Lim conclude their study by stating that

> *... an effort to more accurately estimate the phase from the noisy speech is unwarranted in the context of speech enhancement if the estimate is used to*

[1] Detailed overviews on these recent advances can be found in [32, 33] and more in depth discussions on the contributions relevant to this thesis are given in Chapter 2 and Chapter 3.

[2] End-to-end strategies that process the time domain data directly, such as the recently proposed speech enhancement generative adversarial network (SEGAN) [37], obviously use the full information contained in the signal.

> *reconstruct a signal by combining it with an independently estimated magnitude or to reconstruct the signal using the phase-only signal reconstruction algorithm [5]* [[38] in this thesis, author's note]*. However, if a significantly different approach is used to exploit the phase information such as using the phase estimate to further improve the magnitude estimate, then a more accurate estimation of phase may be important.* [19]

Starting from this vantage point, this thesis aims to investigate the hypothesis that phase information can be exploited to further improve magnitude estimates. Further, we show by means of analytical results as well as by proposing dedicated algorithms how knowledge about the speech spectral phase, or its structure along time and frequency, may be incorporated into estimating the complex-valued short-time spectrum of the clean speech signal. The research question(s) encompassing this thesis hence may be formulated as follows:

What role does the spectral phase play in statistical estimators of the complex-valued spectrum? What tools are needed to find practical solutions to estimating the clean speech signal's complex-valued spectrum? When can we expect performance gains from complex-valued solutions compared to their magnitude-based counterparts?

1.2 Related Doctoral Theses

Single-channel speech enhancement is a well-studied subject and the sheer amount of literature that is constantly published makes it a challenging task to keep track of those publications that are most relevant to one's own work. This thesis was influenced by a vast amount of publications among which the five PhD theses that are shortly summarized in the following are probably the most relevant. The following overview is intended to introduce the basic concepts underlying each of the theses. The conclusion in Chapter 7 then addresses how the contributions in this thesis relate to them.

The work of **Martin Krawczyk-Becker**, who submitted his PhD thesis, "Phase-Aware Single-Channel Speech Enhancement" [39], in 2016 at the Carl von Ossietzky Universität, Oldenburg, is highly relevant to this thesis, particularly to Chapter 4. The main contributions in his work are the development of a novel phase reconstruction method together with clean speech estimators that incorporate the resultant phase information by modeling the spectral phase as a von Mises distributed random variable. Besides the evaluation by means of instrumental metrics that confirms the effectiveness of such approaches, the author of [39] also conducted listening experiments that further validate the hypothesis that phase processing may help to improve speech enhancement quality compared to traditional purely magnitude-based approaches.

Mostly connected to Chapter 4, in 2015, **Matthew C. McCallum** submitted his PhD thesis "Single-Channel Statistical Bayesian Short-Time Fourier Transform Speech Enhancement with Deterministic A Priori Information" [40] at the University of Auckland.

As its key contribution, McCallum's thesis investigates the speech enhancement problem under a stochastic-deterministic signal model for both the speech and noise signals. McCallum justifies the stochastic-deterministic signal model choice by the presence of sinusoidal components in the observed signals that are further interpreted as a non-zero mean for the purpose of statistical modeling the observations. A similar signal model also forms the basis of the derivations in Chapter 4 of the present work. McCallum develops speech enhancement methods that take this signal model into account and also gives a thorough introduction to stochastic-deterministic processes and presents empirical evidence for the selected signal model. The proposed algorithms are reported to come with increased speech quality performance in conjunction with relatively low computational complexity.

Thomas Esch submitted his PhD thesis [41] entitled "Model-Based Speech Enhancement Exploiting Temporal and Spectral Dependencies" in 2012 at the Rheinisch-Westfälische Technische Hochschule, Aachen. In his thesis, Esch develops speech enhancement algorithms that explicitly take into account temporal and spectral dependencies of both the speech and noise DFT expansion coefficients. While the temporal correlations are exploited by means of a modified Kalman filter that relies on a complex-valued subband autoregressive (AR) model, the spectral correlations are incorporated by artificial bandwidth extension techniques. Further, a method for improving the power spectral density (PSD) estimation of harmonic noise in the fashion of minimum statistics is presented. The modified Kalman filter and the corresponding derivations are most relevant for Chapter 5 of the present work, where the inherent AR parameter estimation task is reconsidered under the harmonic plus noise model. Esch reports that all his proposed approaches achieve higher noise suppression and less speech distortions compared to state-of-the art approaches.

The PhD thesis of **Eric Plourde** [42], submitted at the McGill University in Montreal in 2009 is entitled "Bayesian short-time spectral amplitude estimators for single-channel speech enhancement" and can be seen as an attempt to generalize spectral amplitude estimators for speech enhancement. As indicated in the title, this thesis does not explicitly deal with incorporating any knowledge about the spectral phase. However, Plourde develops an analytical unification of the cost function structure of Bayesian spectral amplitude estimators and further generalizes them to the multidimensional case, which allows for modeling correlations of the speech and noise expansion coefficients along frequency. The latter contribution is specifically interesting in the context of Chapter 6 of the present work, where we illustrate that in the case of multidimensional estimators, the statistics of the spectral phase play an important role, even for magnitude-only speech enhancement. Besides the development of algorithms, Plourde's thesis, in its analysis of the proposed estimators, also aims to yield a better understanding of Bayesian spectral amplitude estimators and their parameters.

Finally, the dissertation of **Michael M. Goodwin** [43], entitled "Adaptive Signal Models: Theory, Algorithms, and Audio Applications“, submitted at the University of California, Berkeley in 1997 does not consider the speech enhancement problem but explores signal-adaptive representations of audio signals. A special focus is put on sinusoidal signal models, more specifically on the analysis and synthesis of signals represented by such a model. Further, Goodwin investigates nonstationary signal representations, multiresolutional sinusoidal models in conjunction with adaptive time segmentation, and illustrates that both the sinusoidal model and the wavelet transform benefit from pitch-synchronous segmentation if the analyzed signal is pseudo-periodic. The notion of *physically reasonable partials* [43] in the context of time-frequency analysis, introduced in Goodwin's thesis, is particularly relevant for the contributions in Chapter 4 and Chapter 5 of the present work.

1.3 Contributions

The main body of this thesis is primarily based on the following works:

- [1] J. Stahl and P. Mowlaee, "Iterative harmonic speech enhancement," in *Proc. ITG Symposium on Speech Communication*, Oct 2016, pp. 1–5.
- [2] J. Stahl and P. Mowlaee, "A pitch-synchronous simultaneous detection-estimation framework for speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 436–450, Feb 2018.
- [3] J. Stahl and P. Mowlaee, "A simple and effective framework for a priori SNR estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2018, pp. 5644–5648.
- [4] J. Stahl, S. Wood, and P. Mowlaee, "Overcoming covariance matrix phase sensitivity in single-channel speech enhancement with correlated spectral components," in *Proc. ITG Symposium on Speech Communication*, Oct 2018, pp. 286–290.
- [5] J. Stahl and P. Mowlaee, "Exploiting temporal correlation in pitch-adaptive speech enhancement," *submitted to Speech Communication*, 2018.
- [6] J. Stahl, S. Wood, and P. Mowlaee, "Single-channel speech enhancement with correlated spectral components: Limits - potential," *submitted to IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2018.

For the sake of reproducible research, all implementations that contributed to the main results reported in this thesis and the respective publications are publicly available on `https://gitlab.com/johannesstahl`.

Besides the publications that (partially) became part of this thesis, the author also contributed to the following publications:

- [7] J. Stahl, P. Mowlaee, and J. Kulmer, "Phase-processing for voice activity detection: A statistical approach," in *Proc. European Signal Processing Conf.*, Aug 2016, pp. 1202–1206.

- [8] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice.* John Wiley & Sons, Ltd, 2016.
- [9] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, “Phase-aware signal processing for automatic speech recognition,” in *Proc. Interspeech*, 2016, pp. 3374–3378.
- [10] M. Pirolt, J. Stahl, P. Mowlaee, V. I. Vorobiov, S. Y. Barysenka, and A. G. Davydov, “Phase estimation in single-channel speech enhancement using phase invariance constraints,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2017, pp. 5585–5589.
- [11] P. Mowlaee, J. Stahl, and J. Kulmer, “Iterative joint MAP single-channel speech enhancement given non-uniform phase prior,” *Speech Communication*, vol. 86, pp. 85 – 96, 2017.

1.4 Outline

The rest of this thesis is structured as follows.

Chapter 2 shortly reviews important fundamental concepts in STFT-based single-channel speech enhancement and concludes which assumptions, besides perceptually motivated reasoning, justify the noisy spectral phase as the optimal reconstruction phase.

Chapter 3 presents a literature review on phase processing, introduces important concepts based on harmonic signal modeling that are relevant throughout this thesis, and illustrates how knowledge about the (harmonic) phase has been incorporated into single-channel speech enhancement strategies in the literature.

In Chapter 4, based on the publications in [1,2], a simultaneous detection-estimation approach under the harmonic plus noise model for speech is presented. A pitch-synchronous analysis-synthesis stage is used to enable more accurate estimation of harmonic signal components, and the rather general detection-estimation step fuses estimates obtained from different signal models.

Chapter 5, following the work in [5], is focused on modeling the correlations of the speech spectral coefficients along time by using the statistics of the phase progression along time. In order to facilitate the estimation of the signal model parameters, a pitch-adaptive short-time Fourier transform as proposed in [3] is used. In addition to an analysis of the inter-frame correlations in the pitch-adaptive short-time Fourier transform, as a result of the simplified statistics within this domain, a complex-valued Kalman filter, that relies on the statistics of the spectral phase, is proposed.

Chapter 6 includes and extends the work presented in [4, 6], focused on analyzing the spectral phase’s role when correlation w.r.t. frequency is incorporated into the speech coefficient estimation. The phase of the covariance matrix is found to play a key role in such speech enhancement approaches. Since this quantity is hard to estimate, a linear multidimensional spectral amplitude estimator is proposed, circumventing the need to estimate the covariance phase from the noisy speech data.

Finally, Chapter 7 concludes on the work presented in this thesis and explains in what way it generalizes and extends existing work, with a special focus on the five PhD theses presented in Section 1.2. Further, an attempt to answer the research question based on the results of the present work is made and possible future research directions are discussed.

This thesis is supplemented with five appendices that serve two primary purposes. First, lengthy derivations are outsourced to appendices. Second, some appendices present theoretical and practical background on concepts that need detailed explanations but are not directly relevant to the problems discussed in the main body of the thesis.

2

Single-Channel Speech Enhancement in the STFT Domain

This chapter gives an introduction to single-channel speech enhancement and some of the concepts used therein. Since a myriad of problem formulations and proposed solutions exists, we will mainly focus on a subclass of algorithms formulated in the STFT domain. Within this class of algorithms, a variety of signal models, optimization criteria, and implementation variants exists. This chapter therefore focuses on works that are relevant w.r.t. the contributions of the author. This chapter also aims to determine the conditions the statistical model of speech must fulfill such that the spectral phase information can be neglected for the estimation of the speech STFT coefficients.

2.1 Overview

Most speech enhancement algorithms are derived on the basis of the additive signal model

$$y_{\mathrm{t}}(n) = x_{\mathrm{t}}(n) + d_{\mathrm{t}}(n), \tag{2.1}$$

where $n \in \mathbb{Z}$ is the discrete-time index, the subscript t indicates time domain signals, $x_{\mathrm{t}}(n)$ is the (clean) speech signal, $d_{\mathrm{t}}(n)$ the additive noise, and $y_{\mathrm{t}}(n)$ is the noisy, observed signal. In single-channel speech enhancement, we aim to retrieve an estimate $\hat{x}_{\mathrm{t}}(n)$ of the clean speech signal from the observation $y_{\mathrm{t}}(n)$, where the hat symbol, i.e. $\hat{\cdot}$, denotes estimates in this thesis.

According to [44], speech enhancement algorithms can be coarsely divided into 4 main classes

1. Spectral subtraction-based algorithms, e.g. [12, 20],
2. Statistical model-based algorithms, e.g. [13–15], .
3. Subspace decomposition-based algorithms, e.g. [45],
4. Binary mask-based algorithms, e.g. [46].

Given recent advances, the above list appears incomplete, as for example end-to-end approaches such as the SEGAN method [37], based on generative adversarial learning, are not represented by any of the above algorithm classes.

A key problem of single-channel speech enhancement algorithms is that they typically only improve the perceived quality of the noisy signal, while the intelligibility is degraded [44]. An exception are binary mask-based algorithms, which have been demonstrated to be capable of improving speech intelligibility as well [44]. The main focus of this thesis lies on the class of statistical model-based algorithms, known to balance a trade-off between speech quality improvement and intelligibility degradation. In this regard, incorporating spectral phase information has been reported to improve the capabilities of these algorithms [8, 33].

2.1.1 Analysis, Modification, and Synthesis

Throughout this thesis, the processing chain of speech enhancement algorithms will be divided into three stages, namely the analysis (A), modification (M), and the synthesis (S) stage. We will refer to this high-level model as AMS system [47], described in the following.

The **analysis** stage serves two purposes. First, it divides the input signal $y_\mathrm{t}(n)$ into overlapping segments ℓ of length N by multiplying it with a sliding analysis window function $w(n)$ with support on $n \in [0, N-1]$

$$y_\mathrm{t}(n,\ell) = y_\mathrm{t}(n+\ell L)w(n), \tag{2.2}$$

where ℓ denotes the segment (or frame) index, and L is the frame shift. Second, the frames are optionally transformed into a domain other than the time domain, e.g. via the DFT.

The **modification** stage represents the specific noise reduction scheme that is applied either in the time or transform domain. The modification often relies on a signal model in conjunction with an optimization criterion such as the mean-square error (MSE).

The **synthesis** stage typically consists of the inverse of the analysis operations. This means that it transforms the processed segments back to time domain if needed and recovers a signal stream by using overlap-add.

In this thesis, we focus on speech enhancement algorithms that operate in the STFT domain. Such approaches specify the analysis and synthesis stage of the overall algorithm, discussed in the following. The modification stage, i.e. the the core of many speech enhancement algorithms, is subsequently discussed in more detail.

2.1.2 The Short-Time Fourier Transform

The discrete Fourier transform of a signal $x_{\mathrm{t}}(n)$ is defined as [48]

$$X(k) = \frac{1}{N_{\mathrm{DFT}}} \sum_{n=0}^{N_{\mathrm{DFT}}-1} x_{\mathrm{t}}(n) \mathrm{e}^{-j\omega_k n}, \tag{2.3}$$

where

$$\omega_k = \frac{2\pi k}{N_{\mathrm{DFT}}} \tag{2.4}$$

is the normalized angular DFT frequency with discrete-frequency index $k \in [0, N_{\mathrm{DFT}}-1]$ and N_{DFT} is the DFT length. Applying the DFT on every segment obtained from Equation (2.2) yields the STFT

$$Y(k,\ell) = \frac{1}{N_{\mathrm{DFT}}} \sum_{n=0}^{N_{\mathrm{DFT}}-1} y_{\mathrm{t}}(n,\ell) \mathrm{e}^{-j\frac{2\pi k}{N_{\mathrm{DFT}}}n} \tag{2.5}$$

and due to the linearity of the DFT, the STFT representation of the signal model in Equation (2.1) is

$$Y(k,\ell) = X(k,\ell) + D(k,\ell), \tag{2.6}$$

where capital letters indicate the frequency domain representations of the respective time domain signals, indicated by lower-case letters. The STFT is perhaps the most prominent analysis stage in speech enhancement systems. Among the reasons for the STFT's predominance are the following facts.

1. It can be implemented efficiently by means of the fast Fourier transform (FFT).

2. The resulting DFT expansion coefficients can be considered to be approximately uncorrelated, enabling straightforward optimization of distortion criteria such as the MSE in many cases.

3. Its spectral content may be linked to physical properties of the speech production system [47], making observations in the STFT easy to interpret.

The DFT can be interpreted as a sampled version of the discrete-time Fourier transform (DTFT), given by [48]

$$X(\mathrm{e}^{j\omega}) = \sum_{n=-\infty}^{\infty} x_{\mathrm{t}}(n) \mathrm{e}^{-j\omega n}, \tag{2.7}$$

with normalized angular frequency

$$\omega = 2\pi \frac{f}{f_s}, \tag{2.8}$$

where f is the continuous linear frequency in Hz and f_s is the sampling frequency. In this thesis, the DTFT will be of special interest for conceptual, non-numerical considerations, e.g. when we have an explicit mathematical signal model available and want to analyze the underling signal beyond the DFT "picket fence".

For a signal of length N, the DTFT is a polynomial of order N. Hence, any N samples of the DTFT specify the DTFT and the corresponding time domain signal exactly [43]. Choosing $N_{\mathrm{DFT}} \geq N$ ensures the DFT to be a complete representation of the time domain signal. From this completeness property [49], it follows that $y_{\mathrm{t}}(n, \ell)$ and $Y(k, \ell)$ contain the same information as long as $N_{\mathrm{DFT}} \geq N$, hence, $x_{\mathrm{t}}(n, \ell)$ can be estimated from $Y(k, \ell)$ benefiting from the properties of the STFT.

The DFT represents the special case of N_{DFT} uniformly spaced samples of the DTFT, facilitating the numerical computation of the signal's spectrum. Alternatives to the DFT include the discrete cosine transform (DCT) [50], wavelet transforms [51], and subspace approaches such as the Karhunen-Loève transform (KLT) [45]. However, subspace-based methods do not ensure the completeness of the resulting representation, which is why we will stick to DFT-like transformations throughout this thesis. This does not prevent us from applying signal-adaptive transformations, which will be of particular interest in Chapter 4 and Chapter 5.

By modification of $Y(k, \ell)$, described in the following section, we obtain estimates of the clean speech STFT expansion coefficients. Since in many applications, we aim to reconstruct the time domain signal from these samples, we first apply the inverse DFT and subsequently reconstruct the time domain signal by overlap-add [52]. A large part of the speech enhancement literature is focused on the modification stage, i.e. on the task of finding a suitable model for the observed data and formulating estimators that are optimal w.r.t. this model.

2.2 The Modification Stage

This section focuses on modification paradigms that are based on statistical model-based approaches, i.e. point 2 in the list on page 9. In order to derive optimal estimators of the speech STFT coefficients, two key ingredients are needed [53].

- First, a criterion that represents our definition of optimality is indispensable. This optimization criterion can be mathematically motivated, such as the MSE or the maximum a posteriori (MAP) criterion for example. It may also reflect specific properties of the receiver, e.g. peculiarities in human speech perception in the case of human listeners at the receiver end.

- Second, statistical models for the speech and noise data are needed.

A significant amount of research has explored statistical models for speech and noise, e.g. [17, 54–56]. A seminal work w.r.t. statistically modeling the speech signal in the STFT domain is the work of Ephraim and Malah in [15], which includes assumptions and strategies that are still relevant in very recent publications on speech enhancement, e.g. [57]. In order to emphasize the key differences between the various methods proposed in the literature and those discussed in this thesis, we proceed to explain this very generic statistical model in more detail. This will facilitate the presentation of the main differences between the various methods presented subsequently. Further, due to the generality of this model, the insights w.r.t. to the spectral phase might be relevant to other fields of research employing a similar statistical model.

In [15], we also find the following quote that encompasses what follows in this thesis:

> *..., we note that since the true statistical model* [of speech, author's note] *seems to be inaccessible, the validity of the proposed one can be judged a posteriori on the basis of the results obtained here. In addition, the optimality of the estimators derived here is of course connected with the assumed statistical model.* [15]

2.2.1 The Statistical Model

In attempting to find suitable statistical models for the speech and noise signals, we face the problem that neither speech nor noise can be considered to be stationary or ergodic processes, rendering the meaning of long term statistics to be a controversial topic [15]. Though, several distributions based on histograms have been proposed in the literature, it is important to note that such empirically obtained distributions represent relative frequencies of STFT expansion coefficient appearance rather than their probability density function (pdf). This is why often asymptotic properties of the DFT are used to model speech.

Following [15], we can model the discrete-time speech signal as a sequence of statistically independent, zero-mean random variables (RV) $\{x_{\mathrm{t}}(n,\ell)\}_{n=0}^{N-1}$. Hence, in Equation (2.3), for each DFT bin k, we obtain the corresponding coefficient by a summation of independent RVs. The central limit theorem (CLT) states that the distribution of the sum over independent RVs approaches a Gaussian distribution as the number of RVs that is summed over increases [58]. This Gaussian distribution is characterized by a mean and a variance that both correspond to the (weighted) sum of the means and the variances of the RVs that are summed over. As a result, the asymptotic model yields mutually independent, zero-mean, complex-valued, isotropic Gaussian distributions for the speech STFT coefficients.

The above approximation is only valid if (i) N in the analysis stage approaches infinity and further, if (ii) N is much larger than the correlation span of $x_{\mathrm{t}}(n,\ell)$ [59]. Clearly, both prerequisites cannot be met in practice as an infinite frame length is impossible and also the correlation span is not necessarily much smaller than N, as for example in the case of periodic signals [59]. Although this indicates that the CLT might not be a perfect fit, it comes with the benefit that the resulting statistical model is mathematically tractable and it has proven to be a powerful model in terms of speech enhancement. This is why we will build upon the asymptotic model throughout large parts of this thesis.

The Fourier transform, as an orthogonal transform, asymptotically decorrelates spectral components. In the case of jointly Gaussian distributed RVs, uncorrelatedness implies independence. Hence, the real and imaginary parts of the STFT expansion coefficients are modeled by mutually independent, zero-mean Gaussian RVs. Further we assume that the time domain variances split equally on the real and the imaginary part,[3] meaning that the speech spectral variance is given by

$$\sigma_x^2(k,\ell) = \mathbb{E}\big(X_{\mathrm{R}}^2(k,\ell) + X_{\mathrm{I}}^2(k,\ell)\big) = 2\mathbb{E}\big(X_{\mathrm{R}}^2(k,\ell)\big) = 2\mathbb{E}\big(X_{\mathrm{I}}^2(k,\ell)\big) = \mathbb{E}\big(\,|X(k,\ell)|^2\,\big), \tag{2.9}$$

where $\mathbb{E}(\,\cdot\,)$ denotes the expectation operator and the subscripts “R” and “I” indicate real and the imaginary parts, respectively. The same conjectures can be made for the noise signal, hence, analogously for the noise STFT coefficients we have $\sigma_d^2(k,\ell) = \mathbb{E}\big(\,|D(k,\ell)|^2\,\big)$. Relating the variance to the expected value of the sampled periodogram, i.e. $\mathbb{E}\big(\,|\cdot|^2\,\big)$, means that if we are to estimate the spectral variance, we are estimating the power spectrum of the respective signal [48].

It is important to note that mutual independence of the STFT coefficients is a rather strong assumption, considering that the analysis segments are overlapping, the analyzed signals are far from being stationary, the analysis time spans are chosen correspondingly short, and the speech production itself of course prohibits independent excitation of frequency channels, as all spectral components share the same airflow through the vocal tract. Modeling dependencies w.r.t to time and/or frequency has for example been addressed in [60–64] and will also be of interest in this thesis, specifically in Chapter 5 and Chapter 6. However, in the classical setup, as a consequence of modeling the STFT coefficients to be mutually independent, they are also processed independently. Hence, for the sake of readability, we will drop the frame and the frequency indices wherever possible.

[3] Except for $\omega_k \in \{0,\pi\}$, since for for real-valued signals, the imaginary part is zero at these frequencies.

Following the above argumentation, the pdfs of the real and imaginary parts of speech and noise are given by[4]

$$
\begin{aligned}
p\left(X_{\mathrm{R}}\right) &= \frac{1}{\sqrt{\pi\sigma_x^2}} \mathrm{e}^{-\frac{X_{\mathrm{R}}^2}{\sigma_x^2}}, \quad p\left(X_{\mathrm{I}}\right) = \frac{1}{\sqrt{\pi\sigma_x^2}} \mathrm{e}^{-\frac{X_{\mathrm{I}}^2}{\sigma_x^2}}, \\
p\left(D_{\mathrm{R}}\right) &= \frac{1}{\sqrt{\pi\sigma_d^2}} \mathrm{e}^{-\frac{D_{\mathrm{R}}^2}{\sigma_d^2}}, \quad p\left(D_{\mathrm{I}}\right) = \frac{1}{\sqrt{\pi\sigma_d^2}} \mathrm{e}^{-\frac{D_{\mathrm{I}}^2}{\sigma_d^2}}.
\end{aligned}
\tag{2.10}
$$

Hence, using the independence of the real and imaginary parts, the model for the complex-valued speech STFT coefficients in Cartesian coordinates is simply

$$
p\left(X\right) = p\left(X_{\mathrm{R}}, X_{\mathrm{I}}\right) = p\left(X_{\mathrm{R}}\right) p\left(X_{\mathrm{I}}\right) = \frac{1}{\pi\sigma_x^2} \mathrm{e}^{-\frac{X_{\mathrm{R}}^2 + X_{\mathrm{I}}^2}{\sigma_x^2}} = \frac{1}{\pi\sigma_x^2} \mathrm{e}^{-\frac{|X|^2}{\sigma_x^2}}, \tag{2.11}
$$

and accordingly for the noise STFT coefficients

$$
p\left(D\right) = \frac{1}{\pi\sigma_d^2} \mathrm{e}^{-\frac{D_{\mathrm{R}}^2 + D_{\mathrm{I}}^2}{\sigma_d^2}} = \frac{1}{\pi\sigma_d^2} \mathrm{e}^{-\frac{|D|^2}{\sigma_d^2}} \tag{2.12}
$$

Given the noise model, the likelihood of the noisy observation is given by

$$
p\left(Y|X\right) = \frac{1}{\pi\sigma_d^2} \mathrm{e}^{\frac{|Y-X|^2}{\sigma_d^2}} = \frac{1}{\pi\sigma_d^2} \mathrm{e}^{-\frac{Y_{\mathrm{R}}^2 + Y_{\mathrm{I}}^2 + X_{\mathrm{R}}^2 + X_{\mathrm{I}}^2 - 2\left(Y_{\mathrm{R}} X_{\mathrm{R}} + Y_{\mathrm{I}} X_{\mathrm{I}}\right)}{\sigma_d^2}}. \tag{2.13}
$$

We refer to the above distributions as complex normal distributions. We will further use the short notation $q \sim \mathcal{CN}\left(\mu, \sigma^2\right)$ in order to indicate that a quantity q is modeled by a RV that follows a complex normal distribution with mean μ and variance σ^2.

Transforming the above pdfs from Cartesian to polar coordinates yields an explicit model for the spectral magnitude, which is arguably more perceptually relevant than considering the real and imaginary parts separately. Expressing X and Y in polar coordinates, i.e. in terms of amplitude and phase, yields

$$
\begin{aligned}
X &= A\mathrm{e}^{j\phi}, \\
Y &= R\mathrm{e}^{j\vartheta} = A\mathrm{e}^{j\phi} + |D|\, \mathrm{e}^{j\angle D}.
\end{aligned}
\tag{2.14}
$$

Here, A is the speech short-time spectral amplitude (STSA), ϕ is the speech short-time spectral phase (STSP), R is the noisy STSA, and ϑ is the noisy STSP. Hence, it follows that the joint distribution of speech magnitude and phase is [58]

$$
p\left(A, \phi\right) = \frac{A}{\pi\sigma_x^2} \mathrm{e}^{-\frac{A^2}{\sigma_x^2}} \qquad \text{if } A \geq 0,\ 0 \leq \phi < 2\pi \tag{2.15}
$$

[4] For details on the notation used in this context see Appendix A.2.

where by marginalization we obtain the well-known Rayleigh distribution for the magnitude pdf, i.e. [15]

$$p\left(A\right) = \begin{cases} \frac{2A}{\sigma_x^2}\mathrm{e}^{-\frac{A^2}{\sigma_x^2}} & A \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.16}$$

For the phase, marginalization yields the uniform distribution

$$p\left(\phi\right) = \begin{cases} \frac{1}{2\pi} & 0 \leq \phi < 2\pi, \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

It is important to note that we can factorize Equation (2.15) into the above marginals of the magnitude and phase, i.e.

$$p\left(A, \phi\right) = p\left(A\right) p\left(\phi\right) \tag{2.18}$$

indicating that the two RVs are independent. This property results from the circularity and unimodality of the pdfs in Equation (2.11) and Equation (2.12).[5] However, in the more general case, where the mean of the distribution is non-zero and/or dependencies between real and imaginary part exist, the distributions are non-circular and magnitude and phase are not necessarily independent anymore.

2.2.2 MMSE Estimators for the Classical Gaussian Model

In this section, we shortly present estimators of the complex-valued STFT coefficients, the magnitudes, and the phase, derived from the zero-mean Gaussian signal model. The minimum mean-square error (MMSE) optimal estimate of the STFT coefficient in case of the Gaussian model is the posterior, i.e. conditional, mean

$$\hat{X}_{\mathrm{W}} = \mathbb{E}\big(X|Y\big) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_d^2} Y. \tag{2.19}$$

It will be useful later on to express estimators in terms of gain functions that are multiplied with the noisy observation. The specific gain function resulting from the zero-mean Gaussian model is the ubiquitous Wiener filter [66], i.e.

$$G_{\mathrm{W}} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_d^2} = \frac{\xi}{1 + \xi}, \tag{2.20}$$

[5] Circularity in the context of complex-valued RVs means that its moments are invariant w.r.t. a rotation of the underlying RV in the complex plane [65].

where

$$\xi = \frac{\sigma_x^2}{\sigma_d^2} \tag{2.21}$$

is the so-called *a priori* signal-to-noise ratio (SNR). It is important to note that since $G_{\mathrm{W}} \in \mathbb{R}$, the phase of the noisy coefficient is not affected by the gain function, only the magnitude is as a function of the *a priori* SNR.

Motivated by the relative perceptual importance of the spectral amplitude compared to the spectral phase, Ephraim and Malah in [15] proposed to formulate the MMSE criterion in terms of the spectral magnitude directly instead of the complex-valued STFT coefficients. The posterior of the magnitude given the zero-mean Gaussian model from Section 2.2.1 is a Rician distribution [67], i.e.

$$p\left(A|Y\right) = 2\frac{A}{\sigma_{x|y}} \mathrm{e}^{-\frac{A^2+G_{\mathrm{W}}^2|Y|^2}{\sigma_{x|y}}} I_0\left(2\frac{AG_{\mathrm{W}}\left|Y\right|}{\sigma_{x|y}}\right), \tag{2.22}$$

where $I_\nu\left(\cdot\right)$ is the modified Bessel function of ν^{th} order and the posterior variance $\sigma_{x|y}^2$ is given by [68]

$$\sigma_{x|y}^2 = \frac{\sigma_d^2 \sigma_x^2}{\sigma_x^2 + \sigma_d^2}. \tag{2.23}$$

The posterior variance can be interpreted as the variance of the Wiener filter estimate [68]. The expected value of the Rician distribution yields the MMSE-STSA estimator [15]

$$\begin{aligned} \hat{A}_{\mathrm{EM}} &= \Gamma\left(\frac{3}{2}\right) \sqrt{\frac{\sigma_{x|y}^2}{\sigma_d^2 \zeta}}\ {}_1F_1\left(-\frac{1}{2}; 1; -\sqrt{G_{\mathrm{W}}\zeta}\right) R \\ &= \Gamma\left(\frac{3}{2}\right) \frac{\sqrt{G_{\mathrm{W}}\zeta}}{\zeta} \mathrm{e}^{-\frac{G_{\mathrm{W}}\zeta}{2}} \left(\left(1 + G_{\mathrm{W}}\zeta\right) I_0\left(\frac{G_{\mathrm{W}}\zeta}{2}\right) + G_{\mathrm{W}}\zeta I_1\left(\frac{G_{\mathrm{W}}\zeta}{2}\right)\right) R \\ &= G_{\mathrm{EM}} R \end{aligned} \tag{2.24}$$

where $\Gamma\left(\cdot\right)$ is the gamma function, ${}_1F_1\left(\cdot;\cdot;\cdot\right)$ is the confluent hypergeometric function,[6] G_{EM} refers to the Ephraim-Malah weighting rule, and ζ is the *a posteriori* SNR defined as

$$\zeta = \frac{R^2}{\sigma_d^2}. \tag{2.25}$$

We cannot reconstruct the time domain signal from the spectral magnitude only, hence, a spectral phase estimate is also needed. For the posterior of the phase we have [69]

[6] For definitions of special functions appearing in this thesis we refer to Appendix D.

$$p\left(\phi|Y\right) = \mathrm{e}^{-\frac{R^2 G_{\mathrm{W}}}{\sigma_d^2}} \left(\frac{1}{4\pi^2} + \frac{R\sqrt{G_{\mathrm{W}}}}{2\pi^{\frac{3}{2}}\sigma_d} \cos\left(\vartheta - \phi\right) \mathrm{e}^{\frac{R^2 G_{\mathrm{W}}}{\sigma_d^2}\cos(\vartheta-\phi)^2} Q\left(\sqrt{2}\frac{R\sqrt{G_{\mathrm{W}}}}{\sigma_d}\right)\right), \quad (2.26)$$

where $Q\left(\cdot\right)$ is the Q-function. The expected value of the posterior is the MMSE-optimal phase estimate, which is in this case the noisy phase

$$\hat{\phi}_{\mathrm{MMSE}} = \vartheta. \quad (2.27)$$

This means that independent optimization of the amplitude and phase yields

$$\hat{X}_{\mathrm{EM}} = G_{\mathrm{EM}} Y, \quad (2.28)$$

where $G_{\mathrm{EM}} \in \mathbb{R}$. Hence, although the Wiener filter and the MMSE-STSA gain yield different magnitude estimates, the noisy phase is the optimal estimate in both cases. This is the result of both the chosen prior distributions and the fact that only one observation Y is taken into account in the estimator. As neither the likelihood nor the prior distribution promote any directional orientation in the complex plane, the noisy phase is also optimal w.r.t. the MAP and the ML criterion.

It is often said that the noisy phase is the (MMSE/MAP/ML) optimal choice in the literature. However, from the above discussion we note that this statement should always be accompanied by the statement this is the case *only* for a very specific signal model, which is at its core not even designed on the basis of knowledge about the speech production process but only relies on asymptotic properties of DFT coefficients. Clearly, changing the magnitude prior pdfs does not change this result either, as the likelihood and the joint prior of magnitude and phase remain symmetric around the origin in the complex plane in this case. If however we were to consider more than one observation, the likelihood would be non-circular and hence promoting a certain orientation in the complex plane. We will consider such scenarios in Chapters 4-6.

2.2.3 Noise and Speech PSD Estimation

So far, we assumed all parameters of the distributions under consideration to be known. However, they are not known a priori, hence, they have to be estimated from the noisy mixture. It is common practice to decouple the noise PSD estimation from the actual speech estimation task [18]. Therefore, we find dedicated noise PSD estimators in the literature, developed independently from the subsequent algorithmic steps of the speech enhancement framework.

The most important concepts for noise PSD estimation include methods relying on voice-activity detection (VAD), minima tracking [70], estimators incorporating speech presence probability [71], MMSE estimation [18, 72], or methods relying on subspace decomposition of the DFT representation [73].

Once a noise PSD estimate is obtained, there is still the speech PSD, or spectral variance [48], which is not known and has to be estimated. In practical implementations, it has proven to be advantageous to estimate the *a priori* SNR than directly estimating the speech PSDs. This is mainly since the *a priori* SNR is invariant to signal scaling and its range can be expected to lie within the interval $[-20, 30]$ dB for most noise reduction applications [59].

Given the *a posteriori* SNR estimate, the ML estimate of the *a priori* SNR is given by [15]

$$\hat{\xi}_{\mathrm{ML}}(k,\ell) = \frac{R^2(k,\ell)}{\sigma_d^2(k,\ell)} - 1 = \hat{\zeta}(k,\ell) - 1. \tag{2.29}$$

A second estimate of the *a priori* SNR can be obtained from the preceding frame's speech estimate [15]

$$\hat{\xi}_{\ell-1}(k,\ell) = \frac{|\hat{X}(k,\ell-1)|^2}{\hat{\sigma}_d^2(k,\ell-1)}. \tag{2.30}$$

As a very prominent approach, the decision-directed (DD) estimator linearly combines the two estimates as follows [15]

$$\hat{\xi}_{\mathrm{DD}}(k,\ell) = \alpha_{\mathrm{DD}}\hat{\xi}_{\ell-1}(k,\ell) + (1-\alpha_{\mathrm{DD}})\max[\hat{\xi}_{\mathrm{ML}}(k,\ell), 0], \tag{2.31}$$

where α_{DD} is typically chosen in the range $0.9 \leq \alpha_{\mathrm{DD}} < 1$, depending on the chosen analysis and modification stage parameters. The specific choice of α_{DD} determines the trade-off between tracking-delay and the variance of the estimator. The larger the choice of α_{DD}, the less musical noise artifacts occur at the cost of larger tracking-delay that may be perceived as reverberation in the processed signal. Although the DD approach is an inconsistent estimator [47], it is widely used since it performs well while being simple to implement. It has been thoroughly analyzed in [16, 74] and alternatives or adaptations of the DD approach are presented in e.g. [4, 60, 75–83].

2.2.4 Extensions to the Classical Gaussian Model

The independence of magnitude and phase in Equation (2.18) enables to independently model the prior pdfs of magnitude and phase. The circularity of the joint distributions, i.e. the uniform prior for the phase, has only begun to be questioned in e.g. [24, 27, 84–86]. Meanwhile, a lot of effort has been put in finding prior distributions for the spectral magnitude beyond the Rayleigh model. Prominent examples include the works in [17, 54, 55, 77, 87–89].[7]

[7] Appendix D lists some relevant prior distributions.

Further, in [90] an estimator of the logarithmically compressed magnitude, termed log spectral amplitude (LSA) estimator, has been proposed. The reasoning behind estimating the logarithm of the magnitude rather than the magnitude itself is to take into account the compressive characteristics of the cochlea. As a generalization of estimating compressed magnitudes, the work in [91], suggests so-called β-order MMSE-STSA estimators, derived from evaluating $\mathbb{E}\big(A^{\beta}|Y\big)$. The work in [92] fuses a parameterizable amplitude prior with the β-order distortion measure, yielding a flexible class of estimators that includes many preceding approaches as special cases.

Most importantly for this thesis, independent from the estimation criteria, the amplitude distributions chosen, and the specific distortion measure, the noisy phase remains the optimal estimate, as long as the underlying models do not reconsider the prior phase distribution.

2.3 Conclusion

In this chapter, we have reviewed classical STFT based single-channel speech enhancement algorithms. Particularly, we have also brought to the surface under which conditions, i.e. under which signal model, the noisy phase is the optimal estimate of the speech phase. To summarize, the noisy phase is optimal from an estimation theoretic perspective if

1. The prior distributions of the complex-valued speech STFT coefficients are modeled to be circular.
2. The prior distributions of the complex-valued noise STFT coefficients are modeled to be circular.
3. The signal model implies independence between the STFT coefficients w.r.t. time *and* frequency.

It is important to note that the above criteria assess the validity of the assumption that the noisy phase is optimal only in terms of mathematical reasoning. Perceptual considerations may yield a different result. In the following, we will explicitly present and investigate signal models that violate some of the above conditions.

3 Phase Estimation in Single-Channel Speech Enhancement

This chapter reviews single-channel speech enhancement approaches that incorporate knowledge about the spectral phase into optimal estimation of the STFT coefficients. While on the one hand, estimates of the STFT coefficients' phase may be used for signal reconstruction, i.e. replacing the noisy STFT phase, it is also possible to refine amplitude estimates by incorporating knowledge about the spectral phase. Both approaches can be found in the literature. We overview this topic and further introduce concepts such as the harmonic phase decomposition that will be used throughout the rest of this thesis.

3.1 Overview and Motivation

Chapter 2 concluded on why the classical signal model for STFT-based speech enhancement and many of its adaptations yield the noisy phase as the optimal phase estimate. As explained, the assumptions that lead to this result are rather restrictive. Besides the statistical modeling, it is important to note that a spectrogram corresponds to a time domain signal which is segmented by means of overlapping frames. Hence, overlapping segments are redundant, i.e. by definition of the STFT they have the same content if the weighting of the time domain samples due to the window function is compensated. An STFT representation that fulfills this redundancy constraint is called *consistent*. The independently estimated magnitudes in conjunction with the noisy phase will in general not yield a consistent STFT.

The problem of finding consistent phase functions from magnitude spectrograms is commonly termed phase-retrieval or spectrogram inversion and can also be seen as the problem of reconstructing a time domain signal from a given STFT magnitude. The literature on phase retrieval is widespread, as the underlying problem arises in many diverse scientific fields, e.g. in optics [93], astronomical imaging [94, 95], and X-ray crystallography [96]. The work in [97] investigates conditions under which a signal is uniquely defined by its STFT magnitude and proposes algorithms that can be applied to reconstruct signals from their magnitude spectra. Alternating projection algorithms that iteratively reconstruct the time domain signal and recompute a STFT phase function from the resulting time domain signal are particularly popular. Such methods typically rely on the work of Griffin and Lim in [52] and are often also referred to as Griffin-Lim (GL) algorithms. Various variants of the GL algorithm have been proposed, aiming to make it real-time compatible [98,99], to improve its initialization and convergence speed [100–102], to omit the necessity to repeatedly compute the STFT [103], or to fit the algorithm to specific signal classes [101]. In the context of source separation, the consistency constraint has also been used in conjunction with Wiener filtering in [104].

For Gaussian analysis windows, the log-magnitude and the phase of the STFT are connected by their derivatives along time and frequency, resulting in a coupled pair of first order linear partial differential equations [105, 106]. Recently, based on this connection, the works in [107, 108] propose non-iterative phase reconstruction algorithms that integrate estimates of these derivatives over time and frequency to reconstruct the phase function. This procedure is referred to as Phase Gradient Heap Integration. An overview on recent developments in phase retrieval can be found in [109].

We note that many of the above mentioned phase retrieval algorithms rely on a known spectral amplitude. However, in speech enhancement, we only have estimates of the STFT magnitude available, meaning that finding a consistent STFT phase does not necessarily mean that this phase estimate is perceptually beneficial or is closer to a reasonable STFT phase of speech signals. Most importantly, a consistent STFT phase ensures that the optimal magnitude estimate is preserved throughout the overlap-add reconstruction [104]. However, as the magnitude estimate is only optimal w.r.t. the underlying statistical model, it is also not evident that it is advantageous to ensure the exact preservation of the independently obtained magnitude estimates, as the overlap-add re-introduces correlations due to the implicit averaging of overlapping segments.

The present thesis does not explicitly consider approaches relying on the (in)consistency principle. Instead, this thesis employs the harmonic plus noise model for speech [110–112] in many places, resulting in a mathematical model for the spectral phase that can be exploited for speech enhancement. The harmonicity of voiced speech is a well established property and has been incorporated into speech enhancement in various ways, e.g. in [77, 113, 114]. The recent literature on phase estimation for speech enhancement almost exclusively relies on representing voiced parts of the speech signal as a sum of harmonically related sinusoids, e.g. [24, 28, 115]. Notable work on exploiting the harmonic or more generally speaking, sinusoidal characteristics of voiced speech sounds for speech enhancement is also presented in [116].

The rest of this chapter is organized as follows. First, we introduce the harmonic plus noise model together with the concept of phase normalization that will be important throughout this thesis. Further, we explain recently proposed methodologies that exploit the harmonic plus noise model for single-channel speech enhancement, resulting in optimal phase estimates that are not necessarily equal to the noisy phase.

3.2 The Harmonic plus Noise Model

Originally proposed for speech-synthesis in [112], the harmonic plus noise model for speech is the basis for many speech enhancement algorithms relevant for this thesis. Here, speech is represented as (i) the sum of harmonically related sinusoids $x_{\mathrm{t,H}}(n)$, particularly suitable for voiced speech sounds, and (ii) noise-like contributions $x_{\mathrm{t,S}}(n)$, which represent all components that are not captured by the harmonic model. Hence, the speech signal may be represented as

$$\begin{aligned} x_{\mathrm{t}}(n) &= x_{\mathrm{t,H}}(n) + x_{\mathrm{t,S}}(n) \\ &= \sum_{h=1}^{N_{\mathrm{H}}} A'_h(n) \cos\left(2\pi \int_{t=0}^{n/f_s} f_{\mathrm{c},h}(t)dt + \alpha'_h(n)\right) + x_{\mathrm{t,S}}(n), \end{aligned} \tag{3.1}$$

where $h \in [1, N_{\mathrm{H}}]$ is the harmonic index, N_{H} is the order of the harmonic model, $A'_h(n)$ is the harmonic amplitude, $f_{\mathrm{c},h}(t)$ is the continuous-time harmonic frequency trajectory, and $\alpha'_h(n)$ is the harmonic phase. The harmonic plus noise model is also referred to as stochastic-deterministic model in the speech enhancement literature [56, 86, 117]. We note that for the noise-like (stochastic) contributions we stick to the zero-mean Gaussian model presented in Section 2.2. However, the structure of the harmonic components is not captured by such models.

The purely harmonic signal representation for speech by means of $x_{\mathrm{t,H}}(n)$ is highly compact compared to the DFT, as it concentrates a large amount of signal energy into a relatively small fraction of coefficients, i.e. $N_{\mathrm{H}} \ll N$. However, speech does not only comprise of harmonic content and a good representation of $x_{\mathrm{t,S}}(n)$ is needed, otherwise the harmonic plus noise representation is not complete in the sense that the underlying signal can be perfectly reconstructed from the corresponding coefficients. Among other properties presented earlier, it is the completeness of the DFT that motivates us to stay faithful to the STFT, which is why we will investigate the mapping of the harmonic parameters amplitude, phase, and frequency to the STFT domain in the following.[8] Further, this also enables us to combine any estimation scheme for the harmonic content with classical STFT speech enhancement strategies. Throughout what follows we will assume the fundamental frequency to be given from some dedicated estimator.

[8] The notion of compactness and completeness used here is based on [118], where compact speech representations for speech synthesis are investigated.

3.2.1 Harmonic Signals in the STFT

The windowing of the harmonic components in Equation (3.1) according to Equation (2.2) yields

$$x_{\mathrm{t,H,w}}(n,\ell) = w(n)\sum_{h=1}^{N_{\mathrm{H}}} A'_h(n+\ell L)\cos\left(2\pi\int_{t=0}^{(n+\ell L)/f_s} f_{\mathrm{c},h}(t)dt + \alpha_h(n+\ell L)\right). \tag{3.2}$$

At this point, we will assume that the frame length is chosen such that the harmonic parameter triplet $\{A'_h(n), f'_h(n), \alpha'_h(n)\}$ can be considered to be constant within this time span. Hence, we will only consider sub-sampled versions and substitute for $n \in [0, N-1]$

$$\begin{aligned} A_h(\ell) &= A'_h(n+\ell L), \\ \alpha_h(\ell) &= \alpha'_h(n+\ell L), \\ \omega_h(\ell) &= 2\pi\frac{f_h(\ell)}{f_s} = 2\pi\frac{f_{\mathrm{c},h}\left(\frac{n+\ell L}{f_s}\right)}{f_s}. \end{aligned} \tag{3.3}$$

Note that this simplification renders model mismatches that are not stochastic in nature to be a contribution to the stochastic term $x_{\mathrm{t,S}}(n,\ell)$ in Equation (3.1), e.g. in case of nonstationary voiced speech. Recently, approaches for enhancing voiced speech coping with the non-stationarity of the signal parameters have been proposed in [119, 120]. In this thesis, due to its simplicity and applicability, we stick to the stationary (or constant) model. The notation $\omega_h(\ell)$ allows for an arbitrary choice of the h^{th} partial's frequency. However, in this thesis we implicitly assume a harmonic relationship between the respective partial and a fundamental frequency $\omega_0(\ell)$, i.e. $\omega_h(\ell) = h\omega_0(\ell)$.

By using the sub-sampled harmonic parameters in Equation (3.3), we can rewrite Equation (3.2) as

$$x_{\mathrm{t,H,w}}(n,\ell) = w(n)\sum_{h=1}^{N_{\mathrm{H}}} A_h(\ell)\cos\left(2\pi\left(\int_{t=0}^{\ell L/f_s} f_{\mathrm{c},h}(t)dt + \int_{t=(\ell L)/f_s}^{(n+\ell L)/f_s} f_{\mathrm{c},h}(t)dt\right) + \alpha_h(\ell)\right) \tag{3.4}$$

where, if the fundamental frequency is constant within a frame's time span, we have

$$\begin{aligned} 2\pi\int_{t=\ell L/f_s}^{(n+\ell L)/f_s} f_{\mathrm{c},h}(t)dt &= 2\pi f_{\mathrm{c},h}\left(\frac{n+\ell L}{f_s}\right)\cdot\left(\frac{n+\ell L}{f_s} - \frac{\ell L}{f_s}\right) \\ &= \omega_h(\ell)n. \end{aligned} \tag{3.5}$$

By substituting the harmonic phase progression due to the past fundamental frequency trajectory, i.e.

$$\psi_h(\ell) = 2\pi \int_{t=0}^{\ell L/f_s} f_{\mathrm{c},h}(t)dt, \tag{3.6}$$

we have

$$x_{\mathrm{t,H,w}}(n,\ell) = w(n) \underbrace{\sum_{h=1}^{N_{\mathrm{H}}} A_h(\ell)\cos\left(\omega_h(\ell)n + \phi_h(\ell)\right),}_{x_{\mathrm{t,H}}(n,\ell)} \tag{3.7}$$

with the instantaneous harmonic phase given by

$$\phi_h(\ell) = \psi_h(\ell) + \alpha_h(\ell). \tag{3.8}$$

While $\omega_h(\ell)$ and $\psi_h(\ell)$ are determined by the frequency trajectory of the respective harmonic, $\alpha_h(\ell)$ represents its initial orientation w.r.t. the analysis window, i.e. it is connected to the onset of the harmonic. Therefore, we will refer to $\alpha_h(\ell)$ as the *initial* phase [120] for the rest of this thesis. The initial phase has also been termed unwrapped phase in [24, 115]. Due to its close connection to the onset of the respective harmonic, which is a constant, $\alpha_h(\ell)$ itself can be considered to be slowly time-varying, i.e. we assume the harmonic phase to be coherent over the duration of one voiced syllable. Hence, it is mostly this part of the phase which is processed in one way or the other in the literature that deals with phase-processing for speech enhancement.

Inserting the model of Equation (3.7) into the definition of the DTFT yields

$$\begin{aligned} X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell) &= \sum_{n=-\infty}^{\infty} x_{\mathrm{H}}(n,\ell)w(n)\mathrm{e}^{-j\omega n} \\ &= \sum_{n=-\infty}^{\infty} w(n)\mathrm{e}^{-j\omega n} \sum_{h=1}^{N_{\mathrm{H}}} A_h(\ell)\cos\left(\omega_h(\ell)n + \phi_h(\ell)\right) \\ &= \sum_{h=1}^{N_{\mathrm{H}}} A_h(\ell) \sum_{n=-\infty}^{\infty} w(n)\mathrm{e}^{-j\omega n}\frac{1}{2}\left(\mathrm{e}^{j(\omega_h(\ell)n+\phi_h(\ell))} + \mathrm{e}^{-j(\omega_h(\ell)n+\phi_h(\ell))}\right) \end{aligned} \tag{3.9}$$

which, by applying the modulation theorem [48], can be rewritten as

$$\begin{aligned} X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell) = &\sum_{h=1}^{N_{\mathrm{H}}} \frac{A_h(\ell)}{2}\mathrm{e}^{j\phi_h(\ell)} W\left(\mathrm{e}^{j(\omega-\omega_h(\ell))}\right)\mathrm{e}^{-j(\omega-\omega_h(\ell))\frac{N-1}{2}} \\ &+ \sum_{h=1}^{N_{\mathrm{H}}} \frac{A_h(\ell)}{2}\mathrm{e}^{-j\phi_h(\ell)} W\left(\mathrm{e}^{j(\omega+\omega_h(\ell))}\right)\mathrm{e}^{-j(\omega+\omega_h(\ell))\frac{N-1}{2}} \end{aligned} \tag{3.10}$$

Here, $W(\mathrm{e}^{j\omega})$ is the DTFT of the non-causal, zero-phase window function. Hence, assuming that $w(n)$ is symmetric around $\frac{N-1}{2}$, which is the case for the window functions considered in this thesis, we have the linear phase window

$$\sum_{n=-\infty}^{\infty} w(n)\mathrm{e}^{-j\omega n} = W(\mathrm{e}^{j\omega})\mathrm{e}^{-j\omega\frac{N-1}{2}}. \tag{3.11}$$

By decomposing the instantaneous harmonic phase according to Equation (3.8), we have

$$\begin{aligned} X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell) = \sum_{h=1}^{N_{\mathrm{H}}} \frac{A_h(\ell)}{2}\mathrm{e}^{j\alpha_h(\ell)} W\left(\mathrm{e}^{j(\omega-\omega_h(\ell))}\right)\mathrm{e}^{-j((\omega-\omega_h(\ell))\frac{N-1}{2}-\psi_h(\ell))} \\ + \sum_{h=1}^{N_{\mathrm{H}}} \frac{A_h(\ell)}{2}\mathrm{e}^{-j\alpha_h(\ell)} W\left(\mathrm{e}^{j(\omega+\omega_h(\ell))}\right)\mathrm{e}^{-j((\omega+\omega_h(\ell))\frac{N-1}{2}+\psi_h(\ell))}. \end{aligned} \tag{3.12}$$

Given the analysis window together with the harmonic parameters, the DTFT (and hence the DFT) of the harmonic signal components is completely determined. The parametric representation in terms of the harmonic model is in general more compact than the time domain samples or the DFT expansion coefficients. Therefore, estimating the harmonic parameters is of special interest in many areas of speech processing. In speech enhancement, many approaches imposing the harmonic model are formulated as a parameter estimation task, relying on the hypothesis that estimating the harmonic parameters facilitates enhancing the noisy speech.

Without any approximations, the joint estimation of all harmonic parameters is in general a difficult task. Due to the interaction of the sinusoids, direct optimization is not possible [121], and iterative procedures have therefore been proposed in e.g. [1,122,123]. An alternative approach that simplifies the harmonic model substantially is to neglect the negative-frequency partials, implicitly modeling the speech signal as a complex-valued time domain signal, e.g. [124]. The works in [28, 86, 115] additionally assume that at a specific frequency, only one partial is present. In this case, once the harmonic frequencies are given, estimating the amplitude and phase of each partial can be carried out separately.

Rewriting Equation (3.12) under the above considerations means that each ω is assigned one *dominant* partial $\bar{h}(\omega,\ell)$ only, i.e.

$$X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell) \approx \frac{A_{\bar{h}(\omega,\ell)}(\ell)}{2}\mathrm{e}^{j\alpha_{\bar{h}(\omega,\ell)}(\ell)} W\left(\mathrm{e}^{j(\omega-\omega_{\bar{h}(\omega,\ell)}(\ell))}\right)\mathrm{e}^{-j((\omega-\omega_{\bar{h}(\omega,\ell)}(\ell))\frac{N-1}{2}-\psi_{\bar{h}(\omega,\ell)}(\ell))}. \tag{3.13}$$

Figure 3.1 illustrates the effect of the approximations culminating in Equation (3.13) in the frequency domain for $N_{\mathrm{H}} \in \{1, 2\}$ and for two different window choices, namely the rectangular and the Blackman window. The discrepancy between the dashed and the solid lines in both magnitude and phase indicates that the approximations do not even hold for the case of a signal consisting of a single real-valued sinusoid due to the mutual impact of negative and positive frequency components. However, within the mainlobe of the window around the center frequency, the approximation is accurate. Choosing an analysis window with high sidelobe attenuation, such as the Blackman window, reduces the error introduced by the harmonic approximation.

3.2.2 Phase Decomposition in the STFT

So far, we analyzed the DTFT of a harmonic signal. In practice, by using the FFT we obtain the DFT of the signal, meaning that we only access samples of the DTFT, given at sampling points $\omega_k = \frac{2\pi k}{N_{\mathrm{DFT}}}$, i.e.

$$\begin{aligned} X_{\mathrm{H}}(k,\ell) &= X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell)|_{\omega=\omega_k} \\ &\approx \frac{A_{\bar{h}(\omega_k,\ell)}(\ell)}{2} \mathrm{e}^{j\alpha_{\bar{h}(\omega_k,\ell)}(\ell)} W\left(\mathrm{e}^{j(\omega_k - \omega_{\bar{h}(\omega_k,\ell)}(\ell))}\right) \mathrm{e}^{-j((\omega_k - \omega_{\bar{h}(k,\ell)}(\ell))\frac{N-1}{2} - \psi_{\bar{h}(\omega_k,\ell)}(\ell))}. \end{aligned} \tag{3.14}$$

From the approximate DFT representation of the harmonic signal components in Equation (3.14) we see that given the parameter triplet $\{A_{\bar{h}(\omega_k,\ell)}(\ell), \alpha_{\bar{h}(\omega_k,\ell)}(\ell), \omega_{\bar{h}(\omega_k,\ell)}(\ell)\}$, we can expect a certain structure along frequency and time in the STFT coefficients. On the one hand, we see that phase differences along time are governed by the fundamental frequency trajectory, which determines $\psi_{\bar{h}(\omega_k,\ell)}(\ell)$. On the other hand, the chosen window function will contribute to the phase-progression along frequency. Due to the wrapping of the spectral phase to its principal interval $[-\pi, \pi)$, the structure in the phase is not directly accessible from the DFT. We therefore decompose the phase by subtracting the harmonic phase progression due to the fundamental frequency and the linear phase of the window function. By defining

$$\Psi(k,\ell) = \mathrm{e}^{-j((\omega_k - \omega_{\bar{h}(k,\ell)}(\ell))\frac{N-1}{2} - \psi_{\bar{h}(\omega_k,\ell)}(\ell))}, \tag{3.15}$$

we have

$$\begin{aligned} \widetilde{X}_{\mathrm{H}}(k,\ell) &= X_{\mathrm{H}}(k,\ell)\Psi^*(k,\ell) \\ &= \frac{A_{\bar{h}(\omega_k,\ell)}(\ell)}{2} W\left(\mathrm{e}^{j(\omega_k - \omega_{\bar{h}(\omega_k,\ell)}(\ell))}\right) \mathrm{e}^{j\alpha_{\bar{h}(\omega_k,\ell)}(\ell)} \end{aligned} \tag{3.16}$$

where the tilde sign $\widetilde{\cdot}$ indicates the phase-normalized STFT coefficient, a term originally proposed in [86]. A similar principle is referred to as phase unwrapping, linear phase removal, and phase decomposition in [8, 24, 115, 125].

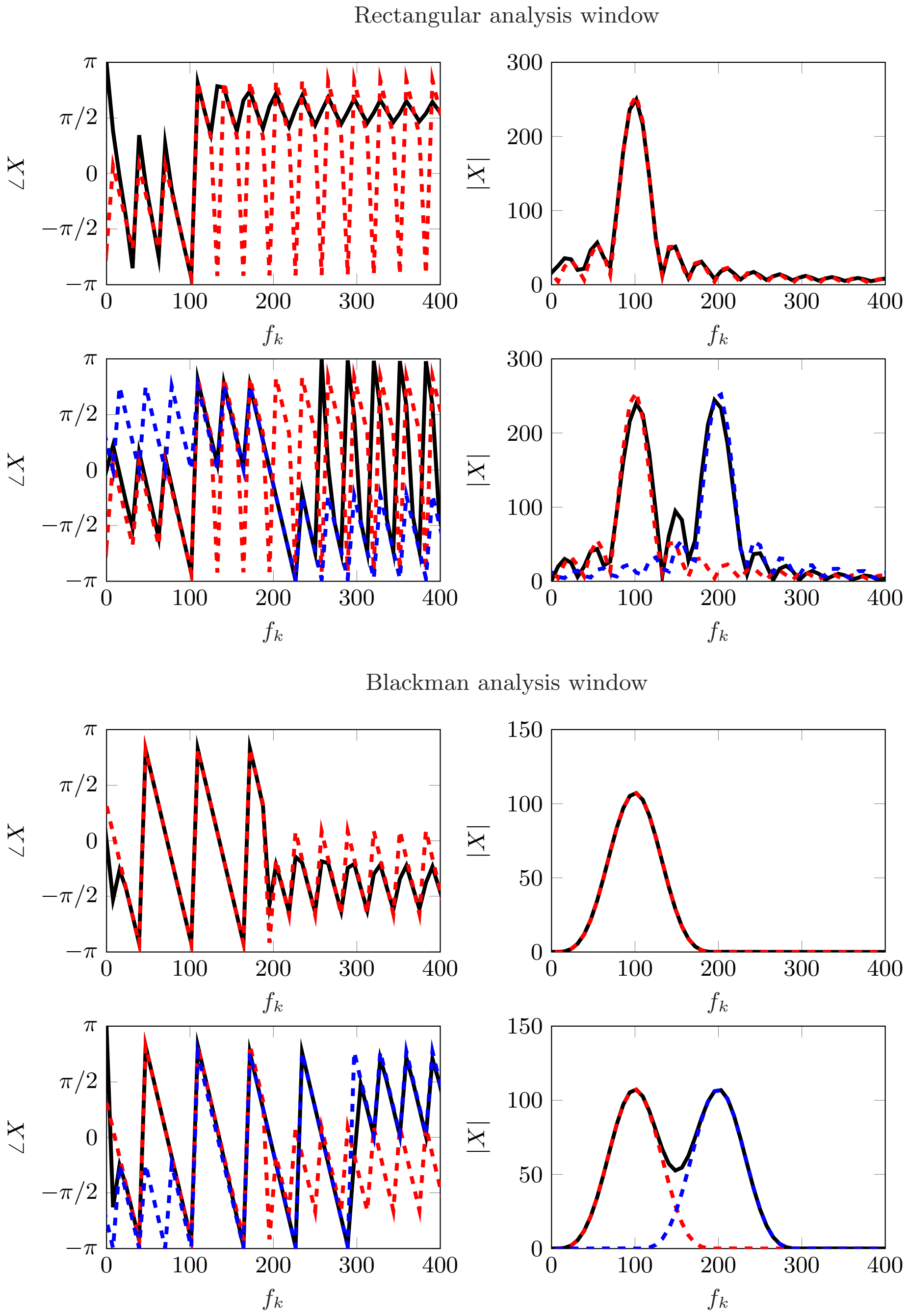

Figure 3.1: The effect of the approximation in Equation (3.13). Black curves: Spectrum of the (sum of) real-valued sinusoid(s). Dashed red and blue curves: individual, complex-valued sinusoids. Left Panels: spectral phase. Right panels: spectral magnitude. Upper panels: rectangular window. Lower panels: Blackman window.

Since $|\Psi(k,\ell)| = 1$, the STFT magnitude is not affected by the phase-normalization. In regions where the approximation in Equation (3.13) is valid, i.e. in the close vicinity of harmonic frequencies, we expect the STFT phase to correspond to the initial phase of the dominant harmonic instead of taking on random values. Further, in the presence of additive noise, depending on the SNR, this structure will be deteriorated. As discussed in Section 3.2.1, the initial harmonic phase can be associated to the initial alignment of the harmonic to the analysis window. While the initial alignment itself is completely random, hence adequately modeled by a uniform distribution, it is reasonable to assume that it does not change rapidly since we consider the harmonic phase to be coherent over the duration of a voiced syllable.

The Linear Phase Window

In the signal analysis stage throughout this thesis, we use linear-phase window functions as in Equation (3.11). Although the linear phase is represented by the term $\mathrm{e}^{-j(\omega_k - \omega_{\bar{h}(k,\ell)}(\ell))\frac{N-1}{2}}$, the linear phase term does not represent jumps of $\pm\pi$ at the roots of the non-causal window's DTFT. To this end, it is important to keep $W\left(\mathrm{e}^{j(\omega_k - \omega_{\bar{h}(\omega_k,\ell)}(\ell))}\right)$ instead of $\left|W\left(\mathrm{e}^{j(\omega_k - \omega_{\bar{h}(\omega_k,\ell)}(\ell))}\right)\right|$, in Equation (3.16). Figure 3.2 illustrates the discontinuities in the window phase function.

If we reconsider the approximation that there is only one dominant harmonic per DFT bin, we may restrict the region associated to a specific partial to the mainlobe width of the analysis window around its center frequency. Hence, we consider a DFT bin to be dominated by a harmonic only if it is "noteworthy" influenced by a harmonic. Figure 3.2 also shows that this approximation is especially reasonable within the mainlobe around a harmonic's frequency. It is important to note that within the mainlobe, we have $\angle W\left(\mathrm{e}^{j(\omega_k - \omega_{\bar{h}(\omega_k,\ell)}(\ell))}\right) = 0$, yielding $\angle \widetilde{X}_{\mathrm{H}}(k,\ell) = \alpha_{\bar{h}(\omega_k,\ell)}(\ell)$ in this case.

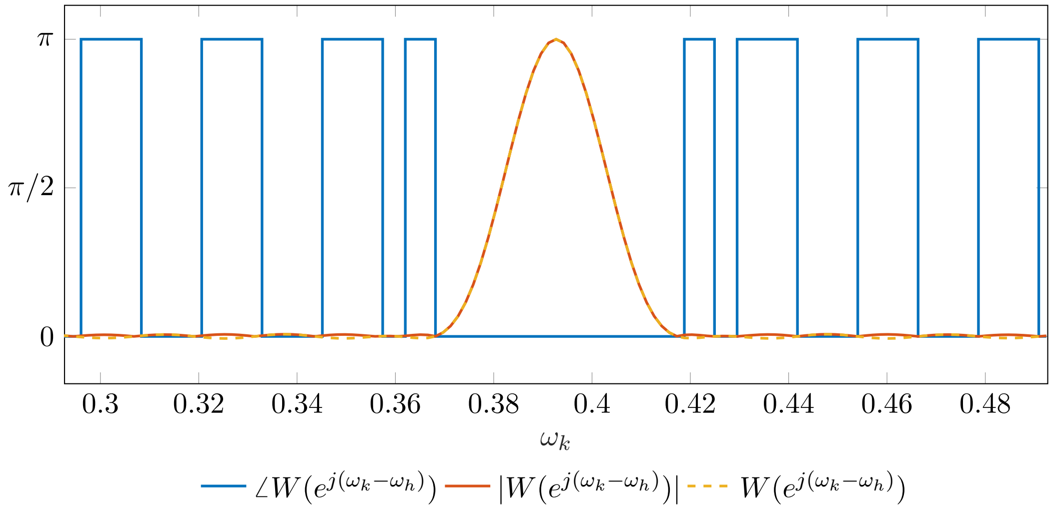

Figure 3.2: The discontinuities of the phase function of a non-causal, symmetric, i.e. zero-phase, window function. Blue: Phase of the window function. The magnitude has been scaled to have a maximum value of π and ω_h has been chosen with $\omega_h = 1000 \cdot 2 \cdot \pi/16000 \approx 0.3927$.

The Harmonic Phase Progression

Reconsidering Equation (3.6) and taking into account that in practice, we only have a sampled version $f_h(\ell) = f_{h,c}(\ell L/f_s)$ of $f_{h,c}(t)$ available, we can compute the phase progression term $\psi_h(\ell)$ recursively under the assumption that it can be approximated by a piece-wise linear function from one frame to the next, i.e.

$$\begin{aligned} \psi_h(\ell) &= \underbrace{2\pi \int_0^{t_1=\frac{(\ell-1)L}{f_s}} f_{h,c}(t)dt}_{\psi_h(\ell-1)} + 2\pi \int_{t_1=\frac{(\ell-1)L}{f_s}}^{t_2=\frac{\ell L}{f_s}} f_{h,c}(t)dt \\ &= \psi_h(\ell-1) + 2\pi \int_{t_1}^{t_2} \Big(f_{h,c}(t_1) + (t-t_1)\frac{f_{h,c}(t_2)-f_{h,c}(t_1)}{t_2-t_1} \Big) dt \\ &= \psi_h(\ell-1) + \frac{L}{2}\big(\omega_h(\ell) + \omega_h(\ell-1)\big). \end{aligned} \tag{3.17}$$

Given the assumption that (i) Equation (3.17) accurately models the harmonic phase progression along time and (ii) harmonics dominate DFT bins around their center frequency, we can expect the speech signal to follow a certain structure along time and frequency. In the following, we will review relevant literature that makes use of this structure to estimate a reconstruction phase from the noisy observation.

3.2.3 Exploiting Harmonic Phase Decomposition for Speech Enhancement

A recent approach to impose the aforementioned phase structure for signal reconstruction was proposed in [28]. The method known as STFT phase improvement (STFTPI) [28] aims to retrieve the harmonic phase structure in voiced speech under the assumption that

$$\bar{h}(\omega_k, \ell) = \arg\min_{h} |\omega_k - \omega_h(\ell)|^2 , \tag{3.18}$$

where we recall $\bar{h}(\omega_k, \ell)$ to be the dominant harmonic at DFT bin k and frame ℓ. T Therefore, each DFT bin is assigned the phase of the closest harmonic, in this case the phase of the noisy signal evaluated at $\omega_{\bar{h}(\omega_k,\ell)}(\ell)$. This procedure is termed "phase reconstruction along frequency" in [28]. The authors of [28] also propose a temporal phase reconstruction routine that aims to reconstruct the harmonic phase progression along time in Equation (3.6). Starting from an initial phase that is set equal to the noisy phase at the beginning of a voiced speech sound, the harmonic frequency is numerically integrated from one frame to the next, aligning the phases of harmonic components over consecutive segments. In [126], the STFTPI method was combined with a least squares (LS) estimate of the initial phase, replacing the noisy initial phase. The "phase reconstruction along time" is conceptually very similar to the algorithms proposed in [127, 128] for sinusoidal signal analysis/synthesis and low bit rate audio coding, respectively. In principle, one could interpret the phase reconstruction along frequency as a method that retrieves a typical group delay (GD) function of speech, while the reconstruction along time algorithm aims to enhance the instantaneous frequency (IF) of the noisy signal.

The works in [25, 129] model the initial or unwrapped harmonic phase as a directional random variable.[9] Under the assumption that additive noise contributes to increased circular variance, the variance of the initial phase is reduced by temporal smoothing of the noisy initial phase. In [115], this approach is extended by formulating a MAP estimator of the initial harmonic phase that enables the phase statistics, i.e. its circular variance, to be incorporated as a parameter that controls the smoothing depth. Following the same argumentation as in Section 3.2.2, the resultant harmonic phase estimates are assigned to DFT bins lying within the mainlobe width around the harmonic frequency.

To summarize, in reconstruction phase estimation, it is particularly important to retrieve a specific *phase structure* imposed by the harmonic model, rather than the exact absolute value of the phase. The next section will present methods that use estimates of the spectral phase, potentially obtained using methods discussed in this section, to refine the estimation of the speech STFT coefficients in general.

[9] For an overview on directional/circular statistics see Appendix C.

3.3 Phase-Sensitive MMSE Estimators

In this section, we assume that *some* a priori information ϕ_μ about the spectral speech phase ϕ is available. Starting from the independence assumption of magnitude and phase in Equation (2.18), a straightforward approach to incorporate this knowledge into statistical model-based speech estimators is to assume a distribution with pdf $p\left(\phi|\phi_\mu\right)$, modeling the uncertainty about this a priori information. In conjunction with β-order compression together with a χ-distribution for the spectral amplitudes we obtain the (C)omplex spectral speech coefficients given (U)ncertain (P)hase information (CUP) estimator proposed by Gerkmann in [27]

$$\begin{aligned}\hat{X}^{(\beta)}_{\mathrm{CUP}} &= \mathbb{E}\big(A^\beta \mathrm{e}^{j\phi}|Y,\phi_\mu\big)\\ &= \frac{\Gamma\left(2\mu+\beta\right)}{\Gamma\left(2\mu\right)}\left(\frac{\xi\sigma_d^2}{2\left(\mu+\xi\right)}\right)^{\frac{\beta}{2}} \frac{\int_0^{2\pi} \mathrm{e}^{j\phi+\frac{\nu^2}{4}} D_{-2\mu-\beta}\left(\nu\right) p\left(\phi|\phi_\mu\right) d\phi}{\int_0^{2\pi} \mathrm{e}^{\frac{\nu^2}{4}} D_{-2\mu}\left(\nu\right) p\left(\phi|\phi_\mu\right) d\phi},\end{aligned} \tag{3.19}$$

where $D_\cdot\left(\cdot\right)$ is the parabolic cylinder function, μ is the shape parameter of the χ-distribution,[10] and

$$\nu = -\sqrt{2\zeta\frac{\xi}{\mu+\xi}}\cos\left(\phi-\vartheta\right). \tag{3.20}$$

The final estimate is obtained by compensating for the β-order compression, i.e.

$$\hat{X}_{\mathrm{CUP}} = \left|\hat{X}^{(\beta)}_{\mathrm{CUP}}\right|^{\frac{1-\beta}{\beta}} \hat{X}^{(\beta)}_{\mathrm{CUP}}. \tag{3.21}$$

The feasibility of solving the integral in Equation (3.19) analytically heavily depends on the choice of $p\left(\phi|\phi_\mu\right)$. In the original work in [27], a von Mises distribution was proposed[11] and in this case, the integral has to be solved numerically as an analytical solution could not be found. The standard assumption of choosing $p\left(\phi|\phi_\mu\right)$ to be a uniform distribution again yields the noisy phase as the optimal estimate [29]. Assuming the prior phase information is deterministic, by inserting $p\left(\phi|\phi_\mu\right) = \delta\left(\phi-\phi_\mu\right)$ into Equation (3.19) and using the sifting property of the delta distribution we obtain [29]

$$\hat{X}^{(\beta)}_{\mathrm{CDP}} = \frac{\Gamma\left(2\mu+\beta\right)}{\Gamma\left(2\mu\right)}\left(\frac{\xi\sigma_d^2}{2\left(\mu+\xi\right)}\right)^{\frac{\beta}{2}} \frac{D_{-2\mu-\beta}\left(-\sqrt{2\zeta\frac{\xi}{\mu+\xi}}\cos\left(\phi_\mu-\vartheta\right)\right)}{D_{-2\mu}\left(-\sqrt{2\zeta\frac{\xi}{\mu+\xi}}\cos\left(\phi_\mu-\vartheta\right)\right)} \mathrm{e}^{j\phi_\mu}. \tag{3.22}$$

[10] For details we refer to Appendix D.

[11] The von Mises distribution together with other relevant circular distributions is discussed in further detail in Appendix C.

Interestingly, the magnitude of the above estimator is also the result of optimizing for the spectral magnitude under the assumption of deterministic knowledge about the clean speech spectral phase proposed in [26]. Further, the case $\beta = 1$ and $\mu = 1$ has already been considered in [130] in 1983, which is surprising considering that the other estimators discussed in this subsection have been proposed rather recently.

Spectral amplitude estimators derived under the assumption of a non-uniform $p\left(\phi|\phi_\mu\right)$ have been further investigated in [29]. Typically, the prior phase information is obtained using harmonic signal modeling and mapping the harmonic phase pattern to the STFT domain as explained in Section 3.2.3. A thorough analysis of the estimators resulting from Equation (3.19) in terms of a subjective listening test can be found in [30], where increased noise reduction performance compared to phase insensitive approaches was reported.

Apart from directly estimating the STFT coefficients, phase-sensitive amplitude estimators have also been incorporated in iterative estimation schemes of the complex STFT coefficients. These estimators rely on a (harmonic) model-based phase estimate. This magnitude estimate is then used to refine the phase estimate and vice versa. While the work in [23] relies on iteratively applying MMSE estimators, the work in [11] aims to approximate a joint MAP solution of the amplitude and phase under the assumption of a von Mises distribution for the spectral phase and a gamma distribution for the spectral amplitudes.

3.4 MMSE Estimators Incorporating the Harmonic Model

The harmonic plus noise model not only provides information about the structure of the spectral phase, but intrinsically assumes harmonically spaced maxima in the spectrum. Hence, the independence of the coefficients w.r.t. frequency is particularly questionable in the case of voiced speech [131]. In [56], a stochastic-deterministic (SD) signal model is proposed, where the deterministic signal components correspond to the summation of harmonically related, complex-valued sinusoids, i.e. the positive frequency components of $x_{\mathrm{t,H}}(n)$ only, and the stochastic components correspond to $x_{\mathrm{t,S}}(n)$ in Equation (3.1). While in [56], hard- and soft-decision estimation schemes are derived under the SD signal model, the work in [132] shows how to combine the two signal models (and potentially more) following the minimum-variance distortionless response (MVDR) filter design principle.

In the literature, the SD signal model is presented in conjunction with the following statistical assumptions. The linearity of the DFT preserves the summation of the components in the frequency domain yielding the stochastic-deterministic model, where the speech STFT coefficients are modeled by RVs that are circularly distributed around the "spectral mean" X_{H}, i.e. $X \sim \mathcal{CN}\left(X_{\mathrm{H}}, \sigma_x^2\right)$. Here, the variance around the mean value represents the stochastic signal components, i.e. $\sigma_x^2 = \mathbb{E}\big(|X - X_{\mathrm{H}}|^2\big)$.

Under the light of the previous section, where $p\left(\phi|\phi_\mu\right)$ was incorporated into MMSE estimation of the speech STFT coefficients, it is interesting to note that the distribution of ϕ conditioned on $|X|$ is a von Mises distribution [27]

$$\begin{aligned} p\left(\phi|\phi_\mu\right) &= p\left(\phi|\left|X\right|, \angle X_\mathrm{H}, \left|X_\mathrm{H}\right|, \sigma_x^2\right) \\ &= \frac{1}{2\pi I_0\left(\kappa\right)} \mathrm{e}^{\kappa \cos\left(\phi-\phi_\mu\right)} \end{aligned} \tag{3.23}$$

with mean phase $\phi_\mu = \angle X_\mathrm{H}$ and concentration parameter

$$\kappa = \frac{2\left|X_\mathrm{H}\right|\left|X\right|}{\sigma_x^2}. \tag{3.24}$$

This result is a justification for using harmonic phase estimates in combination with the estimator in Equation (3.19) [27], as the von Mises distribution can also be seen as a result of harmonic signal modeling. However, the selection of the concentration parameter κ in [27] was not based on Equation (3.24) but on a voicing probability estimate. The connection of a directional RV's concentration, i.e. circular variance, and the voicing probability has also been exploited in [7] for formulating a voice-activity detector.

The Stochastic-Deterministic Wiener Filter

Formulating the posterior speech distribution given the prior as $\mathcal{CN}\left(X_\mathrm{H}, \sigma_x^2\right)$, yields the MMSE estimator of the complex-valued STFT coefficients, termed the SD Wiener Filter in [117]

$$\begin{aligned} \hat{X}_\mathrm{SDW} &= G_\mathrm{W} Y + \left(1 - G_\mathrm{W}\right) X_\mathrm{H} \\ &= \underbrace{\left(G_\mathrm{W} + \left(1 - G_\mathrm{W}\right)\frac{X_\mathrm{H}}{Y}\right)}_{G_\mathrm{SDW}} Y. \end{aligned} \tag{3.25}$$

The Stochastic-Deterministic MMSE-STSA Estimator

Similar to the case of the circular complex normal distribution, in [86], a MMSE-STSA estimator under the assumption of a non-zero spectral mean has been proposed,

$$\hat{A}_\mathrm{SD} = \underbrace{\Gamma\left(\frac{3}{2}\right)\frac{\sqrt{\nu_\mathrm{SD}}}{\zeta}\mathrm{e}^{-\frac{\nu_\mathrm{SD}}{2}}\left(\left(1+\nu_\mathrm{SD}\right) I_0\left(\frac{\nu_\mathrm{SD}}{2}\right) + \nu_\mathrm{SD} I_1\left(\frac{\nu_\mathrm{SD}}{2}\right)\right)}_{G_\mathrm{SD}} R, \tag{3.26}$$

with

$$\nu_{\mathrm{SD}} = \frac{1}{1+\xi}\left(\xi\zeta + \frac{A_{\mathrm{H}}^2}{\sigma_d^2} + 2\sqrt{\frac{A_{\mathrm{H}}^2}{\sigma_d^2}\zeta}\cos\left(\phi_{\mathrm{H}} - \vartheta\right)\right). \tag{3.27}$$

Note that the scaling of magnitude estimates by $\cos\left(\phi_{\mathrm{H}} - \vartheta\right)$ can also be found in Equation (3.19) and Equation (3.22). However, in the stochastic-deterministic case, additional terms that account for the harmonic magnitude are introduced.

One possibility to reconstruct the time domain signal is to combine the magnitude estimate in Equation (3.26) with the noisy phase for reconstruction. The authors of [86] took a different path by evaluating the ML estimator of the phase given the non-circular model, resulting in the phase of SD Wiener filter estimator, i.e.

$$\hat{\phi}_{\mathrm{SDML}} = \angle \hat{X}_{\mathrm{SDW}}. \tag{3.28}$$

Hence, the final, complex-valued STFT coefficient estimator is obtained by

$$\hat{X}_{\mathrm{SD}} = \hat{A}_{\mathrm{SD}} \mathrm{e}^{j\hat{\phi}_{\mathrm{SDML}}}. \tag{3.29}$$

The SD MMSE-STSA estimator has been reported to be highly sensitive to estimation errors in X_{H} [86]. To compensate for the effects of estimation errors in the deterministic components, signal presence uncertainty (SPU) was incorporated. Under the assumption of two hypotheses: speech presence denoted by $\mathcal{H}_1$ and speech absence $\mathcal{H}_0$, this can be achieved by the following weighting [15]

$$\hat{X} = \frac{\Lambda(Y)}{1+\Lambda(Y)}\mathbb{E}\big(X|Y,\mathcal{H}_1\big) \tag{3.30}$$

with the generalized likelihood ratio (GLR) defined as

$$\Lambda(Y) = \frac{p\left(\mathcal{H}_1\right)p\left(Y|\mathcal{H}_1\right)}{p\left(\mathcal{H}_0\right)p\left(Y|\mathcal{H}_0\right)}, \tag{3.31}$$

where $p\left(\mathcal{H}_0\right)$ and $p\left(\mathcal{H}_1\right)$ represent the prior probability of speech absence and presence, respectively. Since speech is either present or not, we have $p\left(\mathcal{H}_0\right) = 1 - p\left(\mathcal{H}_1\right)$. In this thesis, alternatives to the SPU considerations under the harmonic plus noise model and to the harmonic signal component estimation that come with increased accuracy will be discussed.

3.5 Conclusion

This chapter reviewed approaches in single-channel speech enhancement that leverage *phase-aware* processing of the STFT coefficients. The harmonic plus noise signal model is particularly relevant in this thesis. Therefore, we specifically focused on methodologies that rely on harmonic signal modeling and introduced the concept of phase decomposition or phase normalization [86] in Section 3.2.2. The latter concept reveals structure in the STFT phase and will be a key concept throughout this thesis. The estimators presented in this chapter heavily depend on knowledge about the harmonic signal components, which are of course not known a priori. However, the presented estimators are highly sensitive to estimation errors of the harmonic components. Hence, it will be of special interest throughout this thesis to investigate approaches to refine the harmonic signal components' estimates and to find ways to incorporate the resultant insights into speech enhancement methods.

4

Simultaneous Detection-Estimation under the Harmonic plus Noise Model

This chapter presents a simultaneous detection-estimation approach under the harmonic plus noise model for speech. We will show that such an approach necessarily includes a harmonic retrieval problem, which will be of special interest throughout this chapter. All derivations and simulations included in this chapter have been conducted by the author. The idea of exploiting the benefits resulting from pitch-synchronous segmentation was motivated by discussions with Josef Kulmer and Pejman Mowlaee, who used the pitch-synchronous segmentation in the context of harmonic phase estimation for speech enhancement in [24, 115]. Simultaneous detection-estimation approaches are a well known strategy in speech enhancement, e.g. [133, 134], however, extending them to the harmonic plus noise model has been proposed by the author. Large parts of the text in this chapter have been published in [3].

4.1 Overview and Motivation

This chapter consists of three parts.

- The first part presents a rather general explanation of a simultaneous detection-estimation approach that facilitates fusing different signal models for the purpose of speech enhancement. Further, the more specific case of the harmonic plus noise model is considered and the corresponding optimal estimation and decision rules are derived. The statistical model that forms the basis of these derivations has also been used by other authors [40, 56, 132], however, the detection-estimation perspective on the problem was part of the original work in [3]. The derived estimator may be applied to any signal model that relies on the decomposition of a signal into stochastic and deterministic components.
- The second part of this chapter is focused on the estimation of the harmonic signal components from a noisy observation, subsequently needed in the simultaneous detection-estimation procedure. We show that the analysis stage of an AMS system as discussed in Chapter 2 impacts on the achievable accuracy of the harmonic estimators. We find that under certain assumptions, a pitch-synchronous signal

segmentation ensures that the approximations made in Section 3.2.1 hold. We therefore propose to use this segmentation paradigm in the signal analysis stage. Since the pitch-synchronous signal representation comes with interesting properties, explained in Section 4.4.1, we illustrate how they could be exploited in terms of an empirical Bayes approach.

- The third part is dedicated to evaluating the combination of the detection-estimation approach and the pitch-synchronous signal analysis stage in terms of instrumental metrics as well as a subjective listening test. The results indicate that the quality of harmonic signal component estimates may affect the intelligibility improvement capabilities of a speech enhancement system. Further, slight improvements in terms of perceived speech quality are achieved by the pitch-synchronous simultaneous detection-estimation (PSDE) framework.

The block diagram in Figure 4.1 illustrates the proposed speech-enhancement procedure, where we also highlight the three main modules, i.e. analysis, harmonic retrieval, and detection/estimation. Each of the stages can in general be replaced by alternative approaches solving the same task.

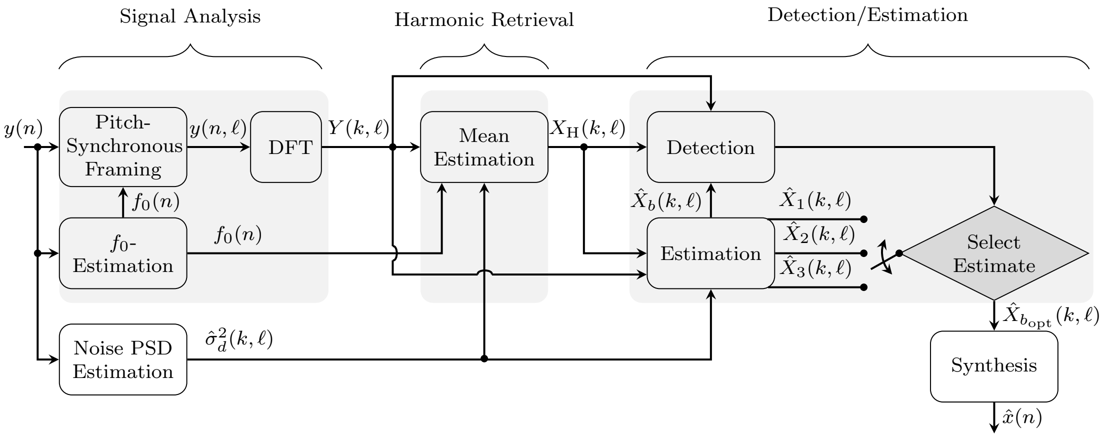

Figure 4.1: Block diagram of the proposed estimation scheme. Areas highlighted in gray indicate separate modules which can be replaced by any other method performing the tasks of signal analysis, harmonic retrieval, or detection/estimation.

4.2 Signal Model

As discussed in Section 3.2, the speech signal $x_{\mathrm{t}}(n,\ell)$ can be decomposed into a harmonic and a noise part, denoted by $x_{\mathrm{t,H}}(n,\ell)$ and $x_{\mathrm{t,S}}(n,\ell)$, respectively. The harmonic part $x_{\mathrm{t,H}}(n,\ell)$ is often referred to as *deterministic* part of the speech signal, e.g. in [40,56,110]. The noise part $x_{\mathrm{t,S}}(n,\ell)$, which is also called *stochastic* part [40, 56, 110], is represented by a zero-mean Gaussian process. Empirical evidence for this signal model was analyzed in [40] and originally, it has been proposed for speech synthesis in [110]. The harmonic model is especially appropriate for modeling voiced speech. In the case of unvoiced speech, the stochastic signal components can compensate for the model misfit that results from only considering $x_{\mathrm{t,H}}(n,\ell)$. Considering additive noise and taking the DFT of the noisy input frames yields

$$Y(k,\ell) = X_{\mathrm{S}}(k,\ell) + X_{\mathrm{H}}(k,\ell) + D(k,\ell). \tag{4.1}$$

A common problem of speech enhancement algorithms that rely on the harmonic model for voiced speech is that they tend to fit artificial higher order harmonics to the noise signal. By artificial we mean that they are not present in the underlying clean speech signal. To reduce such artifacts, in the following, we derive an MMSE estimator of the complex speech DFT coefficients conditioned on the three hypotheses voiced speech ($\mathcal{H}_2$), unvoiced speech ($\mathcal{H}_1$), and speech absence ($\mathcal{H}_0$). This means that we want to detect to which class an observed signal component belongs to and accordingly choose an estimator that is designed to handle the underlying signal class.

Given that we have three distinct hypotheses for the observed signal, each of the hypotheses can be considered to represent a signal model. For each of the hypotheses we may design a dedicated estimator. For the observation model in Equation (4.1) we can summarize the three hypotheses, together with the corresponding signal models as follows.

$$\begin{aligned}
\mathcal{H}_0: &\qquad Y(k,\ell) = D(k,\ell),\\
\mathcal{H}_1: &\qquad Y(k,\ell) = X_{\mathrm{S}}(k,\ell) + D(k,\ell),\\
\mathcal{H}_2: &\qquad Y(k,\ell) = X_{\mathrm{S}}(k,\ell) + X_{\mathrm{H}}(k,\ell) + D(k,\ell).
\end{aligned} \tag{4.2}$$

In the following, we will introduce the detection-estimation framework, and afterwards, we will address the harmonic retrieval problem. In the following, for the sake of readability, frequency and frame indices are dropped.

4.3 Simultaneous Detection and Estimation

To jointly optimize for detection and estimation, each decision is associated with a risk, determined by the chosen estimator $\hat{X}_b$ and the true underlying hypothesis $\mathcal{H}_{b'}$. The combined risk for detection and estimation under multiple hypotheses $\mathcal{R}$ is given by [133,135,136]

$$\mathcal{R} = \sum_{b\in\mathcal{B}} \int_{\mathcal{D}_Y} \int_{\mathcal{D}_X} C_b(X,\hat{X}|\mathcal{H}_{b'}) p(\eta_b|Y) p(Y|X) p(X) dX dY, \tag{4.3}$$

where $\mathcal{B}$ denotes the set of all hypotheses (in our case $\mathcal{B} = \{0, 1, 2\}$), $C_b(\cdot|\mathcal{H}_{b'})$ is a weighted cost function (chosen dependent on the underlying hypothesis $\mathcal{H}_{b'}$), and $\mathcal{D}_Y$ and $\mathcal{D}_X$ denote the domains of the respective variable in the subscript. The term $p(\eta_b|Y)$ corresponds to the probability of decision η_b, given the observation Y.

We are interested in minimizing the joint risk by means of (i) the chosen signal model and (ii) the accordingly selected estimator. Hence, we are searching for

$$\left\{b_{\text{opt}}, \hat{X}_{\text{opt}}\right\} = \arg\min_{b,\hat{X}} \mathcal{R}. \tag{4.4}$$

The above optimization can be carried out in the style of a grid search by means of a three stage procedure [133, 136] that consists of the following three steps.

1. Compute estimates for all hypotheses.

2. Evaluate the corresponding risks.

3. Choose the estimate that yields the minimum risk.

Below, the assumptions and derivations that yield the applicability of this rationale are explained in more detail.

From the above formulation, it is clear that we consider a binary decision framework, meaning that $p(\eta_b|Y)$ is either 0 or 1. Hence, based on the decision that is made, an estimator, derived from the underlying signal model, is applied. Since we have

$$p(X) = \sum_{b' \in \mathcal{B}} p(\mathcal{H}_{b'}) p(X|\mathcal{H}_{b'}) \tag{4.5}$$

we can rewrite Equation (4.3) as

$$\begin{aligned}\mathcal{R} = \int_{\mathcal{D}_Y} \int_{\mathcal{D}_X} p(Y|X) \sum_{b=0}^{2} p(\eta_b|Y) \Big(& C_b(X, \hat{X}|\mathcal{H}_0) p(\mathcal{H}_0) p(X|\mathcal{H}_0) \\ & + C_b(X, \hat{X}|\mathcal{H}_1) p(\mathcal{H}_1) p(X|\mathcal{H}_1) + C_b(X, \hat{X}|\mathcal{H}_2) p(\mathcal{H}_2) p(X|\mathcal{H}_2) \Big) dX dY.\end{aligned} \tag{4.6}$$

By introducing sub-risks of the form

$$r_b(Y|\mathcal{H}_{b'}) = \int_{\mathcal{D}_X} C_b(X, \hat{X}|\mathcal{H}_{b'}) p(Y|X) p(X|\mathcal{H}_{b'}) dX, \tag{4.7}$$

which are always associated to one pair of decision η_b and true hypothesis $\mathcal{H}_{b'}$, we can rewrite Equation (4.6) as

$$\begin{aligned}\mathcal{R} = \int_{\mathcal{D}_Y} & p(\eta_0|Y)(p(\mathcal{H}_0)r_0(Y|\mathcal{H}_0) + p(\mathcal{H}_1)r_0(Y|\mathcal{H}_1) + p(\mathcal{H}_2)r_0(Y|\mathcal{H}_2)) \\ & + p(\eta_1|Y)(p(\mathcal{H}_0)r_1(Y|\mathcal{H}_0) + p(\mathcal{H}_1)r_1(Y|\mathcal{H}_1) + p(\mathcal{H}_2)r_1(Y|\mathcal{H}_2)) \\ & + p(\eta_2|Y)(p(\mathcal{H}_0)r_2(Y|\mathcal{H}_0) + p(\mathcal{H}_1)r_2(Y|\mathcal{H}_1) + p(\mathcal{H}_2)r_2(Y|\mathcal{H}_2))dY\end{aligned} \tag{4.8}$$

Since the hypotheses exclude each other, the joint risk can take three different forms

$$\mathcal{R} = \begin{cases} \int_{\mathcal{D}_Y} \sum_{b'=0}^{2} p(\mathcal{H}_{b'})r_0(Y|\mathcal{H}_{b'})dY & \text{if } p(\eta_0|Y) = 1, \\ \int_{\mathcal{D}_Y} \sum_{b'=0}^{2} p(\mathcal{H}_{b'})r_1(Y|\mathcal{H}_{b'})dY & \text{if } p(\eta_1|Y) = 1, \\ \int_{\mathcal{D}_Y} \sum_{b'=0}^{2} p(\mathcal{H}_{b'})r_2(Y|\mathcal{H}_{b'})dY & \text{if } p(\eta_2|Y) = 1. \end{cases} \tag{4.9}$$

Hence, the optimal decision can be found as follows

$$b_{\text{opt}} = \arg\min_{b} \left\{ p(\mathcal{H}_0)r_b(Y|\mathcal{H}_0) + p(\mathcal{H}_1)r_b(Y|\mathcal{H}_1) + p(\mathcal{H}_2)r_b(Y|\mathcal{H}_2) \right\}. \tag{4.10}$$

Analogously, for each decision η_b we can find an optimal estimate of the speech DFT coefficient X by evaluating

$$\hat{X}_b = \arg\min_{\hat{X}} \left\{ p(\mathcal{H}_0)r_b(Y|\mathcal{H}_0) + p(\mathcal{H}_1)r_b(Y|\mathcal{H}_1) + p(\mathcal{H}_2)r_b(Y|\mathcal{H}_2) \right\}. \tag{4.11}$$

Following the description of the three stage procedure, we compute all solutions to Equation (4.11) and subsequently choose $\hat{X}_{b_{\text{opt}}}$ according to Equation (4.10).

Applying Equation (4.10) and Equation (4.11) sequentially in order to obtain optimal estimates with respect to both detection and estimation is a very general procedure and has been applied to speech enhancement in [133] under the two hypotheses speech absence and presence. Arbitrary signal models and cost functions may be used in the definition of the sub-risks in Equation (4.7). In the following we will have a closer look on the harmonic plus noise model for speech in order to find a statistical model that facilitates the speech estimation task.

4.3.1 Statistical Model

In Section 4.2, we assumed the stochastic speech components $x_{\text{t,S}}(n)$ and the noise signal $d_{\text{t}}(n)$ to be independent zero-mean Gaussian processes. As explained in Chapter 2, their frequency domain representations are often assumed to follow a zero-mean, second order circular, complex Gaussian distribution [15]. The STFT domain representation of the speech signal is given by $X = X_{\text{S}} + X_{\text{H}}$, i.e. the distribution of X is centered around X_{H}.

Hence, incorporating X_{H}, as discussed in Section 3.4, introduces the notion of a spectral mean value [86]. This mean value reflects the voicing state. We model the speech DFT coefficients by the following class-dependent prior distributions

$$p(X|\mathcal{H}_0) = \delta(X), \tag{4.12}$$

$$p(X|\mathcal{H}_1) = p(X_{\mathrm{R}}, X_{\mathrm{I}}|\mathcal{H}_1) = \frac{1}{\pi\sigma_x^2}\mathrm{e}^{-\frac{X_{\mathrm{R}}^2+X_{\mathrm{I}}^2}{\sigma_x^2}}, \tag{4.13}$$

$$p(X|\mathcal{H}_2) = p(X_{\mathrm{R}}, X_{\mathrm{I}}|\mathcal{H}_2) = \frac{1}{\pi\sigma_x^2}\mathrm{e}^{-\frac{(X_{\mathrm{R}}-\bar{X}_{2,\mathrm{R}})^2+(X_{\mathrm{I}}-\bar{X}_{2,\mathrm{I}})^2}{\sigma_x^2}}, \tag{4.14}$$

where $\bar{X}_{2,\mathrm{R}} = \mathrm{Re}\,(X_{\mathrm{H}})$ and $\bar{X}_{2,\mathrm{I}} = \mathrm{Im}\,(X_{\mathrm{H}})$ are the real and imaginary parts of $\bar{X}_2$, respectively.

4.3.2 The Estimation Stage

According to Equation (4.14), in contrast to the scenario with two hypotheses in [133], we have a third hypothesis, namely the scenario of voiced speech corresponding to considering a non-zero spectral mean value. As shown in Section 3.4, when formulating optimal estimators under such models, opposed to the zero-mean model, the resulting phase estimators differ from the noisy spectral phase.

An important component of the sub-risks in Equation (4.7) is the weighted cost function $C_b(X, \hat{X}|\mathcal{H}_{b'})$ that consists of weights $\lambda_{b'}^{b}$, penalizing misclassifications dependent on the pair $\{b, b'\}$, multiplied with the actual cost function $\mathcal{E}_b(X, \hat{X}|\mathcal{H}_{b'})$. In this work, estimation of the complex-valued DFT coefficients is of special interest. As a logical consequence, the cost function is formulated on the squared error of the complex-valued DFT coefficient estimates, i.e.

$$\mathcal{E}_b(X, \hat{X}|\mathcal{H}_{b'}) = \begin{cases} (X - \hat{X}_b)^*(X - \hat{X}_b), & b' = 1, 2 \\ (G_{\min}Y - \hat{X}_b)^*(G_{\min}Y - \hat{X}_b), & b' = 0, \end{cases} \tag{4.15}$$

Setting $G_{\min}$ to zero means applying the maximum noise suppression, i.e. $G = 0$, whenever the algorithm decides for speech absence. This might yield audible artifacts like musical noise in practice because of spectral peaks resulting from the possibly highly selective harsh suppression. To avoid these outliers, the spectral gain is often floored by selecting $G_{\min} > 0$ [16].

Inserting Equation (4.15) into Equation (4.11) results in

$$\begin{aligned}\hat{X}_b = \arg\min_{\hat{X}} \quad & p(\mathcal{H}_0)\lambda_0^b \int_{\mathcal{D}_X} (G_{\min}Y - \hat{X})^*(G_{\min}Y - \hat{X})p(Y|X)p(X|\mathcal{H}_0)dX \\ & + p(\mathcal{H}_1)\lambda_1^b \int_{\mathcal{D}_X} (X - \hat{X})^*(X - \hat{X})p(Y|X)p(X|\mathcal{H}_1)dX \\ & + p(\mathcal{H}_2)\lambda_2^b \int_{\mathcal{D}_X} (X - \hat{X})^*(X - \hat{X})p(Y|X)p(X|\mathcal{H}_2)dX.\end{aligned} \tag{4.16}$$

Considering a multiplicative gain for each class such that the estimated complex-valued STFT coefficients are given by

$$\hat{X}_b = G_b Y, \tag{4.17}$$

the criterion in Equation (4.16) results in the following class-dependent gain function[12]

$$G_b = \frac{\lambda_0^b G_{\min} + \Lambda_1(Y)\lambda_1^b G_{\mathrm{W}} + \Lambda_2(Y)\lambda_2^b G_{\mathrm{SDW}}}{\Lambda_1(Y)\lambda_1^b + \Lambda_2(Y)\lambda_2^b + \lambda_0^b}, \tag{4.18}$$

The Wiener filter G_{W} in Equation (2.20) is the optimal gain under hypothesis $\mathcal{H}_1$, the stochastic-deterministic Wiener filter G_{SDW} [117] from Equation (3.25) is optimal under hypothesis $\mathcal{H}_2$, and accordingly, the optimal gain under hypothesis $\mathcal{H}_0$ is $G_{\min}$. These gain functions are combined according to the respective likelihood ratios

$$\Lambda_b(Y) = \frac{p(\mathcal{H}_b)}{p(\mathcal{H}_0)}\frac{p(Y|\mathcal{H}_b)}{p(Y|\mathcal{H}_0)}, \tag{4.19}$$

with

$$\begin{aligned}p(Y|\mathcal{H}_0) = p(Y_{\mathrm{R}}, Y_{\mathrm{I}}|\mathcal{H}_0) &= \frac{1}{\pi\sigma_d^2}\mathrm{e}^{-\frac{Y_{\mathrm{R}}^2+Y_{\mathrm{I}}^2}{\sigma_d^2}}, \\ p(Y|\mathcal{H}_1) &= \frac{1}{\pi(\sigma_d^2+\sigma_x^2)}\mathrm{e}^{-\frac{Y_{\mathrm{R}}^2+Y_{\mathrm{I}}^2}{\sigma_d^2+\sigma_x^2}}, \\ p(Y|\mathcal{H}_2) &= \frac{1}{\pi(\sigma_d^2+\sigma_x^2)}\mathrm{e}^{-\frac{(Y_{\mathrm{R}}-\bar{X}_{2,\mathrm{R}})^2+(Y_{\mathrm{I}}-\bar{X}_{2,\mathrm{I}})^2}{\sigma_d^2+\sigma_x^2}}.\end{aligned} \tag{4.20}$$

For the case where $\angle Y \neq \angle X_{\mathrm{H}}$, Equation (3.25) and hence also Equation (4.18) result in a complex-valued gain, indicating that $\angle\hat{X}_b \neq \angle Y$. This effect results from the non-circular prior distribution of the speech STFT coefficients, as they have a non-zero mean in the presence of harmonic signal components.

[12] See Appendix E.1 for the detailed derivation.

The class-dependent estimate $\hat{X}_b$ is parameterized by the weighting factors $\lambda_{b'}^b$. Setting them all equal renders the detection operation to be obsolete, the estimators are simply mixed according to their respective *a posteriori* probabilities

$$p(\mathcal{H}_b|Y) = \frac{\Lambda_b(Y)}{1 + \Lambda_1(Y) + \Lambda_2(Y)} \tag{4.21}$$

which is the same weighting as in the soft-decision estimator derived in [56]. In the case of two classes only, i.e. speech presence/absence, this reduces to the well known SPU considerations in [13]. In [86], SPU was incorporated in the estimation of the spectral amplitude under a stochastic-deterministic signal model. For unequal $\lambda_{b'}^b$s, we need to choose from a set of alternative values for $\hat{X}_b$ based on the detection criterion in Equation (4.10). The detection-estimation procedure gives us more flexibility in terms of parameterizing the estimator to not only fulfill estimation theoretic optimization criteria but also take into account perceptual requirements.

4.3.3 The Detection Stage

The detection problem as formulated in Equation (4.10) requires us to solve Equation (4.7) for the sub-risks $r_b(Y|\mathcal{H}_{b'})$ for all $b' \in \{0, 1, 2\}$. The particular outcomes are given by

$$r_b(Y|\mathcal{H}_0) = \lambda_0^b p(Y|\mathcal{H}_0)|G_{\min} - G_b|^2|Y|^2, \tag{4.22}$$

$$r_b(Y|\mathcal{H}_1) = \lambda_1^b p(Y|\mathcal{H}_1)(G_{\mathrm{W}}\sigma_d^2 + |G_{\mathrm{W}} - G_{\mathrm{b}}|^2|Y|^2), \tag{4.23}$$

and

$$r_b(Y|\mathcal{H}_2) = \lambda_2^b p(Y|\mathcal{H}_2)(G_{\mathrm{W}}\sigma_d^2 + |G_{\mathrm{SDW}} - G_b|^2|Y|^2), \tag{4.24}$$

The corresponding derivations are given in Appendix E.2. Given the above equations, Equation (4.10) can be used to evaluate the final decision, i.e. which estimator to choose.

Given estimates of the speech and noise PSDs together with knowledge about the harmonic components, one could apply the above estimation and detection rules right away in any STFT-based speech enhancement framework. A variety of approaches to estimating all three quantities exist, however, in the following, we will have a closer look on retrieving the harmonic components from the noisy observations.

4.4 Estimation of the Harmonic Signal Components

The detection-estimation procedure in Section 4.3 requires us to know $X_{\mathrm{H}}(k,\ell)$ before estimating the speech STFT coefficient itself. Hence, we need to retrieve an estimate of the harmonic components directly from the signal mixture in Equation (4.1). A vast amount of approaches to solve this task exists, many of them aim to estimate the harmonic parameters magnitude, phase, and frequency followed by mapping them to the desired domain, e.g. the time domain. From Equation (3.12) it is obvious that the individual sinusoidal components interact with each other due to the window function. This means that we cannot retrieve the parameters of harmonic h without knowing *all* other harmonics' parameters. There exist several approaches to overcome this problem, e.g. by using an iterative procedure like the expectation-maximization algorithm [1, 122]. However, starting from the assumption that a fundamental frequency estimate is available from some dedicated estimator, we take a different path. The proposed solution intrinsically also solves the problem of interacting sinusoids for the given signal model, although derived from a different perspective.

4.4.1 A Pitch-Synchronous Signal Representation

Considering additive noise and taking the DTFT of the noisy input signal gives us

$$Y(\mathrm{e}^{j\omega},\ell) = X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell) + X_{\mathrm{S}}(\mathrm{e}^{j\omega},\ell) + D(\mathrm{e}^{j\omega},\ell). \tag{4.25}$$

Given the interpretation of the DFT as a sampled version of the DTFT, one may ask if there exists a set of sampling points that facilitates the estimation of $X_{\mathrm{H}}(k,\ell)$. Considering the observed data $Y(k,\ell)$, the stochastic speech components and the additive noise may mask the harmonic spectrum $X_{\mathrm{H}}(k,\ell)$. We capture the amount of masking by a measure similar to the SNR, termed the harmonic to stochastic-plus-noise ratio (HSNR)

$$\mathrm{HSNR}(k,\ell) = \frac{\mathbb{E}\big(\,|X_{\mathrm{H}}(k,\ell)|^2\,\big)}{\mathbb{E}\big(\,|X_{\mathrm{S}}(k,\ell)|^2 + |D(k,\ell)|^2\,\big)} = \frac{|X_{\mathrm{H}}(k,\ell)|^2}{\sigma_x^2(k,\ell) + \sigma_d^2(k,\ell)}. \tag{4.26}$$

The higher the $\mathrm{HSNR}(k,\ell)$, the more reliably we can estimate $X_{\mathrm{H}}(k,\ell)$ from $Y(k,\ell)$. Under the assumption that the noise and speech variance are sufficiently flat across frequency, maximizing $|X_{\mathrm{H}}(k,\ell)|^2$ is equivalent to maximizing $\mathrm{HSNR}(k,\ell)$.

Following the definition of the DFT in Equation (2.3), the spectral sampling points are determined by the choice of N_{DFT}. Therefore, if we stick to the DFT, maximizing $\mathrm{HSNR}(k,\ell)$ w.r.t. the spectral sampling points is equivalent to finding

$$N_{\mathrm{DFT,opt}} = \underset{N_{\mathrm{DFT}}}{\arg\max}\, |X_{\mathrm{H}}(k,\ell)|^2\,. \tag{4.27}$$

Neglecting the negative frequency components in Equation (3.10) and setting $N = N_{\mathrm{DFT}}$ we have

$$X_{\mathrm{H}}(k,\ell) \approx \sum_{h=1}^{N_{\mathrm{H}}} \frac{A_h(\ell)}{2} \mathrm{e}^{j\phi_h(\ell)} W\left(\mathrm{e}^{j(\frac{2\pi k}{N_{\mathrm{DFT}}}-\omega_h(\ell))}\right) \mathrm{e}^{-j(\frac{2\pi k}{N_{\mathrm{DFT}}}-\omega_h(\ell))\frac{N-1}{2}} \tag{4.28}$$

we can use Jensen's inequality for convex functions of sums to define an upper bound for $|X_{\mathrm{H}}(k,\ell)|^2$

$$\begin{aligned} |X_{\mathrm{H}}(k,\ell)|^2 &\approx \left| \sum_{h=1}^{N_{\mathrm{H}}} \frac{A_h(\ell)}{2} \mathrm{e}^{j\phi_h(\ell)} W\left(\mathrm{e}^{j(\frac{2\pi k}{N_{\mathrm{DFT}}}-\omega_h(\ell))}\right) \mathrm{e}^{-j(\frac{2\pi k}{N_{\mathrm{DFT}}}-\omega_h(\ell))\frac{N-1}{2}} \right|^2 \\ &\leq \sum_{h=1}^{N_{\mathrm{H}}} \left| \frac{A_h(\ell)}{2} \right|^2 \left| W\left(\mathrm{e}^{j(\frac{2\pi k}{N_{\mathrm{DFT}}}-\omega_h(\ell))}\right) \right|^2 . \end{aligned} \tag{4.29}$$

Hence, maximizing the above upper bound w.r.t. N_{DFT} means maximizing the DTFT of the window function. In this work, we consider window functions that belong to the Blackman-Harris family [137]. These windows, when modulated with a complex-valued sinusoid, have their maximum at the DFT frequency that is closest to the modulator's frequency, i.e.

$$\arg\max_{N_{\mathrm{DFT}}} |X_{\mathrm{H}}(k,\ell)|^2 = \arg\min_{N_{\mathrm{DFT}}} \left| \frac{2\pi k}{N_{\mathrm{DFT}}} - \omega_h(\ell) \right|^2 , \tag{4.30}$$

Rendering N_{DFT} a function of frame index ℓ

$$N_{\mathrm{DFT}}(\ell) = \frac{2\pi k}{\omega_h(\ell)} = \frac{k f_s}{h f_0(\ell)} = K \frac{f_s}{f_0(\ell)} . \tag{4.31}$$

By selecting a fixed ratio $K = k/h \in \mathbb{Z}$, we ensure that the maxima of $\left|X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell)\right|^2$ are sampled by every K^{th} DFT bin. Further, we have to ensure that $N_{\mathrm{DFT}}(\ell) \in \mathbb{Z}$ by rounding to the next integer and we have

$$N_{\mathrm{DFT}}(\ell) = \mathrm{round}\left(K \frac{f_s}{f_0(\ell)} \right) = \mathrm{round}(K f_s T_0(\ell)) . \tag{4.32}$$

Hence, by choosing the DFT lengths as multiples of the fundamental period we ensure that the harmonic frequencies are a subset of the DFT frequencies and hereby that the maxima of the upper bound as formulated in Equation (4.29) are sampled. By selecting the window length equal to the DFT length, i.e. $N(\ell) = N_{\mathrm{DFT}}(\ell)$, we obtain a pitch-synchronous signal representation which comes with the additional benefit that the inequality in Equation (4.29) is an equality, ensuring that under the given signal model the maxima of $\left|X_{\mathrm{H}}(\mathrm{e}^{j\omega},\ell)\right|^2$ are sampled. This property results from the chosen class of window functions, which ensures that the windows are *DFT-even* [137], i.e. if they are modulated by a sinusoid and the observation length is an integer multiple of the fundamental period T_0 of this sinusoid, the only nonzero samples of their DFT are within the bandwidth of the window's mainlobe around the modulation frequency.[13]

It is important to note that now the window length $N(\ell)$ is a function of the frame index since it depends on the fundamental frequency trajectory. By oversampling the signal before segmentation, we can refine the approximation in Equation (4.32) since the difference between the actual optimal window length and its rounded value decreases with increasing sampling rate. The theoretical result that outside the window's mainlobe around the modulation frequency, the DFT of a sinusoid is zero also renders the approximations made in Equation (3.14) exact. Therefore, under the assumption of (i) perfectly known fundamental frequency and (ii) perfect harmonicity of the signal, DFT bin k is always dominated by *one* harmonic only. Hence, the noisy observation in Equation (4.1) can be represented as

$$Y(k,\ell)=X_{\mathrm{S}}(k,\ell)+D(k,\ell)+\frac{A_{\bar{h}(k)}(\ell)}{2}\mathrm{e}^{j\phi_{\bar{h}(k)}(\ell)}W\left(\mathrm{e}^{j(\frac{2\pi k}{N(\ell)}-\omega_{\bar{h}(k)}(\ell))}\right)\mathrm{e}^{-j(\frac{2\pi k}{N(\ell)}-\omega_{\bar{h}(k)}(\ell))\frac{N(\ell)-1}{2}}, \tag{4.33}$$

where $\bar{h}(k)$ denotes the harmonic index corresponding to the sinusoid which is closest to k. Note that a fundamental property of the resulting representation is that harmonic $\bar{h}(k)$ is always associated to frequency bins occupying the mainlobe width around $k = \bar{h}(k)K+1$. Therefore, depending on the integer factor K, a fixed mapping from frequency bin k to harmonic $\bar{h}(k)$ is possible. This property is fundamentally different to an analysis stage with fixed window lengths.

The Phase Decomposition Stage

The harmonic magnitude and the corresponding initial phase can be considered to be stationary for short time intervals. Later on, we want to make use of this assumption. In order to access the initial harmonic phase, we apply the phase normalization as explained in Section 3.2.2 by multiplying the pitch-synchronous DFT noisy observation with the phase-normalizer $\Psi(k,\ell)$ in Equation (3.15), i.e.

$$\begin{aligned}\widetilde{Y}(k,\ell) &= Y(k,\ell)\Psi^*(k,\ell)\\ &= \widetilde{X}_{\mathrm{H}}(k,\ell)+\widetilde{X}_{\mathrm{S}}(k,\ell)+\widetilde{D}(k,\ell),\end{aligned} \tag{4.34}$$

[13] For frame lengths of 32 ms and a sampling frequency of 16 kHz, this results in a spacing of 31.25 Hz between adjacent DFT bins. In case of a Hamming window, with a mainlobe width of 3 DFT bins [137], this corresponds to a minimum fundamental frequency of 93.75 Hz for Equation (4.29) being an equality, which is a reasonable assumption.

with

$$\widetilde{X}_{\mathrm{H}}(k,\ell) = \frac{A_{\bar{h}(k)}(\ell)}{2} \mathrm{e}^{j\alpha_{\bar{h}(k)}(\ell)} W\left(\mathrm{e}^{j(\frac{2\pi k}{N(\ell)} - \omega_{\bar{h}(k)}(\ell))}\right), \tag{4.35}$$

$$\widetilde{X}_{\mathrm{S}}(k,\ell) = X_{\mathrm{S}}(k,\ell)\Psi^*(k,\ell), \tag{4.36}$$

$$\widetilde{D}(k,\ell) = D(k,\ell)\Psi^*(k,\ell). \tag{4.37}$$

Note that according to Section 3.2.2 the phase of the window function is zero within the mainlobe width around its center frequency (theoretically the only non-zero samples of $\widetilde{X}_{\mathrm{H}}(k,\ell)$ in the pitch-synchronous DFT), i.e. we have

$$\widetilde{X}_{\mathrm{H}}(k,\ell) = \frac{A_{\bar{h}(k)}(\ell)}{2} \mathrm{e}^{j\alpha_{\bar{h}(k)}(\ell)} \left|W\left(\mathrm{e}^{j(\frac{2\pi k}{N(\ell)} - \omega_{\bar{h}(k)}(\ell))}\right)\right|, \tag{4.38}$$

if all model assumptions are fulfilled.

Figure 4.2 illustrates the DFT after phase normalization for the conventional case, i.e. the window length is not an integer multiple of the fundamental period, as well as the case of a pitch-synchronous frame length. The stochastic signal components are set to zero to emphasize on the impact on $X_{\mathrm{H}}(k,\ell)$ only. For the conventional case, the DFT does not always sample the DTFT at its maxima. An amplitude estimate obtained by simple peak-picking results in a biased estimate due to the interaction of the individual harmonics, even if the fundamental frequency is known and the harmonics are well separated. The plots (c) and (d) also indicate that the true sinusoidal phase cannot be obtained in case of the conventional windowing. In the pitch-synchronous case, the interaction of the sinusoids at the harmonic frequencies is minimized, hence, the correct phase value is sampled by the DFT. This is indicated by the common intersection of the DTFT of the fundamental, the DTFT of the sum of all three sinusoids, and the DFT sampling point.

The phase normalization in Equation (4.36) and Equation (4.37) only rotates the DFT coefficients in the complex plane, hence does not affect the respective distributions since they are circular, i.e. their moments are invariant with respect to a rotation in the complex plane [65]. Thus, we have $\widetilde{X}_{\mathrm{S}}(k,\ell) \sim \mathcal{CN}(0,\sigma_x^2(k,\ell))$ and $\widetilde{D}(k,\ell) \sim \mathcal{CN}(0,\sigma_d^2(k,\ell))$. In contrast, given $\widetilde{X}_{\mathrm{H}}(k,\ell) \neq 0$, $\widetilde{Y}(k,\ell)$ follows in general a non-centered and hence non-circular [138] distribution, i.e. $\widetilde{Y}(k,\ell) \sim \mathcal{CN}(\widetilde{X}_{\mathrm{H}}(k,\ell), \sigma_d^2(k,\ell) + \sigma_x^2(k,\ell))$ [40, 86].

In the following, we will illustrate how to exploit the phase-normalization in conjunction with the properties of the pitch-synchronous segmentation to obtain estimates of $X_{\mathrm{H}}(k,\ell)$. In contrast to the voiced speech model, the conventionally assumed statistical properties of the stochastic signal components, i.e. unvoiced speech and noise, are not affected by the non-uniform segmentation as long as we keep the segment lengths short enough for stationarity assumptions to hold and long enough for the application of the CLT to be reasonable.

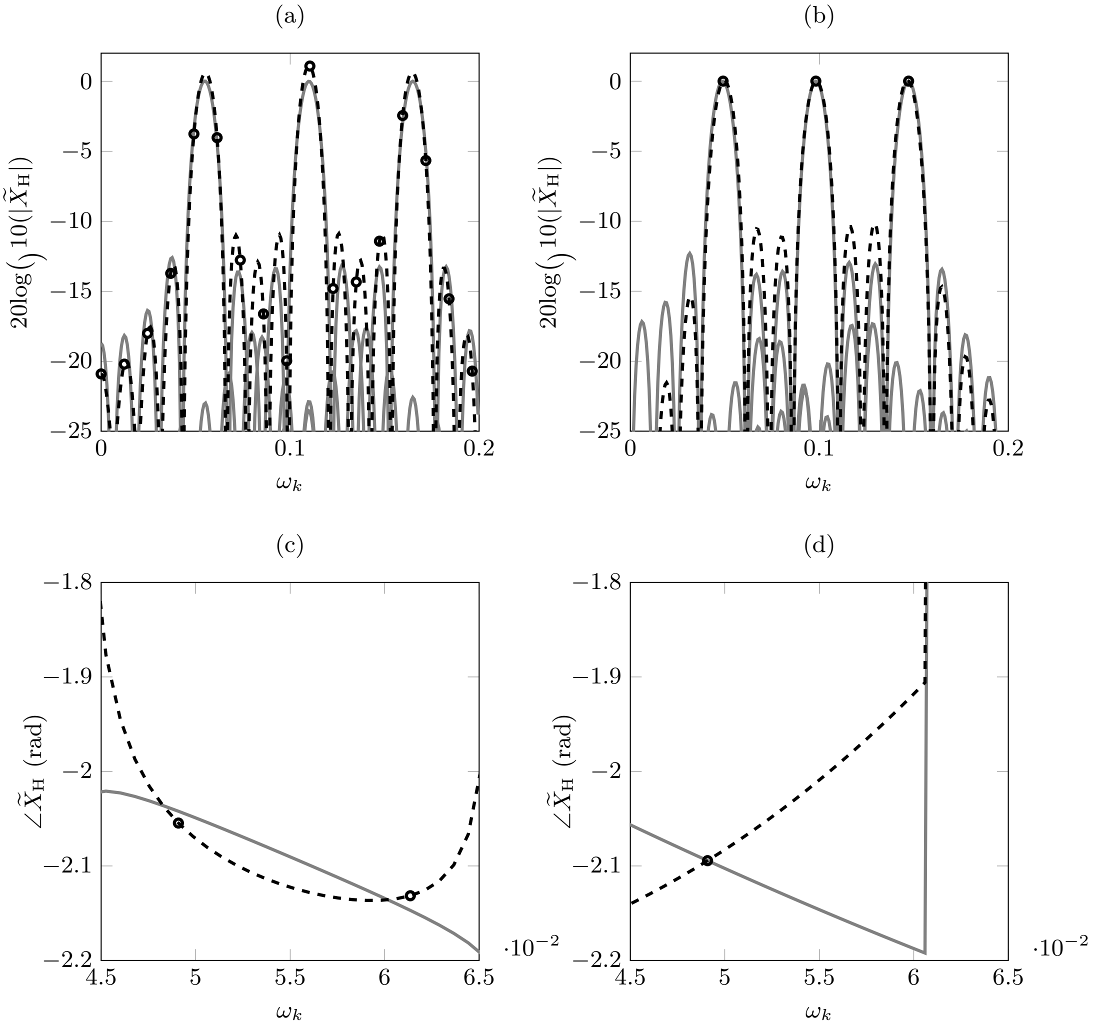

Figure 4.2: Impact of the analysis window length on the DFT representation of three harmonically related sinusoids. The left panels (a) and (c) illustrate the case where $N(\ell) \neq K\frac{f_s}{f_0(\ell)}$, hence, the DTFT is not sampled at its maxima and the sinusoids leak in all frequency bands. Panels (b) and (d) refer to a pitch-synchronous segmentation with $K = 4$ where the leakage problem at the DTFT sampling points is reduced. The gray lines in (a) and (b) represent the individual harmonics, the dashed black line is the summation of the three harmonics, and the black circles are the DFT sample points. In (c) and (d), the gray line is the original phase of the first harmonic.

4.4.2 The Prior Distributions

Since each harmonic $\bar{h}(k)$ in the pitch-synchronous analysis framework is assigned to a fixed set of DFT bins $\mathcal{K}_{\bar{h}(k)}$, not every frequency bin is equally likely to contain (voiced) speech components. The number of DFT bins $|\mathcal{K}_{\bar{h}(k)}|$ in the set $\mathcal{K}_{\bar{h}(k)}$ is determined by the frequency estimation accuracy and the window choice. However, we neglect the frequency estimation accuracy's impact in the following, assuming the fundamental frequency to be perfectly known. As a result, each harmonic theoretically only affects frequency bins within the mainlobe width around the respective DFT bin, which is a generic parameter of the chosen window function. Further, it is reasonable to assume that higher order harmonics contain less energy than lower order harmonics. This was taken into account in related previous works such as [28,33,86] by introducing a maximum voicing frequency, e.g. 4 kHz. For frequencies exceeding this threshold, no harmonic structure was expected. This step was needed to prevent the respective algorithms from fitting harmonic structure to high frequency broad-band noise, resulting in a buzzy sound-quality. In this work, rather than setting a fixed threshold, we incorporate the knowledge about the specific structure of the pitch-synchronous transform into the speech estimation task by means of prior distributions for the harmonic signal components.

In the following, as one possibility to modeling the prior information, we will deploy an empirical Bayesian approach for obtaining and incorporating the prior distributions. In order to follow the nomenclature in Section 4.3, we consider two classes of speech sounds, unvoiced speech, indicated by $\mathcal{H}_1$ and voiced speech, indicated by $\mathcal{H}_2$. Further, due to the statistical assumptions made in the following, we will formulate the priors of the harmonic components in polar coordinates instead of Cartesian ones.

It is reasonable to speculate that without further information, the initial harmonic phase $\alpha_{\mathrm{H}}(k,\ell) = \angle \widetilde{X}_{\mathrm{H}}(k,\ell)$ is arbitrary, hence, uniformly distributed. However, this is not true for the magnitude $A_{\mathrm{H}}(k,\ell) = \left|\widetilde{X}_{\mathrm{H}}(k,\ell)\right|$. We therefore assume $A_{\mathrm{H}}(k,\ell)$ to come from a delta distribution around zero in unvoiced time frequency regions while we expect heavy tailed pdfs[14] in voiced parts, putting more weight on higher magnitudes where we expect the harmonic signal components to contribute to the overall signal energy. We thus select the Gamma distribution [55] to model the amplitude pdf for voiced speech, i.e.

$$p(A_{\mathrm{H}}(k,\ell);\mathcal{H}_b) = \begin{cases} \delta(A_{\mathrm{H}}(k,\ell)), & b=1, \\ \frac{\mu(k)^{\nu(k)}}{\Gamma(\nu(k))} A_{\mathrm{H}}(k,\ell)^{\nu(k)-1} \mathrm{e}^{-\mu(k) A_{\mathrm{H}}(k,\ell)}, & b=2, \end{cases} \tag{4.39}$$

where $\nu(k)$ and $\mu(k)$ denote the frequency dependent shape and inverse scaling parameter, respectively. The joint prior, under the assumption of independent amplitude and

[14] See Appendix D, Equation (D.10).

phase [55], is hence given by the multiplication of the individual pdfs

$$p(A_{\mathrm{H}}(k,\ell),\phi(k,\ell);\mathcal{H}_b)=\begin{cases}\frac{1}{2\pi}\delta(A_{\mathrm{H}}(k,\ell)), & b=1,\\ \frac{\mu(k)^{\nu(k)}}{2\pi\Gamma(\nu(k))}A_{\mathrm{H}}(k,\ell)^{\nu(k)-1}\mathrm{e}^{-\mu(k)A_{\mathrm{H}}(k,\ell)}, & b=2.\end{cases} \tag{4.40}$$

Following the empirical Bayes methodology, the parameters $\nu(k)$ and $\mu(k)$ are obtained by fitting the distributions in Equation (4.40) to empirical distributions obtained from real speech data, explained in the following.

4.4.3 Obtaining the Prior Parameters

For empirically evaluating the frequency dependent parameters of the Gamma distribution we used the TIMIT [139] core database train set containing 72 utterances together with the annotation from [140]. The annotation includes voicing states as well as a fundamental frequency trajectory. For each frequency bin k we estimate $\mu(k)$ and $\nu(k)$. As a consequence of the pitch dependent frame length $N(\ell)$, it is important to normalize the amplitude values by dividing $A(k,\ell)$ by $N(\ell)$ in order to achieve compatible values along time.

It is important to note that although we are dealing with clean speech data and an annotation of the voicing states is available, we cannot access the purely harmonic content directly, since we still observe the mixture

$$\widetilde{X}(k,\ell)=\widetilde{X}_{\mathrm{H}}(k,\ell)+\widetilde{X}_{\mathrm{S}}(k,\ell), \tag{4.41}$$

as by construction we let the stochastic components model all intrinsic model misfits, explained in Section 3.2.1. Since $\widetilde{X}_{\mathrm{S}}(k,\ell)$ is modeled as zero-mean, we obtain an approximation of $\widetilde{X}_{\mathrm{H}}(k,\ell)$ by averaging out $\widetilde{X}_{\mathrm{S}}(k,\ell)$, i.e. for each frequency bin k we take the mean of all frames whose center time fits inside the range $[t(\ell)-20\,\mathrm{ms},t(\ell)]$, where the center time of frame ℓ is given by

$$\begin{aligned}t(\ell)&=\frac{N(\ell)/2+\sum_{\ell'=0}^{\ell-1}\mathrm{round}(\mathrm{HF}\cdot N(\ell'))}{f_s}\\&=t(\ell-1)+\frac{1}{f_s}\left(\mathrm{round}(\mathrm{HF}\cdot N(\ell-1))+\frac{N(\ell)-N(\ell-1)/2}{2}\right),\end{aligned} \tag{4.42}$$

with hop factor (HF) defined as $L/N(\ell)$. Hence, we have

$$A_{\mathrm{H}}(k,\ell)\approx\left|\frac{1}{N_{\mathrm{o}}}\sum_{u=0}^{N_{\mathrm{o}}-1}\widetilde{X}(k,\ell-u)\right|. \tag{4.43}$$

The number of observations considered for the averaging is termed N_{o}, which corresponds to the cardinality of the set of frames that are within the assumed stationarity interval of 20 ms. Following these considerations, a set of observations is obtained which can be used to fit the parameters of the Gamma approximation.

Although we have variable frame lengths, dependent on $f_0(\ell)$ and K, we still want to ensure that they approximately correspond to 30 ms, i.e. resemble the stationarity interval of speech. To this end, we limit the fundamental frequency estimates to the interval $[90, 350]$ Hz and choose $K = 6$, which yields frame lengths around the desired value for a sampling frequency of $f_s = 16$ kHz. We set the overlap of the sliding windows to 87.5% of the corresponding frame length.

To ensure independence of the fitted distributions from overall signal level, it is important to also take into account the variations in signal level. We therefore evaluate the empirical distribution of the spectral magnitude, divided by the constant

$$C(\ell) = \frac{1}{N(\ell)} \sum_{k=1}^{N(\ell)/2} |X(k,\ell)|\,, \tag{4.44}$$

which is proportional to the signal level.

Although not considered to be Gamma distributed in Equation (4.40), we also conducted the same analysis for unvoiced speech in order to compare the resulting Gamma distribution parameters for both classes. For unvoiced speech, we expect the resulting distributions to put more emphasis on magnitudes close to zero. The outcome of this procedure is illustrated in Figure 4.3 for both parameters and both classes. The peaky character of the inverse scaling parameter $\mu(k)$ in case of voiced speech is due to harmonics. Whereas $\mu(k)$ tends to be larger for DFT bins dominated by harmonics, it takes lower values between harmonics and in unvoiced speech, i.e. putting more weight on smaller amplitudes as illustrated in Figure 4.3 (b).

By combining the resulting prior distributions with the observed noisy data by means of the MAP principle, the prior knowledge can be incorporated into the estimation of the harmonic signal components, or, to stick with the nomenclature we used in Section 4.3.1, the spectral mean value.

4.4.4 MAP Estimation of the Spectral Mean

Merging the aforementioned models and using Bayes' theorem to formulate the posterior of the spectral mean, we obtain

$$p(A_{\mathrm{H}}(k,\ell), \alpha_{\mathrm{H}}(k,\ell)|\tilde{\mathbf{y}}(k,\ell); \mathcal{H}_b) = \frac{p(\tilde{\mathbf{y}}(k,\ell)|A_{\mathrm{H}}(k,\ell), \phi_{\mathrm{H}}(k,\ell))p(A_{\mathrm{H}}(k,\ell), \phi_{\mathrm{H}}(k,\ell); \mathcal{H}_b)}{p\left(\tilde{\mathbf{y}}(k,\ell); \mathcal{H}_b\right)} \tag{4.45}$$

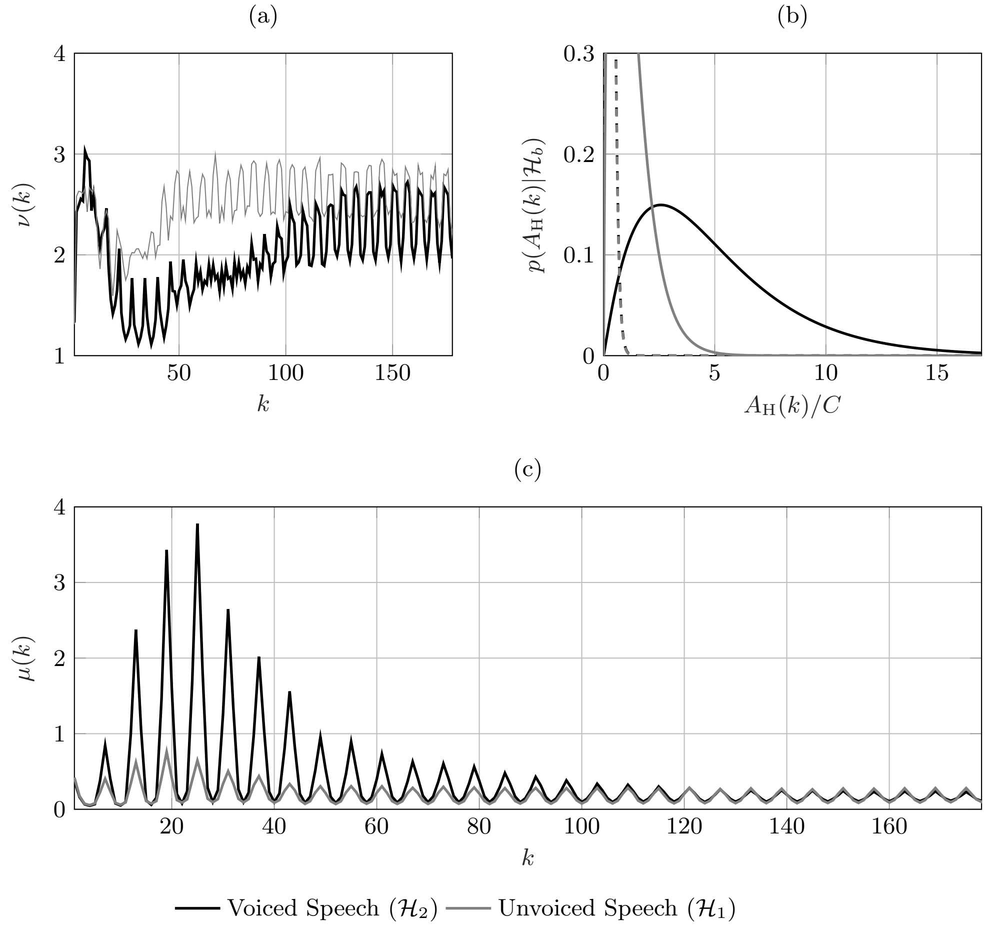

Figure 4.3: (a) and (c): The dependency of the prior distribution parameters on frequency bin k. (b): Inserting the parameters from (a) and (c) for frequency bin $k = 7$ (solid lines, corresponding to harmonic 1) and $k = 5$ (dashed lines) into Equation (4.39) illustrates the dependency of the priors' shape on the DFT bin considered.

with N_{o} observations stacked into the vector

$$\tilde{\mathbf{y}}(k,\ell) = \begin{bmatrix}\widetilde{Y}(k,\ell) & \widetilde{Y}(k,\ell-1) & \ldots & \widetilde{Y}(k,\ell-N_{\mathrm{o}}+1)\end{bmatrix}^T .$$

Considering the observations to be independent and assuming the mean to be stationary within the observed time span, from the statistical model in Section 4.3.1 we have the likelihood

$$p\left(\tilde{\mathbf{y}}(k,\ell)|A_{\mathrm{H}}(k,\ell),\alpha_{\mathrm{H}}(k,\ell)\right) = \prod_{u=0}^{N_{\mathrm{o}}-1} \frac{\mathrm{e}^{-\frac{|\widetilde{Y}(k,\ell-u)-A_{\mathrm{H}}(k,\ell)\mathrm{e}^{j\alpha_{\mathrm{H}}(k,\ell)}|^2}{\sigma_d^2(k,\ell-u)+\sigma_x^2(k,\ell-u)}}}{\pi(\sigma_d^2(k,\ell-u)+\sigma_x^2(k,\ell-u))}. \tag{4.46}$$

Note that the length of the averaging filter in Equation (4.43) should be consistent with the number of observations N_o in Equation (4.46). For the sake of readability we drop the frequency index in the following derivations, however, the dependency of *all* parameters on k maintains.

Finding the MAP estimates of $A_\mathrm{H}(\ell)$ and $\alpha_\mathrm{H}(\ell)$ means maximizing Equation (4.45), or alternatively, the $\log(\,\cdot\,)$ of Equation (4.45) :

$$\begin{aligned}\hat{A}_{b,\mathrm{MAP}}(\ell) &= \underset{A_\mathrm{H}(\ell)}{\arg\max} \log\big(p\left(A_\mathrm{H}(\ell), \alpha_\mathrm{H}(\ell)|\tilde{\mathbf{y}}(\ell); \mathcal{H}_b\right)\big),\\ \hat{\alpha}_{b,\mathrm{MAP}}(\ell) &= \underset{\alpha_\mathrm{H}(\ell)}{\arg\max} \log\big(p\left(A_\mathrm{H}(\ell), \alpha_\mathrm{H}(\ell)|\tilde{\mathbf{y}}(\ell); \mathcal{H}_b\right)\big).\end{aligned} \tag{4.47}$$

For unvoiced speech and frequency bins between harmonics, the spectral mean estimate is equal to zero since we assume the amplitude to follow a delta distribution in this case.

For voiced speech, the MAP phase estimate is:

$$\hat{\alpha}_{2,\mathrm{MAP}}(\ell) = \mathrm{arctan2}\left(\sum_{u=0}^{N_\mathrm{o}-1} \frac{\left|\widetilde{Y}(\ell-u)\right| \sin\left(\angle\widetilde{Y}(\ell-u)\right)}{\sigma_d^2(\ell-u)+\sigma_x^2(\ell-u)}, \sum_{u=0}^{N_\mathrm{o}-1} \frac{\left|\widetilde{Y}(\ell-u)\right| \cos\left(\angle\widetilde{Y}(\ell-u)\right)}{\sigma_d^2(\ell-u)+\sigma_x^2(\ell-u)}\right). \tag{4.48}$$

For $N_\mathrm{o} = 1$, this estimate degenerates to the noisy phase, which is expected from the discussion in Section 2.3, as the prior distribution for the harmonic phase is uniform and for the case $N_\mathrm{o} = 1$ we only take one observation for the estimation of the phase into account. The corresponding MAP amplitude point estimate is given by

$$\hat{A}_{2,\mathrm{MAP}}(\ell) = v(\ell) + \sqrt{v^2(\ell) + (\sigma_d^2(\ell) + \sigma_x^2(\ell))\frac{\nu(k)-1}{2N_\mathrm{o}}}, \tag{4.49}$$

with

$$v(\ell) = \frac{\sigma_{N_\mathrm{o}}^2(\ell)}{2}\left(\sum_{u=0}^{N_\mathrm{o}-1} \frac{\left|\widetilde{Y}(\ell-u)\right| \cos\left(\angle\widetilde{Y}(\ell-u) - \alpha_\mathrm{H}(\ell)\right)}{\sigma_d^2(\ell-u)+\sigma_x^2(\ell-u)} - \frac{\mu(k)}{2C(\ell)}\right) \tag{4.50}$$

and

$$\frac{1}{\sigma_{N_\mathrm{o}}^2(\ell)} = \sum_{u=0}^{N_\mathrm{o}-1} \frac{1}{\sigma_d^2(\ell-u)+\sigma_x^2(\ell-u)}. \tag{4.51}$$

The detailed derivations of Equation (4.48) and Equation (4.49) can be found in Appendix E.3. Note that $\alpha_\mathrm{H}(\ell)$ in Equation (4.50) is not directly accessible. However, since it can be estimated from the noisy observation directly, it is replaced by $\hat{\alpha}_{2,\mathrm{MAP}}$.

Further, whereas $\widetilde{Y}(\ell)$ is the observation and N_{o} is a fixed parameter, all other parameters of this estimator are in principle unknown. For obtaining estimates of $\sigma_d^2(\ell)$ and $\sigma_x^2(\ell)$ we make use of standard approaches from Chapter 2, adapted to our framework, as explained in Section 4.5.1 and Section 4.7. The estimation of $C(\ell)$ will be discussed in Section 4.5.2.

Now, given an estimate of the spectral magnitude and the spectral phase, by re-wrapping the phase again, we can construct a complex-valued STFT, representing the spectral mean value that is needed for the gain function in Equation (4.18)

$$\hat{X}_{\mathrm{H}}(\ell) = \hat{A}_{2,\mathrm{MAP}}(\ell)\mathrm{e}^{j\hat{\alpha}_{2,\mathrm{MAP}}(\ell)}\Psi(\ell). \tag{4.52}$$

4.5 The Pitch-Synchronous Simultaneous Detection-Estimation Framework

Once having obtained the spectral mean estimate according to Equation (4.52) it can be plugged into any estimation scheme relying on a harmonic model for voiced speech, e.g. [56, 86, 117, 132]. Here, we will use the estimation scheme that has been derived in Section 4.3, resulting in the overall framework pictured in Figure 4.1.

Until now, we considered the variances of the DFT coefficients and the normalization constant $C(\ell)$ to be given. In practice they are not. For noise PSD estimation, we use the estimator from [18] throughout this thesis, and the estimation of the other parameters is explained in the following.

4.5.1 Speech Variance Estimation

For the estimation of the speech variance we seek an approach similar to the decision-directed method [15]. As the signal model is different we need to adopt it to our proposed framework. We can express the variance by its definition and by means of the noisy observation resulting from the complex Gaussian model, yielding two expressions:

$$\begin{aligned}\sigma_x^2(\ell) &= \mathbb{E}(|X_{\mathrm{S}}(\ell)|^2) = \mathbb{E}(|X(\ell) - X_{\mathrm{H}}(\ell)|^2),\\ \sigma_x^2(\ell) &= \mathbb{E}(|Y(\ell) - X_{\mathrm{H}}(\ell) - D(\ell)|^2) = \mathbb{E}(|Y(\ell) - X_{\mathrm{H}}(\ell)|^2) - \sigma_d^2(\ell).\end{aligned} \tag{4.53}$$

In practice, the noise variance is unknown and approximated by an estimate $\hat{\sigma}_d^2$ obtained by, e.g. [18]. By introducing a smoothing constant $\alpha_{\mathrm{DD}} \in [0,1]$ we can mix approximations of both terms in Equation (4.53) as follows,

$$\begin{aligned}\hat{\sigma}_x^2(\ell) =&\alpha_{\mathrm{DD}}|\hat{X}_{b_{\mathrm{opt}}}(\ell-1) - \bar{X}_2(\ell-1)|^2\\ &+ (1-\alpha_{\mathrm{DD}})\max\big[|Y(\ell) - \bar{X}_2(\ell)|^2 - \hat{\sigma}_d^2(\ell), 0\big],\end{aligned} \tag{4.54}$$

The estimate of the speech variance in Equation (4.54) differs from the modified decision-directed approach in [86] in subtracting the noise variance estimate $\hat{\sigma}_d^2(\ell)$ in the second term.

4.5.2 Estimation of the Normalization Constant

The normalization constant $C(\ell)$ is needed to obtain the harmonic amplitude estimates at frame ℓ. Since we clearly do not have access to $|X(k,\ell)|$, we assume that $C(\ell)$ changes slowly over time, i.e. $\sum_{k=1}^{N(\ell)/2} \frac{|X(k,\ell)|}{N(\ell)} \approx \sum_{k=1}^{N(\ell-1)/2} \frac{|X(k,\ell-1)|}{N(\ell-1)}$. Using estimates of $|X(k,\ell-1)|$ we obtain

$$\hat{C}(\ell) = \sum_{k=1}^{N(\ell-1)/2} \frac{|\hat{X}_{b_{\mathrm{opt}}}(k,\ell-1)|}{N(\ell-1)}. \tag{4.55}$$

If no previous speech estimates are available, we initialize $\hat{C}(1) = 1$.

4.6 Evaluation with Synthetic Data

In order to gain insights into the potential of the proposed method we first evaluate the proposed estimation scheme by conducting experiments with synthetic data. This includes investigating the maximum achievable performance when all model assumptions are fulfilled as well as the possible impact of model misfits. In a second step, we test the overall algorithm corresponding to the block diagram in Figure 4.1 with real speech data and examine its capability in a speech enhancement scenario. This is achieved by instrumental metrics as well as by a subjective listening test.

The core of this chapter is the combination of a pitch-synchronous sliding window DFT with a detection and estimation framework that takes into account the different characteristics of voiced and unvoiced speech. To evaluate the impact of each component on the estimation performance separately, we conducted Monte-Carlo simulations with synthetic harmonic signals.

We generated the synthetic data as follows. We used 10 cosines, harmonically related with a fundamental frequency of $f_0 = 133.0561\,\mathrm{Hz}$. The amplitudes and phases of the harmonics were both picked from uniform distributions such that $A_h \in [0.5, 1.5]$ and $\alpha_h \in [-\pi, \pi)$. The frame lengths were chosen up to 30 ms at a sampling frequency of $f_s = 16\,\mathrm{kHz}$ in order to resemble a typical setup for speech enhancement. To evaluate the impact of the window length we varied the window length from 11 to 481 samples, where $N = 481 = 4f_s/f_0$ for the given fundamental frequency. Since we apply a Hamming window for the analysis in the subsequent experiments with real speech data, we use the same window function for this experiment. As additive noise we chose white noise and mixed it at global SNRs $\in \{-5, 0, 5, 10, 15\}\,\mathrm{dB}$. The stochastic components of the speech signal were simulated by sampling from a zero-mean Gaussian distribution with fixed variance $\sigma_x^2 = 1$. For each experiment, we averaged results over $N_{\mathrm{MC}} = 1000$ Monte-Carlo runs.

4.6.1 Harmonic Signal Component Estimation Accuracy

Since in Section 4.4.1, we argued that the achievable accuracy in the harmonic retrieval problem is a function of the chosen segment length, we want to enquire how estimation errors in the harmonic signal components are affected by the window length and how these errors propagate through the processing chain. To this end we define the error-to-signal ratio (ESR) for the harmonic components

$$\mathrm{ESR} = 10\log_{10}\Big(\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\frac{\sum_{n=1}^{N(\ell)}(x_{\mathrm{t,H}}(n,\ell)-\hat{x}_{\mathrm{t,H}}(n,\ell))^2}{\sum_{n=1}^{N(\ell)}x_{\mathrm{t,H}}(n,\ell)^2}\Big), \tag{4.56}$$

where $\mathcal{L}$ denotes the set of all frames considered and $|\mathcal{L}|$ corresponds to its cardinality. We obtain $\hat{x}_{\mathrm{t,H}}(n,\ell)$ by sampling the DTFT at known harmonic frequencies, indicated by "DTFT($y_{\mathrm{t}}(n,\ell)$) @ f_h" in the results, and using the resulting amplitude and phase to synthesize the signal according to Equation (3.1). This procedure corresponds to the approximate maximum likelihood solution to the harmonic retrieval problem [121] given that the fundamental frequency is known. Further, since we average over multiple observations to obtain the MAP estimate of $X_{\mathrm{H}}(k,\ell)$ in Equation (4.52), we also vary the length of the averaging filter in order to assess its impact on the ESR. Since the fundamental frequency is known, the ESR is influenced only by the additive noise, the possibly inadequately chosen window length, and the smoothing filter length N_{o} (in samples). We also report the ESR results for the case where no noise is present, i.e. the clean signal's DTFT is sampled, indicated by "DTFT($x_{\mathrm{t,H}}(n,\ell)$) @ f_h". Figure 4.4 illustrates the results for this simulation. The pitch-synchronous framing helps to improve the ESR, especially if additional smoothing across observations is employed. Even if the clean signal's DTFT is sampled, the ESR can further be improved by choosing a pitch-synchronous setup.

4.6.2 Impact of the Harmonic plus Noise Model

In order to evaluate the estimation performance resulting from different model assumptions, we use the normalized mean-square error (NMSE) of the estimated time domain signals. We define the NMSE as follows

$$\mathrm{NMSE} = 10\log_{10}\Big(\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\frac{\sum_{n=1}^{N(\ell)}(x_{\mathrm{t}}(n,\ell)-\hat{x}_{\mathrm{t}}(n,\ell))^2}{\sum_{n=1}^{N(\ell)}x_{\mathrm{t}}(n,\ell)^2}\Big). \tag{4.57}$$

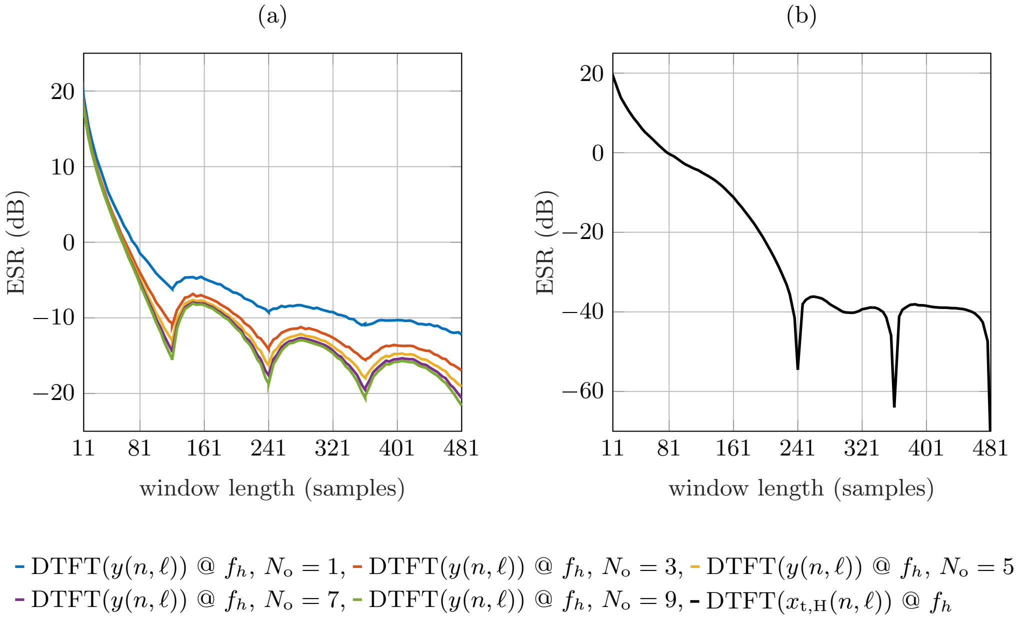

Figure 4.4: ESR dependency on the window length averaged over global SNRs ranging from −5 to 15 dB. (b): illustrates the dependency of the ESR on the window length for the case where the clean signal's DTFT is sampled.

We report the results in Figure 4.5. We note that if harmonic components exist in the signal, the stochastic-deterministic Wiener filter outperforms the purely stochastic Wiener filter for a wide range of ESRs.

4.7 Evaluation

In the previous section we analyzed the first two stages of our algorithm, the signal analysis together with the harmonic retrieval for synthetic data. In the next step we seek to assess the whole framework in terms of its speech enhancement performance.

4.7.1 Benchmark Algorithms

As benchmark algorithms we chose the Wiener filter, the MMSE LSA estimator [90] and the stochastic deterministic MMSE-STSA (SD MMSE-STSA) algorithm from [86], which employs the estimator presented in Section 3.4.[15] In order to additionally assess the impact of the harmonic retrieval block, we combine the mean estimates obtained

[15] I would like to thank Matthew McCallum for sharing his implementation of [86].

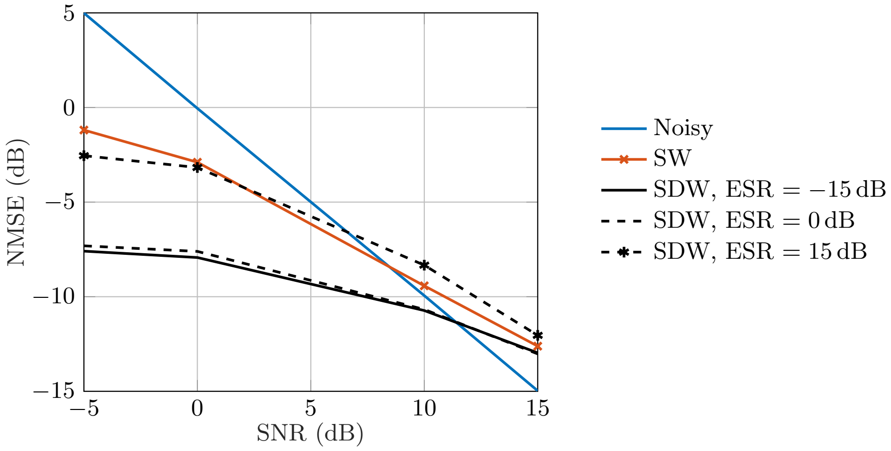

Figure 4.5: An insight into the NMSE performance for the stochastic Wiener filter (G_{W}) compared to the stochastic-deterministic Wiener filter (G_{SDW}) for different accuracies, i.e. values of ESR, of the spectral mean estimate.

from our proposed scheme together with the SD MMSE-STSA algorithm. Since taking into account SPU considerations degraded the performance of the LSA and the Wiener filter in our experiments, we did not include it for these benchmarks. The SD MMSE-STSA algorithm takes into account SPU and the a priori speech absence probability was set to 0.65. Further, the maximum occurring harmonic frequency in the SD MMSE-STSA algorithm was restricted to be below 3 kHz, as recommended in [86]. The proposed algorithm does not need such a restriction, since amplitudes of higher order harmonics are already penalized by the frequency bin dependent Gamma-distribution parameters.

In all derivations, we relied on a given fundamental frequency. This prerequisite cannot be met in practice since we in general do not have access to a ground truth regarding the fundamental frequency. We therefore additionally report results for our proposal employing the annotation of the fundamental frequency from [140] as fundamental frequency estimate. This scenario is a performance upper bound w.r.t. the algorithm's sensitivity on the fundamental frequency estimation accuracy. We use this annotation as a ground truth for the f_0-estimation, hence termed as "oracle f_0" in Figure 4.8 and Figure 4.7.

In the blind scenario, we chose the PEFAC [141] algorithm for estimating the fundamental frequency. The noise variance estimator used in all experiments was the MMSE-based noise estimator from [18]. For the pitch-synchronous methods we chose frame lengths corresponding to $K = 6$ times the fundamental period, consistent with the preceding experiments. The overlap was set to 87.5% of the corresponding frame lengths.

In the detection-estimation step, we set the prior probabilities as follows. The prior probability of voiced speech is set to zero in those frequency regions where we do not expect any influence of harmonics, i.e. all frequency bins outside the mainlobe of the window function around frequency bins $k = hK + 1$, elsewhere we set $p(\mathcal{H}_2) = \frac{1}{6}$ and equally distribute the remaining share on $\mathcal{H}_0$ and $\mathcal{H}_1$,

$$p(\mathcal{H}_1) = p(\mathcal{H}_0) = \frac{1 - p(\mathcal{H}_2)}{2}. \tag{4.58}$$

The overlap-add synthesis in case of the pitch-synchronous framework needs an additional division by the summation of the overlapping synthesis windows to compensate for the varying overlaps. In our experiments we achieved a reconstruction error in the same order as the reconstruction error of a regular overlap-add synthesis using the same window function.

Since we are interested in the impact of each individual step, we report results for the following two variants of our algorithm; i) the stochastic-deterministic Wiener filter applied on the proposed sliding window DFT with pitch-synchronous frame lengths (PS SD Wiener filter), and ii) the full proposal as the outcome of eventually applying the detection and estimation framework (PSDE).

The recommendation in [86] for the SD MMSE-STSA is a Hamming window function, which we consequently selected. For the pitch-synchronous methods we also used the Hamming window function for analysis and synthesis. For the speech variance estimation we employed the adapted DD estimator in Equation (4.54) for the methods which additionally consider the harmonic model. The smoothing constant was set to $\alpha_{\mathrm{DD}} = 0.98$. This standard choice is known to perform well with the benchmark algorithms and also turned out to yield reasonable performance for the proposed method. The mainlobe width of the chosen window function (Hamming) is 3 DFT bins. For the estimation of the spectral mean value we selected the number of observations N_{o} such that it includes frames within the interval $[t_\ell - 20\,\mathrm{ms}, t_\ell]$ (which corresponds roughly to the stationarity interval of speech). The values for $\lambda_{b'}^{b}$ are listed in Table 4.1.

Table 4.1: Assignment of $\lambda_{b'}^{b}$.

b b'	0	1	2
0	1	1.1	1.1
1	0.9	1	0.9
2	0.9	1	1

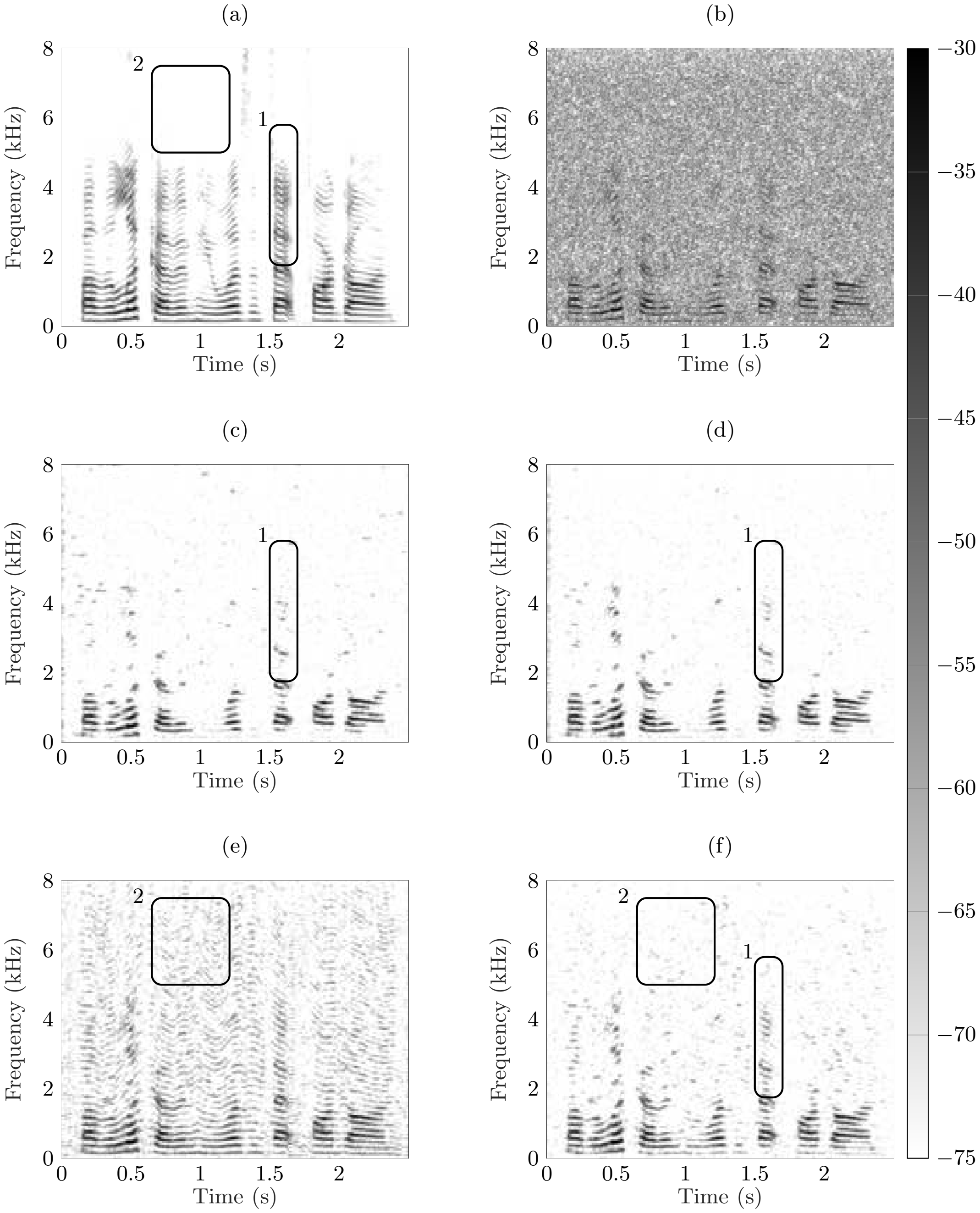

Figure 4.6: Proof-of-concept: Demonstrating the impact of the proposed estimation scheme on a noisy speech signal mixed at 0 dB global SNR with white noise. The clean speech signal is a recording of a female speaker taken from TIMIT saying "By the look of him he wasn't that far gone.". (a) Clean speech signal, (b) noisy speech, (c) conventional Wiener filter, (d) Wiener filter in a pitch-synchronous framework, (e) stochastic-deterministic Wiener filter in the pitch-synchronous domain, (f) applying the simultaneous detection-estimation procedure.

4.7.2 Proof-of-concept

Figure 4.6 depicts spectrograms of a speech signal processed by different algorithms. The conventional Wiener filter results in numerous artifacts such as isolated peaks in the spectrogram representing musical noise, the Wiener filter applied in the pitch-synchronous setup reduces these effects drastically. We attribute this effect to the fact that the DD *a priori* SNR estimator benefits from the property that harmonics and DFT share a fixed mapping, independent of the frame index.[16] However, when compared to the clean speech signal, in both cases low energy harmonics are not very well preserved (see rectangle "1"). The stochastic-deterministic Wiener filter on the other hand introduces artificial harmonics (rectangle "2") resulting in a buzzing residual noise. Applying the detection-estimation step helps dealing with these unwanted components while preserving important but low SNR time-frequency regions. Furthermore, lower frequency harmonics are well retrieved using the proposed method.

4.7.3 Objective Evaluation of Speech Enhancement Performance

Databases and Evaluation Metrics

We used speech utterances from the test set from the TIMIT core database [139] and corrupted the signals with babble, factory, and pink modulated taken from the NOISEX-92 [142] database, as well as a recording of rain drops on a roof, as a noise type with impulsive characteristics, taken from [143]. We mixed the speech signals with noise at global SNRs of -5 to 15 decibels in 5 dB steps, where we follow the SNR definition from [144]. The TIMIT core test set contains 192 utterances and comes with an annotation of the fundamental frequency and the voicing states [140], used as ground truth in our experiments. The signals were sampled at 16 kHz. As evaluation criteria we chose the perceptual evaluation of speech quality (PESQ) [145] and the short-time objective intelligibility (STOI) [146] measure as instrumental predictors of perceived speech quality and speech intelligibility, respectively.[17] We report both PESQ and STOI in terms of Δ improvement over the noisy signal performance.

Results

Figure 4.8 and Figure 4.7 illustrate the results in terms of instrumental metrics. The following observations are made:

- In terms of STOI, in all noise scenarios improvement can be expected for input SNRs up to 10 dB from the proposed methods.

[16] This hypothesis will be examined and explained more thoroughly in Section 5.2.

[17] For an overview on the instrumental evaluation metrics used throughout this thesis we refer to Appendix B.

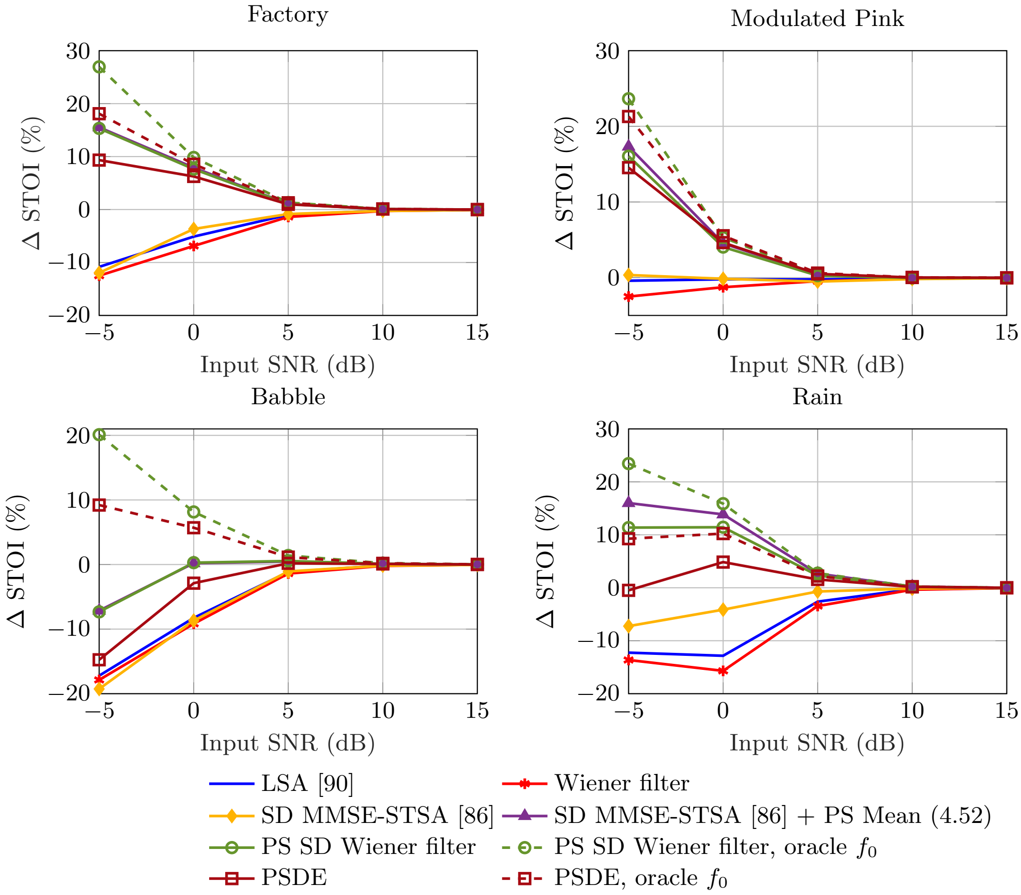

Figure 4.7: Summary of the final results in terms of ΔSTOI versus global SNR.

- We observe that the detection step helps to further improve the algorithm's performance in PESQ while intelligibility is slightly decreased. The PSDE even outperforms the oracle f_0 implementation of the PS SD Wiener filter in PESQ for modulated pink noise and performs equally good for factory noise at SNRs above 0 dB.

- Employing the harmonic model estimate of our proposal as spectral mean value for the SD MMSE-STSA improves the algorithm's performance in STOI but degrades the PESQ at SNRs above 0 dB. This demonstrates that different harmonic model estimates may affect intelligibility and perceived quality differently. Hence, the specific choice of these estimates is an important design step in speech enhancement algorithms that take into account the harmonic model.

- The improvement from the proposed speech enhancement procedure is particularly pronounced at low SNRs. The detection-estimation step helps to retain the PESQ performance also at high SNRs for factory, modulated pink, and babble noise while

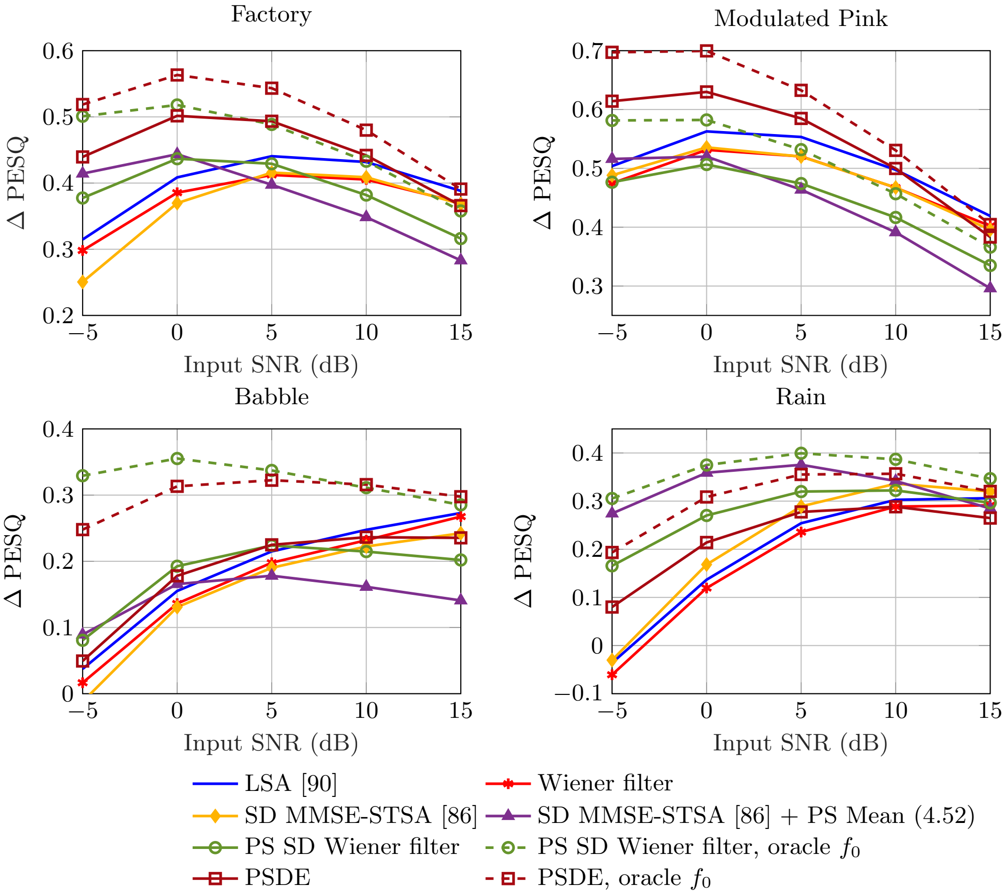

Figure 4.8: Summary of the final results in terms of ΔPESQ versus global SNR for the LSA [90], Wiener filter, SD MMSE-STSA [86], SD MMSE-STSA with mean value estimated from Equation (4.52), PS SD Wiener filter, and the PSDE algortihm.

decreasing the performance gain in STOI at low SNRs. In general, at 15 dB SNR, all methods approximately perform the same. In rain noise, the SD MMSE-STSA estimator [86] together with the pitch-synchronous estimate of the spectral mean performs best (neglecting the oracle f_0 scenarios) in terms of PESQ and STOI for SNRs up to 15 dB.

- In case of non-stationary noise types such as babble noise, the oracle f_0 scenario indicates that improved performance can be achieved if more reliable f_0-estimates are available. This is reasonable, considering that the algorithm in its individual steps strongly relies on accurate f_0-estimation, known to be a difficult task at low SNRs and non-stationary noise [141].

4.7.4 Subjective Evaluation of Speech Enhancement Performance

For the subjective evaluation of the proposed method we followed the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) standard [147]. We included one anchor, obtained by filtering the clean signal by a lowpass filter with a cutoff frequency of 3.5 kHz. 15 listeners participated in the test, which took place in a quite office. As transducer we used AKG-K271 MKII headphones. The listening excerpts were structured as follows; 8 gender balanced randomly selected utterances were taken from the TIMIT core database test set. These were corrupted by factory and babble noise at SNRs of 5 dB and 15 dB each. Each signal was then processed by the LSA [90], SD MMSE-STSA [86], and PSDE algorithms, yielding 6 test signals per utterance together with the noisy utterance, the hidden reference and the anchor. After a training phase, the participants were asked to rate the overall sound quality on a scale from 0 to 100 with the Matlab GUI adapted from [148]. The participants were presented all 6 processing variants of each utterance at once and were allowed to listen to each of the sound files as often as they wanted. The results of this evaluation, averaged over SNRs, are shown in Figure 4.9.

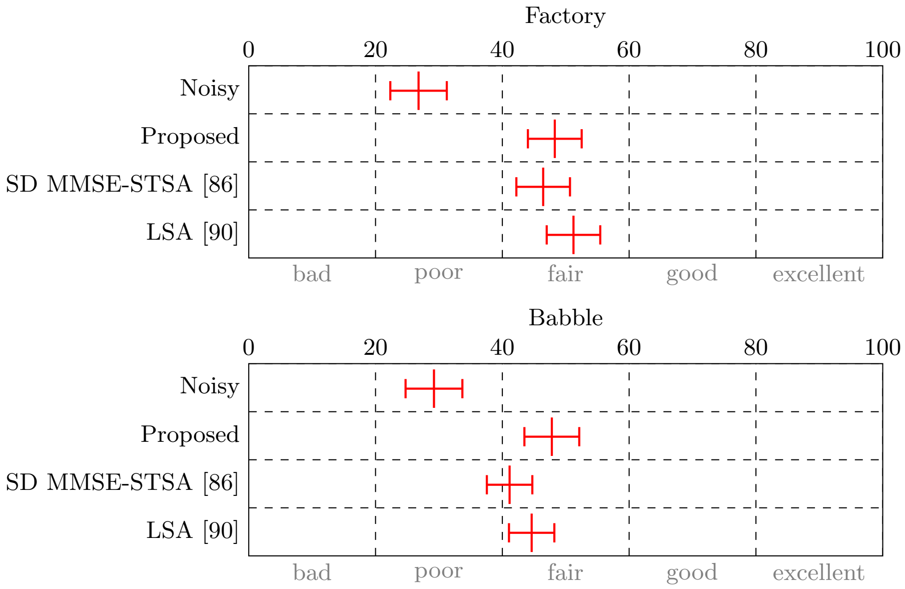

Figure 4.9: Results of the subjective listening test averaged over all participants. The reported confidence interval corresponds to a confidence level of 95%.

Although all participants reported that the algorithms differed in the type of artifacts and the amount of noise reduction they introduced, the three algorithms perform close when averaged over all participants. Pairwise two-tailed t-tests among the methods give the following insights:

- For factory noise, representing a rather stationary real-life noise type, the differences between the algorithms are not significant for a significance level of $p = 0.05$.
- For babble noise, as a non-stationary noise scenario, the preferences of the PSDE algorithm and the LSA algorithm over the SD MMSE-STSA are indicated by p-values of $p = 0.01$ for the LSA and $p = 0.0002$ for the PSDE. The preference of the PSDE over the LSA is quantified by a p-value of $p = 0.04$.

4.8 Conclusion

This chapter presented a pitch-synchronous simultaneous detection-estimation framework under the harmonic plus noise model for speech. The detection-estimation stage is an alternative to existing soft-decision approaches to speech enhancement under the harmonic plus noise model, the pitch-synchronousity is a result of optimizing the signal analysis stage to the estimation of the harmonic signal components. We exploit the specific characteristics of the signal analysis stage by incorporating an empirical Bayes approach for estimating the harmonic components. The resulting scheme is modular, i.e. each of the three steps (i) analysis, (ii) harmonic retrieval, and (iii) detection/estimation can in principle be replaced by any other method performing the respective task.

Methods relying on the same signal model such as [40] or [27, 28] typically solve the problem of harmonic overfitting by incorporating speech presence uncertainty and/or a fixed threshold that limits the maximal harmonic frequency. The empirical Bayes approach circumvents the necessity to explicitly define a maximum voicing frequency, as the prior distribution puts less weight on higher order harmonics. The proposed method positively affects both, the predicted speech quality as well as the speech intelligibility compared to the noisy signal and reference algorithms at low SNRs for the reported noise types.

In the light of the research question posed in Chapter 1, this chapter illustrates that the statement “the noisy phase is optimal” is in fact a result of a commonly used statistical model enforcing circularity of the complex-valued speech DFT coefficient’s prior distribution. If we incorporate the harmonic plus noise signal model, this condition is not fulfilled, resulting in MMSE optimal phase estimates that are not equal to the noisy phase.

The results indicate an overall preference for the PSDE method compared to the benchmarks, in particular the oracle f_0 scenarios perform well in terms of instrumental predictors of perceived speech quality and intelligibility. Also the improved STOI performance of the SD MMSE-STSA [86] in conjunction with the proposed MAP harmonic retrieval scheme is an interesting result, indicating the importance of the harmonic spectrum for intelligibility enhancement. However, the fact that all listening test participants reported that the algorithms were distinguishable but did not decide significantly in favor of the PSDE or the SD MMSE-STSA [86] algorithm compared to the computationally less expensive LSA algorithm poses the question if it is the instantaneous harmonic spectrum that is the key to improved overall single-channel speech enhancement.

5

Exploiting Phase Structure along Time for Single-Channel Speech Enhancement

This chapter is dedicated to analyzing and exploiting correlation of complex-valued STFT coefficients along time for single-channel speech enhancement algorithms. We will present a frequency domain Kalman filter that exploits the simplified statistics resulting from harmonic signal modeling together with a slight modification to the classical STFT framework, termed pitch-adaptive STFT. More specifically, we extend a complex-valued subband Kalman filter by using circular statistics in combination with harmonic signal modeling for the estimation of the corresponding state transition model. All derivations and simulations included in this chapter have been conducted by the author. The complex-valued Kalman filter is part of the submitted work in [5]. The idea of using a pitch-adaptive zero padding was originally intended to improve the DD *a priori* SNR estimator and has been published in [3]. Parts of the text in this chapter have been published in [3].

5.1 Overview and Motivation

The vast majority of speech enhancement algorithms performs the speech estimation step separately for each frame, but there also exist various approaches that exploit inter-frame correlations of STFT coefficients for speech enhancement. The main motivation behind such approaches is that the signal frames are typically correlated since they are overlapping [61]. The amount of correlation however highly depends on the frame shift of the analysis stage. In the context of linear filtering, subband inter-frame correlations have been exploited in [62, 63, 149], where MVDR filters and multiframe Wiener filters have been proposed. These concepts, specifically the task of estimating the inter-frame correlations of speech and noise, have been revisited in [150, 151] and were demonstrated for speech enhancement. Highly relevant for this chapter, the methods presented in [152], [153], and [154] formulate modified Kalman filters based on an AR model along time for STFT coefficients. A Kalman filter that tracks the speech phase in the modulation domain has been proposed in [155].

Interestingly, all the aforementioned works formulate their optimization criteria in terms of the complex-valued STFT coefficients. Independent of the strategy chosen, the parameters, whether they are coefficients of a linear filter or describing an AR model, are not known. Hence, they are either assumed to be known from isolated clean speech data, or they are estimated from some pre-enhanced data. This is problematic, as we will show in Section 5.4 that the low correlation in the spectral phase along time renders the overall correlation of the complex-valued STFT coefficients insignificant.

Given the harmonic plus noise model for speech together with a fundamental frequency estimate, an explicit expression for the phase progression from one frame to the next is available as given in Equation (3.17). However, as the harmonic frequency changes across time, knowledge of the window function is needed to successfully map the harmonic signal components to the STFT. In general, if there is no analytic expression for the window given, this can be obtained by interpolation strategies. This chapter illustrates how a simple zero padding in the time domain may be exploited to align DFT bins that are dominated by the same sinusoid along time, yielding a simple statistical model for the inter-frame correlations that makes the application of a subband Kalman filter straightforward. In order to distinguish the zero padding strategy from the pitch-synchronous method in Chapter 4 we refer to the resulting time-frequency representation as pitch-adaptive short-time Fourier transform (PASTFT). The proposed (P)itch-(A)daptive (Co)mplex-valued Kalman filter (PACO) is evaluated by means of instrumental metrics and a subjective listening test, the latter indicating improved noise reduction performance while preserving the amount of speech distortions.

This chapter consists of three parts:

- First, we introduce, analyze, and explain the features of the PASTFT framework. This analysis also investigates temporal correlation of the spectral phase when harmonic signal modeling is taken into account.

- In a second step, we present an extension of the complex-valued Kalman filter originally proposed in [152]. Specifically, we (i) incorporate harmonic signal modeling into the state-transition model and (ii) further propose a way to estimate the corresponding parameters that exclusively relies on the statistics of the signal's spectral phase.

- Finally, the third part of this chapter presents the objective and subjective evaluation of the proposed method compared to relevant benchmarks.

The idea of applying a pitch-adaptive STFT was originally intended to improve the DD *a priori* SNR estimator. In the following we will sketch the line of thinking that results in the PASTFT and subsequently will go on to analyze its properties w.r.t. the correlations of the PASTFT coefficients along time.

5.2 The Decision-Directed A Priori SNR Estimator: Revisited

Let us recall the DD *a priori* SNR estimator as explained in Section 2.2.3. At its core, it linearly combines two estimates of the *a priori* SNR as follows [15]

$$\hat{\xi}_{\mathrm{DD}}(k,\ell) = \alpha_{\mathrm{DD}}\hat{\xi}_{\ell-1}(k,\ell) + (1-\alpha_{\mathrm{DD}})\max[\hat{\xi}_{\mathrm{ML}}(k,\ell),0]. \tag{5.1}$$

From Equation (5.1), we note that the estimate of the *a priori* SNR for specific $k=k'$ and $\ell=\ell'$, $\hat{\xi}_{\mathrm{DD}}(k',\ell')$ strongly relies on $\hat{\xi}_{\ell-1}(k',\ell')$, which is obtained from the speech estimate $\hat{X}(k',\ell'-1)$ of the preceding frame. However, particularly in the case of larger frame shifts, DFT bin k', which is dominated by speech at frame ℓ', is not necessarily dominated similarly by speech at frame $\ell'-1$ and vice versa.

This has two reasons. First, onsets and offsets induce a change in speech presence/absence from one frame to the next (which is addressed in, e.g., [61, 156]). Second, considering voiced speech as the summation of harmonically related sinusoids has similar consequences. As illustrated in Figure 5.1 (a) and (b), one harmonic does not dominate the same frequency bin for all time instants, since the fundamental frequency changes over time. Hence, assuming that the *a priori* SNR is approximately constant along harmonic trajectories, it is not necessarily $\hat{\xi}_{\ell-1}(k',\ell')$ which approximates $\xi(k',\ell')$ best but potentially any other $\hat{\xi}_{\ell-1}(k,\ell')$ with k close to k' that is dominated by the same harmonic as k'. These considerations are in line with the justification of overcomplete bases given in [43], following the idea of analyzing, and in this case combining, the energy at *physically relevant* partials. Further, it is the same rationale that motivates the harmonic noise PSD tracker developed in [157].

5.3 A Pitch-Adaptive Analysis Stage

Following the above discussion, we aim to recursively smooth the *a priori* SNR estimates along harmonic trajectories instead of fixed DFT frequency bins. Hence, for successive frames, we seek for frequency bins that are dominated by the same harmonic. In order to find potential candidates, we define $k_h(\ell)$, representing the frequency bin k which is closest to the h^{th} harmonic with frequency $f_h(\ell) = hf_0(\ell)$, i.e.,

$$k_h(\ell) = \arg\min_k \left| k - N_{\mathrm{DFT}}(\ell)\frac{hf_0(\ell)}{f_s} \right|. \tag{5.2}$$

We can simplify Equation (5.2) by choosing $N_{\mathrm{DFT}}(\ell)$ dependent on the fundamental frequency

$$N_{\mathrm{DFT}}(\ell) = \mathrm{round}\left(K\frac{f_s}{f_0(\ell)}\right), \tag{5.3}$$

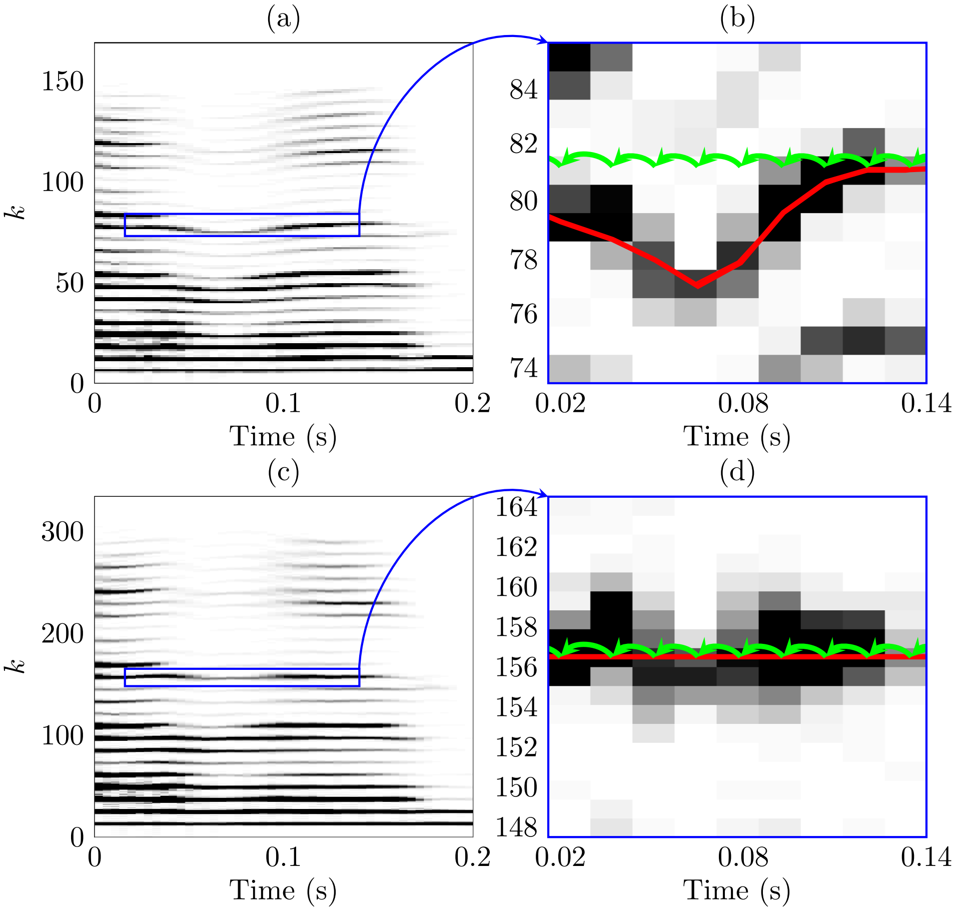

Figure 5.1: (a): Spectrogram of a speech snippet uttered by a female speaker taken from [139]. (b) Zoom into a voiced time-frequency region where the fundamental frequency changes over time. The red solid line indicates the trajectory of harmonic 14. The green arrows indicate the DD smoothing path. (c) The same speech snippet as in (a) analyzed with the PASTFT ($K = 12$ in Equation (5.5)). (d) Now the DD smoothing path at frequency bin $k = 168$ and the trajectory of harmonic 14 coincide.

where K is an integer constant similar to Equation (4.32). The factor K controls the amount of zero padding in the DFT. Inserting Equation (5.3) into Equation (5.2) renders $k_h(\ell)$ to be independent of $N_{\text{DFT}}(\ell)$. Further, using $K\frac{f_s}{f_0(\ell)} \approx \text{round}\left(K\frac{f_s}{f_0(\ell)}\right)$, we obtain

$$\begin{aligned} k_h(\ell) &= \arg\min_k \left| k - \text{round}\left(K\frac{f_s}{f_0(\ell)}\right)\frac{hf_0(\ell)}{f_s}\right| \\ &\approx \arg\min_k |k - Kh| \\ &= Kh. \end{aligned} \tag{5.4}$$

By applying a pitch-adaptive zero padding, $k_h(\ell)$ becomes a constant that does not depend on the frame index ℓ anymore. Hence, $k_h(\ell)$ is always closest to the same harmonic h and its argument ℓ becomes redundant, which is why we drop it.

One harmonic h not only impacts on frequency bin k_h but on all other frequency bins as well. Under the assumption that the speech signal is perfectly harmonic and we know its fundamental frequency, the amount of leakage depends on the chosen window function only. This means that ideally, not only all frequency bins k_h, but also those in-between harmonics are affected similarly by the harmonics at all time instances. By applying the decision-directed approach as defined in Equation (2.31) in the PASTFT framework, obtained by using Equation (5.3) in Equation (2.5), we automatically average along harmonic trajectories instead of fixed frequencies as illustrated in Figure 5.1, panels (c) and (d). The resulting estimator is termed pitch-adaptive decision-directed (PADDi) in [3].

Note that in the pitch-synchronous case in Section 4.4.1, the assumption that each DFT bin is dominated by only one harmonic is true if (i) the signal is perfectly harmonic, (ii) the fundamental frequency is high enough, and (iii) the optimal window length, determined by the pitch period in Equation (5.3) is an integer already before rounding. In the pitch-adaptive case, this is only true if by chance the optimal DFT length $N_{\text{DFT}}(\ell)$ is an integer multiple of the fixed frame length.

5.3.1 The factor K

We can refine the resolution of the DFT by increasing the factor K. However, we cannot select the DFT lengths arbitrarily long if we want to keep the computational effort reasonable. Further, to assure that no non-zero samples in $y_{\text{t}}(n,\ell)$ are neglected for the computation of $Y(k,\ell)$, the DFT length is lower bounded by N. Choosing the window length $N = N_{\text{DFT}}(\ell)$ results in a *pitch-synchronous* representation of the signal, while in the *pitch-adaptive* case, the window length itself is fixed.

Assuming that the fundamental frequency of speech signals lies within the interval $[f_{0,\min}; f_{0,\max}]$, we can express the following bounds for $N_{\text{DFT}}(\ell)$

$$\max\left[N, K\frac{f_s}{f_{0,\max}}\right] \leq N_{\text{DFT}}(\ell) \leq K\frac{f_s}{f_{0,\min}}. \tag{5.5}$$

For the sake of computational efficiency, we select the minimum possible value for the factor K, given by

$$K = \lceil \frac{f_{0,\max}}{f_s} N \rceil, \tag{5.6}$$

where $\lceil\cdot\rceil$ denotes the ceiling operation.

5.3.2 Signal Reconstruction

Given some speech estimate $\hat{X}(k,\ell)$, the time domain frames of the estimated speech signal are computed by applying the inverse DFT

$$\hat{x}_{\mathrm{t}}(n,\ell) = w(n) \sum_{n=0}^{N_{\mathrm{DFT}}(\ell)-1} \hat{X}(k,\ell)\mathrm{e}^{j\frac{2\pi k}{N_{\mathrm{DFT}}(\ell)}n}, \tag{5.7}$$

where the multiplication with the window $w(n)$ ensures that the support of $\hat{x}_{\mathrm{t}}(n,\ell)$ is on $[0, N-1]$. This operation is necessary, as the filtering in the frequency domain corresponds to a circular convolution in the time domain, implying that the support of the filtered signal is in general longer than the support of the original signal. Afterwards, a standard overlap-add routine can be applied.

5.3.3 Phase Decomposition in the PASTFT

Similar to Section 3.2.2 and Section 4.4.1, we can decompose the PASTFT phase into an initial and a progression part yielding the phase normalized observation $\widetilde{Y}(k,\ell)$. Note that the approximation that each DFT bin is dominated by one harmonic only, as opposed to the pitch-synchronous case, is only exact if $N_{\mathrm{DFT}}(\ell)$ obtained from Equation (5.3) happens to be an integer multiple of the *fixed* segment length N, which is in general not the case. Further, in contrast to the classical STFT framework, there is a fixed mapping between h and k. This means that in the close vicinity of a strong harmonic h, if its phase $\alpha_h(\ell)$ and magnitude $A_h(\ell)$ vary slowly over time, this is also true for $\alpha(k,\ell)$ and $A(k,\ell)$.

Figure 5.2 shows scatter plots of successive PASTFT coefficients' instantaneous and initial phases. The instantaneous phase does not reveal any correlation between successive frames. However,the initial phase estimates exhibit high correlation. Motivated by this observation, we suspect that for lag m, the correlation between $\widetilde{X}(k,\ell)$ and $\widetilde{X}(k,\ell-m)$ is now in general increased compared to the correlation between $X(k,\ell)$ and $X(k,\ell-m)$ or even the standard STFT framework. Further, we expect to see significant correlation in the initial phases $\alpha(k,\ell)$ and $\alpha(k,\ell-m)$, not only in the magnitudes $A(k,\ell)$ and $A(k,\ell-m)$ as, e.g. reported in [61] for the standard STFT. We will justify these conjectures by experiments in the following.

5.4 Inter-Frame Correlation Analysis

In this section we compare the temporal correlation of DFT coefficients resulting from the PASTFT framework to the classical STFT. Besides the increased temporal correlation of the spectral magnitude compared to the standard STFT, we can incorporate the phase progression model to access estimates of the initial phase. This provides insights

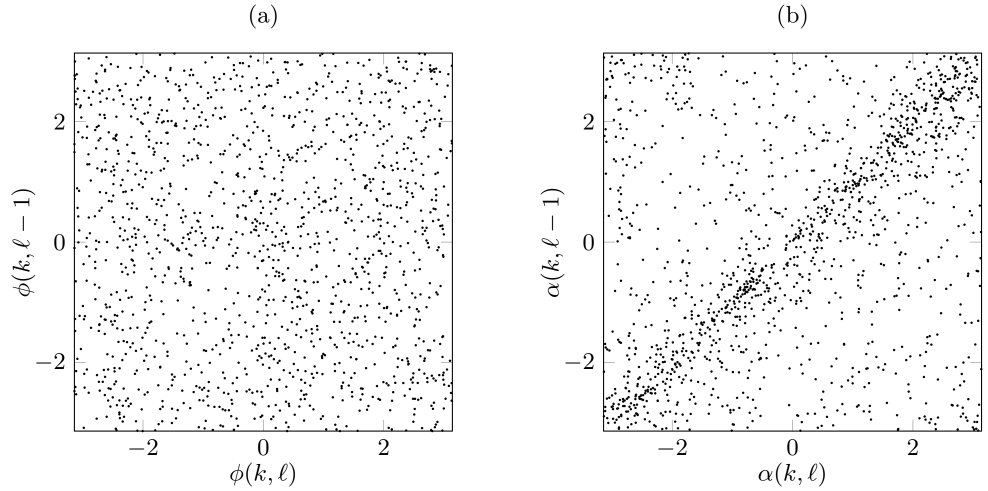

Figure 5.2: Scatter plot of successive PASTFT coefficients' phase, i.e. phase at frame ℓ versus frame $\ell-1$ with $k=36$, corresponding to harmonic 3 for $K=12$. (a) Instantaneous phase, (b) initial phase estimate.

into the correlation along time of the spectral phase. We quantify the amount of correlation between successive DFT bins by means of the normalized sample autocorrelation sequence (ACS). We evaluate the ACS for $A(k,\ell)$, $\phi(k,\ell)$, $X(k,\ell)$, $\alpha(k,\ell)$, and $\widetilde{X}(k,\ell)$. We compute the ACS of a quantity $q(k,\ell)$ at lag m as follows

$$r_{\mathrm{qq}}(m) = \frac{\sum_{\ell\in\mathcal{L}}(q(k,\ell)-\bar{q}(k))(q(k,\ell+m)-\bar{q}(k))^*}{\sum_{\ell\in\mathcal{L}}(q(k,\ell)-\bar{q}(k))(q(k,\ell)-\bar{q}(k))^*}, \tag{5.8}$$

where $\mathcal{L}$ is the set of frames that is considered for the computation of the ACS, and $\bar{q}(k)$ denotes the sample mean of $q(k,\ell)$, given by

$$\bar{q}(k) = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} q(k,\ell), \tag{5.9}$$

with $|\mathcal{L}|$ being the cardinality of $\mathcal{L}$. To ensure that we only take into account DFT bins that contain speech, we define the set of frames $\mathcal{L}$ by thresholding the magnitude as proposed in [61]

$$\mathcal{L} = \{\ell | A(k,\ell) \geq 10^{-30/20} \max_{\ell}\left[A(k,\ell)\right]\}. \tag{5.10}$$

For the following analysis, as for the rest of Section 5.4, we closely followed the procedure in [61], which presents a similar analysis for the standard STFT. As speech data we concatenated seven randomly selected sentences from the TIMIT database [139]. The fundamental frequency estimate needed for choosing the DFT lengths was obtained from the PEFAC estimator [141]. The sampling frequency is 16 kHz, the frame length was chosen with 32 ms, and the frame shift was set to 16 ms, if not noted otherwise. Although of course the fundamental frequency trajectory used for phase decomposition is an estimate itself, we do not use the hat symbol for the initial phase in the following evaluation, since it serves as our ground truth.

5.4.1 Sample Autocorrelation Sequences

In Figure 5.3, we plot the ACSs of all abovementioned quantities for two frequency bins; one, that in theory is always dominated by harmonic 3 and a second one, which is inbetween two harmonics, where the impact of harmonic speech components is very small in theory. For a factor $K = 12$ we hence chose $k \in \{36, 42\}$. Further, to address our speculation that in general, successive DFT bins in the PASTFT exhibit higher correlation than in the standard STFT, we also report the ACSs of the same quantities for the STFT. We selected the ks in the STFT to lie within the same frequency range as the corresponding frequency bins in the PASTFT. This is achieved by averaging over the frequencies that are associated to these frequency bins and choosing the nearest DFT frequencies for a fixed DFT length of 512, resulting in $k \in \{16, 19\}$.

Figure 5.3 indeed confirms that for DFT bins that are dominated by harmonics (i) the correlation of successive DFT bins is higher in the PASTFT than in the STFT framework and that (ii) decomposing the instantaneous phase by multiplying with $\Psi^*(k, \ell)$ from Equation (3.15) not only reveals correlation in the initial phase but also in the complex-valued coefficients $\widetilde{X}(k, \ell)$. However, this does not hold for DFT bins that are not in the direct vicinity of a harmonic, where the inter-frame correlations are even reduced due to different DFT resolutions across frames.

5.4.2 Autocorrelation as a Function of Frequency

As already explained, we suspect the subband inter-frame correlations to highly depend on the frequency bin index in the PASTFT. We illustrate the frequency bin dependent nature of the correlation's extent in Figure 5.4 (a) for lags ranging from $m = 1$ to $m = 4$. Compared to the STFT, see Figure 5.4 (b), by choosing pitch-adaptive DFT lengths, we achieve higher correlation at frequency bins in the vicinity of harmonics and the correlation is better preserved for larger lags m.

5.4.3 Autocorrelation as a Function of Frame Shift

Finally, the latter observation motivates us to study how the frame shift L affects the ACS. To this end, we compare the PASTFT for $k = 36$ with the STFT for $k = 16$. Figure 5.5 depicts the corresponding results for lags $m = 1$ and $m = 2$ for frame shifts ranging from $L = 32$ samples (93.75% overlap) to $L = 256$ samples (50% overlap). We

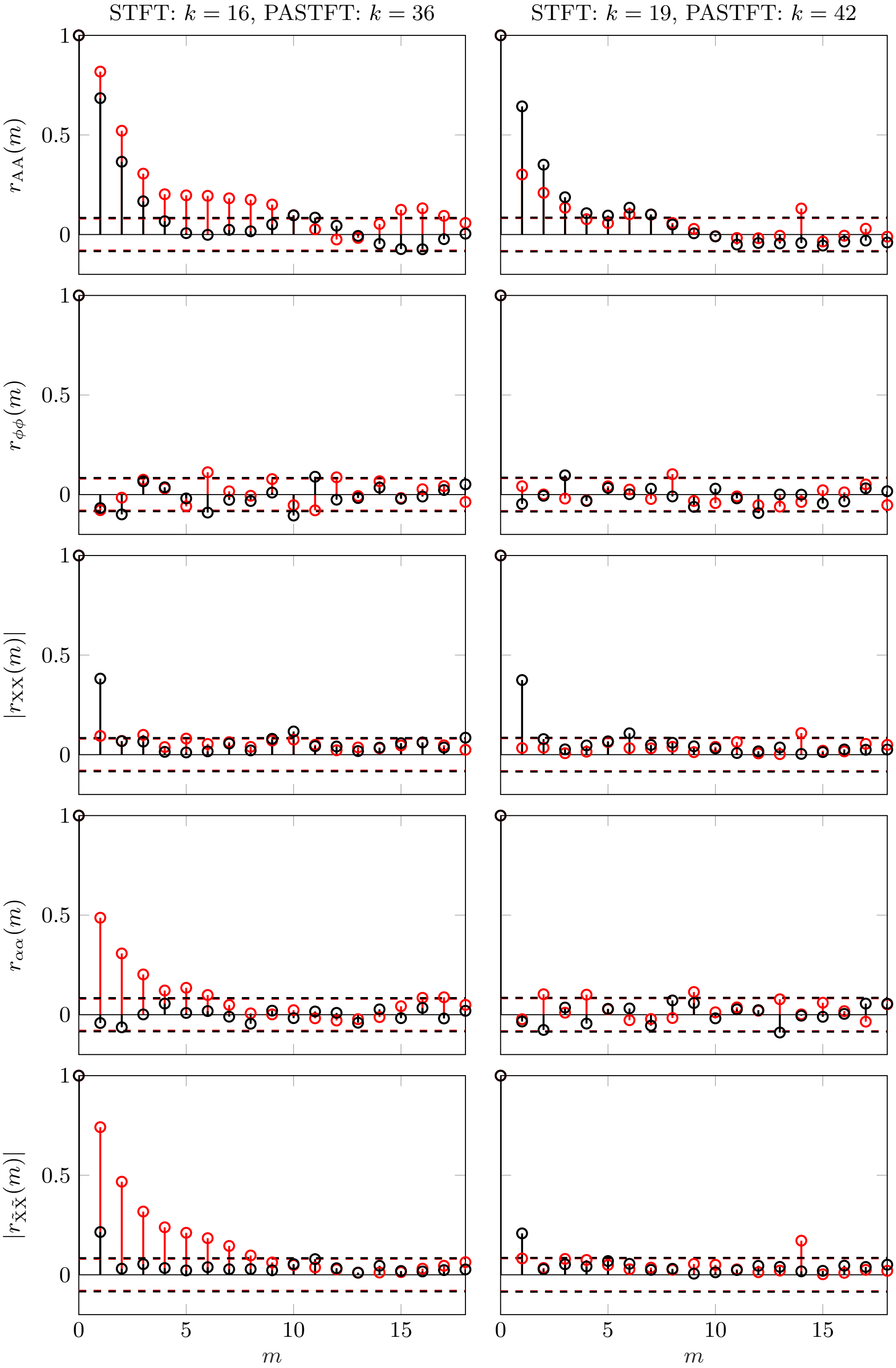

Figure 5.3: Illustration of temporal correlation of the DFT bins by their ACSs. The dotted lines correspond to the 95% confidence limits. Black lines correspond to STFT and red lines correspond to PASTFT, the respective DFT bins can be found in the column titles.

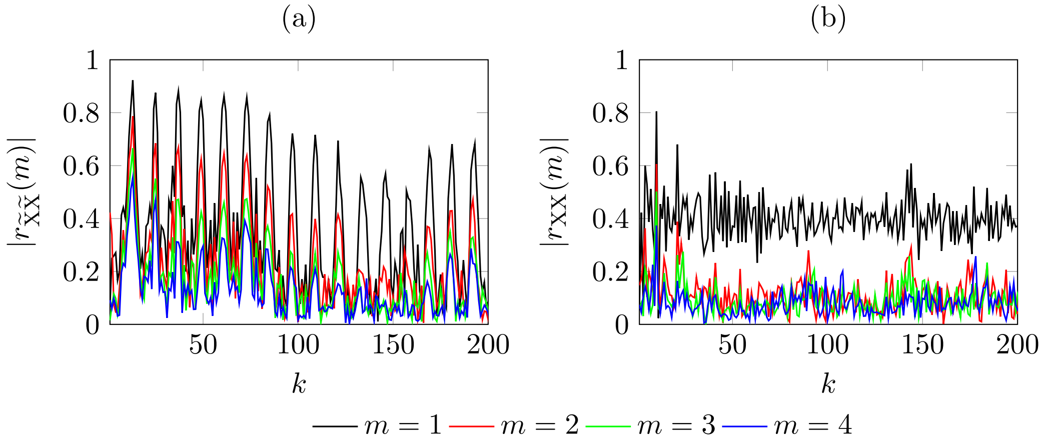

Figure 5.4: (a) Correlation coefficients of the complex-valued DFT coefficients with initial phase used instead of instantaneous phase versus frequency bin in PASTFT. (b) Correlation coefficients versus frequency bin in STFT.

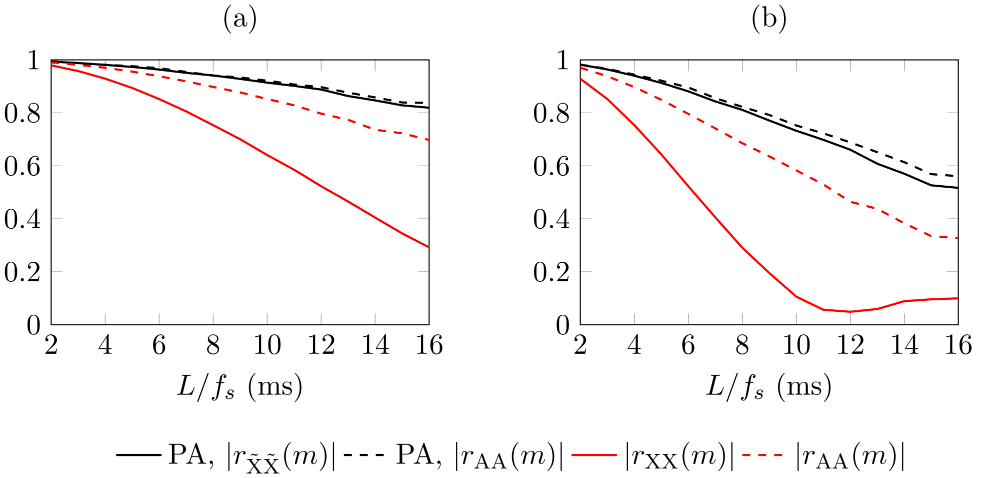

Figure 5.5: (a) Correlation coefficients for PASTFT ($k = 36$) and STFT ($k = 16$) versus frame shift for lag $m = 1$. (b) Same experiment with lag $m = 2$.

observe that (i) the ACS's dependency on the frame shift is lower in the PASTFT than in the STFT and that (ii) in the PASTFT, taking into account the correlation in phase preserves the overall correlation for larger frame shifts, even for the complex-valued DFT coefficients (compare dashed and solid lines).

In the following, motivated by the above analysis, we present one possibility to explicitly take into account inter-frame correlations for speech enhancement. This is achieved by formulating a complex-valued Kalman filter, as described in the following.

5.5 The Complex-Valued Kalman Filter

As discussed in the previous section, even for large frame shifts, successive DFT coefficients in the PASTFT are highly correlated when decomposing the phase into an initial and a progression part. To take into account this correlation, similarly to [152], we model DFT coefficients as a complex-valued AR process, represented by the linear state-space description

$$X(k,\ell) = \mathbf{a}^T(k,\ell)\mathbf{x}(k,\ell-1) + X_{\text{in}}(k,\ell), \tag{5.11}$$

where $\mathbf{a}(k,\ell) \in \mathbb{C}^{N_{\text{AR}}\times 1}$ is a vector containing the complex coefficients of the AR model of order N_{AR}, the previous DFT-coefficients are stacked into the vector

$$\mathbf{x}(k,\ell-1) = \begin{bmatrix} X(k,\ell-1) & X(k,\ell-2) & \dots & X(k,\ell-N_{\text{AR}}) \end{bmatrix}^T,$$

and $X_{\text{in}}(k,\ell)$ represents the innovation, i.e. those parts that are not captured by the AR model. In the following, we explain the pitch-adaptive complex-valued Kalman filter (PACO) which jointly benefits from the PASTFT and AR modeling due to the simplified statistics that arise from the PASTFT compared to the STFT in the light of harmonic signal modeling.

From Equation (5.11) it follows the model for the observed noisy speech

$$Y(k,\ell) = \underbrace{\mathbf{a}^T(k,\ell)\mathbf{x}(k,\ell-1)}_{X_{\text{prop}}(k,\ell)} + X_{\text{in}}(k,\ell) + D(k,\ell), \tag{5.12}$$

where we assume $X_{\text{in}}(k,\ell)$ and $D(k,\ell)$ to be complex, isotropic Gaussian processes with variances $\sigma^2_{x,\text{in}}(k,\ell)$ and $\sigma^2_d(k,\ell)$, respectively. Given this model, we can formulate the standard Kalman filter equations as in [152]. The Kalman filter can be divided into two steps. The propagation step represents the interdependence of successive frames and the update step combines information retrieved from the current observation with the propagated state estimates.

Propagation Step

The state estimate propagation with $\mathbf{x}(k,\ell) \approx \hat{\mathbf{x}}(k,\ell)$ is given by

$$\hat{X}_{\text{prop}}(k,\ell) = \mathbf{a}^T(k,\ell)\hat{\mathbf{x}}(k,\ell-1), \tag{5.13}$$

with previous estimates of the clean speech coefficient stacked into the vector

$$\hat{\mathbf{x}}(k,\ell-1) = \begin{bmatrix} \hat{X}(k,\ell-1) & \hat{X}(k,\ell-2) & \dots & \hat{X}(k,\ell-N_{\text{AR}}) \end{bmatrix}^T.$$

The error covariance is propagated as follows

$$P_{\text{prop}}(k,\ell) = P_{\text{up}}(k,\ell-1)\mathbf{a}^H(k,\ell)\mathbf{a}(k,\ell) + \sigma^2_{x,\text{in}}(k,\ell). \tag{5.14}$$

Update Step

Given the propagated state estimates, we can compute the Kalman gain

$$G_{\text{K}}(k,\ell) = \frac{P_{\text{prop}}(k,\ell)}{P_{\text{prop}}(k,\ell) + \sigma^2_d(k,\ell)}. \tag{5.15}$$

Now, an estimate of the innovation process is achieved by multiplying it with the residual that results from taking into account the state estimate propagation

$$\hat{X}_{\text{in}}(k,\ell) = G_{\text{K}}(k,\ell)\big(Y(k,\ell) - \hat{X}_{\text{prop}}(k,\ell)\big). \tag{5.16}$$

Given the signal model in Equation (5.11), we update the clean speech estimate and the error covariance

$$\begin{aligned} \hat{X}_{\text{up}}(k,\ell) \quad &= \hat{X}_{\text{prop}}(k,\ell) + \hat{X}_{\text{in}}(k,\ell), && (5.17)\\ P_{\text{up}}(k,\ell) \quad &= (1 - G_{\text{K}}(k,\ell))P_{\text{prop}}(k,\ell). && (5.18) \end{aligned}$$

5.5.1 Estimating the AR Coefficients

The propagation and update steps as described above are straightforward once the model parameters, i.e. state transition model, process variance, and noise variance are given. However, the non-stationarity of speech together with the inaccessibility of the clean speech signal renders the problem of applying the above Kalman filter rather difficult. In the following, motivated by the harmonic plus noise model, we include the phase decomposition into the AR model, i.e. we assume that the fundamental frequency trajectory contributes to the AR coefficients' phase following Equation (3.8). The key ingredient of PACO is using this model together with circular statistics to estimate the AR parameters. Given that the concepts described in the following are equally applied for all frequency bins, we drop the frequency index k from now on. Further, we will use the notation convention $q^{(u,v)}(\ell) \equiv q(\ell-u,\ell-v)$ from now on to indicate specific inter-frame relationships of a quantity q.

Typically, the AR coefficient vector is derived from the MMSE criterion. Hence, we have

$$\hat{\mathbf{a}}(\ell) = \arg\min_{\mathbf{a}(\ell)} \mathbb{E}\big(\big(X(\ell) - X_{\text{prop}}(\ell)\big)\big(X(\ell) - X_{\text{prop}}(\ell)\big)^*\big). \tag{5.19}$$

Solving Equation (5.19) yields the well-known result

$$\hat{\mathbf{a}}(\ell) = \mathbf{R}_{\mathrm{XX}}^{-1}(\ell)\mathbf{r}_{\mathrm{XX}}(\ell), \tag{5.20}$$

with autocorrelation vector

$$\mathbf{r}_{\mathrm{XX}}(\ell) = \mathbb{E}\big(\mathbf{x}(\ell-1)X^*(\ell)\big) \in \mathbb{C}^{N_{\mathrm{AR}}\times 1} \tag{5.21}$$

and autocorrelation matrix

$$\mathbf{R}_{\mathrm{XX}}(\ell) = \mathbb{E}\big(\mathbf{x}(\ell-1)\mathbf{x}^H(\ell-1)\big) \in \mathbb{C}^{N_{\mathrm{AR}}\times N_{\mathrm{AR}}}. \tag{5.22}$$

$\mathbf{r}_{\mathrm{XX}}(\ell)$ and $\mathbf{R}_{\mathrm{XX}}(\ell)$ are unknowns in practice and need to be estimated from the signal samples available. Since in speech enhancement we deal with noise corrupted observations, simple sample autocorrelation estimation cannot be applied. In [152], as already proposed in [158], this problem was solved by estimating the AR coefficients from past enhanced samples. Further, in [159] the LSA estimator [15] was used for pre-enhancing the noisy signal before estimating the AR coefficients. In this work, we will exploit the fact that at least for the coherent parts of the of the short-term autocorrelation's phase, as a consequence of harmonic signal modeling, we have an explicit expression available, as depicted in the following.

For $R_{\mathrm{XX}}^{(u,v)}(\ell)$, representing the element in the u^{th} row and v^{th} column of $\mathbf{R}_{\mathrm{XX}}(\ell)$, we have

$$\begin{aligned} R_{\mathrm{XX}}^{(u,v)}(\ell) &= \mathbb{E}\big(X(\ell-u)X^*(\ell-v)\big) \\ &= \mathbb{E}\big(\widetilde{X}(\ell-u)\widetilde{X}^*(\ell-v)\Delta\Psi^{(u,v)}(\ell)\big), \end{aligned} \tag{5.23}$$

where following Section 3.2.2 we define

$$\Delta\Psi^{(u,v)}(\ell) = \Psi(\ell-v)\Psi^*(\ell-u) \tag{5.24}$$

for which we have an estimate available from Equation (3.15) in conjunction with the recursion in Equation (3.17). Hence, inserting the estimated phase terms we have

$$\hat{R}_{\mathrm{XX}}^{(u,v)}(\ell) = \underbrace{\mathbb{E}\big(\widetilde{X}(\ell-u)\widetilde{X}^*(\ell-v)\big)}_{R_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}^{(u,v)}(\ell)}\Delta\Psi^{(u,v)}(\ell), \tag{5.25}$$

where $R_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}^{(u,v)}(\ell)$ denotes the $(u,v)^{\mathrm{th}}$ entry of the autocorrelation matrix $\mathbf{R}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell)$. The corresponding autocorrelation vector $\mathbf{r}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell)$ is constructed from the entries $r_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}^{(u)}(\ell) =$

$R^{(u,0)}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell)$. By defining the diagonal matrix

$$\widehat{\mathbf{\Psi}}(\ell) = \begin{bmatrix} \Delta\Psi^{(0,1)}(\ell) & 0 & \dots & 0 \\ 0 & \Delta\Psi^{(0,2)}(\ell) & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Delta\Psi^{(0,N_{\mathrm{AR}})}(\ell) \end{bmatrix}, \tag{5.26}$$

and by taking into account that we only have estimates of the autocorrelation matrix and vector available, we can rewrite Equation (5.19) as

$$\hat{\mathbf{a}}(\ell) = \widehat{\mathbf{\Psi}}(\ell)\widehat{\mathbf{R}}^{-1}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell)\widehat{\mathbf{r}}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell), \tag{5.27}$$

with

$$\begin{aligned} R^{(u,v)}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell) &= \mathbb{E}\big(\widetilde{X}(\ell-u)\widetilde{X}^*(\ell-v)\big), \\ &= \mathbb{E}\big(A(\ell-u)A(\ell-v)\mathrm{e}^{j(\alpha(\ell-u)-\alpha(\ell-v))}\big), \end{aligned} \tag{5.28}$$

with $\Delta\alpha^{(u,v)}(\ell) = \alpha(\ell-u) - \alpha(\ell-v)$ and assuming that the magnitude and phase terms are independent, we have

$$R^{(u,v)}_{\widetilde{\mathrm{X}}\widetilde{\mathrm{X}}}(\ell) = \underbrace{\mathbb{E}\big(\mathrm{e}^{j\Delta\alpha^{(u,v)}(\ell)}\big)}_{B^{(u,v)}(\ell)}\underbrace{\mathbb{E}\big(A(\ell-u)A(\ell-v)\big)}_{R^{(u,v)}_{\mathrm{AA}}(\ell)}, \tag{5.29}$$

where $R^{(u,v)}_{\mathrm{AA}}(\ell)$ denotes the $(u,v)^{\mathrm{th}}$ element of the magnitude's autocorrelation matrix. Since the phase difference $\Delta\alpha^{(u,v)}(\ell)$ is a directional random variable, $B^{(u,v)}(\ell)$ depends on the angular spread of $\Delta\alpha^{(u,v)}(\ell)$, which in turn is determined by the accuracy and appropriateness in terms of model fitness of the phase decomposition in Section 5.3.3.

In order to draw further conclusions and to represent the aforementioned uncertainties, we model the phase difference $\Delta\alpha^{(u,v)}(\ell)$ as a random variable that follows a wrapped Gaussian distribution, parameterized by mean direction $\mu^{(u,v)}(\ell)$ and standard deviation $\sigma^{(u,v)}(\ell)$ [160]

$$p\big(\Delta\alpha^{(u,v)}(\ell); \mu^{(u,v)}(\ell), \sigma^{(u,v)}(\ell)\big) = \frac{1}{\sqrt{2\pi}\sigma^{(u,v)}(\ell)} \sum_{i=-\infty}^{\infty} \mathrm{e}^{-\frac{(\Delta\alpha^{(u,v)}(\ell)-\mu^{(u,v)}(\ell)+2\pi i)^2}{2\sigma^{(u,v)}(\ell)^2}}. \tag{5.30}$$

The initial phase can be approximated to be constant in case the propagation model in Section 5.3.3 fits the observed signal well. Hence, the mean direction $\mu^{(u,v)}(\ell)$ is zero, as Equation (5.30) essentially represents the uncertainty about the harmonic propagation model. Further, from [160] we know that in this case, the first circular moment, i.e. the circular mean, is given by[18]

$$\begin{aligned} B^{(u,v)}(\ell) &= \mathbb{E}\big(\mathrm{e}^{j\Delta\alpha^{(u,v)}(\ell)}; \mu^{(u,v)}(\ell) = 0, \sigma^{(u,v)}(\ell)\big)\big) \\ &= \mathrm{e}^{-\frac{\sigma^{(u,v)}(\ell)^2}{2}}. \end{aligned} \tag{5.31}$$

Here, the circular mean is real-valued, bounded to the interval $[0, 1]$, and it scales the magnitude autocorrelation in Equation (5.29) according to the angular spread of the phase difference. If the variance of $\Delta\alpha^{(u,v)}(\ell)$ is small, $\left|B^{(u,v)}(\ell)\right|$ is close to one, i.e. the overall correlation is high. For large variances, the wrapped normal distribution approaches the uniform distribution and $\left|B^{(u,v)}(\ell)\right|$ is close to zero. We illustrate this scaling in Figure 5.6.

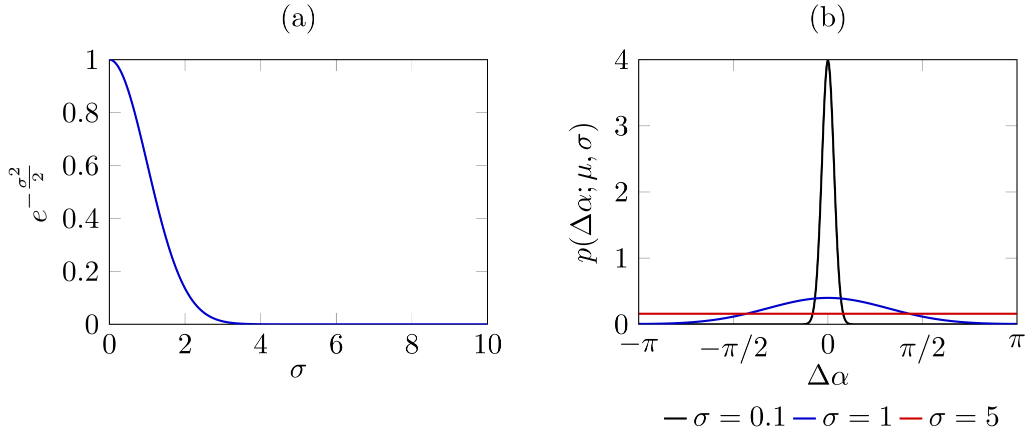

Figure 5.6: (a) The circular mean from Equation (5.31) as a function of the wrapped normal distribution's standard deviation. (b) The wrapped normal distribution for three different standard deviations.

The circular moments can be approximated by the sample mean of the respective quantity, implemented by e.g. a recursive averaging strategy. Hence, we linearly combine an instantaneous estimate of the phase difference with previous circular mean estimates as follows

$$\hat{B}^{(u,v)}(\ell) = \alpha_{\mathrm{B}}\hat{B}^{(u,v)}(\ell-1) + (1-\alpha_{\mathrm{B}})\mathrm{e}^{j\widehat{\Delta\alpha}^{(u,v)}(\ell)}, \tag{5.32}$$

where $\alpha_{\mathrm{B}} \in [0, 1]$.

In order to obtain $\widehat{\Delta\alpha}^{(u,v)}(\ell)$, we need estimates of the previous phase-values, which are easily obtained from previous speech estimates and an estimate of the initial phase of the current frame. Under the given statistical model, the MMSE estimate of the initial phase is given by the phase of the noisy DFT coefficient [15]. This means that we obtain

[18] See Appendix C for details on circular moments.

$\widehat{\Delta\alpha}^{(u,v)}(\ell)$ as follows

$$\widehat{\Delta\alpha}^{(u,v)}(\ell) = \begin{cases} \angle\Big(\frac{Y(\ell-u)}{\hat{X}_{\mathrm{up}}(\ell-v)}\frac{1}{\Delta\Psi^{(u,v)}(\ell)}\Big) & \text{if } u = 0, \\ \angle\Big(\frac{\hat{X}_{\mathrm{up}}(\ell-u)}{\hat{X}_{\mathrm{up}}(\ell-v)}\frac{1}{\Delta\Psi^{(u,v)}(\ell)}\Big) & \text{else.} \end{cases} \tag{5.33}$$

Since according to Equation (5.31), it follows that the expected value of the circular mean's angle is 0, we are only interested in the magnitude of Equation (5.32). Hence, we define the matrix $\widehat{\mathbf{B}}(\ell)$ with $(u,v)^{\mathrm{th}}$ entry $|\hat{B}^{(u,v)}(\ell)|$, as well as the vectors $\mathbf{r}_{\mathrm{AA}}(\ell)$ and $\widehat{\mathbf{b}}(\ell)$, with u^{th} entry $R_{\mathrm{AA}}^{(u,0)}(\ell)$ and $|\hat{B}^{(u,0)}(\ell)|$, respectively. By doing so, we can rewrite Equation (5.20) decomposed into a real-valued part, representing the magnitude of the coefficients in $\hat{\mathbf{a}}(\ell)$, and the diagonal matrix $\widehat{\mathbf{\Psi}}(\ell)$ that represents their phase only, i.e.

$$\hat{\mathbf{a}}(\ell) = \widehat{\mathbf{\Psi}}(\ell)\big(\mathbf{R}_{\mathrm{AA}}(\ell) \circ \widehat{\mathbf{B}}(\ell)\big)^{-1}(\mathbf{r}_{\mathrm{AA}}(\ell) \circ \widehat{\mathbf{b}}(\ell)), \tag{5.34}$$

where "$\circ$" denotes the Hadamard product.

As already discussed, the circular mean length scales the entries of $\mathbf{R}_{\mathrm{AA}}(\ell)$ and $\mathbf{r}_{\mathrm{AA}}(\ell)$. There exist various strategies to estimate $\mathbf{R}_{\mathrm{AA}}(\ell)$ and $\mathbf{r}_{\mathrm{AA}}(\ell)$, e.g. codebook-based approaches similar to [161, 162] could be applied. Considering Figure 5.6, we note that already for standard deviations of $\sigma \geq 3$, the circular mean is close to zero, dominating the outcome of Equation (5.34), which is why we simplify Equation (5.34) as follows

$$\hat{\mathbf{a}}(\ell) = \widehat{\mathbf{\Psi}}(\ell)\widehat{\mathbf{B}}(\ell)^{-1}\widehat{\mathbf{b}}(\ell). \tag{5.35}$$

This means that the complex-valued Kalman filter transition model is essentially a function of the angular spread. If the sinusoidal model does not fit, the circular mean becomes small [7] and successive coefficients are processed independently.

5.5.2 Further Algorithmic Details

So far, we assumed the process variance $\sigma^2_{x,\mathrm{in}}(k,\ell)$ to be known. However, this prerequisite is not realistic since the variance needs to be estimated from the noisy observation. We obtain an estimate of the innovation process variance by adapting the decision-directed approach to the given signal model, resulting in

$$\hat{\sigma}^2_{x,\mathrm{DD,in}}(k,\ell) = \alpha_{\mathrm{DD}}\hat{\sigma}^2_{x,\ell-1,\mathrm{in}}(k,\ell) + (1-\alpha_{\mathrm{DD}})\max[\hat{\sigma}^2_{x,\mathrm{ML,in}}(k,\ell), 0], \tag{5.36}$$

with the ML estimate

$$\hat{\sigma}^2_{x,\mathrm{ML,in}}(k,\ell) = |Y(k,\ell) - \hat{X}_{\mathrm{prop}}(k,\ell)|^2 - \hat{\sigma}^2_d(k,\ell), \tag{5.37}$$

as well as

$$\hat{\sigma}^2_{x,\ell-1}(k,\ell) = |\hat{X}_{\text{in}}(k,\ell-1)|^2. \tag{5.38}$$

Note that the accuracy of the noise PSD estimate together with the propagated signal components will impact on the accuracy of the innovation variance estimate. Since the DFT length is not constant across time, we face the problem of birth and death processes within subbands. To cope with this problem, we set any estimate of speech signal parameters to zero for all $k \in \{k|Y(k,\ell) = 0\}$.

To avoid outliers, it is common in speech enhancement, to floor the effectively applied noise suppression by a minimum gain G_{min}. In our case, we combine the resulting magnitude estimate with the inherent Kalman filter phase estimate, yielding the final speech coefficient estimate

$$\hat{X}(k,\ell) = \max\left[|\hat{X}_{\text{up}}(k,\ell)|, G_{\text{min}}|Y(k,\ell)|\right] \mathrm{e}^{j\angle\hat{X}_{\text{up}}(k,\ell)}. \tag{5.39}$$

The algorithmic in Algorithm 1 summarizes PACO by means of pseudo code.

Algorithm 1 Pseudo code depicting the individual algorithm steps of PACO.

```
After initialization, each signal frame $y_{\text{t}}(n,\ell)$ is processed as follows.
1: Estimate fundamental frequency from $y_{\text{t}}(n,\ell)$.
2: Estimate phase progression $\widehat{\psi}_h(\ell)$ recursively using Equation (3.17).
3: Compute $Y(k,\ell)$ using Equation (5.3).
4: Compute $\hat{\mathbf{a}}(\ell)$ using Equation (5.35), Equation (5.32), and Equation (5.33).
5: Propagate previous state estimates following Equation (5.13).
6: Estimate $\sigma^2_{x,\text{in}}(k,\ell)$ using Equation (5.36) and $\hat{\sigma}^2_d(k,\ell)$ following [18].
7: Propagate error covariance Equation (5.14).
8: Compute Kalman gain Equation (5.15) and obtain $\hat{X}_{\text{in}}(k,\ell)$ Equation (5.16).
9: Update state estimate Equation (5.17) and error covariance Equation (5.18).
10: Obtain $\hat{X}(k,\ell)$ by flooring $\hat{X}_{\text{up}}(k,\ell)$ Equation (5.39).
11: Apply inverse DFT and windowing Equation (5.7), perform overlap-add.
```

5.6 Evaluation

5.6.1 Benchmark Algorithms

Since we want to illustrate the benefits resulting from (i) the PASTFT and (ii) taking into account temporal correlation by Kalman filtering, as benchmarks we chose

- the Wiener filter, which does not exploit any of the two extensions

- the COmplex-valued Kalman (COKA) filter in the STFT domain without phase decomposition proposed by Esch in [152], incorporating temporal correlation in the STFT domain,[19]
- the Wiener filter applied in the PASTFT, referred to as PADDi algorithm [3],
- the multiple frame MVDR filter proposed in [62, 63], incorporating the temporal correlations by linearly filtering STFT subbands.

We summarize the chosen benchmarks in Table 5.1. This table highlights distinguishing features of all reported algorithms, namely the domain and if and how each algorithm incorporates knowledge about temporal correlation of DFT bins. The COKA algorithm was implemented following [152] with a frame length of 20 ms and an AR model order of 3. For the other algorithms, the frame length was chosen with 32 ms and we used a square-root hamming window for all algorithms in all experiments. The works in [62, 63] do not address a fully blind implementation of the MVDR filter, i.e. the noise inter-frame correlation is assumed to be given or estimated from isolated noise data. In order to have a fair comparison to the other algorithms, we used the noise PSD estimate obtained by [18] only, i.e. neglected the inter-frame correlations of the noise signal.

For computing the PASTFT, we set $f_{0,\mathrm{min}} = 90\,\mathrm{Hz}$ and $f_{0,\mathrm{max}} = 350\,\mathrm{Hz}$, respectively. According to Equation (5.6), this choice results in a zero padding factor of $K = 12$. We set $G_{\mathrm{min}} = 0.1$ and $\alpha_{\mathrm{DD}} = 0.98$ for all algorithms. If not mentioned otherwise, the overlap was set to 75%. We chose $\alpha_{\mathrm{B}} = 0.7$ and the AR model order for the PACO algorithm was set to 1 in order to keep the computational effort low and since in terms of speech enhancement performance the differences were rather small between the different model orders. As fundamental frequency estimator for computing the PASTFT we used the PEFAC algorithm from [141].

All experiments were conducted with Matlab® on a workstation equipped with an Intel® Xeon® E5 CPU. With this setup, together with the above explained settings of the algorithms, we measured a real-time factor [163] of 0.80 for PACO, 0.22 for the Wiener filter in the PASTFT, 0.23 for COKA [152], and 4.90 for the MVDR approach [63]. For comparison, the classical Wiener filter results in a real-time factor of 0.03 on this hardware.

5.6.2 Proof-of-concept

In Figure 5.7 we display spectrograms of clean, noisy, and enhanced speech files.

The following observations are made from Figure 5.7:

- As indicated by rectangle 1 in Figure 5.7, low frequency speech components are successfully recovered by the PACO method.

[19] In [152], the AR parameters are estimated from previous, enhanced speech data.

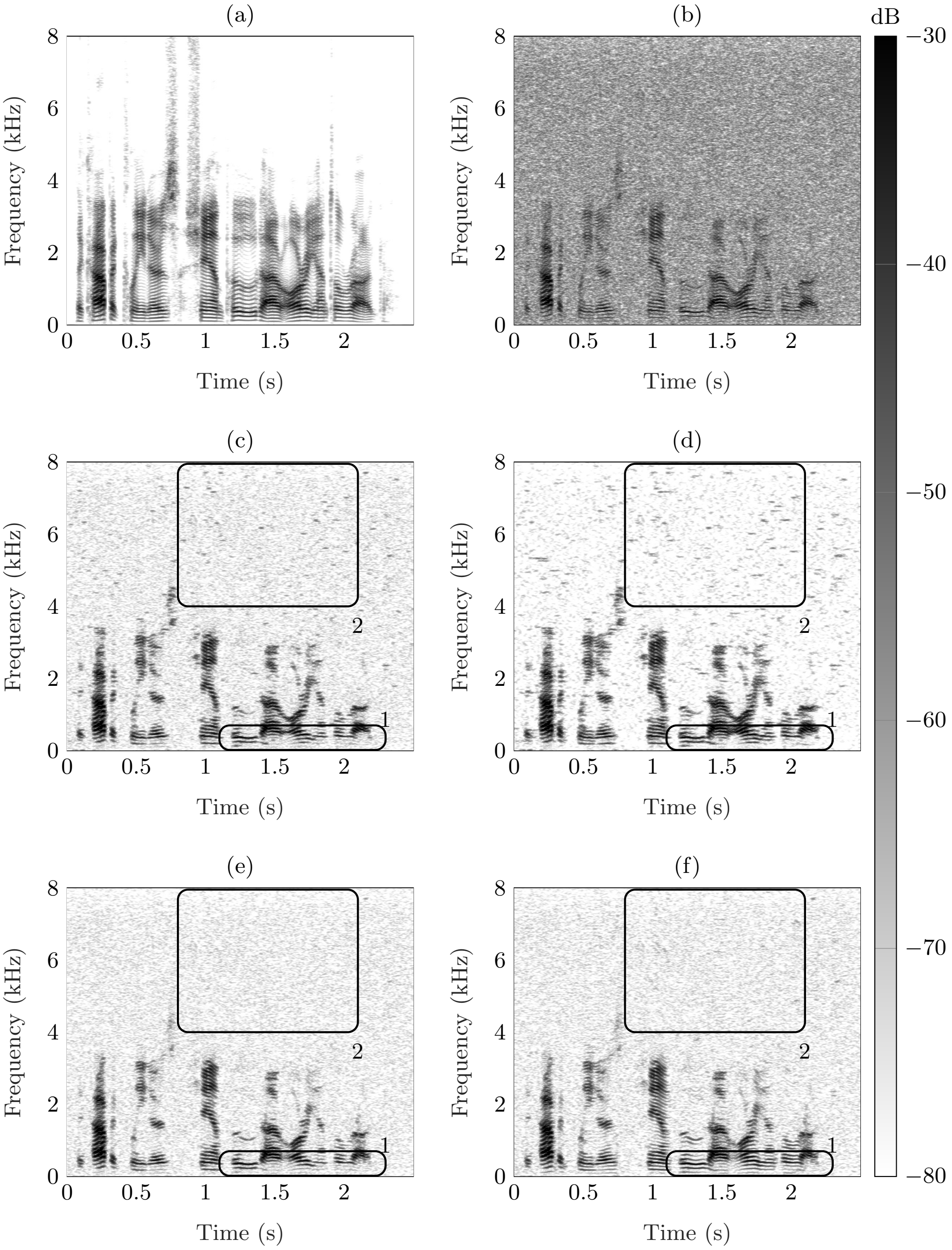

Figure 5.7: Proof-of-concept: Male speaker uttering *The carpet cleaners shampooed our oriental rug.* from the TIMIT core database [139] mixed with white noise, SNR = 5 dB. (a) Clean speech, (b) noisy speech, (c) Wiener Filter, (d) COKA [152], (e) WF + PADDi [3], (f) PACO, proposed.

Table 5.1: Key properties of the tested algorithms.

Method	Domain	Temporal Correlation
WF	STFT	No
WF + PADDi [3]	PASTFT	No
MVDR [63]	STFT	Yes, MVDR
COKA [152]	STFT	Yes, AR
PACO	PASTFT	Yes, AR + Harmonic

- Artifacts such as isolated spectral peaks, traditionally associated to musical noise, are successfully suppressed when processing the noisy speech signal via the PASTFT (rectangle 2 in Figure 5.7).

5.6.3 Objective Evaluation of Speech Enhancement Performance

Databases and Evaluation Metrics

For all simulations in this section we used the test set of the TIMIT core database [139], consisting of 192 utterances. All signals were sampled at 16 kHz and we added one second of noise only at the beginning and at the end of the individual signals. The noisy test data was obtained by mixing the speech signals with noise at SNRs between -5 dB and 15 dB in 5 dB steps following the mixing convention recommended in [164]. The specific noise types chosen in the experiments are white noise, factory noise, and babble noise from the NOISEX-92 database [142], and rain noise from [143]. This selection includes stationary, impulsive, speech-like, and highly non-stationary noise types.

As evaluation criteria, we chose PESQ [145] and STOI [146] as instrumental predictors of perceived speech quality and speech intelligibility, respectively. We report the respective results in terms of Δ improvement over the noisy signal. To obtain further insights into the algorithms' mechanisms, we conducted a black box analysis of all algorithms compared in this chapter. Hence, additionally to ΔPESQ and ΔSTOI we report the weighted log-kurtosis ratio (WLKR) as an indicator of the amount of musical noise in the processed signal and an analysis of $\widetilde{\text{PESQ}}$ across NA_{seg}.[20]

Impact of Fundamental Frequency Estimation Errors

The proposed algorithms heavily depend on fundamental frequency estimates. In order to evaluate the algorithms' sensitivity to fundamental frequency estimation errors, we compare the performance of fully blind implementations of PACO and PADDi to a

[20] In Appendix B we provide details on the evaluation metrics used in this thesis.

scenario, where the fundamental frequency is estimated from isolated, clean speech data. We refer to the latter case as "f_0-oracle" scenario, representing a performance upper bound w.r.t. the chosen fundamental frequency estimator.

Figure 5.8 illustrates the outcome of this experiment, averaged over all noise scenarios. We note that more accurate fundamental frequency estimation mainly impacts on STOI and the achievable noise attenuation. PACO comes with less noise attenuation than the Wiener filter in combination with PADDi in both scenarios but slightly improves ΔPESQ and $\widetilde{\text{PESQ}}$. Further, both perform close in WLKR and in terms of ΔSTOI, the phase-sensitive PACO outperforms its counterpart in the blind as well as in the f_0-oracle scenario.

We note that the phase-aware approach PACO is naturally more sensitive to fundamental frequency estimation errors than PADDi, which is also in line with the observations made in Section 4.7.3 and [24,27,28]. This also indicates the potential in terms of joint quality and intelligibility improvement by phase-sensitive approaches.

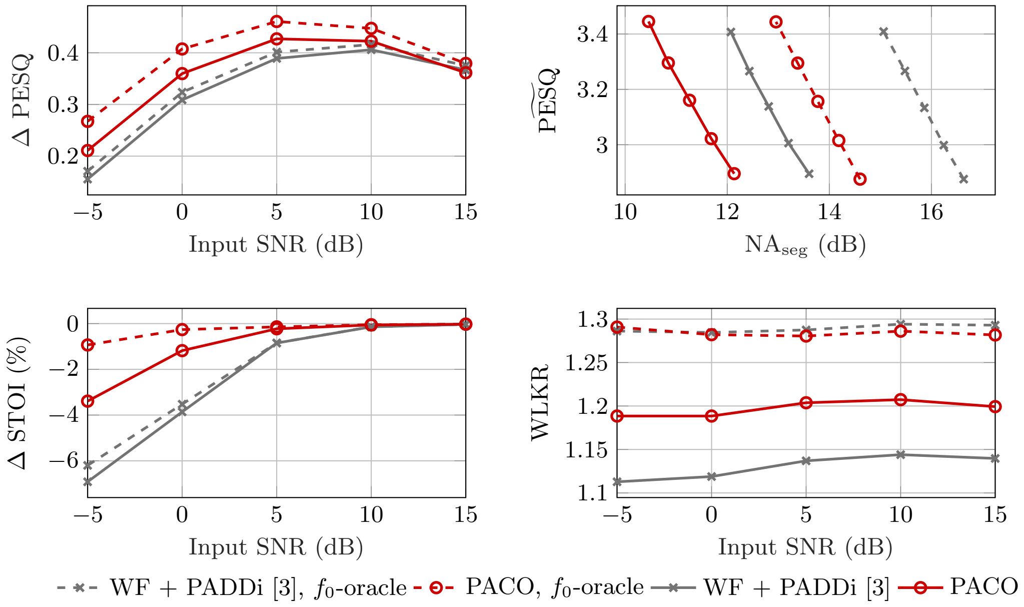

Figure 5.8: Speech enhancement performance of PACO and PADDi with f_0 estimated from isolated speech data, averaged over all noise scenarios.

Full Evaluation Results

From the averaged results in Figure 5.9 we conclude that:

- The two pitch-adaptive approaches PACO and PADDi outperform their non-pitch-adaptive counterparts as well as the other benchmarks in ΔPESQ, ΔSTOI, and WLKR over the full evaluated SNR range.

- Except for ΔPESQ at SNR 15 dB, the Kalman filters in both the STFT and PASTFT outperform their Wiener filter counterparts in the respective domain.

- The $\mathrm{NA}_{\mathrm{seg}}$ results illustrate that the noise attenuation is on average decreased by taking into account inter-frame correlations. However, at the same time, we also note less speech distortions, indicated by increased $\widetilde{\mathrm{PESQ}}$. This effect is particularly pronounced when comparing the Wiener filter and the COKA algorithm.

- The quality improvement of all reported methods comes at the cost of intelligibility degradation compared to the noisy, unprocessed signal. However, PACO and PADDi improve the intelligibility compared to their non pitch-adaptive counterparts. Recent data driven approaches such as deep neural network-based methods [165,166] enable to jointly improve speech quality and intelligibility. However, this work is focused on non data driven methods that on the one hand do not rely on training data and lengthy training phases but which are known to balance a trade-off between quality improvement and intelligibility [167] on the other hand.

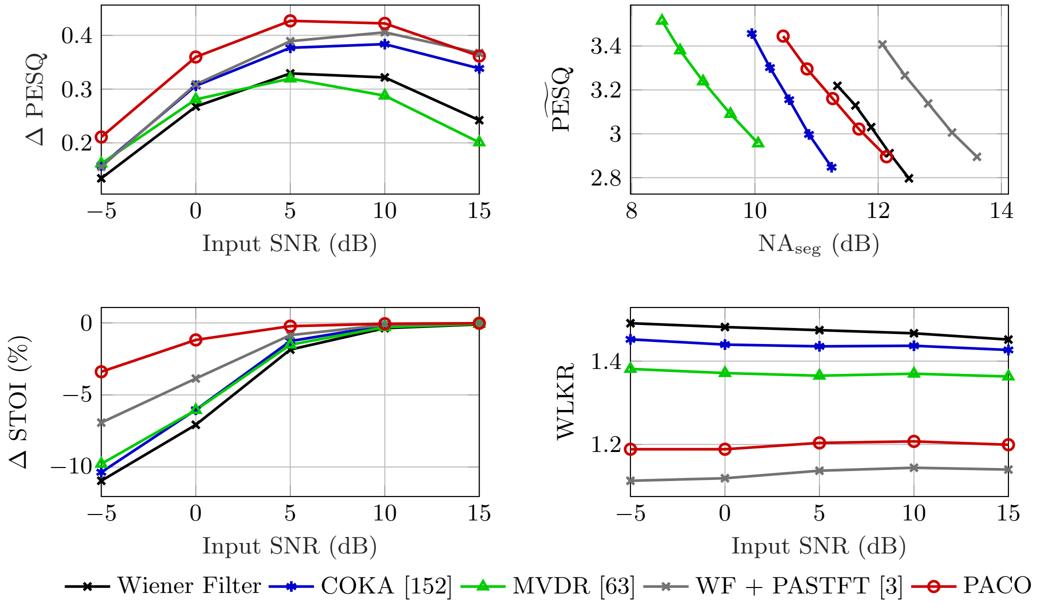

Figure 5.9: Speech enhancement performance, averaged over all noise scenarios.

Figure 5.10 to Figure 5.13 report the detailed results for all noise scenarios and metrics.

- PACO, with few exceptions, consistently outperforms the benchmarks in terms of ΔPESQ and ΔSTOI as can be seen in Figure 5.10 and Figure 5.11, respectively.

- Further, in Figure 5.12 we note that in terms of noise reduction, the PADDi approach in conjunction with the Wiener filter is highly effective, however it comes with slightly reduced PESQ compared to PACO, COKA, and the MVDR approach.
- In Figure 5.13, we also note that the Wiener filter comes with more musical noise artifacts than all other tested algorithms, which was expected. PADDi and PACO are among the top performing methods for all noise scenarios.

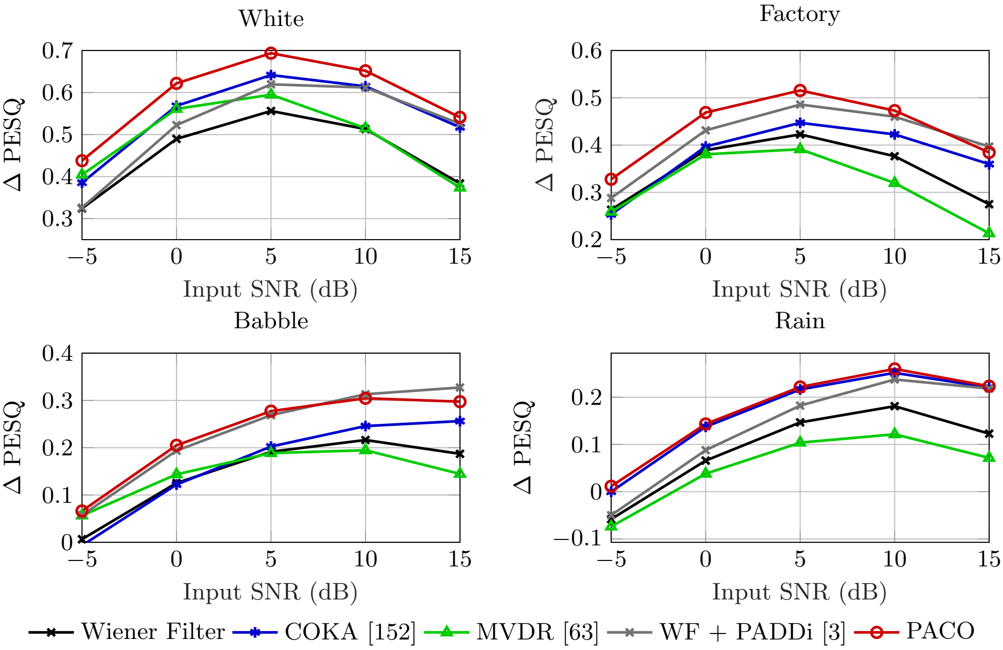

Figure 5.10: Speech enhancement performance in terms ΔPESQ improvement over the noisy signal.

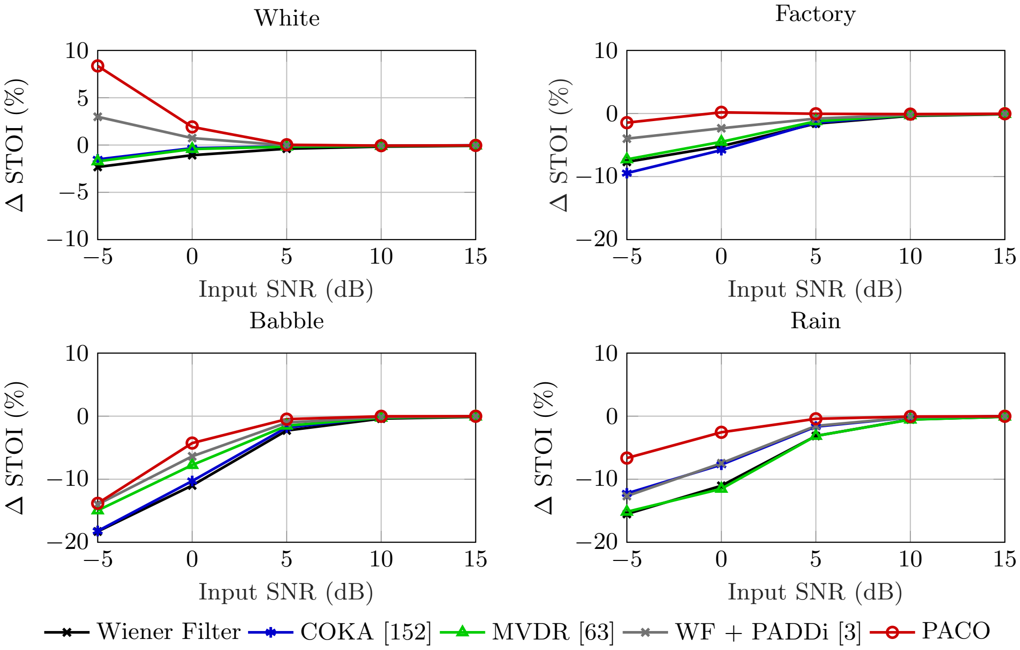

Figure 5.11: Speech enhancement performance in terms ΔSTOI improvement over the noisy signal.

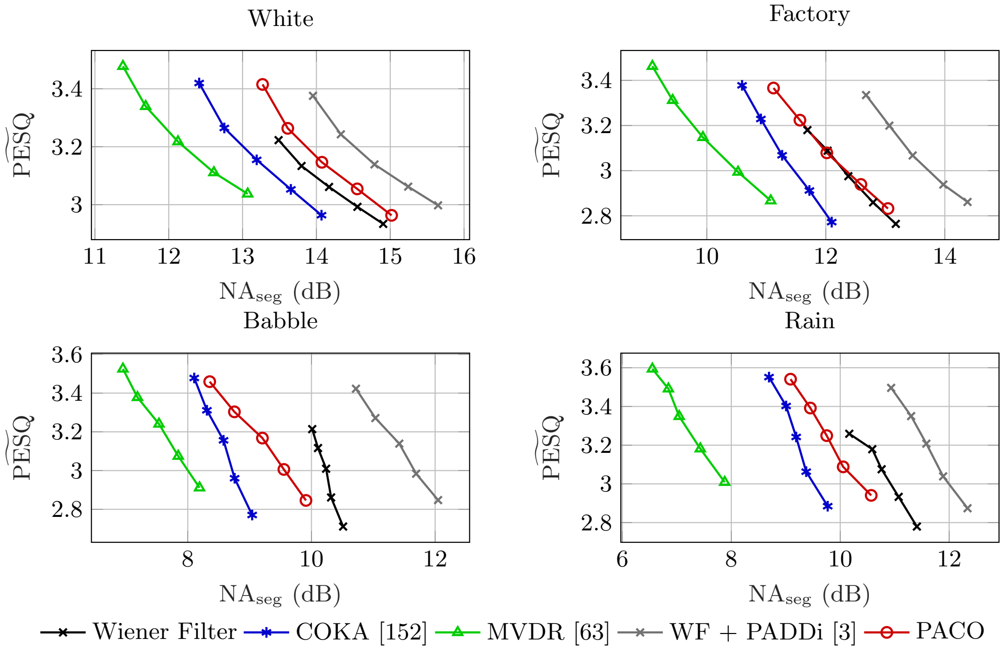

Figure 5.12: Black box analysis of all algorithms in terms of $\widetilde{\text{PESQ}}$ and NA$_{\text{seg}}$.

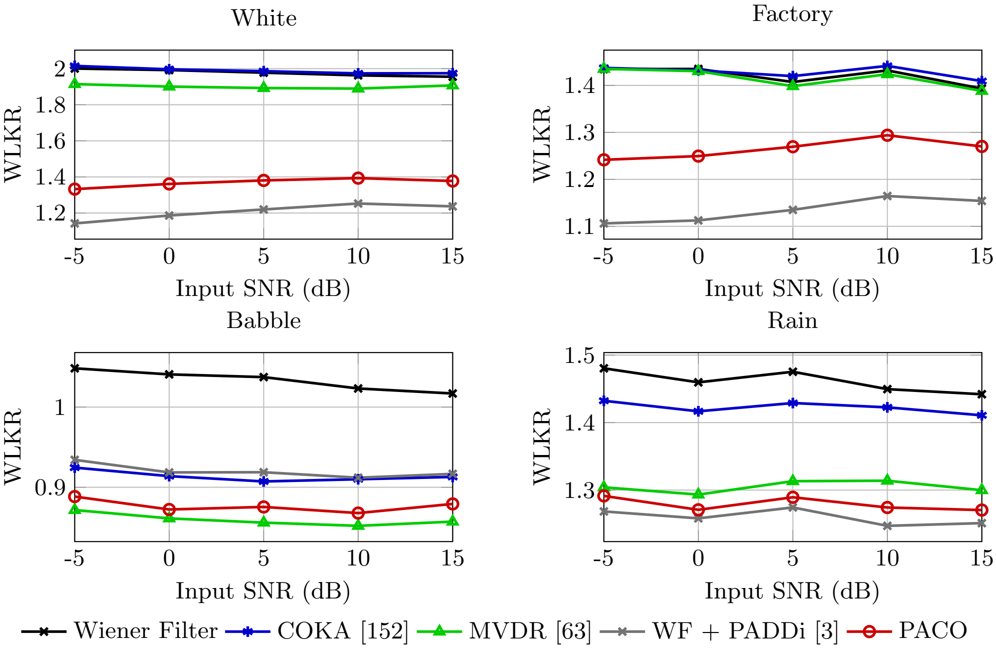

Figure 5.13: Speech enhancement performance in terms of WLKR.

5.6.4 Subjective Evaluation of Speech Enhancement Performance

For the evaluation of the proposed method in terms of subjective listening quality, we chose babble and factory noise from [142] and mixed it with 4 randomly selected speech samples taken from the TIMIT test set [139] at SNRs 5 dB and 10 dB following [164]. Before and after each utterance, we added one second of noise only. Similar to [30], in order to minimize the effect of noise PSD initialization errors, the signals were played to the test subjects starting at 0.5 seconds. The noise segments themselves were selected randomly. All test signals can be found on [168].

The algorithms were presented to the participants by means of a pairwise preference test, i.e. the test subjects were asked to select their preference (better/worse) between two presented methods in terms of (i) noise reduction, (ii) speech distortions, and (iii) overall quality. The signals presented were the (i) noisy signal, (ii) the noisy signal processed by the LSA estimator [90] as a well known benchmark, (iii) the MVDR filter from [63], implemented as described in Section 5.6.3 and (iv) PACO. Hence, 24 pairwise comparisons were presented to each participant. In addition to the samples under test, the clean speech samples were presented as a reference to the participants. We randomized the sequence of the presented samples and the test subjects were allowed to listen to each sample as often as they wanted. We used the framework provided from [169] to set up the graphical user interface and the composition of the test data. The listening test was conducted in a quite room and the signals were presented to the participants using Beyerdynamic DT-770 Pro headphones.

The listening test was performed with 15 self-reported normal-hearing participants, aged between 22 and 38. The results of the listening test, averaged over all noise scenarios, are depicted in Figure 5.14, the detailed results are reported in Figure 5.15. The results illustrate that PACO improves the noise reduction capabilities compared to the benchmark algorithms, while the speech quality is preserved. In this context, it is important to note that the MVDR method [63], as also reported in [150], sounds slightly reverberated when implemented as a blind approach, which is the reason why the participants preferred the other methods in terms of speech and overall quality.

We performed a Friedman Test with Bonferroni correction to draw conclusions about the significance of the difference between the methods under test. We assume significance for a p-value of $p < 0.05$. In terms of noise reduction, the preference of PACO over all reference signals was significant. For speech quality, no significant preference of one method over the others can be observed, while in terms of overall quality, the PACO outperforms the MVDR approach significantly.

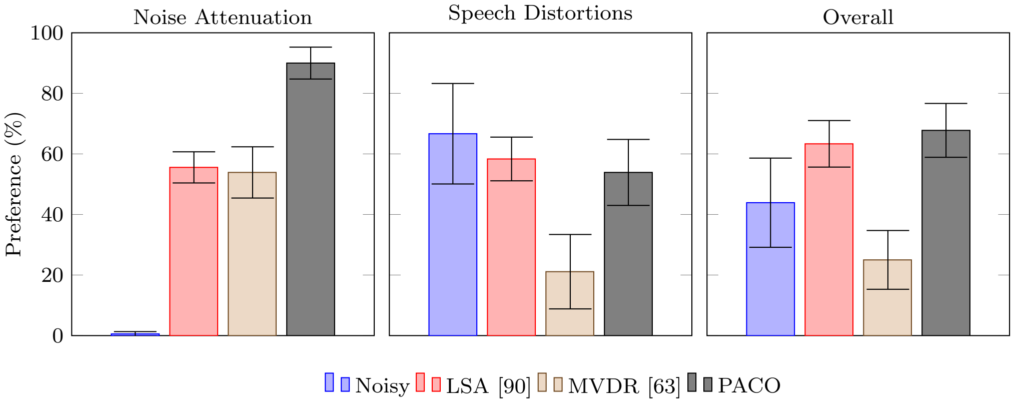

Figure 5.14: Results of the subjective listening test in terms of preference ratings averaged over all noise scenarios together with the corresponding standard deviations.

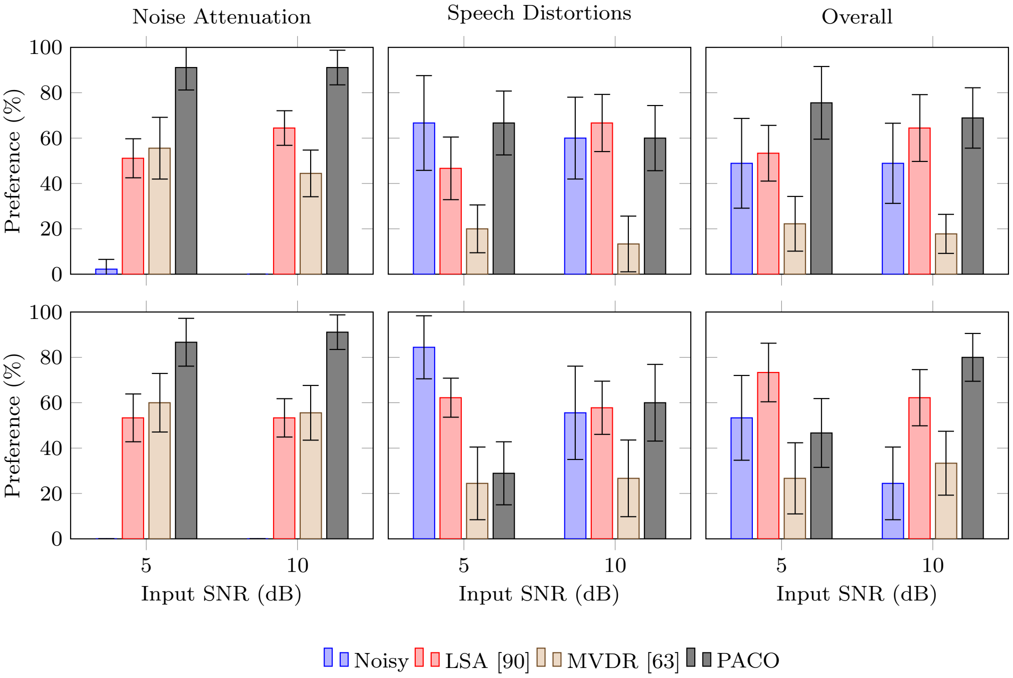

Figure 5.15: Detailed results of the subjective listening test in terms of preference ratings averaged over all SNRs together with the corresponding standard deviations. The upper row corresponds to babble noise, the bottom row to factory noise.

5.7 Conclusion

This chapter presents an approach for incorporating inter-frame correlations into speech enhancement. More specifically, correlations that result from the harmonic plus noise model are exploited to facilitate the estimation of the state transition model of a complex-valued Kalman filter. The derived estimator is simple to implement and solely relies on the statistics of the decomposed spectral phase. In its approximations, this derivation relies on the PASTFT, which is a pitch-adaptive STFT analysis stage that aligns DFT bins that are dominated by the same harmonic with each other.

The PASTFT provides the potential for higher noise attenuation while preserving the level of speech distortions compared to the standard STFT. Further, it decreases the amount of musical noise in the processed signal. By decomposing the spectral phase into an initial and a progression part, we reveal significant correlation of DFT coefficients along time. We showed that in the PASTFT, successive DFT bins dominated by harmonics exhibit higher correlation along time compared to the standard STFT. Taking into account these inter-frame correlations by Kalman filtering comes with improvement in PESQ and STOI compared to the benchmark methods. We further conducted a subjective listening test that indicates that the proposed algorithm improves noise suppression while preserving the level of speech distortions.

6 Exploiting Phase Structure along Frequency for Single-Channel Speech Enhancement

This chapter investigates single-channel speech enhancement algorithms that take into account dependencies w.r.t. frequency. This chapter is based on the author's insight that the problem of estimating the phase of inter-frequency covariances is a key problem of multidimensional STFT speech estimators. All derivations and simulations included in this chapter have been conducted by the author. Parts of the text in this chapter have been published in [4]. Further, some of the results in this chapter are part of the submitted work in [6].

6.1 Overview and Motivation

The derivations presented thus far in this thesis have been based on the assumption that the speech and noise STFT expansion coefficients are independent w.r.t. frequency. This assumption, though omnipresent in the literature, is a strong simplification since speech signal components, even if they are well-separated in frequency, share the same excitation air flow [131]. Their exact interaction is of course unknown and can be seen as a result of the excitation signal and the filtering by the vocal tract. Hence, we may assume the correlations to depend on the specific speaker, the intended speech sound, and some realization noise. However, even if we consider the speech signal as a superposition of multiple independent components, the relatively short analysis segments and the non-stationarity of speech and noise signals may also cause highly correlated frequency bins.

Assuming statistical independence between signal components from different frequency regions results in estimators that are functions of diagonal covariance matrices, i.e. that do not combine observations from different time-frequency regions. Allowing dependencies between DFT coefficients instead yields non-diagonal covariance matrices of speech and noise, where off-diagonal entries are complex-valued in general. The complex-valued nature of the off-diagonal covariance matrix elements will be of special interest in this chapter.

Given dependencies w.r.t. frequency, the joint minimization of the MMSE for all complex-valued Fourier expansion coefficients yields a multidimensional Wiener filter [170] that is a function of the speech and noise covariance matrices. In the literature, also multidimensional MMSE-STSA estimators have been investigated by Plourde and Champagne in [131,171]. Since it is difficult to solve for the optimal STSA estimator analytically, the authors combined lower and upper bounds for the estimator, resulting in a new family of MMSE-STSA estimators. The work in [131] not only considers correlations among adjacent frequency bins but also among frequency bins that are well-separated. This distinguishes it from the work in [172], where a full-band Wiener filter together with an approximation that considers only correlations among adjacent frequency bins has been proposed. Further, in [172], the covariance matrices of the speech and noise signals are assumed to be known and no explicit strategy to estimate them is presented. In [173], Momeni et al. formulated a conditional MMSE estimator that uses inter-frame as well as inter-band correlations, where only the correlations of adjacent frequency bands are exploited in the estimation procedure. A different approach was taken by Fingscheidt et al., where dependencies between frequency channels were incorporated in the *a priori* SNR estimator by modeling the *a priori* SNR as a Markov process w.r.t. frequency [60].

The multidimensional estimators presented in the literature are optimal with respect to *known* statistics of the speech and noise signal. In the case of estimators that do not consider dependencies w.r.t. frequency[21], the second order moments and the resulting estimators are real-valued. In the case of multidimensional estimators, these quantities are complex-valued in general. In the existing work on multidimensional single-channel speech estimators, the second order moment matrices are estimated by approaches similar to the DD approach for *a priori* SNR estimation. However, it is not clear if temporal averaging strategies are also applicable to estimating the complex-valued covariance in the case of multidimensional estimators. In this chapter, we investigate how this may affect a speech enhancement framework that takes into account inter-frequency correlations. More specifically we will discuss

1. the role the covariance matrix phase plays in the estimators that have been derived in the literature so far,
2. why estimation of the covariances' phase by means of traditional approaches like the DD method can be problematic,
3. why the spectral phases can be considered to be mutually independent w.r.t. frequency,
4. an approach to circumvent this problem.

[21] We will also refer to them as one-dimensional estimators.

6.2 Multidimensional Speech Estimators

Reconsidering the signal plus noise model from Equation (2.1) in vector notation we have

$$\mathbf{y}_f(\ell) = \mathbf{x}_f(\ell) + \mathbf{d}_f(\ell), \tag{6.1}$$

where the subscript f indicates frequency domain vectors of the form

$$\mathbf{y}_f(\ell) = \begin{bmatrix} Y(0,\ell) & Y(1,\ell) & \dots & Y(N_{\mathrm{DFT}}-1,\ell) \end{bmatrix}^T \in \mathbb{C}^{N_{\mathrm{DFT}}\times 1}.$$

The covariance matrices of the speech and the noise DFT coefficients, which we assume to be zero-mean, complex-valued, and circularly Gaussian distributed, are given by

$$\begin{aligned} \mathbf{R}_{\mathrm{x}}(\ell) &= \mathbb{E}\big(\mathbf{x}_f(\ell)\mathbf{x}_f^H(\ell)\big) \in \mathbb{C}^{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}}, \\ \mathbf{R}_{\mathrm{d}}(\ell) &= \mathbb{E}\big(\mathbf{d}_f(\ell)\mathbf{d}_f^H(\ell)\big) \in \mathbb{C}^{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}}. \end{aligned} \tag{6.2}$$

If we neglect interdependencies of speech components along frequency, estimators of $X(k_1,\ell)$ are a function of *one* observation $Y(k_1,\ell)$ only, neglecting any information about $X(k_1,\ell)$ that might be contained in $Y(k_2,\ell)$ for $k_2 \neq k_1$. This restriction might be a handicap, since it prevents any estimator from benefiting from a priori knowledge about the instantaneous spectral structure of the speech signal.

The MMSE-optimal estimate of $\mathbf{x}_f(\ell)$ given the speech and noise covariances is found by applying the multidimensional Wiener filter (MD-WF) [170] to the noisy observation

$$\begin{aligned} \hat{\mathbf{x}}_f(\ell) &= \underset{\hat{\mathbf{x}}_f(\ell)}{\arg\min}\, \mathbb{E}\big(\, \|\mathbf{x}_f(\ell) - \hat{\mathbf{x}}_f(\ell)\|_2^2 \,\big) \\ &= \mathbf{G}_{\mathrm{W}}(\ell)\mathbf{y}_f(\ell) \end{aligned} \tag{6.3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and the MD-WF [170] is given as

$$\begin{aligned} \mathbf{G}_{\mathrm{W}}(\ell) &= \mathbf{R}_{\mathrm{x}}(\ell)\big(\mathbf{R}_{\mathrm{x}}(\ell) + \mathbf{R}_{\mathrm{d}}(\ell)\big)^{-1} \\ &= \mathbf{R}_{\mathrm{x}}(\ell)\mathbf{R}_{\mathrm{y}}(\ell)^{-1} \in \mathbb{C}^{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}}. \end{aligned} \tag{6.4}$$

Assuming the DFT bins of both speech and noise to be independent w.r.t. frequency yields diagonal covariance matrices and hence the standard Wiener filter.

In [131], the authors derived a new family of STSA estimators, taking into account the interdependencies of the frequency channels in the STFT. The resulting multi-dimensional STSA (MD-STSA) estimator of the clean speech Fourier expansion coefficients was given by

$$
\begin{aligned}
\hat{\mathbf{x}}_{f,\mathrm{m}}(\ell) &= \underset{|\hat{\mathbf{x}}_f(\ell)|}{\arg\min}\, \mathbb{E}\big(\| |\mathbf{x}_f|\,(\ell) - |\hat{\mathbf{x}}_f(\ell)| \|_2^2 \big) \\
&= \sqrt{|\mathbf{G}_\mathrm{W}(\ell)\mathbf{y}_f(\ell)| + \gamma \mathrm{diag}\left\{\mathbf{G}_\mathrm{W}(\ell)\mathbf{R}_\mathrm{d}(\ell)\right\}} \circ \mathrm{e}^{j\angle \mathbf{y}_f(\ell)},
\end{aligned}
\tag{6.5}
$$

where $\gamma \in [0,1]$ is a mixing factor and $\mathrm{diag}\left\{\cdot\right\}$ accesses the main diagonal of a matrix.

Note that even if the speech magnitude is the quantity being estimated, the gain matrix $\mathbf{G}_\mathrm{W}(\ell)$ in Equation (6.5) is in general complex-valued. Hence, the magnitude estimates in Equation (6.5) rely on a linear combination of the *complex-valued* noisy speech coefficients rather than on the observed magnitudes alone. This is a substantial difference to the one-dimensional case and indicates that by taking spectral dependencies into account, the spectral phase comes into play, even if it is not the quantity one aims to retrieve.

6.3 The Covariance Matrix

The $\{k_1, k_2\}^\mathrm{th}$ entry of the speech covariance matrix $\mathbf{R}_\mathrm{x}(\ell)$ is

$$
\begin{aligned}
R_\mathrm{x}(k_1,k_2,\ell) &= \mathbb{E}\big(X(k_1,\ell)X^*(k_2,\ell)\big) \\
&= \mathbb{E}\big(|X(k_1,\ell)||X(k_2,\ell)|\mathrm{e}^{j(\angle X(k_1,\ell)-\angle X(k_2,\ell))}\big) \\
&= \mathbb{E}\big(A(k_1,\ell)A(k_2,\ell)\mathrm{e}^{j(\phi(k_1,\ell)-\phi(k_2,\ell))}\big).
\end{aligned}
\tag{6.6}
$$

In a practical scenario, we need to estimate $R_\mathrm{x}(k_1,k_2,\ell)$ from the noisy data. It is common in speech enhancement to estimate second moments by using recursive averaging strategies among which the DD approach [15], which was adapted to full covariance matrix estimation in [131], is popular for speech variance estimation. Such temporal averaging strategies require an approximately stationary signal. This might be a reasonable assumption for the magnitude terms in Equation (6.6), but the phase differences fluctuate from one frame to the next, which may yield the interpretation that the underlying process is non-stationary, prohibiting the use of temporal averaging in this case. We illustrate this in Figure 6.1. If the covariance matrix is estimated by means of a recursive smoothing approach, the fluctuations in the phase difference may yield underestimation of the covariance's absolute value due to the incoherent linear combination of successive, complex-valued estimates. Consequently, the resulting gain may substantially deviate from the theoretically optimal $\mathbf{G}_\mathrm{W}(\ell)$ in both absolute value and phase, as we show in Section 6.4.1.

Alternatively, we could model the phase differences as a stationary process with high variance. Assuming the magnitude and phase terms in Equation (6.6) to be independent,

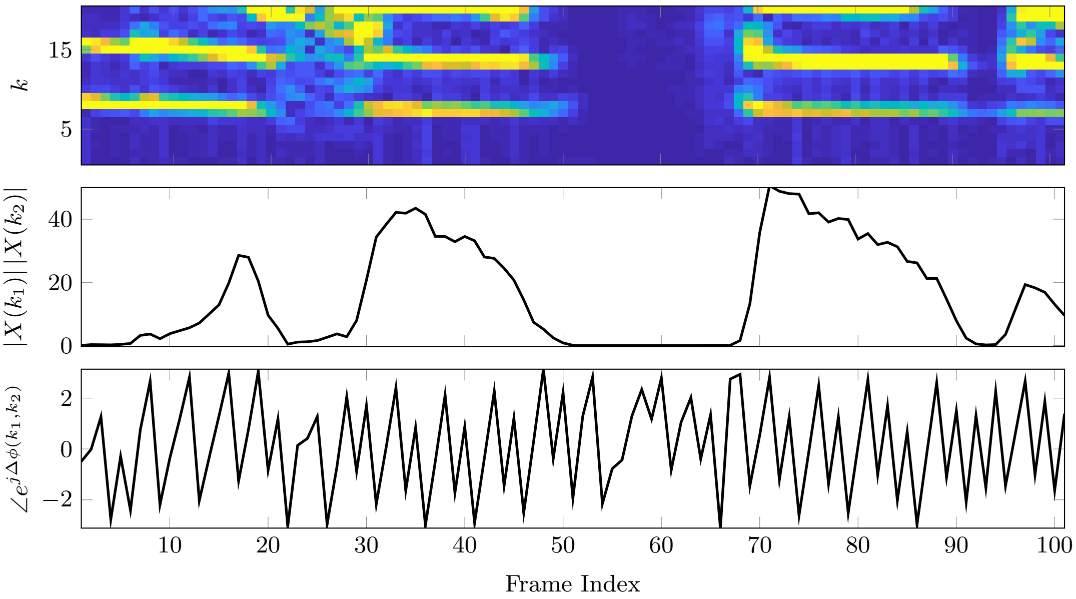

Figure 6.1: Spectrogram snippet (upper panel) of a speech file taken from TIMIT [139] together with the corresponding instantaneous values of magnitude (middle panel) and phase (lower panel) of $X(k_1,\ell)X^*(k_2,\ell)$ with $k_1 = 7$, corresponding to the first harmonic and $k_2 = 14$ corresponding to the second harmonic in the spectrogram.

implying they are uncorrelated, yields

$$\begin{aligned} R_{\mathrm{x}}(k_1,k_2,\ell) &= \mathbb{E}\big(A(k_1,\ell)A(k_2,\ell)\mathrm{e}^{j(\phi(k_1,\ell)-\phi(k_2,\ell))}\big) \\ &= \mathbb{E}\big(A(k_1,\ell)A(k_2,\ell)\big)\mathbb{E}\big(\mathrm{e}^{j(\phi(k_1,\ell)-\phi(k_2,\ell))}\big) \\ &= R_{\mathrm{A}}(k_1,k_2,\ell)\mathbb{E}\big(\mathrm{e}^{j\Delta\phi(k_1,k_2,\ell)}\big), \end{aligned} \tag{6.7}$$

where we define the phase factor covariance as

$$R_\phi(k_1,k_2,\ell) = \mathbb{E}\big(\mathrm{e}^{j\Delta\phi(k_1,k_2,\ell)}\big). \tag{6.8}$$

As explained above, we may assume that the phase difference for $k_1 \neq k_2$ exhibits a large variance, i.e. is uniformly distributed. A uniformly distributed phase is characterized by a circular variance equal to one[22], hence we have

$$\sigma_c^2(k_1,k_2,\ell) = \begin{cases} 1 & \text{if } k_1 \neq k_2, \\ 0 & \text{if } k_1 = k_2 \end{cases} \tag{6.9}$$

[22] See Appendix C.

or equivalently, using

$$\sigma_c^2(k_1,k_2,\ell) = 1 - |R_\phi(k_1,k_2,\ell)| \tag{6.10}$$

we have

$$R_\phi(k_1,k_2,\ell) = \begin{cases} 0 & \text{if } k_1 \neq k_2, \\ 1 & \text{if } k_1 = k_2. \end{cases} \tag{6.11}$$

Since

$$R_\text{x}(k_1,k_2,\ell) = R_\text{A}(k_1,k_2,\ell)R_\phi(k_1,k_2,\ell), \tag{6.12}$$

Equation (6.11) renders Equation (6.6) equal to zero for $k_1 \neq k_2$, implying a diagonal covariance matrix. This in turn would mean that the diagonal Wiener Filter is the optimal gain, since any dependencies in the magnitude are masked by the independence of the phase w.r.t. frequency. Later in this chapter, we derive a speech enhancement framework that takes into account the above observations by relying on the following two assumptions.

Assumption 1. *The amplitude and phase terms in Equation* (6.6) *are independent.*

Assumption 2. *The spectral amplitude exhibits significant dependencies w.r.t. frequency, while the spectral phase does not.*

6.3.1 Empirical Evidence for the Statistical Model

This section provides empirical evidence for Assumption 1 and Assumption 2. In order to quantify the validity of the two assumptions, we first need a suitable measure of mutual independence between random variables. We will then use this measure by evaluating it for real speech data.

The Symmetric Uncertainty Coefficient

The symmetric uncertainty coefficient (SUC) has been used in [174] to quantify the independence between real and imaginary parts as well as amplitude and phase of STFT coefficients. The SUC of two discrete RVs X and Y is defined as [175]

$$U(X,Y) = 2\frac{H(X)+H(Y)-H(X,Y)}{H(X)+H(Y)}, \tag{6.13}$$

where $H(\cdot)$ is the entropy, given by [176]

$$H(X) = \sum_{X \in \mathcal{X}} p\,(x) \log_2\bigl(p\,(x)\bigr), \tag{6.14}$$

where $\mathcal{X}$ is the alphabet of the RV X with probability mass function $p\,(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$. The numerator of Equation (6.13) corresponds to the mutual information $I(X;Y)$ of X and Y. Hence, the SUC can be interpreted as a normalized mutual information with $U(X,Y)$ being close to zero indicating statistical independence of X and Y, while values closer to one indicate dependencies. This can be easily shown, as if X and Y are independent we have the joint entropy

$$H(X,Y) = H(X) + H(Y|X) = H(X) + H(Y), \tag{6.15}$$

which, by inserting Equation (6.15) into Equation (6.13), yields $U(X,Y) = 0$.

In the following experiments, we use histograms obtained from clean speech data sampled at 16 kHz from the TIMIT database test set [139] as empirical probability mass functions. The frame length was chosen to be 32 ms, the overlap was set to 75 %, with a square-root Hamming window. These settings were chosen such that the analysis stage corresponds to the test setup that is used throughout this thesis.

Independence of Magnitude and Phase Terms

To assess the validity of Assumption 1, we evaluate $U(R_{\mathrm{A,inst}}(k_1,k_2,\ell), \Delta\phi(k_1,k_2,\ell))$, i.e.

$$R_{\mathrm{A,inst}}(k_1,k_2,\ell) = A_{\mathrm{N}}(k_1,\ell)A_{\mathrm{N}}(k_2,\ell), \tag{6.16}$$

with subscript N indicating amplitudes that are normalized to the frame signal level similar to Section 4.4.2, i.e.

$$A_{\mathrm{N}}(k,\ell) = \frac{A_{\mathrm{N}}(k,\ell)}{\sum_{k'=1}^{N_{\mathrm{DFT}}/2+1} A(k',\ell)^2}. \tag{6.17}$$

In a second step, we divide the range of $R_{\mathrm{A,inst}}(k_1,k_2,\ell)$ into 10 intervals, all containing the same amount of samples. Thus, we group the values according to their relative contribution to the overall signal level. The rationale behind this strategy is to group samples that can be approximated to be realizations of the same RV. In other words, we aim to make the samples consistent w.r.t. to the distribution model, i.e. their variance [54]. We evaluate the SUC for each of the resulting groups.[23]

The SUC averaged over all groups was $U(R_{\mathrm{A,inst}}(k_1,k_2), \Delta\phi(k_1,k_2)) = 5.28 \times 10^{-4}$. The maximum SUC value achieved by a single group was 1.9×10^{-3}, indicating statistical

[23] An implementation of this experiment can be found on `https://gitlab.com/johannesstahl`.

independence between $R_{\mathrm{A,inst}}(k_1, k_2)$ and $\Delta\phi(k_1, k_2)$ and hence supporting the validity of Assumption 1. We also conducted experiments without normalization and grouping and with alternative normalization and grouping strategies following [17] and [54], all yielding results in the same order of magnitude.

Dependencies of Magnitude and Phase w.r.t. Frequency

To address the validation of Assumption 2, we follow the same experimental strategy as in Section 6.3.1. Figure 6.2 illustrates the average SUC for all possible combinations of k_1 and k_2. For comparison, we also illustrate the results of the same evaluation for white noise. As we are dealing with real-valued time domain data, we only show the frequency range from 0 to $f_s/2$.

While the SUC for the spectral phase is close to 0 for off-diagonal elements, the speech magnitude shows a higher SUC for $k_1 \neq k_2$. Given the independence of the phase w.r.t. frequency, the phase pdfs can be factorized, and we have

$$\mathbb{E}\left(\mathrm{e}^{j(\phi(k_1,\ell)-\phi(k_2,\ell))}\right) = \mathbb{E}\left(\mathrm{e}^{j\phi(k_1,\ell)}\right)\mathbb{E}\left(\mathrm{e}^{-j\phi(k_2,\ell)}\right). \tag{6.18}$$

In Section 6.2, we assumed the DFT coefficients to be zero-mean, complex-valued, circularly Gaussian distributed, meaning that the corresponding phases are uniformly distributed. This implies that $\mathbb{E}\left(\mathrm{e}^{j\phi(k_1,\ell)}\right) = \mathbb{E}\left(\mathrm{e}^{-j\phi(k_2,\ell)}\right) = 0$, hence $R_\phi(k_1, k_2, \ell) = 0$ in Equation (6.11).

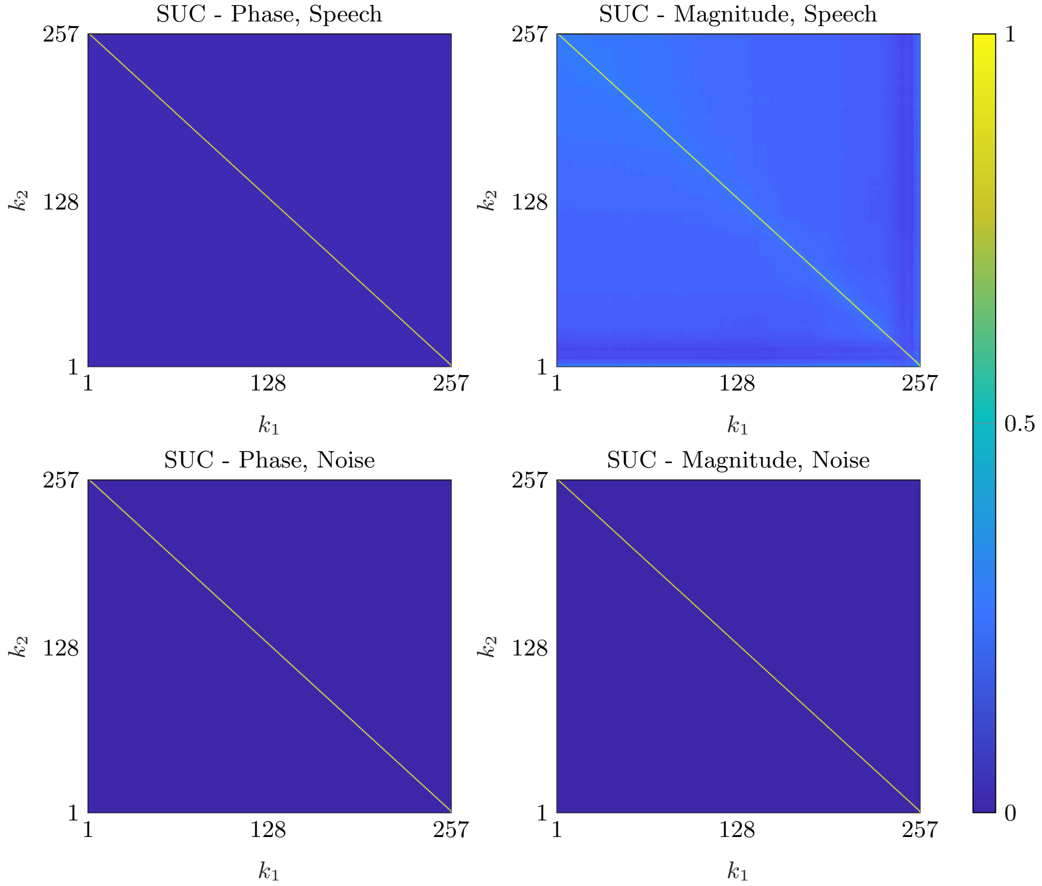

Figure 6.2: Empirical SUC values for (left) spectral phase and (right) spectral magnitude. The upper row refers to speech data while the bottom row corresponds to noise data.

6.4 Proposed Solution to the Covariance Phase Problem

Following Assumption 1 and Assumption 2, this section presents an approach that deliberately neglects phase correlations by introducing a linear multidimensional STSA estimator.

6.4.1 The LMD-STSA Estimator

To circumvent difficulties in estimating the phase of the optimal gain matrix and following the assumption that the spectral phase is independent w.r.t. frequency, we seek a speech estimation framework that does not rely on phase difference estimates while still allowing correlations between frequency channels. To this end, we derive a linear multidimensional MMSE-STSA (LMD-STSA) estimator that is independent of the phase of $\mathbf{y}_f(\ell)$, i.e. we search for [4]

$$\mathbf{G}_{\text{prop}}(\ell) = \underset{\mathbf{G}(\ell)}{\arg\min}\, \mathbb{E}\big(\big\| |\mathbf{x}_f(\ell)| - \mathbf{G}(\ell)\, |\mathbf{y}_f(\ell)| \big\|_2^2 \big). \tag{6.19}$$

In Table 6.1, we relate the above optimization criterion to similar approaches in the literature.

Table 6.1: Summary of multidimensional estimators.

Method	Optimization Criterion	Additional Information
MD-WF [64]	$\mathbb{E}\big(\|\mathbf{x}_f - \mathbf{G}\mathbf{y}_f\|_2^2 \big)$	$\mathbf{G} \in \mathbb{C}^{N_{\text{DFT}} \times N_{\text{DFT}}}$
MD-STSA [131]	$\mathbb{E}\big(\| \lvert\mathbf{x}_f\rvert - \lvert\hat{\mathbf{x}}_f\rvert \|_2^2 \big)$	relies on $\mathbf{G} \in \mathbb{C}^{N_{\text{DFT}} \times N_{\text{DFT}}}$
LMD-STSA [4]	$\mathbb{E}\big(\| \lvert\mathbf{x}_f\rvert - \mathbf{G}\, \lvert\mathbf{y}_f\rvert \|_2^2 \big)$	$\mathbf{G} \in \mathbb{R}^{N_{\text{DFT}} \times N_{\text{DFT}}}$

The optimal gain matrix is obtained by setting the derivative of the MSE in Equation (6.19) w.r.t. $\mathbf{G}(\ell)$ to zero, resulting in

$$\begin{aligned} \mathbf{G}_{\text{prop}}(\ell) &= \mathbb{E}\big(|\mathbf{x}_f(\ell)|\, |\mathbf{y}_f(\ell)|^T \big) \mathbb{E}\big(|\mathbf{y}_f(\ell)|\, |\mathbf{y}_f(\ell)|^T \big)^{-1} \\ &= \mathbf{R}_{|\text{x}||\text{y}|}(\ell) \mathbf{R}_{|\text{y}|}(\ell)^{-1}, \end{aligned} \tag{6.20}$$

with $\{\mathbf{R}_{|\text{x}||\text{y}|}(\ell), \mathbf{R}_{|\text{y}|}(\ell)\} \in \mathbb{R}^{N_{\text{DFT}} \times N_{\text{DFT}}}$, containing the second order (cross-)moments of the quantities in the subscript. In the following, we refer to second order moment matrices instead of covariance matrices if the corresponding random vectors are not necessarily zero-mean.

The Optimal Phase Estimate

In order to reconstruct the time domain signal from the estimated STSA, we also need an STSP estimate. In the context of STSA estimators, the noisy phase $\angle\mathbf{y}_f(\ell)$ is commonly used, since it is optimal if (i) no further additional a priori knowledge about the spectral phase is given and (ii) no dependencies w.r.t. time or frequency are considered. This chapter specifically deals with the latter case. However, since we assumed independence of magnitude and phase terms, any estimate of the spectral phase that relies on dependencies across frequency bins can only be a function of the phase correlation and we so far assumed that this quantity is either not accessible or zero. Hence, the noisy phase is optimal in the MMSE sense for the given signal model. The full DFT representation of the estimated speech signal is given by

$$\hat{\mathbf{x}}_{f,\text{prop}}(\ell) = \big(\mathbf{G}_{\text{prop}}(\ell) |\mathbf{y}_f(\ell)| \big) \circ \mathrm{e}^{j\angle\mathbf{y}_f(\ell)}. \tag{6.21}$$

Theoretical Performance of the LMD-STSA Estimator

This section presents a Monte Carlo simulation to assess the theoretical performance gain that can be achieved by the estimators in Equation (6.4), Equation (6.5), and Equation (6.21) compared to the standard Wiener filter, assuming diagonal second moment matrices. Since the main motivation to derive the LMD-STSA estimator was its insensitivity to covariance phase estimation errors, the impact of the covariance matrices' phase estimation accuracy is also assessed. The true covariance matrices are assumed to be known and their phase is perturbed with an uniformly distributed error d_ϕ, i.e. $d_\phi \sim \mathcal{U}\big(-\frac{\Delta_\phi}{2}, \frac{\Delta_\phi}{2}\big)$. In order to fit the underlying signal model of the analyzed estimators, the simulated speech and noise data were sampled from 2-dimensional, circularly complex-valued, Gaussian distributions characterized by 2×2 complex-valued covariance matrices.

The noise components were selected to be uncorrelated, so that only one covariance matrix phase is perturbed. Since the covariance matrices are Hermitian symmetric, only one cross-correlation coefficient needs to be defined in the 2-D case. We chose $0.99 \cdot \mathrm{e}^{j\frac{\pi}{4}}$ for speech, modeling highly correlated frequency channels. The absolute scaling and the main diagonal entries were sampled according to the input SNR.[24] To illustrate the sensitivity of the different estimator's effectiveness to SNR, we chose SNRs $\in \{-10, 0, 10\}$ dB.

For each experiment, we performed $N_{\mathrm{MC}} = 10^5$ Monte Carlo runs and report the mean square error (MSE) improvement compared to the diagonal Wiener filter estimate, denoted by $\widehat{\mathbf{x}}_{f,\mathrm{WF}}$, in the complex domain and on the estimated magnitude, i.e.

$$\begin{aligned}
\Delta\mathrm{MSE}_{\mathrm{x}} &= 10\mathrm{log}_{10}\left(\frac{\sum_{i=1}^{N_{\mathrm{MC}}} \|\mathbf{x}_{f,i} - \widehat{\mathbf{x}}_{f,i}\|_2^2}{\sum_{i=1}^{N_{\mathrm{MC}}} \|\mathbf{x}_{f,i} - \widehat{\mathbf{x}}_{f,\mathrm{WF},i}\|_2^2}\right), \\
\Delta\mathrm{MSE}_{|\mathrm{x}|} &= 10\mathrm{log}_{10}\left(\frac{\sum_{i=1}^{N_{\mathrm{MC}}} \||\mathbf{x}_{f,i}| - |\widehat{\mathbf{x}}_{f,i}|\|_2^2}{\sum_{i=1}^{N_{\mathrm{MC}}} \||\mathbf{x}_{f,i}| - |\widehat{\mathbf{x}}_{f,\mathrm{WF},i}|\|_2^2}\right).
\end{aligned} \tag{6.22}$$

The outcome of the simulation is shown in Figure 6.3. We see that the performance of the MD-WF and the MD-STSA is hardly affected by the phase error for SNR$= -10$ dB, while with increasing SNR, the impact of the phase error increases. The possible performance gain from the multidimensional estimators compared to the diagonal Wiener filter is more pronounced in $\Delta\mathrm{MSE}_{|\mathrm{x}|}$ than in $\Delta\mathrm{MSE}_{\mathrm{x}}$. As expected, the MSE of the LMD-STSA does not increase with the covariance phase error. Also, with increasing SNR, the LMD-STSA outperforms the MD-WF and the MD-STSA for a wider range of Δ_ϕ.

Despite the simplified nature of these experiments in terms of the number of frequency channels and the independently sampled noise components, we see that even with known

[24] An implementation of this experiment can be found at [177].

covariance magnitudes, their phases can play a crucial role and may deteriorate the performance of the MD-WF and the MD-STSA.

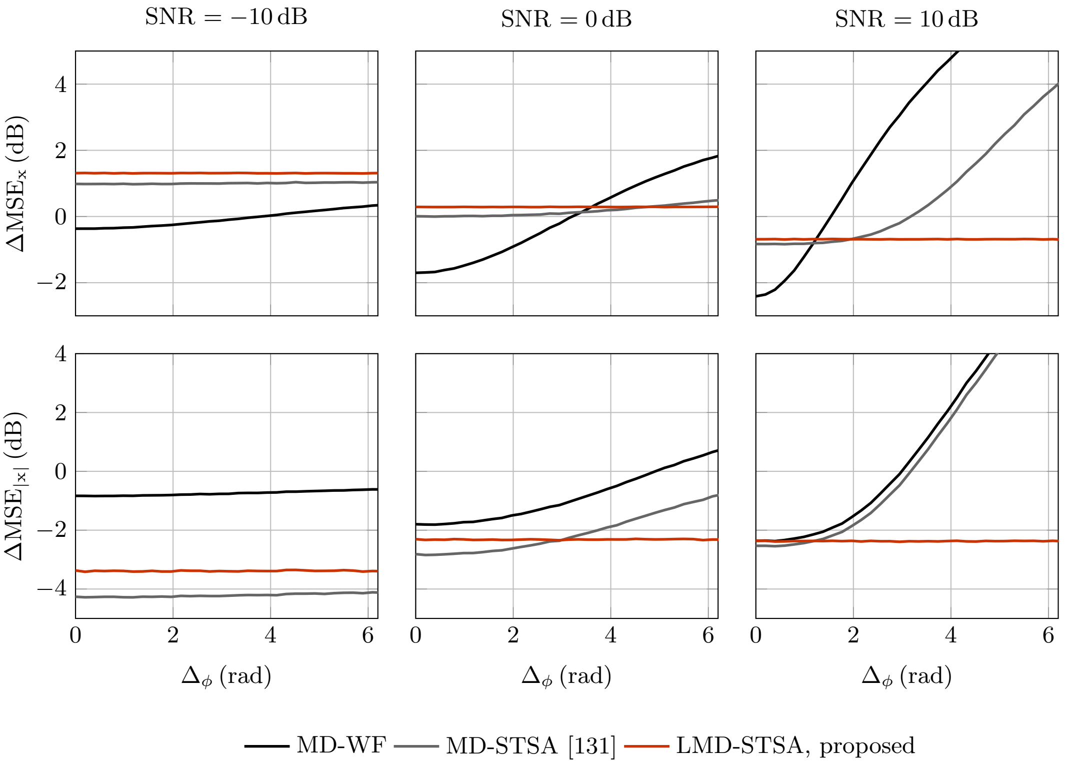

Figure 6.3: MSE improvement (the lower the better) compared to the diagonal Wiener filter as a function of phase error.

Estimating the Second Moments of the Magnitudes

In order to apply the LMD-STSA estimator in a practical scenario, $\mathbf{R}_{|\mathrm{x}||\mathrm{y}|}(\ell)$ and $\mathbf{R}_{|\mathrm{y}|}(\ell)$ have to be estimated from the noisy speech signal. While we can estimate $\mathbf{R}_{|\mathrm{y}|}(\ell)$ directly by temporal averaging, estimating $\mathbf{R}_{|\mathrm{x}||\mathrm{y}|}$ requires knowledge about the exact mixing of $\mathbf{x}_f(\ell)$ and $\mathbf{d}_f(\ell)$, hence, about their phases. Since we lack this knowledge, we analyze upper (ub) and lower bounds (lb) of $\mathbf{R}_{|\mathrm{x}||\mathrm{y}|}$ in the following. These bounds are independent of the phase of $\mathbf{x}_f(\ell)$ and we subsequently use them to formulate an estimate of $\mathbf{R}_{|\mathrm{x}||\mathrm{y}|}(\ell)$ by linearly combining them as follows

$$\widehat{\mathbf{R}}_{|\mathrm{x}||\mathrm{y}|}(\ell) = \alpha_{\mathrm{R}}\mathbf{R}^{\mathrm{ub}}_{|\mathrm{x}||\mathrm{y}|}(\ell) + (1-\alpha_{\mathrm{R}})\mathbf{R}^{\mathrm{lb}}_{|\mathrm{x}||\mathrm{y}|}(\ell), \tag{6.23}$$

where $\alpha_{\mathrm{R}} \in [0,1]$. We can find the upper bounds of the entries of $\mathbf{R}_{|\mathrm{x}||\mathrm{y}|}$ by applying the triangle inequality

$$\begin{aligned} R_{|\mathrm{x}||\mathrm{y}|}(k_1,k_2,\ell) &= \mathbb{E}\big(|X(k_1,\ell)||X(k_2,\ell)+D(k_2,\ell)|\big) \\ &\leq \mathbb{E}\big(|X(k_1,\ell)||X(k_2,\ell)|\big) + \mathbb{E}\big(|X(k_1,\ell)|\big)\mathbb{E}\big(|D(k_2,\ell)|\big) \\ &= R_{|\mathrm{x}|}(k_1,k_2,\ell) + \mu_{|\mathrm{x}|}(k_1,\ell)\mu_{|\mathrm{d}|}(k_2,\ell). \end{aligned} \tag{6.24}$$

We then obtain lower bounds by using Jensen's inequality, i.e. $\mathbb{E}\big(|z|\big) \geq |\mathbb{E}\big(z\big)|$

$$R_{|\mathrm{x}||\mathrm{y}|}(k_1,k_2,\ell) \geq |R_{\mathrm{x}}(k_1,k_2,\ell)|. \tag{6.25}$$

We can rewrite the above upper and lower bounds in matrix-vector notation as

$$\mathbf{R}^{\mathrm{ub}}_{|\mathrm{x}||\mathrm{y}|}(\ell) = \mathbf{R}_{|\mathrm{x}|}(\ell) + \boldsymbol{\mu}_{|\mathrm{x}|}(\ell)\boldsymbol{\mu}_{|\mathrm{d}|}(\ell)^T, \tag{6.26}$$

and

$$\mathbf{R}^{\mathrm{lb}}_{|\mathrm{x}||\mathrm{y}|}(\ell) = |\mathbf{R}_{\mathrm{x}}(\ell)|. \tag{6.27}$$

Following the discussion in Section 6.3, we assume $\mathbf{R}_{\mathrm{x}}(\ell)$ to be diagonal, i.e. its absolute value is also diagonal. This is reasonable, considering that cross-moments equal to zero represent a lower bound of the off-diagonal elements of $\mathbf{R}_{|\mathrm{x}||\mathrm{y}|}(\ell)$ in general. Under this assumption, the ML estimate of $\mathbf{R}^{\mathrm{lb}}_{|\mathrm{x}||\mathrm{y}|}(\ell)$ is given by [15],

$$\widehat{\mathbf{R}}^{\mathrm{lb}}_{|\mathrm{x}||\mathrm{y}|,\mathrm{ML}}(\ell) = \max\big[(\widehat{\mathbf{R}}_{\mathrm{y}}(\ell) - \widehat{\mathbf{R}}_{\mathrm{d}}(\ell)) \circ \mathbf{I}_{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}}, 0\big], \tag{6.28}$$

where $\mathbf{I}_{\cdot\times\cdot}$ is the $\cdot\times\cdot$ identity matrix and the noise covariance $\widehat{\mathbf{R}}_{\mathrm{d}}$ can be estimated in speech pauses or by standard approaches like the minimum statistics method [70] and the MMSE-based noise PSD estimator from [18]. The restriction, that the latter two alternatives are only applicable for the case of a diagonal $\widehat{\mathbf{R}}_{\mathrm{d}}(\ell)$, is not critical for the lower bound of the covariance matrix, as it is anyway a diagonal matrix.

Given the lower bound in Equation (6.28) and an estimate of $\mathbf{R}_{\mathrm{d}}$, we obtain a preliminary speech estimate by applying the standard, diagonal Wiener filter. This pre-filtered speech is then used to estimate the second moment matrix of the speech magnitude by means of a recursive smoothing strategy as follows

$$\widehat{\mathbf{R}}_{|\mathrm{x}|}(\ell) = \alpha_{|\mathrm{x}|}\widehat{\mathbf{R}}_{|\mathrm{x}|}(\ell-1) + (1-\alpha_{|\mathrm{x}|})|\hat{\mathbf{x}}'_f(\ell)||\hat{\mathbf{x}}'_f(\ell)|^T, \tag{6.29}$$

with $\alpha_{|\mathrm{x}|} \in [0,1]$ and $\hat{\mathbf{x}}'_f(\ell) = (\mathbf{G}_{\mathrm{W}}(\ell) \circ \mathbf{I}_{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}})\mathbf{y}_f(\ell)$.

The second moment matrices of the noisy observation are obtained via recursive averaging as well,

$$\begin{aligned} \widehat{\mathbf{R}}_{\mathrm{y}}(\ell) &= \alpha_{\mathrm{y}} \widehat{\mathbf{R}}_{\mathrm{y}}(\ell - 1) + (1 - \alpha_{\mathrm{y}}) \mathbf{y}_f(\ell) \mathbf{y}_f^H(\ell), \\ \widehat{\mathbf{R}}_{|\mathrm{y}|}(\ell) &= \alpha_{|\mathrm{y}|} \widehat{\mathbf{R}}_{|\mathrm{y}|}(\ell - 1) + (1 - \alpha_{|\mathrm{y}|}) |\mathbf{y}_f(\ell)| |\mathbf{y}_f^H(\ell)|, \end{aligned} \tag{6.30}$$

with $\{\alpha_{\mathrm{y}}, \alpha_{|\mathrm{y}|}\} \in [0, 1]$. Note that since $\widehat{\mathbf{R}}_{\mathrm{y}}(\ell)$ is only used for computing the diagonal matrix $\widehat{\mathbf{R}}^{\mathrm{lb}}_{|\mathrm{x}||\mathrm{y}|,\mathrm{ML}}$, the (complex-valued) off-diagonal elements are not required for further computations. Therefore, recursive averaging, unlike for the case of full, complex-valued covariance matrices discussed in Section 6.3, is applicable here. Since we have to invert $\widehat{\mathbf{R}}_{|\mathrm{y}|}(\ell)$ in the LMD-STSA estimator, we perform diagonal loading [178] in order to regularize the matrix,

$$\widehat{\mathbf{R}}_{|\mathrm{y}|}(\ell) = \widehat{\mathbf{R}}_{|\mathrm{y}|}(\ell) + \frac{\delta_{\mathrm{reg}}}{N_{\mathrm{DFT}}} \mathrm{Tr}\{\widehat{\mathbf{R}}_{|\mathrm{y}|}(\ell)\} \mathbf{I}_{N_{\mathrm{DFT}} \times N_{\mathrm{DFT}}}, \tag{6.31}$$

where $\mathrm{Tr}\{\cdot\}$ is the trace operator and δ_{reg} is a parameter that scales the amount of diagonal loading.

The vectors $\boldsymbol{\mu}_{|\mathrm{x}|}(\ell)$ and $\boldsymbol{\mu}_{|\mathrm{d}|}(\ell)$ are obtained by means of the respective second moment matrices' main diagonals. In Section 6.1, we assumed the Fourier expansion coefficients to be zero-mean, complex-valued, circularly Gaussian distributed. Hence, their magnitudes are Rayleigh distributed [15]. The variance of a Rayleigh distributed RV X is given by

$$\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \sigma^2 \frac{4 - \pi}{2}, \tag{6.32}$$

and its expected value is

$$\mathbb{E}(X) = \sigma \sqrt{\frac{\pi}{2}}. \tag{6.33}$$

Isolating σ yields

$$\sigma = \sqrt{\frac{\mathbb{E}(X^2)}{2}}, \tag{6.34}$$

and inserting Equation (6.34) into Equation (6.33) gives us

$$\mathbb{E}(X) = \frac{1}{2} \sqrt{\pi \mathbb{E}(X^2)}. \tag{6.35}$$

We then obtain estimates of $\boldsymbol{\mu}_{|\mathrm{x}|}$ and $\boldsymbol{\mu}_{|\mathrm{d}|}$ as follows

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{|\mathrm{x}|}(\ell) &= \frac{1}{2}\sqrt{\pi \mathrm{diag}\left\{\widehat{\mathbf{R}}_{|\mathrm{x}|}(\ell)\right\}},\\ \hat{\boldsymbol{\mu}}_{|\mathrm{d}|}(\ell) &= \frac{1}{2}\sqrt{\pi \mathrm{diag}\left\{\widehat{\mathbf{R}}_{\mathrm{d}}(\ell)\right\}}.\end{aligned} \tag{6.36}$$

6.5 Evaluation

In the following, we evaluate the proposed LMD-STSA estimator in terms of single-channel speech enhancement performance.

6.5.1 Databases and Evaluation Metrics

For all simulations in this section, we used the test set of the TIMIT core database [139], consisting of 192 utterances. All signals were sampled at 16 kHz and we added one second of noise at the beginning and end of the individual signals. The noisy test data was obtained by mixing the speech signals with noise at SNRs between -5 dB and 15 dB in 5 dB steps following the mixing convention recommended in [164]. The specific noise types chosen in the experiments are white noise, factory noise, and babble noise from the NOISEX-92 database [142], and rain noise from [143]. This selection includes stationary, impulsive, speech-like, as well as highly non-stationary noise types.

We report $\widetilde{\mathrm{PESQ}}$ as a function of $\mathrm{NA}_{\mathrm{seg}}$, following the proposal in [179], explained in detail in Appendix B.2. Both, proof-of-concept and informal listening tests suggested that musical noise is effectively reduced by using multidimensional estimators when compared to their one-dimensional counterparts.[25] To quantify this observation, we report the WLKR [181] as described in Appendix B.3. Further, we report STOI [146] and PESQ [145] both in terms of Δ-improvement over the noisy speech signal.

6.5.2 Benchmark Algorithms

As benchmark algorithms, we present the Wiener filter, the MMSE-STSA estimator from [15], the MD-WF [172], and the MD-STSA estimator [131]. The intention behind this selection is to juxtapose the multidimensional estimators with their respective one-dimensional counterparts. For noise covariance matrix estimation, the estimates obtained from the MMSE-based noise PSD estimator in [18] were used to construct a diagonal matrix $\widehat{\mathbf{R}}_{\mathrm{d}}$, thus neglecting possible correlations across frequency in the noise. For all methods, we used a square-root Hamming window and set the frame length and the frame shift to 32 ms and 8 ms, respectively. The speech covariance matrices for the

[25] Listening examples be found on [180].

MD-WF [172], and the MD-STSA [131] were estimated following [131]. Table 6.2 lists the parameter settings used throughout the following evaluation. For all algorithms, the overall noise suppression was limited to $-20\,\mathrm{dB}$.

Table 6.2: Parameter settings for the MD estimators.

α_{R}	α_{y}	$\alpha_{\lvert\mathrm{y}\rvert}$	$\alpha_{\lvert\mathrm{x}\rvert}$	δ_{reg}	γ
0.5	0.9	0.998	0.5	0.04	0.5

6.5.3 Performance Limits and Potential of the Multidimensional Estimators

We analyze the potential performance gain provided by multidimensional estimators compared to their one-dimensional counterparts. In this section, we are interested in scenarios where the second order moment matrices of the speech STFT coefficients can be assumed to be known. We thus estimate the second moment matrices from isolated clean speech data by means of the DD approach. Similar to the blind experiments, we set $\alpha_{\mathrm{DD}} = 0.98$. However, for the multidimensional estimators, this choice of the forgetting factor resulted in audible reverberations. We therefore adapted the DD approach for the multidimensional estimators as follows

$$\widehat{\mathbf{R}}_{\mathrm{x,oracle}}(\ell) = \alpha_{\mathrm{DD}} \left(\hat{\mathbf{x}}_f(\ell-1)\hat{\mathbf{x}}_f(\ell-1)^H\right) \circ \mathbf{I}_{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}} + (1-\alpha_{\mathrm{DD}})\mathbf{x}_f(\ell)\mathbf{x}_f(\ell)^H, \quad (6.37)$$

and

$$\widehat{\mathbf{R}}_{\lvert\mathrm{x}\rvert,\mathrm{oracle}}(\ell) = \alpha_{\mathrm{DD}} \left(\lvert\hat{\mathbf{x}}_f(\ell-1)\rvert\,\lvert\hat{\mathbf{x}}_f(\ell-1)\rvert^T\right) \circ \mathbf{I}_{N_{\mathrm{DFT}}\times N_{\mathrm{DFT}}} + (1-\alpha_{\mathrm{DD}})\,\lvert\mathbf{x}_f(\ell)\rvert\,\lvert\mathbf{x}_f(\ell)\rvert^T . \quad (6.38)$$

Note that this approach ensures that for all estimators in this experiment, the main diagonals of the second order moment matrices are computed in the same manner, with the same α_{DD}. Since we further used the noise PSD estimate as explained in the previous section, any improvement by the multidimensional estimators will necessarily be a result from taking into account the inter-frequency dependencies of the speech STFT coefficients.

Figure 6.4 illustrates the results of this evaluation, averaged over all noise scenarios. In terms of ΔPESQ and ΔSTOI, we see the high potential of the multidimensional estimators, as they outperform both the Wiener filter and the MMSE-STSA estimator.

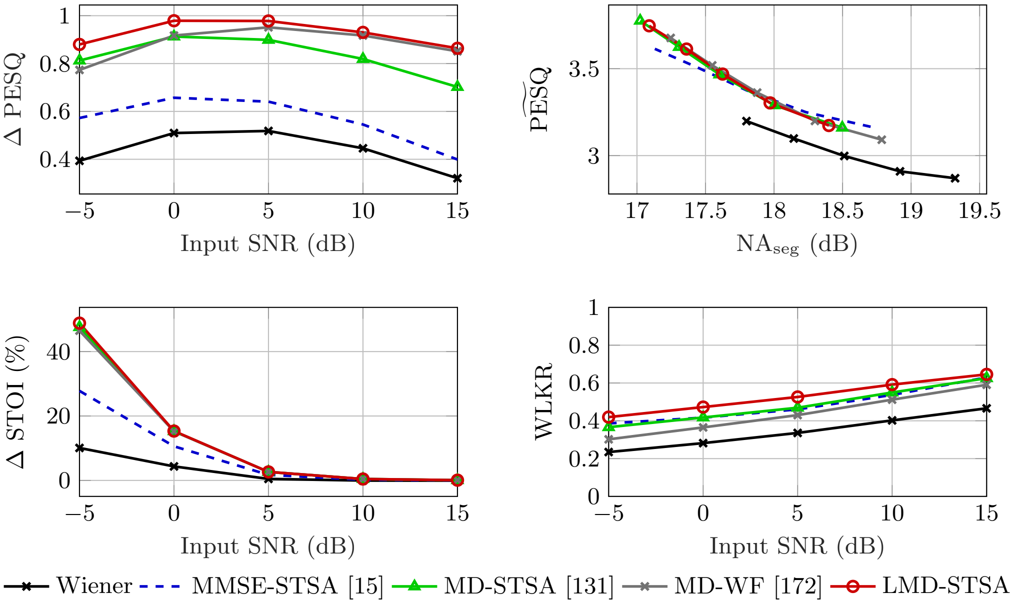

Figure 6.4: Speech enhancement performance of multidimensional and corresponding one-dimensional estimators with second order moments estimated from isolated speech data, averaged over all noise scenarios.

6.5.4 Proof-of-concept

In Figure 6.5, we show spectrograms of clean, noisy, and enhanced speech files. We note that all MD approaches successfully preserve onsets and suppress isolated spectral peaks. Although the MD-STSA estimator [131] and the MD-WF filter [172] are implemented with the same covariance matrix estimation procedure, the MD-STSA estimator better preserves broadband signal components.[26] In this context it is important to recall however that the MD-WF also processes the spectral phase, while the MD-STSA estimator and the LMD-STSA estimator both use the noisy phase for reconstruction. We speculate that the phase estimates resulting from the MD-WF approach lack coherency and hence, the individual spectral components do not add up constructively. This effect has been used under the term phase-randomization [182] for example to reduce auto-focus noise recorded by camera microphones. We also note that compared to the benchmarks, the LMD-STSA estimator better preserves unvoiced parts of the speech signal.

[26] We tried different settings for the DD smoothing constant used for estimating the covariance matrix, all resulting in similar observations.

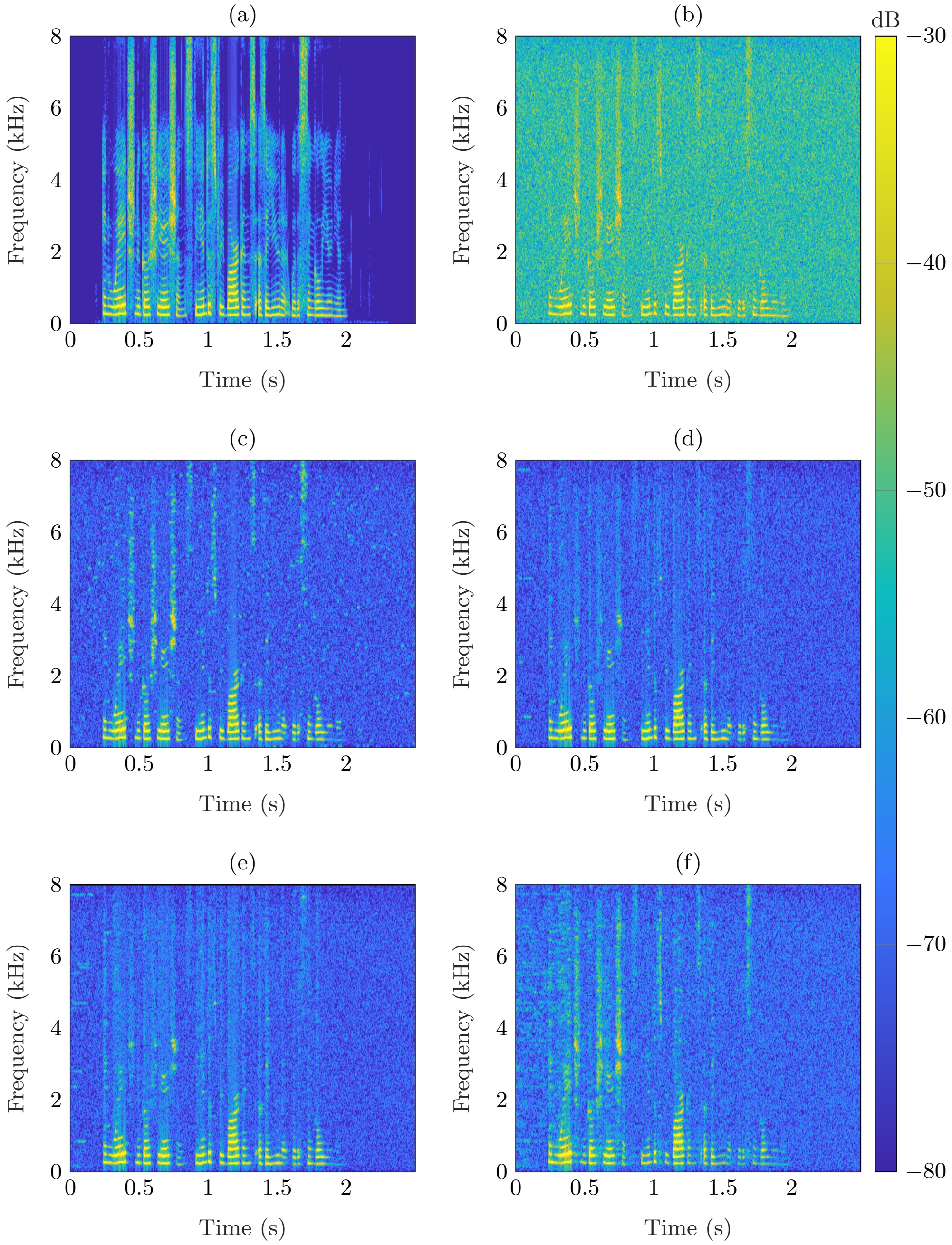

Figure 6.5: Proof-of-concept: Female speaker uttering *In wage negotiations the industry bargains as a unit with a single union.* from the TIMIT core database [139] mixed with white noise, with SNR = 5 dB. (a) Clean speech, (b) Noisy Speech, (c) Wiener Filter, (d) MD-WF [172] + [131], (e) MD-STSA [131], (f) LMD-STSA, proposed.

6.5.5 Objective Evaluation of Speech Enhancement Performance

The ΔPESQ results shown in Figure 6.6 demonstrate that the proposed approach improves the robustness of the multidimensional estimators in a fully blind scenario. Surprisingly, the one-dimensional estimators outperform their multidimensional counterparts in terms of ΔPESQ, except for babble noise, where the LMD-STSA estimator performs best. However, evaluating $\widetilde{\mathrm{PESQ}}$ instead of ΔPESQ yields a different picture. Figure 6.7 illustrates the evaluation in terms of $\mathrm{NA}_{\mathrm{seg}}$ and $\widetilde{\mathrm{PESQ}}$. We note that

- the multidimensional speech estimators outperform their one-dimensional counterparts for a large range of noise and SNR scenarios and
- the LMD-STSA estimator results in increased noise reduction compared to the benchmarks in all scenarios while preserving the level of speech distortions, as indicated by $\widetilde{\mathrm{PESQ}}$.

The STOI results are shown in Figure 6.8. In low SNR conditions, the LMD-STSA estimator improves the predicted intelligibility compared to the benchmarks, except for rain noise, where the MMSE-STSA estimator performs best. However, compared to the multidimensional benchmarks, the LMD-STSA estimator consistently performs best in terms of STOI, indicating improved robustness in a fully blind scenario.

Informal listening tests suggested that the residual noise floor of the proposed method sounds more natural compared to the reference methods though slight reverberation is introduced. In particular, musical noise was barely perceivable, which is also indicated by the WLKR results in Figure 6.9. All multidimensional estimators seem preferable in terms of musical noise artifacts compared to one-dimensional methods. This is due to the fact that taking into account dependencies w.r.t. frequency intrinsically prevents isolated spectral peaks.

The results averaged over all noise scenarios are depicted in Figure 6.10, summarizing the observations made in this section. In terms of STOI and WLKR, the LMD-STSA estimator performs best among all tested algorithms.

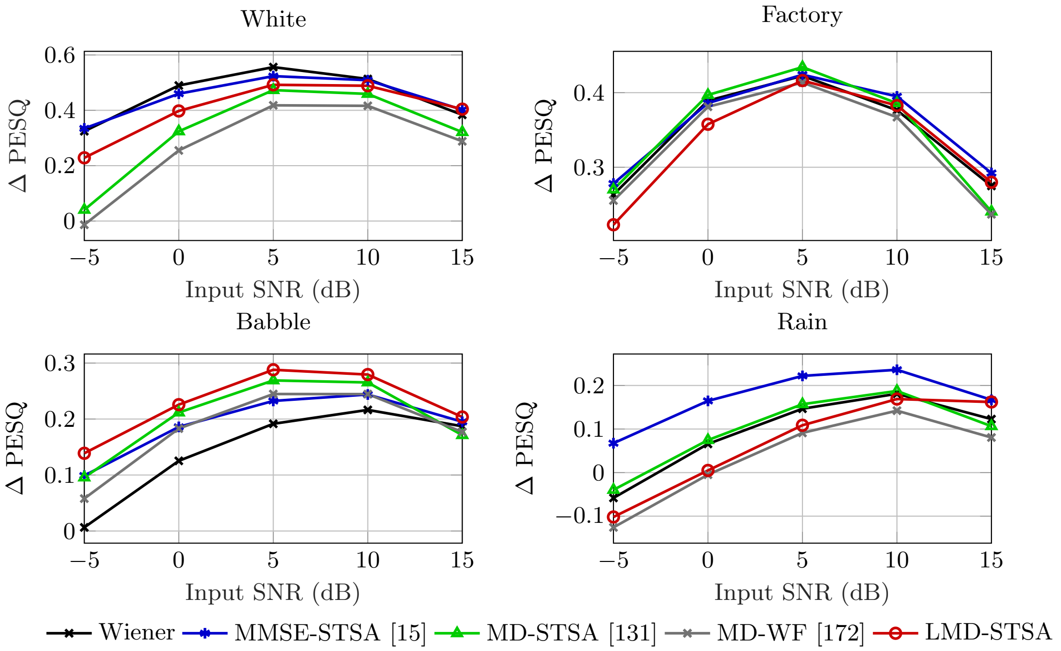

Figure 6.6: Speech enhancement performance of the LMD-STSA algorithm in terms of ΔPESQ-improvement over the noisy input speech for segmental SNR.

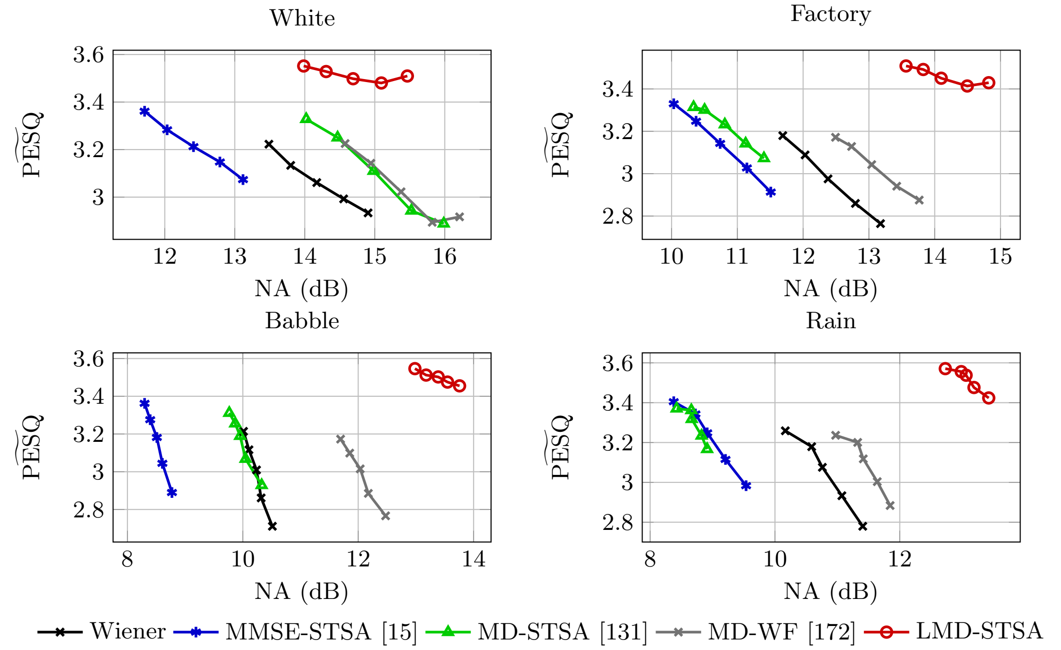

Figure 6.7: Black box evaluation of speech enhancement performance in of multidimensional estimators in terms of NA_{seg} and $\widetilde{\text{PESQ}}$.

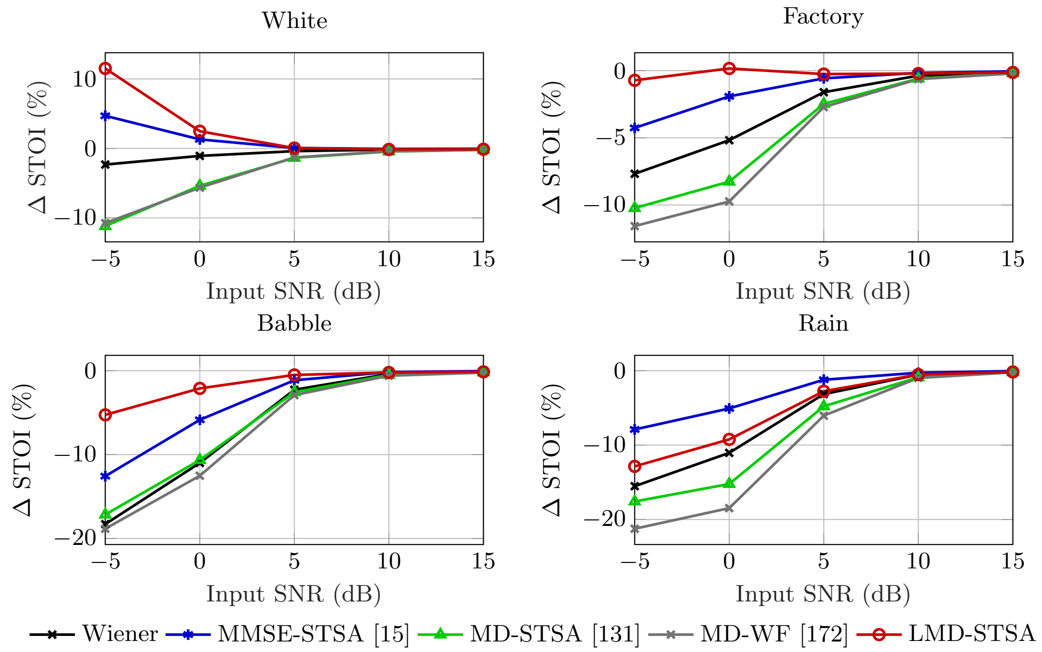

Figure 6.8: Speech enhancement performance of the LMD-STSA algorithm in terms of ΔSTOI-improvement over the noisy input speech.

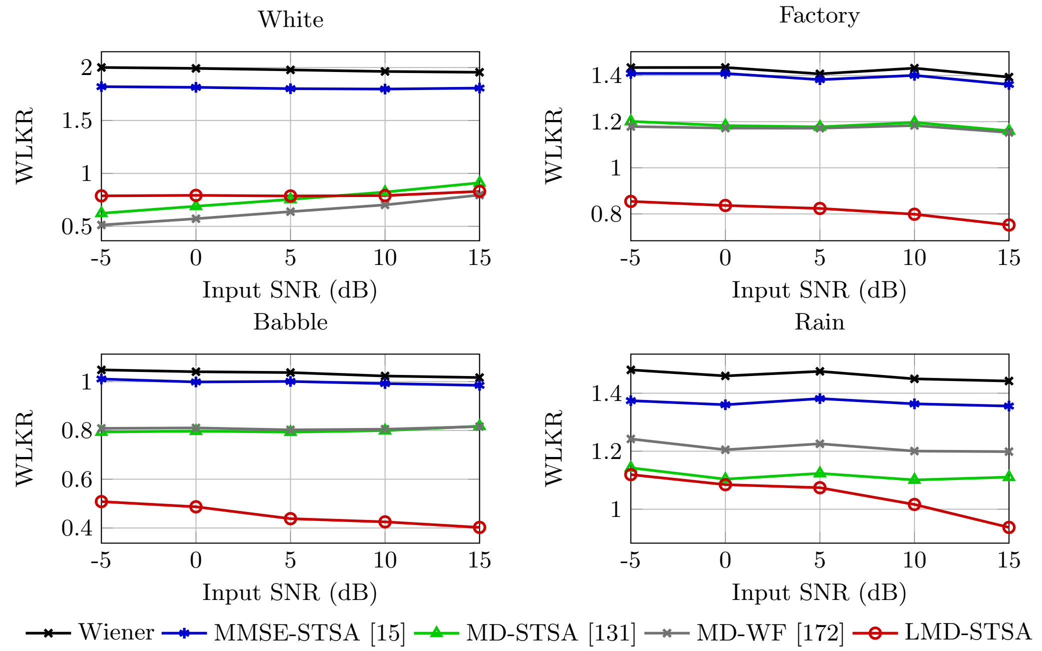

Figure 6.9: Musical noise evaluation in terms of WLKR. Lower WLKR values correspond to less musical noise.

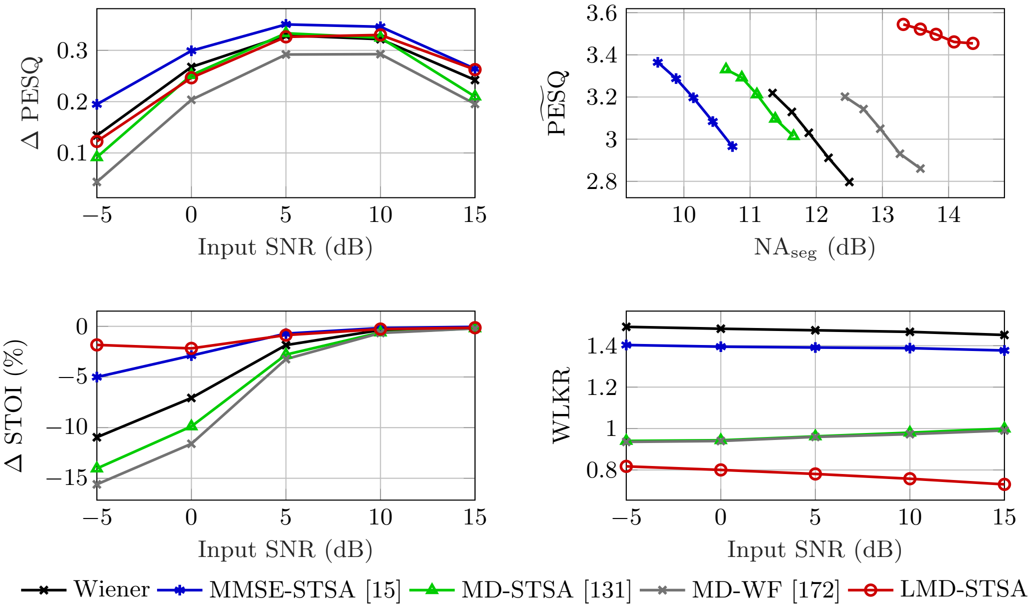

Figure 6.10: Speech enhancement performance of multidimensional and corresponding one-dimensional estimators, averaged over all noise scenarios.

6.6 Conclusion

This chapter presents an analysis of speech enhancement algorithms that take inter-frequency channel correlations of the speech signal into account. An important result is that such approaches typically also imply a phase estimation task, since the covariance matrices of speech and noise are complex-valued. In terms of speech enhancement performance, the estimation of the full covariance matrices is a critical task that is not adequately solved by standard approaches like the DD method. The estimation of the covariances' phase is particularly problematic. The contributions contained in this chapter are threefold: (i) we analyze the impact of the covariance phase on spectral coefficient estimation performance, (ii) we illustrate that it is reasonable to assume that the spectral phase is independent w.r.t. frequency while the spectral amplitude is not, and (iii) we propose a linear multidimensional MMSE-STSA estimator that is independent of the covariance phase. Further, the proposal is compared to relevant benchmark methods, indicating improved noise reduction while preserving both speech quality and intelligibility. In an oracle scenario, where the second order moment matrices are computed from isolated speech data, we demonstrate the significant potential of multidimensional estimators w.r.t. joint speech intelligibility and quality improvement. The results indicate that the proposal of neglecting inter-frequency phase dependencies helps to improve the robustness of the multidimensional approaches in a fully blind scenario. However, it is the belief of the author that data-driven approaches to estimating the second order moments have the potential to further push the performance of the proposed multidimensional estimator towards its ideal performance. In this respect, an advantage of the proposed multidimensional estimator is that only a positive-valued quantity has to be learned from the data, compared to existing approaches, where complex-valued covariance matrices would have to be learned. In this respect, an advantage of the proposed multidimensional estimator is that only positive-valued quantities have to be learned, thus opening the door to compositional model learning.

7 Conclusion

This thesis presented three scenarios in STFT-based speech enhancement, where the spectral phase needs to be considered in a different way than in classical speech enhancement. More specifically, we studied the impact of (i) incorporating the harmonic signal model, (ii) correlations of STFT coefficients w.r.t. time, and (iii) correlations of STFT coefficients w.r.t. frequency.

Chapter 4 illustrates that given an estimate of harmonic speech signal components, the MMSE-optimal gain is complex-valued in general. A joint detection-estimation approach under this premise is derived and a pitch-synchronous signal analysis stage is shown to be beneficial for estimating the harmonic signal components, a problem that has to be solved prior to the detection-estimation stage. To avoid a fixed choice for the maximum harmonic frequency, an empirical Bayesian approach that facilitates parameterizing the prior probability of harmonics being present in a specific time-frequency region is derived. The pitch-synchronous detection-estimation framework comes with slight perceived speech quality improvements, indicated by subjective and objective tests, and the instrumental intelligibility predictor STOI yields promising results, particularly in low SNR conditions. Further, the intelligibility performance of the benchmark proposed in [86] is improved by replacing the originally proposed harmonic retrieval stage with the proposed Bayesian pitch-synchronous approach, indicating that the specific harmonic retrieval approach chosen may impact dramatically on the speech enhancement performance of estimators relying on the harmonic plus noise model.

In the context of the PhD theses relevant to the present thesis it is worth noting that the proposed simultaneous detection-estimation approach generalizes the stochastic-deterministic Wiener filter presented in [40]. Further, the pitch-synchronous harmonic retrieval block has been shown to improve the speech-enhancement performance of the SD MMSE-STSA estimator from [40]. In this respect, it is important to note that the harmonic retrieval approach presented in Chapter 4 benefits from incorporating empirical priors, as they circumvent the need to set a fixed maximum harmonic frequency, which was needed in [39, 40].

In **Chapter 5**, we consider a signal representation that aligns harmonic frequencies in the time-frequency plane. This representation facilitates modeling temporal correlation in the STFT coefficients. By representing the instantaneous spectral phase as the composition of a linear phase progression part plus a zero-mean wrapped Gaussian stochastic phase component, the state transition model of a subband Kalman filter can be obtained by applying circular statistics. Instrumental metrics indicate that the pitch-adaptive signal analysis stage in conjunction with the Kalman filter comes with increased noise reduction and preserves the speech distortion level at the same time. We confirmed this conjecture by a subjective listening test which shows a significant preference of the proposed method in terms of noise reduction compared to the benchmark methods. The speech distortion level is rated similarly across the tested algorithms.

The complex-valued Kalman filter extends the work in [41], as it provides an explicit analytical model for the AR parameters in voiced speech and further uses this model to formulate the propagation step of the Kalman filter. Further, the departure to pitch-adaptive and pitch-synchronous signal representations can be seen as an extension of general considerations w.r.t. sinusoidal signal models presented in [43], applied to the problem of single-channel speech-enhancement.

Chapter 6 investigates the role of the spectral phase in modeling inter-frequency correlations in speech and noise signals. Existing multidimensional estimators rely on complex-valued covariance matrices for the estimation of the speech STFT coefficients. Without any further assumptions, the phases of the covariance matrices are challenging to estimate. We therefore derive a linear multidimensional STSA estimator which circumvents the need to estimate them. Further, we show that it is reasonable to assume that the spectral phase is independent w.r.t. frequency, which is not the case for the spectral magnitude. The proposed linear multidimensional STSA estimator facilitates incorporating inter-frequency correlations into a classical STFT speech enhancement framework. When the speech second order moment matrices are estimated from isolated speech data, all multidimensional estimators outperform their one-dimensional counterparts in terms of instrumentally predicted speech quality and intelligibility. The experimental results indicate that in a fully blind scenario, focusing on the magnitude improves the robustness of the multidimensional estimators.

The results of Chapter 6 are particularly interesting in the context of the related PhD theses in Section 1.2. In the conclusion of his PhD thesis [42], Plourde points out that he observed the complex-valued STFT coefficients to be less correlated w.r.t. frequency than their magnitudes. As a consequence, he mentions the possibility to derive an estimator under the assumption of correlated magnitudes but uncorrelated phases. The LMD-STSA estimator and the corresponding phase estimate in Chapter 6 are, although motivated differently, derived from this statistical model.

In the light of the **research question(s)** posed in Chapter 1, Chapter 4 exemplifies that from a mathematical point of view, any non-circular statistical model for the STFT coefficients renders the noisy phase to be a suboptimal estimate of the clean speech spectral phase. Incorporating harmonic signal modeling is one source of non-circularity in the STFT coefficients' distributions. However, estimators derived from such a signal model heavily rely on the accuracy of the respective parameters' estimates. For the harmonic model, tools such as iterative algorithms, e.g. [1], or adaptations to the signal analysis stage as in [2] are needed to increase the accuracy compared to directly optimizing criteria such as the least squares, ML, or MAP criteria. An essential goal of phase-sensitive speech enhancement strategies is the preservation or the retrieval of spectral phase structure along time and frequency. The harmonic model can indicate what structure to expect in voiced speech segments. Further, the STFT coefficients are highly interdependent w.r.t. time and frequency due to the speech production process itself, the finite window length, and the overlapping signal frames. As illustrated in Chapter 5 and Chapter 6 estimating the cross-moments of the respective complex-valued coefficients is the key problem to solve before speech enhancement algorithms can profit from the interdependencies w.r.t. time and frequency. While again the harmonic signal model may be used to simplify this task, we also show that specific spectral phase properties render the estimation of its cross-moments unreliable. This is why under such circumstances it can be beneficial to resort to magnitude-only solutions.

Appendices

A
Some Remarks on the Notation in this Thesis

A.1 Continuous-Time and Discrete-Time Signals

While this thesis mainly deals with discrete-time signals, for certain considerations, continuous-time signals will be of interest. The discrete-time signal is obtained by sampling the corresponding continuous-time signal $x_c(t)$ with sampling rate f_s, i.e.

$$x(n) = x_\mathrm{c}(nT_\mathrm{s}), \tag{A.1}$$

where $n \in \mathbb{Z}$ is the discrete-time index, $T_\mathrm{s} = 1/f_\mathrm{s}$ is the sampling period, and the continuous-time is $t \in \mathbb{R}$. While in the literature, e.g. [48], we often find [ ] to enclose the independent variable of discrete-variable functions and ( ) to enclose the independent variable of continuous-variable functions, in this thesis we leave the task of distinguishing between the two cases solely to the subscript "c".

A.2 Random Variables

Conceptually it is of course important to distinguish random variables and realizations of random variables. While in many works, this is done by using notation indicators such as capital and lowercase letters, in this work, we follow [121], where such a notation convention is omitted and hope that the meaning will be clear from the context. By using this notation, we also omit the subscript that identifies the RV that corresponds to the pdf under consideration. We leave this task to the pdf argument only.

B

Evaluation Metrics

In this appendix, we shortly review the instrumental metrics we chose for the assessment of the speech enhancement performance of the algorithms discussed in this thesis.

B.1 Objective Speech Quality and Speech Intelligibility Predictors

Speech Quality

As objective measures for perceived speech quality prediction, we report the perceptual evaluation of speech quality (PESQ) measure [144, 145, 183]. The implementation was taken from [44]. It is important to note that the PESQ metric was originally proposed for the evaluation of speech coding algorithms. Hence, it is not designed for handling artifacts that are specific to speech enhancement such as musical noise. Nevertheless, it is widely used and has been found to be highly correlated to the human rating of the processed signals' overall quality [184]. An alternative, that is also based on the PESQ measure is presented in Section B.2.

Speech Intelligibility

For instrumental prediction of speech intelligibility, we used the short-time objective intelligibility (STOI) measure proposed in [146, 185], which has been proposed for intelligibility prediction of time-frequency weighted noisy speech. In order to map the STOI values to word accuracy rates (WAR), we applied a nonlinear mapping function as proposed for the English database in [146]. This strategy has, for example, also been used in [29, 132]. Since this mapping strongly depends on the speech database under test, the so-obtained values should be seen as an indicator of relative potential of the evaluated methods.

B.2 The Black Box Approach

To obtain insights into details of the tested noise suppression mechanism, we analyze how an algorithm affects both the speech and noise components individually. One possibility to achieve this is to log the weights that are applied to the noisy speech and separately apply them to the isolated speech and noise signals. Such a methodology is called white box approach in the literature [186–188][27] and suffers from the disadvantage that it is highly intrusive in the sense that it is applicable only if the parameters of speech enhancement system are perfectly known [189]. The black box methodology does not suffer from this drawbacks and emulates the system under test by approximating it as a spectral gain function that is computed as follows [190]

$$\hat{G}(k,\ell) = \min\left(\frac{\left|\hat{X}(k,\ell)\right|}{|Y(k,\ell)|}, 1\right) e^{j\angle\frac{\hat{X}(k,\ell)}{Y(k,\ell)}}, \tag{B.1}$$

where limiting the amplitude gain to a maximum value of 1 has been introduced in [190] for the sake of increased robustness of the above approximation. By applying $\hat{G}(k,\ell)$ separately on the isolated clean speech and noise STFTs and transforming the resulting STFTs back to time domain we obtain $\tilde{x}_{\mathrm{t}}(n,\ell)$ and $\tilde{d}_{\mathrm{t}}(n,\ell)$, respectively. To then quantify the impact of the system under test on the noise signal only, we analyze the segmental noise attenuation ($\mathrm{NA}_{\mathrm{seg}}$) given by [179]

$$\mathrm{NA}_{\mathrm{seg}} = 10\log_{10}\left(\frac{1}{|\mathcal{L}_{\mathrm{NA}}|}\sum_{\ell\in\mathcal{L}_{\mathrm{NA}}}\frac{\sum_{n=0}^{N-1} d_{\mathrm{t}}^2(n,\ell)}{\sum_{n=0}^{N-1} \tilde{d}_{\mathrm{t}}^2(n,\ell)}\right), \tag{B.2}$$

where $\mathcal{L}_{\mathrm{NA}}$ corresponds to the set of frames that is considered in the evaluation and $|\mathcal{L}_{\mathrm{NA}}|$ denotes its cardinality. The speech distortions can be evaluated by means of the segmental speech to speech distortion ratio ($\mathrm{SSDR}_{\mathrm{seg}}$) that is computed framewise by [179]

$$\mathrm{SSDR}_{\mathrm{seg}} = \frac{1}{|\mathcal{L}_{\mathrm{SSDR}}|}\sum_{\ell\in\mathcal{L}_{\mathrm{SSDR}}}\min\left(\mathrm{SSDR}(\ell), 30\right), \tag{B.3}$$

with

$$\mathrm{SSDR}(\ell) = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} x_{\mathrm{t}}^2(n,\ell)}{\sum_{n=0}^{N-1}\left(\tilde{x}_{\mathrm{t}}(n,\ell) - x_{\mathrm{t}}(n,\ell)\right)^2}\right), \tag{B.4}$$

and $\mathcal{L}_{\mathrm{SSDR}} = \{\ell|\mathrm{SSDR}(\ell) > -10\,\mathrm{dB}\}$ [179].

[27] The term “white” thus refers to a scenario where the internal processing of the noise reduction scheme is exactly known.

To investigate the perceptual relevance of the introduced speech distortions, we can construct the time domain signal $\tilde{x}_{\mathrm{t}}(n)$ and evaluate it by means of the PESQ score [190]. We denote the resulting values by $\widetilde{\mathrm{PESQ}}$. Note that in the context of speech enhancement, evaluating $\widetilde{\mathrm{PESQ}}$ is better justified than directly evaluating the PESQ of the filtered noisy signal, since the PESQ metric originally has been proposed for the purpose of evaluating speech codecs and is not designed to handle residual noise as it occurs in the case of speech enhancement algorithms [191].

B.3 Musical Noise Assessment

Besides the speech component quality and the level of noise suppression, a further measure one may be interested in is the quality of the residual noise floor. A well known artifact of speech enhancement algorithms is so-called musical noise [16], a phenomenon that results from isolated peaks in the spectrogram, perceived as highly fluctuating tones. In order to quantify the amount of musical noise, the works in [192–194] consider the change in kurtosis of the (residual) noise before and after processing the noisy speech signal. The rationale behind these approaches is that the kurtosis, i.e. the "tailedness", of an empirical distribution indicates the sparseness of the underlying data [195]. Low kurtosis means that events scatter uniformly while high kurtosis represents less frequent but more extreme events. Hence, high kurtosis of the residual noise after processing compared to the kurtosis of the unprocessed noise can be interpreted as an indicator for musical noise artifacts. The works in [192–194] perform the musical noise analysis of noise reduction algorithms analytically by means of a white box scenario that is only applicable to a highly limited number of algorithms. In contrast, the works in [181, 196, 197] consider a black box approach that can be applied to any system whose internal processing does not need to be known. In this thesis, we use the weighted log kurtosis ratio (WLKR) as proposed in [197]

$$\mathrm{WLKR} = \log\left(\frac{\mathrm{WK}_{\tilde{d}}}{\mathrm{WK}_{d}}\right), \tag{B.5}$$

where $\mathrm{WK}_{\tilde{d}}$ and WK_{d} are the weighted kurtosis values of $\tilde{d}_{\mathrm{t}}(n)$ and $d_{\mathrm{t}}(n)$, respectively. Here, the kurtosis of a signal refers to the sample kurtosis of its (weighted) spectral magnitudes, calculated from the STFT. The overall WK values are obtained by averaging the kurtosis over frames that are not dominated by speech, as proposed in [196].

B.4 Summary

Table B.1 lists all metrics that have been used to assess the performance of the speech enhancement algorithms presented in this thesis. Please note that the variety of metrics used throughout this thesis is a result from discussions at conferences and reviews that the author received throughout the past years. These steadily helped to improve the instrumental evaluation framework.

Table B.1: Summary of the evaluation metrics used in this thesis.

Metric	Evaluated Quantity	Reference(s)
PESQ	Perceptual speech quality of processed noisy speech	[44, 144, 145, 183]
STOI	Intelligibility of the processed noisy speech	[146]
$\mathrm{NA}_{\mathrm{seg}}$	Noise attenuation of the algorithm under test	[179]
$\widetilde{\mathrm{PESQ}}$	Perceptual speech quality of the filtered speech signal only	[144, 190]
WLKR	Extent of musical noise	[197]

C Circular Statistics

The spectral phase can be seen as directional data in the complex plane. To statistically describe and analyze the spectral phase, we hence use circular/directional statistics in parts of this thesis. This appendix explains fundamental concepts such as circular moments and selected circular distributions.

C.1 Measures of Location and Dispersion - Circular Moments

Representation of Circular Data

We define circular data points as data points on the unit circle, i.e. we may represent a random angle ϕ by means of a complex exponential with unit length

$$z = \mathrm{e}^{j\phi}. \tag{C.1}$$

The angle of this complex exponential is measured in degrees or, e.g. in this thesis, in radians. The usefulness of the representation in Equation (C.1) becomes clear when we consider the problem of taking the sample mean of observed data. Given for example two angles $\phi_1 = \pi + \epsilon$ and $\phi_2 = \pi - \epsilon$ with $0 < \epsilon < \pi/2$, the result of taking the the two angles' sample mean directly will strongly depend on the chosen principal interval, i.e. the interval that angle measurements are wrapped to. We denote the wrapped angle as the pincipal value, which is accessed by applying the $\mathrm{ARG}\left(\cdot\right)$ operator. If we define the principal value of an angle to lie within the interval $[0, 2\pi)$, for the sample mean we have

$$\bar{\phi}_a = \frac{\mathrm{ARG}\left(\phi_1\right) + \mathrm{ARG}\left(\phi_2\right)}{2} = \frac{\pi + \epsilon + \pi - \epsilon}{2} = \pi. \tag{C.2}$$

In contrast, defining the principal interval on $[-\pi, \pi)$ yields

$$\bar{\phi}_b = \frac{\mathrm{ARG}\left(\phi_1\right) + \mathrm{ARG}\left(\phi_2\right)}{2} = \frac{-\pi + \epsilon + \pi - \epsilon}{2} = 0. \tag{C.3}$$

Clearly, the first result ϕ_a is a much more reasonable (or at least practically more relevant) mean value for the observed data, since it gives us an indication about a directional preference in the observed data.

The representation as a complex exponential according to Equation (C.1) is much more practical in this context. We define the sample mean resultant vector as

$$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i. \tag{C.4}$$

While the angle of $\bar{z}$ corresponds to a mean direction, its absolute value (or length) gives us information about the concentration of the data around the mean direction, i.e. the larger $|\bar{z}|$, the higher the concentration, with maximal concentration indicated by $|\bar{z}| = 1$ if all data points come from the same direction. For the aforementioned measurements, we have $\phi_c = \angle z = \phi_a$.

The above discussion is especially relevant in the context of optimization criteria, as in the case of circular data, they should not be formulated on ϕ but on $\mathrm{e}^{j\phi}$, since it enables to straightforwardly take into account the wrapping around the unit circle. In the following, we will present measures of population location and dispersion of circular data.

The First Circular Moment

The mean resultant (vector), or first circular moment, of a random angle ϕ is defined as [160]

$$B = \mathbb{E}\big(z\big) = \mathbb{E}\big(\mathrm{e}^{j\phi}\big). \tag{C.5}$$

It is the population-equivalent to Equation (C.4).

The Mean Direction

The mean angle (or direction) of the random angle ϕ is the angle of the first circular moment, i.e.

$$\mu = \angle B. \tag{C.6}$$

The Mean Resultant Length

The mean resultant length of the random angle ϕ, i.e. the absolute value of the first circular moment, represents the concentration of ϕ around the mean direction μ and is given by $|B|$. The closer $|B|$ is to one, the more concentrated ϕ is around μ.

The Circular Variance

The circular variance is closely connected to the mean resultant length and is defined as [160]

$$\sigma_c^2 = 1 - |B| . \tag{C.7}$$

Hence, similar to $|B|$, it is bounded to the interval $[0, 1]$. However, as it is a variance, ϕ can be considered to be maximally concentrated if $\sigma_c^2 = 0$. If $\sigma_c^2 = 1$, ϕ is uniformly distributed.

C.2 Selected Distributions

According to [160], the pdf $p(\phi)$ of an absolutely continuous circular distribution fulfills the following constraints

1. $p(\phi) \geq 0$ almost everywhere on $(-\infty, \infty)$,
2. $p(\phi + 2\pi) = p(\phi)$ almost everywhere on $(-\infty, \infty)$,
3. $\int_0^{2\pi} p(\phi)\, d\phi = 1$.

C.2.1 The Uniform Distribution

The uniform distribution on the unit circle has the pdf

$$p(\phi) = \frac{1}{2\pi}. \tag{C.8}$$

An important property of the uniform distribution is that the sum of two random variables ϕ_1 and ϕ_2 is uniformly distributed if one of the two is itself uniformly distributed, independent of the other random variable's distribution [160].

Circular Moments

For the uniform distribution, no direction in the complex plane is promoted, i.e. for the circular variance we have

$$\sigma_c^2 = 1. \tag{C.9}$$

C.2.2 The Wrapped Gaussian Distribution

By wrapping any distribution around the circumference of the unit circle, we obtain the corresponding wrapped distribution. For the case of the normal distribution, we hence can compute the pdf of the wrapped normal distribution as [160]

$$p\left(\phi;\mu,\sigma\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{k=-\infty}^{\infty}\mathrm{e}^{-\frac{(\phi-\mu+2\pi k)^2}{2\sigma^2}}, \tag{C.10}$$

with mean direction μ and variance σ^2. For any random variable on the line that has a pdf, the above principle can be used to formulate the pdf of the corresponding wrapped distribution [160].

Circular Moments

The first circular moment of the wrapped normal distribution is given by [160]

$$B=\mathrm{e}^{-\frac{\sigma^2}{2}}\mathrm{e}^{j\mu}, \tag{C.11}$$

hence, the mean direction is μ and the circular variance is given by [160]

$$\sigma_c^2=1-\mathrm{e}^{-\frac{\sigma^2}{2}}. \tag{C.12}$$

C.2.3 The Von Mises Distribution

The pdf of a von Mises distributed random variable is given by [160]

$$p\left(\phi;\mu,\kappa\right)=\frac{1}{2\pi I_0\left(\kappa\right)}\mathrm{e}^{\kappa\cos(\phi-\mu)}, \tag{C.13}$$

with mean direction μ and concentration parameter κ. $I_\nu\left(\cdot\right)$ is the modified Bessel function of ν^{th} order as defined in Equation (D.5). For large κ, the von Mises distribution can be approximated by a normal distribution with variance $\sigma^2=1/\kappa$ [198].

Circular Moments

The first circular moment of the von Mises distribution is given by [160]

$$B=\frac{I_1\left(\kappa\right)}{I_0\left(\kappa\right)}\mathrm{e}^{j\mu}, \tag{C.14}$$

the mean direction is μ and the circular variance is given by [160]

$$\sigma_c^2 = 1 - \frac{I_1\left(\kappa\right)}{I_0\left(\kappa\right)}. \tag{C.15}$$

D

Special Functions and Probability Density Functions

The Gamma Function

The Gamma, or factorial, function can be expressed by Euler's integral [199, Eq. 6.1.1]

$$\Gamma\left(z\right) = \int_0^\infty t^{z-1} \mathrm{e}^{-t} dt, \qquad \text{with } \mathrm{Re}\left(z\right) > 0. \tag{D.1}$$

For $n \in \mathbb{N}$, it fulfills [199, Eq. 6.1.5]

$$\Gamma\left(n+1\right) = n!. \tag{D.2}$$

Bessel Functions

The Bessel function of the first kind is defined by the series [200, Eq. 8.402]

$$J_\nu\left(z\right) = \frac{z^\nu}{2^\nu} \sum_{k=0}^{\infty} \left(-1\right)^k \frac{z^{2k}}{2^{2k} k! \Gamma\left(\nu + k + 1\right)}, \qquad \text{with } \left|\angle z\right| < \pi. \tag{D.3}$$

The modified Bessel function of the first kind is given by [200, Eq. 8.406]

$$I_\nu\left(z\right) = \mathrm{e}^{-j\frac{\pi}{2}\nu} J_\nu\left(\mathrm{e}^{j\frac{\pi}{2}} z\right), \qquad \text{with } -\pi < \angle z \leq \frac{\pi}{2}, \tag{D.4}$$

with integral representation [200, Eq. 8.431]

$$I_\nu\left(z\right) = \frac{1}{\pi} \int_0^\pi \cos\left(\nu\phi\right) \mathrm{e}^{z\cos(\phi)} d\phi - \frac{\sin\left(\nu\pi\right)}{\pi} \int_0^\infty \mathrm{e}^{-z\cosh(t-\nu t)} dt, \tag{D.5}$$

with $\left|\angle z\right| \leq \pi/2$ and $\mathrm{Re}\left(\nu\right) > 0$.

Arctan2

The four-quadrant inverse tangent is defined as

$$\operatorname{arctan2}\left(z_{\mathrm{I}}, z_{\mathrm{R}}\right)=\begin{cases}\arctan\left(\frac{z_{\mathrm{I}}}{z_{\mathrm{R}}}\right) & \text{if } z_{\mathrm{R}}>0, \\ \arctan\left(\frac{z_{\mathrm{I}}}{z_{\mathrm{R}}}\right)+\pi & \text{if } z_{\mathrm{R}}<0 \text{ and } z_{\mathrm{I}} \geq 0, \\ \arctan\left(\frac{z_{\mathrm{I}}}{z_{\mathrm{R}}}\right)-\pi & \text{if } z_{\mathrm{R}}<0 \text{ and } z_{\mathrm{I}}<0, \\ \frac{\pi}{2} & \text{if } z_{\mathrm{R}}=0 \text{ and } z_{\mathrm{I}}>0, \\ -\frac{\pi}{2} & \text{if } z_{\mathrm{R}}=0 \text{ and } z_{\mathrm{I}}<0, \\ \text{indeterminate} & \text{if } z_{\mathrm{R}}=0 \text{ and } z_{\mathrm{I}}=0.\end{cases} \tag{D.6}$$

The Parabolic Cylinder Function

$$D_{-2\mu}(\nu)=\frac{\mathrm{e}^{-\frac{\nu^2}{4}}}{\Gamma(2\mu)} \int_0^{\infty} \mathrm{e}^{-x\nu-\frac{x^2}{2}} x^{2\mu-1} dx \tag{D.7}$$

The Confluent Hypergeometric Function

The confluent hypergeometric function is defined by the series [200, Eq. 9.210]

$$ {}_1F_1(\alpha;\gamma;z)=1+\tfrac{\alpha}{\gamma}\tfrac{z}{1!}+\tfrac{\alpha(\alpha+1)}{\gamma(\gamma+1)}\tfrac{z^2}{2!}+\tfrac{\alpha(\alpha+1)(\alpha+2)}{\gamma(\gamma+1)(\gamma+2)}\tfrac{z^3}{3!}+\ldots \tag{D.8}$$

In the context of this thesis it is relevant for solving integrals including Bessel functions such as [200, Eq. 6.631]

$$\int_0^{\infty} x^{\mu} \mathrm{e}^{-\alpha x^2} J_{\nu}(\beta x)\, dx=\frac{\beta^{\nu}\Gamma\left(\frac{\nu+\mu+1}{2}\right)}{2^{\nu+1}\alpha^{\frac{\mu+\nu+1}{2}}\Gamma(\nu+1)}\, {}_1F_1\left(\tfrac{\nu+\mu+1}{2};\nu+1;-\tfrac{\beta^2}{4\alpha}\right), \tag{D.9}$$

with $\operatorname{Re}(\alpha)>0$ and $\operatorname{Re}(\mu+\nu)>-1$, occurring in STFT magnitude estimators.

The Generalized Gamma Distribution

The pdf of the generalized Gamma distribution is given by [55]

$$p(A)=\frac{\gamma\mu^{\nu}}{\Gamma(\nu)} A^{\gamma\nu-1} \mathrm{e}^{-\mu A^{\gamma}} \qquad \mu>0, \gamma>0, \nu>0, A \geq 0. \tag{D.10}$$

Dependent on the parameter pair $\{\gamma, \nu\}$, special cases of the generalized Gamma distribution can be found. For $\{\gamma = 2, \nu = 1/2\}$, Equation (D.10) corresponds to a Rayleigh density as in [15] and for $\{\gamma = 1, \nu\}$ we have the Gamma density that has been used in [3, 201] and in Equation (4.39) in this thesis.

The χ-Distribution

The pdf of the χ-distribution is given by [27]

$$p\left(A\right) = \frac{2}{\Gamma\left(\mu\right)} \left(\frac{\mu}{\sigma_x^2}\right)^{\mu} A^{2\mu-1} \mathrm{e}^{-\frac{\mu}{\sigma_x^2} A^2}. \tag{D.11}$$

E

Derivations for Chapter 4

E.1 Derivation of the Estimates and Wirtinger Calculus

Inserting Equation (4.15) into Equation (4.11) results in

$$\begin{aligned}\hat{X}_b = \underset{\hat{X}}{\arg\min} \quad & p(\mathcal{H}_0)\lambda_0^b \int_{\mathcal{D}_X} (G_{\min}Y - \hat{X})^*(G_{\min}Y - \hat{X}) p(Y|X) p(X|\mathcal{H}_0) dX \\ & + p(\mathcal{H}_1)\lambda_1^b \int_{\mathcal{D}_X} (X - \hat{X})^*(X - \hat{X}) p(Y|X) p(X|\mathcal{H}_1) dX \\ & + p(\mathcal{H}_2)\lambda_2^b \int_{\mathcal{D}_X} (X - \hat{X})^*(X - \hat{X}) p(Y|X) p(X|\mathcal{H}_2) dX.\end{aligned} \tag{E.1}$$

In the following we will make use of Wirtinger calculus for minimizing Equation (E.1) with respect to the complex variable $\hat{X}$ [202]. By doing so, we avoid considering the real and imaginary part of $\hat{X}$ separately, yielding the same result but a much more straightforward derivation. Since the cost function $f(\hat{z}) = (z - \hat{z})^*(z - \hat{z})$ with $\{z, \hat{z}\} \in \mathbb{C}$ is a real-valued function, its derivative and its conjugate derivative are complex conjugates of each other, i.e., $\left(\frac{\partial f}{\partial \hat{z}}\right)^* = \frac{\partial f}{\partial \hat{z}^*}$ [138]. Consequently, we can compute stationary points of f by simply setting either $\frac{\partial f}{\partial \hat{z}}$ or $\frac{\partial f}{\partial \hat{z}^*}$ to zero, both yielding the same result for $\hat{z}$. This is why we minimize Equation (E.1) by setting its partial derivative with respect to $\hat{X}^*$ to zero while treating $\hat{X}$ as a constant:

$$\begin{aligned} & p(\mathcal{H}_1)\lambda_1^b \int_{\mathcal{D}_X} (X - \hat{X}) p(X|\mathcal{H}_1) p(Y|X) dX + \\ & p(\mathcal{H}_2)\lambda_2^b \int_{\mathcal{D}_X} (X - \hat{X}) p(X|\mathcal{H}_2) p(Y|X) dX + \\ & p(\mathcal{H}_0)\lambda_0^b \int_{\mathcal{D}_X} (G_{\min}Y - \hat{X}) p(X|\mathcal{H}_0) p(Y|X) dX = 0, \end{aligned} \tag{E.2}$$

where the integral over $\mathcal{D}_X$ corresponds to integrating over the real and imaginary parts of X. Further, by using $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(X|\mathcal{H}_0)p(Y|X) dX_{\mathrm{R}} dX_{\mathrm{I}} = p(Y|\mathcal{H}_0)$ and $p(X|\mathcal{H}_0) =$

$\delta(X)$ we obtain

$$\begin{aligned}
&\hat{X}_b\left[\Lambda_1(Y)\lambda_1^b+\Lambda_2(Y)\lambda_2^b+\lambda_0^b\right]= \\
&\Lambda_1(Y)\lambda_1^b\underbrace{\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}Xp(X|\mathcal{H}_1)p(Y|X)dX_{\mathrm{R}}dX_{\mathrm{I}}}{p(Y|\mathcal{H}_1)}}_{\mathbb{E}(X|Y,\mathcal{H}_1)} \\
&+\Lambda_2(Y)\lambda_2^b\underbrace{\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}Xp(X|\mathcal{H}_2)p(Y|X)dX_{\mathrm{R}}dX_{\mathrm{I}}}{p(Y|\mathcal{H}_2)}}_{\mathbb{E}(X|Y,\mathcal{H}_2)} \\
&+\lambda_0^bG_{\min}Y,
\end{aligned} \tag{E.3}$$

with likelihood ratios

$$\Lambda_b(Y)=\frac{p(\mathcal{H}_b)}{p(\mathcal{H}_0)}\frac{p(Y|\mathcal{H}_b)}{p(Y|\mathcal{H}_0)}. \tag{E.4}$$

The conditional expectations $\mathbb{E}(X|Y,\mathcal{H}_1)$ and $\mathbb{E}(X|Y,\mathcal{H}_2)$ can be solved by inserting the distributions resulting from the statistical model in Section 4.3.1:

$$p(X|\mathcal{H}_1)=p(X_{\mathrm{R}},X_{\mathrm{I}}|\mathcal{H}_1)=\frac{1}{\pi\sigma_x^2}\mathrm{e}^{-\frac{X_{\mathrm{R}}^2+X_{\mathrm{I}}^2}{\sigma_x^2}}, \tag{E.5}$$

$$p(X|\mathcal{H}_2)=p(X_{\mathrm{R}},X_{\mathrm{I}}|\mathcal{H}_2)=\frac{1}{\pi\sigma_x^2}\mathrm{e}^{-\frac{(X_{\mathrm{R}}-\bar{X}_{2,\mathrm{R}})^2+(X_{\mathrm{I}}-\bar{X}_{2,\mathrm{I}})^2}{\sigma_x^2}}, \tag{E.6}$$

$$p(Y|X)=p(Y_{\mathrm{R}},Y_{\mathrm{I}}|X_{\mathrm{R}},X_{\mathrm{I}})=\frac{1}{\pi\sigma_d^2}\mathrm{e}^{-\frac{(Y_{\mathrm{R}}-X_{\mathrm{R}})^2+(Y_{\mathrm{I}}-X_{\mathrm{I}})^2}{\sigma_d^2}}, \tag{E.7}$$

where the subscripts “R” and “I” denote the real and imaginary part of a complex-valued variable $z=z_{\mathrm{R}}+jz_{\mathrm{I}}$ and $\bar{X}_2=X_{\mathrm{H}}$ in the case where the mean value is introduced due to harmonic components. In the case of unvoiced speech, we have circularly complex Gaussian distributed speech DFT coefficients known to result in the noisy signal, filtered by the Wiener Filter (G_{W}) [14], as the conditional expectation value:

$$\mathbb{E}(X|Y,\mathcal{H}_1)=\frac{\sigma_x^2}{\sigma_x^2+\sigma_d^2}Y=G_{\mathrm{W}}Y. \tag{E.8}$$

In case of $\mathcal{H}_2$, the conditional expectation value is the noisy observation times the stochastic-deterministic Wiener Filter [40, 117] from Equation (3.25), which also takes into account the harmonic components of the signal

$$\mathbb{E}(X|Y,\mathcal{H}_2)=G_{\mathrm{SDW}}Y. \tag{E.9}$$

Solving for real and imaginary parts separately, finally yields the solution of Equation (E.3) for the complex-valued DFT coefficients,

$$\hat{X}_b = \underbrace{\frac{\lambda_0^b G_{\min} + \Lambda_1(Y)\lambda_1^b G_{\mathrm{W}} + \Lambda_2(Y)\lambda_2^b G_{\mathrm{SDW}}}{\Lambda_1(Y)\lambda_1^b + \Lambda_2(Y)\lambda_2^b + \lambda_0^b}}_{G_b} Y. \tag{E.10}$$

E.2 Derivation of the Subrisks

In the following, we derive the subrisk for a general $\mathcal{H}_{b'}$ with $b' \in \{1, 2\}$ and subsequently insert the correpsonding class-specific parameters to obtain the individual subrisks. This is possible because they share the same cost function. Inserting the cost function together with the distributions into Equation (4.7) yields

$$\begin{aligned}
r_b(\tilde{Y}|\mathcal{H}_{b'}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda_{b'}^b(\tilde{X} - \hat{\tilde{X}}_b)^*(\tilde{X} - \hat{\tilde{X}}_b) \\
&\quad \times \frac{1}{\pi\sigma_d^2}\mathrm{e}^{-\frac{\tilde{Y}_\mathrm{R}^2 + \tilde{Y}_\mathrm{I}^2 + \tilde{X}_\mathrm{R}^2 + \tilde{X}_\mathrm{I}^2 - 2(\tilde{Y}_\mathrm{R}\tilde{X}_\mathrm{R} + \tilde{Y}_\mathrm{I}\tilde{X}_\mathrm{I})}{\sigma_d^2}} \\
&\quad \times \frac{1}{\pi\sigma_x^2}\mathrm{e}^{-\frac{\tilde{X}_\mathrm{R}^2 + \tilde{X}_\mathrm{I}^2 + \bar{X}_{b',\mathrm{R}}^2 + \bar{X}_{b',\mathrm{I}}^2 - 2(\tilde{X}_\mathrm{R}\bar{X}_{b',\mathrm{R}} + \tilde{X}_\mathrm{I}\bar{X}_{b',\mathrm{I}})}{\sigma_x^2}} d\tilde{X}_\mathrm{I} d\tilde{X}_\mathrm{R} \\
&= \frac{\lambda_{b'}^b}{\pi^2\sigma_d^2\sigma_x^2}\mathrm{e}^{-\frac{\tilde{Y}_\mathrm{R}^2 + \tilde{Y}_\mathrm{I}^2}{\sigma_d^2} - \frac{\bar{X}_{b',\mathrm{R}}^2 + \bar{X}_{b',\mathrm{I}}^2}{\sigma_x^2}} \\
&\quad \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda_{b'}^b\Big(\tilde{X}_\mathrm{R}^2 + \tilde{X}_\mathrm{I}^2 + \hat{\tilde{X}}_{b,\mathrm{R}}^2 + \hat{\tilde{X}}_{b,\mathrm{I}}^2 - 2(\tilde{X}_\mathrm{R}\hat{\tilde{X}}_{b,\mathrm{R}} + \tilde{X}_\mathrm{I}\hat{\tilde{X}}_{b,\mathrm{I}})\Big) \\
&\quad \times \mathrm{e}^{-(\tilde{X}_\mathrm{R}^2 + \tilde{X}_\mathrm{I}^2)\frac{\sigma_d^2 + \sigma_x^2}{\sigma_d^2\sigma_x^2} + 2\tilde{X}_\mathrm{R}\frac{\tilde{Y}_\mathrm{R}\sigma_x^2 + \bar{X}_{b',\mathrm{R}}\sigma_d^2}{\sigma_d^2\sigma_x^2} + 2\tilde{X}_\mathrm{I}\frac{\tilde{Y}_\mathrm{I}\sigma_x^2 + \bar{X}_{b',\mathrm{I}}\sigma_d^2}{\sigma_d^2\sigma_x^2}} d\tilde{X}_\mathrm{I} d\tilde{X}_\mathrm{R}
\end{aligned} \tag{E.11}$$

For brevity we substitute:

$$\begin{aligned}
&p = \frac{\sigma_d^2 + \sigma_x^2}{\sigma_d^2\sigma_x^2}, \quad q_\mathrm{R} = \frac{\tilde{Y}_\mathrm{R}\sigma_x^2 + \bar{X}_{b',\mathrm{R}}\sigma_d^2}{\sigma_d^2\sigma_x^2}, \quad \zeta = \frac{\tilde{Y}_\mathrm{R}^2 + \tilde{Y}_\mathrm{I}^2}{\sigma_d^2}, \\
&q_\mathrm{I} = \frac{\tilde{Y}_\mathrm{I}\sigma_x^2 + \bar{X}_{b',\mathrm{I}}\sigma_d^2}{\sigma_d^2\sigma_x^2}, \quad \gamma = \frac{\bar{X}_{b',\mathrm{R}}^2 + \bar{X}_{b',\mathrm{I}}^2}{\sigma_x^2}.
\end{aligned}$$

Hence, after multiplying out we have

$$\begin{aligned}
r_b(\tilde{Y}|\mathcal{H}_{b'}) &= \frac{\lambda_{b'}^b}{\pi^2\sigma_d^2\sigma_x^2}\mathrm{e}^{-\zeta-\gamma} \\
&\times \Big[\int_{-\infty}^{\infty}\mathrm{e}^{-\tilde{X}_{\mathrm{I}}^2p+2\tilde{X}_{\mathrm{I}}q_{\mathrm{I}}}\int_{-\infty}^{\infty}\tilde{X}_{\mathrm{R}}^2\mathrm{e}^{-\tilde{X}_{\mathrm{R}}^2p+2\tilde{X}_{\mathrm{R}}q_{\mathrm{R}}}d\tilde{X}_{\mathrm{R}}d\tilde{X}_{\mathrm{I}} \\
&+ \int_{-\infty}^{\infty}\mathrm{e}^{-\tilde{X}_{\mathrm{R}}^2p+2\tilde{X}_{\mathrm{R}}q_{\mathrm{R}}}\int_{-\infty}^{\infty}\tilde{X}_{\mathrm{I}}^2\mathrm{e}^{-\tilde{X}_{\mathrm{I}}^2p+2\tilde{X}_{\mathrm{I}}q_{\mathrm{I}}}d\tilde{X}_{\mathrm{I}}d\tilde{X}_{\mathrm{R}} \\
&+ (\hat{\tilde{X}}_{b,\mathrm{R}}^2+\hat{\tilde{X}}_{b,\mathrm{I}}^2)\int_{-\infty}^{\infty}\mathrm{e}^{-\tilde{X}_{\mathrm{R}}^2p+2\tilde{X}_{\mathrm{R}}q_{\mathrm{R}}}\int_{-\infty}^{\infty}\mathrm{e}^{-\tilde{X}_{\mathrm{I}}^2p+2\tilde{X}_{\mathrm{I}}q_{\mathrm{I}}}d\tilde{X}_{\mathrm{I}}d\tilde{X}_{\mathrm{R}} \\
&- 2\hat{\tilde{X}}_{b,\mathrm{R}}\int_{-\infty}^{\infty}\mathrm{e}^{-\tilde{X}_{\mathrm{I}}^2p+2\tilde{X}_{\mathrm{I}}q_{\mathrm{I}}}\int_{-\infty}^{\infty}\tilde{X}_{\mathrm{R}}\mathrm{e}^{-\tilde{X}_{\mathrm{R}}^2p+2\tilde{X}_{\mathrm{R}}q_{\mathrm{R}}}d\tilde{X}_{\mathrm{R}}d\tilde{X}_{\mathrm{I}} \\
&- 2\hat{\tilde{X}}_{b,\mathrm{I}}\int_{-\infty}^{\infty}\mathrm{e}^{-\tilde{X}_{\mathrm{R}}^2p+2\tilde{X}_{\mathrm{R}}q_{\mathrm{R}}}\int_{-\infty}^{\infty}\tilde{X}_{\mathrm{I}}\mathrm{e}^{-\tilde{X}_{\mathrm{I}}^2p+2\tilde{X}_{\mathrm{I}}q_{\mathrm{I}}}d\tilde{X}_{\mathrm{I}}d\tilde{X}_{\mathrm{R}}\Big]
\end{aligned} \tag{E.12}$$

From [200] (3.323, 3.462) we have:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\mathrm{e}^{-px^2+2qx}dx = \sqrt{\frac{\pi}{p}}\mathrm{e}^{\frac{q^2}{p}} \\
&\int_{-\infty}^{\infty}x\mathrm{e}^{-px^2+2qx}dx = \frac{q}{p}\sqrt{\frac{\pi}{p}}\mathrm{e}^{\frac{q^2}{p}} \\
&\int_{-\infty}^{\infty}x^2\mathrm{e}^{-px^2+2qx}dx = \frac{1}{2p}\sqrt{\frac{\pi}{p}}(1+2\frac{q^2}{p})\mathrm{e}^{\frac{q^2}{p}}
\end{aligned}$$

Since all integrals we have to solve take one of the above forms we can simply insert the solutions:

$$\begin{aligned}
r_b(\tilde{Y}|\mathcal{H}_{b'}) &= \frac{\lambda_{b'}^b}{\pi^2\sigma_d^2\sigma_x^2}\mathrm{e}^{-\zeta-\gamma+\frac{q_{\mathrm{R}}^2+q_{\mathrm{I}}^2}{p}}\frac{\pi}{p} \\
&\times \Big[\frac{1}{p}(1+\frac{q_{\mathrm{R}}^2+q_{\mathrm{I}}^2}{p}) - 2\Big(\hat{\tilde{X}}_{b,\mathrm{R}}\frac{q_{\mathrm{R}}}{p}+\hat{\tilde{X}}_{b,\mathrm{I}}\frac{q_{\mathrm{I}}}{p}\Big)+\hat{\tilde{X}}_{b,\mathrm{R}}^2+\hat{\tilde{X}}_{b,\mathrm{I}}^2\Big] \\
&= \lambda_{b'}^b\underbrace{\frac{1}{\pi(\sigma_d^2+\sigma_x^2)}\mathrm{e}^{-\frac{|\tilde{Y}-\bar{X}_{b'}|^2}{\sigma_x^2+\sigma_d^2}}}_{p(\tilde{Y}|\mathcal{H}_{b'})} \\
&\times \Big[\underbrace{\frac{1}{p}(1+\frac{q_{\mathrm{R}}^2}{p}+\frac{q_{\mathrm{I}}^2}{p})}_{G_{\mathrm{W}}\sigma_d^2+|G_{\mathrm{W}}\tilde{Y}+G_{\mathrm{DW}}\bar{X}_{b'}|^2} -2\Big(\hat{\tilde{X}}_{b,\mathrm{R}}\frac{q_{\mathrm{R}}}{p}+\hat{\tilde{X}}_{b,\mathrm{I}}\frac{q_{\mathrm{I}}}{p}\Big)+|\hat{\tilde{X}}_b|^2\Big]
\end{aligned} \tag{E.13}$$

with $q_\mathrm{R}/p = G_\mathrm{W}\tilde{Y}_\mathrm{R} + G_\mathrm{DW}\bar{X}_{b',\mathrm{R}}$ (analogously for q_I/p) we have

$$\begin{aligned} r_b(\tilde{Y}|\mathcal{H}_{b'}) &= \lambda^b_{b'} p(\tilde{Y}|\mathcal{H}_{b'})\Big(G_\mathrm{W}\sigma_d^2 + |G_\mathrm{W}\tilde{Y} + G_\mathrm{DW}\bar{X}_{b'} - \hat{\tilde{X}}_b|^2\Big) \\ &= \lambda^b_{b'} p(\tilde{Y}|\mathcal{H}_{b'})\Big(G_\mathrm{W}\sigma_d^2 + |G_\mathrm{W} + G_\mathrm{DW}G_{b',\mathrm{mean}} - G_b|^2|\tilde{Y}|^2\Big). \end{aligned} \tag{E.14}$$

By inserting $b' = 1$ and $b' = 2$ we have $G_{1,\mathrm{mean}} = 0$ as well as $G_\mathrm{SDW} = G_\mathrm{W} + G_\mathrm{DW}G_{2,\mathrm{mean}}$ and obtain Equation (4.23) and Equation (4.24), respectively.

E.3 MAP Estimates of Mean Amplitude and Phase

For the sake of a clearer notation, in the following we will again drop the frequency index k, while of course the dependency of all distribution parameters on this quantity remains. First, using Bayes theorem and the fact that the evidence is independent of the parameters we optimize together with the independence assumption of amplitude and phase we search for

$$\underset{\{A_\mathrm{H}(\ell),\alpha_\mathrm{H}(\ell)\}}{\arg\max}\; p\left(\tilde{\mathbf{y}}(\ell)|A_\mathrm{H}(\ell),\alpha_\mathrm{H}(\ell)\right) p\left(A_\mathrm{H}(\ell),\alpha_\mathrm{H}(\ell);\mathcal{H}_2\right), \tag{E.15}$$

i.e.

$$\underset{\{A_\mathrm{H}(\ell),\alpha_\mathrm{H}(\ell)\}}{\arg\max} \frac{\mu(k)^{\nu(k)}}{2\pi\Gamma(\nu(k))} A_\mathrm{H}(k,\ell)^{\nu(k)-1} \frac{\mathrm{e}^{-\mu(k)A_\mathrm{H}(k,\ell)}}{2\pi} \prod_{u=0}^{N_\mathrm{o}-1} \frac{\mathrm{e}^{-\frac{|\widetilde{Y}(k,\ell-u) - A_\mathrm{H}(k,\ell)\mathrm{e}^{j\alpha_\mathrm{H}(k,\ell)}|^2}{\sigma_d^2(k,\ell-u)+\sigma_x^2(k,\ell-u)}}}{\pi(\sigma_d^2(k,\ell-u) + \sigma_x^2(k,\ell-u))}. \tag{E.16}$$

E.3.1 Mean Amplitude

Taking the logarithm of Equation (E.16), differentiating with respect to $A_\mathrm{H}(\ell)$, and setting the so-obtained expression to zero yields

$$\begin{aligned} 0 &= \frac{\partial \log\big(p(A_\mathrm{H}(\ell),\alpha_\mathrm{H}(\ell)|\tilde{\mathbf{y}}(\ell);\mathcal{H}_2)\big)}{\partial A_\mathrm{H}(\ell)} \\ 0 &= -\sum_{q=0}^{N_\mathrm{o}-1} \frac{2A_\mathrm{H}(\ell) - 2\left|\widetilde{Y}(\ell-q)\right|\cos\left(\angle\widetilde{Y}(\ell-q) - \alpha_\mathrm{H}(\ell)\right)}{\sigma_d^2(\ell-q) + \sigma_x^2(\ell-q)} + \frac{\nu-1}{A_\mathrm{H}(\ell)} - \frac{\mu}{C(\ell)} \\ 0 &= -\frac{A_\mathrm{H}^2(\ell)}{\sigma_{N_\mathrm{o}}^2(\ell)} + A_\mathrm{H}(\ell)\sum_{q=0}^{N_\mathrm{o}-1} \frac{\left|\widetilde{Y}(\ell-q)\right|\cos\left(\angle\widetilde{Y}(\ell-q) - \alpha_\mathrm{H}(\ell)\right)}{\sigma_d^2(\ell-q) + \sigma_x^2(\ell-q)} + \frac{\nu-1}{2} - A_\mathrm{H}(\ell)\frac{\mu}{2C(\ell)} \end{aligned} \tag{E.17}$$

where

$$\frac{1}{\sigma^2_{N_\mathrm{o}}(\ell)} = \sum_{q=0}^{N_\mathrm{o}-1} \frac{1}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)} \tag{E.18}$$

By defining

$$2u(\ell) = \sigma^2_{N_\mathrm{o}}(\ell) \left(\sum_{q=0}^{N_\mathrm{o}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \cos\left(\angle\widetilde{Y}(\ell-q) - \alpha_\mathrm{H}(\ell)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)} - \frac{\mu}{2C(\ell)} \right), \tag{E.19}$$

we obtain

$$A^2_\mathrm{H}(\ell) - 2A_\mathrm{H}(\ell)u(\ell) - \sigma^2_{N_\mathrm{o}}(\ell)\frac{\nu-1}{2} = 0, \tag{E.20}$$

which is a simple quadratic equation to solve for $A_\mathrm{H}(\ell)$, resulting in Equation (4.49).

E.3.2 Mean Phase

Similarily we can maximize the posterior for $\alpha_\mathrm{H}(\ell)$, i.e.

$$0 = \frac{\partial \log\big(p(A_\mathrm{H}(\ell), \alpha_\mathrm{H}(\ell)|\tilde{\mathbf{y}}(\ell);\mathcal{H}_2)\big)}{\partial \alpha_\mathrm{H}(\ell)}, \tag{E.21}$$

using the identity

$$\sin(x-y) = \sin(x)\cos(y) - \cos(x)\sin(y) \tag{E.22}$$

we have

$$\sum_{q=0}^{N_\mathrm{o}-1} \frac{2A(\ell)\left|\widetilde{Y}(\ell-q)\right|}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)} \left(\sin\left(\angle\widetilde{Y}(\ell-q)\right)\cos(\alpha_\mathrm{H}(\ell)) - \cos\left(\angle\widetilde{Y}(\ell-q)\right)\sin(\alpha_\mathrm{H}(\ell))\right) = 0, \tag{E.23}$$

since both $\alpha_{\mathrm{H}}(\ell)$ and $\mathrm{A}_{\mathrm{H}}(\ell)$ are independent of q we can rewrite the above equation as

$$\begin{aligned}
\cos\left(\alpha_{\mathrm{H}}(\ell)\right) \sum_{q=0}^{N_{\mathrm{o}}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \sin\left(\angle\widetilde{Y}(\ell-q)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)} &= \sin\left(\alpha_{\mathrm{H}}(\ell)\right) \sum_{q=0}^{N_{\mathrm{o}}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \cos\left(\angle\widetilde{Y}(\ell-q)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)}, \\
\frac{\sin\left(\alpha_{\mathrm{H}}(\ell)\right)}{\cos\left(\alpha_{\mathrm{H}}(\ell)\right)} &= \frac{\sum_{q=0}^{N_{\mathrm{o}}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \sin\left(\angle\widetilde{Y}(\ell-q)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)}}{\sum_{q=0}^{N_{\mathrm{o}}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \cos\left(\angle\widetilde{Y}(\ell-q)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)}}, \\
\tan\left(\alpha_{\mathrm{H}}(\ell)\right) &= \frac{\sum_{q=0}^{N_{\mathrm{o}}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \sin\left(\angle\widetilde{Y}(\ell-q)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)}}{\sum_{q=0}^{N_{\mathrm{o}}-1} \frac{\left|\widetilde{Y}(\ell-q)\right| \cos\left(\angle\widetilde{Y}(\ell-q)\right)}{\sigma_d^2(\ell-q)+\sigma_x^2(\ell-q)}},
\end{aligned} \tag{E.24}$$

which is easily solved for $\alpha_{\mathrm{H}}(\ell)$ by applying $\arctan2\left(\cdot,\cdot\right)$ on the right hand side of Equation (E.24).

Bibliography

[1] J. Stahl and P. Mowlaee, "Iterative harmonic speech enhancement," in *Proc. ITG Symposium on Speech Communication*, Oct 2016, pp. 1–5.

[2] ——, "A pitch-synchronous simultaneous detection-estimation framework for speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 436–450, Feb 2018.

[3] ——, "A simple and effective framework for a priori SNR estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2018, pp. 5644–5648.

[4] J. Stahl, S. U. N. Wood, and P. Mowlaee, "Overcoming covariance matrix phase sensitivity in single-channel speech enhancement," in *Proc. ITG Symposium on Speech Communication*, Oct 2018, pp. 286–290.

[5] J. Stahl and P. Mowlaee, "Exploiting temporal correlation in pitch-adaptive speech enhancement," *submitted to Speech Communication*, 2018.

[6] J. Stahl, S. U. N. Wood, and P. Mowlaee, "Single-channel speech enhancement with correlated spectral components: Limits - potential," *submitted to IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2018.

[7] J. Stahl, P. Mowlaee, and J. Kulmer, "Phase-processing for voice activity detection: A statistical approach," in *Proc. European Signal Processing Conf.*, Aug 2016, pp. 1202–1206.

[8] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice.* John Wiley & Sons, Ltd, 2016.

[9] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, "Phase-aware signal processing for automatic speech recognition," in *Proc. Interspeech*, 2016, pp. 3374–3378.

[10] M. Pirolt, J. Stahl, P. Mowlaee, V. I. Vorobiov, S. Y. Barysenka, and A. G. Davydov, "Phase estimation in single-channel speech enhancement using phase invariance constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2017, pp. 5585–5589.

[11] P. Mowlaee, J. Stahl, and J. Kulmer, "Iterative joint map single-channel speech enhancement given non-uniform phase prior," *Speech Commun.*, vol. 86, pp. 85 – 96, 2017.

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[13] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr 1980.

[14] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[15] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[16] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[17] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sept 2005.

[18] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[19] D. Wang and J. Lim, “The unimportance of phase in speech enhancement,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, August 1982.

[20] M. R. Weiss, A. E. Aschkenasy, and T. W. Parsons, “Study and development of the intel technique for improving speech intelligibility,” Nicolet Scientific Corp., Tech. Rep., 1974.

[21] K. Paliwal, K. Wòjcicki, and B. Shannon, “The importance of phase in speech enhancement,” *Speech Commun.*, vol. 53, no. 4, pp. 465 – 494, 2011.

[22] P. Mowlaee and R. Martin, “On phase importance in parameter estimation for single-channel source separation,” in *Proc. International Workshop on Acoustic Signal Enhancement.*, Sept 2012, pp. 1–4.

[23] P. Mowlaee and R. Saeidi, “Iterative closed-loop phase-aware single-channel speech enhancement,” *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec 2013.

[24] P. Mowlaee and J. Kulmer, “Phase estimation in single-channel speech enhancement: Limits-potential,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1283–1294, Aug 2015.

[25] ——, “Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1521–1532, Sept 2015.

[26] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, Feb 2013.

[27] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, Aug 2014.

[28] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, Dec 2014.

[29] M. Krawczyk-Becker and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2251–2262, Dec 2016.

[30] ——, "An evaluation of the perceptual quality of phase-aware single-channel speech enhancement," vol. 140, pp. 364–369, Oct. 2016.

[31] ——, "A study on the benefits of phase-aware speech enhancement in challenging noise scenarios," in *Latent Variable Analysis and Signal Separation*, Y. Deville, S. Gannot, R. Mason, M. D. Plumbley, and D. Ward, Eds. Cham: Springer International Publishing, 2018, pp. 407–416.

[32] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Commun.*, vol. 81, pp. 1 – 29, 2016, Phase-Aware Signal Processing in Speech Communication.

[33] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, March 2015.

[34] M. Krawczyk-Becker and T. Gerkmann, "On speech enhancement under PSD uncertainty," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1144–1153, June 2018.

[35] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[36] M. Kolbæk, Z. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," 2018.

[37] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Aug. 2017, pp. 3642–3646.

[38] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 6, pp. 672–680, December 1980.

[39] M. Krawczyk-Becker, "Phase-aware single-channel speech enhancement," Ph.D. dissertation, 2016.

[40] M. C. McCallum, "Single-channel statistical Bayesian short-time Fourier transform speech enhancement with deterministic *A Priori* information," Ph.D. dissertation, University of Auckland, 2015.

[41] T. Esch, "Model-based speech enhancement exploiting temporal and spectral dependencies," Ph.D. dissertation, RWTH Aachen University, 2012.

[42] E. Plourde, "Bayesian short-time spectral amplitude estimators for single-channel speech enhancement," Ph.D. dissertation, McGill University, Montreal, 2009.

[43] M. M. Goodwin, "Adaptive signal models: Theory, algorithms, and audio applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[44] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[45] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul 1995.

[46] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[47] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art.* Morgan & Claypool Publishers, 2013, vol. 9, no. 1.

[48] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[49] W. B. Kleijn, "A frame interpretation of sinusoidal coding and waveform interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2000, pp. 1475–1478.

[50] P. Mowlaee, M. Blass, and W. B. Kleijn, "New results in modulation-domain single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2125–2137, Nov 2017.

[51] T. Gülzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, vol. 64, no. 1, pp. 5 – 19, 1998.

[52] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr 1984.

[53] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct 1992.

[54] T. Gerkmann and R. Martin, “Empirical distributions of DFT-domain speech coefficients based on estimated speech variances,” in *Proc. International Workshop on Acoustic Signal Enhancement.*, 2010.

[55] 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 Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug 2007.

[56] R. C. Hendriks, R. Heusdens, and J. Jensen, “An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 406–415, Feb 2007.

[57] R. Rehr and T. Gerkmann, “On the importance of super-Gaussian speech priors for machine-learning based speech enhancement,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 357–366, Feb 2018.

[58] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, ser. McGraw-Hill series in electrical engineering: Communications and signal processing. McGraw-Hill, 2002.

[59] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment.* John Wiley & Sons, 2006.

[60] T. Fingscheidt, C. Beaugeant, and S. Suhadi, “Overcoming the statistical independence assumption w.r.t. frequency in speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, March 2005, pp. 1081–1084.

[61] I. Cohen, “Relaxed statistical model for speech enhancement and a priori SNR estimation,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[62] J. Benesty and Y. Huang, “A single-channel noise reduction MVDR filter,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2011, pp. 273–276.

[63] Y. A. Huang and J. Benesty, “A multi-frame approach to the frequency-domain single-channel noise reduction problem,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[64] H. Huang, L. Zhao, J. Chen, and J. Benesty, “A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction,” *Digital Signal Processing*, vol. 33, pp. 169–179, 2014.

[65] E. Ollila, “On the circularity of a complex random variable,” *IEEE Signal Process. Lett.*, vol. 15, pp. 841–844, 2008.

[66] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, June 1978.

[67] P. J. Wolfe and S. J. Godsill, “Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement,” *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 10, p. 910167, Sep 2003.

[68] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement.* John Wiley & Sons Ltd, 2018.

[69] L. Ying, *Phase Unwrapping.* American Cancer Society, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780471740360.ebs1356

[70] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[71] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[72] R. C. Hendriks, R. Heusdens, and J. Jensen, “MMSE based noise PSD tracking with low complexity,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4266–4269.

[73] R. C. Hendriks, J. S. Erkelens, and R. Heusdens, “Comparison of complex-DFT estimators with and without the independence assumption of real and imaginary parts,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 4033–4036.

[74] C. Breithaupt and R. Martin, “Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-snr and transient conditions,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 2, pp. 277–289, Feb 2010.

[75] I. Cohen, “Speech enhancement using a noncausal a priori SNR estimator,” *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, Sept 2004.

[76] C. Plapous, C. Marro, and P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2098–2108, Nov 2006.

[77] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar 2008, pp. 4897–4900.

[78] S. Suhadi, C. Last, and T. Fingscheidt, “A data-driven approach to a priori SNR estimation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 186–195, Jan 2011.

[79] M. Djendi and P. Scalart, “Reducing over- and under-estimation of the a priori SNR in speech enhancement techniques,” *Digital Signal Processing*, vol. 32, no. Supplement C, pp. 124 – 136, 2014.

[80] Z. Chen and V. Hohmann, “Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1904–1916, Nov 2015.

[81] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, “An iterative speech model-based a priori SNR estimator,” in *Proc. Interspeech*, 2015, pp. 1740–1744.

[82] A. Chinaev and R. Haeb-Umbach, “A priori SNR estimation using a generalized decision directed approach,” in *Proc. Interspeech*, 2016, pp. 3758–3762.

[83] A. Chinaev, J. Heitkaemper, and R. Haeb-Umbach, “A priori SNR estimation using Weibull mixture model,” in *Proc. ITG Symposium on Speech Communication*, 2016, pp. 1–5.

[84] J. Benesty, J. Chen, and Y. Huang, “A widely linear distortionless filter for single-channel noise reduction,” *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 469–472, May 2010.

[85] J. Benesty, J. Chen, and Y. A. Huang, “On widely linear Wiener and tradeoff filters for noise reduction,” *Speech Commun.*, vol. 52, no. 5, pp. 427 – 439, 2010.

[86] M. McCallum and B. Guillemin, “Stochastic-deterministic MMSE STFT speech enhancement with general a priori information,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 7, pp. 1445–1457, July 2013.

[87] T. Lotter and P. Vary, “Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model,” *EURASIP J. on Advances in Signal Processing*, no. 7, May 2005.

[88] J. Jensen and R. Heusdens, “Improved subspace-based single-channel speech enhancement using generalized super-Gaussian priors,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 862–872, March 2007.

[89] T. Gerkmann and R. Martin, “On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling,” *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov 2009.

[90] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[91] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, July 2005.

[92] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 4037–4040.

[93] A. Walther, "The question of phase retrieval in optics," *Optica Acta: International Journal of Optics*, vol. 10, no. 1, pp. 41–49, 1963.

[94] J. Fienup, "Phase retrieval algorithms: a comparison," *J. Opt. Soc. Am. A*, vol. 21, pp. 2758–69, 08 1982.

[95] J. C. Dainty and J. Fienup, "Phase retrieval and image reconstruction for astronomy," *Image Recovery: Theory Appl.*, vol. 13, 1987.

[96] R. P. Millane, "Phase retrieval in crystallography and optics," *J. Opt. Soc. Am. A*, vol. 7, no. 3, pp. 394–411, Mar 1990.

[97] S. Nawab, T. Quatieri, and J. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 4, pp. 986–998, August 1983.

[98] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[99] V. Gnann and M. Spiertz, "Inversion of short-time fourier transform magnitude spectrograms with adaptive window lengths," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 2009, pp. 325–328.

[100] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency."

[101] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 421–424, May 2010.

[102] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 1, pp. 178–185, Jan 2013.

[103] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction." in *Proc. Interspeech*, 2008.

[104] J. L. Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, March 2013.

[105] M. Portnoff, “Magnitude-phase relationships for short-time Fourier transforms based on Gaussian analysis windows,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, April 1979, pp. 186–189.

[106] S. Shimauchi, S. Kudo, Y. Koizumi, and K. Furuya, “On relationships between amplitude and phase of short-time Fourier transform,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2017, pp. 676–680.

[107] Z. Průša, P. Balazs, and P. L. Søndergaard, “A noniterative method for reconstruction of phase from STFT magnitude,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1154–1164, May 2017.

[108] Z. Průša and N. Holighaus, “Non-iterative filter bank phase (re)construction,” in *Proc. European Signal Processing Conf.*, Aug 2017, pp. 922–926.

[109] K. Jaganathan, Y. C. Eldar, and B. Hassibi, “Phase retrieval: An overview of recent developments,” *CoRR*, vol. abs/1510.07713, 2015. [Online]. Available: http://arxiv.org/abs/1510.07713

[110] I. Stylianou, “Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification,” Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, 1996.

[111] Y. Stylianou, “Decomposition of speech signals into a deterministic and a stochastic part,” in *Proc. Int. Conf. Spoken Language*, vol. 2, Oct 1996, pp. 1213–1216 vol.2.

[112] ——, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan 2001.

[113] J. Jensen and J. H. L. Hansen, “Speech enhancement using a constrained iterative sinusoidal model,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 731–740, Oct 2001.

[114] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, “Instantaneous a priori SNR estimation by cepstral excitation manipulation,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug 2017.

[115] J. Kulmer and P. Mowlaee, “Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 2015, pp. 5063–5067.

[116] S. M. Nørholm, “Enhancement of speech signals-with a focus on voiced speech models,” Ph.D. dissertation, Aalborg Universitet, 2015.

[117] M. McCallum and B. Guillemin, “Joint stochastic-deterministic Wiener filtering with recursive Bayesian estimation of deterministic speech,” in *Proc. Interspeech*, 2013, pp. 460–464.

[118] W. B. Kleijn and D. Talkin, "Compact speech representations for speech synthesis," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.*, Sept 2002, pp. 35–38.

[119] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement and noise statistics estimation for non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 645–658, April 2016.

[120] ——, "Enhancement of non-stationary speech using harmonic chirp filters," in *Proc. Interspeech*, 2015.

[121] S. Kay, "Fundamentals of statistical processing, vol. i: Estimation theory," *America: Prentice Hall PTR*, 1993.

[122] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the estimation of the speech harmonic model," in *Proc. Interspeech*, 2008.

[123] Y. Pantazis, "Decomposition of am-fm signals with applications in speech processing," Ph.D. dissertation, University of Crete, 2010.

[124] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, 03 2009, vol. 5.

[125] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 38, Oct 2014.

[126] S. M. Nørholm, M. Krawczyk-Becker, T. Gerkmann, S. Par, J. R. Jensen, and M. G. Christensen, "Least squares estimate of the initial phases in STFT based speech enhancement," in *Proc. Interspeech*, 2015.

[127] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, August 1986.

[128] K. N. Hamdy, M. Ali, and A. H. Tewfi, "Low bit rate high quality audio coding with combined harmonic and wavelet representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 1996, pp. 1045–1048 vol. 2.

[129] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May 2015.

[130] Y. Ephraim and D. Malah, "Speech enhancement using vector spectral subtraction amplitude estimation," 03 1983, p. 13th IEEE Convention of Electrical & Electronics Engineers.

[131] E. Plourde and B. Champagne, "Multidimensional STSA estimators for speech enhancement with correlated spectral components," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3013–3024, July 2011.

[132] M. Krawczyk-Becker and T. Gerkmann, “Fundamental frequency informed speech enhancement in a flexible statistical framework,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 940–951, May 2016.

[133] A. Abramson and I. Cohen, “Simultaneous detection and estimation approach for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2348–2359, Nov 2007.

[134] H. Momeni, H. R. Abutalebi, and A. Tadaion, “Joint detection and estimation of speech spectral amplitude using noncontinuous gain functions,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 8, pp. 1249–1258, 2015.

[135] D. Middleton and R. Esposito, “Simultaneous optimum detection and estimation of signals in noise,” *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 434–444, May 1968.

[136] A. Fredriksen, D. Middleton, and V. VandeLinde, “Simultaneous signal detection and estimation under multiple hypotheses,” *IEEE Trans. Inf. Theory*, vol. 18, no. 5, pp. 607–614, Sep 1972.

[137] F. J. Harris, “On the use of windows for harmonic analysis with the discrete Fourier transform,” *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan 1978.

[138] T. Adali and P. J. Schreier, “Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation,” *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 112–128, Sept 2014.

[139] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT acoustic phonetic continuous speech corpus CDROM,” 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[140] S. Gonzalez, “Pitch of the core timit database set.” [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip

[141] S. Gonzalez and M. Brookes, “PEFAC - a pitch estimation algorithm robust to high levels of noise,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[142] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX–92 Study on the Effect of Additive Noise on Automatic Speech Recognition,” *Technical Report, DRA Speech Research Unit*, 1992.

[143] M. Haberkorn, “Raindrops on plastic.” [Online]. Available: https://www.freesound.org/people/mmorast/sounds/192149/

[144] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” Tech. Rep., 2001.

[145] "ITU-T P.862.2 Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Tech. Rep., 2007.

[146] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[147] "ITU-R BS.1534-2 Method for the subjective assessment of intermediate quality level of audio systems," Tech. Rep., 2014.

[148] E. Vincent, "MUSHRAM: A MATLAB interface for mushra listening tests," Tech. Rep., 2005.

[149] D. Fischer, S. Doclo, E. A. P. Habets, and T. Gerkmann, "Combined single-microphone Wiener and MVDR filtering based on speech interframe correlations and speech presence probability," in *Proc. ITG Symposium on Speech Communication*, Oct 2016, pp. 1–5.

[150] A. Schasse and R. Martin, "Online inter-frame correlation estimation methods for speech enhancement in frequency subbands," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2013, pp. 7482–7486.

[151] ——, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1355–1365, Sept 2014.

[152] T. Esch and P. Vary, "Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 4877–4880.

[153] ——, "Exploiting temporal correlation of speech and noise magnitudes using a modified Kalman filter for speech enhancement," in *Proc. ITG Symposium on Speech Communication*, Oct 2008, pp. 1–4.

[154] E. Zavarehei, S. Vaseghi, and Q. Yan, "Temporal modelling and Kalman filtering of DFT trajectories for enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2006, pp. 481–484.

[155] N. Dionelis and M. Brookes, "Phase-aware single-channel speech enhancement with modulation-domain Kalman filtering," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 937–950, May 2018.

[156] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2064–2074, Nov 2006.

[157] T. Esch, M. Rüngeler, F. Heese, and P. Vary, "Combined reduction of time varying harmonic and stationary noise using frequency warping," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov 2010, pp. 533–537.

[158] G. Doblinger, “Smoothing of noisy AR signals using an adaptive Kalman filter,” in *Proc. European Signal Processing Conf.*, Sept 1998, pp. 1–4.

[159] N. Dionelis and M. Brookes, “Speech enhancement using modulation-domain Kalman filtering with active speech level normalized log-spectrum global priors,” in *Proc. European Signal Processing Conf.*, pp. 2309–2313.

[160] K. V. Mardia and P. E. Jupp, *Directional Statistics.* John Wiley & Sons, Ltd, 2004.

[161] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, “Binaural speech enhancement using a codebook based approach,” in *Proc. International Workshop on Acoustic Signal Enhancement.*, Sept 2016, pp. 1–5.

[162] ——, “Model based binaural enhancement of voiced and unvoiced speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2017, pp. 666–670.

[163] N. D. Gaubitch, H. W. Loellmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, “Performance comparison of algorithms for blind reverberation time estimation from speech,” in *Proc. International Workshop on Acoustic Signal Enhancement.*, Sept 2012, pp. 1–4.

[164] “ITU-T P.835 Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,” Tech. Rep., 2011.

[165] D. S. Williamson and D. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, July 2017.

[166] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, “End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, Sept 2018.

[167] P. C. Loizou and G. Kim, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 47–56, Jan 2011.

[168] J. Stahl and P. Mowlaee, “Exploiting temporal correlation in pitch-adaptive speech enhancement.” [Online]. Available: http://www2.spsc.tugraz.at/people/pmowlaee/PACO

[169] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, “webMUSHRA – A Comprehensive Framework for Web-based Listening Tests,” *Journal of Open Research Software*, vol. 6, no. 1, p. 8, 2018.

[170] C. Li and S. V. Andersen, “A block-based linear MMSE noise reduction with a high temporal resolution modeling of the speech excitation,” *EURASIP J. on Applied Signal Processing*, no. 18, Nov 2005.

[171] E. Plourde and B. Champagne, “A family of Bayesian STSA estimators for the enhancement of speech with correlated frequency components,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4766–4769.

[172] J. Chen and J. Benesty, “Single-channel noise reduction in the STFT domain based on the bifrequency spectrum,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2012, pp. 97–100.

[173] H. Momeni, H. R. Abutalebi, and E. A. P. Habets, “Conditional MMSE-based single-channel speech enhancement using inter-frame and inter-band correlations,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2016, pp. 5215–5219.

[174] I. Andrianakis and P. White, “Speech spectral amplitude estimators using optimally shaped gamma and chi priors,” *Speech Commun.*, vol. 51, no. 1, pp. 1 – 14, 2009.

[175] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing.* New York, NY, USA: Cambridge University Press, 1988.

[176] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, July 2006.

[177] J. Stahl, S. U. N. Wood, and P. Mowlaee, “Overcoming covariance matrix phase sensitivity in single-channel speech enhancement with correlated spectral components.” [Online]. Available: https://gitlab.com/johannesstahl/LMDSTSA/

[178] B. D. Carlson, “Covariance matrix estimation errors and diagonal loading in adaptive arrays,” *IEEE Aerosp. Electron. Syst. Mag.*, vol. 24, no. 4, pp. 397–401, Jul 1988.

[179] T. Fingscheidt, S. Suhadi, and S. Stan, “Environment-optimized speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[180] J. Stahl, S. U. N. Wood, and P. Mowlaee, “Overcoming covariance matrix phase sensitivity in single-channel speech enhancement with correlated spectral components.” [Online]. Available: http://www2.spsc.tugraz.at/people/pmowlaee/ITG2018.html

[181] H. Yu and T. Fingscheidt, “Black box measurement of musical tones produced by noise reduction systems,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2012, pp. 4573–4576.

[182] A. Sugiyama and R. Miyahara, “Phase randomization - a new paradigm for single-channel signal enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2013, pp. 7487–7491.

[183] “ITU-T P.862 Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” Tech. Rep., 2001.

[184] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[185] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4214–4217.

[186] T. Fingscheidt and S. Suhadi, “Quality assessment of speech enhancement systems by separation of enhanced speech, noise, and echo,” in *Proc. Interspeech*, 2007, pp. 818–821.

[187] T. Lotter, “Single and multimicrophone speech enhancement for hearing aids,” Ph.D. dissertation, RWTH Aachen University, 2004.

[188] S. Gustafsson, R. Martin, and P. Vary, “Combined acoustic echo control and noise reduction for hands-free telephony,” *Signal Processing*, vol. 64, no. 1, pp. 21 – 32, 1998.

[189] K. Steinert, S. Suhadi, T. Fingscheidt, and M. Schönle, “Instrumental speech distortion assessment of black box speech enhancement systems,” *Proc. International Workshop on Acoustic Signal Enhancement.*, 2008.

[190] T. Fingscheidt, S. Suhadi, and K. Steinert, “Towards objective quality assessment of speech enhancement systems in a black box approach,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 273–276.

[191] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, “DNN-supported speech enhancement with cepstral estimation of both excitation and envelope,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2460–2474, Dec 2018.

[192] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, “Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics,” in *Proc. International Workshop on Acoustic Signal Enhancement.*, 2008.

[193] ——, “Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 2009, pp. 4433–4436.

[194] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, “Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 6, pp. 1770–1779, Aug 2011.

[195] D. J. Field, “What is the goal of sensory coding?” *Neural Comput.*, vol. 6, no. 4, pp. 559–601, Jul. 1994.

[196] H. Yu and T. Fingscheidt, “A figure of merit for instrumental optimization of noise reduction algorithms,” in *Proc. of 5th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems, Kiel, Germany*, 2011.

[197] ——, “A weighted log kurtosis ratio measure for instrumental musical tones assessment in wideband speech,” in *Proc. ITG Symposium on Speech Communication*, Sept 2012, pp. 1–4.

[198] N. I. Fisher, T. Lewis, and B. J. J. Embleton, *Statistical analysis of spherical data.* Cambridge University Press, 1987.

[199] M. I. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications, 1972.

[200] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Elsevier/Academic Press, Amsterdam, 2007.

[201] I. Andrianakis and P. R. White, “MMSE speech spectral amplitude estimators with chi and gamma speech priors,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, May 2006, pp. 1068–1071.

[202] P. Schreier and L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals.* Cambridge University Press, 2010.