# Speech Signal Enhancement by Information Combining

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

vorgelegt von

Diplom-Ingenieur

**Florian Kurt Wolfgang Paul Heese**

aus Darmstadt

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

Tag der mündlichen Prüfung: 22. September 2016

**AACHENER BEITRÄGE ZU DIGITALEN NACHRICHTENSYSTEMEN**

Herausgeber:

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

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

1. Auflage Aachen:
Wissenschaftsverlag Mainz in Aachen
(Aachener Beiträge zu digitalen Nachrichtensystemen, Band 44)
ISSN 1437-6768
ISBN 978-3-958896-125-1



Wissenschaftsverlag Mainz
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34
www.Verlag-Mainz.de

Gedruckt auf chlorfrei gebleichtem Papier

D 82 (Diss. RWTH Aachen University, 2016)

# Acknowledgments

This thesis was written during my time as research assistant at the Institute of Communication Systems at the RWTH Aachen University. It is my great pleasure to take the opportunity to thank the people who contributed to the success of this work.

First, I would like to express my sincere gratitude to my supervisor Prof. Dr.-Ing. Peter Vary for his continuous support, his numerous ideas and suggestions, as well as for inspiring discussions. I would also like to thank Prof. Dr.-Ing. Rainer Martin for being my co-supervisor and for his interest in my thesis. In addition, I want to express my appreciation to Prof. Dr.-Ing. Peter Jax who took over the position as head of the Institute in April 2015.

Furthermore, I would like to thank all my current and former colleagues as well as the permanent staff at the Institute for fruitful collaboration, the intense scientific and technical discussions, proof-reading this work, and the enjoyable working environment. I want to thank Christiane Antweiler, Thomas Esch, Benedikt Eschbach, Bernd Geiser, Daniel Haupt, Hauke Krüger, Stefan Kühl, Sebastian Nagel, Christoph Nelke, Markus Niermann, Matthias Rüngeler, Thomas Schlien, Tim Schmitz, Matthias Schrammen, Magnus Schäfer, Simone Sedwick, Sylvia Sieprath, and Andreas Welbers. I would also like to express my appreciation to all the students I worked with during my time at the Instutite and who supported my work.

This work was accompanied by projects with Nokia, Ericsson, Rovi MainConcept, and IENT RWTH Aachen University. I want to thank our former project partners for the good and friendly collaboration. In addition, I would like to thank the Speech and Signal Processing Group at the Bar-Ilan University Israel for sharing the Speech & Acoustic Lab, for numerous intense discussions, and for making my visit so pleasant.

Finally, I thank my friends and family, especially my parents for giving me the fundament for everything.

Aachen, September 2016 *Florian Heese*

# Abstract

Mobile phones as well as tablets are omnipresent and belong to everyday life. Today audiovisual communication takes place at different locations and in a large variety of acoustic environments. In consequence, the intelligibility as well as the quality of speech may significantly be degraded by ambient background noise. In order to improve speech intelligibility and to ensure a convenient communication with high audio quality, speech enhancement techniques are required. In this thesis all critical components contributing to the enhancement of the up-link signal are addressed:

- signal capturing at the acoustic front-end with a *new near field beamformer*,
- *new codebook based speech and noise estimation procedure* generating and exploiting reliability information, and
- actual *noise reduction exploiting spectral dependencies* of human speech.

For the acoustic front-end of the digital processing chain a novel concept for the filter optimization of a near field beamformer is introduced. The optimization scheme allows to closely approximate a predefined reception characteristic which can be freely chosen according to the application. The output of the beamformer provides a pre-enhanced signal with improved SNR for subsequent single-microphone based speech enhancement.

Single-microphone noise reduction usually relies on statistical properties of speech and noise. In general, the noise is assumed to be stationary or only slightly time-varying, which is in practice often not fulfilled. Due to imprecise noise estimation, single-microphone systems are prone to unpleasant artifacts that are called *musical tones*. In this context different *Information Combining* methods, merging various estimates, are presented which address specifically the problem of non-stationary noise signals, leading to a significant improved estimation accuracy.

On the one hand, the proposed *Information Combining* is used with respect to spectral dependencies of human speech. On the other hand, it merges the best of several speech and noise estimates depending on their reliability. The necessary estimates are provided by a new statistical noise estimator as well as a codebook driven speech and noise estimation algorithm. The achieved estimation quality opens up the possibility to close the gap between the conflicting goals of high noise attenuation, low speech distortion, and the prevention of undesired *musical tone* artifacts. Finally, the practical aspects of the proposed enhancement systems are considered and discussed with two implemented real-time demonstrators.

# Contents

# Chapter 1

# Introduction

Speech is one of the most important manners of human interaction. The invention of the telephone enables a communication with persons all around the world which is a matter of course nowadays. As a result of continuous technological progress and economical interests, wireless communication as well as the internet have been evolving. Mobile phones and tablets are omnipresent and belong to everyday life. The Internet enables new multi-modal communication services such as video conferencing, gaining more and more importance, e. g., for international cooperation of companies, for home office or for social communication.

Mobile communication takes place at different locations and in a large variety of acoustic environments. In consequence, the intelligibility as well as the quality of speech signals may be significantly degraded in case of background noise such as, e. g., traffic, engine, wind, babble, and office noise.

In order to ensure the speech intelligibility and even to improve the listening comfort with high audio quality, speech enhancement techniques are required. These algorithms aim at reducing echos, reverberation and background noise without affecting the underlying speech signal. A typical application scenario is depicted in Fig. 1.1. A clean speech signal $s$ is disturbed by surrounding noise sources $n$ which are captured by the microphones of a mobile phone. Before the signal is transmitted over the radio channel, speech enhancement is applied. In this thesis the focus is on the problem of noise reduction.

Noise reduction systems can be subdivided into two classes: single-microphone systems and multi-microphone systems. Single-microphone systems usually rely on statistical properties of speech and noise for signal enhancement. In general, the noise is assumed to be stationary or only slightly time-varying. In practice, however, this assumption is often not fulfilled. Consequently, single-microphone systems suffer frequently from unpleasant artifacts due to imprecise noise estimation. These artifacts are called *musical tones*. Adding a second microphone allows to exploit the coherence for improved noise estimation. In contrast, multi-microphone speech enhancement systems are designed to exploit additionally spatial information as the desired and interfering audio signals are usually spatially separated. Utilizing the spatial information, a beamformer with usually more than two microphones, for example, can amplify a target speaker efficiently while simultaneously damping other speakers and background noise. Hence, an appropriately designed microphone array allows to achieve a substantial improvement of the *signal-to-noise ratio*

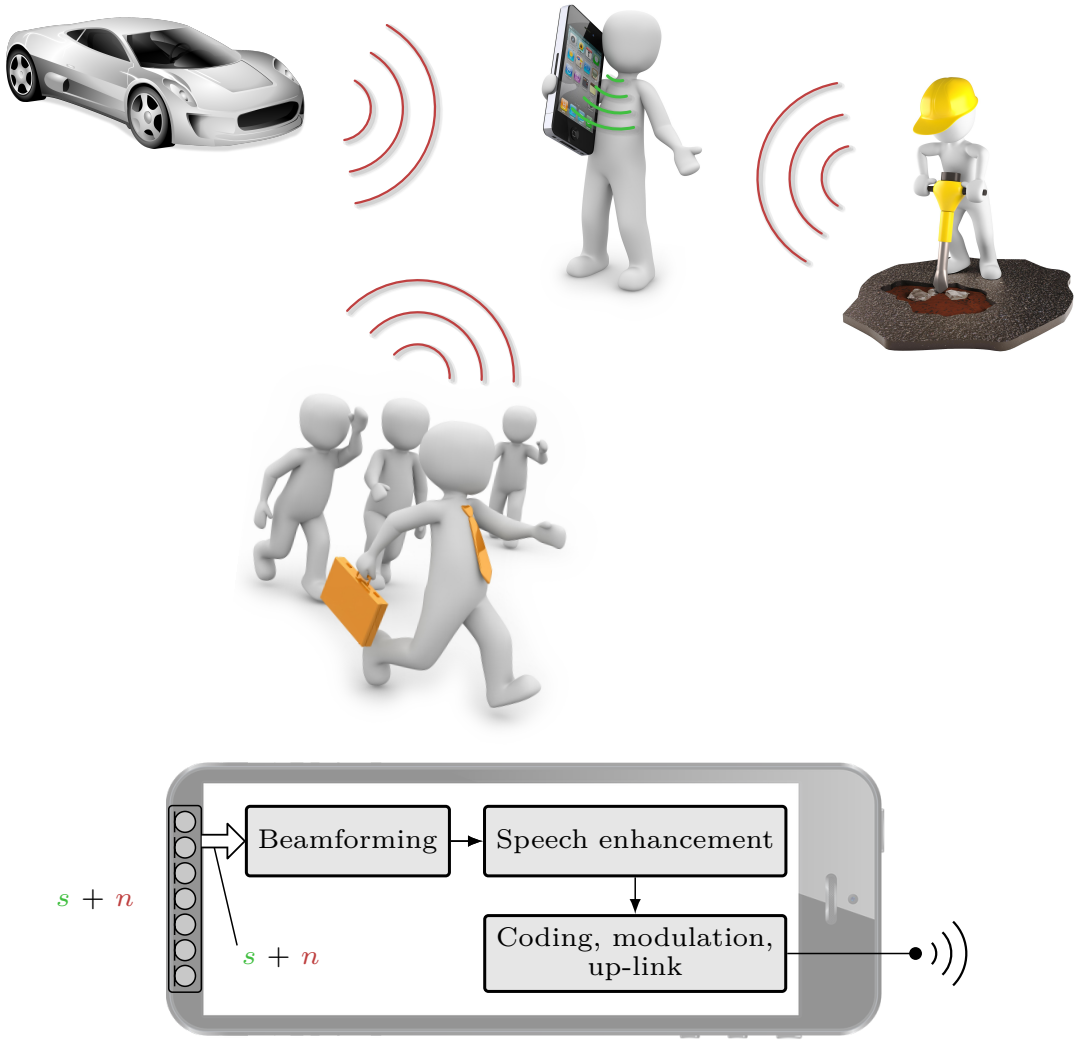



**Figure 1.1:** Application scenario for speech enhancement: Mobile communication in noisy environments.

(SNR) already at the acoustic front-end. Quite often both concepts, multi- and single-microphone systems, are concatenated for improved speech enhancement.

Addressing multi-microphone speech enhancement at the acoustic front-end, a novel concept for the optimization of filter coefficients for a near field beamformer is presented. The optimization scheme allows to closely approximate a predefined reception characteristic which can be freely chosen according to the application. The output of the beamformer might be subject to further single-channel based speech enhancement systems.

Strong emphasis of this thesis is on single-microphone (single-channel) noise reduction. New approaches are developed which exploit different aspects of so called *Information Combining* utilizing different estimates. In the context of this work the term *Information Combining* is used in a different manner as usually known from channel coding, wireless diversity receivers, and information theory [Huber & Huettinger 2003; Land et al. 2005; Land & Huber 2006]. On the one hand, *Information Combining* is used with respect to spectral dependencies of human

speech between the low band (50 Hz – 4 kHz) and the high band (4 kHz – 7 kHz) yielding a refined noise estimate. On the other hand, information about the reliability of different estimates of speech and noise is generated and exploited. Depending on this information the estimates are combined resulting in refined estimates of speech and noise, enabling advanced speech enhancement. The necessary different estimates of speech and noise are provided by a novel codebook based algorithm and a new developed low complexity noise estimator called *Baseline Tracing*. It turns out, that the use of codebook driven speech and noise estimation together with *Information Combining* is able to model and cope with highly non-stationary noise. It is of special interest that the occurrence of undesired artifacts such as *musical tones* is reduced tremendously.

Furthermore, in a video conferencing application, multi-modal *Information Combining* is carried out. The proposed near field beamformer is embedded in a high quality video conferencing client. Exploiting information provided from audio and video analysis, the activity of speakers is determined in terms of soft decision values as a function of space *and* time. On this basis, the most active speakers can be identified and separated.

## 1.1 Related Works

In literature, a vast amount of proposals for speech enhancement can be found. A comprehensive overview including the historical evolution up to state-of-the-art approaches for the estimation of the short-term noise *power spectral density* (PSD), the input SNR, and different weighting gain functions is presented in, e. g., [Benesty et al. 2009, 2007; Hänsler & Schmidt 2006, 2008; Loizou 2013; Vary et al. 1998; Vary & Martin 2006]. The first practical implementations date back to the year 1965. In [Schroeder 1965] the first patent on spectral subtraction was published for an analog circuit implementation.

In the digital era, *digital signal processor*s (DSPs) prepared the ground to implement adaptive and more sophisticated noise reduction algorithms. The key digital signal processing approaches can be found in [Boll 1979; Lim & Oppenheim 1979; McAulay & Malpass 1980] and are based on spectral subtraction or the Wiener filter [Wiener 1949] method.

For real-time speech enhancement, the noisy input signal is segmented into overlapping frames. Usually these segments are transformed into a domain, in which speech and noise are better separable, e. g., the *short-term Fourier domain* (STFD) or the cepstral domain. This procedure is called analysis. Model based processing is carried out in the transform domain. A statistical estimation framework is usually applied exploiting certain assumptions about the statistics of speech and noise. While a Gaussian model is often used for noise, speech is modeled by either a Gaussian or super-Gaussian distribution. Specific solutions are detailed, e. g., in [Breithaupt et al. 2007, 2008; Ephraim & Malah 1984, 1985; Erkelens et al. 2007; Lotter & Vary 2005; Martin 2005; Vary 1985]. After manipulation, the enhanced signal is transformed back into the time domain which is called synthesis.

In particular, the precise estimation of the time varying noise spectrogram remains the most crucial part in speech enhancement and is a prerequisite for noise reduction by adaptive time and frequency dependent filtering. If the noise is stationary or only slowly varying with time, a short-term noise PSD estimate can either be obtained during speech pauses or by continuously tracking the magnitude minima in the STFD. Further processing and updating over time is necessary. Several methods have been proposed for the estimation of noise, e. g., [Baasch et al. 2014; Cohen 2003; Doblinger 1995; Dörbecker & Ernst 1996; Gerkmann & Hendriks 2011; Hendriks et al. 2010; Jeub et al. 2011; Martin 2001, 2006; Nelke et al. 2013].

Specialized solutions are, e. g., [Chen et al. 2009; Esch et al. 2010c] for rapidly time-varying harmonic car engine noise, [Godsill et al. 2015; Talmon et al. 2013] facing an abrupt or impulsive noise sound which is typical for keyboard typing or door knocking. Wind noise is covered, e. g., in [Nelke et al. 2015; Nelke & Vary 2015] and references therein.

## 1.2 Structure of this Thesis

The thesis is subdivided into six chapters which are supplemented by a number of appendices.

In Chap. 2 the concept of microphone array processing is introduced. The difference between the far and near field is emphasized motivated by a video conference application scenario. On this basis an optimization scheme for near field beamforming is derived. The optimization scheme allows to closely approximate a predefined reception characteristic which can be freely chosen according to the application. Finally, the novel concept for filter optimization is assessed in a free field scenario as well as in a reverberant room.

The basic principles of statistical noise reduction are introduced in Chap. 3. Subsequently, *Baseline Tracing*, a novel short-term noise PSD estimator, is presented. The basic idea consists of a constrained logarithmic magnitude tracing of the noisy observation separately for each frequency bin. The new short-term noise PSD estimator is an inherently unbiased estimator and does not need correction terms. A detailed performance analysis is provided covering the noise estimation performance as well as the application embedded in a conventional noise reduction system. Furthermore, the estimator is evaluated also on pure speech signals.

In addition, an approach to wideband (50 Hz – 7 kHz) noise reduction is presented. Spectral dependencies between the low band (50 Hz – 4 kHz) and the high band (4 kHz – 7 kHz) of speech signals are investigated. An analysis of meaningful and noise robust features is carried out. Applying techniques known from artificial bandwidth extension, features from the enhanced low band signal are extracted and used to improve the noise estimate in high band. Spectral weighting gains determined from this noise estimate are adaptively combined with conventional gains obtained in addition for the high band. This *combining* in the high band is possible employing a pre-trained SNR dependent look-up table.

Codebook based speech and noise estimation is detailed in Chap. 4. *A priori*

knowledge about speech and noise allows to model and to cope with highly non-stationary noise environments. A brief overview of the development and the fundamental principles is presented. Starting point is a brute force codebook matching approach, which provides the upper performance bound and serves as reference. The basic concept is a superposition of scaled speech and noise codebook entries. At first, the *a priori* assumptions of speech and noise are analyzed with respect to practical application scenarios. While the speech codebook is pre-trained in advance, the noise codebook is adapted to new noise types online. Thus, the system becomes independent of *a priori* knowledge regarding noise. Training vectors for online noise codebook updates are identified using a *voice activity detector* (VAD) and a codebook mismatch measure. For this purpose, a novel noise robust VAD is developed which depends only on *a priori* knowledge about speech.

In Chap. 5 a generic theoretical analysis of the joint speech and noise estimation problem is carried out given the noisy observation. The analysis considers an arbitrary number of different speech and noise estimates. An analytic solution is formulated which minimizes the estimation error power with respect to the noisy observation. This procedure is called *Information Combining* and provides optimal mixing coefficients of the different speech and noise estimates. On this basis two main restrictions of codebook based speech and noise estimation are addressed.

Missing *a priori* codebook knowledge regarding noise is compensated utilizing an additional noise estimate as automatic fallback, e. g., provided by the new proposed statistical noise estimator *Baseline Tracing*. In a second step this procedure is generalized to additionally provide a refined speech estimate.

With respect to practical application scenarios, a substantial complexity reduction is necessary. Utilizing the *Information Combining* procedure in this context, the brute force codebook driven speech and noise estimates can be replaced by two cascades of gain shape *vector quantizer* (VQ) estimates, i. e., the determination of the spectral shape using a codebook in a first step and the calculation of the corresponding gain in a second step. The chapter closes with a comprehensive evaluation including all presented aspects of codebook driven noise reduction.

In Chap. 6 two application examples are presented. The near field beamformer detailed in Chap. 2 is utilized in a high quality video conferencing scenario in order to determine the most active speakers as a function of time *and* space. In the second part of the chapter, the codebook driven speech enhancement system is analyzed and a further complexity reduction is carried out, for both the codebook matching as well as the VAD. Utilizing a software based *private branch exchange* (PBX) a proof of concept implementation on a lightweight embedded computing platform is created. Finally, the results of this thesis are discussed in Chap. 7.

Parts of this thesis have been presented in the following references published by the author: [Heese et al. 2010; Heese et al. 2011; Heese et al. 2012a; Heese et al. 2012b; Heese et al. 2013; Heese et al. 2014a; Heese & Vary 2015; Heese et al. 2015; Esch et al. 2010a; Esch et al. 2010b; Esch et al. 2010c; Esch et al. 2012; Schlien et al. 2013; Schäfer et al. 2012; Bulla et al. 2013; Hadad et al. 2014; Niermann et al. 2015]. Throughout this thesis, these references are marked by underlining the year.

# Chapter 2

# Near Field Beamforming

The aim of sensor array signal processing is to estimate a desired signal which may be degraded by additive noise using temporal and spatial information from array sensors [Haykin & K. R. Liu 2010]. The design of such systems is an ongoing topic of research with many applications in the radio frequency domain [Haykin et al. 1993] as well as in the acoustic domain [Brandstein & Ward 2001]. Especially in the acoustic domain the class of linear microphone arrays received attention since they can easily be integrated into communication systems such as video conferencing terminals. If the desired and the interfering audio signals are spatially separated, an appropriately designed microphone array allows to achieve a substantial *signal-to-noise ratio* (SNR) gain already at the acoustic front-end.

Spatial separation of audio source is often present, e. g., in a conferencing scenario or a typical office room. Moreover, the reverberation as well as the level of diffuse background noise are usually quite low in these environments. Hence, speech enhancement techniques utilizing multichannel microphone arrays, such as beamformer algorithms, are appropriate to amplify a target speaker efficiently while simultaneously damping other competing speakers and background noise.

Beamformer algorithms can be subdivided into fixed and adaptive approaches [Brandstein & Ward 2001; Haykin & K. R. Liu 2010]. Fixed beamforming algorithms are independent of the input signals and can realize robust directional gains with moderate numerical complexity. Typical representatives are the (weighted) delay-and-sum as well as the filter-and-sum beamformer. Adaptive beamforming algorithms are well suited for cancelling moving interferers. Among various adaptive beamforming categories, the *minimum variance distortionless response* (MVDR), the *multichannel Wiener filter* (MWF), the *linearly constrained minimum variance* (LCMV) beamformer, and the *generalized sidelobe canceller* (GSC) are the most commonly used [Griffiths & Jim 1982; Markovich Golan et al. 2009; Van Veen & Buckley 1988].

Furthermore beamformers can be realized operating in the time domain or the sub-band domain, e. g., [De Haan et al. 2001; Lorenzelli et al. 1996; Nordholm et al. 2008; Zhao et al. 2011]. Using a sub-band beamformer offers several advantages compared to a full-band beamformer such as an overall lower filter degree or an improved reception characteristic with respect to the operational frequency.

When designing beamformers, a specific spatial and if necessary frequency-dependent reception characteristic is usually the desired goal. For the far field case,

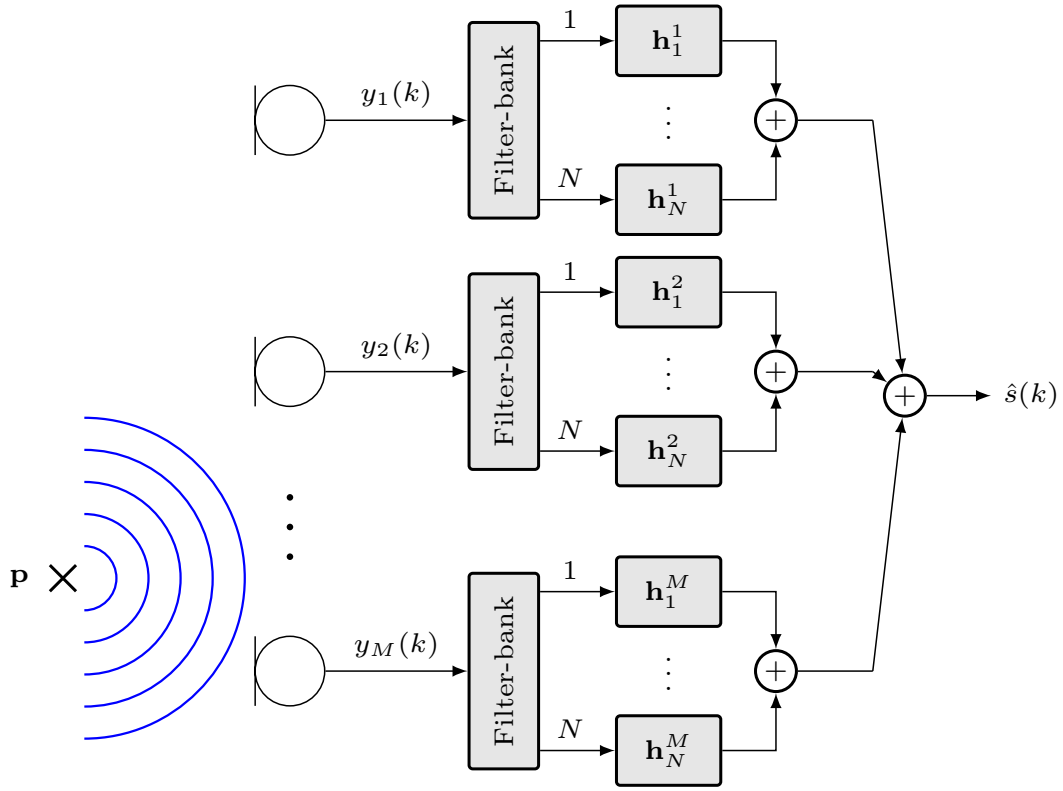



**Figure 2.1:** Filter-and-sum beamformer with $M$ microphones and $N$ non-uniform sub-bands

i. e., if the distance to the array is significantly larger than its geometric setup, many beamformer design methods are known, e.g., [Doclo & Moonen 2003; Ward et al. 1995]. There are also procedures known, aiming specifically at the near field, where the far field assumption provides only an approximation in the best case, see e. g., [Doclo & Moonen 2003; Fisher & Rafaely 2011; Kennedy et al. 1996; Ryan & Goubran 2000] and references therein. However, these approaches optimize the reception characteristic limited by several design constraints, e. g., only on a (semi-)circular arc at one specific distance from the array.

To circumvent this limitations, an alternative new design strategy is considered [Heese et al. 2013; Schäfer et al. 2012]. The reception characteristic is optimized for a certain predefined two-dimensional target area in the near field, simultaneously for different distances and angles. The work in [Schäfer et al. 2012] considered a weighted delay-and-sum array with full-band processing as basis for the optimization while in [Heese et al. 2013] a more generalized approach using sub-bands and a filter-and-sum beamformer is applied which is presented in the following.

## 2.1 Non-Uniform Near Field Sub-band Filter-and-Sum Beamformer

A simplified block diagram of the proposed beamformer system is depicted in Fig. 2.1. It consists of $M$ microphones followed by non-uniform filter-banks [Löllmann 2011] each comprising $N$ sub-bands. Subsequently, all sub-bands are processed by different filter-and-sum units represented by the impulse responses $\mathbf{h}_n^m$, $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots, N\}$, where $n$ denotes the sub-band index and $m$ the microphone index. Finally, the summation of the output signals of the filter-and-sum units result in the all-over beamformer output signal $\hat{s}(k)$.

The microphone signal samples $y_m(k)$ are obtained by analog-digital conversion with a sampling frequency of $f_\mathrm{s}$, where $k$ represents the discrete time index. A point source $s(k)$ is assumed to be at position $\mathbf{p}$ on an appropriately chosen two-dimensional spatial grid, e. g.,

- in a polar coordinate system: $\mathbf{p} = (r\ \varphi)^T$ or
- in a Cartesian coordinate system $\mathbf{p} = (x\ y)^T$.

Given the impulse responses $h_{\mathbf{p}m}(k)$, $m \in \{1, \ldots, M\}$, from the point source $\mathbf{p}$ to each of the $M$ microphones, each microphone signal $y_m(k)$ can be expressed as:

$$y_m(k) = h_{\mathbf{p}m}(k) * s(k)\,, \tag{2.1}$$

where $*$ denotes the linear discrete convolution operator. The output $\hat{s}(k)$ thereby depends on the source location $\mathbf{p}$ and can be calculated according to:

$$\hat{s}(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * \left( h_n^{\mathrm{FB}}(k) * y_m(k) \right)\,, \tag{2.2}$$

where $h_n^{\mathrm{FB}}(k)$ represents the bandpass impulse responses of the filterbank and $h_n^m(k)$ the *finite impulse response* (FIR) sub-band filters of length $L$ to be determined by numerical optimization.

## 2.2 Numerical Optimization

The optimization of the filter-and-sum units is carried out in frequency sub-bands to decouple the optimization procedure. Furthermore the frequency resolution is chosen following the human auditory system. The principle of the numerical optimization procedure for each sub-band is depicted in Fig. 2.2. An iterative minimization of an error measure is carried out. The error measure is a function of the level difference between a predefined target reception characteristic and the simulated one. The simulated reception characteristic is calculated on the current state of the filter coefficients and the impulse responses, modeling the acoustic path between the source positions and the microphones, in each iteration.

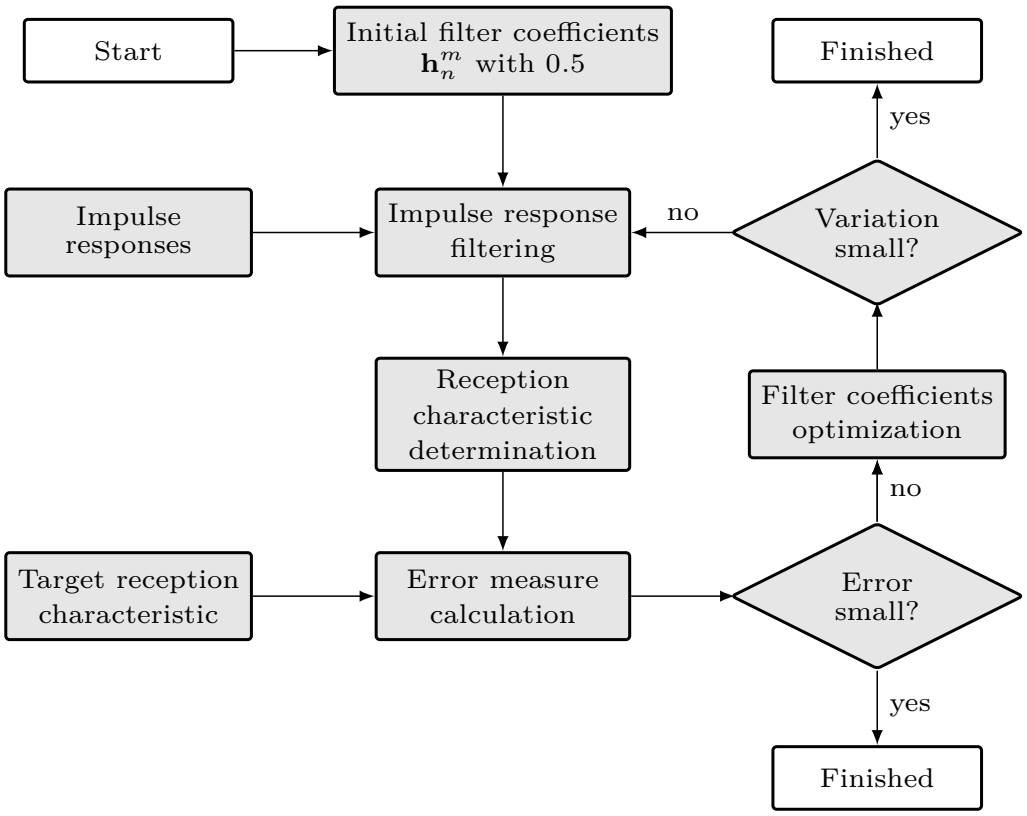



**Figure 2.2:** Optimization process for each frequency sub-band

Since the optimization is performed in exactly the same manner for simulated impulse responses as well as for measured ones, it is very flexible referring to different practical application scenarios.

### 2.2.1 Definition of the Target

In order to determine the filter coefficients of the beamformer, a target area in front of the microphone array with different amplification or attenuation gains has to be defined. Hence, the target reception characteristic $S_{\mathbf{p}}(f)$ is defined as a spatial distribution of areas with defined amplification or attenuation in front of the microphone array.

The spatial target reception characteristic $S_{\mathbf{p}}(f)$ can be defined as frequency-dependent but a frequency-independent target, i. e., $S_{\mathbf{p}}(f) = S_{\mathbf{p}}$, is suitable in many applications. The target speaker should be in the amplified region $\mathbb{P}_{\text{high}}$ (target amplification gain $S_{\text{high}}$) while the attenuated area $\mathbb{P}_{\text{low}}$ (target attenuation gain $S_{\text{low}}$) is chosen to contain all interfering signals. This corresponds to a given SNR improvement compared to a single omnidirectional microphone for the received signal. Hence, the target reception characteristic is defined as,

$$S_{\mathbf{p}} = \begin{cases} S_{\text{high}} & \text{for } \mathbf{p} \in \mathbb{P}_{\text{high}} \\ S_{\text{low}} & \text{for } \mathbf{p} \in \mathbb{P}_{\text{low}} \,. \end{cases} \tag{2.3}$$

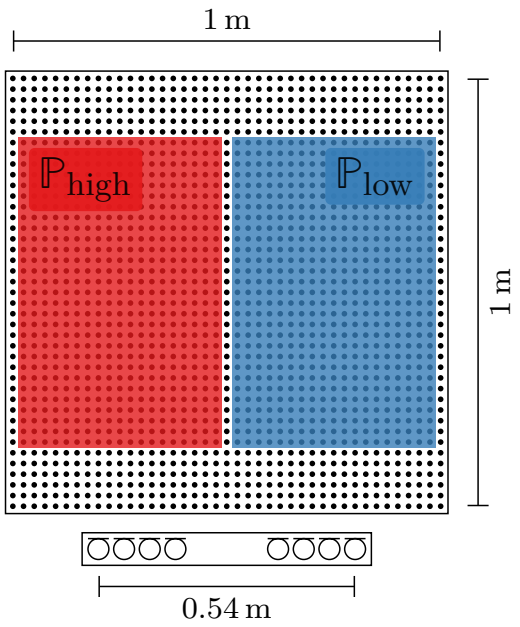



**Figure 2.3:** Example of target reception characteristic. The black dots indicate the two-dimensional spatial grid in front of the microphone array. The amplified region $\mathbb{P}_{\text{high}}$ and the attenuated region $\mathbb{P}_{\text{low}}$ are represented by the red and blue rectangle, respectively

The precise choice of the areas and gains depends on *a priori* knowledge from the application, e. g., in a conference scenario the target speaker in $\mathbb{P}_{\text{high}}$ should be amplified by gain $S_{\text{high}}$, while all interfering sources in $\mathbb{P}_{\text{low}}$ shall be attenuated by gain $S_{\text{low}}$. An example of a typical target reception characteristic within an conferencing scenario is given in Fig. 2.3.

### 2.2.2 Determination of the Reception Characteristic in the Near Field

With respect to the iterative optimization procedure of the filter coefficients $\mathbf{h}_n^m$, the predefined target reception characteristic $S_{\mathbf{p}}$ is compared to the current state of the simulated reception characteristic $\widehat{S}_{\mathbf{p}}(f)$. Applying the following three steps, the reception characteristic $\widehat{S}_{\mathbf{p}}(f)$ based on the current state of the filter coefficients is calculated:

- simulating or measuring impulse responses between all grid positions within the target region ($\mathbf{p} \in \{\mathbb{P}_{\text{high}} \cup \mathbb{P}_{\text{low}}\}$) and all $M$ microphones,
- processing these impulse responses with the filter-and-sum beamformer (see Fig. 2.1 and Sec. 2.1) to get an overall filter including all microphones for every point in the near field, and
- calculating the amplification or attenuation for every point from these overall filters.

Since the output signal $\hat{s}(k)$ for each source location $\mathbf{p}$ can be expressed as a linear superposition of the filtered version of the source signal:

$$\hat{s}(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * h_n^{\mathrm{FB}}(k) * h_{\mathbf{p}m}(k) * s(k), \tag{2.4}$$

the overall filter $g_{\mathbf{p}}(k)$ can be calculated as:

$$g_{\mathbf{p}}(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * h_n^{\mathrm{FB}}(k) * h_{\mathbf{p}m}(k). \tag{2.5}$$

Applying the frequency transform of the overall filter $g_{\mathbf{p}}(k)$ yields:

$$\mathcal{G}_{\mathbf{p}}(f) \overset{\mathcal{F}}{\bullet\!\!-\!\!\circ} g_{\mathbf{p}}(k)\,. \tag{2.6}$$

Finally, the reception characteristic $\widehat{S}_{\mathbf{p}}(f)$ in dB of the beamformer at frequency $f$ for every point in the target region ($\mathbf{p} \in \{\mathbb{P}_{\mathrm{high}} \cup \mathbb{P}_{\mathrm{low}}\}$ of the microphone array is obtained by:

$$\widehat{S}_{\mathbf{p}}(f) = 20 \cdot \log_{10} \left|\mathcal{G}_{\mathbf{p}}(f)\right|\,. \tag{2.7}$$

### 2.2.3 Error Function

The reception characteristic $\widehat{S}_{\mathbf{p}}(f)$ represents the intermediate reception characteristic which is realized by the respective filter coefficients in each iteration step. By variation of the sub-band filters, the distance between the predefined target $S_{\mathbf{p}}$ and $\widehat{S}_{\mathbf{p}}(f)$ is minimized in terms of the summed quadratic gain difference $\Delta_S(n)$ for each sub-band $n$. For all points, where $S_{\mathbf{p}}$ is defined according to Eq. (2.3), and over all frequencies $f_i$ with $i \in \{i_{\min}, \ldots, i_{\max}\}$, for which the reception characteristic shall be optimized, the sum of the quadratic gain differences is calculated according to:

$$\Delta_S(n) = \sum_{i=i_{\min}}^{i_{\max}} \sum_{\mathbf{p} \in \left\{\mathbb{P}_{\mathrm{high}} \cup \mathbb{P}_{\mathrm{low}}\right\}} \left(S_{\mathbf{p}} - \widehat{S}_{\mathbf{p}}(f_i)\right)^2, \tag{2.8}$$

where $f_{i_{\min}}$ and $f_{i_{\max}}$ denote the lower and upper edge frequencies of sub-band $n$.

### 2.2.4 Optimization Procedure

The optimum filter coefficients for each microphone and each sub-band $n$ are determined in a *minimum mean-square error* (MMSE) sense by:

$$\left[\mathbf{h}_n^1, \ldots, \mathbf{h}_n^M\right]_{\mathrm{opt}} = \arg\min_{\mathbf{h}} \Delta_S(n)\,. \tag{2.9}$$

For the optimum search, the iterative interior-point algorithm according to [Byrd et al. 2000] is employed with the constraint that the filter coefficients range within $-1$ and 1. Since this constraint only limits the maximum amplification that is achievable by the array itself and does not change the relation between the filter coefficients, subsequent scaling of the output $\hat{s}(k)$ can be applied to map the reception characteristic to a desired gain.

## 2.3 Performance Example

The proposed new design strategy using decoupled sub-band filters for the optimization of the reception characteristic in the near field is demonstrated in two steps. At first the proposed new design strategy is compared with two other beamforming approaches. In a possible application scenario, e. g., a video conferencing system, the simulation of the impulse responses can be rather simple since conference rooms are usually not highly reverberant. In this case, a simple mirror-image approach or even the approximation by a free field model is suitable. Hence, the assessment is based here (without loss of generality) on a free field setup since this allows for a clearer evaluation of the impact of the filter coefficients. In a second example the proposed beamformer is evaluated using measured impulse responses of a reverberant room.

### 2.3.1 Free Field

In this assessment the proposed new design strategy is compared with two other beamforming approaches. A beamformer which also allows to optimizes the reception characteristic in the near field and a conventional unoptimized one. The performance is evaluated by comparing the reception characteristic of the different methods.

As representative for an unoptimized beamformer the Chebyshev weighting approach [Harris 2004] is utilized. This is a fair comparison since the Chebyshev weighting allows to specify a minimum attenuation for all sidelobes while at the same time also minimizing the width of the main lobe. Hence, this combination allows to maximize the SNR between a target area and a diffuse noise field.

As a second reference, the near field full-band optimized weighted delay-and-sum beamformer from [Schäfer et al. 2012] is considered as an optimized beamformer candidate. In order to demonstrate the benefit of the proposed sub-band processing, the full-band weighted delay-and-sum beamformer [Schäfer et al. 2012] is modified utilizing a filter-and-sum unit instead of the weighted delay-and-sum unit. In the following this beamformer is referred as modified full-band beamformer.

The reception characteristics are evaluated for a one square meter sized area in front of the microphone array and the density of the spatial grid is set to 0.01 m for both dimensions ($x$ and $y$). Since the simulation is based on an acoustic free field, the impulse responses $h_{\mathbf{p}m}(k)$ from each point of the spatial grid to the microphone array represent the corresponding delays. All beamformer setups are parameterized

such that they are supposed to achieve a level difference between the amplified and damped area of 40 dB.

For the two optimized beamformer setups (proposed sub-band and modified full-band) the target area can be defined explicitly and is depicted in Fig. 2.3. The microphone arrays are designed to amplify sources on the left ($\mathbb{P}_{\text{high}}$: $-0.5\,\text{m} \leq x < 0\,\text{m} \wedge 0.2\,\text{m} < y \leq 0.8\,\text{m}$) while attenuating sources on the right ($\mathbb{P}_{\text{low}}$: $0\,\text{m} < x \leq 0.5\,\text{m} \wedge 0.2\,\text{m} < y \leq 0.8\,\text{m}$). Due to the specified spatial grid resolution this leads to 3000 points in $\mathbb{P}_{\text{high}}$ and $\mathbb{P}_{\text{low}}$, respectively.

The sample rate $f_\text{s}$ is set to 48 kHz and the microphone line array consists of $M = 8$ omnidirectional sensors which have a spacing of 3 cm between the sensors and a gap of 30 cm in the middle of the array, e. g., for camera mounting in a video conferencing application. The microphone array is centered at the origin of the coordinate system as depicted in Fig. 2.3. Spatial aliasing can be expected for frequencies higher than approximately 5600 Hz due to the microphone spacing. For the proposed system the number of sub-bands is set to $N = 6$ using a non-uniform filter bank according to the human auditory system [Löllmann 2011]. The frequency range of each sub-band can be seen in Table 2.1. For simplicity all sub-band filters have been realized as FIR filters. The degree of the filter-and-sum units $\mathbf{h}_n^m$ is set to $L = 8$ resulting in 48 filter coefficients to be optimized. Thus, the modified full-band beamformer based on [Schäfer et al. 2012] is also set up with a filter length of 48[1].

A comparison of the three reception characteristics is given for two different operating frequencies:

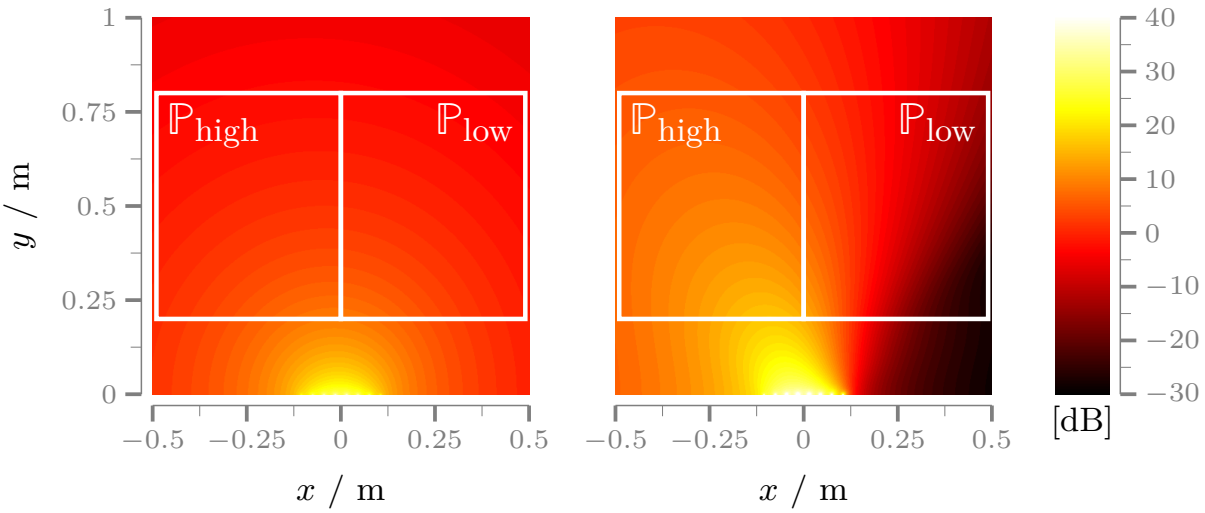
- $f_i = 500\,\text{Hz}$ as a representative for the lower frequencies for which the microphone array can be utilized,
- $f_i = 2000\,\text{Hz}$ as a frequency that is right in the center of the operational frequency range of the microphone array.

In Fig. 2.4 the two-dimensional reception characteristics in front of the microphone array are depicted for the Chebyshev weighting. Looking at the operational frequency of 2000 Hz in Fig. 2.4b there is a notable level difference between the amplified area $\mathbb{P}_{\text{high}}$ and the damped area $\mathbb{P}_{\text{low}}$ on average. However, the desired reception characteristic within $\mathbb{P}_{\text{low}}$ is only achieved at the bottom right corner of

**Table 2.1:** Filterbank sub-bands

| Band | Frequency range / Hz | Band | Frequency range / Hz |
|---|---|---|---|
| 1 | 1 - 268 | 4 | 1549 - 2614 |
| 2 | 268 - 839 | 5 | 2614 - 4731 |
| 3 | 839 - 1549 | 6 | 4731 - 12049 |

[1] The optimized filter coefficients are listed in Appendix A.1.1 for the sub-band beamformer and in Appendix A.1.2 for the full-band approach.



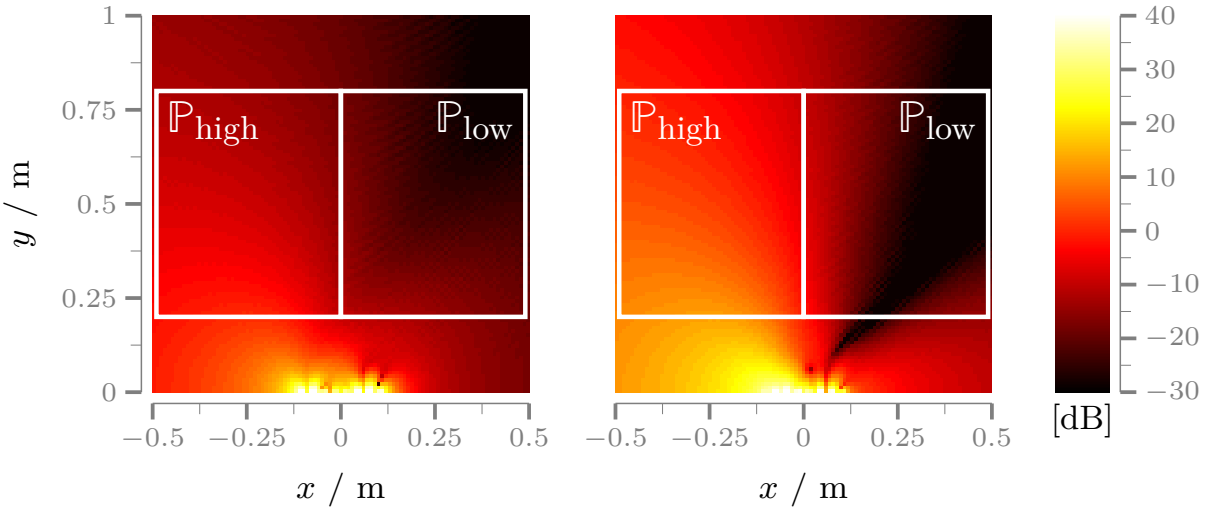

**(a)** 500 Hz **(b)** 2000 Hz

**Figure 2.4:** Reception characteristic of the microphone array with Chebyshev weighting [Harris 2004] at different operational frequencies

$\mathbb{P}_{\text{low}}$. For the 500 Hz case which is depicted in Fig. 2.4a the reception characteristic of the microphone array is resembling the behavior of a single omnidirectional microphone which is located in the origin of the coordinate system. Hence, no level difference between $\mathbb{P}_{\text{high}}$ and $\mathbb{P}_{\text{low}}$ is visible.

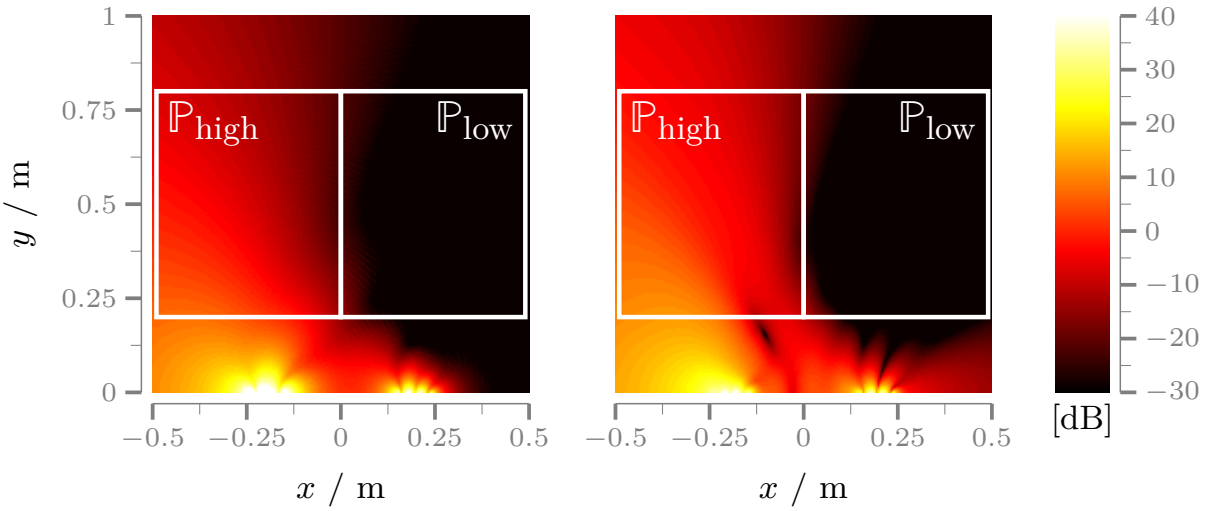
Fig. 2.5 and Fig. 2.6 present the reception characteristics for the optimized beamformer algorithms. The performance based on optimized weighting according to [Schäfer et al. 2012] is shown for the 500 Hz case in Fig. 2.5a. A noticeable level difference between the areas $\mathbb{P}_{\text{high}}$ (right side) and $\mathbb{P}_{\text{low}}$ (left side) can be seen. However, the target for the damped area is only partly achieved, yet it is better compared to the Chebyshev weighting. The reception characteristic for the proposed system (see Fig. 2.6a) fits the previously defined areas of attenuation and amplification very well even at this low frequency.

Comparing the performance for the 2000 Hz case (Fig. 2.5b and 2.6b) the difference of the reception characteristics is smaller. Both algorithms provide a significant level difference between $\mathbb{P}_{\text{high}}$ and $\mathbb{P}_{\text{low}}$ and outperform the result given by the Chebyshev weighting (cf. Fig. 2.4b). However, especially in the critical border region at $x = 0$ m the new beamformer matches the predefined target reception characteristic even better.




**(a)** 500 Hz **(b)** 2000 Hz

**Figure 2.5:** Reception characteristics of the microphone array employing the modified full-band optimized filters at different operational frequencies




**(a)** 500 Hz **(b)** 2000 Hz

**Figure 2.6:** Reception characteristics of the microphone array employing the proposed sub-band optimized filters at different operational frequencies

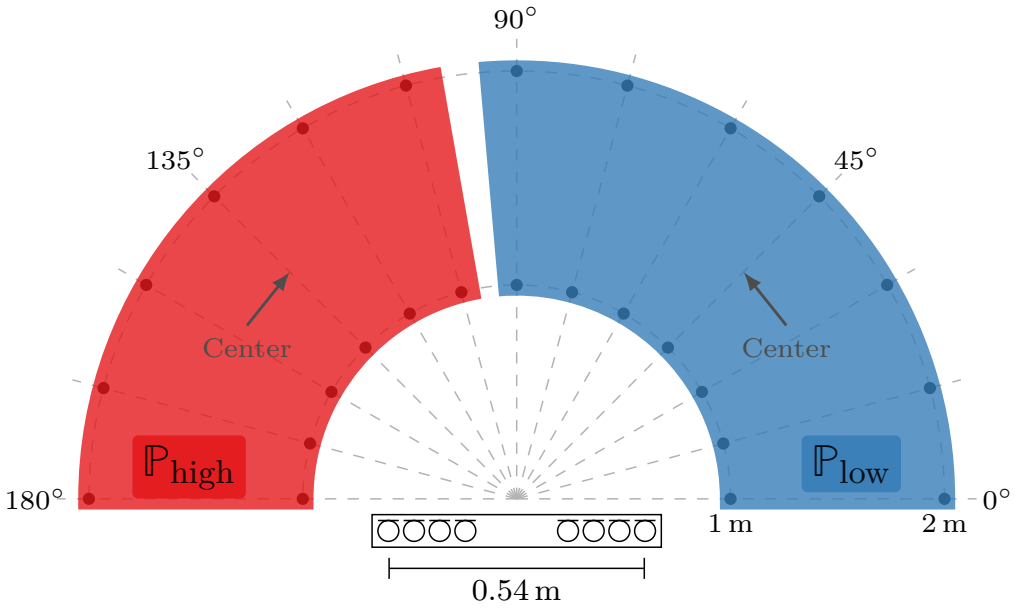



**Figure 2.7:** Geometric setup. The nodes of the spatial grid for the determination of the impulse responses are indicated by the dots. The amplified region $\mathbb{P}_{\text{high}}$ and the attenuated region $\mathbb{P}_{\text{low}}$ are represented by the red and blue shape, respectively

### 2.3.2 Reverberant Room

As a real performance example the proposed algorithm was also evaluated using 26 measured impulse responses for the optimization procedure[2]. Therefore the Speech & Acoustics Lab of the Faculty of Engineering at Bar-Ilan University, with controllable reverberation time, was utilized to create an audio-database using an 8-channel microphone array.[3]

According to a typical scenario, e. g., a video conference, the reverberation time was set to 160 ms. The density of the spatial grid (in polar coordinates) for this setup as depicted in Fig. 2.7 is given by angles from 0° to 180° in 15° steps for 1 m and 2 m radii. $\mathbb{P}_{\text{low}}$ in polar coordinates maps to all radii for angles from 0° to 90°, $\mathbb{P}_{\text{high}}$ from 91° to 180°, respectively. The remaining parameters are not changed comparing to the previous Sec. 2.3.1. Fig. 2.8 presents the magnitude transfer function for the center of regions $\mathbb{P}_{\text{high}}$ and $\mathbb{P}_{\text{low}}$. A significant level difference (in average approx. 19.4 dB) can be seen over the complete frequency range up to the spatial alias frequency of approx. 5600 Hz. This also confirms the performance of the proposed system independently of the operational frequency.

[2] The optimized filter coefficients are listed in Appendix A.2

[3] The audio-database consists of eight channel impulse responses. The measurements were taken for different microphone setups and different reverberation times at different locations on a spatial grid as depicted in Fig. 2.7. For details, cf. [Hadad et al. 2014].

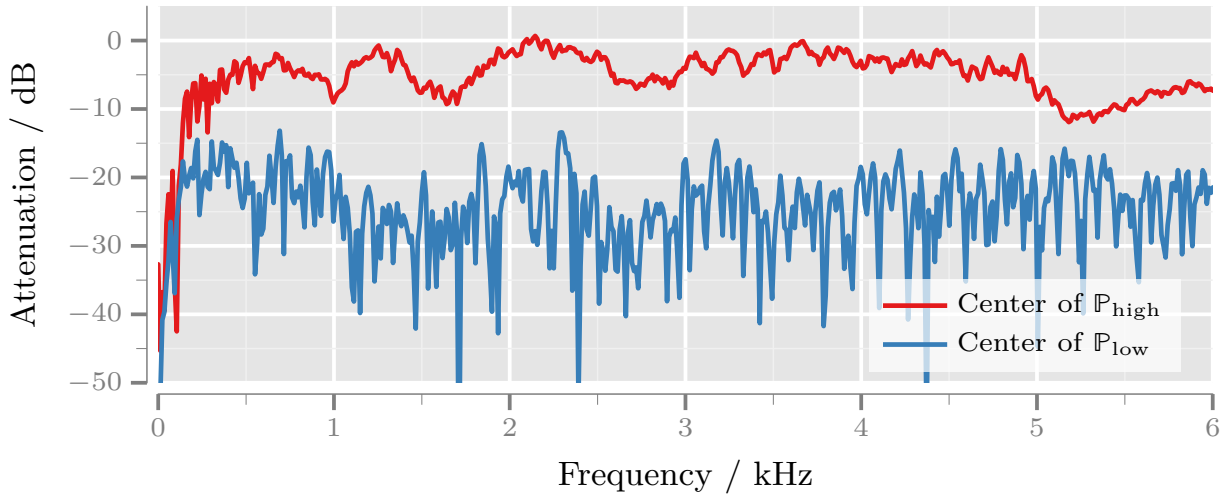



**Figure 2.8:** Transfer function of the proposed microphone array with sub-band optimized filters presented for the center of regions $\mathbb{P}_{\text{high}}$ and $\mathbb{P}_{\text{low}}$

## 2.4 Summary

A novel concept for filter optimization of a filter-and-sum beamformer based on numerical near field optimization is presented. The beamformer consists of a non uniform filterbank with FIR filters in the sub-bands which are optimized to achieve improved beamforming. The proposed design strategy combines the advantages of decoupled sub-band filter optimization with a frequency resolution according to the human auditory system. The optimization scheme allows to closely approximate a predefined reception characteristic which can be freely chosen according to the application. The proposed system provides a distinct spatial selectivity independent of the operational frequency. Switching between different reception characteristics, e. g., for speaker selection in a conference scenario, can be easily achieved by several pre-computed filter sets. A further demonstration of the novel beamformer is included in Sec. 6.1 where it is employed for a spatial *voice activity detector* (VAD) in order to determine the active speakers in a video conference scenario.

# Chapter 3

# Statistical Noise Reduction in the Frequency Domain

This chapter introduces the basic principles of statistical noise reduction in the frequency domain which are required in the sequel of this thesis. A general overview about statistical noise reduction techniques including state-of-the-art approaches for the estimation of the short-term noise *power spectral density* (PSD), the input *signal-to-noise ratio* (SNR), and different gain functions is provided. For a more detailed insight into statistical noise suppression techniques the reader is referred to the literature, e. g., [Benesty et al. 2007, 2009; Hänsler & Schmidt 2006, 2008; Vary & Martin 2006].

When it comes to the transmission of speech signals in communication systems, the original speech signal is often impaired by annoying background noise. Noise reduction algorithms aim at suppressing the background noise while keeping the speech signal as natural as possible. Since more than 30 years, noise reduction is covered in literature and is still an ongoing topic, e. g., [Boll 1979; Ephraim & Malah 1984, 1985; Vary & Martin 2006]. The noise reduction approaches can be subdivided into two classes: single-microphone systems and multi-microphone systems. Systems comprising multiple microphones are able to employ statistical and spatial information about speech and noise. Single microphone systems usually rely on (temporal) statistical properties of the speech and noise signal components for noise reduction. Depending on the application, the environment, the number of microphones, the noise type and source signals, different approaches are used in practice. Specialized solutions are, e. g., [Chen et al. 2009; Esch et al. 2010c] for rapidly time-varying harmonic (car engine) noise, [Godsill et al. 2015; Talmon et al. 2013] facing an abrupt change or impulsive noise sound which is typical for keyboard typing or door knocking.

Throughout this thesis all developed algorithms are tailored to real-time processing of single-microphone audio signals. This covers typical applications including hearing-aid or mobile-phone scenarios. With this constraint only causal modifications of the recorded audio signal are possible, i. e., only signal properties from the current point in time and the past are available but no information from the future.

A commonly used approach to perform single-microphone speech enhancement utilizes the so called spectral decomposition exploiting statistical techniques to separate speech and noise from the noisy observation. In order to transform the

noisy signal into the frequency domain, the signal is segmented into frames which are subsequently transformed utilizing the *short-term Fourier domain* (STFD) transformation. Individual adaptive gains are applied to each STFD coefficient to perform the actual noise suppression. If the SNR for a specific STFD coefficient is high an absolute gain close to one is chosen. In the opposed case where the SNR is low an absolute gain close to zero is applied. The gain function minimizes a specific distortion measure between the clean speech and the speech estimate. Usually, the gain function requires knowledge about the short-term noise PSD and the input SNR, which are in general not known *a priori* and have to be estimated. Thereafter, the processed spectrum is transformed back into the time domain.

The first part of this chapter is organized following the signal flow of a typical statistical noise reduction system. After introducing the signal model, the analysis – synthesis framework (Sec. 3.2.1) is described including the transformation into and from the short-term frequency domain. A conventional noise suppression system is detailed in Sec. 3.4 including the estimation of the short-term noise PSD (Sec. 3.4.1), the short-term SNR (Sec. 3.4.2) and the spectral weighting gain calculation (Sec. 3.4.3). Subsequently, a new statistical short-term noise PSD estimator is presented in Sec. 3.5. In Sec. 3.6 a wideband (50 Hz – 7 kHz) noise suppression approach is presented exploiting spectral dependencies between the low- and high-band. Conclusions are drawn in Sec. 3.7.

## 3.1 Problem Formulation

In Fig. 1.1 the problem of capturing speech signals in the presence of noise is illustrated for a mobile phone scenario. In the following the speech enhancement problem is discussed for single-microphone systems. In general, the microphone of a mobile phone does not only record the desired speech signal $s(k)$ but also a superposition of surrounding noise signals. The samples from the microphone signal $y(k)$ are obtained by analog-digital conversion with a sampling frequency of $f_\mathrm{s}$. The noisy input signal $y(k)$ is modeled by a clean speech component $s(k)$ which is degraded by additive noise components $n_j(k)$ according to:

$$y(k) = s(k) + \sum_j n_j(k) = s(k) + n(k), \tag{3.1}$$

where $k$ is the discrete time index and $j$ the index of the noise sources. The speech and noise signals are modeled as uncorrelated and zero-mean random processes. The aim of noise reduction is to estimate the clean speech signal having only the noisy observation $y(k)$ available. This is achieved by attenuating the noise as much as possible while keeping the speech distortion as low as possible at the same time using adaptive filtering. The resulting speech signal estimate at the output of the noise reduction system is denoted by $\hat{s}(k)$. A further requirement of the speech enhancement system is to allow a convenient conversation without notable delay of the recorded signal $y(k)$. With this constraint only causal modifications are possible, i. e., only signal properties of the past can be taken into account to estimate $\hat{s}(k)$.

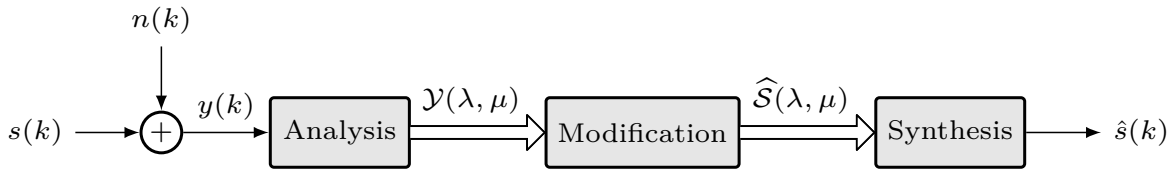



**Figure 3.1:** Generic block diagram of a speech enhancement system

## 3.2 System Overview

Throughout this thesis all considered noise reduction approaches are covered by the block diagram depicted in Fig. 3.1. Wide arrows indicate multi-channel signals and thin arrows single-channel signals, respectively.

For the derivation of most statistical noise reduction algorithms, speech and noise are considered as stationary processes. Hence, the resulting filter coefficients would be fixed over time and could be applied using simple *finite impulse response* (FIR) or *infinite impulse response* (IIR) filters. However, regarding noise, the assumption of stationarity strongly depends on the noise type and in case of speech it does not hold at all. The human speech production is a time varying process and especially plosive sounds, created by sudden pressure rises in the vocal tract, result in a highly non-stationary speech signal. However, segmenting the speech signal into short-time segments of 10 – 100 ms, speech can be assumed as short-term stationary within the segment [Rabiner & Schafer 1978].

In order to respect the short-term stationarity of speech, the noisy input signal $y(k)$ is subdivided into short-time frames and the processing of the noisy input signal is carried out framewise. Hence, the delay of the system results in one frame which is below the threshold of perception in the context of speech communication [Cox 1984; Kitawaki & Itoh 1991]. The temporal changes of speech and noise are considered for updating the filter coefficients continuously on a frame by frame basis. The frames are transformed into a domain in which the desired and the noise signal are better separable. Usual transformation domains are the frequency or cepstral domain. Using the frequency domain is a widely accepted technique for speech enhancement as it is very similar to the processing taking place in the human auditory system [Zwicker & Fastl 1990]. Therefore, the *discrete Fourier transform* (DFT) frequency domain is used as transfer domain in this thesis. The procedure including segmenting, windowing and transformation is called analysis. After manipulation in the transform domain the enhanced segments are transformed back into the time domain and combined which is called synthesis. Hence, an analysis – synthesis framework with perfect reconstruction forms the basis for the speech enhancement system.

### 3.2.1 Analysis and Synthesis

A block diagram of the analysis and synthesis framework used in this work is depicted in Fig. 3.2. As mentioned before, the input signal $y(k)$ is segmented due

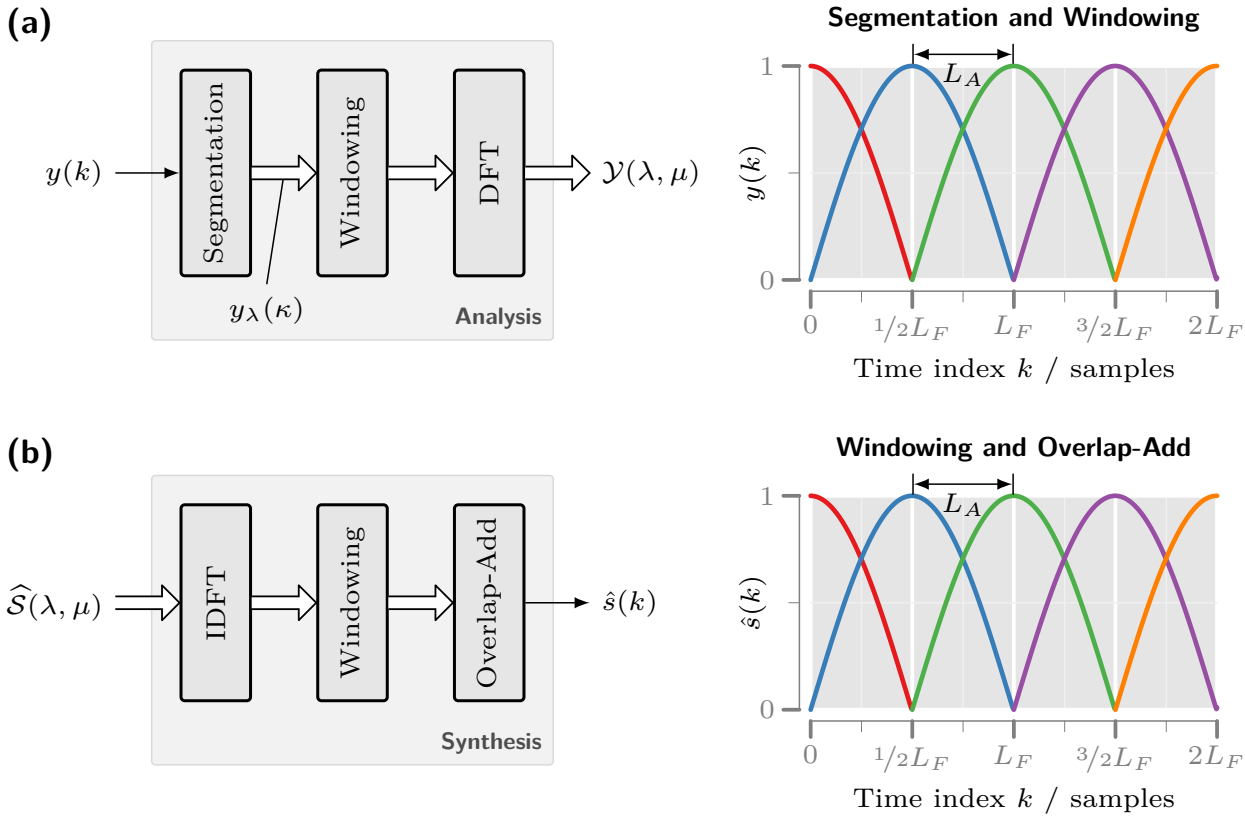



**Figure 3.2:** Analysis – synthesis framework: **(a)** Analysis block including segmentation, windowing and DFT, **(b)** Synthesis consisting of IDFT, windowing and overlap-add.

to short-term (quasi) stationarity into frames of $L_F$ samples which may overlap according to

$$y_\lambda(\kappa) = y(\lambda \cdot L_A + \kappa) \quad \text{with } \kappa \in \{0, \ldots, L_F - 1\},\ \lambda \in \mathbb{N}_0, \tag{3.2}$$

where $\lambda$ is the frame index, $L_A$ the frame advance in samples, and $\kappa$ is the sample position within one frame. If no overlap is required $L_A$ equals $L_F$. In order to avoid major discontinuities at the frame edges and to counteract the spectral leakage effect, a tapered window function $w(\kappa)$ is applied to each frame [Vary & Martin 2006]. The effect of the window function is a fade in and fade out of the frame. In addition, this reduces the unavoidable cyclic effects of DFT domain processing. In consequence, windowing requires a frame overlap to ensure perfect reconstruction of the frames during synthesis. An example of a window can be seen in Fig. 3.3 (e. g., one of the colored curves). Suitable and frequently deployed window functions are the Hann window, Hamming window or Blackmann window [Oppenheim et al. 1989]. Typical values for the frame length in speech processing are $T_F \in \{5\,\text{ms}, \ldots, 40\,\text{ms}\}$ [Paliwal et al. 2010; Vary & Martin 2006], which results in a frame-size in samples of

$$L_F = \lfloor T_F \cdot f_\text{s} \rfloor, \tag{3.3}$$

with a frame overlap of typical 50 % – 75 % [Benesty et al. 2007; Loizou 2013]. Arbitrary frame overlaps require a window function whose shifted versions according to $L_A$ add up to at least a constant value or to one for perfect reconstruction[1]. The Hann window fulfills this requirement at least for a subset of frame overlaps. Due to the symmetric bell shape of the window function the maximum value is located in the center of the window. Hence, the frame overlap has to be greater or equal than 50 % for perfect reconstruction. To circumvent this limitation the so called flat top Hann window is introduced here, which allows arbitrary frame overlaps and perfect reconstruction. The window is composed of three parts: a half Hann window, a series of ones and the second half of the Hann window. The size of each part depends on the frame size and the frame advance. The flat top Hann window $w_{\text{ftHann}}(\kappa)$ for frame size $L_F$ and frame advance $L_A$, where $L_A \leq L_F$, is calculated according to:

$$N_{w/2} = L_F - \left\lceil \frac{L_F}{2L_A} \right\rceil \cdot L_A \tag{3.4}$$

$$w_{\text{Hann}/2}(\kappa) = \frac{1}{2}\left(1 - \cos\left(\frac{\pi\kappa}{N_{w/2} - 1}\right)\right) \tag{3.5}$$

$$w_{\text{ftHann}}(\kappa) = \begin{cases} w_{\text{Hann}/2}(\kappa) & \text{if } 0 \leq \kappa < N_{w/2} \\ 1 & \text{if } N_{w/2} \leq \kappa < L_F - N_{w/2} \\ w_{\text{Hann}/2}(L_F - 1 - \kappa) & \text{if } L_F - N_{w/2} \leq \kappa < L_F \\ 0 & \text{otherwise .} \end{cases} \tag{3.6}$$

From 100 % down to 50 % frame overlap the flat top Hann window is identical to a regular Hann window

$$w_{\text{Hann}}(\kappa) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi\kappa}{L_F - 1}\right)\right). \tag{3.7}$$

With decreasing overlap but less than 50 % the center of the window is filled with ones and in the border case where $L_F = L_A$, i. e., no frame overlap, the flat top Hann window results in a rectangular window.

As depicted in Fig. 3.2 the window function is applied in the considered framework during analysis and synthesis. The window function is applied twice for the following reasons. On the one hand negative effects mainly caused by cyclic convolution which are introduced due to spectral modifications are reduced and on the other hand the spectral modifications are cross-faded in the overlapping parts of the windows [Marin-Hurtado & Anderson 2011]. Doing so, the square root function is applied to the window function due to the multiplicative concatenation of analysis and synthesis which yields an allover perfect reconstruction. The resulting square root Hann windows are visualized on the right hand side of Fig. 3.2.

After segmenting and windowing and, if necessary, zero-padding each resulting noisy short-time frame $y_\lambda(\kappa)$ is transformed into the frequency domain using a

[1] Note if the constant is not equal to one, a normalization has to be applied within the synthesis.

short-time *discrete Fourier transform* (DFT)[2] of length $N_{\mathrm{DFT}}$. Zero-padding is required if $N_{\mathrm{DFT}} > L_F$. The spectrum of the input signal $y(k)$ at frequency bin $\mu$ and frame $\lambda$ is given by:

$$y_\lambda(\kappa) \cdot w(\kappa) \quad \overset{\mathrm{DFT}}{\circ\!\!-\!\!\bullet} \quad \mathcal{Y}(\lambda,\mu) = \mathcal{S}(\lambda,\mu) + \mathcal{N}(\lambda,\mu), \tag{3.8}$$

where $\mathcal{S}(\lambda,\mu)$ and $\mathcal{N}(\lambda,\mu)$ correspond to the spectral coefficients of the speech and noise signal, respectively. Note that the frequency domain representation of the input, the speech and noise signal already includes the effect of windowing. Thereafter, the actual speech enhancement takes place in the frequency domain which is detailed in Sec. 3.4.

In order to obtain the enhanced signal in the time domain the operations which are applied in the analysis stage are reversed. As presented in Fig. 3.2b the enhanced frames $\widehat{\mathcal{S}}(\lambda,\mu)$ are transformed into time domain segments $\hat{s}_\lambda(\kappa)$ using the IDFT. Subsequently, the window function is applied. Since it is possible that the windows add up to a constant greater than one (due to the overlap) a normalization factor $g_w$ within the synthesis procedure is necessary which is calculated by:

$$N_w = \left\lceil \frac{L_F}{L_A} \right\rceil , \tag{3.9}$$

$$g_w = \left\lfloor \frac{N_w}{2} \right\rfloor + N_w \bmod 2 , \tag{3.10}$$

where $N_w$ specifies the number of beginning windows within one frame. Finally, the enhanced signal in the time domain $\hat{s}(k)$ is constructed by overlap-add [Crochiere 1980] of the normalized and windowed segments

$$\widehat{\mathcal{S}}(\lambda,\mu) \quad \overset{\mathrm{IDFT}}{\bullet\!\!-\!\!\circ} \quad \hat{s}_\lambda(\kappa), \tag{3.11}$$

$$\hat{s}(k) = \sum_{\lambda \in \mathbb{N}_0} \frac{1}{g_w} \cdot \hat{s}_\lambda(\kappa) \cdot w(\kappa), \tag{3.12}$$

with $\kappa = \lambda \cdot L_A - k$. Note that the enhanced signal frames in the time domain $\hat{s}_\lambda(\kappa)$ and the window function $w(\kappa)$ are zero for $0 < \kappa > L_F - 1$.

An example of the overall windowing including the analysis and synthesis stage is depicted in Fig. 3.3 for successive frames. The effective windows are colored and the sum of the windows is indicated by the black curve. Note for perfect reconstruction at least $N_w$ overlapping windows are necessary. Hence, perfect reconstruction can not be achieved at the beginning and the end of $\hat{s}(k)$. If not stated otherwise the square-root Hann window is used and a frame length of 20 ms is applied with an frame overlap of 50 % throughout this thesis.

Both, the analysis and the synthesis stage are not subject of this work. Different solutions can be found for an analysis – synthesis framework, e. g., by a filter-bank

[2]Throughout this thesis the *fast Fourier transform* (FFT) [Cooley & Tukey 1965] is used as efficient implementation of the DFT. A prerequisite for applying the FFT is a frame size of a power of two.

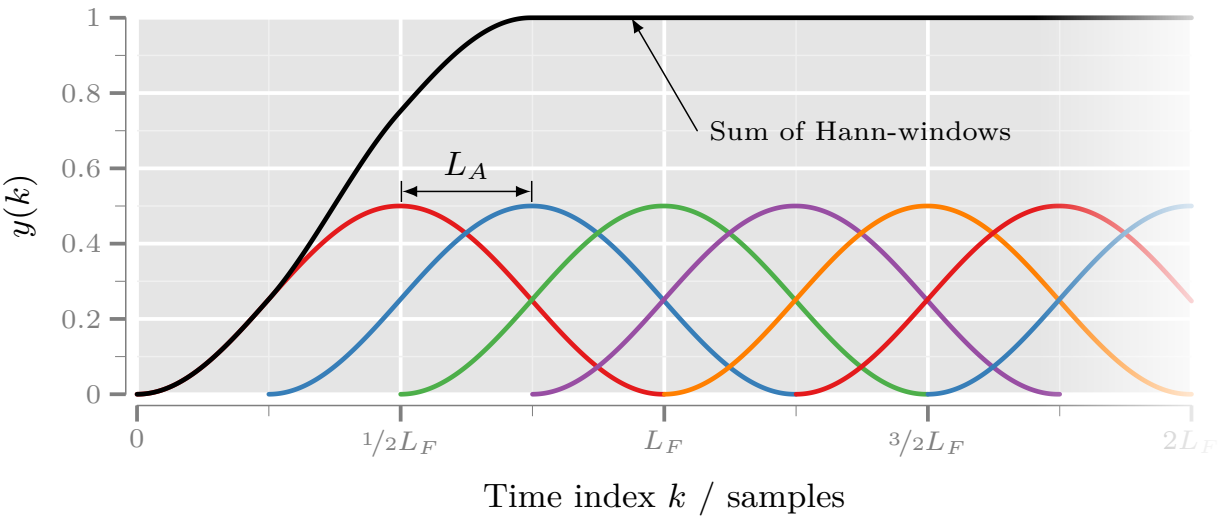



**Figure 3.3:** Windowing example with $L_A = 1/4 L_F$

structure (see [Löllmann 2011] and references therein). This work focuses on the modification block depicted in Fig. 3.1, i. e., the detection of speech and noise from the noisy input and the subsequent enhancement of the degraded speech signal.

## 3.3 Signal Properties

Quite often, only a single realization of a stochastic process can be observed. Subsequently, it is not possible to estimate its statistics by averaging over an ensemble of observations. If the true statistics of a stationary random process may be obtained from the time averages over single observations within time, the random process is called ergodic, i. e., the state space of the random process is completely covered over the time. Given an ergodic random process the ensemble averages can be replaced by time averages, e. g., [Papoulis & Pillai 2002].

In the context of speech enhancement, statistical quantities can only be estimated by time averages. Since speech and noise signals are not stationary and thus not ergodic, the signals are subdivided into short time segments which are considered as stationary. Hence, to apply time averages the underlying signals are required to be short-term stationary. Therefore, it is beneficial to define certain expectation operators. The expectation of $x(k)$ is defined by

$$\mathbb{E}\{x(k)\} = \lim_{K\to\infty} \frac{1}{2K+1} \sum_{\kappa=-K}^{K} x(\kappa). \tag{3.13}$$

The mean short-term expectation of $x(k)$ is defined by

$$\overline{\mathbb{E}}_K\{x(k)\} = \frac{1}{K} \sum_{\kappa=0}^{K-1} x(k+\kappa) \tag{3.14}$$

The recursive short-term expectation of $x(k)$ is defined by

$$\tilde{\mathbb{E}}_{\alpha}\left\{x(k)\right\} = (1-\alpha)\cdot x(k) + \alpha \cdot x(k-1). \tag{3.15}$$

The parameters $K$ and $\alpha$ control the smoothing properties of the respective short-term expectation estimator. Assuming an uncorrelated signal $x(k)$, a relation between the two parameters can be found by equating the variance of the mean short-term expectation $\overline{\mathbb{E}}_{K}\left\{\cdot\right\}$ and recursive short-term expectation $\tilde{\mathbb{E}}_{\alpha}\left\{\cdot\right\}$ estimators. It can be shown that the equivalent rectangular window length of the recursive short-term expectation estimator is given by

$$\alpha = \frac{K-1}{K+1}, \tag{3.16}$$

in terms of the samples $K$ which are used for the moving average estimator and vice versa

$$K = \frac{1+\alpha}{1-\alpha}. \tag{3.17}$$

Refer to Appendix B for a detailed derivation.

With regard to speech enhancement, most of the algorithms are derived based on *power spectral density* (PSD) $\Phi(\mu)$, short-term PSD $\overline{\Phi}(\lambda,\mu)$ or power signal quantities $|\cdot|^2$. The computation of power quantities should be normalized to the frame-size for a correct physical definition, but will be neglected as it is usually done in literature. This is possible as within a specific speech enhancement system the used frame-size and frame advance are fixed and therefore no normalization is necessary. Moreover, power quantities are almost always used in ratios, e. g., for SNR computation. Hence, the dependency of the normalization factor on the frame-size is canceled out. Thus, the PSD of $x(k)$ is defined as

$$\Phi_{xx}(\mu) = \mathbb{E}\left\{\left|\mathcal{X}(\lambda,\mu)\right|^2\right\}, \tag{3.18}$$

where $\mathcal{X}(\lambda,\mu)$ is the frequency representation of $x(k)$ according to Sec. 3.2.1. The short-term PSD is given by

$$\overline{\Phi}_{xx}(\lambda,\mu) = \overline{\mathbb{E}}_{K}\left\{\left|\mathcal{X}(\lambda,\mu)\right|^2\right\}. \tag{3.19}$$

## 3.4 Conventional Noise Suppression

Most state-of-the-art noise reduction systems are realized in a framework as depicted in Fig. 3.1 employing the presented or a similar analysis – synthesis framework (see Sec. 3.2.1). In the following the functionality of the intermediate *modification block* in Fig. 3.1 is described in detail including state-of-the-art examples. A common approach to enhance degraded speech is presented in Fig. 3.4. It consists of a detection or estimation part on the left hand side and the actual speech enhancement

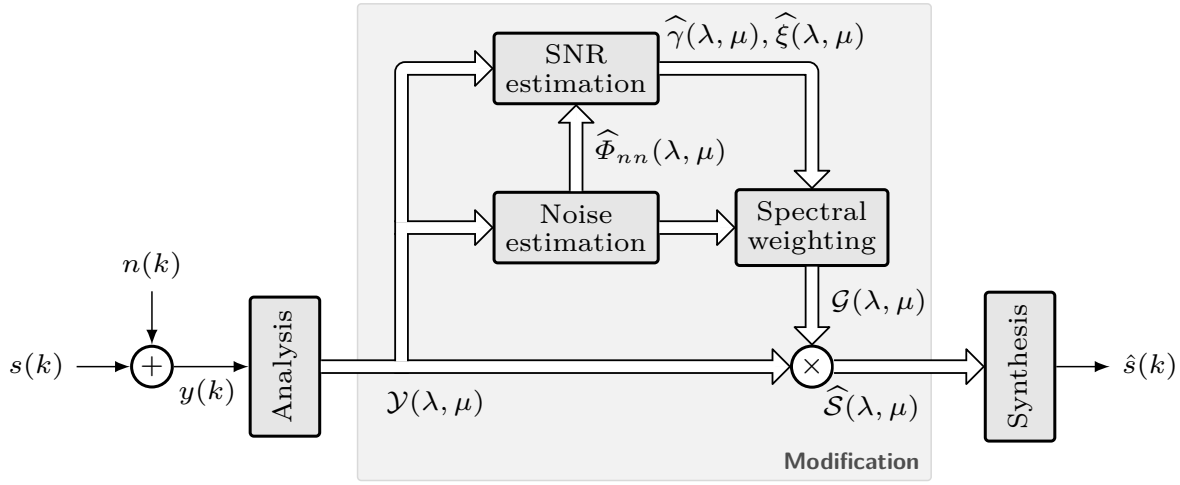



**Figure 3.4:** Block diagram of standard noise reduction system working in the frequency domain

on the right hand side. The estimate[3] of the short-term noise PSD $\widehat{\Phi}_{nn}(\lambda,\mu)$ is the basis from which the *a priori* SNR estimate $\widehat{\xi}(\lambda,\mu)$ and the *a posteriori* SNR estimate $\widehat{\gamma}(\lambda,\mu)$ can be calculated. A weighting gain $\mathcal{G}(\lambda,\mu)$ is computed which aims to minimize a specific distortion measure between the clean speech $\mathcal{S}(\lambda,\mu)$ and the speech estimate $\widehat{\mathcal{S}}(\lambda,\mu)$ signal as a function of one or more of the estimated quantities $\widehat{\Phi}_{nn}(\lambda,\mu)$, $\widehat{\xi}(\lambda,\mu)$ and $\widehat{\gamma}(\lambda,\mu)$. The actual noise reduction is carried out by spectral weighting, i. e., multiplying the noisy input $\mathcal{Y}(\lambda,\mu)$ with the spectral gain $\mathcal{G}(\lambda,\mu)$ and results in the clean speech estimate $\widehat{\mathcal{S}}(\lambda,\mu)$. Frequency bins of the noisy input signal $\mathcal{Y}(\lambda,\mu)$ which contain mostly noise shall be damped while frequency bins comprising mainly speech shall pass. Utilizing the synthesis stage the corresponding enhanced time domain signal $\hat{s}(k)$ is created.

### 3.4.1 Noise Estimation

All speech enhancement systems covered in this thesis rely on knowledge about the short-term noise PSD. The estimation of the short-term noise PSD $\widehat{\Phi}_{nn}(\lambda,\mu)$ remains a crucial and challenging task in every noise reduction system, especially in case of non-stationary noise. Noise estimation algorithms usually rely on the assumption that speech and noise have different temporal statistics which can be used to estimate the noise from the noisy input signal. Overestimation of the noise leads likely to over-attenuation of the speech signal resulting in strong speech distortions. On the other hand, high remaining levels of noise is the consequence of noise underestimation.

If the noise is stationary or only slowly varying in time, a short-term noise PSD estimate can either be obtained during speech pauses or by continuously tracking versus time the magnitude minima in the short-time Fourier domain.

[3] An estimated signal or parameter is indicated by the hat symbol $\widehat{\ }$ throughout this thesis

Further processing and updating over time is necessary. Several methods have been proposed for the estimation of the short-term noise PSD by tracking and post-processing the magnitude minima in the short-time Fourier domain, e. g., [Baasch et al. 2014; Cohen 2003; Cohen & Berdugo 2002; Doblinger 1995; Gerkmann & Hendriks 2011; Hendriks et al. 2010; Martin 2001, 2006].

A comparison of state-of-the-art single microphone short-term noise PSD estimators can be found in [Taghia et al. 2011]. The most important methods are briefly presented in the following.

### Voice Activity Detection

One of the first approaches known from literature for the estimation of the short-term noise PSD, e. g., [McAulay & Malpass 1980; Van Compernolle 1989] is based on a *voice activity detector* (VAD). A short-term noise PSD estimate is obtained by updating the noise PSD only in phases of speech absence. A simple noise PSD estimate is provided using a first order recursive system with $0 < \alpha_\Phi < 1$ given by

$$\widehat{\Phi}_{nn}(\lambda, \mu) = \alpha_\Phi \cdot \widehat{\Phi}_{nn}(\lambda - 1, \mu) + (1 - \alpha_\Phi) \cdot |\mathcal{Y}(\lambda, \mu)|^2 , \tag{3.20}$$

while speech is absent and kept constant during speech presence, i. e., $\alpha_\Phi = 1$. However, the quality of VAD is limited by the input SNR leading to unreliable VAD estimates for low input SNR conditions [Vary & Martin 2006]. Hence, a suitable short-term noise PSD estimate is only possible for moderate SNR conditions and rather stationary background noise. In recent years more sophisticated methods were developed which update the noise PSD also during speech presence.

### Minimum Tracker

In [Doblinger 1995] the noise spectrum is estimated for each frequency bin based on a temporally smoothed periodogram of the noisy observation by nonlinear temporal minima tracking. If the last noise PSD estimate is smaller than the current noisy observation the tracking is realized by a weighted average of the last and current noisy frame. In the other case the current noisy observation serves as new noise PSD estimate.

### Minimum Statistics

The *Minimum Statistics* [Martin 1994, 2001, 2006] method is based on two assumptions:

- speech and noise are statistically independent and
- the power of the noisy signal often decays to the power level of the noise.

Using a smoothed periodogram of the noisy signal, it is possible to track a minimum separately in each frequency bin within a certain sliding time window to obtain a

short-term noise PSD estimate $\widehat{\Phi}_{nn}(\lambda, \mu)$. The smoothed signal power is also given by a first order recursive system according to

$$\left|\overline{\mathcal{Y}}(\lambda, \mu)\right|^2 = \alpha_{\mathrm{MS}}(\lambda, \mu) \cdot \left|\overline{\mathcal{Y}}(\lambda - 1, \mu)\right|^2 + (1 - \alpha_{\mathrm{MS}}(\lambda, \mu)) \cdot \left|\mathcal{Y}(\lambda, \mu)\right|^2, \quad (3.21)$$

with $0 < \alpha_{\mathrm{MS}}(\lambda, \mu) < 1$ denoting a time and frequency dependent smoothing factor. The smoothing factor $\alpha_{\mathrm{MS}}(\lambda, \mu)$ minimizes the conditional *minimum mean-square error* (MMSE) between the true noise $\widehat{\Phi}_{nn}(\lambda, \mu)$ and the smoothed signal power $\left|\overline{\mathcal{Y}}(\lambda, \mu)\right|^2$. The smoothing factor can be expressed as a function of the smoothed *a posteriori* SNR [Martin 2001]. Afterwards the minimum within a sliding time window of the past $L_{\mathrm{MS}}$ frames is computed separately for each frequency bin by

$$\left|\overline{\mathcal{Y}}(\lambda, \mu)_{\min}\right|^2 = \min_{\tilde{\lambda}} \left|\overline{\mathcal{Y}}(\tilde{\lambda}, \mu)\right|^2, \quad (3.22)$$

with $\tilde{\lambda} \in \{\lambda - L_{\mathrm{MS}} + 1, \ldots, \lambda\}$ representing the frame index of the sliding time window. The duration of the time window for the minimum search states a trade-off between fast noise tracking and remaining speech portions in the noise PSD estimate. A typical value for the time window length corresponds to 1.5 s. As the minimum is always smaller or equal to the mean noise power a bias correction $B(\lambda, \mu)$ is necessary. The bias correction is mainly dependent on the length of the minimum search interval and on the variance of the noisy input periodogram and thus dependent on the smoothing parameter $\alpha_{\mathrm{MS}}(\lambda, \mu)$ of the periodogram. The short-term noise PSD estimate is finally given by

$$\widehat{\Phi}_{nn}(\lambda, \mu) = B(\lambda, \mu, \alpha_{\mathrm{MS}}(\lambda, \mu)) \cdot \left|\overline{\mathcal{Y}}(\lambda, \mu)_{\min}\right|^2. \quad (3.23)$$

*Minimum Statistics* performs well in stationary and slowly changing noise conditions as the minimum at each frequency bin within the search time window provides a good estimate of the actual noise power.

#### Noise power estimation based on the probability of speech presence (SPP)

Given a reliable VAD, the aforementioned VAD based noise estimator updates the short-term noise PSD estimate only in phases of speech absence. In contrast, the *SPP* algorithm [Gerkmann & Hendriks 2011, 2012], which is a further development of [Hendriks et al. 2010], estimates the noise PSD for each frequency by a smoothed linear combination of the current observed noisy short-term PSD and the last estimate of the noise PSD weighted by the speech presence and speech absence probability, respectively. The *speech presence probability* (SPP) is a time and frequency dependent soft value for the speech activity ranging between zero and one. Assuming a Gaussian distribution for the real and imaginary components of the noise and speech spectral coefficients the SPP can be formulated. Applying Bayes's theorem, the probability $p$ of speech presence $H_1$ can be expressed given the noisy observation $\mathcal{Y}(\lambda, \mu)$ and a noise PSD estimate $\widehat{\Phi}_{nn}(\lambda, \mu)$ according to

[Cohen & Berdugo 2001] by[4]

$$p(H_1|\mathcal{Y}(\lambda,\mu)) = \left(1 + (1+\xi_{\text{opt}})\exp\left(-\frac{|\mathcal{Y}(\lambda,\mu)|^2}{\widehat{\Phi}_{nn}(\lambda,\mu)} \cdot \frac{\xi_{\text{opt}}}{\xi_{\text{opt}}+1}\right)\right)^{-1}, \quad (3.24)$$

where the fixed optimal *a priori* SNR $\xi_{\text{opt}}$ is chosen as $10\log_{10}(\xi_{\text{opt}}) = 15\,\text{dB}$ under the constraint that the true *a priori* SNR is less or equal to 20 dB [Gerkmann et al. 2008]. Moreover, the speech presence and speech absence is modeled equiprobable, i. e.,

$$p(H_1) = p(H_0) = 0.5. \quad (3.25)$$

If the short-term noise PSD estimate $\widehat{\Phi}_{nn}(\lambda,\mu)$ underestimates the true short-term noise power, the SPP $p(H_1|\mathcal{Y}(\lambda,\mu))$ is overestimated since the denominator of Eq. (3.24) gets smaller due to the dominant ratio $-|\mathcal{Y}(\lambda,\mu)|^2/\widehat{\Phi}_{nn}(\lambda,\mu)}$ in $\exp(\cdot)$. In the extreme case, i. e., $\widehat{\Phi}_{nn}(\lambda,\mu) << |\mathcal{Y}(\lambda,\mu)|^2$, the SPP $p(H_1|\mathcal{Y}(\lambda,\mu))$ tends to one although $|\mathcal{Y}(\lambda,\mu)|^2$ is small with respect to the true, but unknown, noise power. In order to avoid stagnation for SPP values close to one post-processing of $p(H_1|\mathcal{Y}(\lambda,\mu))$ is applied, including recursive smoothing and bounding the smoothed SPP to an upper limit. According to [Gerkmann & Hendriks 2011] the SPP can be interpreted as frequency and time dependent soft VAD which is suitable to control the update of the noise periodogram leading to

$$\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2 = p(H_0|\mathcal{Y}(\lambda,\mu))\,|\mathcal{Y}(\lambda,\mu)|^2 + p(H_1|\mathcal{Y}(\lambda,\mu))\widehat{\Phi}_{nn}(\lambda-1,\mu) \quad (3.26)$$

with the probability of speech absence given by

$$p(H_0|\mathcal{Y}(\lambda,\mu)) = 1 - p(H_1|\mathcal{Y}(\lambda,\mu)). \quad (3.27)$$

Finally, the spectral noise power estimate is computed by temporal smoothing of the noise periodogram according to

$$\widehat{\Phi}_{nn}(\lambda,\mu) = 0.8 \cdot \widehat{\Phi}_{nn}(\lambda-1,\mu) + 0.2 \cdot \left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2. \quad (3.28)$$

The evaluation carried out in [Gerkmann & Hendriks 2012] confirmed a good performance for this noise PSD tracking algorithm also in challenging noise environments, i. e., in case of at least slowly time-varying noise.

### 3.4.2 Signal-to-noise ratio Estimation

The *signal-to-noise ratio* (SNR) is an important measurement for speech enhancement and is exploited by many algorithms. Various spectral weighting rules can be formulated as a function of the SNR. Two important SNR quantities are the

[4] In a realistic system the noise estimate $\widehat{\Phi}_{nn}(\lambda,\mu)$ is approximated by the noise estimate from the previous frame $\widehat{\Phi}_{nn}(\lambda-1,\mu)$ to estimate the SPP.

*a priori* SNR $\xi$ and the *a posteriori* SNR $\gamma$. Their estimates are defined in [McAulay & Malpass 1980]. The *a posteriori* SNR is defined as the ratio between the noisy periodogram and the short-term noise PSD as given by

$$\overline{\gamma}(\lambda,\mu) = \frac{|\mathcal{Y}(\lambda,\mu)|^2}{\overline{\Phi}_{nn}(\lambda,\mu)} = \frac{|\mathcal{Y}(\lambda,\mu)|^2}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}}, \tag{3.29}$$

where $\overline{\mathbb{E}}_K\{\cdot\}$ represents the short-term mean expectation operator, i. e., the short-term average of its argument in this context as defined in Sec. 3.3. Given an estimate of the short-term noise PSD $\widehat{\Phi}_{nn}(\lambda,\mu)$ the *a posteriori* SNR can easily be measured. In contrast, the *a priori* SNR defined by

$$\overline{\xi}(\lambda,\mu) = \frac{\overline{\Phi}_{ss}(\lambda,\mu)}{\overline{\Phi}_{nn}(\lambda,\mu)} = \frac{\overline{\mathbb{E}}_K\left\{|\mathcal{S}(\lambda,\mu)|^2\right\}}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}}, \tag{3.30}$$

is more challenging to estimate since the short-term PSD of speech $\overline{\Phi}_{ss}(\lambda,\mu)$ is necessary. In general, $\overline{\Phi}_{ss}(\lambda,\mu)$ is not known *a priori*. Using the relation

$$\mathcal{Y}(\lambda,\mu) = \mathcal{S}(\lambda,\mu) + \mathcal{N}(\lambda,\mu), \tag{3.31}$$

and assuming again that speech and noise are uncorrelated the cross terms

$$\overline{\Phi}_{sn}(\lambda,\mu) = \overline{\Phi}_{ns}(\lambda,\mu) = 0, \tag{3.32}$$

are close to zero and the *a priori* SNR can now be formulated in terms of the *a posteriori* SNR according to:

$$\overline{\xi}(\lambda,\mu) = \frac{\overline{\Phi}_{ss}(\lambda,\mu)}{\overline{\Phi}_{nn}(\lambda,\mu)} = \frac{|\mathcal{Y}(\lambda,\mu)|^2}{\overline{\Phi}_{nn}(\lambda,\mu)} - 1 = \overline{\gamma}(\lambda,\mu) - 1. \tag{3.33}$$

The *decision-directed* approach is a widely accepted method in literature to estimate the *a priori* SNR $\xi(\lambda,\mu)$ and was suggested by [Ephraim & Malah 1984]. It is assumed that a speech estimate $\widehat{\mathcal{S}}(\lambda-1,\mu)$ of the previous frame is available and furthermore that $\mathcal{S}(\lambda,\mu) \approx \mathcal{S}(\lambda-1,\mu)$, which is true for a quasi-stationary speech sound but less valid for, e. g., transient sounds. Now, the *a priori* SNR is computed by a linear combination of speech and noise estimates from the last frame and an instantaneous realization of the *a posteriori* SNR

$$\hat{\xi}(\lambda,\mu) = \alpha_\xi \frac{\left|\widehat{\mathcal{S}}(\lambda-1,\mu)\right|^2}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda-1,\mu)|^2\right\}} + (1-\alpha_\xi)\max\left(\overline{\gamma}(\lambda,\mu)-1,\ 0\right), \tag{3.34}$$

where $\max(\cdot,\cdot)$ returns the maximum of the two arguments. The choice of $\alpha_\xi$ states a tradeoff between noise reduction and speech distortion. A typical value for $\alpha_\xi$ lies in the range $0.9 < \alpha_\xi < 0.99$. In this work $\alpha_\xi = 0.98$ is chosen as suggested in [Ephraim & Malah 1984].

### 3.4.3 Spectral Weighting

As depicted in Fig. 3.4 the actual noise reduction is achieved by spectral weighting yielding the enhanced speech estimate $\widehat{\mathcal{S}}(\lambda, \mu)$ in the frequency domain by

$$\widehat{\mathcal{S}}(\lambda, \mu) = \mathcal{G}(\lambda, \mu) \cdot \mathcal{Y}(\lambda, \mu) = \mathcal{G}(\lambda, \mu) \cdot |\mathcal{Y}(\lambda, \mu)| \, \mathrm{e}^{\mathrm{i}\vartheta_{\mathcal{Y}}(\lambda,\mu)}, \tag{3.35}$$

where $|\mathcal{Y}(\lambda, \mu)|$ is the noisy magnitude and $\vartheta_{\mathcal{Y}}(\lambda, \mu)$ the corresponding phase at frequency bin $\mu$ and frame $\lambda$. The weighting gain is updated in each frame and the calculation is usually a function of the previously introduced short-term noise PSD estimate $\widehat{\Phi}_{nn}(\lambda, \mu)$ and the SNR estimates $\widehat{\xi}(\lambda, \mu)$ and $\widehat{\gamma}(\lambda, \mu)$. Typically, the optimization of the weighting gain function aims to minimize a specific mathematical cost function between the clean speech signal $\mathcal{S}(\lambda, \mu)$ and its estimate $\widehat{\mathcal{S}}(\lambda, \mu)$ assuming certain statistical characteristics about speech and noise. Often used cost functions are the MMSE, the *maximum likelihood* (ML) or *maximum* a posteriori (MAP) criterion. In general, the weighting gains can be complex-valued. As the human auditory system is rather insensitive w. r. t. to phase distortions [Vary 1985; Wang & Lim 1982] most weighting gain rules modify only the spectral magnitudes of the noisy DFT coefficients. Doing so, $\mathcal{G}(\lambda, \mu)$ is real-valued and lies in the range between zero and one. Hence, the noisy phase is applied during synthesis to obtain the enhanced speech signal in the time domain.

In the following the well-known Wiener filter weighting rule [Lim & Oppenheim 1979; Vaseghi 1996] is presented. The Wiener filter minimizes the MMSE between the clean speech DFT coefficients $\mathcal{S}(\lambda, \mu)$ and the enhanced DFT coefficients $\widehat{\mathcal{S}}(\lambda, \mu)$ independently for each frequency bin $\mu$ assuming Gaussian *probability density function*s (PDFs) for both. Using Eq. (3.35) it follows for the MMSE expression:

$$\mathbb{E}\left\{\left|\mathcal{S}(\lambda, \mu) - \widehat{\mathcal{S}}(\lambda, \mu)\right|^2\right\} = \mathbb{E}\left\{|\mathcal{S}(\lambda, \mu) - \mathcal{G}(\lambda, \mu) \cdot \mathcal{Y}(\lambda, \mu)|^2\right\} \stackrel{!}{=} \min. \tag{3.36}$$

Assuming that the DFT coefficients are independent, it can be shown that the partial derivation of Eq. 3.36 with respect to the real and imaginary parts of $\mathcal{G}(\lambda, \mu)$

$$\frac{\partial\, \mathbb{E}\left\{\left|\mathcal{S}(\lambda, \mu) - \widehat{\mathcal{S}}(\lambda, \mu)\right|^2\right\}}{\partial\, \mathrm{Im}\{\mathcal{G}(\lambda, \mu)\}} = 0, \quad \frac{\partial\, \mathbb{E}\left\{\left|\mathcal{S}(\lambda, \mu) - \widehat{\mathcal{S}}(\lambda, \mu)\right|^2\right\}}{\partial\, \mathrm{Re}\{\mathcal{G}(\lambda, \mu)\}} = 0 \tag{3.37}$$

yields [Vaseghi 1996]

$$\mathrm{Im}\{\mathcal{G}(\lambda, \mu)\} = 0, \tag{3.38}$$

$$\mathrm{Re}\{\mathcal{G}(\lambda, \mu)\} = \frac{\mathbb{E}\left\{|\mathcal{S}(\lambda, \mu)|^2\right\}}{\mathbb{E}\left\{|\mathcal{S}(\lambda, \mu)|^2 + |\mathcal{N}(\lambda, \mu)|^2\right\}} = \frac{\widehat{\Phi}_{ss}(\lambda, \mu)}{\widehat{\Phi}_{ss}(\lambda, \mu) + \widehat{\Phi}_{nn}(\lambda, \mu)} \tag{3.39}$$

where Re{·} and Im{·} denote the real and imaginary parts, respectively. Hence, the Wiener filter weighting is real valued and it can also be expressed in terms of the *a priori* SNR estimate $\widehat{\xi}$ as

$$\mathcal{G}_W(\lambda,\mu) = \frac{\widehat{\xi}(\lambda,\mu)}{1+\widehat{\xi}(\lambda,\mu)}. \tag{3.40}$$

Another popular approach is called spectral subtraction and was proposed by [Boll 1979]. The noise reduction is achieved by subtracting an estimate of the noise magnitude spectrum from the noisy speech magnitude spectrum according to

$$\left|\widehat{\mathcal{S}}(\lambda,\mu)\right| = |\mathcal{Y}(\lambda,\mu)| - \mathbb{E}\left\{|\mathcal{N}(\lambda,\mu)|\right\}, \tag{3.41}$$

which leads to the weighting gain calculation rule:

$$\mathcal{G}_S(\lambda,\mu) = 1 - \frac{\mathbb{E}\left\{|\mathcal{N}(\lambda,\mu)|\right\}}{|\mathcal{Y}(\lambda,\mu)|}. \tag{3.42}$$

In [Hansen 1991] a gain rule was proposed which generalizes the approach of [Boll 1979] introducing the two parameters $\alpha_{\mathcal{G}}$ and $\beta_{\mathcal{G}}$ and using the noise estimate $\widehat{\mathcal{N}}(\lambda,\mu)$. The gain rule is given by

$$\mathcal{G}(\lambda,\mu) = \sqrt{\left[1 - \left(\frac{\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2}{|\mathcal{Y}(\lambda,\mu)|^2}\right)^{\beta_{\mathcal{G}}}\right]^{\alpha_{\mathcal{G}}}}. \tag{3.43}$$

The parameters $\alpha_{\mathcal{G}}$ and $\beta_{\mathcal{G}}$ can be either fixed or adaptive incorporating the characteristics of speech and noise over the time. Using $\alpha_{\mathcal{G}} = 2$ and $\beta_{\mathcal{G}} = 0.5$ yields the spectral subtraction rule by [Boll 1979], whereas $\alpha_{\mathcal{G}} = \beta_{\mathcal{G}} = 1$ leads to the power subtraction rule. Setting $\alpha_{\mathcal{G}} = 2$ and $\beta_{\mathcal{G}} = 1$ results in the Wiener filter weighting gain (refer Eq. 3.39).

## 3.5 Noise Estimation by Logarithmic Baseline Tracing

A novel noise PSD estimator for disturbed speech signals that operates in the short-time Fourier domain is presented [Heese & Vary 2015]. A short-term noise PSD estimate is provided by constrained tracing with time the noisy observation separately for each frequency bin. The constraint is a limitation of the logarithmic magnitude change between successive time frames. Since speech onsets are assumed as sudden rises in the noisy observation, a fixed and an adaptive tracing parameter $\beta$ will be derived to track the contained noise while preventing speech leakage to the noise PSD estimate. In other words, the new estimator is explicitly designed to estimate all signal components with a lower dynamic than speech. Hence, the remaining signal estimate is considered as noise. The constraint frequency

dependent magnitude change causes inertia of the noise estimate over time which models the different temporal and frequency dependent statistics of speech and noise.

The experimental evaluation and comparison with state-of-the-art algorithms, *SPP* [Gerkmann & Hendriks 2011] and *Minimum Statistics* [Martin 2001, 2006], confirms a lower logarithmic noise estimation error and superior speech enhancement rated in a standard noise reduction system. The proposed concept has an extremely low computational complexity and memory consumption. Thus, it is well suited for applications where processing power and memory is limited.

The property of the new estimator to largely prevent speech leakage to the noise estimate along with the low computational complexity is an important feature for information combining as detailed in Sec. 5 of different speech and noise short-term PSDs.

### 3.5.1 Signal Model

For the derivation of the new short-term noise PSD estimator it is assumed that the speech and noise signals have zero mean and are independent so that

$$\mathbb{E}\left\{|\mathcal{Y}(\lambda,\mu)|^2\right\} = \mathbb{E}\left\{|\mathcal{S}(\lambda,\mu)|^2\right\} + \mathbb{E}\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}. \tag{3.44}$$

### 3.5.2 Definition of the Noise Signal Baseline

In most derivations of (short-term) noise PSD estimators speech and noise are assumed as uncorrelated and the noise is modeled as a stationary process [Gerkmann & Hendriks 2012; Martin 2006]. Hence, applying the expectation operator in the derivations cancels the speech-noise cross-terms out which simplifies the estimation problem. As noise estimators and speech enhancement systems operate on a frame-by-frame basis, this simplifications do not hold. In this section the aforementioned simplifications are analyzed relaxing the requirement of the noise to be stationary and by formulating the estimation problem including the speech-noise cross-terms. The final estimation term can be expressed in terms of a baseline which is equivalent to the short-term noise PSD.

In the following consideration an arbitrary but stationary noise only signal is assumed. Since most speech enhancement algorithms are derived based on the noise signal *power spectral density* (PSD) $\Phi_{nn}(\mu)$, the determination of the noise signal PSD or at least the short-term noise PSD $\overline{\Phi}_{nn}(\lambda,\mu)$ is the objective of a noise estimator. Quite often, it is not possible to observe more than a single realization, i. e., a noise signal frame, of a stochastic noise signal process. Then, the estimation of the PSD by averaging over an ensemble of observations is not possible. The periodogram $|\mathcal{N}(\lambda,\mu)|^2$ [Schuster 1898] is a commonly utilized non-parametric simple estimator for the noise PSD $\Phi_{nn}(\mu)$ resulting in a so called short-term PSD estimate $\widehat{\Phi}_{nn}(\lambda,\mu) = |\mathcal{N}(\lambda,\mu)|^2$. Since the periodogram in contrast to the PSD is calculated from a finite segment length, e. g., a signal frame $n_\lambda(\kappa)$, it is suitable for block processing which is demanded in the context of real-time signal

enhancement. However, the difference between $\widehat{\Phi}_{nn}(\lambda,\mu)$ and $\Phi_{nn}(\mu)$ is caused by the finite frame length, which is in the order of 20 ms in speech enhancement, used for the calculation of the periodogram. Hence, the finite frame length and the random nature of most noise signals cause the obtained short-term periodograms of consecutive signal frames to vary randomly around the true average spectrum, i. e., the PSD $\Phi_{nn}(\mu)$. In order to reduce the variance of the short-term noise PSD estimate $\widehat{\Phi}_{nn}(\lambda,\mu) = |\mathcal{N}(\lambda,\mu)|^2$ possibly adaptive temporal smoothing is applied to the periodograms, which yields a refined short-term noise PSD estimate

$$\begin{aligned}\widehat{\Phi}_{nn}(\lambda,\mu) &= \tilde{\mathbb{E}}_{\alpha}\left\{\mathcal{N}(\lambda,\mu)\mathcal{N}^*(\lambda,\mu)\right\} \\ &= \alpha_{\Phi}(\lambda,\mu)\cdot\widehat{\Phi}_{nn}(\lambda-1,\mu) + (1-\alpha_{\Phi}(\lambda,\mu))\cdot|\mathcal{N}(\lambda,\mu)|^2,\end{aligned} \tag{3.45}$$

as close approximation of the true noise PSD $\Phi_{nn}(\mu)$.

With regard to more realistic noise scenarios, the requirement for the noise signal is relaxed allowing both, stationary or slowly varying short-term stationary noise which is denoted by $\overline{\Phi}_{nn}(\lambda,\mu)$ (where $\lambda$ is the frame index). Then, the true noise PSD $\Phi_{nn}(\mu)$ would be a sub-optimal noise power estimate with respect to an arbitrary signal frame $\lambda$, since it does not provide temporal information. Assuming a short-term stationary noise signal, the desired short-term PSD estimate $\widehat{\Phi}_{nn}(\lambda,\mu)$ is an approximation of the short-term noise PSD $\overline{\Phi}_{nn}(\lambda,\mu)$ which is defined as the average over all $K$ signal frames centered around the current frame $\lambda$ which are considered as stationary:

$$\overline{\Phi}_{nn}(\lambda,\mu) = \overline{\mathbb{E}}_K\left\{\mathcal{N}(\tilde{\lambda},\mu)\cdot\mathcal{N}^*(\tilde{\lambda},\mu)\right\}, \quad \text{with } \tilde{\lambda} = \lambda - \left\lceil\frac{K}{2}\right\rceil. \tag{3.46}$$

In general, the smoothing parameter $\alpha_{\Phi}(\lambda,\mu)$ is set such that the resulting smoothed estimate is also short-term stationary. The choice of the smoothing parameter states the trade off between estimation delay and noise power over- and under-estimation.

Another interpretation of the smoothing procedure is to decompose the noise signal into a baseline $\mathcal{B}(\lambda,\mu)$, which is equivalent to the short-term noise PSD $\overline{\Phi}_{nn}(\lambda,\mu)$ and the remaining fast temporal fluctuations $\mathcal{F}(\lambda,\mu)$ according to

$$\mathcal{B}(\lambda,\mu) = \overline{\Phi}_{nn}(\lambda,\mu) \approx \widehat{\Phi}_{nn}(\lambda,\mu) \tag{3.47}$$

$$|\mathcal{N}(\lambda,\mu)|^2 = \mathcal{B}(\lambda,\mu) + \mathcal{F}(\lambda,\mu). \tag{3.48}$$

In practice, $\mathcal{B}(\lambda,\mu)$ can be approximated by the above mentioned smoothed version of the noise signal power $\widehat{\Phi}_{nn}(\lambda,\mu)$ described by Eq. (3.45). In Fig. 3.5 an example of pure noise $|\mathcal{N}(\lambda,\mu)|^2$ its baseline $\mathcal{B}(\lambda,\mu)$ and the remaining fast variations $\mathcal{F}(\lambda,\mu)$ is depicted as a function of time for a single frequency bin $\mu = 101$ ($f_{\mathrm{s}} = 16\,\mathrm{kHz}$, $N_{\mathrm{DFT}} = 512$).

A common aim of statistical noise PSD estimators is the determination of the smoothed short-term noise PSD estimate $\widehat{\Phi}_{nn}(\lambda,\mu)$ given only the noisy observation $\mathcal{Y}(\lambda,\mu) = \mathcal{S}(\lambda,\mu) + \mathcal{N}(\lambda,\mu)$. In the context of slowly varying noise, the presented baseline $\mathcal{B}(\lambda,\mu)$ as a function of time and frequency provides the desired smoothed

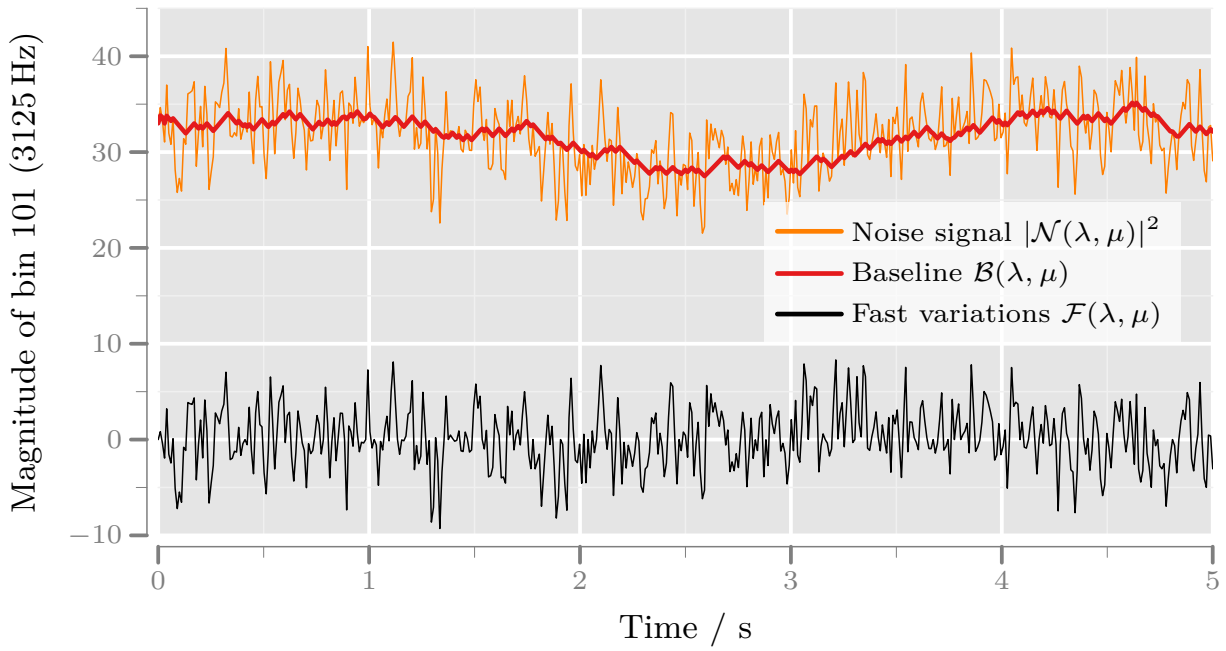



**Figure 3.5:** Definition of the baseline. The orange line represents the magnitude of pure noise periodogram for a single frequency bin $\mu = 101$ over frames $\lambda$. The noise can be decomposed into its slow varying baseline $\mathcal{B}(\lambda,\mu)$ and the remaining fast variations $\mathcal{F}(\lambda,\mu)$.

short-term noise PSD estimate. In order to obtain the baseline $\mathcal{B}(\lambda,\mu)$, the signal model is now given by

$$\mathcal{Y}(\lambda,\mu) = |\mathcal{S}(\lambda,\mu)| \cdot e^{\mathrm{i}\vartheta_{\mathcal{S}}(\lambda,\mu)} + \sqrt{\mathcal{B}(\lambda,\mu) + \mathcal{F}(\lambda,\mu)} \cdot e^{\mathrm{i}\vartheta_{\mathcal{N}}(\lambda,\mu)}, \tag{3.49}$$

where $\vartheta_{\mathcal{S}}$ and $\vartheta_{\mathcal{N}}$ denote the phase of speech and noise, respectively. Hence, the short-term noise PSD estimation problem is expressed by the periodogram of the noisy observation. For the sake of brevity the time-frequency index $(\lambda,\mu)$ is omitted in the following equation, yielding the noisy periodogram

$$|\mathcal{Y}|^2 = \underbrace{|\mathcal{S}|^2 + 2\,|\mathcal{S}|\,\sqrt{\mathcal{B}+\mathcal{F}}\cos(\vartheta_{\mathcal{S}} - \vartheta_{\mathcal{N}}) + \mathcal{F}}_{\text{removed by smoothing and post-processing}} + \mathcal{B}. \tag{3.50}$$

Since speech is modeled as highly non-stationary and $\mathcal{F}(\lambda,\mu)$ is by definition the rapidly changing proportion of the noise signal, the baseline $\mathcal{B}(\lambda,\mu)$ of the underlying noise signal can be obtained by smoothing and post-processing of $|\mathcal{Y}(\lambda,\mu)|^2$.

In the presence of speech, a short-term noise PSD estimate can either be obtained during speech pauses or by continuously tracking the magnitude in the short-term Fourier domain and apply further speech aware post-processing. Usually, statistical noise PSD estimators such as [Doblinger 1995; Gerkmann & Hendriks 2011; Martin 2006] apply temporal nonlinear and recursive smoothing of either the noisy input periodogram or the noise estimate itself in order to compute the noise signal baseline

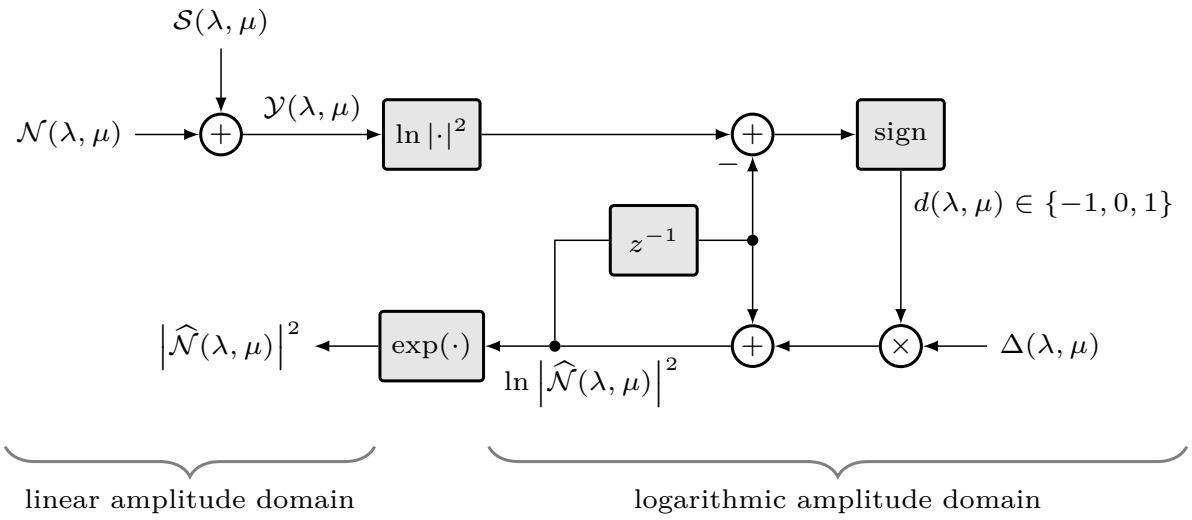



**Figure 3.6:** Equivalent block diagram of proposed noise estimator

$\mathcal{B}(\lambda,\mu)$. An alternative approach for the determination of baseline $\mathcal{B}(\lambda,\mu)$ is to limit the magnitude change between successive time frames by a fixed or adaptive step size. This concept will be explained in the next sections.

### 3.5.3 Concept of Baseline Tracing

The noise estimation problem is formulated in the logarithmic amplitude domain, while the actual implementation is carried out with linear amplitudes. This procedure is beneficial for the following reasons:

- the linear domain processing is computationally less complex than in the logarithmic domain,
- the logarithmic domain estimator is inherently unbiased, as shown below, and does not need correction terms like, e. g., *Minimum Statistics* [Martin 2001, 2006],
- the logarithmic domain formulation of the proposed estimator does not need explicit amplitude normalization,
- the logarithmic domain corresponds to the perception of the human auditory system.

The equivalent logarithmic domain block diagram of the proposed short-term noise PSD estimator is depicted in Fig. 3.6. The estimator can be explained in terms of delta modulation with an adaptive step size $\Delta(\lambda,\mu)$. For each fixed frequency bin $\mu$, the variable step size $\Delta(\lambda,\mu)$ is deliberately adjusted such that the estimate $\ln\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2$ follows the baseline of the logarithmic noisy sub-band, which is called *Baseline Tracing*. The noise estimate as depicted in Fig. 3.6 is obtained in the

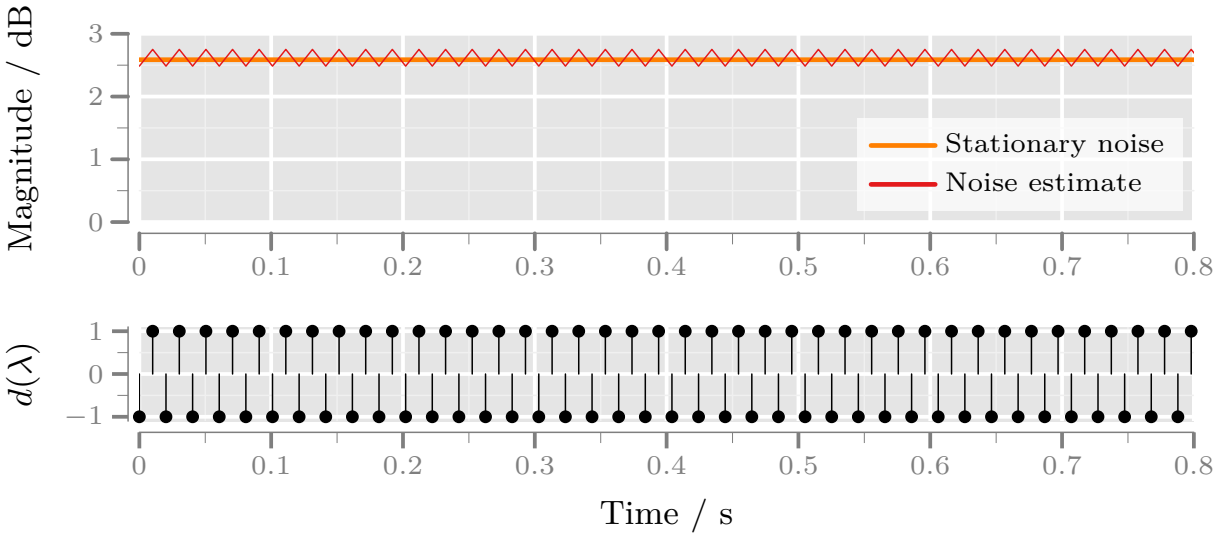



**Figure 3.7:** Insights into *Baseline Tracing* of a stationary noise component for one specific frequency bin over frames. The upper part depicts the stationary noise component and its estimate while in the lower part the corresponding signum series $d(\lambda)$ is plotted.

logarithmic domain by

$$\ln\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2 = \ln\left|\widehat{\mathcal{N}}(\lambda-1,\mu)\right|^2 + \Delta(\lambda,\mu)\cdot d(\lambda,\mu) \tag{3.51}$$

$$d(\lambda,\mu) = \operatorname{sign}\left(\ln|\mathcal{Y}(\lambda,\mu)|^2 - \ln\left|\widehat{\mathcal{N}}(\lambda-1,\mu)\right|^2\right), \tag{3.52}$$

where $\operatorname{sign}(\cdot)$ denotes the signum function.

In a first order delta modulator, the input signal is traced by an estimate which increases or decreases with a linear slope, which is determined by the step size $\Delta$ and the sign of the error between the input and the estimate. By adaptive control of the step size, the delta modulator is operated here in the slope overload mode [Jayant & Noll 1984] such that the estimate follows the baseline, which is determined by the short-term noise PSD $\overline{\Phi}_{nn}(\lambda,\mu)$. Due to the additive noise, the magnitudes of the speech component frequently decay to the level of the noise component. This is also exploited by *SPP* [Gerkmann & Hendriks 2011] and *Minimum Statistics* [Martin 2001, 2006].

In the upper part of Fig. 3.7 a stationary noise component is depicted as a function of time for a fixed frequency bin as bold orange curve (——). The derived noise estimate employing $\Delta(\lambda,\mu) = 0.1\,\text{dB}$ is indicated in red (——). The corresponding signum series $d(\lambda)$ is plotted in the lower part of the figure. By means of this stationary noise component it can be seen, that the signum series $d(\lambda) \in \{-1, 0, 1\}$ alternates with each time step $\lambda$ and is zero mean on average, i. e., $\mathbb{E}\{d(\lambda)\} = 0$. Thus, the proposed estimator is unbiased in the logarithmic

domain except of the granular noise known from delta modulation. In contrast to delta modulation $d(\lambda) = 0$ is allowed, which is favorable as the noise estimation may exactly match the, e. g., constant input.

For complexity reasons, the logarithmic short-term noise PSD estimator is implemented in the linear amplitude domain. The resulting equations (3.54) and (3.55) are in parts similar to [Baasch et al. 2014]. However, the adaptation mechanism proposed in this thesis is speech dependent and the control is effective in the logarithmic amplitude domain.

Given a noise estimate $\left|\widehat{\mathcal{N}}(\lambda-1,\mu)\right|^2$ from the last frame, the current estimate $\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2$ is calculated by stretching or compressing the last estimate with the tracing factor $\beta(\mu)$ in each frequency bin. The tracing factor $\beta$ is equivalent to

$$\beta(\lambda,\mu) = e^{\Delta(\lambda,\mu)}, \tag{3.53}$$

and can be realized frequency dependent or independent. A further option is to use a time varying $\beta(\lambda,\mu)$ in analogy to the adaptive step size control in delta modulation [Jayant & Noll 1984; Proakis & Salehi 2001]. As criterion for stretching or compressing, the signum function is used. If the difference between the current noisy observation $\mathcal{Y}(\lambda,\mu)$ and the last estimate $\widehat{\mathcal{N}}(\lambda-1,\mu)$ is greater than zero, $\widehat{\mathcal{N}}(\lambda-1,\mu)$ will be stretched by $\beta$ and compressed by $1/\beta$ in the other case. The estimation step, which is equivalent to the "Delta Modulation Algorithm" in the logarithmic amplitude domain of Fig. 3.6, is described by the following equations in the linear amplitude

$$\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2 = \left|\widehat{\mathcal{N}}(\lambda-1,\mu)\right|^2 \cdot \beta(\lambda,\mu)^{D(\lambda,\mu)}, \tag{3.54}$$

$$D(\lambda,\mu) = \operatorname{sign}\left(\ln|\mathcal{Y}(\lambda,\mu)|^2 - \ln\left|\widehat{\mathcal{N}}(\lambda-1,\mu)\right|^2\right), \tag{3.55}$$

$$= \operatorname{sign}\left(|\mathcal{Y}(\lambda,\mu)|^2 - \left|\widehat{\mathcal{N}}(\lambda-1,\mu)\right|^2\right), \tag{3.56}$$

with the initialization of the first estimate $\left|\widehat{\mathcal{N}}(1,\mu)\right|^2 = |\mathcal{Y}(1,\mu)|^2$.

A proof of concept example for a single frequency bin $\mu = 59$ corresponding to a frequency of 1816 Hz is depicted in Fig. 3.8 as a function of time. Here, a noisy signal consisting of factory1 noise [Varga et al. 1992] and a female speaker randomly taken from the NTT database [NTT-Corporation 1994] at 5 dB input SNR was processed with a frequency independent tracing factor

$$\beta(\lambda,\mu) = 1.05 \approx 0.4\,\text{dB}, \tag{3.57}$$

which corresponds to approximately 5 % change in $\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2$ from frame to frame in this example. In the upper plot the clean speech (——) and noise signal (——)

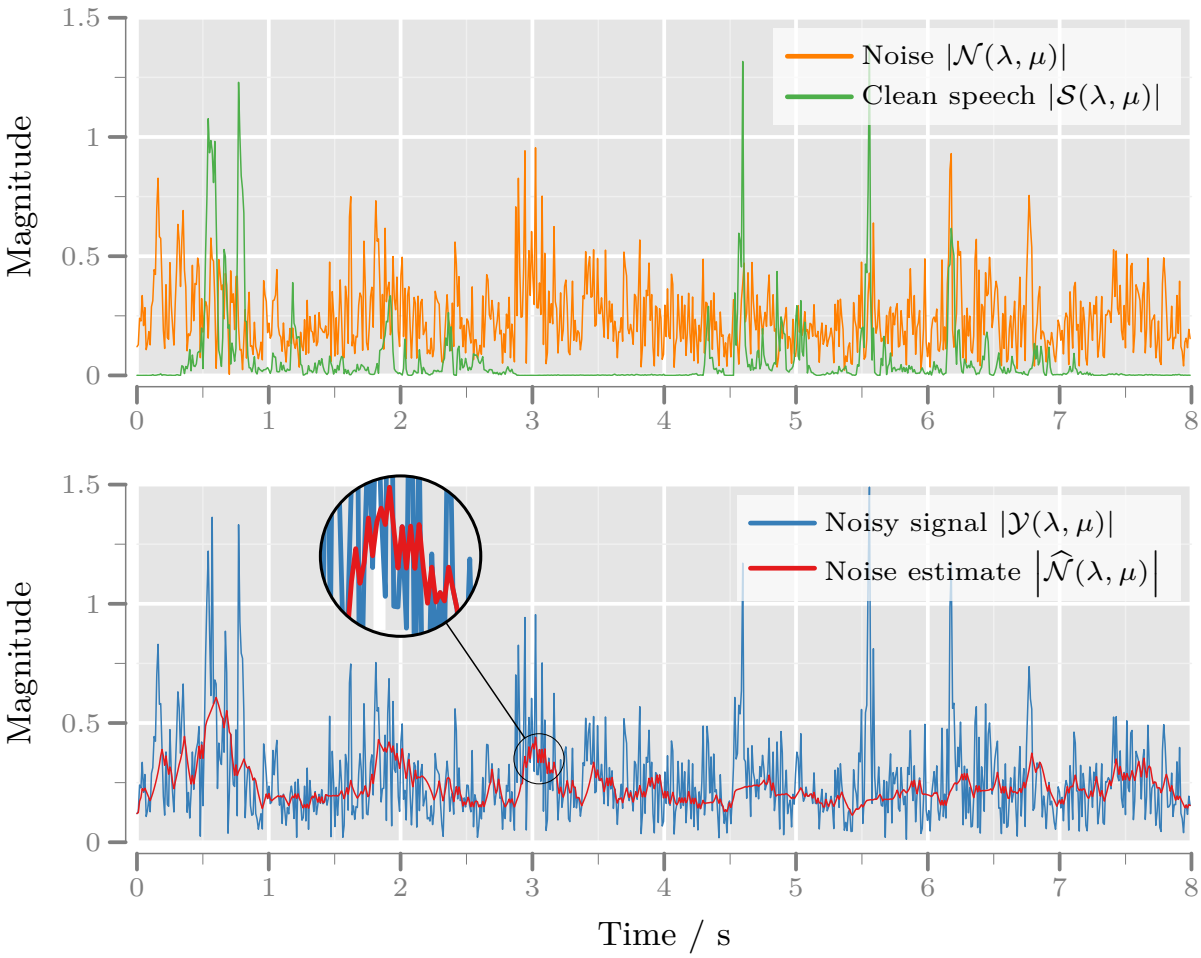



**Figure 3.8:** *Baseline Tracing* proof of concept example. As reference, the upper plot presents the magnitude of pure noise and clean speech depicted over frames $\lambda$ and for a single frequency bin $\mu = 59$ (1816 Hz, $f_s = 16$ kHz). The corresponding noisy mixture and the noise estimate are shown in the lower plot.

can be seen, while in the lower plot the noisy mixture (——) and the short-term noise PSD estimate (——) are depicted. It is visible that the simple concept of the new estimator is able to track the short-term noise PSD.

### 3.5.4 Tracing Factor $\beta$

Although the choice of $\beta = 1.05$ in the previous example (Fig. 3.8) works properly, it seems reasonable to define a frequency and time dependent scaling factor $\beta$ yielding

$$\beta(\lambda,\mu) = e^{\Delta(\lambda,\mu)} = 1 + \alpha(\lambda)\phi(\mu), \tag{3.58}$$

where $\alpha$ represents the time and $\phi$ the frequency dependent component. Since compression or stretching is realized by multiplication and division, $\beta(\lambda,\mu)$ has to be greater than one. In addition, the choice of the tracing factor $\beta(\lambda,\mu)$ depends

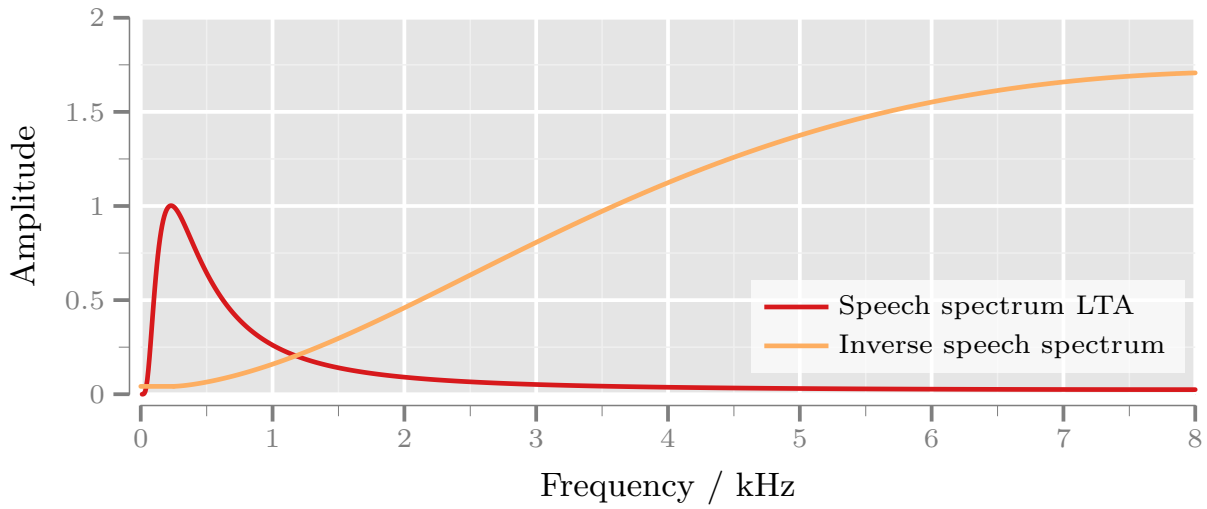



**Figure 3.9:** Long-term speech spectrum $\mathrm{LTA}(f)$ in the linear amplitude representation, normalized for clarity to a maximum of one and its inverse.

on the dynamic of the noise signal relative to the speech component. Facing a high noise signal dynamic, $\beta(\lambda, \mu)$ should be high allowing fast noise tracking and vice versa.

### Speech Dependent Scaling $\phi(\mu)$ over Frequency

If $\beta$ is too large, $\left|\widehat{\mathcal{N}}(\lambda, \mu)\right|^2$ follows unintentionally also the speech signal and the noise PSD estimate thus contains parts of speech. In order to prevent that speech contributes to the noise PSD estimate, the tracking speed for speech relevant frequencies is decreased while allowing faster tracking at the remaining frequencies.

Therefore, $\alpha(\lambda)$ is chosen proportional to the inverse of the *long-term speech spectrum average* (LTA) as shown in Fig. 3.9 with the definition of the LTA [ITU-T Recommendation P.50 1999] given by

$$\begin{aligned}\mathrm{LTA}(f)|_{\mathrm{dB}} = &-376.44 + 465.439 \log_{10}(f) \\ &- 157.745 \log_{10}^2(f) + 16.7124 \log_{10}^3(f),\end{aligned} \tag{3.59}$$

where $f$ is the frequency in Hz. A piece-wise approximation of the inverse long-term speech spectrum average $\mathrm{LTA}^{-1}(\mu)$ is introduced

$$\mathrm{LTA}^{-1}(\mu) = \begin{cases} \left(10^{\mathrm{LTA}\left(\frac{f_\mathrm{s}}{N_\mathrm{DFT}}\mu\,\mathrm{Hz}\right)/20}\right)^{-1} & \text{if } \frac{f_\mathrm{s}}{N_\mathrm{DFT}}\mu \geq 230\,\mathrm{Hz} \\ \left(10^{\mathrm{LTA}(230\,\mathrm{Hz})/20}\right)^{-1} & \text{if } \frac{f_\mathrm{s}}{N_\mathrm{DFT}}\mu < 230\,\mathrm{Hz}, \end{cases} \tag{3.60}$$

which ensures a smooth transition at low frequencies. In the next step, the new

speech aware and frequency dependent $\phi(\mu)$ is specified as:

$$\phi(\mu) = \frac{N_{\mathrm{DFT}} \cdot \mathrm{LTA}^{-1}(\mu)}{\sum_{i=0}^{N_{\mathrm{DFT}}-1} \mathrm{LTA}^{-1}(i)}. \tag{3.61}$$

Note $\phi(\mu)$ is thus normalized to a mean of one. Both, the normalized long-term speech spectrum (——) and its normalized inverse $\phi(\mu)$ (——) are depicted in Fig. 3.9.

### Fixed Scaling Parameter $\alpha$ with the Time

As mentioned above, a large $\beta$ leads to an erroneous noise PSD estimate including also speech. As $\phi(\mu)$ is one on average, $\beta(\lambda, \mu)$ may be too large in many cases and $\left|\widehat{\mathcal{N}}(\lambda, \mu)\right|^2$ changes excessively in successive frames, which can be solved by an appropriate choice of $\alpha(\lambda)$. According to Fig. 3.9 the main part of speech energy is distributed up to approx. 3.4 kHz. Allowing a change of $p$ % on average every 10 ms at this frequency range yields to a fixed $\alpha(\lambda)$ of:

$$\alpha(\lambda) = \frac{p \cdot L_A \left( \left\lfloor \frac{3.4\,\mathrm{kHz} \cdot N_{\mathrm{DFT}}}{f_{\mathrm{s}}} \right\rfloor + 1 \right)}{f_{\mathrm{s}} \cdot \sum_{i=0}^{\left\lfloor \frac{3.4\,\mathrm{kHz} \cdot N_{\mathrm{DFT}}}{f_{\mathrm{s}}} \right\rfloor} \phi(i)}, \tag{3.62}$$

where $L_A$ is the frame advance in samples. Setting $p$, e. g., to 5 % as in the presented example in Fig. 3.8 yields $\alpha(\lambda) \approx 0.13 \,\hat{\approx}\, 0.4\,\mathrm{dB}/10\,\mathrm{ms}$.

### Adaptive Scaling $\alpha(\lambda)$ with the Time

A further option is an adaptive $\alpha(\lambda)$ as a function of the frame *a posteriori* SNR. If the *a posteriori* SNR is extremely high, the adaptive $\alpha(\lambda)$ should be very small, resulting in small changes of $\left|\widehat{\mathcal{N}}(\lambda, \mu)\right|^2$ in successive frames. Whereas with decreasing SNR, $\alpha(\lambda)$ should grow, allowing a faster tracking of the noise. In order to prevent error propagation, the adaptive $\alpha(\lambda)$ is chosen as a function of the segmental mean SNR with an upper limit of $\gamma_{max}$ defined as

$$\overline{\gamma}_{\mathrm{seg}}(\lambda) = \min \left( \frac{1}{N_{\mathrm{DFT}}} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \frac{|\mathcal{Y}(\lambda-1, \mu)|^2}{\left|\widehat{\mathcal{N}}(\lambda-1, \mu)\right|^2},\ \gamma_{max} \right), \tag{3.63}$$

controlled by a second independent *a posteriori* SNR estimate

$$\overline{\gamma}_{\mathrm{2nd}}(\lambda) = \frac{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1} |\mathcal{Y}(\lambda, \mu)|^2}{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left|\widehat{\mathcal{N}}_{\mathrm{2nd}}(\lambda, \mu)\right|^2}, \tag{3.64}$$

where $\widehat{\mathcal{N}}_{\mathrm{2nd}}(\lambda, \mu)$ is provided by a second *Baseline Tracer* with a large fixed $\alpha_{\mathrm{2nd}}$ according to Eq. (3.62), resulting in a fast but rough noise tracking. The reason

| Parameter | Settings |
|---|---|
| Sampling frequency $f_\mathrm{s}$ | 16 kHz |
| Frame length $L_F$ | 320 ($\hat{=}$ 20 ms) |
| FFT length $N_\mathrm{DFT}$ | 512 (including zero-padding) |
| Frame overlap | 50 % ($L_A = 160 \,\hat{=}\, 20$ ms) |
| Window function | $\sqrt{\text{Hann}}-$ window |
| SNR estimation | Decision-directed approach |

**Table 3.1:** Simulation system settings

behind $\gamma_\mathrm{2nd}$ is to reduce the tracking speed in case of sudden increase of the speech component. Combining both SNR estimates, the adaptive $\alpha(\lambda)$ is now specified as

$$\alpha(\lambda) = \frac{1 - \overline{\gamma}_\mathrm{seg}(\lambda)/\gamma_\mathrm{max}}{\overline{\gamma}_\mathrm{2nd}(\lambda)}, \tag{3.65}$$

where the denominator provides fast and robust scaling of $\alpha(\lambda)$ which is refined by the nominator and $\gamma_{max}$ defines the upper limit for noise tracking.

### 3.5.5 Evaluation

The evaluation is carried out in three steps. At first the noise estimation performance itself is rated. Afterwards the new estimator is applied to a clean speech signal as boundary experiment for infinity input SNR. In a third step the new estimator is evaluated embedded in a standard noise reduction system. Different objective speech enhancement scores serve as indirect performance measures. In the following, a standard speech enhancement system which is depicted in Fig. 3.4 serves as benchmark platform. The corresponding simulation parameters are summarized in Tab. 3.1.

The proposed noise PSD estimator *Baseline Tracing* is compared in two different configurations for $\beta(\lambda, \mu)$ with three state-of-the-art methods: *Minimum Tracking* [Doblinger 1995], *Minimum Statistics* [Martin 2006] and the *SPP* noise tracker [Gerkmann & Hendriks 2011].

The first configuration of the new baseline tracing algorithm employs a frequency dependent $\phi(\mu)$ according to the inverse long-term speech average spectrum (Sec. 3.5.4) and a fixed $\alpha(\lambda) = 0.4\,\mathrm{dB}/10\,\mathrm{ms}$, while in the second configuration $\alpha(\lambda)$ is *a posteriori* SNR dependent (Sec. 3.5.4) with $\gamma_{max} \hat{=} 15\,\mathrm{dB}$ and $\alpha_\mathrm{2nd} = 1.6\,\mathrm{dB}/10\,\mathrm{ms}$. The parameters of the *Minimum Tracking*, *Minimum Statistics* and *SPP* algorithm are chosen as suggested in [Doblinger 1995; Martin 2006; Gerkmann & Hendriks 2011], respectively.

The comparison is performed for all permutations of the following parameters:

- the input SNR varies from $-10$ to 35 dB in 5 dB steps[5] and

[5]The mixing procedure is detailed in Sec. C.1. Note that for the calculation of the scaling

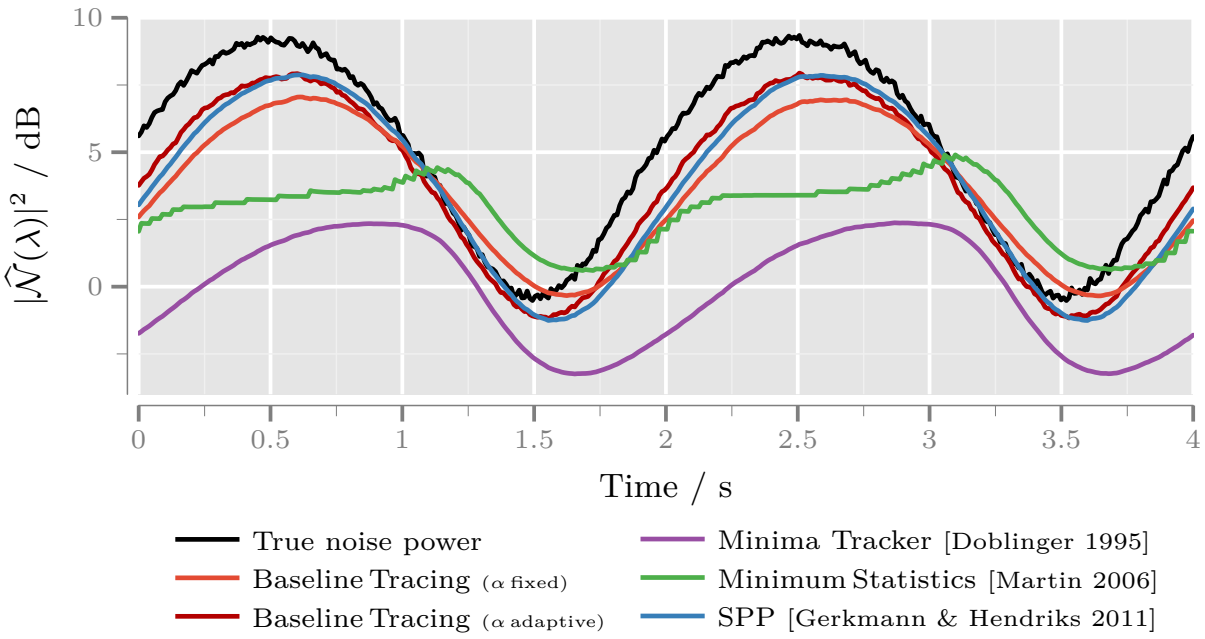



**Figure 3.10:** Noise power estimations for pure modulated white Gaussian noise (no speech present). The presented results are averaged across frequency and over 15 modulation periods.

- 15 male and 15 female english speakers (randomly taken from the NTT database) are mixed with
- seven different stationary and non-stationary noise types (f16, factory1, babble, buccaneer1 [Varga et al. 1992], modulated Gaussian noise, vacuum cleaner, passing cars).

The Gaussian noise is modulated with $f_{\text{mod}} = 0.5\,\text{Hz}$ according to

$$f(k) = 1 + \frac{1}{2}\sin\left(2\pi k \frac{f_{\text{mod}}}{f_{\text{s}}}\right). \tag{3.66}$$

### Noise PSD Estimation Performance

The evaluation is carried out by the logarithmic noise PSD estimation errors: $\text{LogErr}^{\text{Over}}$, $\text{LogErr}_{\text{Under}}$, and LogErr which are related among each other according to

$$\text{LogErr} = \text{LogErr}^{\text{Over}} + \text{LogErr}_{\text{Under}}. \tag{3.67}$$

Lower values of the respective measure indicate a better performance. The errors describe the mismatch between the estimated and the real noise short-term PSD and are defined in Appendix C.3. In applications such as speech enhancement an

factor to adjust the input SNR only speech and noise signal sections with speech presence are considered.

| **Estimator** | LogErr | $\text{LogErr}_{Under}$ | $\text{LogErr}^{Over}$ |
|---|---|---|---|
| Baseline Tracing (adaptive $\alpha$) | 2.94 dB | 1.19 dB | 1.76 dB |
| Baseline Tracing (fixed $\alpha$) | 4.11 dB | 1.65 dB | 2.46 dB |
| SPP | 3.93 dB | 1.45 dB | 2.48 dB |
| Minimum Statistics | 5.09 dB | 2.55 dB | 2.54 dB |
| Minima Tracker | 5.35 dB | 4.00 dB | 1.35 dB |

**Table 3.2:** Noise PSD logarithmic error measures for different short-term noise PSD estimators. The input signal consists exclusively of modulated white Gaussian noise.

overestimation of the true noise power, as indicated by $\text{LogErr}^{\text{Over}}$, likely results in an attenuation of the speech and thus in speech distortions. On the other hand, a noise power underestimation, pointed out by the $\text{LogErr}^{\text{Under}}$ causes probable lower noise attenuation.

In a first boundary experiment the noise estimators are analyzed applying them to a non-stationary noise signal, i. e., without the influence of speech. The synthetically composed noise sequence consists of 62 s of a modulated white Gaussian noise signal ($f_{\text{mod}} = 0.5\,\text{Hz}$, Eq. (3.66)). The first 2 seconds, i. e., the first modulation period, of the noise estimate results are discarded due to initialization operations. The remaining 60 s are subdivided into 15 periods of 4 seconds length and averaged. In addition the results are also averaged across frequency in order to reduce the variance and to provide a compact representation. The short-term noise PSD estimates as indicated by the colored curves and the true noise power marked by the black color of averaged noise signal periods are depicted in Fig. 3.10. It is obvious that all noise estimation algorithms are not able to follow closely the true noise power (——) in this example. The *Minima Tracker* (——) consequently underestimates the true noise power. In contrast, the *Minimum Statistics* method (——) underestimates only the rising edge of the noise signal and is able to follow more closely the falling edge. This behavior can be expected due to the length of the sliding minimum window of approx. 1.5 s. The remaining noise estimators perform similar. Analyzing the *Baseline Tracing* estimator utilizing a fixed $\alpha$ (——) it is apparent that it has a slightly worse performance and a tendency to overestimate the noise power at the falling edges of the noise signal compared with the *SPP* estimator. If the *Baseline Tracing* is using the adaptive $\alpha(\lambda)$ (——), the algorithm is able to follow the true noise power more precisely than the *SPP* approach. The presented results are confirmed by the objective LogErr measures applied to the whole noise signal sequence and are summarized in Tab. 3.2.

In the following, the noise estimation performance is evaluated in a more realistic scenario under the influence of speech, different noise types and various SNRs as described above. The averaged results in terms of the LogErr are presented in Fig. 3.11. As in the previous example the *Minimum Tracker* (——) marks the

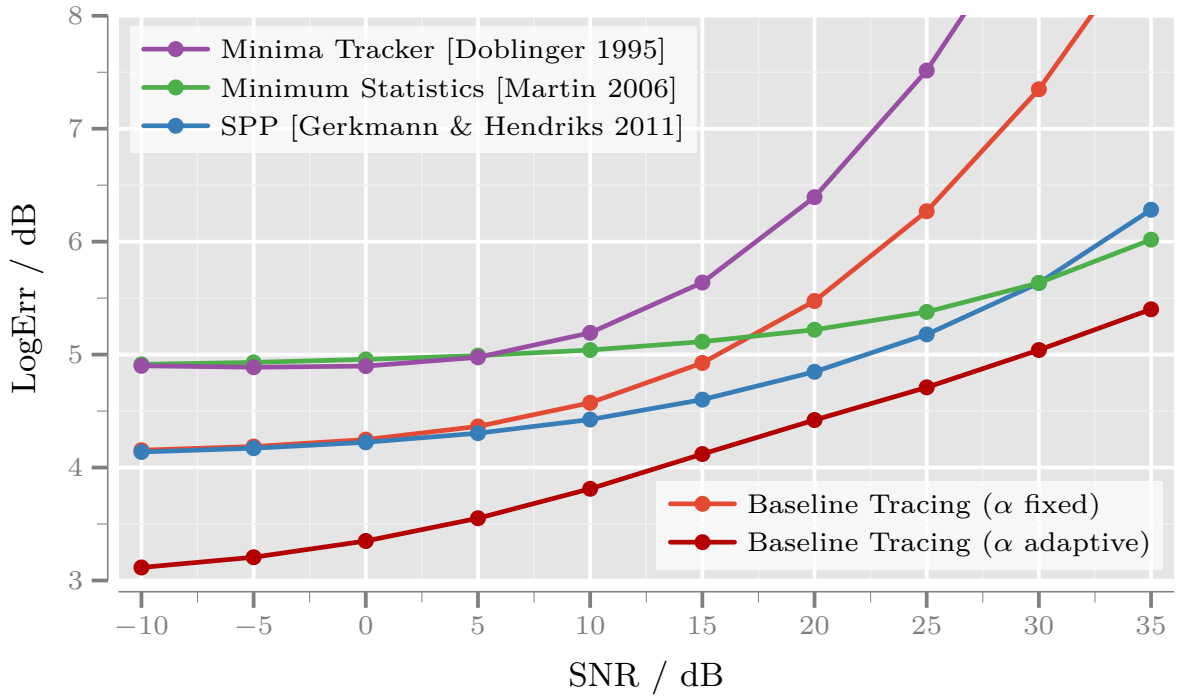



**Figure 3.11:** Logarithmic error measure of different short-term noise PSD estimators. The error is depicted over the input SNR and averaged over 30 speakers and 7 noise types.

lower bound of the noise estimation performance over the complete SNR range, especially for SNR values greater than 15 dB. In contrast, the *Minimum Statistics* algorithm (—●—) performs almost constant over the complete SNR range and provides scores around 5 dB LogErr. In low SNR conditions the *SPP* algorithm (—●—) is able to provide significantly better scores compared to *Minimum Statistics*. However, with increasing SNR the distance becomes smaller. Up to 10 dB SNR the proposed *Baseline Tracing* with fixed $\alpha$ (—●—) performs similar to SPP. Beyond, a performance loss is visible. This can be explained due to the fixed operating point of the algorithm, i. e., the frequency dependent but fixed magnitude change in successive frames of the short-term noise PSD estimate. With increasing SNR the magnitude change is significantly larger than the dynamic of the true noise floor, which leads to an alternating under- and over- estimation of the noise. The proposed *Baseline Tracing* utilizing the adaptive $\alpha(\lambda)$ (—●—) performs best over the complete SNR range with a distinct distance to all competitors. By incorporating the frame *a posteriori* SNR, it is possible to adapt the magnitude change in successive frames to match the dynamic of the underlying noise signal. Hence, the alternating under- and over- estimation of the noise floor is significantly reduced for high SNR values.

In Fig. 3.12 the averaged results are summarized for selected noise types at various SNRs. Comparing the proposed *Baseline Tracing* with fixed $\alpha$ (▮) to the best state-of-the-art algorithm, i. e., *SPP* (▮), the performance is quite similar for all noise types and SNR conditions, except for babble noise at 10 and 15 dB, where *SPP* performs slightly better. Also the relation between under and over

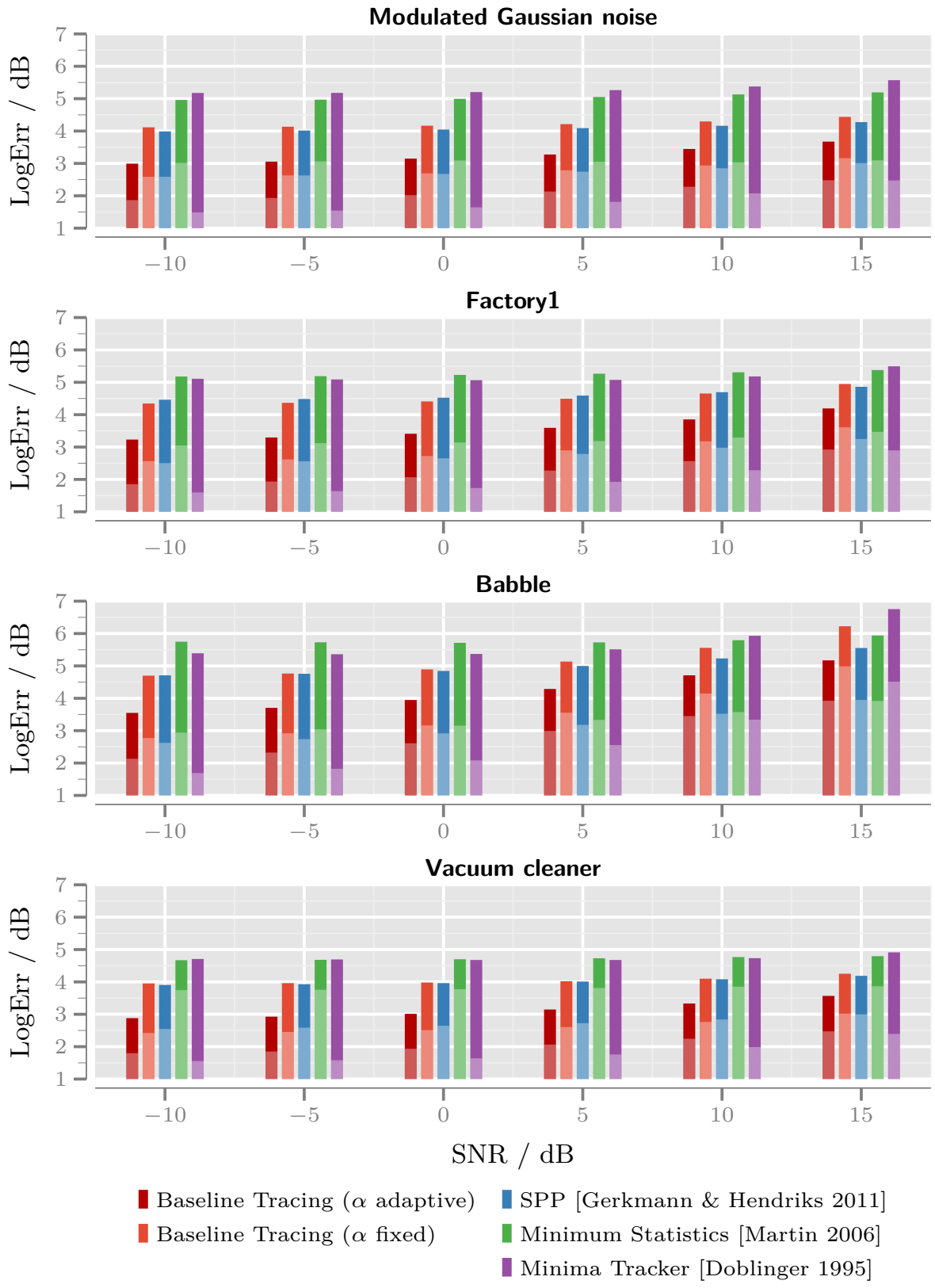



**Figure 3.12:** Logarithmic error measure averaged over 30 speakers taken from the NTT database at various SNRs for selected noise types. The lower part of the bars indicate the $\mathrm{LogErr}^{\mathrm{Over}}$, while the upper part represents $\mathrm{LogErr}_{\mathrm{Under}}$. The total height corresponds to LogErr.

estimation is similar. The *Minimum Statistics* (▮) and *Minimum Tracking* (▮) have a comparable performance regarding the total LogErr measure and perform 0.59 dB worse on average compared to *SPP* and the proposed estimator with fixed $\alpha$. In contrast to *Minimum Statistics*, the LogErr analysis of *Minimum Tracking* confirmed a dominant underestimation of the short-term noise PSD, indicating lower performance in terms of noise reduction. For all noises and SNR conditions, the proposed estimator *Baseline Tracing* with adaptive $\alpha(\lambda)$ (▮) holds the best performance in all error measures with an advance up to 1.1 dB and 0.71 dB on average.

### Short-term Noise PSD Estimation on a Clean Speech Signal

In this experiment the noise estimators are applied to the randomly chosen clean speech signals, i. e., without noise. This reflects on the one hand the border case of infinite input SNR. On the other hand, the tendency of the respective estimator to estimate erroneously speech as noise can be studied. If speech contributes to the short-term noise PSD estimate, speech distortions in terms of speech attenuation during the noise reduction process will likely occur. In addition, the performance of the mentioned information combining, detailed in Sec. 5, of different noise and speech short-term PSD estimates is significantly degraded if one of the noise estimates contains speech.

The noise short-term PSD estimates of the four best approaches are depicted in terms of spectrograms in Fig. 3.13 exemplarily for one speech signal and confirm that the *Minimum Statistics* algorithm and the proposed *Baseline Tracing* noise estimator in both configurations for $\alpha(\lambda)$ deliver an almost perfect noise estimate, i. e., $\widehat{\Phi}_{nn}(\lambda, \mu)$ is very close to zero. However, the *SPP* approach shows isolated significant contributions in the noise estimate. The LogErr measures normalized to the best noise estimator, i. e., *Baseline Tracing* (adaptive), and averaged over all 30 speakers result in

- *Baseline Tracing* (adaptive): 0 dB ΔLogErr,
- *Minimum Statistics*: 2.22 dB ΔLogErr,
- *Baseline Tracing* (fixed): 3.93 dB ΔLogErr,
- *SPP* 6.37 dB ΔLogErr, and
- *Minimum Tracker*: 15.33 dB ΔLogErr.

Note that for the calculation of the LogErr the noise floor reference is set to -80 dB which corresponds to the most silent part of the clean speech signals.

### Noise Reduction Performance

In addition, the performance of the different noise estimators is also rated in terms of objective speech enhancement scores. Therefore, the noise estimators are utilized

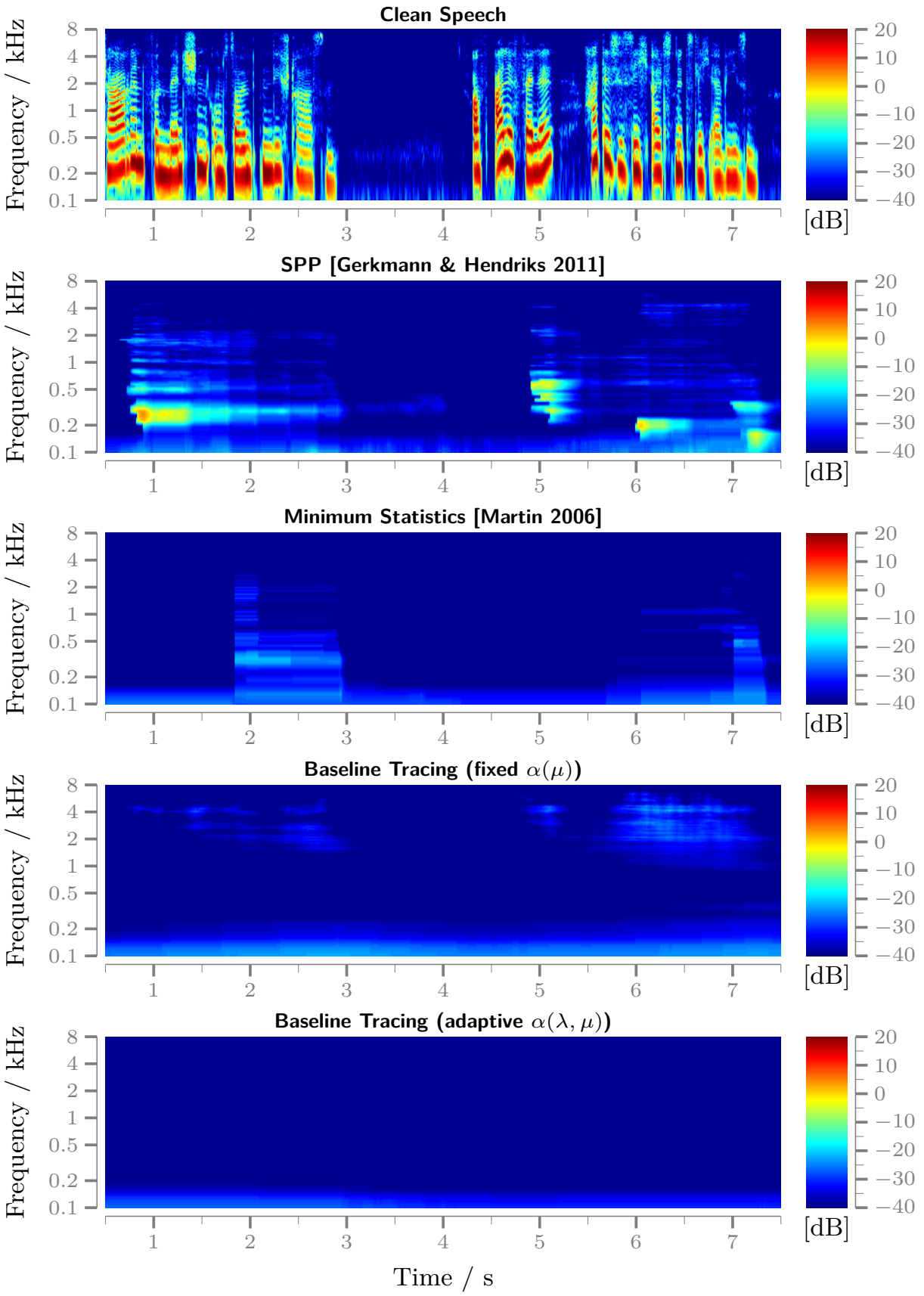



**Figure 3.13:** Noise short-term PSD estimates from clean speech depicted as spectrograms.

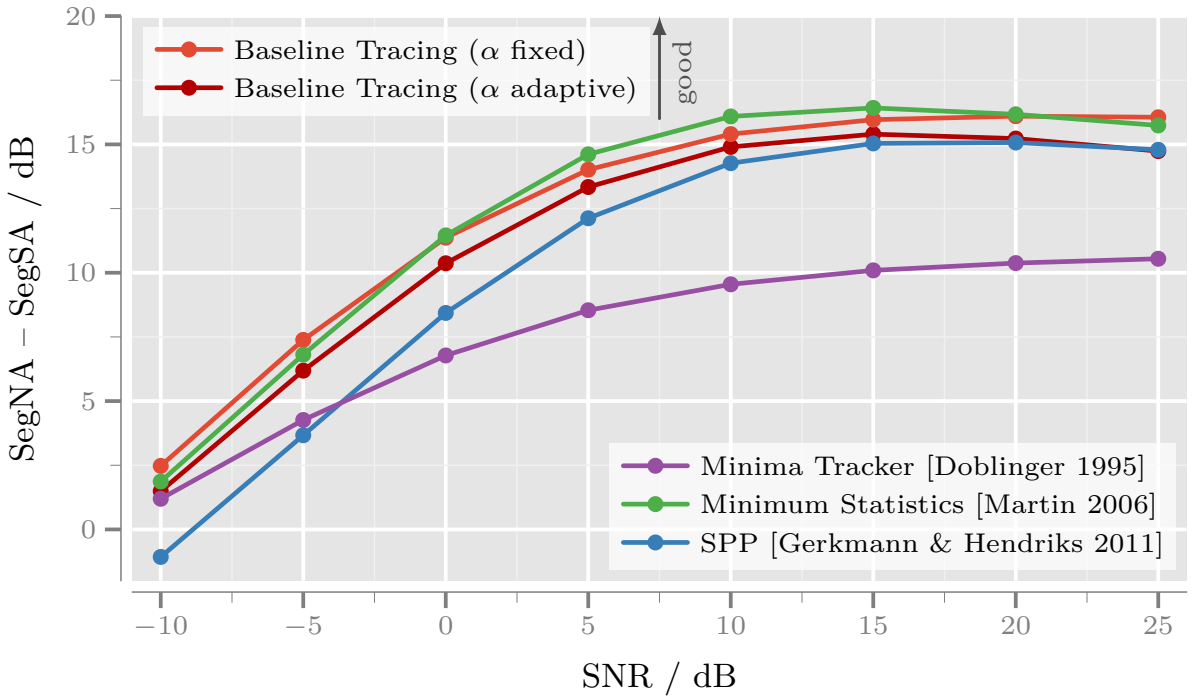



**Figure 3.14:** Difference between SegNA and SegSA over the input SNR.

in a standard speech enhancement system as depicted in Sec. 3.4 by Fig. 3.4. The estimates of the *a priori* SNR and *a posteriori* SNR are provided by the decision-directed approach [Ephraim & Malah 1984]. For the spectral gains, the Wiener filter is utilized which depends on the *a priori* SNR estimate. The enhanced time domain signal $\hat{s}(k)$ is obtained by applying an IDFT, windowing using a square-root Hann window and overlap-add.

The employed objective scores[6] are the *segmental speech attenuation* (SegSA), *segmental noise attenuation* (SegNA) as well as the *cepstral distance* (CD) and the *perceptual evaluation of speech quality* (PESQ). Regarding the CD, lower values indicate a lower speech distortion. The difference between SegNA and SegSA corresponds to the noise reduction performance. Hence, larger values indicate better performance. The PESQ measure provides an objective measure of the perceived audio quality that predicts the results of a subjective listening test. PESQ compares the original clean speech signal $s(k)$ with the enhanced speech signal $\hat{s}(k) = \tilde{y}(k)$. The resulting PESQ values are analogous to the *mean-opinion score* (MOS) and range from one (bad) to 4.5 (no distortion).

Fig. 3.14 shows the results for the SegNA – SegSA scores. As indicated in the previous section by the consequent noise underestimation, the *Minimum Tracking* (—●—) achieves the lowest SegNA – SegSA performance over the input SNR. Since the noise is underestimated significantly, the resulting speech distortion should be low, which is confirmed by the CD measure up to 10 dB depicted in Fig. 3.15. While the *Minimum Statistics* (—●—) and the proposed system with fixed (—●—) and adaptive $\alpha$ (—●—) perform in the SegNA – SegSA measure similar over the

[6] The objective scores are described in Appendix C.2

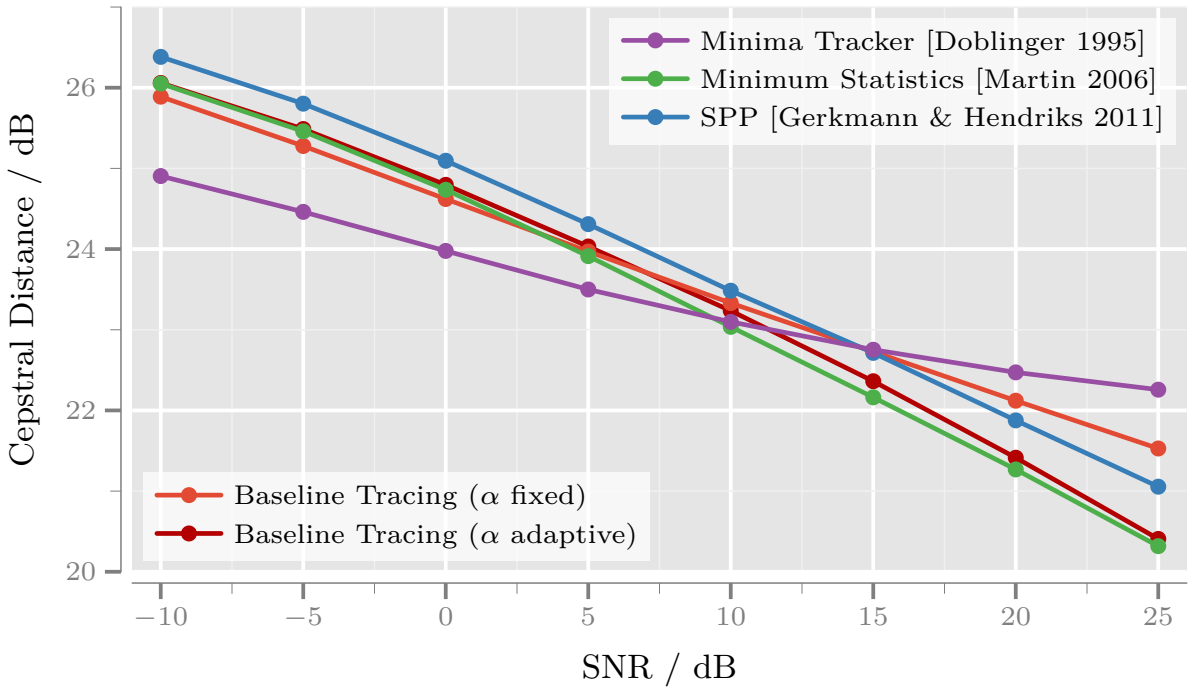



**Figure 3.15:** The *cepstral distance* (CD) is depicted over the input SNR.

complete input SNR, the *SPP* method (—●—) has a lower score of approx. 3.5 dB at −10 dB SNR reaching a similar performance starting with 10 dB SNR. Except the *Minimum Tracking* for high SNR, the *SPP* has a slightly higher CD over the input SNR, where the proposed estimator with adaptive $\alpha$ and *Minimum Statistics* perform similar with the best scores on average. Up to 10 dB SNR, the *Baseline Tracing* with fixed $\alpha$ performs also likewise.

The PESQ measure is presented in Fig. 3.16 and reflects the previously observed objective scores. Again, the lower bound of the performance is defined by the *Minimum Tracker* algorithm (—●—). *Minimum Statistics* (—●—) and *SPP* (—●—) exhibit very similar PESQ scores over the complete input SNR range. Concerning the *Baseline tracing* method with fixed $\alpha$ (—●—) a turning point is visible. Up to 15 dB the performance is slightly better compared to *Minimum Statistics* and *SPP*. Again this behavior is explainable with the fixed $\alpha$ which causes a significant alternating over- and under-estimation of the true noise power starting with 15–20 dB input SNR. Utilizing the *Baseline tracing* algorithm with adaptive $\alpha$ (—●—) the low cepstral distance as well as the property not to estimate speech as noise is verified by the best PESQ scores over the complete input SNR. This confirms the superior LogErr performance also in the noise reduction task for both new *Baseline Tracing* estimators, as they provide a high noise attenuation at simultaneously low speech distortion.

### Computational complexity

The *Baseline Tracing* with fixed $\beta$ consists only of four arithmetic operations per frame and frequency bin: one (complex) magnitude operation, one difference

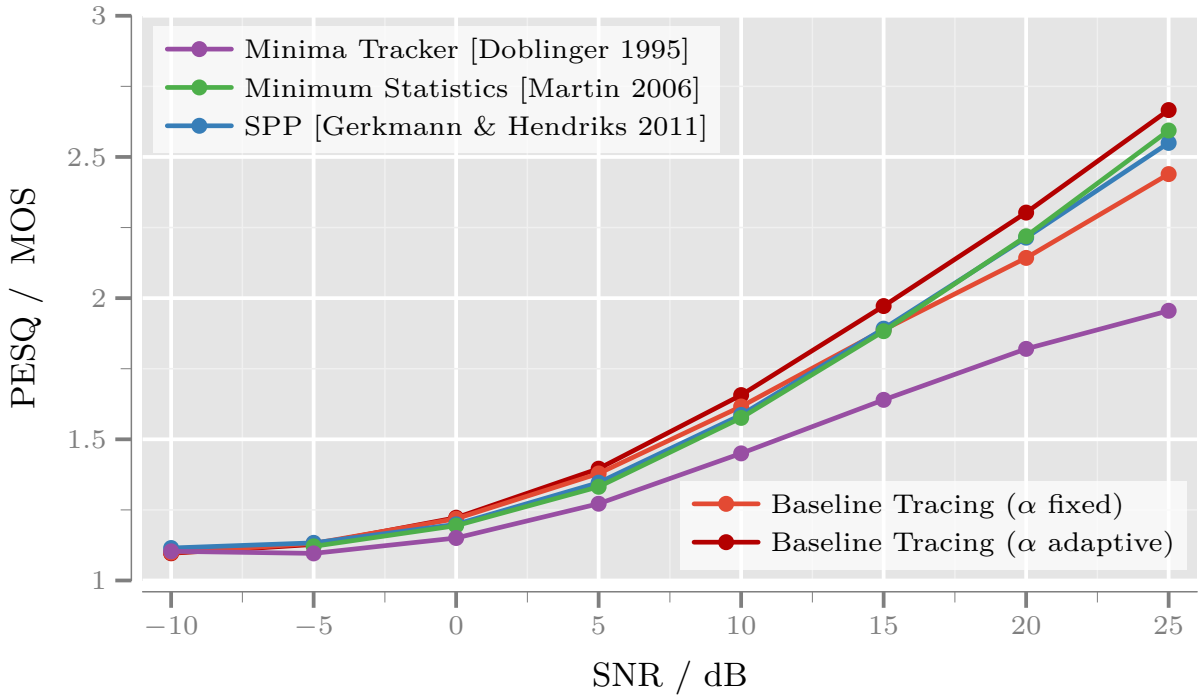



**Figure 3.16:** PESQ over the input SNR.

operation, an "if else" statement, and an multiplication or division operation. It only relies on the last short-term noise PSD estimate and the scaling factor $\beta$, resulting in low memory consumption and low computational complexity. The memory consumption consists of two times $N_{\mathrm{DFT}}/2 + 1$ values belonging to the last short-term noise PSD estimate and the frequency dependent $\beta(\lambda)$. Using the adaptive *Baseline Tracing*, the complexity extends by a second non-adaptive *Baseline Tracing*, the calculation of two frame *a posteriori* SNR values ($N_{\mathrm{DFT}}/2 + 1$ multiplications and summations, two times $N_{\mathrm{DFT}}/2+1$ summations, and one division operation) and the final calculation of $\beta(\lambda, \mu)$ according to Eq. (3.58) and (3.65) (one difference, two division, and one multiplication), which is still low.

## 3.6 Noise Reduction by Information Combining Exploiting Spectral Dependencies

Most of the speech enhancement systems proposed in literature have been derived for narrow band signals (50 Hz – 4 kHz) using certain assumptions about the statistics of the speech and noise signals, e. g., [Ephraim & Malah 1984, 1985; Gerkmann & Hendriks 2012; Lim & Oppenheim 1979; Lotter & Vary 2005]. In case of noise reduction for wideband signals (50 Hz – 7 kHz), a common method is to double the sampling rate and the transform length and to apply the low band algorithms also for the wider frequency range. Thereby, neither the unequal spectral energy distribution of speech and noise signals nor the properties of the human auditory system are considered. In Fig. 3.17 the relative cumulative speech energy distribution over the

| Noise type | Average ratio of low band SNR to high band SNR | |
|---|---|---|
| | male speakers | female speakers |
| Cockpit | +15.39 dB | +13.98 dB |
| Babble | +0.55 dB | −0.86 dB |
| Factory | +12.55 dB | +11.14 dB |
| Buccaneer | +15.64 dB | +14.23 dB |
| White Gaussian noise | +26.81 dB | +25.39 dB |

**Table 3.3:** SNR deviation of the low band from the high band for different noise types. For the measurement, six speech signals (three male and three female speakers) obtained from the NTT database [NTT-Corporation 1994] are used. The noise signals have been taken from the NOISEX-92 database [Varga et al. 1992].

frequency is depicted for the LTA defined in [ITU-T Recommendation P.50 1999]. In addition, the relative cumulative speech energy distribution is measured for a random set of speakers taken from the [NTT-Corporation 1994] database. It is notable, that 99 % of the speech energy is located in the base-band frequency range between 50 Hz and 3.4 kHz and only 1 % of the energy belongs to the remaining higher frequency range.

Mostly the energy of speech signals decays stronger than the energy of noise signals beyond 3 kHz. Hence, the SNR in the low band is higher than in the high band. Table 3.3 shows quantitative examples of how much the SNR in the low band is better than in the high band, exemplarily for different speakers and different noise environments.

In most cases the SNR significantly degrades in the high band which leads to imprecise SNR estimation and thus fluctuating weighting gains. This results in

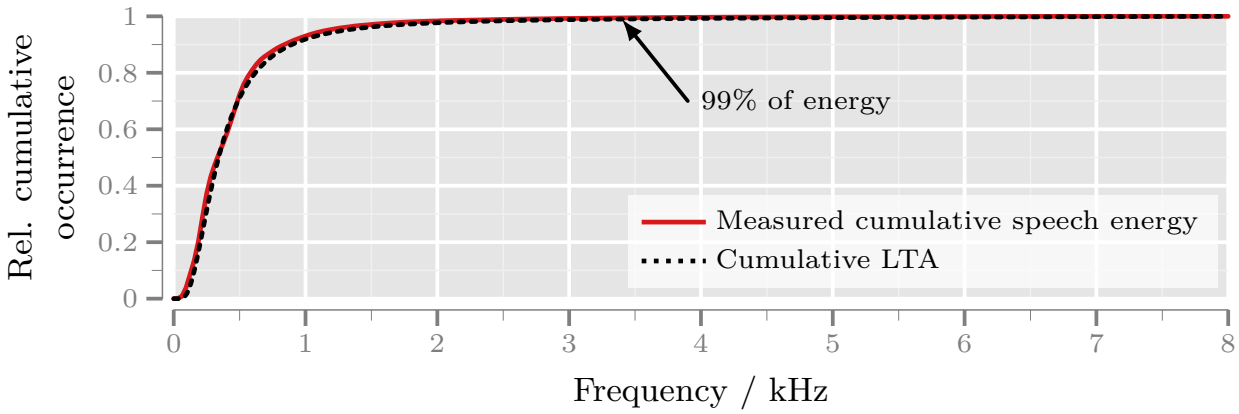



**Figure 3.17:** Cumulative speech energy distribution.

increasing occurrence of so called *musical tones*, especially at higher frequencies. So far, only a limited number of proposals have been made which take into account the aforementioned aspects when enhancing wideband speech signals, e. g., [Esch et al. 2010a; Heese et al. 2010] and [Beaugeant et al. 2006].

From another field of speech enhancement it is known, that the spectral dependencies of speech signals can be exploited to estimate missing high frequencies by analyzing the low band speech signal. This technique is called *artificial bandwidth extension* (BWE), e. g., [Geiser et al. 2007; Heese et al. 2012a; Jax & Vary 2006]. With respect to noise reduction, techniques from the BWE can be used to improve the estimation of the weighting gains in the high band.

Additionally to conventional calculated weighting gains, an intermediate enhanced low band signal is used to provide a second set of high band weighting gains utilizing techniques from BWE. The weighting gains are combined using an SNR dependent *information combining* approach.

### 3.6.1 Wideband Noise Reduction System Overview

To counteract the mentioned problems when it comes to wideband noise reduction, a joint noise reduction system [Esch et al. 2010a; Heese et al. 2010] is presented. It uses different noise reduction schemes for the low and high band and makes use of the spectral dependencies in speech signals similar to techniques known from BWE. In the following the sub-index "LB" indicates the low frequency band and "HB" the high band. The block diagram of the proposed system is depicted in Fig. 3.18.

The input signal $y(k)$ is decomposed into its low band $y_{\mathrm{LB}}$ and high band $y_{\mathrm{HB}}$ components applying a two-channel *infinite impulse response* (IIR) *quadrature mirror filter* (QMF) with critical sampling and near perfect reconstruction [Löllmann et al. 2009]. Subsequently, the filtered signals are down-sampled by a factor of 2, where $k'$ represents the discrete time index in the sub-sampled domain. Individual analysis – synthesis structures allow the re-use of existing low band noise reduction systems[7]. The noise reduction is carried out in the frequency domain by spectral weighting for both bands. For this purpose $y_{\mathrm{LB}}(k')$ and $y_{\mathrm{HB}}(k')$ are segmented into overlapping frames and transformed to the spectral domain as stated in Sec. 3.2.1. Thus, the spectral coefficients of the noisy input signal at frequency bin $\mu$ and frame $\lambda$ are given by:

$$\mathcal{Y}_{\mathrm{LB}}(\lambda,\mu) = \mathcal{S}_{\mathrm{LB}}(\lambda,\mu) + \mathcal{N}_{\mathrm{LB}}(\lambda,\mu), \tag{3.68}$$

$$\mathcal{Y}_{\mathrm{HB}}(\lambda,\mu) = \mathcal{S}_{\mathrm{HB}}(\lambda,\mu) + \mathcal{N}_{\mathrm{HB}}(\lambda,\mu), \tag{3.69}$$

where $\mathcal{S}_{\mathrm{LB}}(\lambda,\mu)$, $\mathcal{S}_{\mathrm{HB}}(\lambda,\mu)$ and $\mathcal{N}_{\mathrm{LB}}(\lambda,\mu)$, $\mathcal{N}_{\mathrm{HB}}(\lambda,\mu)$ represent the spectral coefficients of the speech and noise component of the low and high band, respectively. While a conventional noise suppression, operating in the frequency domain, is used in the low band (50 Hz – 4 kHz), a joint noise suppression approach is applied in the high band (4 kHz – 7 kHz). Using spectral features from the intermediate

[7] Note that the DFT length $N_{\mathrm{DFT}}$ is defined for each band after downsampling.

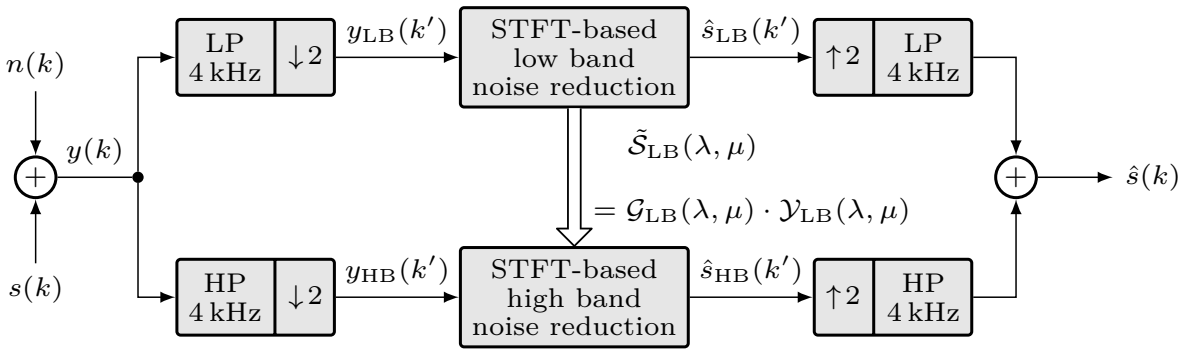



**Figure 3.18:** Wideband noise reduction system using different schemes in the low and high band exploiting the spectral dependencies of speech.

enhanced low band signal $\tilde{\mathcal{S}}_{\mathrm{LB}}(\lambda,\mu)$, the high band noise reduction is supported by techniques known from BWE.

In general, limiting the weighting gains to a lower bound $\max\{\mathcal{G}(\lambda,\mu), g_{\min}\}$ allows to control the tradeoff between noise attenuation and speech distortion. The favored tradeoff depends on the application. Since noise disturbs the application of BWE techniques, a stronger noise attenuation for the intermediate enhanced low band signal $\tilde{\mathcal{S}}_{\mathrm{LB}}(\lambda,\mu) = \mathcal{G}_{\mathrm{LB}}(\lambda,\mu) \cdot \mathcal{Y}_{\mathrm{LB}}(\lambda,\mu)$ using a small $g_{\min}$ is desirable. Whereas, a small amount of speech attenuation is favored in case of actual speech enhancement utilizing a higher $g_{\min}$.

Finally, both enhanced signals $\hat{s}_{\mathrm{LB}}(k')$ and $\hat{s}_{\mathrm{HB}}(k')$ are combined by a QMF synthesis in order to obtain the enhanced wideband signal $\hat{s}(k)$.

## 3.6.2 Joint Noise Reduction in the High Band

The principle of the combined high band noise reduction system is illustrated in Fig. 3.19. Since the analysis – synthesis framework remains the same as for the low band showed in Fig. 3.18, only the processing blocks in the spectral domain are depicted. Two separate noise reduction schemes are performed in parallel to the noisy high band spectrum $\mathcal{Y}_{\mathrm{HB}}(\lambda,\mu)$. The results are two gain estimates, conventional gains $\mathcal{G}_{\mathrm{conv}}$ and novel gains $\mathcal{G}_{\mathrm{bwe}}$ which exploit spectral dependencies between the low and high band. For the following reasons the frequency resolution of the weighting gains in the high band is decreased:

- The properties of the human auditory system are taken into account, i.e., the frequency selectivity decreases with higher frequencies [Zwicker 1982].
- Due to the aforementioned imprecise SNR estimation in the high band the resulting weighting gains exhibit a high variance over time and frequency, which results likely in *musical tones*. Decreasing the frequency resolution by combining neighboring frequency bins limits the temporal fluctuations of the weighting gains and reduces their variance over time. This yields a better suppression of *musical tones*.

- Since the estimation accuracy of the BWE is limited to the spectral envelope of the high band, the determination of $\mathcal{G}_{\mathrm{bwe}}$ is bounded inherently to subbands.

Hence, the frequency resolution is decreased from $N_{\mathrm{DFT}}$ to $N'_{\mathrm{DFT}}$ by combining adjacent frequency-bins using 50% overlapping Hann windows of the same lengths. The decimated frequency index is denoted by $\mu'$, where $N'_{\mathrm{DFT}} < N_{\mathrm{DFT}}$.

The conventional approach consists of noise power estimation, SNR estimation and the calculation of the weighting gain $\mathcal{G}_{\mathrm{conv}}(\mu)$ as described in Sec. 3.2. The subsequent post-processing decreases the frequency resolution as described above. The determination of the novel weighting gains $\mathcal{G}_{\mathrm{bwe}}(\mu')$ will be detailed in Section 3.6.2. The final weighting gain $\mathcal{G}_{\mathrm{HB}}(\mu')$ for the high band is obtained by adaptive combining the two independent weighting gains:

$$\mathcal{G}_{\mathrm{HB}}(\lambda,\mu') = \alpha_{\mathcal{G}}(\lambda,\mu')\mathcal{G}_{\mathrm{bwe}}(\lambda,\mu') + \left(1-\alpha_{\mathcal{G}}(\lambda,\mu')\right)\mathcal{G}_{\mathrm{conv}}(\lambda,\mu'), \tag{3.70}$$

where $\alpha_{\mathcal{G}} \in [0,1]$ represents a reliability factor which is frame and frequency dependent and will be explained later.

Finally, the frequency resolution of the resulting high band weighting gains $\mathcal{G}_{\mathrm{HB}}(\lambda,\mu')$ is interpolated to its original resolution from $N'_{\mathrm{DFT}}$ to $N_{\mathrm{DFT}}$ using overlap-add of 50% overlapping scaled Hann windows. Spectral weighting of the noisy high band coefficients according to

$$\widehat{\mathcal{S}}_{\mathrm{HB}}(\lambda,\mu) = \mathcal{Y}_{\mathrm{HB}}(\lambda,\mu)\cdot\mathcal{G}_{\mathrm{HB}}(\lambda,\mu) \tag{3.71}$$

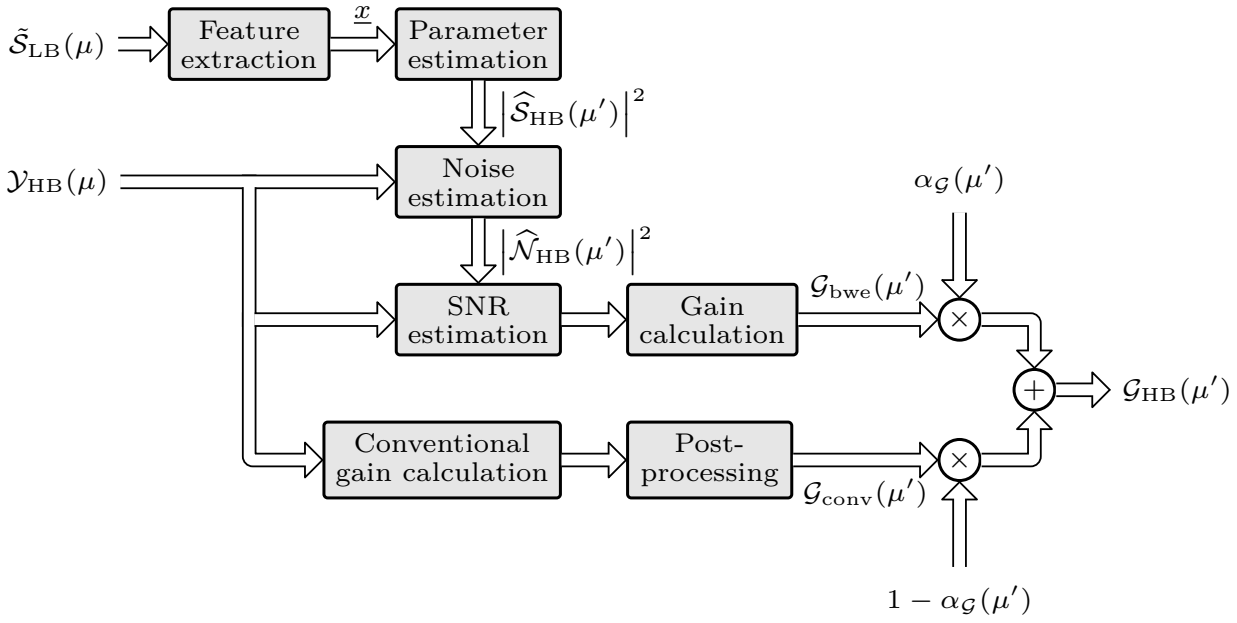



**Figure 3.19:** Highband noise reduction scheme exploiting spectral dependencies between low and high band (applied to each frame $\lambda$)

yields an estimate of the clean high band coefficients $\widehat{\mathcal{S}}_{\mathrm{HB}}(\lambda, \mu)$. The enhanced signal $\hat{s}(k')$ in the time domain is obtained by applying an inverse DFT and overlap-add.

### Noise Reduction Exploiting Spectral Dependencies

Statistical dependencies between the low band (50 Hz – 4 kHz) and the high band (4 kHz – 7 kHz) are exploited using techniques known from BWE. The method that is used here is partly included in [Geiser et al. 2007]. The concept is to estimate high band signal parameters based on meaningful features which are extracted only from the intermediate enhanced low band signal applying a trained *hidden markov model* (HMM).

Therefore, spectral features from the processed enhanced low band signal $\tilde{\mathcal{S}}_{\mathrm{LB}}(\mu)$ are calculated. Usually, representations of the spectral envelope of the low band signal serve as features and are extracted on a frame-by-frame basis [Jax & Vary 2004]. In the classical BWE application the *mel frequency cepstral coefficients* (MFCC) and the *zero-crossing rate* (ZCR) [Rabiner & Schafer 1978] have been proven as suitable features. In a first approach those features are chosen for the estimation of the clean speech high band parameters [Esch et al. 2010a]. Since the enhanced low band signal still contains noise, more appropriate features allow to improve the BWE estimation performance in this context. Hence, the feature vector $\underline{x}$ derived from the low band consists of $N_{\mathrm{C}}$ *relative spectral transform - perceptual linear prediction* (RASTA-PLP) coefficients and the ZCR [Rabiner & Schafer 1978] of the low band signal $\tilde{S}_{\mathrm{LB}}$. RASTA [Hermansky & Morgan 1994] is a technique that applies a filter in each frequency sub-band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel. The PLP [Hermansky 1990] algorithm preserves the important speech information by warping spectra to minimize the differences between speakers.

As mentioned before and in contrast to a classical BWE application where an undisturbed input signal is assumed and the HMM can be trained with clean speech, the processed enhanced low band signal, which serves here as input for the BWE, will still contain remaining background noise. This fact is taken into account and incorporated into the training process of the HMM. White Gaussian noise serves here as model for residual noise. Hence, the training data used to determine the low band features is disturbed by white Gaussian noise with an SNR of 0 dB to cope even with strongly impaired signals. Subsequently, a conventional noise reduction, e. g., Sec. 3.2, is applied using a strong noise suppression with $g_{\min}$ close to zero, e. g., -20 dB. Doing so, typical processing artifacts are integrated in the training process.

As in [Geiser et al. 2007], a trained HMM is used to estimate the parameter vector $\underline{v}$ representing $\mu'$ clean speech sub-band energies of the high band. Let $\underline{X} = \{\underline{x}(1), \ldots, \underline{x}(\lambda)\}$ denote a sequence of feature vectors starting with frame one to $\lambda$. The MMSE estimation of a parameter vector $\underline{v}$ of the current frame with

given observations $\underline{X}$ can be formulated as

$$\mathbb{E}\left\{||\underline{v}-\hat{\underline{v}}||^2|\underline{X}\right\} \stackrel{!}{=} \min, \tag{3.72}$$

where $\hat{\underline{v}}$ is the respective estimate. The solution to this optimization problem is the conditional expectation $\underline{v}_{\mathrm{MMSE}} = \mathbb{E}\{\underline{v}|\underline{X}\}$. Given a pre-computed codebook $\mathcal{C} = \{\hat{\underline{v}}_1, \ldots, \hat{\underline{v}}_{M_\mathrm{C}}\}$ for the vectors $\underline{v}$ this MMSE estimate can be approximated as

$$\hat{\underline{v}}_{\mathrm{MMSE}} = \sum_{\hat{\underline{v}}_i \in \mathcal{C}} \hat{\underline{v}}_i \cdot P(\hat{\underline{v}}_i|\underline{X}), \tag{3.73}$$

which is basically a weighted sum over the centroids of the codebook. The weights $P(\hat{\underline{v}}_i|\underline{X})$ comprise *a posteriori* probabilities which can be determined using HMM techniques [Geiser et al. 2007]. The codebook $\mathcal{C}$ is obtained creating a large amount of training vectors which are then used for the training of a *vector quantizer* (VQ). The result of the VQ training corresponds to the codebook. In this work the LBG algorithm [Linde et al. 1980] with the MMSE distance measure is employed.

Once the clean speech energies $\hat{\underline{v}}_{\mathrm{MMSE}} = \{\left|\hat{\mathcal{S}}_{\mathrm{HB}}(0)\right|^2, \cdots, \left|\hat{\mathcal{S}}_{\mathrm{HB}}(N'_{\mathrm{DFT}}-1)\right|^2\}$ of the $\mu'$ sub-bands have been estimated, they are together with the noisy observation used to estimate the noise power in the high band for each frame $\lambda$:

$$|\widehat{\mathcal{N}}_{\mathrm{HB}}(\mu')|^2 = \max\left(|\mathcal{Y}_{\mathrm{HB}}(\mu')|^2 - |\widehat{\mathcal{S}}_{\mathrm{HB}}(\mu')|^2, 0\right). \tag{3.74}$$

Finally, the *a posteriori* SNR $\gamma(\mu')$ and *a priori* SNR $\xi(\mu')$ can be estimated and expressed according to:

$$\hat{\gamma}_{\mathrm{HB}}(\mu') = \frac{|\mathcal{Y}_{\mathrm{HB}}(\mu')|^2}{|\widehat{\mathcal{N}}_{\mathrm{HB}}(\mu')|^2} \quad \text{and} \quad \hat{\xi}_{\mathrm{HB}}(\mu') = \frac{|\widehat{\mathcal{S}}_{\mathrm{HB}}(\mu')|^2}{|\widehat{\mathcal{N}}_{\mathrm{HB}}(\mu')|^2}. \tag{3.75}$$

Based on the SNR estimates a Wiener filter or any state-of-the-art weighting rule $\mathcal{G}_{\mathrm{bwe}}$ can be calculated.

### Information Combining by Cross-Fading

As mentioned before, the information of the two high band estimates, in terms of the weighting gains $\mathcal{G}_{\mathrm{conv}}(\lambda,\mu')$ and $\mathcal{G}_{\mathrm{bwe}}(\lambda,\mu')$, is adaptively combined using the cross-fading-factor $\alpha_{\mathcal{G}}(\lambda,\mu')$. Assuming optimal weighting gains $\mathcal{G}_{\mathrm{opt}}$, which are derived from the ideal *a posteriori* SNR $\gamma(\mu')$ and *a priori* SNR $\xi(\mu')$ also determined at the reduced frequency resolution by combining adjacent frequency bins as before

$$\gamma_{\mathrm{HB}}(\mu') = \frac{|\mathcal{Y}_{\mathrm{HB}}(\mu')|^2}{|\mathcal{N}_{\mathrm{HB}}(\mu')|^2} \quad \text{and} \quad \xi_{\mathrm{HB}}(\mu') = \frac{|\mathcal{S}_{\mathrm{HB}}(\mu')|^2}{|\mathcal{N}_{\mathrm{HB}}(\mu')|^2}, \tag{3.76}$$

the oracle cross-fading factor $\alpha_{\mathcal{G},\mathrm{oracle}}(\mu')$ can be formulated as

$$\alpha_{\mathcal{G},\mathrm{oracle}}(\mu') = \frac{\left(\mathcal{G}_{\mathrm{opt}}(\mu') - \mathcal{G}_{\mathrm{conv}}(\mu')\right)^2}{\left(\mathcal{G}_{\mathrm{opt}}(\mu') - \mathcal{G}_{\mathrm{conv}}(\mu')\right)^2 + \left(\mathcal{G}_{\mathrm{opt}}(\mu') - G_{\mathrm{bwe}}(\mu')\right)^2}, \tag{3.77}$$

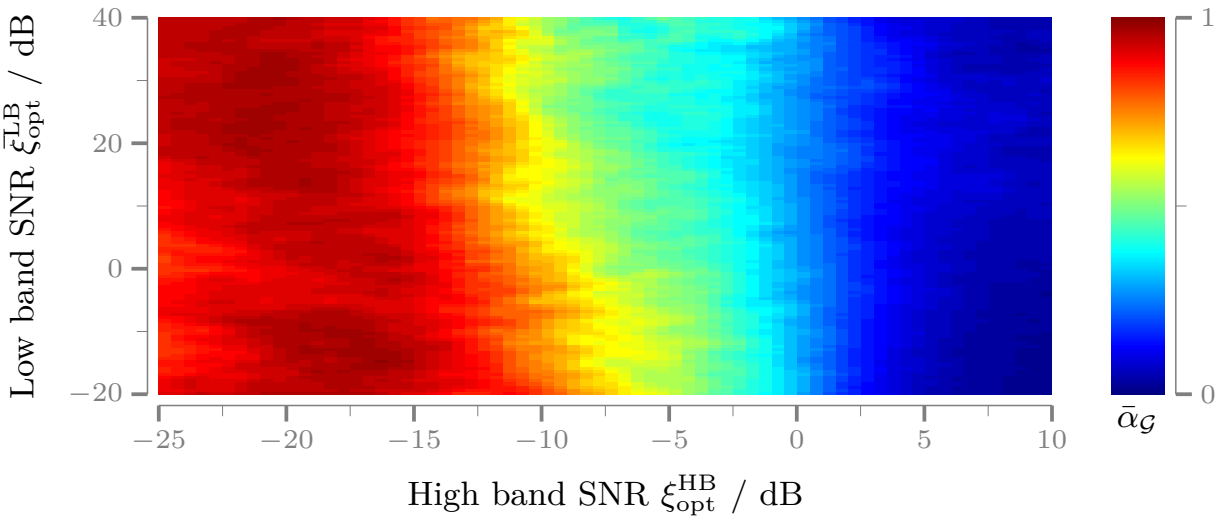



**Figure 3.20:** Example of a look-up table for the determination of $\bar{\alpha}_{\mathcal{G}}$ ($\mu' = 1$).

and is normalized to one. This oracle cross-fading factor minimizes the distance of $\mathcal{G}_{\text{conv}}$ to the optimal weighting gains $\mathcal{G}_{\text{opt}}$. If the conventional noise suppression technique performs better than the BWE approach, i. e., $(G_{\text{opt}} - G_{\text{conv}})^2 < (G_{\text{opt}} - G_{\text{bwe}})^2$, $\alpha_{\text{ref}}$ tends to smaller values leading to a stronger weighting of $G_{\text{conv}}$ and vice versa.

Since $\mathcal{G}_{\text{opt}}$ is not available in a realistic scenario the cross-fading factor of the weighting gains, which is a reliability indicator for $\mathcal{G}_{\text{conv}}$ and $\mathcal{G}_{\text{bwe}}$, has to be estimated from given quantities. Utilizing the averaged low band and the sub-band SNR of the respective high band, the cross-fading factor can be estimated which is realized here by means of a look-up table. In a training process, where all ideal quantities are available, $\alpha_{\mathcal{G},\text{oracle}}(\mu')$ is recorded for every frame $\lambda$ and every sub-band $\mu'$ together with the respective sub-band SNR $\xi_{\text{opt}}^{\text{HB}}(\mu')$ of the high band and the averaged SNR $\bar{\xi}_{\text{opt}}^{\text{LB}}$ of the low band

$$\bar{\xi}_{\text{opt}}^{\text{LB}} = \frac{1}{N_{\text{DFT}}} \sum_{\mu=0}^{N_{\text{DFT}}-1} \frac{|\mathcal{S}_{\text{LB}}(\mu)|^2}{|\mathcal{N}_{\text{LB}}(\mu)|^2}. \tag{3.78}$$

Based on the training data, a look-up table for the estimation of $\alpha_{\mathcal{G}}(\mu')$ is generated for every sub-band. Therefore, $\xi_{\text{opt}}^{\text{HB}}(\mu')$ and $\bar{\xi}_{\text{opt}}^{\text{LB}}$ are quantized (e. g., 1 dB step size) and the corresponding values for $\alpha_{\mathcal{G},\text{oracle}}(\mu')$ are averaged within the quantization levels. An example of a look-up table for sub-band $\mu' = 1$ is depicted in Fig. 3.20. At the end, the final look-up table provides one estimate $\bar{\alpha}_{\mathcal{G}}(\mu')$ for each quantized combination of $\xi_{\text{opt}}^{\text{HB}}(\mu')$ and $\bar{\xi}_{\text{opt}}^{\text{LB}}$. In a real application, $\xi_{\text{opt}}^{\text{HB}}$ and $\bar{\xi}_{\text{opt}}^{\text{LB}}$ are not available. Here, the respective SNR estimates of the conventional noise suppression techniques in the low band and high band are utilized to determine $\bar{\alpha}_{\mathcal{G}}(\mu')$ using the pre-trained look-up table for each sub-band $\mu'$.

| Parameter | Settings |
|---|---|
| Sampling frequency $f_\mathrm{s}$ | 16 kHz |
| Frame length $L_F$ | 160 ($\hat{=}$ 20 ms due to downsampling) |
| FFT length $N_\mathrm{DFT}$ | 256 (including zero-padding) |
| Frame overlap | 50 % (Hann window) |
| Input SNR | −10 dB ... 35 dB (step size: 5 dB) |
| Noise estimation | Minimum Statistics [Martin 2006] |
| SNR estimation | decision-directed (Sec. 3.4.2) |
| Number sub-bands $\mu'$ | 24 |
| Number RASTA-PLPs $N_\mathrm{C}$ | 13 |
| Codebook size $M_\mathrm{C}$ | 128 (training based on 1.5 h speech) |
| Gain limitation ($g_\mathrm{min}$ / $\tilde{g}_\mathrm{min}$) | (0.2857 / 0.01) |

**Table 3.4:** System settings.

### 3.6.3 Experimental Results

Any conventional noise reduction system can be applied for the low band and to estimate the conventional weighting gains $\mathcal{G}_\mathrm{conv}$ in the high band. Since the focus of the evaluation is on the joint noise reduction in the high band, the choice of the used conventional noise estimator plays a minor role. For better comparability with other conventional noise reduction systems the noise is estimated by *Minimum Statistics* [Martin 2006], the SNR is estimated by the *decision-directed* approach [Ephraim & Malah 1984] and the well-known Wiener filter [Lim & Oppenheim 1979] is utilized as weighting gain rule for $\mathcal{G}_\mathrm{conv}$ and $\mathcal{G}_\mathrm{bwe}$.

The proposed joint noise suppression technique employing different configurations is compared with the conventional case, where only the Wiener filter weighting gains are applied to both the low band and the high band. In the first configuration

- the features consists of MFCCs and the ZCR with the use of $\alpha_{\mathcal{G},\mathrm{oracle}}$ [Esch et al. 2010a]. The HMM training is based only on clean speech.
- The second configuration comprises RASTA-PLP and the ZCR as features with the use of $\alpha_{\mathcal{G},\mathrm{oracle}}$ and $\bar{\alpha}_\mathcal{G}$. In addition, the HMM is trained based on enhanced $\tilde{\mathcal{S}}_\mathrm{LB}$ speech which has been disturbed by additive white Gaussian noise in advance with an SNR of 0 dB and employing an aggressive weighting gain utilizing $\tilde{g}_\mathrm{min}$.

For the objective evaluation of the different noise reduction systems the simulation setup as described in Appendix C is utilized. The simulation parameters which are used for evaluation are listed in Tab. 3.4. The $N'_\mathrm{DFT}$ look-up tables which are required for the estimation of $\alpha_{\mathcal{G},\mathrm{oracle}}$ are generated based on 10 min of clean speech from the NTT database [NTT-Corporation 1994] disturbed by white Gaussian noise at different input SNR values varying from −10 dB to 35 dB in 5 dB steps. White Gaussian noise is utilized as background noise model for the

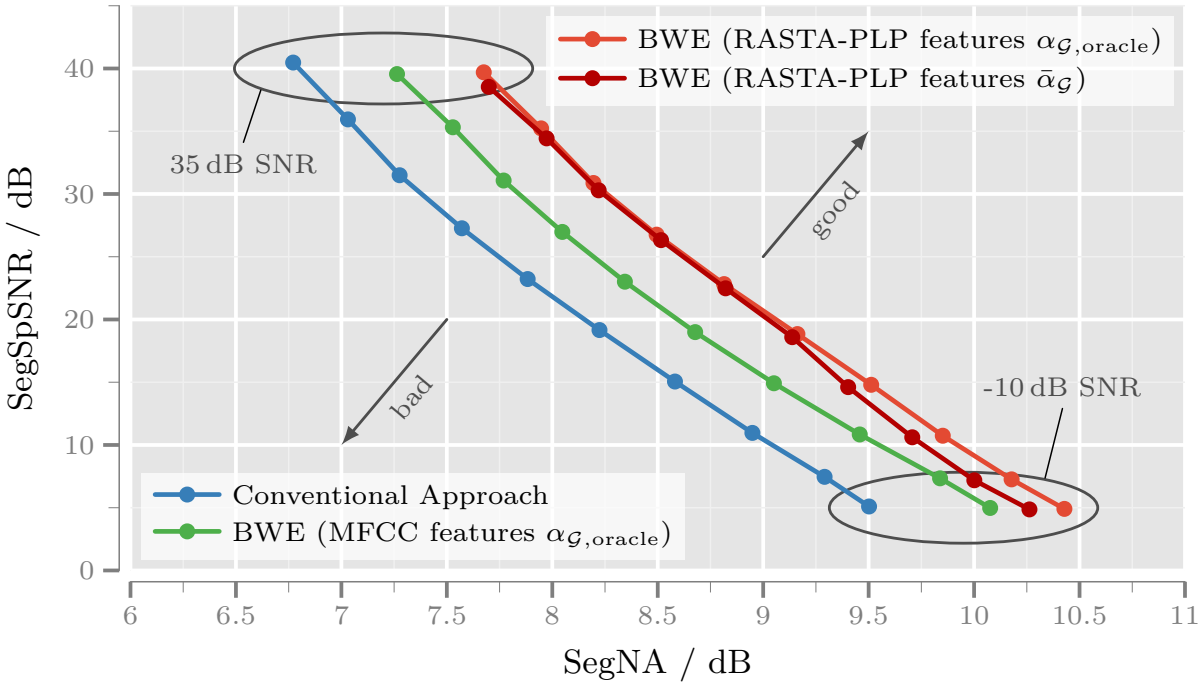



**Figure 3.21:** The *segmental speech SNR* (SegSpSNR) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter.

training of the cross-fading factor in order to prevent memorization or fitting to specific background noise types. For the comparison, seven different speech signals (four male, three female) from the NTT speech database are each degraded by four different noise types (cockpit, babble, factory1, buccaneer), taken from the NOISEX-92 database [Varga et al. 1992] at different input SNRs varying from −10 to 35 dB in 5 dB steps[8]. Note that the speech signals used for the evaluation are not included in the training data for the HMM and the look-up tables. The performance of the rated systems is evaluated in terms of the SegNA, the SegSA and the *segmental speech SNR* (SegSpSNR).

Figure 3.21 depicts the averaged results for SegSpSNR plotted over SegNA with the input SNR as variable parameter. Hence, a fair comparison with respect to the tradeoff noise attenuation versus speech distortion is possible. The points of best performance would be placed as much as possible in the upper right corner of the figure. The objective measurements (Fig. 3.21) show that the additional use of the artificial BWE in the high band (—●—, —●—, —●—) outperforms the results of conventional noise suppression techniques (—●—) consistently. In the employed simulation framework the oracle cross-fading factor $\alpha_{\mathcal{G},\text{oracle}}$ is available which allows the best combination of the weighting gains in the high-band. Hence, the influence of the used features and the improved HMM training can be investigated

[8]The mixing procedure is detailed in Sec. C.1. Note that for the calculation of the scaling factor to adjust the input SNR only speech and noise signal sections with speech presence are considered.

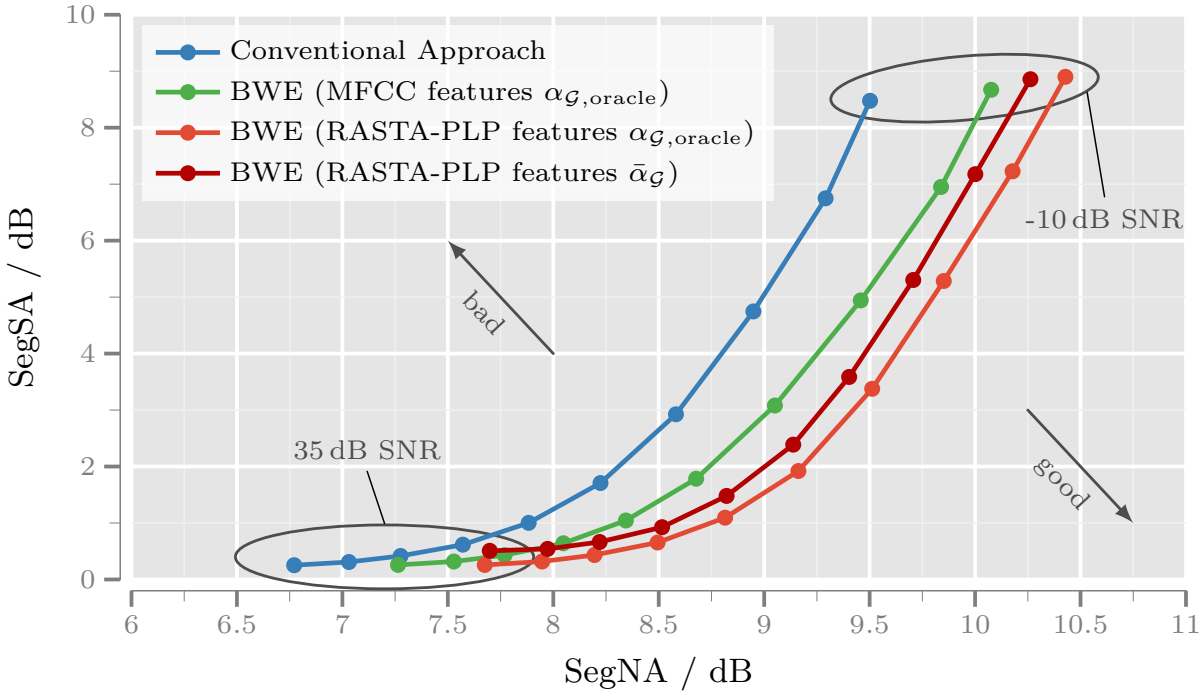



**Figure 3.22:** The *segmental speech attenuation* (SegSA) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter.

insulated. Comparing the MFCC (—●—) to the matched RASTA-PLP (—●—) features employing the new HMM training, it is obvious that new features and training further improves the performance over the complete range. In the realistic case using the estimated $\bar{\alpha}_{\mathcal{G}}$ factor (—●—) it can be seen that the performance is very similar to the oracle experiment. At low SNR values the proposed method benefits from the enhanced noisy training process, while the used RASTA-PLP features improve the SegSpSNR at higher SNR values. The results are confirmed by the averaged results presented in Fig. 3.22 which depict the objective measures SegSA over SegNA. Here, best performance would be in the upper left corner of the plot. In addition, informal listening tests confirmed the instrumental measurements and showed that the occurrence of *musical tones* is reduced by the joint noise reduction in the high band.

## 3.7 Summary

Besides the basic principles of statistical noise reduction utilizing the *short-term Fourier domain* (STFD), the novel short-term noise PSD estimator *Baseline Tracing* is presented. The basic idea consists of a constrained logarithmic magnitude tracing of the noisy observation separately for each frequency bin $\mu$. This constraint magnitude change causes slow evolution (inertia) of the noise estimate over time which models the different temporal statistics of speech and noise. In addition

the inertia performs implicit smoothing. Hence, no further smoothing of the noisy observation or the noise estimate itself is necessary, which simplifies the computation and reduces the number of algorithm parameters to only one parameter $\beta$. The estimator can be explained in terms of delta modulation with an adaptive step size, operated in the slope overload mode. In the linear domain, the noise PSD of the current frame is calculated by a simple scaling of the last noise estimate with a certain frequency and time dependent $\beta$. Stretching or compressing is decided according to the sign of the difference between the last short-term noise PSD estimate and the current noisy frame. Doing so, the estimator aims to follow the noisy observation. Since speech onsets are assumed as sudden rises in the noisy observation, $\beta$ has to be selected to only follow the noise. A fixed as well as an adaptive $\beta(\lambda, \mu)$ are presented which consider the long-term speech spectrum average and frame SNR. The new short-term noise PSD estimator is an inherently unbiased estimator in the logarithmic domain and does not need correction terms. This is also valid for the linear amplitude domain except of granular noise known from delta modulation. Compared to state-of-the-art systems, the new *Baseline Tracing* algorithm with adaptive $\beta(\lambda, \mu)$ has a superior performance with respect to the noise PSD error measure while performing similar to the *SPP* using a fixed $\beta(\mu)$. The noise reduction performance is characterized by a low cepstral distance, i. e., low speech distortion and high SegNA – SegSA measures resulting in a high noise attenuation.

In addition, an approach to wideband speech enhancement is presented that exploits spectral dependencies between the low band (50 Hz – 4 kHz) and the high band (4 kHz – 7 kHz) of speech signals in order to improve the noise reduction in the high band. While a conventional noise suppression takes place in the low band, a joint noise suppression approach is applied in the high band. Features from the processed and enhanced low band signal are extracted and used to estimate sub-band energies of the high band using techniques known from artificial bandwidth extension. The utilized RASTA-PLP features for the HMM are more robust against short-term noise variations compared to MFCC features and minimize the speaker difference. The weighting gains determined from these energy estimates are adaptively combined with conventional gains obtained in addition for the high band. This information combining in the high band is possible employing a pre-trained look-up table which is dependent on the average low band and the respective high band SNR. In order to increase the perceived speech quality if only a noisy low band signal has been received, a slightly modified version of the system can additionally be used to perform a joint noise reduction and artificial bandwidth extension.

# Chapter 4

# Codebook Based Noise Suppression

Single microphone noise reduction systems usually rely on different statistical properties of speech and noise [Boll 1979; Ephraim & Malah 1984, 1985; Lotter & Vary 2005]. In addition, it is assumed that the ambient background noise is stationary or only slightly time-varying [Martin 2001; Hendriks et al. 2010; Gerkmann & Hendriks 2011; Heese & Vary 2015] which is usually not fulfilled in practice. In consequence, statistical state-of-the-art noise estimators provide an estimate for the short-term noise *power spectral density* (PSD) in the best case. If the underlying noise signal exhibits a reasonable variance, the spectral fine-structure over frequency and time is estimated inadequately. Hence, statistical noise reduction systems are only able to remove the short-term mean of the noise which likely results in unpleasant artifacts that are called *musical tones*.

In contrast, the class of codebook based speech enhancement systems [Sreenivas & Kirnapure 1996; Srinivasan et al. 2006, 2007; Rosenkranz 2010; Rosenkranz & Puder 2012a; Sigg et al. 2012; Hao & Bao 2015; Deng & Bao 2016] faces the aforementioned constraints by using *a priori* knowledge about speech and/or noise and also allows to model and thus cope with highly non-stationary noise environments. Hence, the aim is to estimate *short-term power spectra* (STPSs) instead of short-term PSD quantities. Additionally, the codebook driven noise reduction systems have the potential to reduce the occurrence of *musical tones*, since the instantaneous speech *and* noise is estimated jointly over frequency and time.

One of the first proposals for codebook based noise reduction consists of an iterative Wiener Filter which relies on spectral constraints given by *a priori* speech knowledge [Sreenivas & Kirnapure 1996]. The block diagram of the basic concept is depicted in Fig. 4.1. A Wiener filter $\mathcal{G}(\lambda, \mu)$ is applied to the noisy input $\mathcal{Y}(\lambda, \mu)$ yielding a clean speech estimate $\widehat{\mathcal{S}}(\lambda, \mu)$. This speech estimate is converted to *linear prediction coefficient*s (LPCs) and refined using the best matching entry of a pre-trained speech codebook. The Wiener Filter in the next iteration utilizes the refined speech estimate. For the first iteration, the Wiener Filter $\mathcal{G}(\lambda, \mu)$ is initialized with $\mathcal{G}_0(\lambda, \mu) = 1$. The iteration process will be finished if the same codebook entry is chosen in two consecutive iterations.

Recent approaches employ *a priori* knowledge about both speech and noise. Spectral speech and noise estimates are obtained on a frame-by-frame basis in the frequency domain by a linear combination or a weighted sum of entries from *pre-*

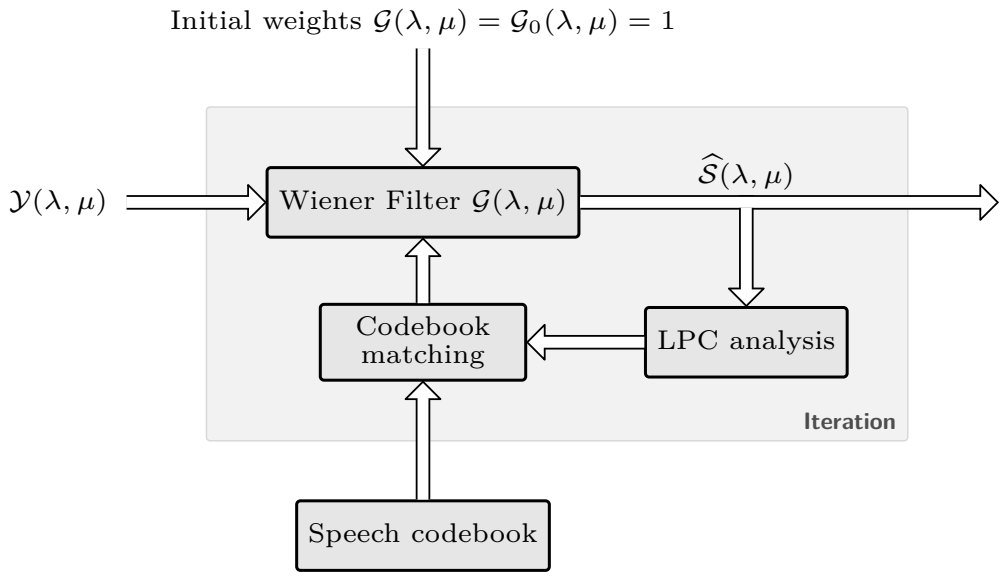



**Figure 4.1:** Block diagram of iterative codebook constrained Wiener filtering approach [Sreenivas & Kirnapure 1996].

*trained* gain-normalized codebooks. Since no closed-form solution for the optimal gain calculation of the codebook entries exists, the gains have to be approximated. The concept is illustrated in Fig. 4.2. The proposed algorithms [Srinivasan et al. 2006, 2007; Rosenkranz 2010; Rosenkranz & Puder 2012a] mainly differ in the methods of gain estimation and in the features which are stored in the codebooks. Since the employed features describe only the spectral envelope of a frame, the speech $\left|\widehat{\mathcal{S}}(\lambda,\mu)\right|^2$ and noise $\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2$ estimates exhibit a limited spectral resolution. Hence, the weighting gain rule which is derived from these estimates is also spectrally smoothed. This leads to severe speech distortion especially in voiced speech parts since the harmonic structure of speech is not modeled by spectral envelopes. By using an adaptive comb-filter as in [Rosenkranz 2010; Yoshioka et al. 2010], the spectral fine-structure can be recreated. However, this filter requires an accurate estimate of the fundamental frequency which is challenging. Since current pitch estimators are only able to reliably estimate the fundamental frequency in *signal-to-noise ratio* (SNR) ranges above approximately 10 dB [Shahnaz et al. 2005; Gonzalez & Brookes 2014], this solution is restricted to few realistic scenarios. In [Rosenkranz & Puder 2012a], the authors propose a more accurate gain estimation based on Newton's method [Bronstein et al. 1999] and use an envelope model which is based on the real-valued cepstrum for the codebook entries. In addition, a new weighting gain rule is proposed which depends only on the noise estimate, obtained by the codebook processing, and the noisy input. Doing so, the spectral fine structure is implicitly somewhat modeled by the noisy input. Due to the codebook-based spectrally smooth noise estimate, sharp spectral peaks of the noise are not accurately modeled. In turn, this generates an increased occurrence of *musical tones*.

However, the performance of codebook matching is mainly limited either by

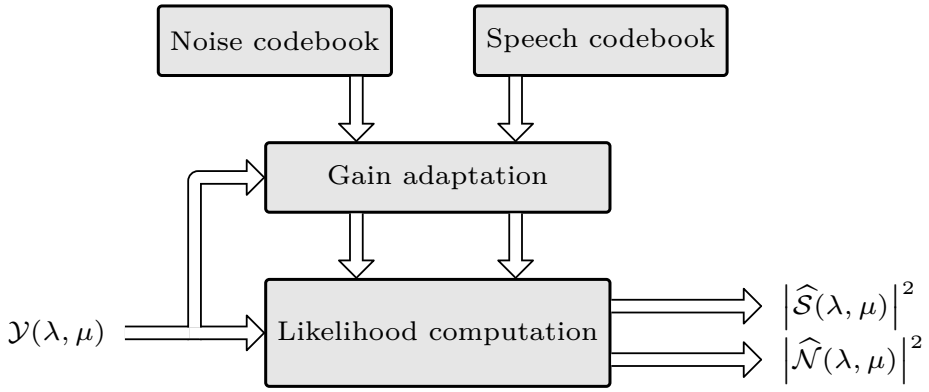



**Figure 4.2:** Basic concept of codebook driven noise reduction [Srinivasan et al. 2006, 2007; Rosenkranz 2010; Rosenkranz & Puder 2012a].

missing *a priori* knowledge, especially with respect to noise, or deviations due to the signal transmission path, i. e., changing acoustic and electrical (recording equipment, microphone) path. In [Rosenkranz 2010] the speech and noise codebook entries are adjusted (equalized) to compensate the influence of the transmission path similar to cepstral mean subtraction [Westphal 1997; Veth & Boves 1998] while in [Rosenkranz & Puder 2012b] fixed delta codebooks between the actual noise and a conventional noise estimate (e. g., [Martin 2006; Hendriks et al. 2010; Gerkmann & Hendriks 2011]) are employed to reduce the effect of missing *a priori* noise knowledge.

In summary, the main issues of codebook based speech and noise estimation are the limited spectral resolution of the codebooks and missing or unknown *a priori* knowledge regarding noise.

In the following, a novel codebook based speech and noise estimation system is presented which tackles the aforementioned problems. The basic concept of the proposed codebook speech enhancement system is the superposition of a scaled speech and noise codebook entry on a frame-by-frame basis. While the speech codebook is pre-trained offline using a representative data basis, the noise codebook is adapted quickly to new noise types online. Thus, the system is independent of *a priori* noise knowledge. Training vectors for noise codebook updates are identified using a *voice activity detector* (VAD) and a codebook mismatch measure. The VAD is realized as part of the codebook matching but utilizes only *a priori* knowledge on speech. A Wiener filter or any state-of-the-art weighting rule can be applied subsequently for speech enhancement, cf. Sec. 3.4.3. For the sake of speaker independence, the speech codebook also comprises spectral envelopes, while the noise codebook exhibits the full spectral resolution. Hence, the speech and noise estimate after codebook matching exhibit different spectral resolutions. Since no closed-form solution for optimal gain calculation exists, a brute force approach[1]

[1] The brute force approach enumerates all possible candidates for the noisy observation, i. e., all possible combinations of speech and noise codebook entries scaled by all possible gains, and evaluates which combination matches best the noisy observation.

serves as reference codebook processing platform.

However, with respect to speech enhancement applications, e. g., for a mobile phone scenario, a dramatic reduction of complexity is necessary. This is accomplished by replacing the brute force codebook matching with a cascade of gain shape estimates. This reduces the complexity significantly but provides various speech estimates and various noise estimates. Compared with the brute force search, these estimates of speech and noise are somewhat inaccurate estimates. Furthermore, the different estimates have to be merged in order to provide the final estimates of speech and noise and to improve estimation quality. The adaptive combination of different speech and noise estimates is subject to the next chapter (Chap. 5). Moreover, the entire evaluation of codebook-based speech enhancement is also presented in Chap. 5.

The remainder of this chapter is organized as follows. In Sec. 4.1 the signal model is introduced and the basic codebook matching algorithm is presented. A refined SNR estimation considering speech and noise estimates is presented in Sec. 4.2. The speech and noise codebook training is detailed in Sec. 4.3 while in Sec. 4.4 the speech codebook driven VAD including a comprehensive evaluation is described. In Sec. 4.5 the online noise codebook adaptation is explained. A summary and conclusion are presented in Sec. 4.6.

## 4.1 Speech and Noise Estimation

A simplified block diagram of the proposed codebook estimation system is given in Fig. 4.3. As in the chapter about statistical noise reduction (Sec. 3.2), it is assumed that the noisy input signal $y(k)$ consists of a clean speech signal $s(k)$ degraded by an additive noise component $n(k)$. Since the processing takes place in the *short-term Fourier domain* (STFD), the noisy input signal $y(k)$ is segmented into overlapping frames, followed by windowing and subsequent transformation into the frequency domain[2]. The spectral coefficients of the segmented and windowed

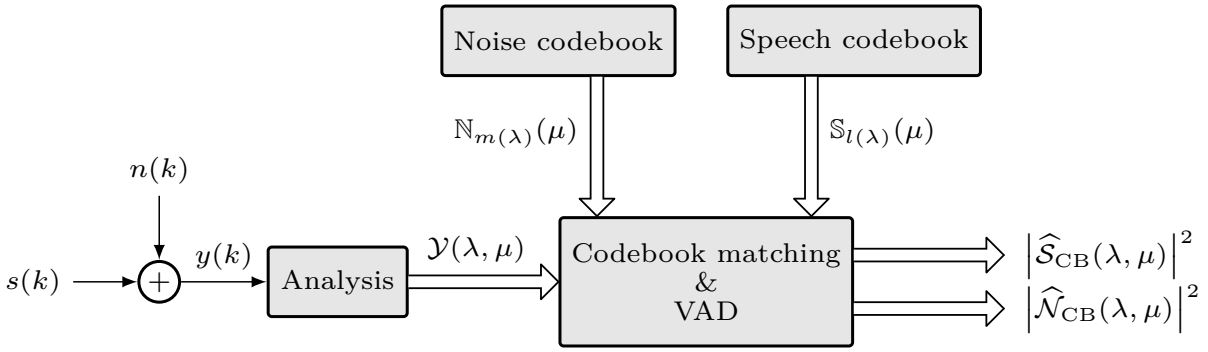



**Figure 4.3:** Proposed codebook based speech and noise estimation system

[2]Refer to Sec. 3.2.1 for the detailed analysis procedure. Note that the frequency domain representation of the respective signals already includes the effect of windowing.

input signal $y_\lambda(\kappa)$ at frequency bin $\mu$ and frame $\lambda$ are given by

$$y_\lambda(\kappa) = s_\lambda(\kappa) + n_\lambda(\kappa) \quad \overset{\text{DFT}}{\circ\!\!-\!\!\bullet} \quad \mathcal{Y}(\lambda,\mu) = \mathcal{S}(\lambda,\mu) + \mathcal{N}(\lambda,\mu), \tag{4.1}$$

where $\mathcal{S}(\lambda,\mu)$ and $\mathcal{N}(\lambda,\mu)$ correspond to the spectral coefficients of the clean speech signal and the noise signal, respectively. Speech and noise are estimated from the noisy observation by minimizing the difference between the noisy input frame $\mathcal{Y}(\lambda,\mu)$ and its estimate $\widehat{\mathcal{Y}}(\lambda,\mu)$. The estimation of the noisy input frame $\widehat{\mathcal{Y}}(\lambda,\mu)$ is modeled by a scaled superposition of a speech $\mathbb{S}_l(\mu)$ and a noise $\mathbb{N}_m(\mu)$ codebook entry according to

$$\widehat{\mathcal{Y}}(\lambda,\mu) = \underbrace{\sigma_s(\lambda)\mathbb{S}_{l(\lambda)}(\mu)}_{\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)} + \underbrace{\sigma_n(\lambda)\mathbb{N}_{m(\lambda)}(\mu)}_{\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)} \tag{4.2}$$

where $l \in \{1,\ldots,L\}$, $m \in \{1,\ldots,M\}$ denote the codebook indices and $\sigma_s$, $\sigma_n$ the gain factors of speech and noise, respectively. The codebook entries $\mathbb{S}_l(\mu)$ and $\mathbb{N}_m(\mu)$ are normalized to one with respect to their power.

With regard to speech enhancement, most of the algorithms are derived based on *power spectral density* (PSD), short-term PSD or power signal quantities. The computation of power quantities should be normalized to the frame-size for a correct physical definition, but will be neglected as it is usually done in literature. This is possible as within a specific speech enhancement system the frame-size and frame advance are fixed and therefore no normalization is necessary. Moreover, power quantities are almost always used in relation to each other, e. g., for SNR computation. Hence, the dependency on the frame-size is canceled out.

### 4.1.1 Codebook Matching by Distance Minimization

With regard to the minimization procedure, Eq. (4.2) exhibits too many degrees of freedom and should therefore be simplified. Because most of the algorithms in speech enhancement are derived on *short-term power spectrum* (STPS) quantities, it is sufficient to provide estimates for the STPS of speech and noise. Therefore, the minimization will be carried out on the STPS $|\mathcal{Y}(\lambda,\mu)|^2$, which will be derived in the following. According to the additive signal model, the STPS of the noisy observation $\mathcal{Y}(\mu)$ of the current frame $\lambda$ can be expressed as

$$\begin{aligned} |\mathcal{Y}(\mu)|^2 &= \left|\left(|\mathcal{S}(\mu)|e^{\mathrm{i}\vartheta_{\mathcal{S}}(\mu)} + |\mathcal{N}(\mu)|e^{\mathrm{i}\vartheta_{\mathcal{N}}(\mu)}\right)\right|^2 \\ &= |\mathcal{S}(\mu)|^2 + |\mathcal{N}(\mu)|^2 + 2|\mathcal{S}(\mu)||\mathcal{N}(\mu)|\cos\left(\vartheta_{\mathcal{S}}(\mu) - \vartheta_{\mathcal{N}}(\mu)\right), \end{aligned} \tag{4.3}$$

where $\vartheta_{\mathcal{S}}$ and $\vartheta_{\mathcal{N}}$ denote the phase of speech and noise, respectively. In terms of the speech and noise codebooks, the STPS estimate $\left|\widehat{\mathcal{Y}}(\mu)\right|^2$ for the current frame $\lambda$ is thus formulated by,

$$\begin{aligned} \left|\widehat{\mathcal{Y}}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2 &= \sigma_s^2\,|\mathbb{S}_l(\mu)|^2 + \sigma_n^2\,|\mathbb{N}_m(\mu)|^2 \\ &\quad + 2\sigma_s\sigma_n|\mathbb{S}_l||\mathbb{N}_m|\cos\left(\vartheta_{\mathcal{S}}(\mu) - \vartheta_{\mathcal{N}}(\mu)\right). \end{aligned} \tag{4.4}$$

Equation (4.4) describes a mapping between the set of parameters $[l, m, \sigma_s, \sigma_n]$ and the estimate $\left|\widehat{\mathcal{Y}}(\mu)\right|^2$, where $\left|\widehat{\mathcal{Y}}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2$ forms a set of all possible estimations. Given an accurate estimate for the true STPS $|\mathcal{Y}(\mu)|^2$, speech and noise are implicitly estimated due to the frequency independent gains $\sigma_s$ and $\sigma_n$. To obtain a precisely estimate for the true STPS $|\mathcal{Y}(\mu)|^2$ from the set $\left|\widehat{\mathcal{Y}}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2$, the optimal parameters $l_{\text{opt}}, m_{\text{opt}}, \sigma_{s,\text{opt}}, \sigma_{n,\text{opt}}$ can be found by minimizing:

$$\arg \min_{l,m,\sigma_s,\sigma_n} \text{dist}\left(|\mathcal{Y}(\mu)|^2, \left|\widehat{\mathcal{Y}}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2\right). \tag{4.5}$$

Equation 4.5 describes the very general approach for the codebook matching. In the following simplifications will be presented in order to reduce the computational complexity.

### 4.1.2 Model Assumptions and Simplifications

The phase difference $\vartheta(\mu) = \vartheta_{\mathcal{S}}(\mu) - \vartheta_{\mathcal{N}}(\mu)$ is unknown *a priori*. According to measurements with plain speech and noise, $\vartheta(\mu)$ is considered to be an equally uniformly distributed random variable on the interval $[0, 2\pi)$. Since $\mathbb{E}\{\cos(\vartheta(\mu))\} = 0$ due to averaging over time as, e. g., in the SNR estimation stage, the cross-term in Eq. (4.4) is omitted in the following. Experiments have confirmed that the additional estimation error of speech and noise introduced by omitting the cross-term is orders of magnitude below the true estimation error. Additional experiments have confirmed that the influence by omitting the cross-term on the performance of noise reduction is negligible (see Appendix D for further details). With this assumption Eq. (4.4) simplifies to

$$\left|\widehat{\mathcal{Y}}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2 = \sigma_s^2 |\mathbb{S}_l(\mu)|^2 + \sigma_n^2 |\mathbb{N}_m(\mu)|^2 \tag{4.6}$$

$$= \left|\widehat{\mathcal{S}}_{\text{CB}}(\mu)\right|^2 + \left|\widehat{\mathcal{N}}_{\text{CB}}(\mu)\right|^2. \tag{4.7}$$

The codebook entries $\mathbb{S}_l(\mu)$, $\mathbb{N}_m(\mu)$ are normalized to one with respect to their power. Thus, the gain factors $\sigma_s^2$ and $\sigma_n^2$ represent the short-term power of speech and noise. Applying the constraint that the power of the noisy observation is equal to the power of the optimal estimate,

$$\sum_{\mu} \left|\widehat{\mathcal{Y}}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2 \approx \sum_{\mu} |\mathcal{Y}(\mu)|^2 =: \sigma_y^2, \tag{4.8}$$

and further exploiting that the codebook entries are normalized, the speech gain $\sigma_s^2$ can be substituted and Eq. (4.6) simplifies to

$$\left|\widehat{\mathcal{Y}}_{l,m,\sigma_n}(\mu)\right|^2 = \left(\sigma_y^2 - \sigma_n^2\right) |\mathbb{S}_l(\mu)|^2 + \sigma_n^2 |\mathbb{N}_m(\mu)|^2, \tag{4.9}$$

which reduces the number of parameters to be optimized and thus the computational expense. Techniques known from gain shape *vector quantizer* (VQ) to determine the codebook entries and the gain $\sigma_n^2$ independently are not applicable here. The optimization of the gain $\sigma_n^2$ for a fixed but arbitrary combination of speech and noise codebook entries, carried out in the *minimum mean-square error* (MMSE) sense, shows that $\sigma_n^2 \geq 0$ is not guaranteed (cf. Appendix E for derivation) and in case of $\sigma_n^2 < 0$ the model assumption is violated, i. e., $\sigma_n^2$ represents the short-term power of noise. Hence, all permutations of the parameters $l$, $m$, and $\sigma_n$ must be taken into account, which can be realized by a quantization of $\sigma_n$ according to:

$$\sigma_n = \frac{i}{N_q - 1}\sigma_y, \quad i = 0, ..., N_q - 1. \tag{4.10}$$

Finally, the optimal parameters $l_{\text{opt}}, m_{\text{opt}}, \sigma_{n,\text{opt}}$ can be found by minimizing:

$$\arg \min_{l,m,\sigma_n} \text{dist}\left( \left|\mathcal{Y}(\mu)\right|^2, \left|\widehat{\mathcal{Y}}_{l,m,\sigma_n}(\mu)\right|^2 \right), \tag{4.11}$$

where $\text{dist}(\cdot,\cdot)$ represents an arbitrary distance measure. Hence, the codebook estimate of $|\mathcal{Y}(\mu)|^2$ for any frame $\lambda$ yields,

$$\left|\widehat{\mathcal{Y}}(\mu)\right|^2 = \left|\widehat{\mathcal{Y}}_{l_{\text{opt}},m_{\text{opt}},\sigma_{n,\text{opt}}}(\mu)\right|^2 \tag{4.12}$$

$$= \underbrace{\left(\sigma_y^2 - \sigma_{n,\text{opt}}^2\right)\left|\mathbb{S}_{l_{\text{opt}}}(\mu)\right|^2}_{\left|\widehat{\mathcal{S}}_{\text{CB}}(\mu)\right|^2} + \underbrace{\sigma_{n,\text{opt}}^2\left|\mathbb{N}_{m_{\text{opt}}}(\mu)\right|^2}_{\left|\widehat{\mathcal{N}}_{\text{CB}}(\mu)\right|^2}. \tag{4.13}$$

In the following, the subscript "opt" will be omitted for the sake of brevity. In order to reduce the speaker dependence, only spectral envelopes are stored as speech codebook entries.

### 4.1.3 Distance Measures

For the implementation of Eq. (4.11) a suitable distance measure is necessary. In this section possible distance measures are presented. The notation of the distance operator $\text{dist}|\ $ is illustrated by,

$$\text{dist}\left|{}^{\mathcal{P},\widehat{\mathcal{P}}}_{\text{Algorithm}} = f\left(\mathcal{P}(\mu), \widehat{\mathcal{P}}(\mu)\right). \tag{4.14}$$

While the operands from which the distance is calculated are denoted at the top of the vertical bar symbol, the actual distance algorithm is indicated at the bottom. If no algorithm is specified, the dist operator serves as place holder for an arbitrary distance measure. Note that in the expression the operands $\mathcal{P}$ and $\widehat{\mathcal{P}}$ are assumed to be power quantities. The distance measures express the difference between an original spectrum $\mathcal{P}(\mu)$ and the estimation or approximation $\widehat{\mathcal{P}}(\mu)$ of that spectrum as a function $f$, describing the employed algorithm.

### Mean-Square Error difference

$$\operatorname{dist}\Big|_{\mathrm{MSE}}^{\mathcal{P},\widehat{\mathcal{P}}} = \frac{1}{N_{\mathrm{DFT}}} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left(\sqrt{\mathcal{P}(\mu)} - \sqrt{\widehat{\mathcal{P}}(\mu)}\right)^2 . \tag{4.15}$$

### Relative Power difference

$$\operatorname{dist}\Big|_{\mathrm{REL}}^{\mathcal{P},\widehat{\mathcal{P}}} = \frac{1}{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \mathcal{P}(\mu)} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left|\mathcal{P}(\mu) - \widehat{\mathcal{P}}(\mu)\right| . \tag{4.16}$$

### Itakura-Saito distance

The Itakura–Saito distance [Itakura & Saito 1968] is often used for speech coding and speech quality assessment. It is not designed as perceptual measure, but it reflects subjective meaningful distortion for the spectral shape of speech. Due to its asymmetric nature the Itakura–Saito distance is more sensitive to spectral peaks than spectral valleys [Wei & Gibson 2000] and is defined as

$$\operatorname{dist}\Big|_{\mathrm{IS}}^{\mathcal{P}(\mu),\widehat{\mathcal{P}}(\mu)} = \frac{1}{N_{\mathrm{DFT}}} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left[\frac{\mathcal{P}}{\widehat{\mathcal{P}}(\mu)} - \log\frac{\mathcal{P}}{\widehat{\mathcal{P}}(\mu)} - 1\right] . \tag{4.17}$$

## 4.2 Modified Decision-Directed SNR Estimation

In case of statistical speech enhancement, the *a priori* SNR estimation is usually carried out by the *decision-directed* approach [Ephraim & Malah 1984] which is detailed in Sec. 3.4.2. The *decision-directed* SNR $\widehat{\xi}_{\mathrm{Stat}}(\lambda,\mu)$ only depends on a noise estimate and the previous enhanced frame from the output of the speech enhancement system. The *a priori* SNR estimate is formulated by a linear combination of speech and noise estimates from the last frame and an instantaneous realization of the *a posteriori* SNR $\overline{\gamma}(\lambda,\mu)$,

$$\overline{\gamma}(\lambda,\mu) = \frac{|\mathcal{Y}(\lambda,\mu)|^2}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}} \approx \frac{|\mathcal{S}(\lambda,\mu)|^2}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}} + \frac{|\mathcal{N}(\lambda,\mu)|^2}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}}, \tag{4.18}$$

$$\widehat{\xi}_{\mathrm{Stat}}(\lambda,\mu) = \alpha_\xi \underbrace{\frac{\left|\widehat{\mathcal{S}}(\lambda-1,\mu)\right|^2}{\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda-1,\mu)|^2\right\}}}_{\mathrm{SNR_{DD}}(\lambda-1,\mu)} + (1-\alpha_\xi)\underbrace{\max\left(\overline{\gamma}(\lambda,\mu)-1,\ 0\right)}_{\mathrm{SNR_i}(\lambda,\mu)} . \tag{4.19}$$

Conceptually, the *decision-directed* SNR can be interpreted as a weighted sum of two *a priori* SNR estimates, the decision-directed $\mathrm{SNR_{DD}}$ and instantaneous $\mathrm{SNR_i}$. The $\mathrm{SNR_{DD}}$ estimate is a refined version of the *a priori* SNR of the previous frame

incorporating the *decision* of the speech enhancement system. The $\mathrm{SNR_i}$ is an *a priori* SNR estimate utilizing the instantaneous realization of the *a posteriori* SNR $\overline{\gamma}(\lambda,\mu)$ which only depends on a noise estimate. Since $\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\}$ may differ significantly from $|\mathcal{N}(\lambda,\mu)|^2$ the simplification $\overline{\mathbb{E}}_K\left\{|\mathcal{N}(\lambda,\mu)|^2\right\} = |\mathcal{N}(\lambda,\mu)|^2$ does not hold and results in an imprecise *a priori* SNR $\overline{\gamma}(\lambda,\mu)-1$. As a workaround, the estimate $\overline{\gamma}(\lambda,\mu)-1$ is limited to values greater or equal to zero applying the max operator.

In contrast to statistical speech enhancement, the codebook driven approach can additionally exploit a speech estimate in the current frame. Hence, the *a priori* SNR estimate $\mathrm{SNR_i}$ in Eq. (4.19) is replaced by $\mathrm{SNR_{CB}}$ which exploits both, the speech $\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ and noise $\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)$ estimate. Considering the high temporal resolution of the codebook matching, the *a priori* SNR estimate according to the *decicison-directed* approach is now given in terms of STPS by,

$$\widehat{\xi}_{\mathrm{CB}}(\lambda,\mu) = \alpha_\xi \underbrace{\frac{\left|\widehat{\mathcal{S}}(\lambda-1,\mu)\right|^2}{\left|\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda-1,\mu)\right|^2}}_{\mathrm{SNR_{DD}}(\lambda-1,\mu)} + (1-\alpha_\xi)\underbrace{\frac{\left|\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)\right|^2}{\left|\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)\right|^2}}_{\mathrm{SNR_{CB}}(\lambda,\mu)}. \tag{4.20}$$

As mentioned before, the evaluation of the modified Decision-Directed SNR estimation is presented in the next chapter in Sec. 5.8.3.

## 4.3 Codebook Training

A crucial point is the generation of suitable speech and noise codebook entries which form the codebooks. As mentioned before, the noise codebook entries exhibit the full spectral resolution while the speech codebook consists of spectral envelopes in order to reduce the speaker dependence. Thus, the noise codebook entries are stored as *short-term power spectrum* (STPS). There exist several compact representations of the spectral envelope which are based on *auto-regressive* (AR) modeling [Itakura 1975; Kleijn & Paliwal 1995; Murthi & Rao 2000; Soong & Juang 1984] like the *linear prediction coefficient*s (LPCs), the *line spectral frequencies* (LSF) or the *minimum variance distortionless response* (MVDR) representation. For simplicity of the simulation framework, the speech codebook entries are also stored as STPS. Hence, the codebook entry vectors for speech and noise are defined in vector notation as

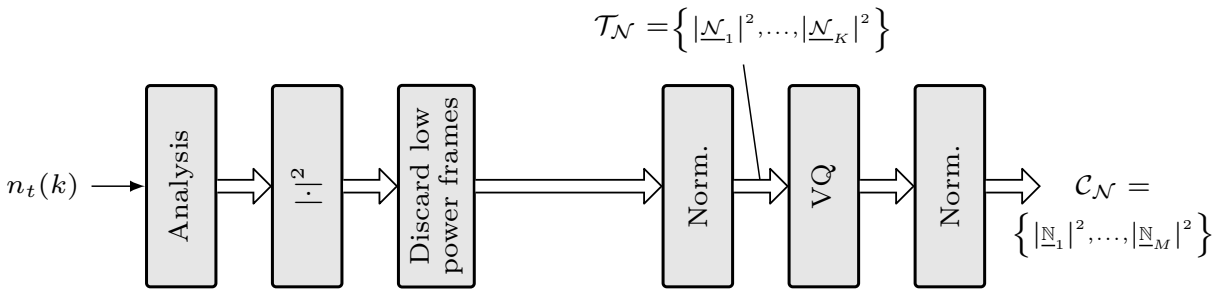
$$|\underline{\mathbb{S}}_l|^2 = \left(|\mathbb{S}_l(\mu=0)|^2,\ldots,|\mathbb{S}_l(\mu=N_{\mathrm{DFT}}-1)|^2\right)^{\intercal}, \tag{4.21}$$

$$|\underline{\mathbb{N}}_m|^2 = \left(|\mathbb{N}_m(\mu=0)|^2,\ldots,|\mathbb{N}_m(\mu=N_{\mathrm{DFT}}-1)|^2\right)^{\intercal}, \tag{4.22}$$

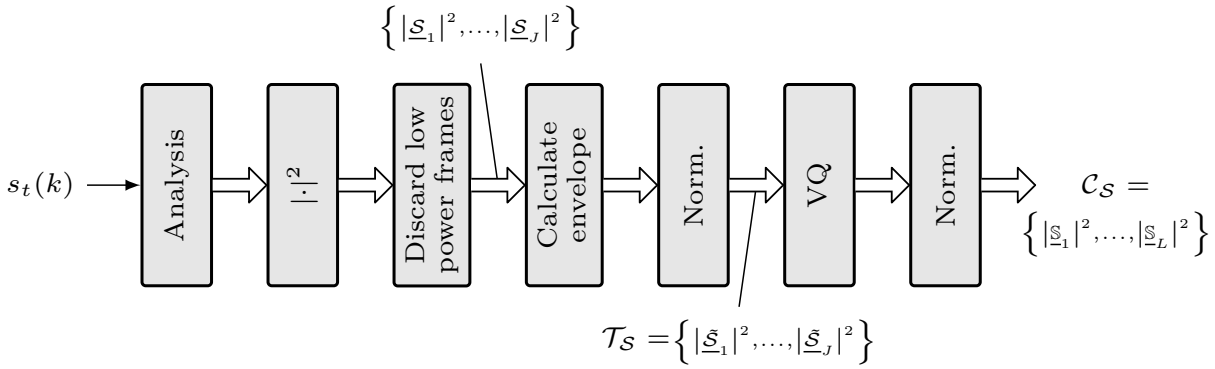
each containing a STPS. The respective speech and noise codebooks,

$$\mathcal{C}_{\mathcal{S}} = \left\{|\underline{\mathbb{S}}_1|^2,\ldots,|\underline{\mathbb{S}}_L|^2\right\}, \tag{4.23}$$

$$\mathcal{C}_{\mathcal{N}} = \left\{|\underline{\mathbb{N}}_1|^2,\ldots,|\underline{\mathbb{N}}_M|^2\right\}, \tag{4.24}$$




**(a)** Noise




**(b)** Speech

**Figure 4.4:** Generation of speech (a) and noise (b) codebook entries

consist of collections of entry vectors. Since the noise codebook $\mathcal{C}_\mathcal{N}$ shall be adapted online it is further necessary to divide the codebook into a fixed part consisting of $M_\square \in \mathbb{N}$ base codebook entries and an adaptive part with a maximum number of $M_\circ \in \mathbb{N}_0$ entries, which are created and updated online.

In the following, the codebook training process is explained which is valid for offline creation as well as online adaptation. Figure 4.4a depicts the block diagram of the noise codebook training process while Fig. 4.4b presents the generation of the speech codebook. Each training sequence as indicated by the subscript $t$ is normalized to $-26\,\mathrm{dBov}$ according to ITU P.56 [ITU-T Recommendation P.56 1993] standardization. Following, the sequence is segmented into overlapping frames, windowed and transformed into the frequency domain according to Sec. 3.2.1. For a compact representation, the spectral coefficients of the speech and noise training frames are denoted by,

$$\underline{\mathcal{N}}_\lambda = (\mathcal{N}(\lambda, \mu = 0), \ldots, \mathcal{N}(\lambda, \mu = N_\mathrm{DFT} - 1))^\intercal, \tag{4.25}$$

$$\underline{\mathcal{S}}_\lambda = (\mathcal{S}(\lambda, \mu = 0), \ldots, \mathcal{S}(\lambda, \mu = N_\mathrm{DFT} - 1))^\intercal. \tag{4.26}$$

After applying the magnitude square operation, all resulting STPS frames below a certain power threshold are discarded. Since the input sequence is normalized, this threshold can be adjusted independently of the training input. On the one hand, doing so removes silent parts of the training data which may be over-represented

in the subsequent vector quantization. On the other hand, it prevents frames with up-scaled recording noise (self noise of recording equipment, e. g., microphone, amplifier, and analog-to-digital converter) after power normalization. As indicated before, different training schemes are applied for the speech and noise training sequences. In the following, the particularities of speech and noise are investigated.

### 4.3.1 Noise Codebook

The training sequence $n_t(k)$ consists of plain noise. Applying the aforementioned procedure, a large number of $K$ STPS input vectors $\mathcal{T}_\mathcal{N} = \left\{ |\underline{\mathcal{N}}_1|^2, \ldots, |\underline{\mathcal{N}}_K|^2 \right\}$ exist which are used for the training of a VQ. The result of the VQ training is used as codebook. For the fixed part of the noise codebook the VQ is configured to return $M_\square$ codebook entries while the adaptive part of size $M_\circ$ is constructed by sub-codebooks consisting of $M_\Delta$ entries with $r \cdot M_\Delta \leq M_\circ$ and $r \in \mathbb{N}_0$ is the number of sub-codebooks. In this work, the LBG algorithm [Linde et al. 1980] is employed together with the Itakura Saito distance (Eqn. 4.17) as distance measure. As it is not assured that the output of the VQ training is still normalized, subsequent normalization to one with respect to the power is applied again.

To obtain a fixed noise codebook which contains several noise types, it has been proven useful, to concatenate individual sub-codebooks each trained with meaningful prototype sequences of the particular noise type.

### 4.3.2 Speech Codebook

To keep the speech codebook as generic as possible, the speaker dependence of the speech codebook entries is reduced in order to contain mainly information about the spoken phonemes. According to the source-filter model of human speech production [Vary & Martin 2006], speech is created by an excitation signal which has a flat spectral shape and a subsequent vocal tract filter which forms the spectral shape of the specific phoneme. The excitation signal's counterpart in the human speech production system consists of the lungs and larynx, while the vocal tract filter models the neck, nasal cavity, and the mouth. From the areas of low bitrate speech coding and speech recognition it is known, that the excitation signal is significantly speaker dependent while the vocal tract filter is rather similar among different speakers. Voiced sounds are constructed by an excitation signal consisting of periodic pulses caused by the larynx. The frequency of these pulses is called fundamental pitch frequency $f_\mathrm{p}$ and is very specific among various humans. In particular, the pitch of men is in the range of 50 – 250 Hz and typically lower than for women with a pitch of 120 – 500 Hz. In contrast, unvoiced sounds like "s" or "ch" are caused by a white noise excitation signal whereas plosives like "p" and "k" are created by sudden pressure-rises in the vocal tract. Thus, voiced sounds are more critical with respect to speaker-dependence. The spectrum of a voiced excitation signal is characterized by a harmonic structure with the fundamental pitch frequency as distance between the spectral peaks. In order to significantly reduce the speaker dependency, the speaker-dependent excitation signal is removed, i. e., the spectral

envelope has to be calculated. Doing so, a training set $\mathcal{T}_{\mathcal{S}} = \left\{ \left|\underline{\tilde{\mathcal{S}}}_1\right|^2, \ldots, \left|\underline{\tilde{\mathcal{S}}}_J\right|^2 \right\}$ of STPSs consisting of the speaker-independent spectral envelopes emerges. Different methods for obtaining the spectral envelope are known from literature [Rosenkranz 2012]. Popular ones are based on AR modeling and linear prediction or the cepstral smoothing. The AR model has a pitch dependency since with significant higher model order also the pitch harmonics are included in the envelope estimate. In turn, the cepstral model separates the spectral envelope and pitch strictly and thus a more accurate estimate of the spectral envelope is possible [Rosenkranz 2012]. Therefore, the cepstral smoothing is preferred and will be used in the following.

### Cepstral Processing

The speaker-dependent pitch frequency $f_\mathrm{p}$ of the excitation is assumed to be in the range between 50 Hz and 500 Hz [Vary & Martin 2006]. As mentioned before, a cepstral approach, like in [Rosenkranz 2010], is applied to separate the spectral envelope and the excitation. Therefore, the clean speech STPSs $\left|\underline{\mathcal{S}}_j\right|^2$ of the training data are frame-wise transformed to the cepstral domain:

$$\mathbb{C}_{\left|\underline{\mathcal{S}}_j\right|^2}(q) = \frac{1}{2} \sum_{\mu=0}^{N_\mathrm{DFT}-1} \log\left(\left|\mathcal{S}_j(\mu)\right|^2\right) \mathrm{e}^{\mathrm{i}2\pi \frac{\mu \mathrm{q}}{\mathrm{N_{DFT}}}}, \quad q = 0, \ldots, N_\mathrm{DFT} - 1, \tag{4.27}$$

where $q$ represents the cepstral bin index (quefrency). A pitch frequency $f_\mathrm{p}$ is represented in the cepstrum as a peak in the cepstral bin $q_p = \left\lfloor \frac{f_\mathrm{s}}{f_\mathrm{p}} \right\rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor rounding operator [Martin et al. 2008; Rosenkranz 2010]. Assuming that the pitch frequencies are bounded to be lower than 500 Hz and considering the symmetry of the cepstral coefficients, the range $q_\mathrm{p} < q < N_\mathrm{DFT} - q_\mathrm{p}$ is called the excitation part in the following.

The speaker-dependent excitation is removed from the training sequence $\mathcal{T}_{\mathcal{S}}$ by setting the corresponding cepstral coefficients of the excitation part to zero:

$$\mathbb{C}_{\left|\underline{\tilde{\mathcal{S}}}_j\right|^2}(q) = \begin{cases} 0 & \text{if } q_\mathrm{p} < q < M - q_\mathrm{p} \\ \mathbb{C}_{\left|\underline{\mathcal{S}}_j\right|^2}(q) & \text{else.} \end{cases} \tag{4.28}$$

Afterwards, the modified cepstrum $\mathbb{C}_{\left|\underline{\tilde{\mathcal{S}}}_j\right|^2}(q)$ is transformed back to the spectral domain:

$$\left|\tilde{\mathcal{S}}_j(\mu)\right|^2 = \exp\left( 2 \cdot \sum_{q=0}^{N_\mathrm{DFT}-1} \mathbb{C}_{\left|\underline{\tilde{\mathcal{S}}}_j\right|^2}(q) e^{-\mathrm{i}2\pi \frac{\mu \mathrm{q}}{\mathrm{N_{DFT}}}} \right). \tag{4.29}$$

A subsequent normalization of $\left|\tilde{\mathcal{S}}_j(\mu)\right|^2$ to one with respect to the power is applied to obtain the training set $\mathcal{T}_{\mathcal{S}} = \left\{ \left|\underline{\tilde{\mathcal{S}}}_1\right|^2, \ldots, \left|\underline{\tilde{\mathcal{S}}}_J\right|^2 \right\}$ consisting of $J$ spectral envelope STPSs,

$$\left|\underline{\tilde{\mathcal{S}}}_j\right|^2 = \left( \left|\tilde{\mathcal{S}}_j(\mu = 0)\right|^2, \ldots, \left|\tilde{\mathcal{S}}_j(\mu = N_\mathrm{DFT} - 1)\right|^2 \right)^\mathsf{T}. \tag{4.30}$$

The same VQ configuration as for the noise codebook is used for the speech codebook creation. After VQ training of $\mathcal{T}_\mathcal{S}$, finally, a codebook $\mathcal{C}_\mathcal{S}$ with $L$ entries is created. Each codebook entry STPS is normalized to a power of one since the output vectors of VQ process are not ensured to be normalized.

### 4.3.3 Codebook Training Quality Measure

In order to obtain new training sequences for the online adaptation of the noise codebook, a suitable mismatch measure $\mathrm{Q}\,|_M$ and a threshold $\mathrm{Q}\,|_\mathcal{C}$ are required. The mismatch measure describes the ability to approximate the current noisy observation $\mathcal{Y}(\lambda, \mu)$ by means of the speech and noise codebooks during the codebook matching process. The computation of the threshold $\mathrm{Q}\,|_\mathcal{C}$ will be formulated in terms of the speech codebook training quality $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ and the noise codebook training quality $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{N}}$. Both the mismatch measure and the threshold are derived in Sec. 4.5. In addition, the speech codebook training quality $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ serves also as indicator for an adequate speech codebook training.

The codebook training quality is defined as the ability of the respective codebook to represent its training data. Hence, the training quality measure is calculated for each codebook according to,

$$\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}} = \overline{\mathrm{dist}}\,\Big|^{\mathcal{C}_\mathcal{S},\mathcal{T}_\mathcal{S}}\,, \tag{4.31}$$

$$\overline{\mathrm{dist}}\,\Big|^{\mathcal{C}_\mathcal{S},\mathcal{T}_\mathcal{S}} = \frac{1}{J} \sum_{|\underline{\tilde{\mathcal{S}}}_i|^2 \in \mathcal{T}_\mathcal{S}} \min_l \left\{ \mathrm{dist}\left( |\underline{\mathbb{S}}_l|^2 , \left|\underline{\tilde{\mathcal{S}}}_i\right|^2 \right) \;\middle|\; l \in (1, \ldots, L) \right\}, \tag{4.32}$$

$$\mathrm{Q}\,|_{\mathcal{C}_\mathcal{N}} = \overline{\mathrm{dist}}\,\Big|^{\mathcal{C}_\mathcal{N},\mathcal{T}_\mathcal{N}}\,, \tag{4.33}$$

$$\overline{\mathrm{dist}}\,\Big|^{\mathcal{C}_\mathcal{N},\mathcal{T}_\mathcal{N}} = \frac{1}{K} \sum_{|\underline{\mathcal{N}}_i|^2 \in \mathcal{T}_\mathcal{N}} \min_l \left\{ \mathrm{dist}\left( |\underline{\mathbb{N}}_m|^2, |\underline{\mathcal{N}}_i|^2 \right) \;\middle|\; m \in (1, \ldots, M) \right\}. \tag{4.34}$$

The codebook training quality is basically the mean of the distance between each training input vector to the closest codebook entry vector, where dist is the same distance measure which is employed during the codebook matching process.

### 4.3.4 Evaluation of Speech Codebook Training Quality

Since the speech codebook consists of *a priori* speech knowledge, it is created in advance and therefore needs to be dimensioned appropriately. For the generation of the speech codebook two degrees of freedom are available, the training length, i. e., the number of training frames $J$ and the number of speech codebook entries $L$.

While the LBG algorithm, which is used for the generation of the speech codebook entries, uses the Itakura Saito distance as distance measure, different distance measures can be applied during the application of the speech codebook. Hence, the speech codebook training is analyzed with respect to the employed distance measures, i. e., the Itakura Saito distance or the relative power distance. As detailed in Sec. 4.3.3, the speech codebook training quality $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ is defined as

| **Parameter** | **Settings** |
|---|---|
| Sampling frequency $f_\mathrm{s}$ | 16 kHz |
| Frame length $L_F$ | 320 ($\hat{=}$ 20 ms) |
| Frame advance $L_A$ | 160 ($\hat{=}$ 10 ms) |
| FFT length $N_\mathrm{DFT}$ | 512 (including zero-padding) |
| Frame overlap | 50 % ($\sqrt{\text{Hann-window}}$) |
| Maximum pitch frequency $f_\mathrm{p}$ | 500 Hz |

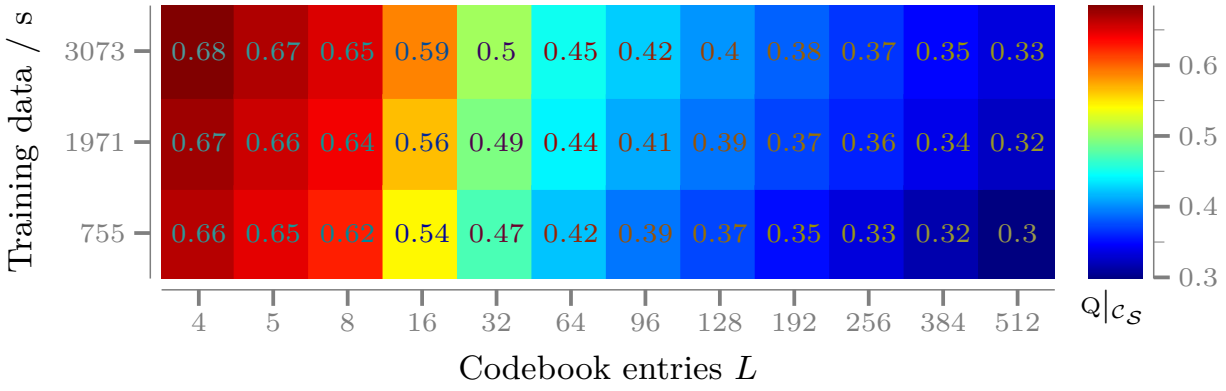
**Table 4.1:** Speech codebook training parameters

the ability of the respective codebook to represent its training data utilizing the chosen distance measure.

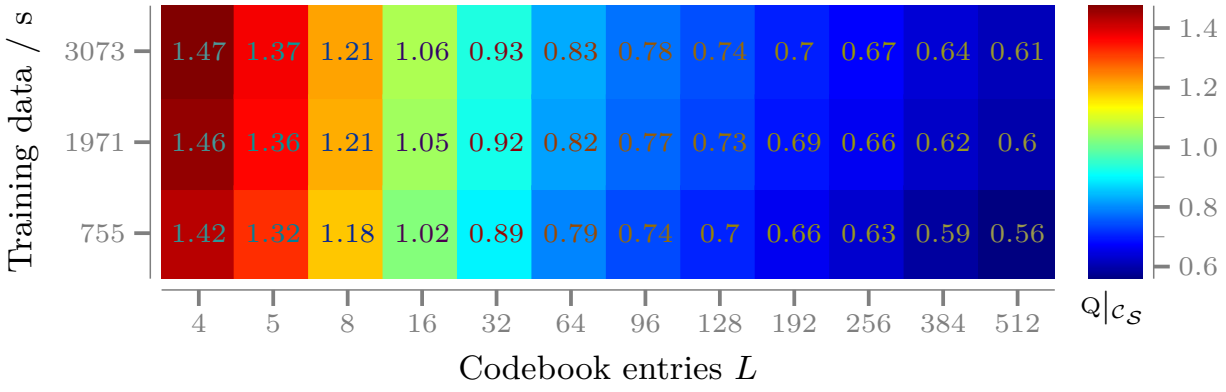
The parameters for the speech codebook training are summarized in Table 4.1. The training data consists of a randomly chosen subset from the test set of the TIMIT database [Garofolo & Consortium 1993]. The speech codebook training is carried out according to Sec. 4.3.2. The results in terms of $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ are calculated according to Sec. 4.3.3 and depicted regarding the relative power distance in Fig. 4.5a and for the Itakura Saito distance in Fig. 4.5b. The color of the respective plot denotes the codebook quality $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ from blue (good) to red (bad).

In general, a greater codebook size ensures a better performance. Both distance measures perform very similar with respect to the gradation for both, the codebook size and the training length, as indicated by the color. In addition, both configurations exhibit a slight quality degradation with increasing training length given a fixed but arbitrary codebook size. This is caused due to the higher variance of the training data for a increasing number of training data. Since the TIMIT database consists of 42 phonemes a saturation of the quality measures can be expected for $L >= 42$. This is confirmed as with a codebook size of $L = 64$ entries both quality measures start to saturate.

If not stated otherwise, the speech codebook is created from a 3073 s training sequence from the test set of the TIMIT database and exhibits $L = 128$ codebook entries. This choice states a good compromise between numerical complexity during the application of the codebook and the codebook quality $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ for both used distance measures.




**(a)** Relative power difference




**(b)** Itakura Saito distance

**Figure 4.5:** The speech codebook training quality $\mathrm{Q}|_{\mathcal{CS}}$ is depicted for various codebook entry sizes $L$, different training data lengths and distance metrics. The quality measure $\mathrm{Q}|_{\mathcal{CS}}$ describes the ability of the codebook to represent its training data and can be calculated for different distance metrics. Hence, small values indicate a better performance.

## 4.4 Speech Codebook based VAD

The objective of a *voice activity detector* (VAD) is to detect the presence or absence of human speech in, e. g., a microphone signal which might be degraded by background noise. As mentioned before, a robust VAD with respect to highly non-stationary background noise is required for online noise codebook adaptation.

Early VAD systems extract simple energy features such as SNR estimations, that respond while speech is present, and compare the quantified values to a fixed or adaptive threshold for a VAD decision, e. g., [McAulay & Malpass 1980; Van Compernolle 1989; Vary & Martin 2006]. In the GSM cellular radio system the VAD [ETSI Recommendation GSM 06.32 1996] is basically an energy detector whose

accuracy is improved by adaptive filtering to increase the speech-to-noise ratio. Since the encountered noise in mobile environments may be constantly changing with time and frequency, the adaptive filter is only updated when three conditions are fulfilled: speech is absent, the signal seems stationary, and does not include a pitch component which is inherent in voiced speech.

However, energy based techniques do not work reliably under adverse acoustic conditions, e.g., at signal-to-noise ratios of 0 dB or below. Recent systems mainly employ statistical models, also including additional features like the zero crossing rate, pitch, tone, complex-signal correlation, and the energy levels of frequency bands [Cho & Kondoz 2001; Ghosh et al. 2011; Sohn et al. 1999; Vähätalo & Johansson 1999]. By adding more microphones, the voice activity detection accuracy can be improved, e. g., [Rosca et al. 2002; Taghizadeh et al. 2011]. All these approaches cope with moderate, mainly stationary noise. However, for many applications, they are not sufficiently robust with respect to highly non-stationary noise.

Sohn [Sohn et al. 1999] proposes a likelihood ratio test, combined with a Markov process, that models speech occurrences in order to obtain a VAD. Cho [Cho & Kondoz 2001] analyzes this method and improves some fundamental problems at speech offset regions using a smoothed likelihood ratio for the adaptation of the noise variance, resulting in an improved decision of voice activity. Tan [Tan et al. 2010] employs a likelihood ratio test and modifies the handling of voiced frames by selecting exclusively the harmonic components for computing. Ghosh [Ghosh et al. 2011] introduces a "long-term signal variability measure" which represents the degree of non-stationarity. Combined with the assumption that speech is significantly less stationary than noise, this measure discriminates between noise and noisy speech, resulting in a more robust VAD performance.

Here, a new approach is presented that is operating in the short-time *discrete Fourier transform* (DFT) domain and provides soft VAD decisions. The proposed algorithm is a continued development of [Heese et al. 2015]. Acoustically degraded speech signals are frame-wise compared with a speech codebook. Doing so, a similarity measure between the input signal and typical spectral speech compositions is determined and further processed to obtain a soft speech presence indicator. This new technique is robust to highly non-stationary noise types and reliably detects speech also in adverse SNR conditions of -5 dB. Since the speech codebook is designed speaker-independently and the algorithm does not rely on a noise codebook, the algorithm is not restricted to known speakers and independent to different noise types.

### 4.4.1 Codebook VAD Overview

The VAD algorithm is carried out using a speech codebook as *a-priori* knowledge. An overview of the algorithm is depicted in Fig. 4.6. A possibly degraded speech signal $\mathcal{Y}(\lambda, \mu)$ is frame-wise compared with a speech codebook by utilizing gain shape vector quantization. A modified version of the speech codebook is adapted in every frame to the current speaker by combining in the cepstral domain the current

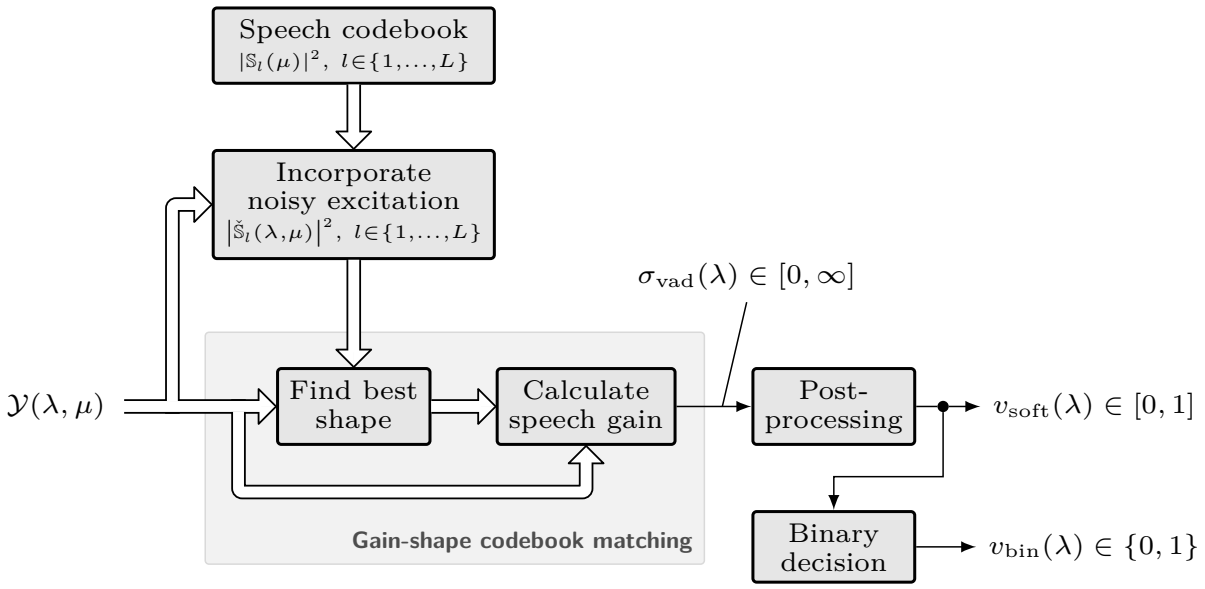



**Figure 4.6:** Speech codebook based VAD

noisy speech frame with the codebook entries as explained in the next section (Sec. 4.4.2). Soft VAD values $v_{\mathrm{soft}}(\lambda)$ ranging from zero to one are calculated by post-processing of the speech gain $\sigma_{\mathrm{vad}}(\lambda)$. If desired, a binary VAD $v_{\mathrm{bin}}(\lambda)$ can be calculated from the soft VAD values $v_{\mathrm{soft}}(\lambda)$, e. g., by applying a threshold.

### 4.4.2 Gain Shape Codebook Matching

In contrast to a joint codebook matching of speech and noise as in Sec. 4.1.1, the concept of codebook driven VAD employs only a speech codebook. Thus, gain shape codebook matching is possible, i. e., the determination of the spectral shape using gain normalized codebooks in a first step and subsequently the calculation of the speech gain in a second step.

In contradiction to the noisy input frames, the speaker-independent codebook $\mathcal{C}_{\mathcal{S}}$ contains only entries with spectral envelopes $|\mathbb{S}_l|^2$. The envelopes $|\mathbb{S}_l|^2$ have to be modified according to Eq. (4.35), as detailed below, in order to re-established their harmonic structure caused by the excitation of the source-filter model (cf. Sec. 4.3.2). This improves the determination of the spectral shape during the gain shape matching process. The fundamental principle is to compare the noisy speech signal $|\mathcal{Y}(\lambda,\mu)|^2$ frame-wise with modified speech codebook entries $\left|\check{\mathbb{S}}_l(\mu)\right|^2$ in order to find the entry $\left|\check{\mathbb{S}}_{l_{opt}}(\mu)\right|^2$ which fits best the current noisy frame.

The adapted codebook entries $\left|\check{\mathbb{S}}_l\right|^2$ have a comb-like structure whose pitch frequency $f_{\mathrm{p}}$ equals the one of the current input speech frame. In the cepstral domain the comb-like harmonic structure is mapped into one pitch specific cepstral bin. Thus, the power of this cepstral bin is assumed to be significantly above the noise floor of neighboring bins. Hence, the codebook adaptation is realized by means of a cepstral approach, i. e., the excitation part from the noisy STPS

$|\mathcal{Y}(\lambda,\mu)|^2$ is extracted and incorporated into each codebook entry $|\mathbb{S}_l(\mu)|^2$. This procedure is repeated for each input frame.

The cepstral representation $\mathbb{C}_{\underline{\mathbb{S}}_l}(q)$ of the codebook entries $|\underline{\mathbb{S}}_l|^2$ is calculated analogously to Eq. (4.27) and $\mathbb{C}_{|\mathcal{Y}(\lambda,\mu)|^2}(q)$ is the cepstrum of the noisy speech signal. The envelope of $\mathbb{C}_{\underline{\mathbb{S}}_l}(q)$ and the pitch from $\mathbb{C}_{|\mathcal{Y}(\lambda,\mu)|^2}(q)$ are combined according to:

$$\mathbb{C}_{\left|\check{\mathbb{S}}_l(\lambda,\mu)\right|^2}(q) = \begin{cases} \mathbb{C}_{|\mathcal{Y}(\lambda,\mu)|^2}(q) & q_\mathrm{p} < q < N_\mathrm{DFT} - q_\mathrm{p} \\ \mathbb{C}_{|\mathbb{S}_l(\mu)|^2}(q) & \text{else,} \end{cases} \tag{4.35}$$

where $q_\mathrm{p} < q < N_\mathrm{DFT} - q_\mathrm{p}$ represents the excitation part and $q_p = \left\lfloor \frac{f_\mathrm{s}}{f_\mathrm{p}} \right\rfloor$ is the cepstral bin corresponding to a pitch frequency of $f_\mathrm{p}$. Afterwards $\mathbb{C}_{\left|\check{\mathbb{S}}_l(\lambda,\mu)\right|^2}(q)$ is transformed to the spectral representation analogously to Eq. (4.29) and normalized to a power of one. The result $\left|\check{\mathbb{S}}_l(\lambda,\mu)\right|^2$ is a codebook entry which is adapted to the current speaker with a corresponding harmonic frequency structure.

Finally, the optimal speech codebook entry $l_\mathrm{opt}$ for the current frame $\lambda$ can be found by minimizing:

$$\arg\min_l \operatorname{dist}\left(\frac{1}{\sigma_y^2(\lambda)}\,|\mathcal{Y}(\lambda,\mu)|^2, \left|\check{\mathbb{S}}_l(\lambda,\mu)\right|^2\right), \tag{4.36}$$

with $\sigma_y^2(\lambda) = \sum_{\mu=0}^{N_\mathrm{DFT}-1} |\mathcal{Y}(\lambda,\mu)|^2$. Since the speech codebook entries are normalized, a distance measure is required whose mapping and order is only dependent on the spectral shape and independent to a scaling of $\left|\check{\mathbb{S}}_l(\lambda,\mu)\right|^2$. Thus in contrast to the joint speech and noise codebook matching, the Itakura Saito distance is not applicable here. The relative power distance $\operatorname{dist}\Big|_{\mathrm{REL}}^{\mathcal{P},\widehat{\mathcal{P}}}$ is used as distance measure which turned out to be the best of the proposed metrics.

After determining the optimal codebook entry $\left|\check{\mathbb{S}}_{l_\mathrm{opt}}(\lambda,\mu)\right|^2$, the speech gain $\sigma_\mathrm{vad}$ which represents the speech power is calculated. The speech gain scales the found codebook entry $\left|\check{\mathbb{S}}_{l_\mathrm{opt}}(\lambda,\mu)\right|^2$ to the correct power resulting in the speech estimate $\left|\widehat{\mathcal{S}}(\lambda,\mu)\right| = \sigma_\mathrm{vad}(\lambda)\cdot\left|\check{\mathbb{S}}_{l_\mathrm{opt}}(\lambda,\mu)\right|$. From speech coding it is known that the optimal gain $\sigma_\mathrm{vad}(\lambda)$ can be found by minimizing the distance between the speech estimate $\widehat{\mathcal{S}}(\lambda,\mu)$ and the true speech $\mathcal{S}(\lambda,\mu)$ for the current frame $\lambda$. Since noisy speech is explicitly assumed as input to the algorithm, the gain derivation is, in contrast, carried out based on the distance between $\widehat{\mathcal{S}}(\lambda,\mu)$ and the noisy observation $\mathcal{Y}(\lambda,\mu)$. The relation of $\sigma_\mathrm{vad}$ to the true speech power is analyzed afterwards. Hence, the optimization is calculated in the MMSE sense for the current frame $\lambda$ according to[3]:

$$\operatorname{dist}\Big|_{\mathrm{MSE}}^{\mathcal{Y},\widehat{\mathcal{S}}} = \sum_{\mu=0}^{N_\mathrm{DFT}-1} \left(|\mathcal{Y}(\mu)| - \sigma_\mathrm{vad}\check{\mathbb{S}}_{l_\mathrm{opt}}(\mu)\right)^2 \stackrel{!}{=} \min. \tag{4.37}$$

[3]Since the codebook entries are real-valued and positiv (cf. Sec. 4.3), the absolute value operator of $\check{\mathbb{S}}_{l_\mathrm{opt}}$ is omitted in the following for the sake of clarity.

Building the partial derivation of dist $\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{S}}}$ with respect to $\sigma_{\text{vad}}$ and setting to zero yields the extremum of the distance given by

$$\frac{\partial}{\partial\sigma_{\text{vad}}}\left(\text{dist}\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{S}}}\right) = \sum_{\mu=0}^{N_{\text{DFT}}-1}\frac{\partial}{\partial\sigma_{\text{vad}}}\left(|\mathcal{Y}(\mu)| - \sigma_{\text{vad}}\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)\right)^2 \stackrel{!}{=} 0 \tag{4.38}$$

$$= \sum_{\mu=0}^{N_{\text{DFT}}-1} 2\cdot\left(|\mathcal{Y}(\mu)| - \sigma_{\text{vad}}\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)\right)\left(-\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)\right) \tag{4.39}$$

$$= -2\cdot\sum_{\mu=0}^{N_{\text{DFT}}-1}|\mathcal{Y}(\mu)|\,\check{\mathbb{S}}_{l_{\text{opt}}}(\mu) + 2\cdot\sigma_{\text{vad}}\sum_{\mu=0}^{N_{\text{DFT}}-1}\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu) \tag{4.40}$$

Hence, Eq. (4.40) can be transformed and $\sigma_{\text{vad}}$ is expressed as:

$$\sigma_{\text{vad}} = \frac{\sum_{\mu=0}^{N_{\text{DFT}}-1}|\mathcal{Y}(\mu)|\,\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)}{\sum_{\mu=0}^{N_{\text{DFT}}-1}\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu)} = \frac{\sum_{\mu=0}^{N_{\text{DFT}}-1}|\mathcal{S}(\mu)+\mathcal{N}(\mu)|\,\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)}{\sum_{\mu=0}^{N_{\text{DFT}}-1}\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu)}\,. \tag{4.41}$$

Since the second partial derivation of Eq. (4.40) with respect to $\sigma_{\text{vad}}$ yields

$$\frac{\partial^2}{\partial^2\sigma_{\text{vad}}}\left(\text{dist}\Big|_{\text{MSE}}^{\mathcal{Y}(\mu),\widehat{\mathcal{S}}(\mu)}\right) = 2\cdot\sum_{\mu=0}^{N_{\text{DFT}}-1}\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu) > 0, \tag{4.42}$$

and is greater than zero, the found extremum is in fact a minimum of dist $\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{S}}}$.

In the following it is analyzed to what extend $\sigma_{\text{vad}}$ is related to the speech power of a frame although the minimization is carried out on the noisy observation $|\mathcal{Y}(\mu)|^2 = |\mathcal{S}(\mu)+\mathcal{N}(\mu)|^2$. Since the denominator of Eq. (4.41) is independent of the noisy observation, the gain $\sigma_{\text{vad}}$ is mainly determined by the numerator. A further evaluation of the numerator of Eq. (4.41) leads to an expression describing the gain $\sigma_{\text{vad}}$ separated into a speech, a noise and a speech-noise (cross-term) dependent contribution of $\sigma_{\text{vad}}$ according to,

$$|\mathcal{S}(\mu)+\mathcal{N}(\mu)|\,\check{\mathbb{S}}_{l_{\text{opt}}}(\mu) = \sqrt{\begin{aligned}&|\mathcal{S}(\mu)|^2\,\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu) + |\mathcal{N}(\mu)|^2\,\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu)\\ &+2\,|\mathcal{S}(\mu)|\,|\mathcal{N}(\mu)|\,\check{\mathbb{S}}_{l_{\text{opt}}}^2(\mu)\cos(\vartheta_{\mathcal{S}}(\mu)-\vartheta_{\mathcal{N}}(\mu))\end{aligned}}, \tag{4.43}$$

where $\vartheta_{\mathcal{S}}(\mu)$ and $\vartheta_{\mathcal{N}}(\mu)$ denote the phase of speech and noise, respectively. First, two special cases with respect to the noisy input signal $\mathcal{Y}(\mu)$ are considered:

**Speech only** In the case where the input signal $\mathcal{Y}(\mu)$ consists only of speech, i. e., $\mathcal{Y}(\mu) = \mathcal{S}(\mu)$, a codebook entry $\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)$ with an excellent matching spectral

shape can be found and the gain $\sigma_{\text{vad}}$ yields

$$\sigma_{\text{vad},\mathcal{S}} = \frac{\sum_{\mu=0}^{N_{\text{DFT}}-1} |\mathcal{S}(\mu)|\, \check{\mathbb{S}}_{l_{\text{opt}}}(\mu)}{\sum_{\mu=0}^{N_{\text{DFT}}-1} \check{\mathbb{S}}^2_{l_{\text{opt}}}(\mu)}. \tag{4.44}$$

The gain $\sigma_{\text{vad},\mathcal{S}}$ represents the correct frame speech power $\sigma_s$ in a very good approximation since $|\mathcal{S}(\mu)|$ and $\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)$ are highly correlated over frequency.

**Noise only** In the opposite case where the input signal consists of noise only, i. e., $\mathcal{Y}(\mu) = \mathcal{N}(\mu)$, no suitable codebook entry is available in general. Thus, the spectral envelopes of the speech codebook and the observed noise frame differ significantly and the gain $\sigma_{\text{vad}}$ is expressed by

$$\sigma_{\text{vad},\mathcal{N}} = \frac{\sum_{\mu=0}^{N_{\text{DFT}}-1} |\mathcal{N}(\mu)|\, \check{\mathbb{S}}_{l_{\text{opt}}}(\mu)}{\sum_{\mu=0}^{N_{\text{DFT}}-1} \check{\mathbb{S}}^2_{l_{\text{opt}}}(\mu)}. \tag{4.45}$$

If the noise $\mathcal{N}(\mu)$ and the selected speech codebook entry $\check{\mathbb{S}}_{l_{\text{opt}}}$ have no significant spectral overlap (low correlation) $\sigma_{\text{vad},\mathcal{N}} \ll \sigma_{\text{vad},\mathcal{S}}$.

Combining the boundary cases described above models a realistic scenario including speech pauses as well as occurring background noise, i. e., $\mathcal{Y}(\mu) = \mathcal{S}(\mu) + \mathcal{N}(\mu)$. In this case the gain $\sigma_{\text{vad}}$ is determined by Eq. (4.43). Again, the noise component $\mathcal{N}(\mu)$ and the selected speech codebook entry $\check{\mathbb{S}}_{l_{\text{opt}}}$ as well as the noise component and the current speech component $\mathcal{S}(\mu)$ are assumed to be (almost) uncorrelated. Thus, Eq. (4.43) is dominated by the addend $|\mathcal{S}(\mu)|^2\, \check{\mathbb{S}}^2_{l_{\text{opt}}}(\mu)$ and the gain results in $\sigma_{\text{vad}} \approx \sigma_{\text{vad},\mathcal{S}}$. Hence, the speech gain $\sigma_{\text{vad}}$ is used as speech presence indicator.

However, in practical scenarios spectral overlaps between $|\mathcal{N}(\mu)|$, $\check{\mathbb{S}}_{l_{\text{opt}}}(\mu)$ and $|\mathcal{N}(\mu)|$, $|\mathcal{S}(\mu)|$ occur, i. e., speech and noise are not strictly uncorrelated. Thus, a noise floor in the gain $\sigma_{\text{vad}}$ depending on the noise signal is observed since $|\mathcal{N}(\mu)|^2\, \check{\mathbb{S}}^2_{l_{\text{opt}}}(\mu) > 0$ and $|\mathcal{S}(\mu)|\, |\mathcal{N}(\mu)|\, \check{\mathbb{S}}^2_{l_{\text{opt}}}(\mu) \cos(\vartheta_{\mathcal{S}}(\mu) - \vartheta_{\mathcal{N}}(\mu)) \neq 0$. Further post-processing is necessary to obtain a reliable VAD measure.

### 4.4.3 Speech Gain Post-Processing

Due to remaining noise and sudden outliers the speech gain $\sigma_{\text{vad}}(\lambda)$ fluctuates. Thus, in a first step of the post-processing, recursive smoothing is applied to the speech gain by:

$$\overline{\sigma^2}_{\text{vad}}(\lambda) = \left[\alpha_\sigma \sqrt{\overline{\sigma^2}_{\text{vad}}(\lambda - 1)} + (1 - \alpha_\sigma) \sqrt{\sigma^2_{\text{vad}}(\lambda)}\right]^2. \tag{4.46}$$

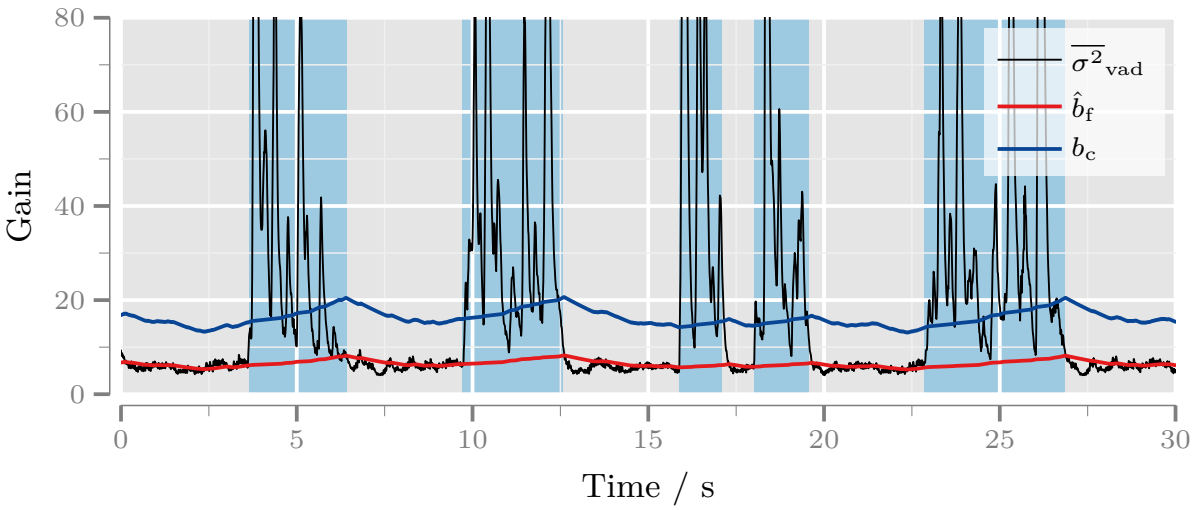



**Figure 4.7:** Example of smoothed gain $\overline{\sigma^2}_{\text{vad}}$, floor estimation $\hat{b}_{\text{f}}$ and ceiling estimation $b_{\text{c}}$ for male speech and jackhammer noise (SNR = 5 dB). A blue background indicates true speech activity.

The smoothing parameter $0 < \alpha_\sigma < 1$ determines the smoothing intensity and has to be chosen such that the system still adapts quickly to changes while larger fluctuations are leveled out. In order to control the on- and offset behavior of the voice activity individually, $\alpha_\sigma$ is chosen differently for rising or falling values of $\overline{\sigma^2}_{\text{vad}}(\lambda)$ according to:

$$\alpha_\sigma = \begin{cases} \alpha_{\sigma\uparrow} & \text{if } \sigma^2_{\text{vad}}(\lambda) \geq \overline{\sigma^2}_{\text{vad}}(\lambda - 1) \\ \alpha_{\sigma\downarrow} & \text{if } \sigma^2_{\text{vad}}(\lambda) < \overline{\sigma^2}_{\text{vad}}(\lambda - 1). \end{cases} \tag{4.47}$$

The smoothed speech gain $\overline{\sigma^2}_{\text{vad}}(\lambda)$ is a reliable speech presence indicator with a range of values in $[0, \infty)$. In Figure 4.7 an example of $\overline{\sigma^2}_{\text{vad}}(\lambda)$ is presented depicted by the black curve (——) for a male speech signal which is disturbed by highly non-stationary jackhammer noise with an input SNR of 5 dB. The highlighted background (■) indicates true speech activity. It is obvious that the speech gain $\overline{\sigma^2}_{\text{vad}}(\lambda)$ exhibits a noise floor during speech pauses and considerably higher values while speech is present (blue background, ■).

However, soft VAD values $v_{\text{soft}}(\lambda)$ between zero and one are desired. This requires further processing of $\overline{\sigma^2}_{\text{vad}}(\lambda)$ in order to remove the observed noise floor on the one hand and to map $\overline{\sigma^2}_{\text{vad}}(\lambda)$ into the desired range of values on the other hand. The principle is to provide the soft VAD by an interpolation between the noise floor and an adaptive upper bound, called the gain ceiling, which will be derived in the following.

The threshold between speech absence and arising speech presence is defined by the noise floor $b_{\text{f}}(\lambda)$ of $\overline{\sigma^2}_{\text{vad}}(\lambda)$. Hence a noise floor estimation $\hat{b}_{\text{f}}(\lambda)$ is necessary. Methods known from speech enhancement can be utilized to obtain the noise floor estimate $\hat{b}_{\text{f}}(\lambda)$ by tracing the baseline of the speech gain $\overline{\sigma^2}_{\text{vad}}(\lambda)$. This is possible

since the noise floor detection is similar to noise estimation in speech enhancement as the speech gain can be decomposed into a slowly changing noise floor component (the baseline of the speech gain) and the remaining strong fluctuations due to speech activity. Thus, a baseline tracing of the noise floor $b_\mathrm{f}(\lambda)$ similar to the noise estimator presented in Sec. 3.5 is carried out according to

$$\hat{b}_\mathrm{f}(\lambda) = \hat{b}_\mathrm{f}(\lambda - 1) + \operatorname{sign}\left(\overline{\sigma^2}_\mathrm{vad}(\lambda) - \hat{b}_\mathrm{f}(\lambda - 1)\right) \Delta'(\lambda). \tag{4.48}$$

In each frame, the noise floor $\hat{b}_\mathrm{f}(\lambda)$ is updated by shifting $\pm\Delta'(\lambda)$ in order to follow $\overline{\sigma^2}_\mathrm{vad}(\lambda)$ slowly.

Total speech presence, i. e., $v_\mathrm{soft}(\lambda) = 1$, is assumed if the speech gain $\overline{\sigma^2}_\mathrm{vad}(\lambda)$ exceeds the speech gain ceiling $b_\mathrm{c}(\lambda)$. The speech gain ceiling $b_\mathrm{c}(\lambda)$ is derived from the noise floor estimate $\hat{b}_\mathrm{f}(\lambda)$ and is defined with the adaptive factor $\eta(\lambda)$ according to

$$b_\mathrm{c}(\lambda) = \max\left(\eta(\lambda) \cdot \hat{b}_\mathrm{f}(\lambda),\, b_{c,\min}\right), \tag{4.49}$$

where $b_{c,\min}$ defines a minimum value for the ceiling $b_\mathrm{c}(\lambda)$ if no substantial noise floor is present. The factor $\eta(\lambda)$ is dependent on the speech gain SNR and bounded to $\eta_\mathrm{min} \leq \eta(\lambda) \leq \eta_\mathrm{max}$. Therefore, the local past of the speech gain and the noise floor estimate $\hat{b}_\mathrm{f}(\lambda)$ are stored in a sliding time window of length $T_\mathrm{w}$. The fraction of the mean of the sliding time windows provides a speech gain SNR estimate. It is only updated in phases where total speech presence is indicated, i. e., $v_\mathrm{soft}(\lambda) = 1$. In addition, the same recursive smoothing as for speech gain, c.f. Eq. (4.46), is applied which yields the speech gain SNR estimate $\eta(\lambda)$.

An example of the noise floor estimate $\hat{b}_\mathrm{f}(\lambda)$ and the ceiling $b_\mathrm{c}(\lambda)$ is depicted in Fig. 4.7 by the red (——) and blue curve (——), respectively. Finally, soft VAD values for $\overline{\sigma^2}_\mathrm{vad}$ between $\hat{b}_\mathrm{f}(\lambda)$ and $b_\mathrm{c}(\lambda)$ are interpolated linearly according to

$$v_\mathrm{soft}(\lambda) = \max\left(\min\left(\frac{\overline{\sigma^2}_\mathrm{vad}(\lambda) - \hat{b}_\mathrm{f}(\lambda)}{b_\mathrm{c}(\lambda) - \hat{b}_\mathrm{f}(\lambda)}, 1\right), 0\right). \tag{4.50}$$

Gains lower or equal to the noise floor are mapped to zero, whereas gains higher or equal to the ceiling $b_\mathrm{c}(\lambda)$ are clipped and mapped to one. The resulting soft values are robust to different noise floor levels in the speech gain which may result from low input SNR and varying noise types.

In order to be independent of system parameters like the sampling frequency $f_\mathrm{s}$ or the frame advance $L_A$, a relative shift $\Delta$ is introduced with dimension $\frac{\%}{time}$ such that $\frac{L_A}{f_\mathrm{s}}\Delta$ is the relative change per frame. Moreover, it is desirable to update the noise floor mainly in cases of speech absence, yielding the absolute shift to

$$\Delta'(\lambda) = \begin{cases} \frac{L_A}{f_\mathrm{s}} \cdot \Delta \cdot \hat{b}_\mathrm{f}(\lambda - 1) & \text{for } \overline{\sigma^2}_\mathrm{vad}(\lambda) \leq b_\mathrm{c}(\lambda - 1) \\ \frac{L_A}{f_\mathrm{s}} \cdot \Delta \cdot \hat{b}_\mathrm{f}(\lambda - 1) \cdot \beta_\mathrm{sp} & \text{for } \overline{\sigma^2}_\mathrm{vad}(\lambda) > b_\mathrm{c}(\lambda - 1). \end{cases} \tag{4.51}$$

If the speech gain exceeds the ceiling $b_\mathrm{c}$, total speech presence is assumed and the tracing speed is reduced by the factor $0 < \beta_\mathrm{sp} < 1$. It is not set to zero to prevent

| **Parameter** | **Settings** |
|---|---|
| Sampling frequency $f_\mathrm{s}$ | 16 kHz |
| Frame length $L_F$ | 320 ($\hat{=}$ 20 ms) |
| Frame advance $L_A$ | 160 ($\hat{=}$ 10 ms) |
| FFT length $N_\mathrm{DFT}$ | 512 (including zero-padding) |
| Frame overlap | 50 % ($\sqrt{\text{Hann}}$-window) |
| Speech codebook entries $L$ | 128 |
| Smoothing parameters $\alpha_{\sigma\uparrow}$ \| $\alpha_{\sigma\downarrow}$ | 0.8 \| 0.91 |
| Gain ceiling factor bounds $\eta_\mathrm{min}\|\eta_\mathrm{max}$ | 3 dB \| 15 dB |
| Ceiling minimum $b_{c,\mathrm{min}}$ | 3 |
| Relative shift $\Delta$ | $0.2\,\mathrm{s}^{-1}$ |
| Speech presence factor $\beta_\mathrm{sp}$ | $\frac{1}{4}$ |
| Speech gain SNR window length $T_\mathrm{w}$ | 0.1 s $\left(\hat{=}\left\lceil \frac{f_\mathrm{s}}{L_A} T_\mathrm{w} \right\rceil = 10 \text{ frames}\right)$ |

**Table 4.2:** Simulation system settings

that the system gets stuck in case of a completely wrong floor and ceiling estimation. Experiments confirmed that the relative shift over time $\Delta$ should be in the range between $\frac{0.2\,\%}{20\,\mathrm{ms}}$ and $\frac{0.8\,\%}{20\,\mathrm{ms}}$, i. e., the noise floor changes by the given percentage during 20 ms, a time period in which speech is considered to be stationary [Vary & Martin 2006].

If a binary VAD is desired, it can be calculated by a simple comparison with a threshold $0 < thr < 1$ according to

$$v_\mathrm{bin}(\lambda) = \begin{cases} 0 & \text{if } v_\mathrm{soft}(\lambda) < thr \\ 1 & \text{if } v_\mathrm{soft}(\lambda) \geq thr \ . \end{cases} \tag{4.52}$$

### 4.4.4 Evaluation

The proposed speech codebook based VAD system is assessed in a benchmark with four reference methods proposed by [Sohn et al. 1999], [Tan et al. 2010], [Ghosh et al. 2011] and the GSM VAD [ETSI Recommendation GSM 06.32 1996]. All algorithms except the GSM VAD provide soft VAD values. Since the objective scores require a binary VAD, Eq. (4.52) is utilized applying different thresholds varying between zero and one.

The parameters for the simulation are listed in Tab. 4.2. The speech codebook is trained according to Sec. 4.3 with randomly chosen speech files from the training set of the TIMIT database [Garofolo & Consortium 1993], resulting in a total training sequence length of 3073 s, cf. 4.3.4. The configuration of the remaining algorithms are chosen as suggested in [ETSI Recommendation GSM 06.32 1996; Ghosh et al. 2011; Sohn et al. 1999; Tan et al. 2010].

The benchmark is performed for all permutations of the following parameters:

- the input SNR ranges from from $-5\,\mathrm{dB}$ to $20\,\mathrm{dB}$ in $5\,\mathrm{dB}$ steps[4],
- 24 randomly chosen sentences belonging to 12 male and 12 female, randomly chosen speakers from the test set of the TIMIT database [Garofolo & Consortium 1993] are selected and concatenated. The test set is not included in the training set. Three seconds of silence are inserted at the beginning and the end of the sequence as well as between the sentences.
- The resulting 160 s of speech sequences are mixed with 11 types of noise (pink, jackhammer, canteen, wind, outside traffic, midsize car, inside train, train station, nature, pub noise, indoor soccer) from the ETSI database [ETSI EG 202 396-1 2009] resulting in 66 different noisy signals, respectively 176 minutes.
- The threshold for the binary VAD calculation varies for all tested soft VAD algorithms in 39 steps from zero to one.

An objective evaluation is performed which is based on a numerical comparison of the binary VAD $v_{\mathrm{bin}}(\lambda)$ with a ground truth binary VAD $v_{\mathrm{true}}(\lambda)$. As mentioned before, $v_{\mathrm{bin}}(\lambda)$ is provided by Eq. (4.52) applying different thresholds varying between zero and one for each soft VAD value. In this simulation, the clean speech and the scaled noise, from which the noisy signal is additively generated, are separately available. The objective measurement of active speech level according to ITU P.56 standardization [ITU-T Recommendation P.56 1993] provides a reliable binary VAD based on clean speech signals. Hence, this measure is applied to the clean speech signal in order to provide the ground truth reference VAD $v_{\mathrm{true}}(\lambda)$. The numerical evaluation is performed in terms of three VAD measures,

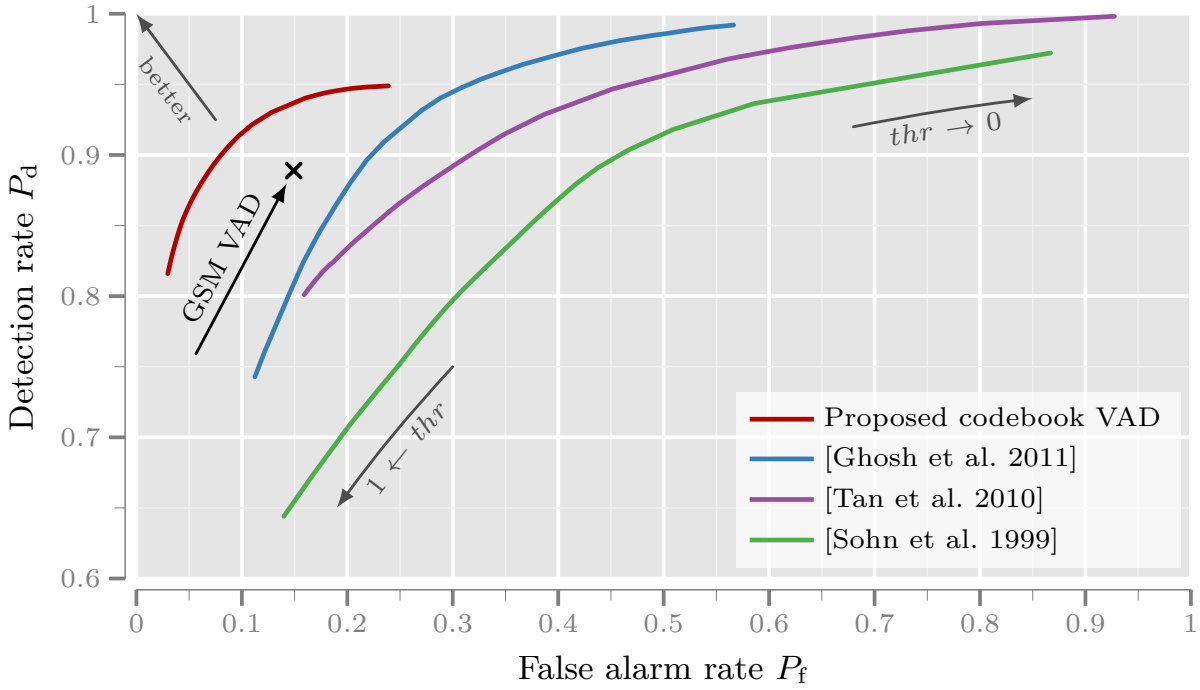
- *Accuracy rate* $P_{\mathrm{a}}$: Percentage of speech frames with correct VAD estimation;
- *Detection rate* (or true positive rate) $P_{\mathrm{d}}$: Fraction of active speech frames that are detected correctly;
- *False alarm rate* (or false positive rate) $P_{\mathrm{f}}$: Fraction of speech frames without speech that are classified erroneously as speech.

The objective measures are detailed in Appendix C.4. Note the first 160 frames, i. e., 1.6 s, are not included in the evaluation to ignore transient effects.

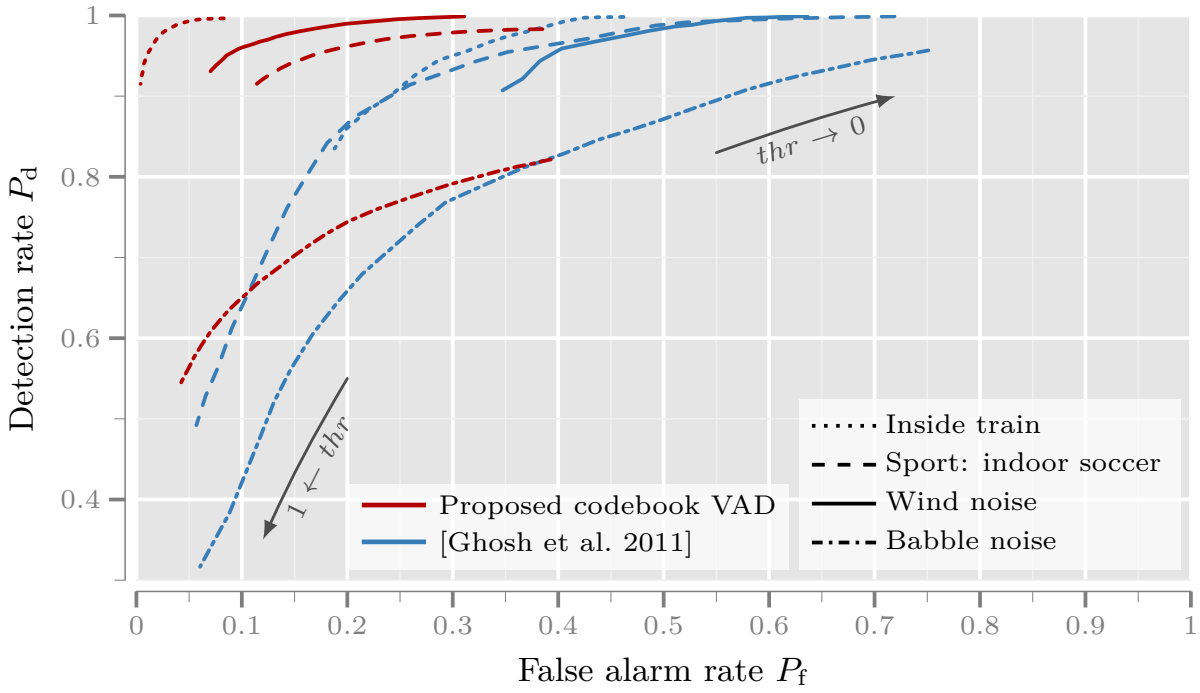
When applying a VAD, a compromise between detection-rate and false-alarm-rate has to be made by choosing an appropriate threshold. This compromise can be visualized, utilizing a ROC curve as a function of varying thresholds[5]. A fixed but arbitrary threshold corresponds to a specific point on the ROC curve. In Fig. 4.8 different aspects of the above mentioned compromise are detailed in terms of ROC curves. Fig. 4.8a presents a ROC curve which is generated by averaging

[4]The mixing procedure is detailed in Appendix C.1. Note that for the calculation of the scaling factor to adjust the input SNR only speech and noise signal sections with speech presence are considered.

[5]For the sake of clarity, the thresholds $thr \in \{0, 1\}$ are discarded in the presented figures.




**(a)** Averaged over all noise types




**(b)** Four exemplary noise types

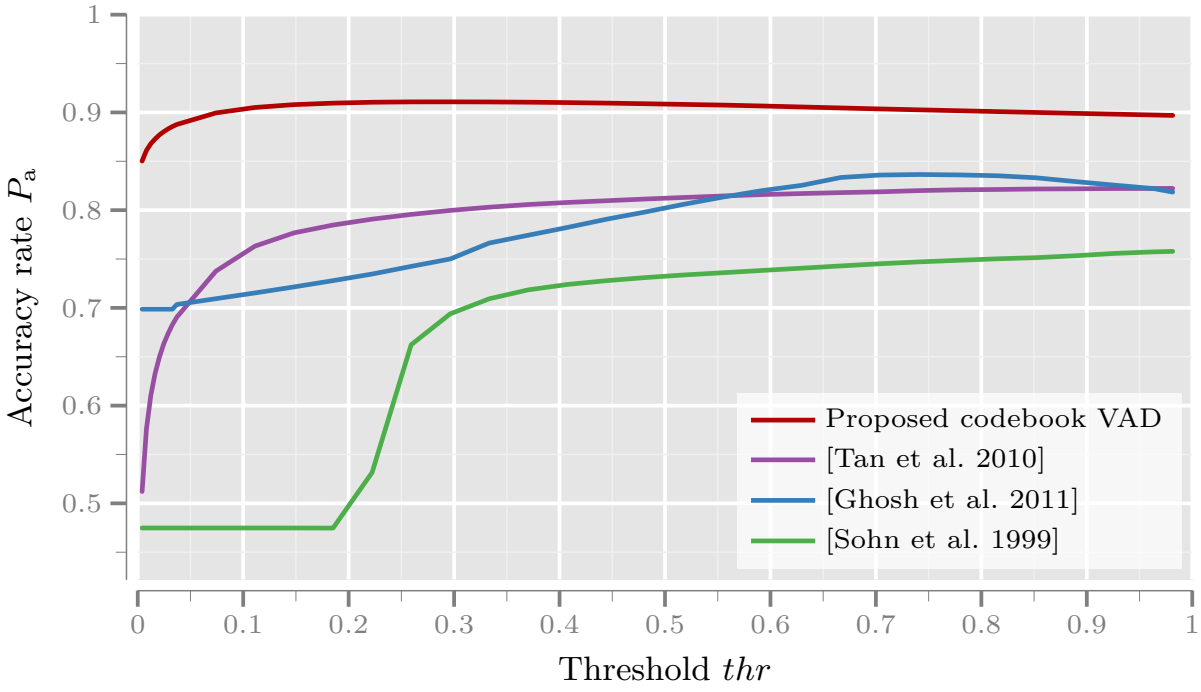
**Figure 4.8:** The upper part depicts the ROC curves for a varying threshold *thr*. The ROC curve for four exemplary types of noise is shown in the lower plot for the proposed and best reference algorithm [Ghosh et al. 2011] (——) at a varying threshold.

the objective scores detection-rate and false-alarm-rate for all permutations of the input SNR and noise types, separately for each threshold *thr*. Thus, it presents the achievable combinations of detection-rate and false-alarm-rate that result from varying the threshold. In addition, the binary GSM VAD [ETSI Recommendation GSM 06.32 1996] is depicted as reference and marked by the cross sign. For the proposed speech codebook VAD system (——), it is obvious that it holds the best relationship between the false-alarm-rate and the detection-rate. The false-alarm-rate never exceeds 24 % with a maximum detection-rate of 95 %. In order to achieve the same detection-rate, significantly higher false-alarm-rates of 32 % (Ghosh, ——), 45 % (Tan, ——) or 70 % (Sohn, ——) must be tolerated. However, the reference VAD systems (——, ——, ——) achieve a higher maximum-detection-rate compared to the proposed VAD (——), but at the expense of a significantly higher false-alarm-rate.

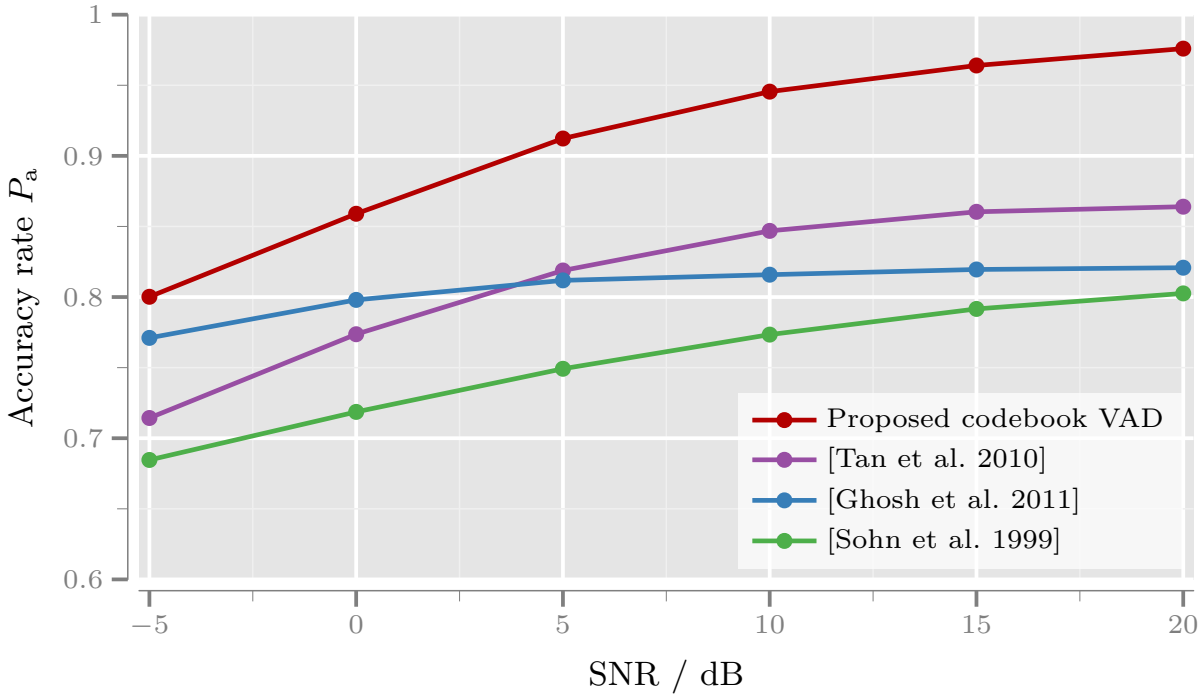
In Fig. 4.8b the averaged results are summarized for selected noise types: inside train (·····), indoor soccer (- - -), wind noise (——), and babble noise (-·-·-). Hence, the influence of stationary and instationary noise can be analyzed. For the sake of clarity, only the proposed VAD (——) and the best reference method, i. e., [Ghosh et al. 2011] (——), are visualized. The superior performance of the proposed codebook VAD is confirmed. For all noise types, the proposed method (——) yields the best performance. Moreover, the proposed algorithm performs well for stationary noise types, e. g., inside train (·····) as well as for instationary noise types like indoor soccer (- - -) and wind noise (——). Comparing wind noise, the proposed VAD (——) achieves approximately 30 % better false-alarm-rate than [Ghosh et al. 2011] (——) and similar detection-rate scores. However, a reliable voice detection during babble noise (-·-·-) is not possible because this sort of noise is very similar to the speech codebook entries. Hence, babble noise is frequently classified as speech, leading to a high false alarm rate, yet better than [Ghosh et al. 2011] (-·-·-).

The VAD accuracy is analyzed in Fig. 4.9. In order to examine the influence of the threshold, the results are averaged over the input SNR and noise types and plotted as a function of the threshold in Fig. 4.9a. Also in this VAD measure, the codebook based VAD (——) clearly provides the best scores over the complete threshold range, especially for thresholds up to 0.4. The advance to the second best algorithm [Ghosh et al. 2011] (——) for $thr > 0.4$ is approximately 10 % accuracy. One advantage of the proposed technique is the flatness of the accuracy measure. Because of that, it is possible to set any desired working point on the ROC curves depicted in Fig. 4.8 by adjusting the threshold without losing accuracy. The accuracy of the reference VAD algorithms (——, ——, ——) increases with the threshold. [Tan et al. 2010] (——) and [Ghosh et al. 2011] (——) achieve similar performance for $thr > 0.5$ while [Sohn et al. 1999] (——) has the worst performace over the complete threshold range, approximately 10 % worse than [Ghosh et al. 2011] (——).

To gain more insights into the behavior of the VAD algorithms at different SNR conditions, the best thresholds for each algorithm are selected from Fig. 4.9a.




**(a)** Averaged over the SNR




**(b)** Selected thresholds over the SNR

**Figure 4.9:** In the upper plot, the average accuracies for varying thresholds are depicted, while the lower plot shows the average accuracies over the SNR. For each algorithm, the most favorable threshold is chosen.

Using those thresholds, the accuracy is averaged for all noise types and each VAD algorithm. The results are depicted over the input SNR in Fig. 4.9b. As might be expected, the performance of all algorithms gets better with increasing SNR. Again, the proposed speech codebook based VAD (—●—) provides the best performance, starting with 80 % accuracy at −5 dB SNR and achieving nearly 100 % at 20 dB SNR. Comparing the reference VAD algorithms (—●—, —●—, —●—), [Ghosh et al. 2011] (—●—) performs best at low SNR values, while [Tan et al. 2010] (—●—) performs best for SNR values greater or equal than 5 dB. As indicated by the previous results in Fig. 4.9a [Sohn et al. 1999] (—●—) achieves the lowest accuracy over the complete SNR range.

With respect to the online noise codebook training process as described in Sec. 4.5, the new proposed speech codebook based VAD algorithm is well suited. It achieves the best scores in all VAD measures. Especially in the critical SNR range around 5 dB, the proposed VAD provides excellent accuracy rates in the range of 90 % and is thus 10 % better compared to the second best algorithm [Ghosh et al. 2011]. The new VAD does not rely on noise *a-priori* information, which makes it robust also to highly non-stationary noise and adverse SNR conditions, e. g., down to -5 dB. The new algorithm is characterized by higher detection-rates at a significantly lower false-alarm-rate compared to state-of-the-art systems [Ghosh et al. 2011; Sohn et al. 1999; Tan et al. 2010]. In addition, it is possible to adjust the compromise between a higher detection-rate versus a higher false-alarm-rate by changing the threshold without increasing the total number of miss-detections.

## 4.5 Online Noise Codebook Adaptation

Since the noise environment is unknown *a priori*, an online training and adaptation of the noise codebook is required. Hence, a training sequence acquired from the noisy observation $y(k)$ of the new and unknown noise type is necessary. Assuming speech pauses in $y(k)$ the training sequences can be found if speech is absent and a mismatch $\mathrm{Q}\,|_M\,(\lambda)$ during the codebook matching process is recognized. The mismatch $\mathrm{Q}\,|_M\,(\lambda)$ is defined between the noisy observation $\mathcal{Y}(\mu)$ and its codebook based approximation $\widehat{\mathcal{Y}}(\mu)$ and given for the current frame $\lambda$ by

$$\mathrm{Q}\,|_M\,(\lambda) = \mathrm{dist}\left( |\mathcal{Y}(\lambda,\mu)|^2 , \left|\widehat{\mathcal{Y}}(\lambda,\mu)\right|^2 \right). \tag{4.53}$$

Hence, the mismatch measure $\mathrm{Q}\,|_M\,(\lambda)$ describes the ability to estimate the noisy observation on the current state of the speech and noise codebooks.

Based on the training quality measures $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}$ and $\mathrm{Q}\,|_{\mathcal{C}_\mathcal{N}}$ of the respective codebooks, a lower bound of the mismatch $\mathrm{Q}\,|_M$ is estimated by

$$\mathrm{Q}\,|_{\mathcal{C}}\,(\lambda) = \frac{\sigma_s^2(\lambda)}{\sigma_s^2(\lambda)+\sigma_n^2(\lambda)} \cdot \mathrm{Q}\,|_{\mathcal{C}_\mathcal{S}}\,(\lambda) + \frac{\sigma_n^2(\lambda)}{\sigma_s^2(\lambda)+\sigma_n^2(\lambda)} \cdot \mathrm{Q}\,|_{\mathcal{C}_\mathcal{N}}\,(\lambda), \tag{4.54}$$

assuming that $\sigma_s^2$ and $\sigma_n^2$ are reliable also in the case of decent actual noise codebook

| Parameter | Settings |
|---|---|
| Training frames $L_T$ | 40 |
| VQ output size $M_\Delta$ | 4 codebook entries |
| Hangover VAD margin $L_H$ | 60 frames |
| Adaption margin $L_M$ | 40 frames |
| Hit rate $T$ | 80 % |
| Speech codebook size $L$ | 128 entries |
| Histogram window $L_W$ | 500 frames |

**Table 4.3:** Codebook algorithm parameters

mismatch. Comparing $\mathrm{Q}\,|_M$ with $\mathrm{Q}\,|_\mathcal{C}$, the effective codebook mismatch can be quantified incorporating the speech and noise codebook training quality.

In particular the following conditions must match in order to acquire new noise training sequences:

- Training frames must not contain speech, which requires a robust VAD measure. A robust speech codebook based VAD is presented in Sec. 4.4. In addition, a hangover frame distance to the last VAD frame of $L_H$ is introduced.
- A frame is classified as new noise type if the mismatch measure exceeds the threshold, i. e., $\mathrm{Q}\,|_M > \mathrm{Q}\,|_\mathcal{C}$.
- The distance measure evaluation of the last $L_T$ frames must have detected an unknown noise sound, i. e., $T$ percent of the last $L_T$ frames exceed the distance threshold $\mathrm{Q}\,|_M > \mathrm{Q}\,|_\mathcal{C}$.
- A safety margin between two adaptions of frame length $L_M$ has to be kept.

Given at least $L_T$ frames in the past which satisfy these conditions the same vector training as in Sec. 4.3 is utilized to obtain $M_\Delta$ new adaptive codebook entries which are then combined with the noise codebook. If the maximum defined noise codebook size $M = M_\square + M_\circ$ is exceeded where $M_\circ = r \cdot M_\Delta$ and $r$ is the number of codebook updates, the less used entries from the variable codebook part of the last $L_W$ frames are discarded.

### 4.5.1 Performance Example

An example of the online noise codebook adaptation is illustrated in Fig. 4.10. A noisy input signal is generated consisting of five, different six seconds long stationary and non-stationary noise types mixed with five male and female english speakers taken from the TIMIT database [Garofolo & Consortium 1993] at a SNR of 0 dB. Since the noise codebook is initialized with a single white noise codebook entry,

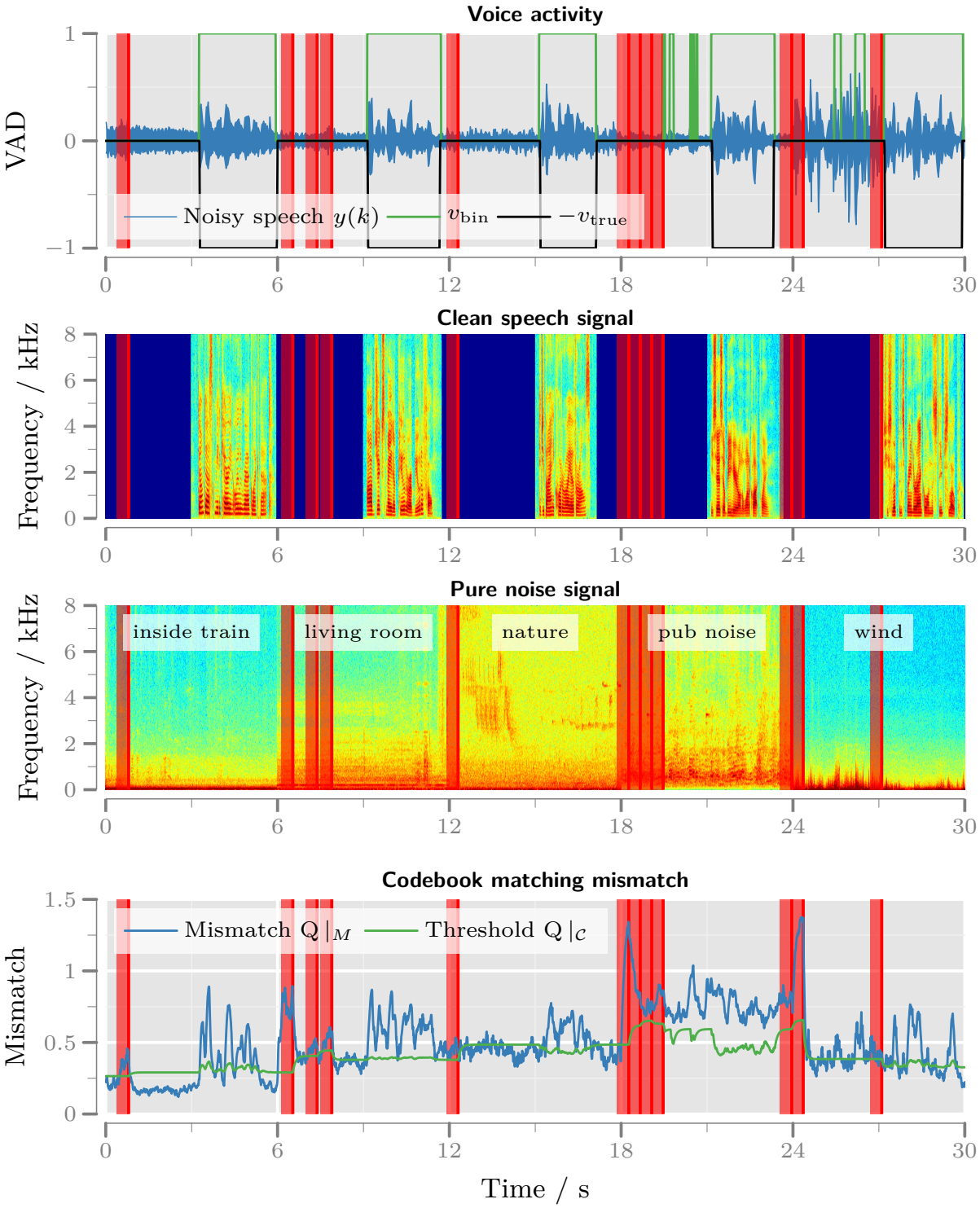



**Figure 4.10:** Example of online noise codebook adaption, learning five unknown noise types at 0 dB SNR. The upper plot shows the VAD performance where $v_{\text{bin}}$ is the computed VAD according to Sec. 4.4 and $v_{\text{true}}$ represents the ground truth VAD derived from the clean speech signal as reference. The lower plot presents the codebook mismatch measure and the mismatch threshold. Red areas ▮ indicate codebook update events.

it has to be adapted every six seconds. The VAD measure is provided by the proposed speech-codebook-based VAD algorithm detailed in the previous Sec. 4.4. The parameters for the simulation and the VAD setup are presented in Tab. 4.2. The codebook algorithm settings are summarized in Tab. 4.3, with a maximum noise codebook size of $M = 28$. The upper plot of Fig. 4.10 depicts the waveform of the noisy input signal to emphasize the performance of the VAD in terms of $v_{\text{bin}}$ (——) and the ground truth reference $v_{\text{true}}$ (——). The ground truth reference VAD is calculated according to ITU P.56 [ITU-T Recommendation P.56 1993] standardization from the clean speech signal which is available in the simulation system. Apart from "pub noise" (around 20 s) and "wind noise" (around 26 s), the presented new VAD algorithm provides reliable decisions. Vertical red lines indicate a codebook adaption which uses each time the past 40 frames as training sequence, indicated by the light red background. The lower plot depicts the codebook matching mismatch measure $\mathrm{Q}\,|_M$ (——) and the adaptive threshold $\mathrm{Q}\,|_{\mathcal{C}}$ (——). The spectrograms of the clean speech signal and the noise-only component are depicted for reference. It is obvious that each noise change is detected and the noise codebook is adapted accordingly. By means of the stationary noise types "inside train" and "nature" it is demonstrated, that a single adaptation of the noise codebook is sufficient while repeatedly adapting is necessary in the remaining cases, which reflects the fast changing characteristic of the noise signals. This observation is supported by the course of $\mathrm{Q}\,|_M$ (——). It is also apparent that adaptation takes exclusively place in speech pauses while a certain safety distance to speech activity frames is always maintained which avoids speech leaking into the noise codebook.

## 4.6 Summary and Conclusion

Most state-of-the-art noise reduction systems can be explained by means of noise estimation, spectral SNR estimation, and spectral weighting. In contrast, the codebook-based approach also incorporates a speech estimate. *A priori* knowledge about speech and noise allows to model and cope with highly non-stationary noise environments. A new modified decision-directed *a priori* SNR estimate $\widehat{\xi}_{\text{mod}}$ is proposed incorporating the codebook driven speech estimate.

In a first step, the concept of the proposed codebook speech and noise estimation is based on superposition of scaled speech and noise codebook entries. For the sake of speaker independence, the speech codebook consists of spectral envelopes, while the noise codebook comprises the full spectral resolution. Since no closed-form solution for optimal gain calculation of the speech and noise codebook entries exists, a brute force approach serves as reference codebook processing scheme. While the speech codebook is pre-trained in advance, the noise codebook is adapted to new noise types online. Thus, the system is independent of *a priori* knowledge on noise. Training vectors for online noise codebook updates are identified using a *voice activity detector* (VAD) and a codebook mismatch measure.

The VAD is realized as part of the codebook matching, but utilizes only *a priori* knowledge on speech. A speech power gain is provided in each frame. This gain

provides a reliable speech indicator and may contain a noise floor, especially at low SNRs. By means of a baseline tracing algorithm, known from noise reduction, the noise floor is removed and subsequently the gain is mapped to soft VAD values between zero and one. Instrumental measurements confirmed a consistent improvement in comparison to state-of-the-art systems [Ghosh et al. 2011; Sohn et al. 1999; Tan et al. 2010], resulting in better detection rates at significant lower false alarm rates, especially for low input SNR, e. g., $-5\,\mathrm{dB}$ SNR.

Although the noise codebook is updated online, it is not guaranteed that an appropriate codebook entry is available for each noisy observation. A noise codebook update is prevented, for example, if the ambient noise changes while speech is still present. In such cases, the noise estimation is restricted. Moreover, with respect to feasible applications, e. g., mobile phones, a significant complexity reduction is necessary which demands to replace the brute force codebook matching. This two remaining restrictions of codebook based speech and noise estimation are tackled in the next chapter.

# Chapter 5

# Information Combining

A major advantage of codebook driven speech and noise estimation is its ability to model highly non-stationary speech and noise processes. However, the estimation accuracy is limited by the *a priori* knowledge of the codebooks, especially with respect to missing noise *a priori* knowledge. Although the noise codebook is adaptive, it does not guarantee that an appropriate codebook entry is available for the current noisy observation. For example, if speech is present during a sequence where the noise type changes, a noise codebook update is not possible. Hence, in phases of missing *a priori* noise knowledge an alternative independent noise estimate, e. g., provided by a statistical noise estimator (cf., 3.4.1), should be considered. Moreover an *adaptive combination* of both noise estimates is desirable, resulting in a refined noise estimate.

In order to carry out this *adaptive combination*, a reliability measure is necessary. Utilizing the codebook speech estimate, it is possible to create all permutations of the speech estimate and the noise estimates which provide different estimates for the noisy observation. The distance between the different estimates and the noisy observation itself serves as reliability measure. Afterwards, the noise estimates are combined by a weighted sum according to the obtained distances separately for each frequency bin, which yields the refined final noise estimate. Given a second speech estimate, e. g., from the last enhanced frame of the speech enhancement system, the adaptive combining procedure can be generalized and provides in addition a refined speech estimate. Doing so, it is possible to recreate the spectral fine-structure in the final speech estimate. This *adaptive combination* procedure is called *information combining* in the following. By *information combining*, the speech and noise estimates are significantly improved.

As mentioned before, a substantial complexity reduction of the codebook matching process is necessary for the application of codebook based speech enhancement. Utilizing the *information combining* procedure, the codebook driven speech and noise estimates can be replaced by somewhat inaccurate estimates. Hence, the brute force search of the codebook matching is replaced by a cascade of gain shape estimates, which provides various speech and noise estimates. Compared with the brute force search, the cascade of gain shape estimates plus subsequent *information combining* allows a huge complexity reduction without notable quality loss. Thus, *information combining* improves estimation quality and provides complexity reduction.

The remainder of this chapter is organized as follows. The concept of *information combining* is outlined in Sec. 5.1. In Sec. 5.2 the joint estimation problem of speech and noise is formulated. While in Sec. 5.3 the constraints of combining speech and noise are detailed, the resulting estimation error is derived in Sec. 5.4. The minimization of the total estimation error is carried out in Sec. 5.5 and a closed-form solution for the total estimation error power is given in Sec. 5.6. On the basis of the developed *information combining* approach a complexity reduction of the codebook matching is outlined utilizing gain shape techniques in Sec. 5.7. The entire evaluation of codebook based speech enhancement is presented in Sec. 5.8 and conclusions are drawn in Sec. 4.6.

## 5.1 Concept of Information Combining

The term *information combining* is known from channel coding and information theory [Huber & Huettinger 2003; Land et al. 2005; Land & Huber 2006]. If the same data sequence is transmitted in parallel over independent channels or several times sequentially over the same channel, the independent observations can be combined at the receiver. The concept of *information combining* is to merge different independent estimates of a quantity into one in order to improve the overall estimation performance. The overall mutual information represents a combination of the mutual information of the independent estimates. The simplest realization of *information combining* would be the average of the different estimates. In general, averaging does not necessarily ensure an enhancement of the estimation performance. However, if reliability information related to the different estimates is available, the estimation quality is improved by applying an automatic weighted averaging of the estimates depending on their reliability yielding the refined final estimate. A special application of *information combining* is known from mobile radio transmission technology as maximum ratio diversity combining [Brennan 2003] and has been successfully used to improve the *signal-to-noise ratio* (SNR) given several antenna receiver signals. To the best knowledge of the author, no approach is known yet in the literature covering noise reduction that exploits the concept of *information combining* using different speech and noise estimates.

## 5.2 Estimation Problem Formulation

In the addressed example of *information combining* in mobile radio transmission, the noisy observations from the antenna receivers consist of the desired source target signal and additive noise. Hence, knowledge about the noise, e. g., in terms of the *signal-to-noise ratio* (SNR), enables to provide the required reliability measure for *information combining*.

In contrast to radio transmission technology or channel coding, two different estimation targets can be identified in speech enhancement, namely the speech and the noise signal. Furthermore, both estimation targets are included in the noisy

observation. Hence, it is more challenging to derive the estimation error, i. e., the reliability measure of either the speech or the noise estimate separately.

With respect to speech enhancement, the relation between the desired target quantity, its estimates and the associated estimation errors is given by

$$\mathcal{S}(\lambda,\mu) = \widehat{\mathcal{S}}_s(\lambda,\mu) + E_s^{\mathcal{S}}(\lambda,\mu) \qquad \text{with } 1 \leq s \leq N_s, \tag{5.1}$$

$$\mathcal{N}(\lambda,\mu) = \widehat{\mathcal{N}}_n(\lambda,\mu) + E_n^{\mathcal{N}}(\lambda,\mu) \qquad \text{with } 1 \leq n \leq N_n, \tag{5.2}$$

for speech and noise, respectively. The number of speech estimates is given by $N_s$, the number of noise estimates by $N_n$, and the particular subscripts $s, n$ indicate a specific estimate. The corresponding estimation errors are denoted by $E_s^{\mathcal{S}}(\lambda,\mu)$ and $E_n^{\mathcal{N}}(\lambda,\mu)$. In the following it is assumed that at least two speech estimates $\widehat{\mathcal{S}}_s(\lambda,\mu)$ and two noise estimates $\widehat{\mathcal{N}}_n(\lambda,\mu)$ exist.

In general, only the noisy observation including speech and noise is available. In consequence, two different estimation problems are included in the noisy observation. Hence, it is not possible to obtain the estimation errors $E_s^{\mathcal{S}}(\lambda,\mu)$ and $E_n^{\mathcal{N}}(\lambda,\mu)$ directly, associated with each of the different speech $\widehat{\mathcal{S}}_s(\lambda,\mu)$ and noise $\widehat{\mathcal{N}}_n(\lambda,\mu)$ estimates. However, combining the speech and noise estimates it is possible to compute several estimates of the noisy observation which are denoted by $\widehat{\mathcal{Y}}_i(\lambda,\mu)$. The subscript $i$ corresponds to all permutations of the speech and noise estimates and is detailed later. Employing the signal model

$$\mathcal{Y}(\lambda,\mu) = \mathcal{S}(\lambda,\mu) + \mathcal{N}(\lambda,\mu), \tag{5.3}$$

the noisy observation can be written using Eq. (5.1) and (5.2) as combination of the speech and noise estimates and their estimation errors according to

$$\mathcal{Y}(\lambda,\mu) = \underbrace{\widehat{\mathcal{S}}_s(\lambda,\mu) + \widehat{\mathcal{N}}_n(\lambda,\mu)}_{\widehat{\mathcal{Y}}_i(\lambda,\mu)} + \underbrace{E_s^{\mathcal{S}}(\lambda,\mu) + E_n^{\mathcal{N}}(\lambda,\mu)}_{E_i^{\mathcal{Y}}(\lambda,\mu)}. \tag{5.4}$$

Hence, the former desired target quantities speech and noise are combined and mapped into one target quantity. The new target quantity is given by the noisy observation $\mathcal{Y}(\lambda,\mu)$ which is estimated by

$$\widehat{\mathcal{Y}}_i(\lambda,\mu) = \widehat{\mathcal{S}}_s(\lambda,\mu) + \widehat{\mathcal{N}}_n(\lambda,\mu), \tag{5.5}$$

and the corresponding estimation error $E_i^{\mathcal{Y}}(\lambda,\mu) = \mathcal{Y}(\lambda,\mu) - \widehat{\mathcal{Y}}_i(\lambda,\mu)$ yields

$$E_i^{\mathcal{Y}}(\lambda,\mu) = E_s^{\mathcal{S}}(\lambda,\mu) + E_n^{\mathcal{N}}(\lambda,\mu). \tag{5.6}$$

In contrast to $E_s^{\mathcal{S}}(\lambda,\mu)$ and $E_n^{\mathcal{N}}(\lambda,\mu)$, the estimation error $E_i^{\mathcal{Y}}(\lambda,\mu)$ can be computed given the noisy observation $\mathcal{Y}(\lambda,\mu)$ and its estimate $\widehat{\mathcal{Y}}_i(\lambda,\mu)$.

For a complete specification of Eq. (5.5), a mapping rule from the noisy observation estimate index $i$ to the speech and noise estimate indices $s$ and $n$ is

necessary, which covers all permutations. Hence, the mapping is defined in terms of the estimates subscripts according to

$$i = \{1, \ldots, N_s \cdot N_n\}, \tag{5.7}$$

$$s = ((i-1) \bmod N_s) + 1, \tag{5.8}$$

$$n = \left\lceil \frac{i}{N_s} \right\rceil. \tag{5.9}$$

To provide a more intuitive overview, the permutation of the speech and noise estimates is visualized as matrix operation. With the noisy observation, the speech and noise estimates in vector notation denoted by

$$\widehat{\mathbb{Y}} = \left(\widehat{\mathcal{Y}}_1(\lambda,\mu), \ldots, \widehat{\mathcal{Y}}_{N_s \cdot N_n}(\lambda,\mu)\right)^{\mathsf{T}}, \tag{5.10}$$

$$\widehat{\mathbb{S}} = \left(\widehat{\mathcal{S}}_1(\lambda,\mu), \ldots, \widehat{\mathcal{S}}_{N_s}(\lambda,\mu)\right)^{\mathsf{T}}, \tag{5.11}$$

$$\widehat{\mathbb{N}} = \left(\widehat{\mathcal{N}}_1(\lambda,\mu), \ldots, \widehat{\mathcal{N}}_{N_n}(\lambda,\mu)\right)^{\mathsf{T}}, \tag{5.12}$$

the permutation is written as

$$\underset{(N_s \cdot N_n \times 1)}{\begin{pmatrix} \widehat{\mathbb{Y}} \end{pmatrix}} = \underset{(N_s \cdot N_n \times 1)}{\begin{pmatrix} \widehat{\mathbb{S}}_{N_s \times 1} \\ \widehat{\mathbb{S}}_{N_s \times 1} \\ \vdots \\ \widehat{\mathbb{S}}_{N_s \times 1} \end{pmatrix}} + \underset{(N_s \cdot N_n \times N_n)}{\begin{pmatrix} \mathbb{1}_{N_s \times 1} & \mathbb{0}_{N_s \times 1} & \cdots & \mathbb{0}_{N_s \times 1} \\ \mathbb{0}_{N_s \times 1} & \mathbb{1}_{N_s \times 1} & \cdots & \mathbb{0}_{N_s \times 1} \\ \vdots & \mathbb{0}_{N_s \times 1} & \ddots & \vdots \\ \mathbb{0}_{N_s \times 1} & \cdots & \mathbb{0}_{N_s \times 1} & \mathbb{1}_{N_s \times 1} \end{pmatrix}} \cdot \underset{(N_n \times 1)}{\begin{pmatrix} \widehat{\mathbb{N}} \end{pmatrix}} \tag{5.13}$$

where $\mathbb{0}_{N_s \times 1}$ and $\mathbb{1}_{N_s \times 1}$ denote the zero and one matrix, respectively. An example employing two speech and noise estimates results thus in four permutations according to[1]

$$\begin{pmatrix} \widehat{\mathcal{Y}}_1 \\ \widehat{\mathcal{Y}}_2 \\ \widehat{\mathcal{Y}}_3 \\ \widehat{\mathcal{Y}}_4 \end{pmatrix} = \begin{pmatrix} \widehat{\mathcal{S}}_1 + \widehat{\mathcal{N}}_1 \\ \widehat{\mathcal{S}}_2 + \widehat{\mathcal{N}}_1 \\ \widehat{\mathcal{S}}_1 + \widehat{\mathcal{N}}_2 \\ \widehat{\mathcal{S}}_2 + \widehat{\mathcal{N}}_2 \end{pmatrix} = \begin{pmatrix} \widehat{\mathcal{S}}_1 \\ \widehat{\mathcal{S}}_2 \\ \widehat{\mathcal{S}}_1 \\ \widehat{\mathcal{S}}_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \widehat{\mathcal{N}}_1 \\ \widehat{\mathcal{N}}_2 \end{pmatrix} \tag{5.14}$$

[1] The frame and frequency index $(\lambda, \mu)$ is omitted for the sake of brevity

## 5.3 Constraint Combining of Speech and Noise Estimates

Since only $E_i^{\mathcal{Y}}(\lambda,\mu)$ is measurable given the noisy observation $\mathcal{Y}(\lambda,\mu)$, the *information combining* of the estimates $\widehat{\mathcal{S}}_s(\lambda,\mu)$ and $\widehat{\mathcal{N}}_n(\lambda,\mu)$ is carried out indirectly in terms of $\widehat{\mathcal{Y}}_i(\lambda,\mu)$. Hence, the *information combining* of the different estimates $\widehat{\mathcal{Y}}_i(\lambda,\mu)$ is performed by a weighted averaging, utilizing the weights $c_i(\lambda,\mu)$,

$$\widehat{\mathcal{Y}}(\lambda,\mu) = \sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot\widehat{\mathcal{Y}}_i(\lambda,\mu) \tag{5.15}$$

$$= \underbrace{\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot\mathcal{Y}(\lambda,\mu)}_{\overset{!}{=}\mathcal{Y}(\lambda,\mu)} - \underbrace{\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot E_i^{\mathcal{Y}}(\lambda,\mu)}_{E^{\mathcal{Y}}(\lambda,\mu)}, \tag{5.16}$$

which yields the enhanced estimate $\widehat{\mathcal{Y}}(\lambda,\mu)$. In order to also model the enhanced estimate $\widehat{\mathcal{Y}}(\lambda,\mu)$ in terms of the noisy observation and an estimation error $E^{\mathcal{Y}}(\lambda,\mu)$, the weights are constraint by

$$\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu) = 1, \tag{5.17}$$

which yields the relation $\widehat{\mathcal{Y}}(\lambda,\mu) = \mathcal{Y}(\lambda,\mu) - E^{\mathcal{Y}}(\lambda,\mu)$. Using Eq. (5.5) in Eq. (5.15), the enhanced speech and noise estimates are finally given in terms of the weights by

$$\widehat{\mathcal{S}}(\lambda,\mu) = \sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot\widehat{\mathcal{S}}_s(\lambda,\mu), \quad \text{with } s = ((i-1) \bmod N_s) + 1, \tag{5.18}$$

$$\widehat{\mathcal{N}}(\lambda,\mu) = \sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot\widehat{\mathcal{N}}_n(\lambda,\mu), \quad \text{with } n = \left\lceil \frac{i}{N_s} \right\rceil. \tag{5.19}$$

In order to exploit the available *information* of the different estimates $\widehat{\mathcal{S}}_s(\lambda,\mu)$ and $\widehat{\mathcal{N}}_n(\lambda,\mu)$, the weights $c_i(\lambda,\mu)$ should be dependent on the measurable estimation error $E_i^{\mathcal{Y}}(\lambda,\mu)$. Moreover, the weights $c_i(\lambda,\mu)$ should minimize the total estimation error power $\left|E^{\mathcal{Y}}\right|^2$. Hence, an expression for the total estimation error power dependent on the weights $c_i(\lambda,\mu)$ is necessary.

## 5.4 Estimation Error

In this section an expression for the total estimation error power $\left|E^{\mathcal{Y}}\right|^2$ is derived which depends on the weights $c_i(\lambda,\mu)$. In addition it is analyzed to which extend

the total estimation error power $\left|E^{\mathcal{Y}}\right|^2$ is related to the estimation error power of the speech and noise estimates.

Assuming that $E^{\mathcal{Y}}(\lambda,\mu)$ is an ergodic process, the estimation error power is given by

$$\left|E^{\mathcal{Y}}\right|^2 = \mathbb{E}\left\{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\left|\mathcal{Y}(\lambda,\mu)-\widehat{\mathcal{Y}}(\lambda,\mu)\right|^2\right\} = \mathbb{E}\left\{\left|E^{\mathcal{Y}}(\lambda)\right|^2\right\}, \tag{5.20}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator with respect to time, i. e., the frame index $\lambda$. Utilizing Eq. (5.15) and (5.17) the error power $\left|E^{\mathcal{Y}}(\lambda)\right|^2$ for each frame $\lambda$ is formulated by

$$\left|E^{\mathcal{Y}}(\lambda)\right|^2 = \sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\left|\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot E_i^{\mathcal{Y}}(\lambda,\mu)\right|^2 \tag{5.21}$$

and the total error power $\left|E^{\mathcal{Y}}\right|^2$ yields

$$\left|E^{\mathcal{Y}}\right|^2 = \mathbb{E}\left\{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\left|\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot E_i^{\mathcal{Y}}(\lambda,\mu)\right|^2\right\} \tag{5.22}$$

$$= \sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\mathbb{E}\left\{\left|\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)\cdot E_i^{\mathcal{Y}}(\lambda,\mu)\right|^2\right\} \tag{5.23}$$

$$= \sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\left|E^{\mathcal{Y}}(\mu)\right|^2 \tag{5.24}$$

With respect to the minimizing procedure of the total estimation error power, it is sufficient to minimize the estimation error power with respect to the frequency index $\left|E^{\mathcal{Y}}(\mu)\right|^2$. Evaluating $\left|E^{\mathcal{Y}}(\mu)\right|^2$ and separating the auto estimation error terms from the double sum yields

$$\left|E^{\mathcal{Y}}(\mu)\right|^2 = \mathbb{E}\left\{\left(\sum_{i=1}^{N_s\cdot N_n} c_i(\lambda,\mu)E_i^{\mathcal{Y}}(\lambda,\mu)\right)\left(\sum_{j=1}^{N_s\cdot N_n} c_j(\lambda,\mu)E_i^{\mathcal{Y}}(\lambda,\mu)\right)^*\right\} \tag{5.25}$$

$$= \sum_{i=1}^{N_s\cdot N_n}\mathbb{E}\left\{|c_i(\lambda,\mu)|^2\cdot\left|E_i^{\mathcal{Y}}(\lambda,\mu)\right|^2\right\} + \tag{5.26}$$

$$\sum_{i=1}^{N_s\cdot N_n}\sum_{\substack{j=1\\ j\neq i}}^{N_s\cdot N_n}\mathbb{E}\left\{c_i(\lambda,\mu)c_j(\lambda,\mu)^*\cdot E_i^{\mathcal{Y}}(\lambda,\mu)E_j^{\mathcal{Y}}(\lambda,\mu)^*\right\}. \tag{5.27}$$

Assuming that the estimation errors of speech and noise are uncorrelated from each

other, i. e.,

$$0 = \mathbb{E}\left\{E_s^{\mathcal{S}}(\lambda,\mu)E_{\tilde{s}}^{\mathcal{S}}(\lambda,\mu)^*\right\} \qquad \forall\ s \neq \tilde{s} \tag{5.28}$$

$$0 = \mathbb{E}\left\{E_n^{\mathcal{N}}(\lambda,\mu)E_{\tilde{n}}^{\mathcal{N}}(\lambda,\mu)^*\right\} \qquad \forall\ n \neq \tilde{n} \tag{5.29}$$

$$0 = \mathbb{E}\left\{E_s^{\mathcal{S}}(\lambda,\mu)E_n^{\mathcal{N}}(\lambda,\mu)^*\right\} \qquad \forall\ s, n \tag{5.30}$$

$$\tag{5.31}$$

the estimation errors $E_i^{\mathcal{Y}}(\lambda,\mu)$ are also uncorrelated from each other, i. e.,

$$0 = \mathbb{E}\left\{E_i^{\mathcal{Y}}(\lambda,\mu)E_j^{\mathcal{Y}}(\lambda,\mu)^*\right\} \quad \forall\ j \neq i \tag{5.32}$$

and Eq. (5.25) simplifies to

$$\left|E^{\mathcal{Y}}(\mu)\right|^2 = \sum_{i=1}^{N_s \cdot N_n} c_i^2(\mu) \cdot \left|E_i^{\mathcal{Y}}(\mu)\right|^2. \tag{5.33}$$

Hence, the total estimation error power is basically a weighted sum over the error power of the different estimates $\widehat{\mathcal{Y}}_i(\lambda,\mu)$. A further evaluation of Eq. (5.33) utilizing Eq. (5.6) yields

$$\left|E_i^{\mathcal{Y}}(\mu)\right|^2 = \mathbb{E}\left\{\left|E_i^{\mathcal{Y}}(\lambda,\mu)\right|^2\right\} = \mathbb{E}\left\{\left|E_s^{\mathcal{S}}(\lambda,\mu) + E_n^{\mathcal{N}}(\lambda,\mu)\right|^2\right\} \tag{5.34}$$

$$= \mathbb{E}\left\{\left|E_s^{\mathcal{S}}(\lambda,\mu)\right|^2 + \left|E_n^{\mathcal{N}}(\lambda,\mu)\right|^2\right\} + \tag{5.35}$$

$$\mathbb{E}\left\{E_s^{\mathcal{S}}(\lambda,\mu)E_n^{\mathcal{N}}(\lambda,\mu)^* + E_s^{\mathcal{S}}(\lambda,\mu)^*E_n^{\mathcal{N}}(\lambda,\mu)\right\}$$

Since the estimation errors of speech and noise are assumed to be uncorrelated, the previous equation simplifies to

$$\left|E_i^{\mathcal{Y}}(\mu)\right|^2 = \left|E_s^{\mathcal{S}}(\mu)\right|^2 + \left|E_n^{\mathcal{N}}(\mu)\right|^2. \tag{5.36}$$

Finally, the total estimation error is expressed in terms of the speech and noise estimation errors given by

$$\left|E^{\mathcal{Y}}\right|^2 = \sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\left(\sum_{i=1}^{N_s \cdot N_n} c_i^2(\mu) \cdot \left(\left|E_s^{\mathcal{S}}(\mu)\right|^2 + \left|E_n^{\mathcal{N}}(\mu)\right|^2\right)\right), \tag{5.37}$$

with the indices given by $s = ((i-1) \bmod N_s) + 1$ and $n = \left\lceil \frac{i}{N_s} \right\rceil$.

## 5.5 Total Estimation Error Power Minimization

It was shown in the last section that the total estimation error power $\left|E^{\mathcal{Y}}\right|^2$ is dependent on the weights $c_i(\lambda,\mu)$ and the measurable estimation error $E_i^{\mathcal{Y}}(\lambda,\mu)$ of the noisy observation estimates. Moreover, minimizing the total estimation error

power $\left|E^{\mathcal{Y}}\right|^2$ also minimizes the estimation error power of the speech $E_s^{\mathcal{S}}(\lambda,\mu)$ and noise $E_n^{\mathcal{N}}(\lambda,\mu)$ estimates as related by Eq. (5.24) and (5.37).

Hence, the *information* of the different estimates $\widehat{\mathcal{S}}_s(\lambda,\mu)$ and $\widehat{\mathcal{N}}_n(\lambda,\mu)$ is optimally combined by minimizing the estimation error power $\left|E^{\mathcal{Y}}\right|^2$ of the enhanced estimate $\widehat{\mathcal{Y}}(\lambda,\mu)$ of the noisy observation. As mentioned before, it is sufficient to minimize $\left|E^{\mathcal{Y}}(\mu)\right|^2$ in Eq. (5.24). This yields a constrained optimization problem of the estimation error power $\left|E^{\mathcal{Y}}(\mu)\right|^2$, which can be solved by the *Lagrange multipliers* method [Bertsekas 1996; Bronstein et al. 1999]. In the following the frequency index $(\mu)$ will be omitted for the sake of brevity.

With Eq. (5.33) describing the estimation error power and the constraint given by Eq. (5.17), the Lagrange function is defined by

$$\Lambda(c_1,\ldots,c_{N_s\cdot N_n},\psi) = \sum_{i=1}^{N_s\cdot N_n} c_i^2 \left|E_i^{\mathcal{Y}}\right|^2 + \psi\cdot\left(1-\sum_{i=1}^{N_s\cdot N_n} c_i\right). \tag{5.38}$$

Building the partial derivation of $\Lambda(c_1,\ldots,c_{N_s\cdot N_n},\psi)$ with respect to the weights $c_i$ yields

$$\frac{\partial\Lambda c_1,\ldots,c_{N_s\cdot N_n},\psi}{\partial c_i} = 2c_i\left|E_i^{\mathcal{Y}}\right|^2 - \psi, \quad \text{with } 1\le i\le N_s\cdot N_n, \tag{5.39}$$

and the partial derivation with respect to the Lagrange multiplier $\psi$ results in

$$\frac{\partial\Lambda c_1,\ldots,c_{N_s\cdot N_n},\psi}{\partial\psi} = 1-\sum_{i=1}^{N_s\cdot N_n} c_i. \tag{5.40}$$

Setting the partial derivations Eq. (5.39) and Eq. (5.40) to zero and equating yields the following system of equations,

$$c_i = \frac{\psi}{2\left|E_i^{\mathcal{Y}}\right|^2} \quad \text{with } 1\le i\le N_s\cdot N_n. \tag{5.41}$$

Using Eq. (5.41) in Eq. (5.40) and solving the equation with respect to $\psi$ yields

$$\psi = \frac{2}{\sum\limits_{i=1}^{N_s\cdot N_n}\frac{1}{\left|E_i^{\mathcal{Y}}\right|^2}}. \tag{5.42}$$

Substituting $\psi$ in Eq. (5.41) yields the weights $c_i$ according to

$$c_i = \frac{1}{\left|E_i^{\mathcal{Y}}\right|^2\sum\limits_{j=1}^{N_s\cdot N_n}\frac{1}{\left|E_j^{\mathcal{Y}}\right|^2}}. \tag{5.43}$$

Since the second partial derivations of Eq.(5.38) with respect to the weights $c_i$

$$\frac{\partial^2 \Lambda c_1, \ldots, c_{N_s \cdot N_n}, \psi}{\partial^2 c_i} = 2 \left| E_i^{\mathcal{Y}} \right|^2 \quad \text{with } 1 \leq i \leq N_s \cdot N_n \tag{5.44}$$

and with respect to the Lagrange multiplier $\psi$

$$\frac{\partial^2 \Lambda c_1, \ldots, c_{N_s \cdot N_n}, \psi}{\partial^2 \psi} = 0 \tag{5.45}$$

are greater or equal to zero, the found extremum is in fact an absolute minimum. Adding the frequency index again to Eq. (5.43), the weights are finally given by,

$$c_i(\mu) = \frac{1}{\left| E_i^{\mathcal{Y}}(\mu) \right|^2 \sum_{j=1}^{N_s \cdot N_n} \frac{1}{\left| E_j^{\mathcal{Y}}(\mu) \right|^2}} \tag{5.46}$$

Hence, the determined weights according to Eq. (5.46) minimize the total error power $\left| E^{\mathcal{Y}} \right|^2 = \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left| E^{\mathcal{Y}}(\mu) \right|^2$ and thereby also the total error power of the speech and noise estimates.

## 5.6 Total Estimation Error Power

A closed solution for the total estimation error power $\left| E^{\mathcal{Y}} \right|^2$ is found by substituting Eq. (5.46) in Eq. (5.33) which yields the resulting error power after weighted averaging according to

$$\left| E_{\min}^{\mathcal{Y}}(\mu) \right|^2 = \sum_{i=1}^{N_s \cdot N_n} \left( \frac{1}{\left| E_i^{\mathcal{Y}}(\mu) \right|^2 \sum_{j=1}^{N_s \cdot N_n} \frac{1}{\left| E_j^{\mathcal{Y}}(\mu) \right|^2}} \right)^2 \left| E_i^{\mathcal{Y}}(\mu) \right|^2 \tag{5.47}$$

$$= \sum_{i=1}^{N_s \cdot N_n} \frac{1}{\left| E_i^{\mathcal{Y}}(\mu) \right|^2 \left( \frac{1}{\sum_{j=1}^{N_s \cdot N_n} \frac{1}{\left| E_j^{\mathcal{Y}}(\mu) \right|^2}} \right) \left( \frac{1}{\sum_{j=1}^{N_s \cdot N_n} \frac{1}{\left| E_j^{\mathcal{Y}}(\mu) \right|^2}} \right)} \tag{5.48}$$

$$= \sum_{i=1}^{N_s \cdot N_n} c_i(\mu) \frac{1}{\sum_{j=1}^{N_s \cdot N_n} \frac{1}{\left| E_j^{\mathcal{Y}}(\mu) \right|^2}} \tag{5.49}$$

$$\left| E_{\min}^{\mathcal{Y}} \right|^2 = \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left| E_{\min}^{\mathcal{Y}}(\mu) \right|^2 \tag{5.50}$$

$$= \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \frac{1}{\sum_{i=1}^{N_s \cdot N_n} \frac{1}{\left| E_i^{\mathcal{Y}}(\mu) \right|^2}} = \frac{1}{\sum_{i=1}^{N_s \cdot N_n} \frac{1}{\left| E_s^{\mathcal{S}}(\mu) \right|^2 + \left| E_n^{\mathcal{N}}(\mu) \right|^2}}, \tag{5.51}$$

with $s = ((i-1) \bmod N_s) + 1$ and $n = \left\lceil \frac{i}{N_s} \right\rceil$. By a closer examination of Eq. (5.51) it turns out that the equation describes a similar relation compared to the overall resistance of a parallel circuit of resistors. Hence, the total estimation error power after weighting $\left|E_{\min}^{\mathcal{Y}}\right|^2$ is less than the minimum of the error power of the individual estimates $\left|E_i^{\mathcal{Y}}\right|^2 = \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left|E_i^{\mathcal{Y}}(\mu)\right|^2$, i. e., $\left|E_{\min}^{\mathcal{Y}}\right|^2 < \left|E_i^{\mathcal{Y}}\right|^2 \forall\ i$.

Since the estimation error $\left|E_i^{\mathcal{Y}}\right|^2$ varies in practice over the time, the weights $c_i(\mu)$ are thus calculated in each frame $\lambda$.

## 5.7 Complexity Reduction

With respect to speech enhancement applications, e. g., for a mobile phones scenario, a dramatic complexity reduction of the codebook matching process is necessary. According to Sec. 4.1.2 Eq. (4.11) the brute force search, considering a speech and a noise codebook, consists of all combinations of the three parameters $l, m, \sigma_n$. Hence the computational effort grows exponentially with any of the parameters. Techniques known from gain shape *vector quantizer* (VQ) to determine the codebook entries and the gains independently are not applicable here. The optimization of the gains for a fixed but arbitrary combination of speech and noise codebook entries does not guarantee positive gains which violates the model assumption, i. e., the gains represents the short-term power of noise and speech (cf. Appendix E).

However, utilizing a *voice activity detector* (VAD) and the *information combining* as explained in Chap. 5, it is possible to replace the brute force codebook matching partly by a gain shape VQ or a cascade of gain shape VQs. The concept of gain shape VQ is the determination of the spectral shape using a gain normalized codebook in a first step and subsequently the calculation of the corresponding gain in a second step. The employed VQ is similar to the one introduced in Sec. 4.4.2. The optimal codebook entry for the current frame $\lambda$ of either speech $l_{\mathrm{opt}}$ or noise $m_{\mathrm{opt}}$ can be found by minimizing

$$\arg\min_{m} \operatorname{dist}\left(\frac{1}{\sigma_y^2(\lambda)}\left|\mathcal{Y}(\lambda,\mu)\right|^2, \left|\mathbb{N}_m(\lambda,\mu)\right|^2\right), \tag{5.52}$$

$$\arg\min_{l} \operatorname{dist}\left(\frac{1}{\sigma_y^2(\lambda)}\left|\mathcal{Y}(\lambda,\mu)\right|^2, \left|\mathbb{S}_l(\lambda,\mu)\right|^2\right), \tag{5.53}$$

with $\sigma_y^2(\lambda) = \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left|\mathcal{Y}(\lambda,\mu)\right|^2$. Since the codebook entries are gain normalized, a distance measure is required whose mapping and order is only dependent on the spectral shape and is independent to a scaling of $\left|\mathbb{S}_l(\lambda,\mu)\right|^2$ or $\left|\mathbb{N}_m(\lambda,\mu)\right|^2$. Thus, the Itakura Saito distance is not applicable here in contrast to the joint brute force speech and noise codebook matching. The relative power distance $\operatorname{dist}_{\mathrm{REL}}^{\mathcal{P},\widehat{\mathcal{P}}}$ is used as distance measure instead which turned out to be the best metric.

After determining the optimal codebook entry of either $\left|\mathbb{N}_{m_{\mathrm{opt}}}(\lambda,\mu)\right|^2$ or $\left|\mathbb{S}_{l_{\mathrm{opt}}}(\lambda,\mu)\right|^2$, the corresponding gain $\sigma_n$ or $\sigma_s$ which represents the noise or

speech power is calculated. The gain scales the found codebook entry to the correct power, resulting in the estimate $\left|\widehat{\mathcal{N}}(\lambda,\mu)\right| = \sigma_n(\lambda)\cdot\left|\mathbb{N}_{m_{\text{opt}}}(\lambda,\mu)\right|$ or $\left|\widehat{\mathcal{S}}(\lambda,\mu)\right| = \sigma_s(\lambda)\cdot\left|\mathbb{S}_{l_{\text{opt}}}(\lambda,\mu)\right|$. The optimal gain can be found by minimizing the distance between the selected codebook entry and the noisy observation $\mathcal{Y}(\lambda,\mu)$. Hence, the optimization is calculated in the *minimum mean-square error* (MMSE) sense for the current frame $\lambda$ according to:

$$\text{dist}\left|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{N}}} = \sum_{\mu=0}^{N_{\text{DFT}}-1}\left(|\mathcal{Y}(\mu)| - \sigma_n\mathbb{N}_{m_{\text{opt}}}(\mu)\right)^2 \stackrel{!}{=} \min\,, \tag{5.54}$$

$$\text{dist}\left|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{S}}} = \sum_{\mu=0}^{N_{\text{DFT}}-1}\left(|\mathcal{Y}(\mu)| - \sigma_s\mathbb{S}_{l_{\text{opt}}}(\mu)\right)^2 \stackrel{!}{=} \min\,. \tag{5.55}$$

and results in

$$\sigma_n = \frac{\sum\limits_{\mu=0}^{N_{\text{DFT}}-1}|\mathcal{Y}(\mu)|\,\mathbb{N}_{m_{\text{opt}}}(\mu)}{\sum\limits_{\mu=0}^{N_{\text{DFT}}-1}\left|\mathbb{N}_{m_{\text{opt}}}(\mu)\right|^2}\,, \tag{5.56}$$

$$\sigma_s = \frac{\sum\limits_{\mu=0}^{N_{\text{DFT}}-1}|\mathcal{Y}(\mu)|\,\mathbb{S}_{l_{\text{opt}}}(\mu)}{\sum\limits_{\mu=0}^{N_{\text{DFT}}-1}\left|\mathbb{S}_{l_{\text{opt}}}(\mu)\right|^2}. \tag{5.57}$$

Further details of the derivation and an analysis to what extend $\sigma_s, \sigma_n$ are related to the true speech or noise power are included in Sec. 4.4.2.

### 5.7.1 Using VAD

In a first step the computational complexity is reduced in phases of speech absence. Utilizing a binary VAD $v_{\text{bin}}(\lambda)$, the brute force search according to Sec. 4.1.2 employing a speech and a noise codebook can be replaced by a gain shape VQ employing a noise codebook. The corresponding block diagram is depicted in Fig. 5.1 using configuration ⓐ in the codebook matching block. While speech is present, the brute force codebook matching block is selected, i. e., $\widehat{\mathcal{S}}_{\text{CB}}(\lambda,\mu) = \widehat{\mathcal{S}}_{\text{BF}}(\lambda,\mu)$ and $\widehat{\mathcal{N}}_{\text{CB}}(\lambda,\mu) = \widehat{\mathcal{N}}_{\text{BF}}(\lambda,\mu)$. In the opposite case, a noise codebook gain shape approach is utilized to determine the noise estimate, $\widehat{\mathcal{N}}_{\text{CB}}(\lambda,\mu) = \widehat{\mathcal{N}}_{\text{GS}}(\lambda,\mu)$. Since speech is absent, the speech estimate is set to zero, i. e., $\widehat{\mathcal{S}}_{\text{CB}}(\lambda,\mu) = 0$.

### 5.7.2 Employing Information Combining

Using the information from the VAD, the brute force search, using a speech and a noise codebook, is only necessary in phases of speech activity. Hence, a

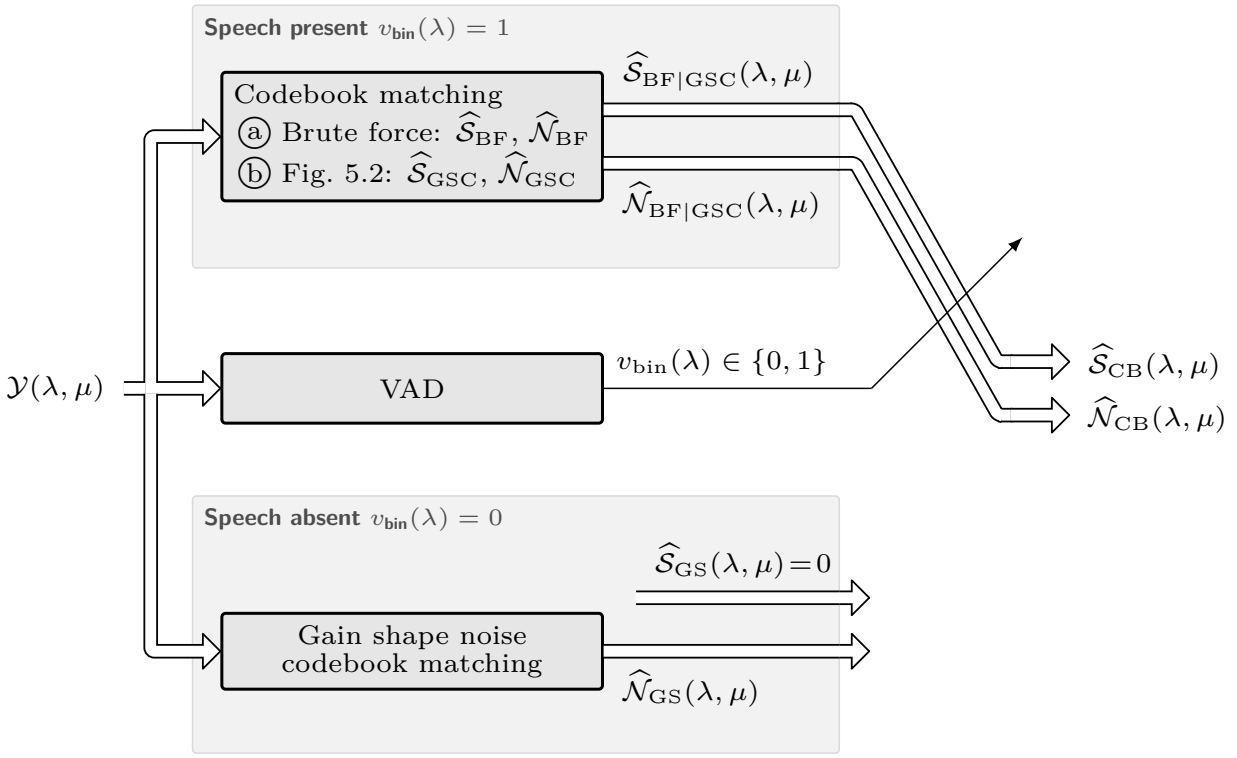



**Figure 5.1:** Complexity reduction based on VAD

further complexity reduction is necessary during speech activity, i. e., the codebook matching block utilizing the brute force search in Fig. 5.1 (configuration ⓐ) yielding $\widehat{\mathcal{N}}_{\text{BF}}(\lambda, \mu)$ and $\widehat{\mathcal{S}}_{\text{BF}}(\lambda, \mu)$ has to be replaced.

With respect to gain shape VQ, two scenarios exist which allow to replace the brute force codebook matching. Given a very high SNR, i. e., $\mathcal{N}(\lambda, \mu)$ very close to zero, the brute force search can be replaced by gain shape VQ utilizing a speech codebook and setting the noise estimate to zero. In the opposite case, where the SNR is very low, a gain shape VQ employing a noise codebook is utilized and the speech estimate is set to zero. Assuming the theoretical special case of orthogonal speech and noise shapes (no spectral overlap) in each frame $|\mathcal{Y}(\lambda, \mu)|^2$, two gain shape VQ units, employing a speech and a noise codebook, can be used to estimate a reliable speech and noise estimate.

Concerning realistic scenarios, neither orthogonal speech and noise shapes nor infinite high or low SNR can be expected. However, depending on the SNR two different cascades of gain shape VQ as depicted in Fig. 5.2 provide suitable speech and noise estimates depending on the SNR. The upper cascade GS1 is subject to a gain shape VQ unit utilizing a noise codebook. Subsequently, the determined noise estimate $\widehat{\mathcal{N}}_{\text{GS1}}(\lambda, \mu)$ is subtracted from the noisy observation and bounded to be greater or equal to zero. Afterwards, a gain shape VQ employing a speech codebook is applied to determine the speech estimate $\widehat{\mathcal{S}}_{\text{GS1}}(\lambda, \mu)$. In the cascade GS2 the codebooks are interchanged.

Doing so, a sequential optimization of first the noise estimate and following the

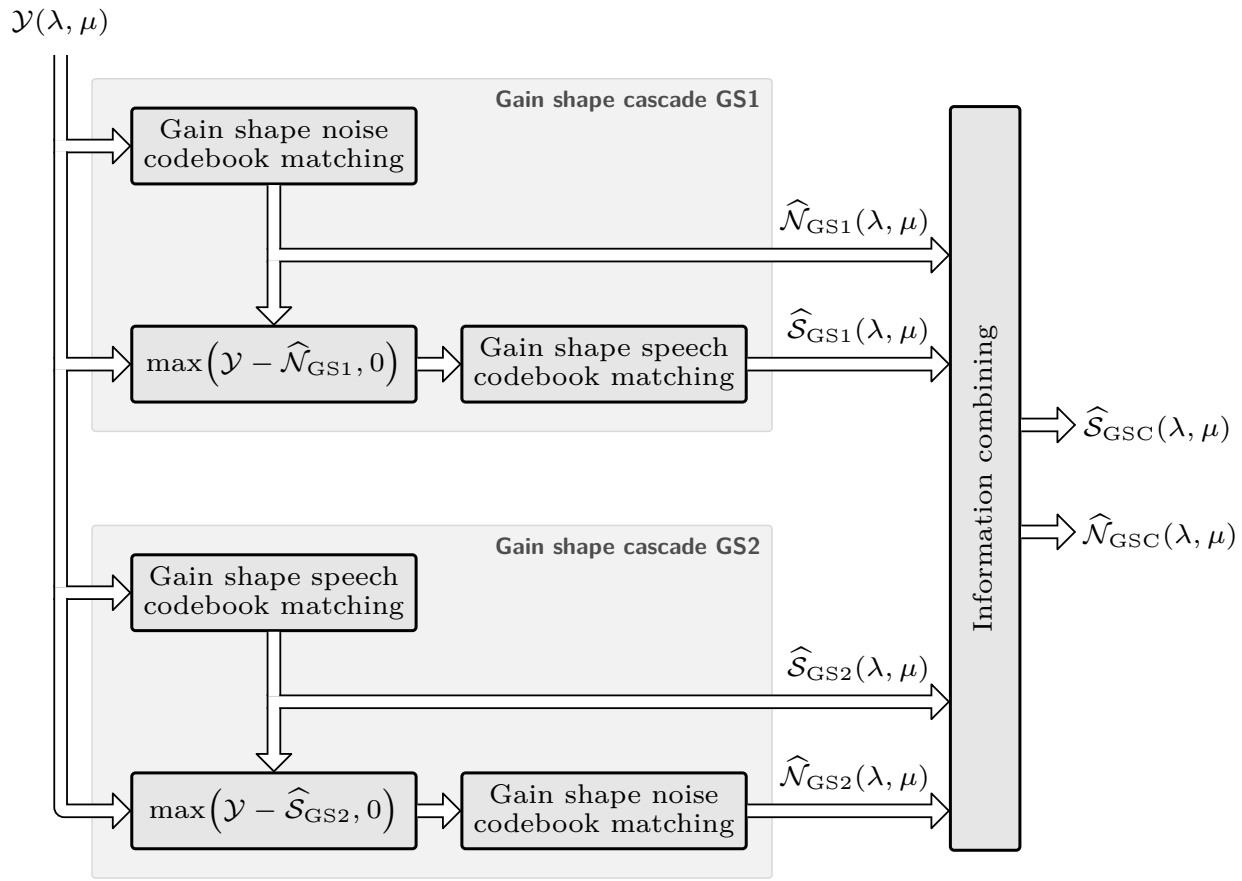



**Figure 5.2:** Complexity reduction employing sequential optimization and information combining

speech estimate is carried out in GS1 and with respect to first the speech estimate followed by the noise estimate in GS2, respectively. Note that the determined speech as well as the noise estimates exhibit different reliability. Compared with the brute force search, a sub-optimal solution is thereby provided in general. However, cascade GS1 is expected to provide a reliable noise estimate $\widehat{\mathcal{N}}_{\mathrm{GS1}}(\lambda,\mu)$ given a low SNR, while cascade GS2 provides a reliable speech estimate $\widehat{\mathcal{S}}_{\mathrm{GS2}}(\lambda,\mu)$ for high SNR. Finally, the different speech $\widehat{\mathcal{S}}_{\mathrm{GS1}}(\lambda,\mu)$, $\widehat{\mathcal{S}}_{\mathrm{GS2}}(\lambda,\mu)$ and noise $\widehat{\mathcal{N}}_{\mathrm{GS1}}(\lambda,\mu)$, $\widehat{\mathcal{N}}_{\mathrm{GS2}}(\lambda,\mu)$ estimates are merged independently for each frequency bin utilizing the *information combining* approach introduced in Chap. 5. The final estimate of speech is denoted by $\widehat{\mathcal{S}}_{\mathrm{GSC}}(\lambda,\mu)$ and the noise estimate is given by $\widehat{\mathcal{N}}_{\mathrm{GSC}}(\lambda,\mu)$.

Applying the sequential optimization with subsequent *information combining* (Fig. 5.2) instead of the brute force approach (Sec. 4.1.2) reduces the complexity from $\mathcal{O}(M \cdot L)$ to $\mathcal{O}(M + L)$.

In order to obtain the complete complexity reduction, the codebook matching block in Fig. 5.1 is set to configuration (b). Hence, the speech $\widehat{\mathcal{S}}_{\mathrm{GSC}}(\lambda,\mu)$ and noise $\widehat{\mathcal{N}}_{\mathrm{GSC}}(\lambda,\mu)$ estimates employing sequential optimization and *information combining* are used instead of the joint brute force estimates $\widehat{\mathcal{S}}_{\mathrm{BF}}(\lambda,\mu)$ and $\widehat{\mathcal{N}}_{\mathrm{BF}}(\lambda,\mu)$.

## 5.8 Evaluation

The codebook driven speech and noise estimation is evaluated in different configurations by means of a noise reduction system. Different objective speech enhancement scores serve as performance measures. Since the performance of the speech enhancement system may strongly depend on the individual performance of the respective sub systems, at first, a reference codebook system employing the brute force search as explained in Sec. 4.1.2 and fixed speech and noise codebooks is defined. Using this reference platform,

- the new modified decision-directed SNR (cf., Sec. 4.2) as well as
- the *information combining* method (cf., Chap. 5, Sec. 5.5)

are analyzed. In the second part of the evaluation, the reference system is compared with the proposed system, comprising all features, i. e., the new modified decision-directed SNR, the adaptive online noise codebook learning, the *information combining*, and the complexity reduction.

The benchmarks are performed for all noisy input signals which are obtained from the permutation of the following parameters:

- The input SNR varies from $-10\,\mathrm{dB}$ to $35\,\mathrm{dB}$ in $5\,\mathrm{dB}$ steps[2].
- 30 randomly chosen sentences, spoken by 15 male and 15 female speakers, are selected from the test set of the TIMIT database [Garofolo & Consortium 1993]. Note the test set is not included in the training set for the speech codebook. Three seconds of silence are inserted at the beginning of each sequence.
- The resulting speech sequences are mixed with 12 different stationary and non-stationary types of noise (F16, living room, train station, inside train, highway inside car, outside traffic road, wind, jackhammer, forest, pub noise, indoor soccer, modulated Gaussian noise). The Gaussian noise is modulated with $f_{\mathrm{mod}} = 0.5\,\mathrm{Hz}$ and generated according to Eq. (3.66). This results in 3600 different noisy speech data permutations, respectively 6 hours, 38 minutes and 40 seconds.

The performance of the rated systems is evaluated by the objective scores[3] *segmental noise attenuation* (SegNA), *segmental speech attenuation* (SegSA), as well as the *cepstral distance* (CD). Regarding the CD, lower values indicate a lower speech distortion. A high SegNA is desired while at the same time a low SegSA is favored.

For the purpose of evaluation, a modular noise reduction system is created, covering the different configurations, including the complexity reduction of the codebook matching process and the *information combining* procedure.

[2]The mixing procedure is detailed in Appendix C.1. Note that for the adjustment of the input SNR only speech and noise signal sections with speech presence are considered.

[3]The objective scores are described in detail in Appendix C.2.

### 5.8.1 Overview of Evaluation System

The evaluation system is based on a standard noise reduction system as depicted in Fig. 3.4 consisting of analysis, spectral weighting for speech enhancement and synthesis. The block diagram of the proposed codebook driven noise reduction system is illustrated in Fig. 5.3. The analysis and synthesis is carried out as presented in Sec. 3.2.1. After analysis, the processing of the noisy input signal $\mathcal{Y}(\lambda, \mu)$ takes place in the frequency domain. The codebook matching block provides estimates for the *short-term power spectrum* (STPS) of speech $\left|\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$ and noise $\left|\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$. The codebook matching is implemented as depicted in Fig. 5.1. The VAD is estimated as introduced in Sec. 4.4 and the corresponding VAD algorithm parameters are summarized in Tab. 5.3. Using configuration ⓐ, the joint brute force search (cf. Sec. 4.1.2) is employed while speech is active to obtain $\left|\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$ and $\left|\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$, whereas the sequential optimization using a cascade of gain shape VQs with subsequent *information combining* (cf. Sec. 5.5, Sec. 5.7.2) is carried out using configuration ⓑ. In the following configuration ⓐ is referred to as ***brute force*** and configuration ⓑ is referred to as ***GSC*** in the legend of the respective plots. By setting $v_{\mathrm{bin}}(\lambda) = 1$ in Fig. 5.1 the VAD can be disabled.

In parallel, a second STPS noise estimate is computed by a statistical noise estimator, indicated by $\left|\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda, \mu)\right|^2$, e. g., *SPP* [Gerkmann & Hendriks 2011], *Minimum Statistics* [Martin 2006] or *Baseline Tracing*, cf. Sec. 3.4.1. Furthermore, a second estimate for the STPS of speech $\left|\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda - 1, \mu)\right|^2$ is provided which is detailed later. All speech and noise estimates are fed into the *information combining* block which merges the respective estimates according to Sec. 5.5, to provide enhanced estimates for the STPS of speech $\left|\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda, \mu)\right|^2$ and noise $\left|\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda, \mu)\right|^2$, respectively[4]. If the *information combining* block is disabled, the speech and noise estimates yield $\left|\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda, \mu)\right|^2 = \left|\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$ and $\left|\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda, \mu)\right|^2 = \left|\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$, respectively. The expression ***information combining*** indicates in the legend of related plots the enabled operation of the information combining block.

Subsequently, different SNR estimates are computed from the refined speech and noise estimates. The estimate $\widehat{\gamma}(\lambda, \mu)$ of the *a posteriori* SNR is calculated according to Eq. (3.29). Two different *a priori* SNR estimates are calculated: the estimate $\widehat{\xi}(\lambda, \mu)$ is computed by the decision-directed approach [Ephraim & Malah 1984] and the new estimate $\widehat{\xi}_{\mathrm{mod}}(\lambda, \mu)$ is determined as introduced in Sec. 4.2. Note $\widehat{\xi}(\lambda, \mu)$ is a function of $\left|\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda, \mu)\right|^2$, $\left|\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda - 1, \mu)\right|^2$ and $\left|\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda - 1, \mu)\right|^2$, while

[4] Note that in case of operating the codebook matching block in configuration ⓑ the *information combining* algorithm is applied twice. First, while codebook matching using the estimates provided by the cascade of gain shape VQs and second in the information combining block employing $\left|\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$, $\left|\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda, \mu)\right|^2$, $\left|\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda, \mu)\right|^2$, $\left|\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda - 1, \mu)\right|^2$.

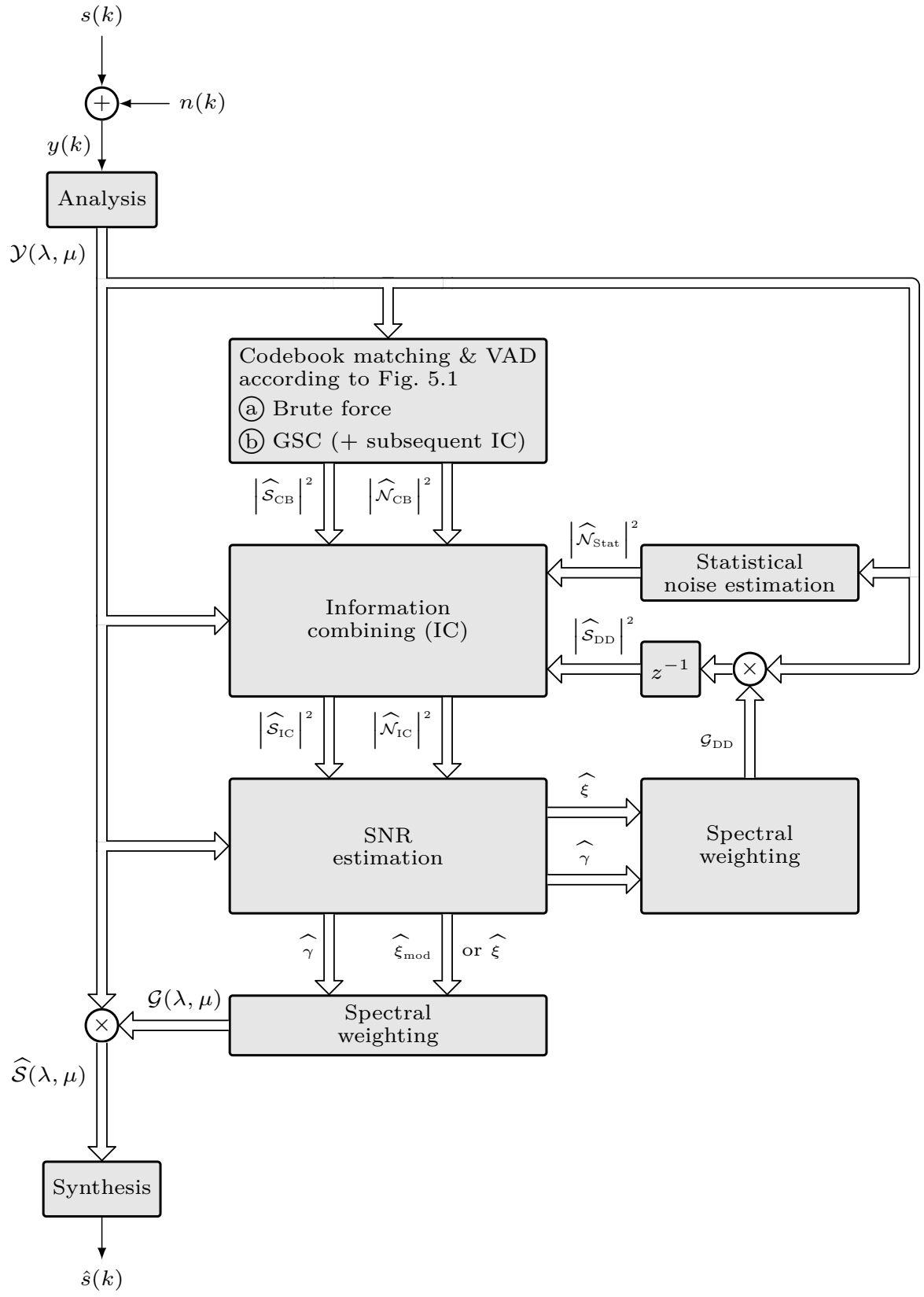



**Figure 5.3:** Block diagram of codebook based noise reduction system working in the frequency domain including information combining

| Parameter | Settings |
|---|---|
| Sampling frequency $f_\mathrm{s}$ | 16 kHz |
| Frame length $L_F$ | 320 ($\hat{=}$ 20 ms) |
| Frame advance $L_A$ | 160 ($\hat{=}$ 10 ms) |
| FFT length $N_\mathrm{DFT}$ | 512 (including zero-padding) |
| Frame overlap | 50 % ($\sqrt{\text{Hann-window}}$) |
| Speech codebook entries $L$ | 128 (training sequence 3073 s, cf. Sec. 4.3.4) |
| Decision-directed SNR factor | $\alpha_\xi = 0.98$ |
| Spectral weights $\mathcal{G}(\lambda, \mu)$ | Wiener Filter (cf. Sec. 3.4.3) |
| Spectral weights $\mathcal{G}_\mathrm{DD}(\lambda, \mu)$ | Wiener Filter (cf. Sec. 3.4.3) |

**Table 5.1:** Simulation system settings

$\widehat{\xi}_\mathrm{mod}(\lambda, \mu)$ uses $\left|\widehat{\mathcal{N}}_\mathrm{IC}(\lambda, \mu)\right|^2$, $\left|\widehat{\mathcal{N}}_\mathrm{IC}(\lambda - 1, \mu)\right|^2$, $\left|\widehat{\mathcal{S}}(\lambda - 1, \mu)\right|^2$ and additionally takes the speech estimate $\left|\widehat{\mathcal{S}}_\mathrm{IC}(\lambda, \mu)\right|^2$ into account. Based on the SNR estimates, two different weighting gains are calculated. As mentioned before, the speech estimate $\left|\widehat{\mathcal{S}}_\mathrm{DD}(\lambda, \mu)\right|^2$ is utilized by the *information combining* block and is provided by multiplying the spectral weighting gain $\mathcal{G}_\mathrm{DD}(\lambda, \mu)$ with $\mathcal{Y}(\lambda, \mu)$. Due to causality, only the previous frame $\left|\widehat{\mathcal{S}}_\mathrm{DD}(\lambda - 1, \mu)\right|^2$ can be used in the *information combining* block. Hence, it is called decision-directed speech estimate. Since the chain of *information combining*, SNR estimation and speech estimation $\widehat{\mathcal{S}}_\mathrm{DD}$ forms a loop, the chance of error propagation with respect to the speech estimate $\left|\widehat{\mathcal{S}}_\mathrm{IC}(\lambda, \mu)\right|^2$ exists. This can be prevented, if the SNR estimate from which the weighting gain $\mathcal{G}_\mathrm{DD}(\lambda, \mu)$ is calculated is independent of the speech estimate $\left|\widehat{\mathcal{S}}_\mathrm{IC}(\lambda, \mu)\right|^2$. Hence, $\mathcal{G}_\mathrm{DD}(\lambda, \mu)$ is calculated from the *a priori* SNR estimate $\widehat{\xi}(\lambda, \mu)$. For the actual speech enhancement, another spectral weighting $\mathcal{G}(\lambda, \mu)$ is utilized which depends on the new *a priori* SNR estimate $\widehat{\xi}_\mathrm{mod}(\lambda, \mu)$. The enhanced time domain signal $\hat{s}(k)$ is obtained by applying an *inverse DFT* (IDFT), windowing using a square root Hann-window and overlap-add. The common parameters of the simulation system are detailed in Tab. 5.1. Note that the employed speech codebook is pre-trained as detailed in Sec. 4.3.2 and evaluated in Sec. 4.3.4. The selected fixed speech codebook comprises spectral envelopes and is used for all codebook driven algorithms in this section.

Additionally, a representative of a conventional statistical based noise reduction system is included in the benchmark and serves as anchor. Therefore, the noise reduction system as depicted in Fig. 3.4 is utilized. The noise estimate is provided by the *speech presence probability* (SPP) algorithm which is parameterized as suggested in [Gerkmann & Hendriks 2011]. The estimate of the *a priori* SNR and *a posteriori*

SNR is provided by the decision-directed approach [Ephraim & Malah 1984]. The speech enhancement is carried out by spectral weighting which depends on the *a priori* SNR estimate and is also implemented as Wiener Filter (cf. Sec. 3.4.3). The enhanced time domain signal $\hat{s}(k)$ is obtained by the same procedure as in the proposed evaluation system. This system is referred to as ***SPP*** in the following.

## 5.8.2 Reference Codebook Implementation

The reference codebook matching system is based on fixed speech and noise codebooks. It utilizes a brute force search for the determination of the optimal parameters $l_{\text{opt}}, m_{\text{opt}}, \sigma_{n,\text{opt}}$ as explained in Sec. 4.1.2 using $N_q = 16$ quantization levels for $\sigma_{n,\text{opt}}$, i. e., configuration ⓐ in Fig. 5.3. Since no VAD information is available, i. e., $v_{\text{bin}}(\lambda) = 1$, the brute force search is performed in each frame $\lambda$ resulting in the speech estimate $\left|\widehat{\mathcal{S}}_{\text{CB}}(\lambda,\mu)\right|^2$ and in the noise estimate $\left|\widehat{\mathcal{N}}_{\text{CB}}(\lambda,\mu)\right|^2$. Two different reference codebook matching systems are defined, representing different degrees of *a priori* knowledge with respect to noise.

**A** This configuration exhibits a pre-trained large noise codebook, consisting of four entries for each of the 11 noise types. Sub-codebooks for each type of noise are trained as described in Sec. 4.3.1. The final noise codebook consists of a concatenation of the respective sub-codebooks. In total, $M = 44$ entries are created from 10 s training sequence for each noise type. Since "pub noise" is very similar to speech, it is excluded from the noise codebook. Hence, the noise codebook exists of 11 types of noise. Since all noise types (except "pub noise") to be evaluated are included in the training this configuration is referred to as: ***A large fixed ref. CB*** ($M = 44$)[5].

**B** In contrast to configuration A, the training sequences for each noise type are concatenated and a codebook with only $M = 4$ entries as representative for a small fixed codebook is created. This is equivalent to strongly averaged and imprecise *a priori* knowledge. This configuration is named: ***B small fixed ref. CB*** ($M = 4$).

## 5.8.3 Modified Decision-Directed SNR Estimation

The new decision-directed SNR estimate $\widehat{\xi}_{\text{mod}}(\lambda,\mu)$ is compared to the conventional decision-directed approach $\widehat{\xi}(\lambda,\mu)$ using the reference codebook implementation in both configurations and the standard noise reduction system utilizing the SPP noise estimator. This is achieved by selecting either $\widehat{\xi}_{\text{mod}}(\lambda,\mu)$ or $\widehat{\xi}(\lambda,\mu)$ in Fig. 5.3 for the calculation of the weighting gain $\mathcal{G}(\lambda,\mu)$. The performance is evaluated by means of the objective scores. Since the *information combining* block in Fig. 5.3 is disabled

[5] This choice is a tradeoff between the accuracy of *a priori* knowledge on noise and numerical complexity. With respect to the employed computer cluster of 150 nodes and the processing time, $M$ is set to 44, yielding $L \cdot M \cdot N_q = 90112$ distance calculations per frame.

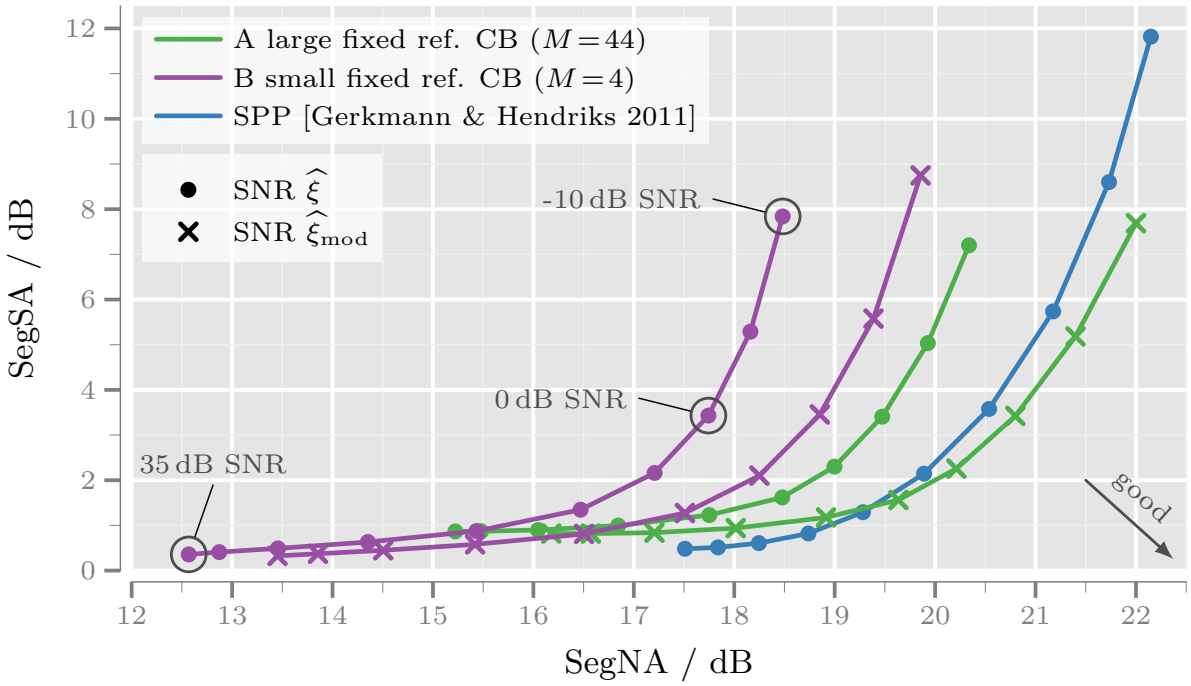



**Figure 5.4:** The *segmental speech attenuation* (SegSA) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter (Setup: Fig. 5.3 configuration ⓐ, i. e., brute force codebook search, disabled *information combining* block)

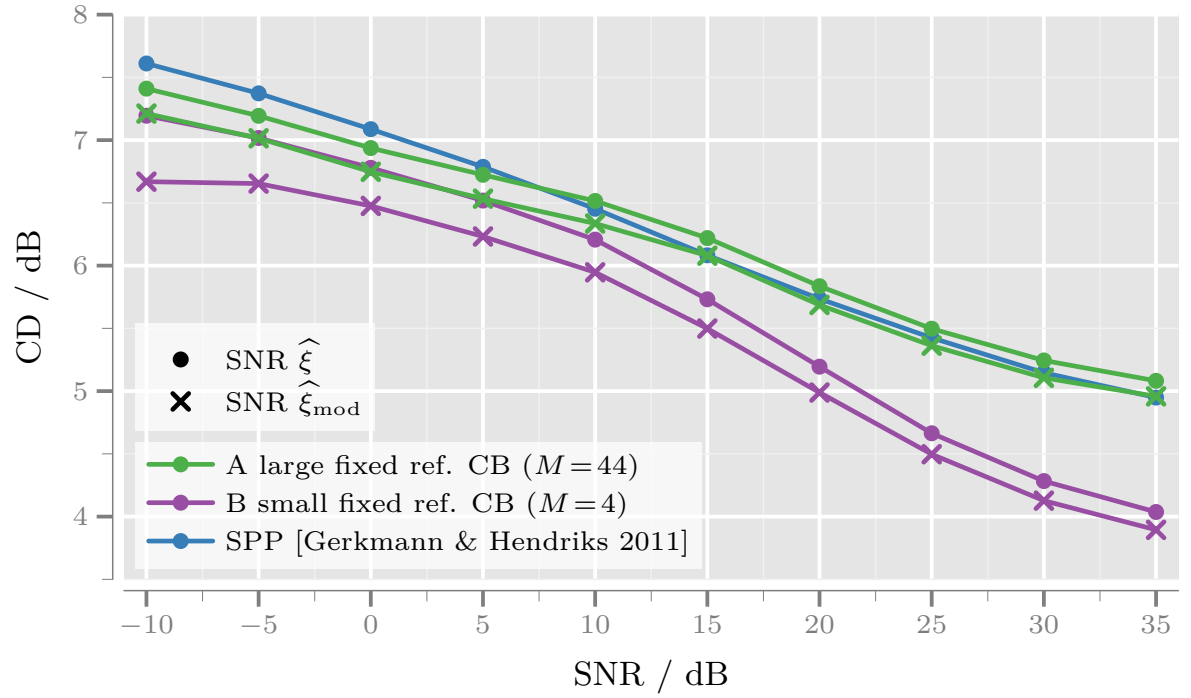



**Figure 5.5:** The *cepstral distance* (CD) is depicted over the input SNR (Setup: Fig. 5.3 configuration ⓐ, i. e., brute force codebook search, disabled *information combining* block).

in this investigation, the speech and noise estimates yield $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu) = \widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ and $\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda,\mu) = \widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)$, respectively.

Figure 5.4 depicts the averaged results for SegSA plotted over SegNA with the input SNR as variable parameter. Hence, a fair comparison with respect to the tradeoff SegNA versus SegSA is possible. The points of best performance would be placed in the lower right corner of that figure. At first, the results based on the conventional SNR estimates (—●—, —●—, —●—) are analyzed amongst each other, allowing a fair comparison of the codebook based noise reduction systems (—●—, —●—) with the statistical based one (—●—). As expected by the very condensed *a priori* knowledge on noise, the codebook based system using configuration B (—●—) marks the lower bound of the performance. Compared with *SPP* (—●—), a lower SegSA is observed over the complete input SNR range, but at the expense of a significantly lower SegNA. The good SegSA performance is confirmed by the best performance of the CD measure (—●—) presented in Fig. 5.5. Comparing the codebook enhancement system using configuration A (—●—) with the *SPP* based conventional system (—●—), a reduced performance regarding the SegNA measure is visible, yet significantly better than the codebook approach using configuration B (—●—). This is plausible since the noise codebook exhibits four codebook entries of each occurring noise type. Regarding the SegSA scores, the good performance of approach B is reflected, achieving the best scores for low input SNR.

In general, the use of the speech estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu) = \widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ for the calculation of the new SNR estimate $\widehat{\xi}_{\mathrm{mod}}(\lambda,\mu)$ is beneficial over the complete input SNR range (—×—, —×— vs. —●—, —●—). The noise attenuation performance increases while the speech attenuation holds approximately the same, except from an outlier at −10 dB input SNR for both configurations. Moreover, the codebook approach using configuration A (—×—) utilizing $\widehat{\xi}_{\mathrm{mod}}(\lambda,\mu)$ exhibits a similar SegNA performance compared with the *SPP* based system (—●—). The performance gain from the new SNR estimate $\widehat{\xi}_{\mathrm{mod}}(\lambda,\mu)$ is also reflected in the CD measure, cf. Fig. 5.5. Over the complete input SNR range a lower CD is observed. It is notable that although the SegNA is increased the speech distortion is reduced at the same time.

### 5.8.4 *Information Combining*

In this section the *information combining* procedure is analyzed by means of the reference codebook implementations, i. e., configuration ⓐ in Fig. 5.3 with enabled *information combining* block. Since no VAD information is available, the brute force codebook search is performed in each frame $\lambda$. Several speech and noise estimates are merged to obtain a refined speech estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu)$ and noise estimate $\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda,\mu)$. As depicted in Fig. 5.3, the speech estimates are provided by the codebook matching unit $\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ and the decision-directed speech estimate $\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda-1,\mu)$ of the last frame. Note that the determination of $\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda-1,\mu)$ is independent of the refined speech estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu)$ in order to prevent error propagation. The noise estimates comprise the estimate $\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)$ from the codebook matching as well as an independent statistically based noise estimate $\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda,\mu)$, e. g., Sec. 3.4.1.

### Example of *Information Combining*

A noisy input signal is generated consisting of ten different, six seconds long stationary and non-stationary noise types mixed with five male and female English speakers taken from the TIMIT database [Garofolo & Consortium 1993] at 0 dB SNR. In order to increase the contrast between the statistical $\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda,\mu)$ and the codebook based $\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)$ noise estimate, $\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda,\mu)$ is provided by *Minimum Statistics* [Martin 2006] and parameterized as suggested by the author. The parameters for the simulation remain the same as for the benchmark (cf., Tab. 5.1).

The result of the *information combining* procedure is summarized in terms of spectrograms in Fig. 5.6 for noise and in Fig. 5.7 for speech, respectively. For each plot, the most meaningful time section of the 60 s example is depicted. The two spectrograms placed in the middle of each figure depict the two input estimates which yield the refined estimate depicted in the lower spectrogram after *information combining*. In the upper plot either the noise or clean speech signal is presented as reference and marks the upper bound for the estimates $\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda,\mu)$ and $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu)$.

While the codebook driven noise estimate $\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)$ in Fig. 5.6 exhibits a reasonable performance with respect to the temporal structure, the statistical based estimate $\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda,\mu)$ is reliable regarding the stationary noise components. However, significant estimation errors occur occasionally in $\widehat{\mathcal{N}}_{\mathrm{CB}}(\lambda,\mu)$, e. g., at position 45.4 s, 46.5 s and while "wind" noise is present for frequencies greater than 2 kHz. Due to the sliding time window of *Minimum Statistics*, a significant underestimation of the noise is often caused by the transition of noise types, e. g., from "wind" to "jackhammer" noise at 54 s. The spectrogram of $\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda,\mu)$, demonstrates that the *information combining* procedure is able to combine the best of both noise estimates, yielding a refined noise estimate which exhibits a precise spectral and temporal structure. Moreover, dominant estimation errors are compensated by the respective other noise estimate.

Since the speech codebooks exhibit spectral envelopes, the speech estimate $\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ is spectrally smooth as illustrated in the second spectrogram of Fig. 5.7. While this yields a sub-optimal estimate for voiced sounds, unvoiced sounds are in contrast estimated reliably, e. g., at position 4.9 s or 15.7 s. On the other hand, the decision-directed estimate $\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda,\mu)$ is a rather aggressive estimate of speech which tends to especially underestimate unvoiced sounds. However, voiced sounds are precisely estimated including the spectral fine-structure caused by the speaker dependent pitch, e. g., at 3.5 s or 9.5 s. Similar to the *information combining* of the noise estimates, the speech estimate after *information combining* merges the best of both speech estimates as shown in the corresponding spectrogram in Fig. 5.7. More insights into the combining behavior are gained by analyzing the respective *information combining* weights, which are visualized by means of spectrograms in Fig. 5.8. Blue areas indicate a weight close to zero, while red areas denote weights close to one. In general, it is observed that the magnitudes minima between the pitch harmonics of voiced sounds are canceled out preferring the speech estimate $\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda,\mu)$. Whereas at the pitch lines either $\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ or $\widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda,\mu)$ is selected,

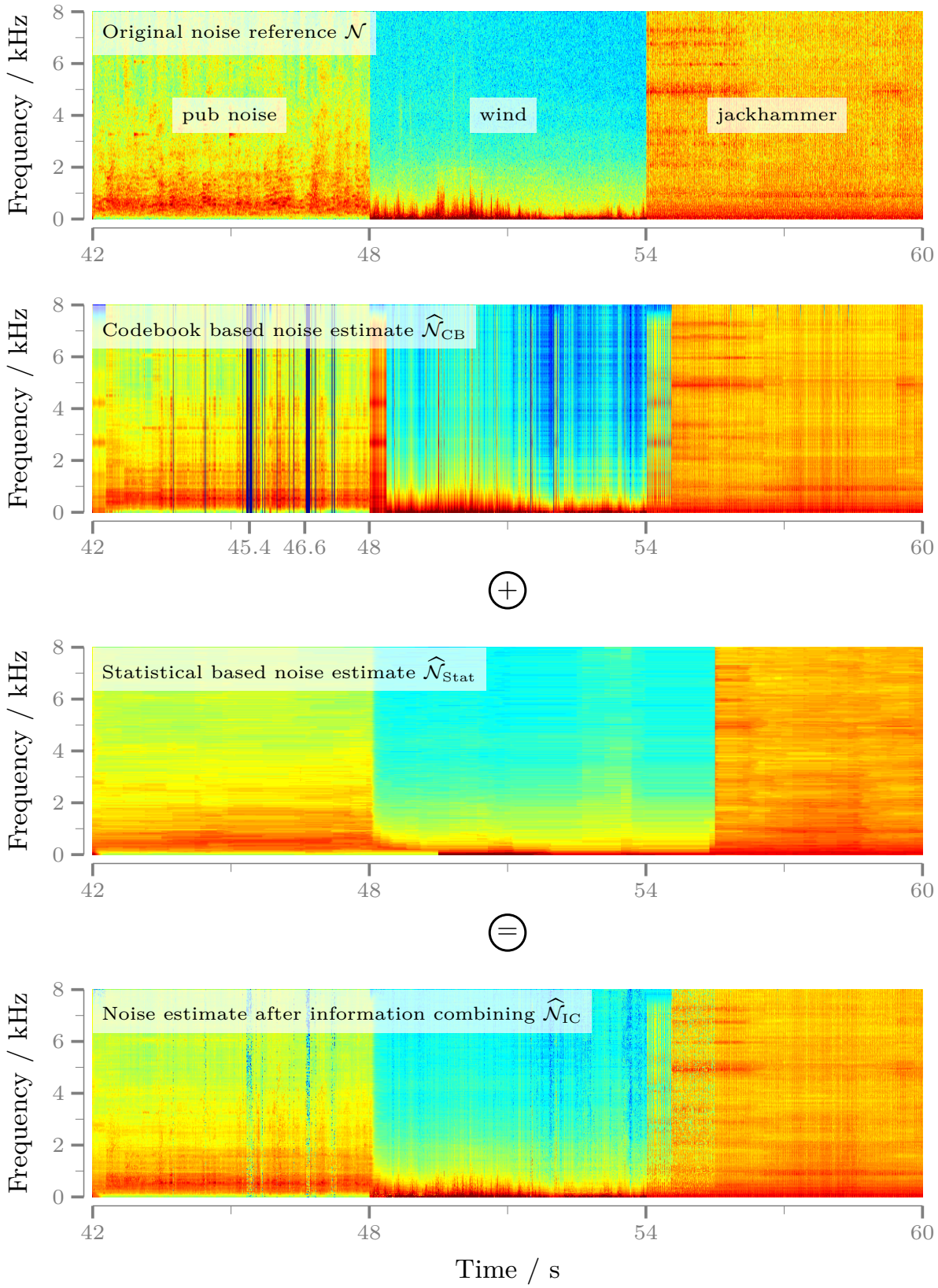



**Figure 5.6:** Example of *information combining* depicted as spectrograms. Two different noise estimates $\widehat{\mathcal{N}}_{\mathrm{CB}}$ and $\widehat{\mathcal{N}}_{\mathrm{Stat}}$ are combined yielding the enhanced estimate $\widehat{\mathcal{N}}_{\mathrm{IC}}$. The true noise $\mathcal{N}$ is depicted as reference in the upper spectrogram.

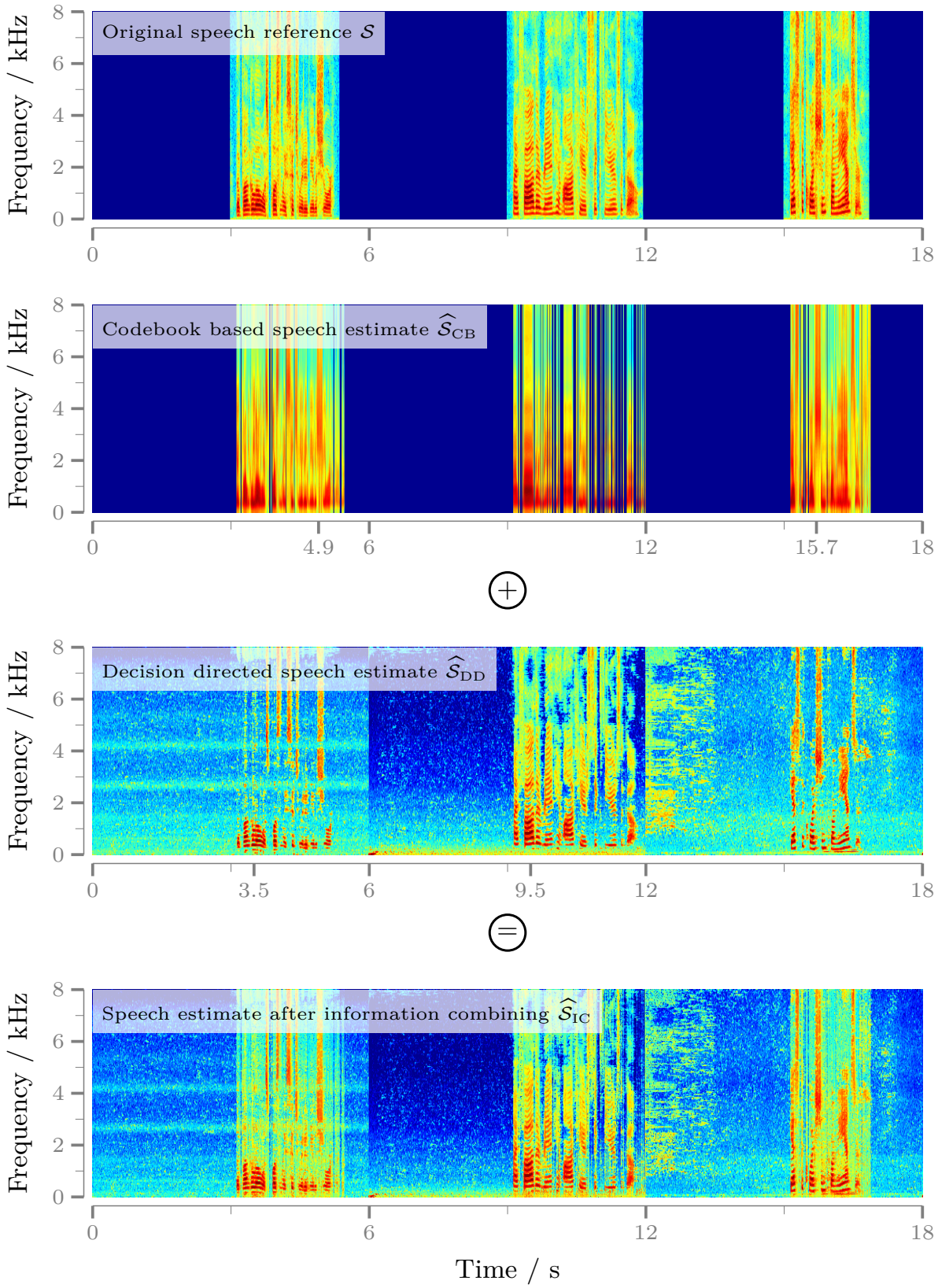



**Figure 5.7:** Example of *information combining* depicted as spectrograms. Two different speech estimates $\widehat{\mathcal{S}}_{\mathrm{CB}}$ and $\widehat{\mathcal{S}}_{\mathrm{DD}}$ are combined yielding the enhanced estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}$. The true speech $\mathcal{S}$ is depicted as reference in the upper spectrogram.

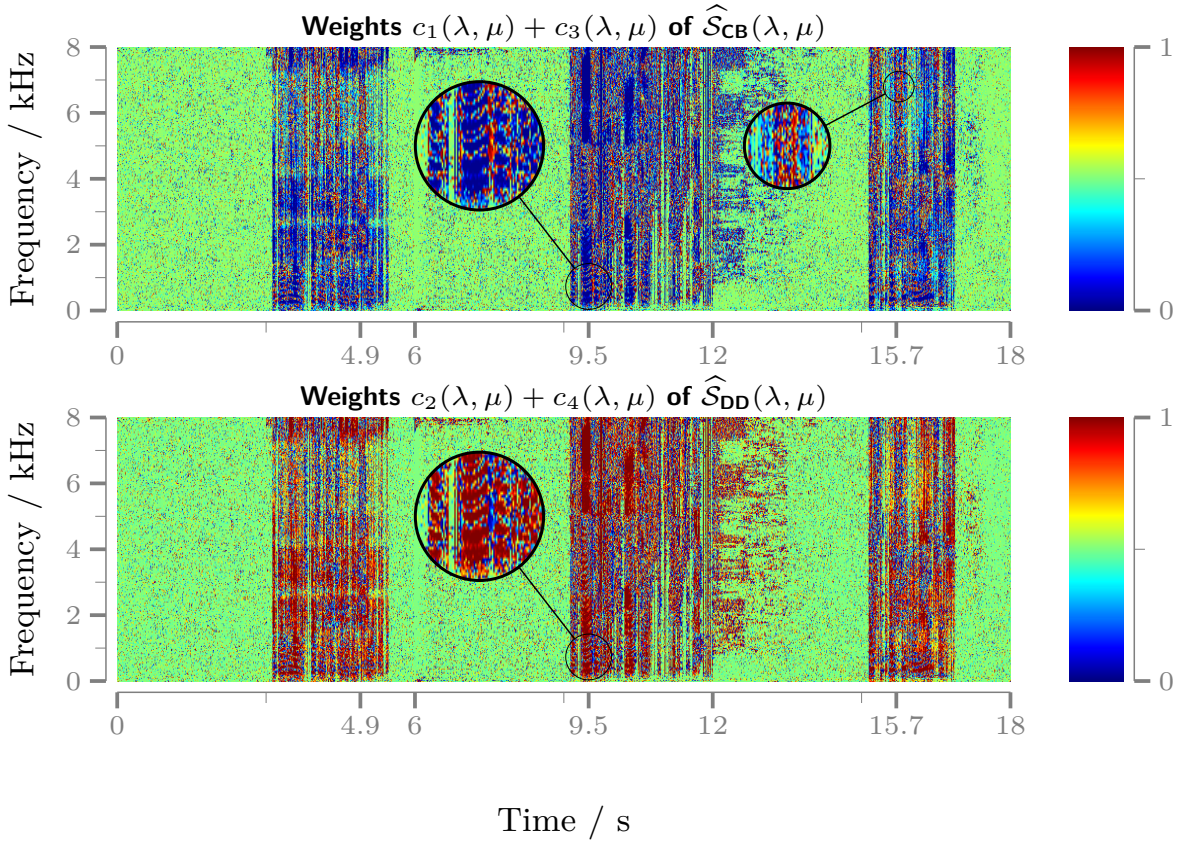



**Figure 5.8:** Example of *information combining* weights of the speech estimates $\widehat{\mathcal{S}}_{\mathrm{CB}}$ and $\widehat{\mathcal{S}}_{\mathrm{DD}}$ corresponding to Fig. 5.7 are depicted as spectrograms. According to Eq. (5.18) the speech estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu)$ after information combining yields $[c_1(\lambda,\mu) + c_3(\lambda,\mu)] \cdot \widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu) + [c_2(\lambda,\mu) + c_4(\lambda,\mu)] \cdot \widehat{\mathcal{S}}_{\mathrm{DD}}(\lambda,\mu)$ with $\widehat{\mathcal{S}}_1 = \widehat{\mathcal{S}}_{\mathrm{CB}}$ and $\widehat{\mathcal{S}}_2 = \widehat{\mathcal{S}}_{\mathrm{DD}}$.

e. g., at 9.5 s. This indicates that the spectral envelope of voice sounds is in general estimated reliably by $\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$. Moreover, voiced sounds are more frequently selected from the codebook driven estimate $\widehat{\mathcal{S}}_{\mathrm{CB}}(\lambda,\mu)$ as demonstrated at, e. g., position 4.9 s, 15.2 s or 15.7 s.

### Noise Reduction Performance

The performance is also rated in terms of objective speech enhancement scores. For the calculation of the weighting gain $\mathcal{G}(\lambda,\mu)$ the conventional decision-directed approach $\widehat{\xi}(\lambda,\mu)$ is used. The statistical noise estimate $\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda,\mu)$ for *information combining* is provided by *Baseline Tracing*, cf. Sec. 3.5. The parameters for the simulation and the VAD setup remain the same as detailed in Sec. 5.8.1.

It should be noted that the focus of this section is to emphasize the performance gain by applying *information combining* to the codebook driven enhancement systems and to create a reference for the evaluation of the online noise codebook

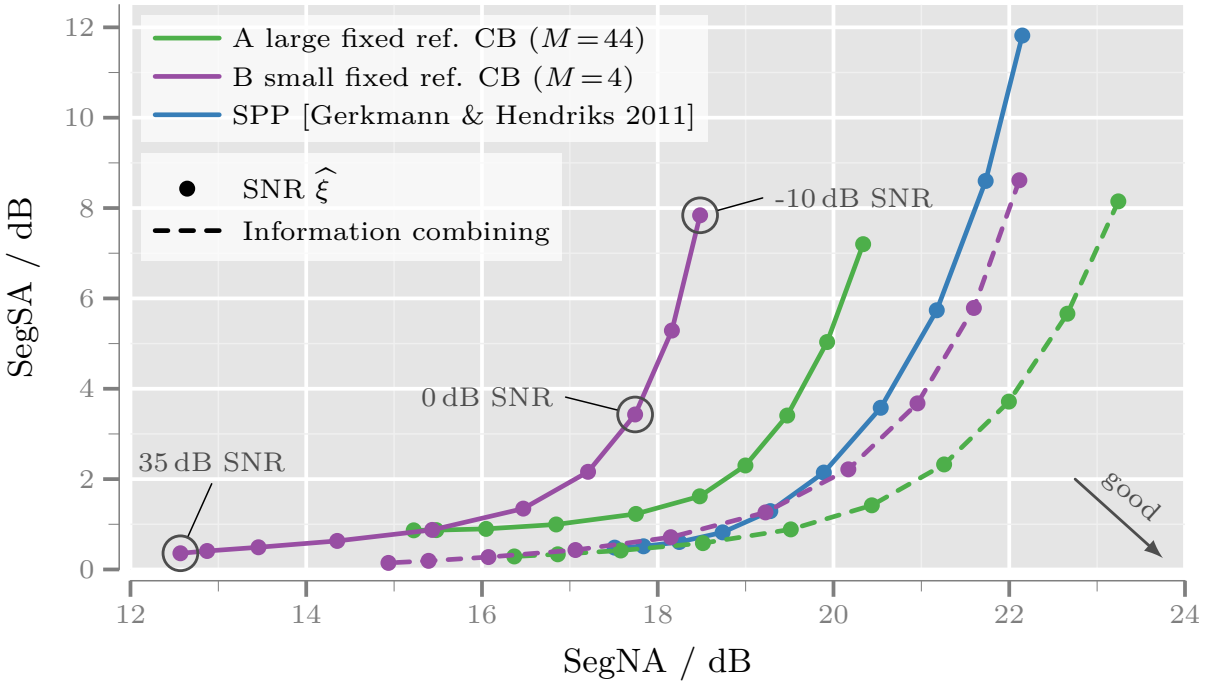



**Figure 5.9:** The *segmental speech attenuation* (SegSA) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter (Setup: Fig. 5.3 configuration ⓐ, i. e., brute force codebook search).

adaptation detailed later in Sec. 5.8.5. It is not intended as an unfair comparison with the conventional *SPP* based enhancement system.

In Fig. 5.9 the averaged results are presented for SegSA plotted over SegNA with the input SNR as variable parameter. Dashed lines (– – –) indicate the noise reduction systems utilizing the *information combining* block in Fig. 5.3, while solid lines (——) depict the regular approach known from Sec. 5.8.3 without *information combining*. Again, the *SPP* based enhancement system (—●—) is depicted as representative for conventional speech enhancement. As indicated by the previous examples, the use of *information combining* (—●—, —●— vs. – –●– –, – –●– –) yields a tremendous performance gain. This is also reflected in the objective scores where a considerable enhancement is observed for both configurations of the codebook matching. Evaluating configuration B, employing the condensed noise codebook, the use of *information combining* (—●— vs. – –●– –) provides a significant improvement regarding SegNA, with a lead up to 3.7 dB. While for low input SNR (−10 dB to 0 dB) the SegSA measure performs slightly worse, a better performance in SegSA is observed for the remaining SNR range. Hence, for SNRs greater than 0 dB both measures are improved by *information combining* simultaneously. Consulting the results from the CD measures depicted in Fig. 5.10, this observation is confirmed. Utilizing *information combining* (– –●– –), the speech distortion, as indicated by CD, performs better starting with 0 dB SNR.

A similar behavior is noticeable for configuration A (—●— vs. – –●– –), using the

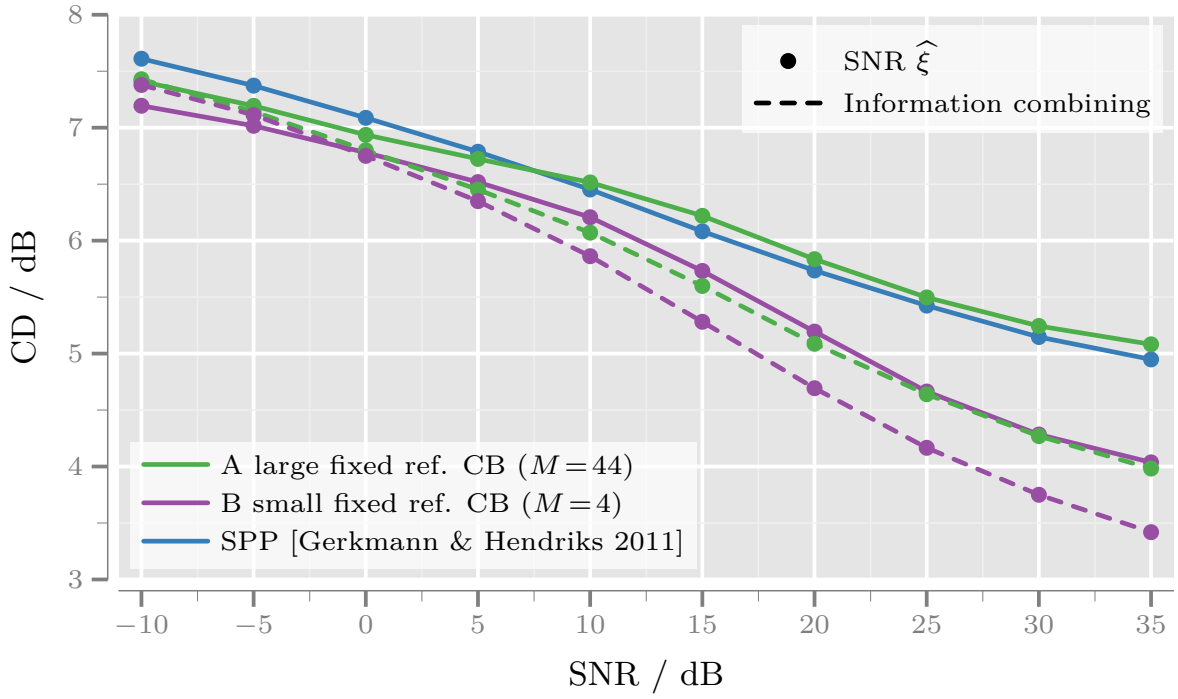



**Figure 5.10:** The *cepstral distance* (CD) is depicted over the input SNR (Setup: Fig. 5.3 configuration ⓐ, i. e., brute force codebook search).

large noise codebook. Since the *a priori* knowledge on noise is better, this approach performs inherently better compared with configuration B (—●—, - -●- -). While the performance gain regarding SegNA is likewise utilizing *information combining*, the improvement of SegSA for high input SNR is more pronounced. This is reflected in terms of speech distortion as the CD enhancement utilizing *information combining* is prominent over the complete input SNR range and exhibits a more distinct improvement.

In comparison with the conventional system (—●—), both codebook driven enhancement systems (-●-, - -●- -) are clearly superior utilizing *information combining*. While the SegNA is greater (except for very high input SNR), the SegSA and CD are simultaneously lower.

### Analysis of the Estimation Errors

From the benchmark of the previous section, the estimation errors of the respective speech and noise estimates are investigated. At first, a lower bound for the estimation errors of speech and noise is defined. Given the true speech $\mathcal{S}(\lambda, \mu)$ and noise $\mathcal{N}(\lambda, \mu)$ signals, which are available in the simulation environment, it is possible to calculate the true power of the estimation errors regarding the different speech $\left|E_s^{\mathcal{S}}(\mu)\right|^2$ and noise $\left|E_n^{\mathcal{N}}(\mu)\right|^2$ estimates, respectively. Optimal *information combining* depending on the true error powers is carried out independently for the speech and noise estimates by applying Eq. (5.15) and Eq. (5.46) likewise. The estimates obtained from this procedure are denoted by $\widehat{\mathcal{S}}_{\mathrm{IC,opt}}(\lambda, \mu)$ and

$\widehat{\mathcal{N}}_{\mathrm{IC,opt}}(\lambda,\mu)$ and are considered as the best possible estimates. These estimates are compared to the regularly computed estimates $\widehat{\mathcal{S}}_{\mathrm{IC}}(\lambda,\mu)$, $\widehat{\mathcal{N}}_{\mathrm{IC}}(\lambda,\mu)$, the individual estimates $\widehat{\mathcal{S}}_{\mathrm{CB}}$, $\widehat{\mathcal{S}}_{\mathrm{DD}}$, $\widehat{\mathcal{N}}_{\mathrm{CB}}$, $\widehat{\mathcal{N}}_{\mathrm{Stat}}$, as well as a representative of a simple *information combining* by averaging the individual estimates given by

$$\widehat{\mathcal{S}}_{\mathrm{mean}}(\lambda,\mu)=\frac{1}{N_s}\sum_{s=1}^{N_s}\widehat{\mathcal{S}}_s(\lambda,\mu),\quad \widehat{\mathcal{N}}_{\mathrm{mean}}(\lambda,\mu)=\frac{1}{N_n}\sum_{i=1}^{N_n}\widehat{\mathcal{N}}_n(\lambda,\mu). \tag{5.58}$$

The estimation errors of speech: $E^{\mathcal{S}}_{\mathrm{IC}}$, $E^{\mathcal{S}}_{\mathrm{CB}}$, $E^{\mathcal{S}}_{\mathrm{DD}}$, $E^{\mathcal{S}}_{\mathrm{mean}}$, and noise: $E^{\mathcal{N}}_{\mathrm{IC}}$, $E^{\mathcal{N}}_{\mathrm{CB}}$, $E^{\mathcal{N}}_{\mathrm{Stat}}$, $E^{\mathcal{N}}_{\mathrm{mean}}$ are calculated in the MSE sense for each frame of the benchmark. For the sake of clarity, a normalization with respect to the maximum occurred error from the optimal estimates $E^{\mathcal{S}}_{\mathrm{IC,opt}}$ or $E^{\mathcal{N}}_{\mathrm{IC,opt}}$ is carried out, respectively. For a clear presentation, the delta errors, given by

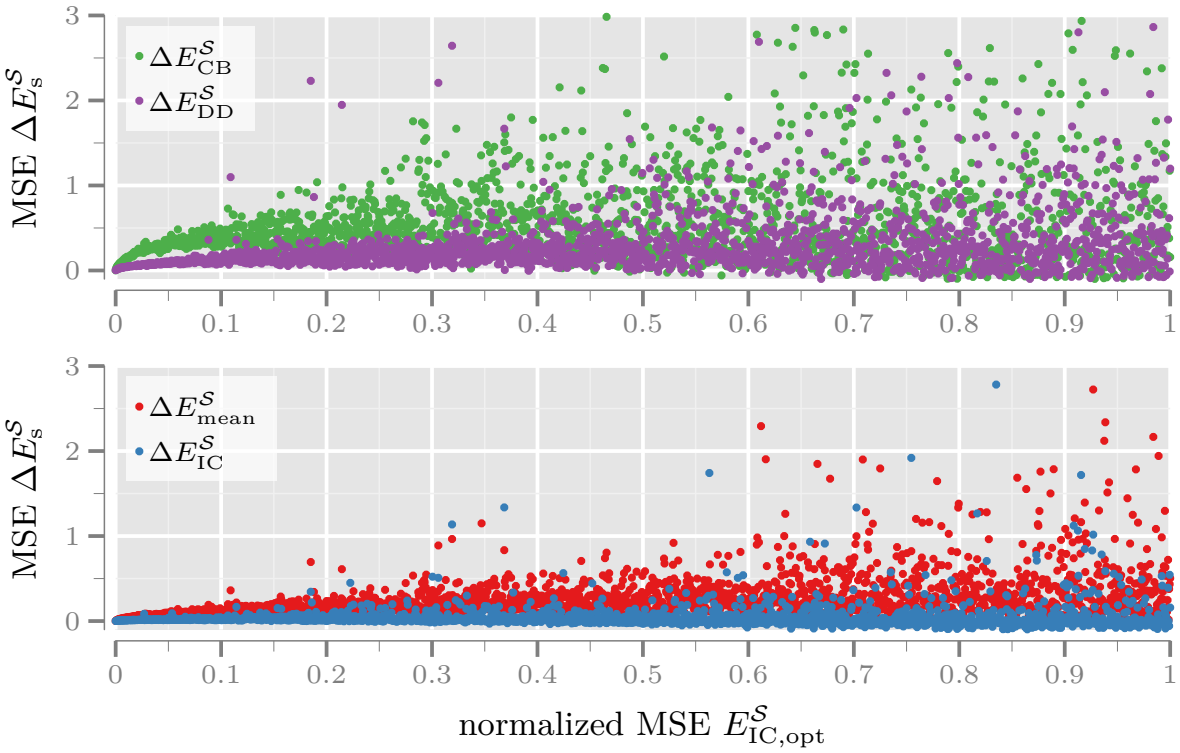
$$\Delta E^{\mathcal{S}}_{\mathrm{s}}=E^{\mathcal{S}}_{\mathrm{s}}-E^{\mathcal{S}}_{\mathrm{IC,opt}},\qquad s\in\{\mathrm{IC},\mathrm{DD},\mathrm{mean}\}, \tag{5.59}$$

$$\Delta E^{\mathcal{N}}_{\mathrm{n}}=E^{\mathcal{N}}_{\mathrm{n}}-E^{\mathcal{N}}_{\mathrm{IC,opt}},\qquad n\in\{\mathrm{IC},\mathrm{Stat},\mathrm{mean}\}, \tag{5.60}$$

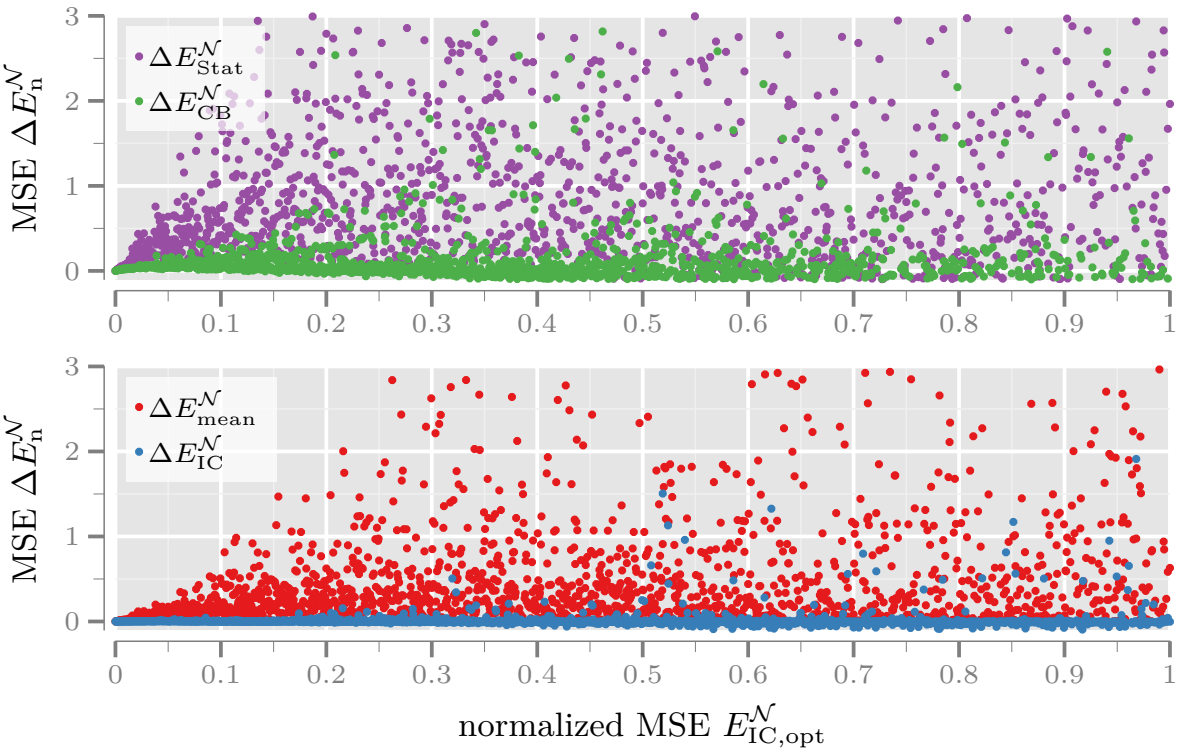
are computed and summarized in Fig. 5.11 separately for speech and noise. The respective delta errors ($\Delta E^{\mathcal{S}}_{\mathrm{s}}$, $\Delta E^{\mathcal{N}}_{\mathrm{n}}$) are depicted over the normalized estimation error of the optimal estimate ($E^{\mathcal{S}}_{\mathrm{IC,opt}}$, $E^{\mathcal{N}}_{\mathrm{IC,opt}}$). Hence, the abscissa of the plots range from zero to one. Ordinate values greater than zero indicate an additional error compared to the optimal estimate, which is considered as lower bound of the estimation errors. The ordinate intercept is chosen from zero to three and presents the most meaningful section.

In Fig. 5.11a the estimation errors of speech are outlined. The upper plot depicts the delta estimation errors of the individual estimates, i. e., $\Delta E^{\mathcal{S}}_{\mathrm{CB}}$ and $\Delta E^{\mathcal{S}}_{\mathrm{DD}}$. Since the codebook based speech estimate $\widehat{\mathcal{S}}_{\mathrm{CB}}$ provides only spectral envelopes, which are sub-optimal estimates, the largest delta errors $\Delta E^{\mathcal{S}}_{\mathrm{CB}}$ among all methods are observed. In contrast, the delta estimation errors of the decision-directed speech estimates $\Delta E^{\mathcal{S}}_{\mathrm{DD}}$ performs significantly better. In the lower plot, the performance of the combined speech estimates is shown. While the simple *information combining* method by averaging $\Delta E^{\mathcal{S}}_{\mathrm{mean}}$ is able to outperform the codebook driven speech estimate, a rather similar performance is observed regarding $\Delta E^{\mathcal{S}}_{\mathrm{DD}}$, yet yielding a smaller variance. The proposed *information combining* methods clearly performs best. In addition, the mean and the variance of the delta errors of each method are summarized in Tab. 5.2. It is notable that the mean and variance of the delta estimation error $\Delta E^{\mathcal{S}}_{\mathrm{IC}}$ for the proposed method is close to zero.

The estimation errors regarding the noise estimates are depicted in Fig. 5.11b. As the statistical based noise estimate $\widehat{\mathcal{N}}_{\mathrm{Stat}}(\lambda,\mu)$ is not able to follow non-stationary noise, the worst performance is expected and confirmed by $\Delta E^{\mathcal{N}}_{\mathrm{Stat}}$. In turn, the codebook based noise estimate $\Delta E^{\mathcal{N}}_{\mathrm{CB}}$ is also able to follow non-stationary noise and thus clearly outperforms the statistically based estimate. The difference between both methods is considerably pronounced compared to the individual speech estimation methods. For this reason, the simple *information combining*




**(a)** Speech




**(b)** Noise

**Figure 5.11:** Insights into the estimation error of indirect *information combining*. The delta *mean-square error* (MSE) $\Delta E_{\mathrm{IC}}^{\mathcal{S},\mathcal{N}}$ is depicted over the normalized reference MSE $E_{\mathrm{IC,opt}}^{\mathcal{S},\mathcal{N}}$, which is obtained from *information combining* utilizing the true estimation error powers of speech and noise.

| | delta error $\Delta E_{\mathrm{s}}^{\mathcal{S}}$ of speech | | | | delta error $\Delta E_{\mathrm{n}}^{\mathcal{N}}$ of noise | | | |
|---|---|---|---|---|---|---|---|---|
| | $\widehat{\mathcal{S}}_{\mathrm{CB}}$ | $\widehat{\mathcal{S}}_{\mathrm{DD}}$ | $\widehat{\mathcal{S}}_{\mathrm{mean}}$ | $\widehat{\mathcal{S}}_{\mathrm{IC}}$ | $\widehat{\mathcal{N}}_{\mathrm{CB}}$ | $\widehat{\mathcal{N}}_{\mathrm{Stat}}$ | $\widehat{\mathcal{N}}_{\mathrm{mean}}$ | $\widehat{\mathcal{N}}_{\mathrm{IC}}$ |
| mean | 0.67 | 0.35 | 0.29 | **0.07** | 0.31 | 3.38 | 0.98 | **0.03** |
| variance | 0.86 | 1.12 | 0.15 | **0.10** | 2.18 | 127.48 | 9.56 | **0.02** |

**Table 5.2:** Mean and variance of the delta estimation errors $\Delta E_{\mathrm{s}}^{\mathcal{S}}$ and $\Delta E_{\mathrm{n}}^{\mathcal{N}}$ of the different speech and noise estimates.

method by averaging $\Delta E_{\mathrm{mean}}^{\mathcal{N}}$ performs worse compared to the codebook based noise estimate $\Delta E_{\mathrm{CB}}^{\mathcal{N}}$. Again, the proposed method $\Delta E_{\mathrm{IC}}^{\mathcal{N}}$ clearly achieves the best scores, with a delta error very close to zero. The results are confirmed by Tab. 5.2.

Although the *information combining* of the proposed method is carried out indirectly, the performance is virtually unaffected compared to the *information combining* method exhibiting perfect knowledge on the estimation errors of speech and noise. In case of noise estimation, the proposed method performs even better compared to speech estimation.

### 5.8.5 Online Noise Codebook Adaptation

The investigations of the previous sections were based on the reference codebook system which employs a pre-trained noise codebook. Since the noise codebook training includes all considered noise types, the system covers unrealistic use cases in general. In this section the online noise codebook adaptation as explained in Sec. 4.5 is evaluated. The corresponding parameters for the VAD and online learning are summarized in Tab. 5.3. Since the information about speech activity is inherently available, the brute force codebook search is only applied in phases of speech activity. Hence, the codebook matching and VAD block of Fig. 5.3 is setup with configuration ⓐ as depicted in Fig. 5.1. The fixed part of the noise codebook uses the very condensed noise codebook ($M_{\square} = 4$) from configuration B of the reference system and the *maximum* number of codebook entries is defined as $M = 12$, i. e., $M_{\circ} = 8$ adaptive codebook entries. Hence, after $r = 2$ online codebook updates the less used entries of the last $L_W$ frames are replaced in the adaptive part of the codebook.

Figure 5.12 presents the averaged results for SegSA plotted over SegNA with the input SNR as variable parameter. For reference, the *SPP* based enhancement system (—●—) is depicted as well as the reference codebook system with configuration A ($M = 44$, –●–, –✖–). Utilizing the online noise codebook adaptation (–●–, –✖–) improves the SegNA enormously over the complete input SNR compared with all reference systems (—●—, –●–, –✖–), e. g., up to 5.8 dB comparing –✖– with –✖– at 0 dB input SNR. Utilizing the modified decision-directed SNR estimate $\widehat{\xi}_{\mathrm{mod}}$ (–✖–) is advantageous again.

If the input SNR is very low, the occurrence of false positives during VAD is increased (cf. Sec. 4.4.4). Hence, speech contributes occasionally to the noise

| | Parameter | Settings |
|---|---|---|
| Online CB adaptation | Training frames $L_T$ | 40 |
| | VQ output size $M_\Delta$ | 4 codebook entries |
| | Hangover VAD margin $L_H$ | 60 frames |
| | Adaption margin $L_M$ | 40 frames |
| | Hit rate $T$ | 80 % |
| | Speech codebook size $L$ | 128 entries |
| | Histogram window $L_W$ | 500 frames |
| VAD | Smoothing parameters $\alpha_{\sigma\uparrow} \mid \alpha_{\sigma\downarrow}$ | 0.8 \| 0.91 |
| | Gain ceiling factor bounds $\eta_{\min} \mid \eta_{\max}$ | 3 dB \| 15 dB |
| | Ceiling minimum $b_{c,\min}$ | 3 |
| | Relative shift $\Delta$ | $0.2\,\mathrm{s}^{-1}$ |
| | Speech presence factor $\beta_{\mathrm{sp}}$ | $1/4$ |
| | Speech gain SNR window length $T_{\mathrm{w}}$ | $0.1\,\mathrm{s}$ $\left(\hat{=}\left\lceil \frac{f_{\mathrm{s}}}{L_A} T_{\mathrm{w}} \right\rceil = 10 \text{ frames}\right)$ |
| | Binary VAD threshold *thr* | 0.5 |

**Table 5.3:** Algorithm specific parameters

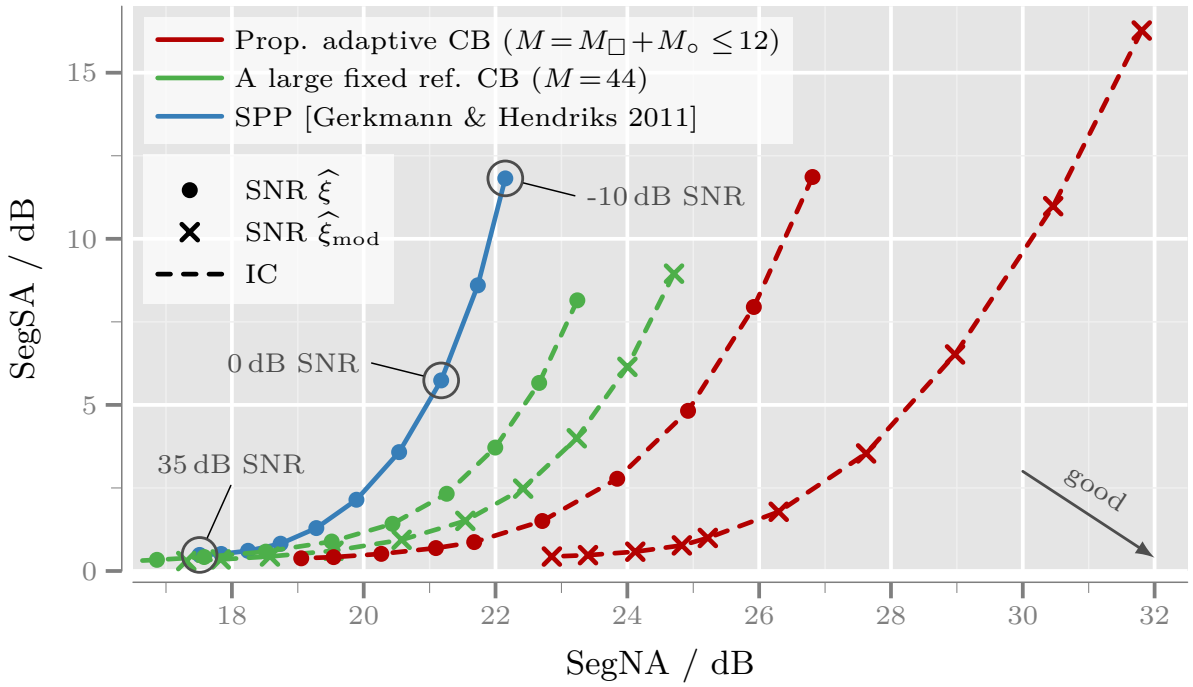



**Figure 5.12:** The *segmental speech attenuation* (SegSA) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter (IC: *information combining*, Setup: Fig. 5.3 configuration ⓐ, i. e., brute force codebook search).

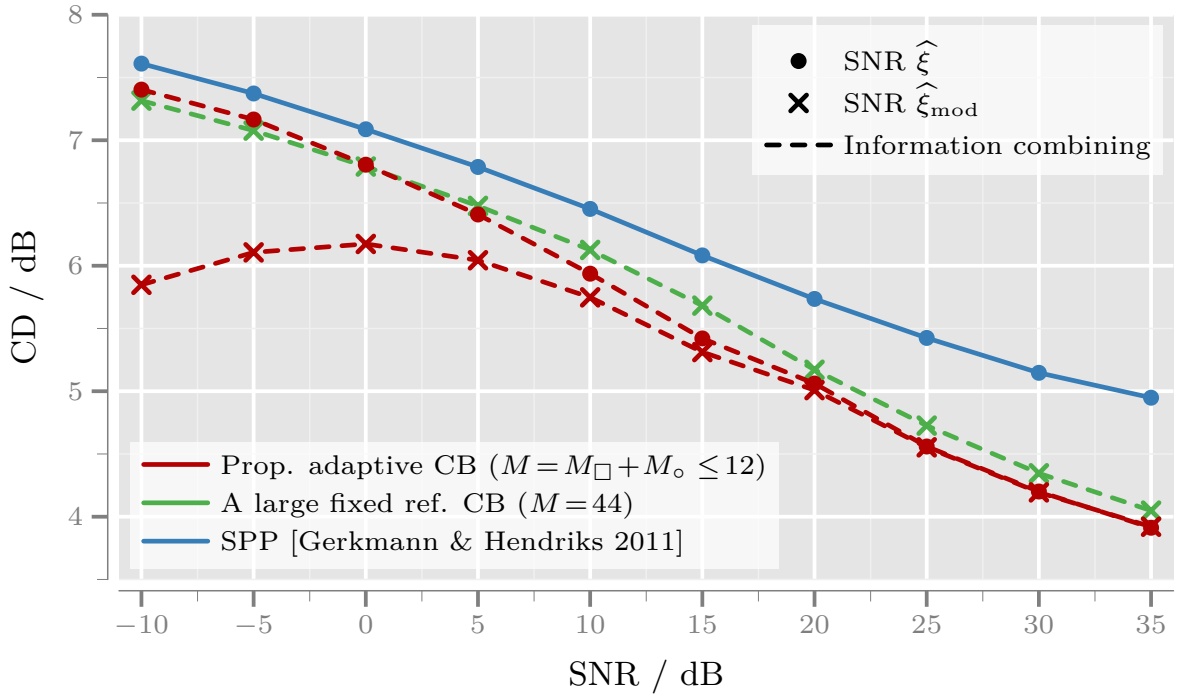



**Figure 5.13:** The *cepstral distance* (CD) is depicted over the input SNR (Setup: Fig. 5.3 configuration ⓐ, i. e., brute force codebook search).

codebook and results eventually in an unfavorable *a priori* SNR estimate. In consequence, speech is wrongly attenuated which is visible in the SegSA measure for low input SNR up to 0 dB. Starting with 0 dB input SNR both proposed methods (−●−, −✕−) perform likewise compared with *SPP* (—●—) regarding SegSA. For high input SNR the SegSA performance is comparable with the reference codebook system (−●−, −✕− vs. −●−, −✕−) achieving the best scores.

With respect to speech distortion the CD is depicted in Fig. 5.13. Again the use of the modified decision-directed SNR estimate $\widehat{\xi}_{\mathrm{mod}}$ (−✕−) performs superior over the complete input SNR range, converging for very high input SNR. Both proposed systems (−●−, −✕−) outperform *SPP* (—●—) over the complete input SNR range. The objective scores of the reference codebook system using $\widehat{\xi}_{\mathrm{mod}}$ (−✕−) are already achieved by the proposed system utilizing $\widehat{\xi}$ (−●−) while the proposed system using $\widehat{\xi}_{\mathrm{mod}}$ (−✕−) clearly scores best.

Although the reference codebook system with configuration A exhibits four noise codebook entries per noise type, the proposed online noise codebook learning method scores significantly better. Hence, specific noise learning in the local past is beneficial and allows to provide a very precise estimate of the underlying noise signal.

However, a local maximum regarding CD is observed at 0 dB input SNR for the proposed system utilizing $\widehat{\xi}_{\mathrm{mod}}$ (−✕−). Since in Eq. (4.4) the cross-term, $2\sigma_s\sigma_n\,|\mathbb{S}_l|\,|\mathbb{N}_m|\cos(\vartheta_{\mathcal{S}}(\mu) - \vartheta_{\mathcal{N}}(\mu))$, is neglected for the codebook matching procedure, the strongest inherent estimation error can be expected for 0 dB SNR. See

Appendix D for further details. Moreover, as the input SNR increases or decreases, the estimation error decreases which results in decreased speech distortion and is confirmed by the course of the CD curve (–✖–). Since the speech estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}$ is not used for the *a priori* SNR estimate $\widehat{\xi}$, the error resulting from $\max\left(\overline{\gamma}(\lambda,\mu)-1,\ 0\right)$ in Eq. (4.19) appears to dominate. Hence, the effect of the local maximum is not visible in the CD curve (–●–) of the proposed system utilizing $\widehat{\xi}$. The same applies for the reference codebook system –✖–. Although the new SNR estimate $\widehat{\xi}_{\mathrm{mod}}$ is used, the imprecise *a priori* knowledge on noise causes the dominant error in the SNR estimation.

### 5.8.6 Complexity Reduction

The previous section confirmed a superior performance of the proposed enhancement system utilizing online noise codebook learning. This is due to the very precise estimate of the underlying noise signal. With respect to speech enhancement applications, a dramatic complexity reduction of the codebook matching process is necessary. Utilizing the VAD and *information combining* algorithm, the brute force codebook search is replaced by a cascade of gain shape VQs as suggested in Sec. 5.7.2. Hence, the codebook matching and VAD block of Fig. 5.3 is setup with configuration (b) which is detailed in Fig. 5.1. The online noise codebook learning remains. This configuration is referred as *gain shape cascade* (***GSC***) in the legend of the respective plots. Note that the legend entry IC belongs to the activated *information combining* block depicted in Fig. 5.3[6].

At first, the number of distance calculations which are necessary for either of the methods are investigated. While for the brute force codebook search the number of distance calculations is given by

$$N_{\mathrm{D,BF}} = p \cdot \underbrace{(L \cdot M \cdot N_q)}_{\text{brute force}} + (1-p) \cdot \underbrace{(M+L+2)}_{\text{gain shape + VAD}} \quad , \tag{5.61}$$

the gain shape VQ based approaches[7] need

$$N_{\mathrm{D,GSC}} = p \cdot \underbrace{(2 \cdot (L+M+2))}_{\text{gain shape cascade}} + (1-p) \cdot \underbrace{(L+M+2)}_{\text{gain shape + VAD}} \quad , \tag{5.62}$$

distance calculations. Hence, the reference codebook system with configuration A comprising the large noise codebook without VAD knowledge ($p=1$), needs $N_{\mathrm{D,BF,refFixedA}} = 90112$ distance calculations per frame. Assuming 50 % speech presence ($p=0.5$), the proposed codebook system utilizing online noise codebook adaptation and the brute force search requires only up to $N_{\mathrm{D,BF,adaptive}} = 12359$

[6] Combining GSC and IC, the *information combining* algorithm is applied twice. First, while codebook matching using the estimates provided by the cascade of gain shape VQs and second in the *information combining* block.

[7] Note that the gain calculation is taken into account by adding +2 inside the brackets as the gain calculation is similarly computational complex to the distance calculation.

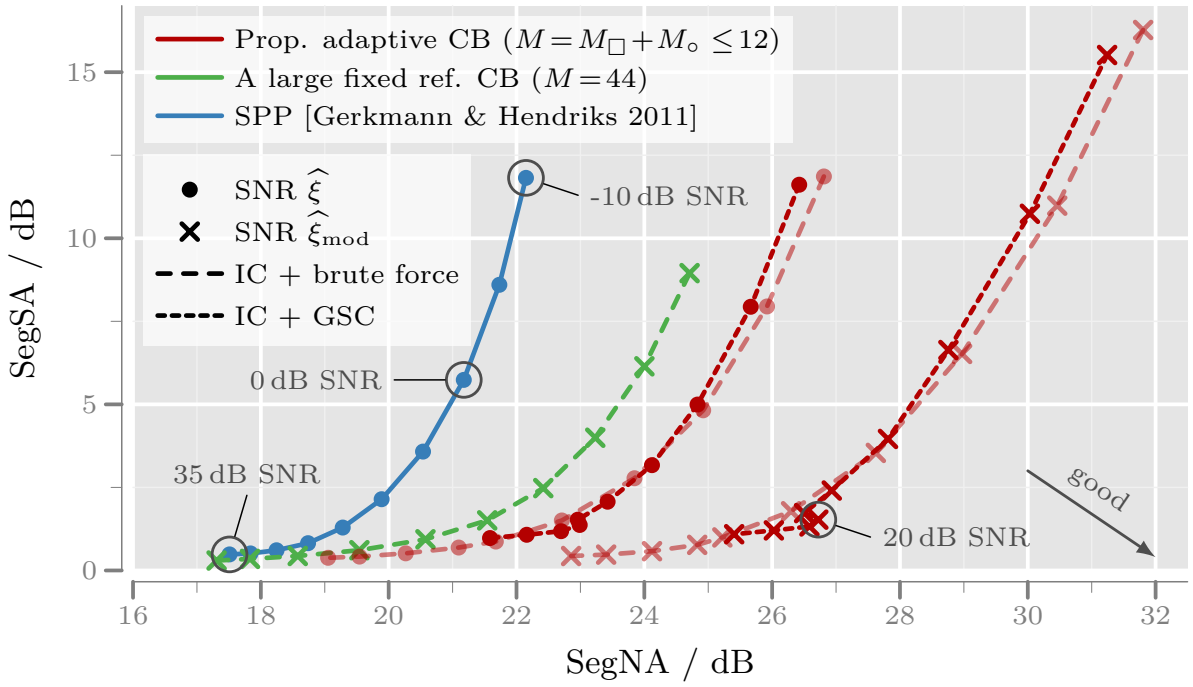



**Figure 5.14:** The *segmental speech attenuation* (SegSA) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter (IC: *information combining* block, GSC: gain shape cascade).

distance calculations on average due to the significantly reduced codebook size and the use of the VAD. In contrast, the proposed codebook system utilizing *information combining* and the full complexity reduction requires merely $N_{\mathrm{D,GSC,adaptive}} = 213$ distance calculations per frame on average, which is a tremendous complexity reduction.

In Fig. 5.14 the averaged results for SegSA plotted over SegNA with the input SNR as variable parameter are depicted. Again, the *SPP* based enhancement system (—●—) as well as the reference codebook system with the large noise codebook (–×–) are depicted for reference. For low input SNRs of up to 0 dB, the proposed system utilizing gain shape VQ cascade and the system employing the brute force search (--●--, --×-- and –●–, –×–) perform likewise independently of the SNR estimation method. However, with increasing SNR the performance diverges. For the proposed systems comprising the gain shape VQ cascade (--●--, --×--) the SegNA increases but at the expense of SegSA. This is observed for both *a priori* SNR estimates. Consulting the results from the CD measure depicted in Fig. 5.15, the increased SegSA is reflected as increased speech distortion.

The unsteady course of the curve (–×–) in Fig. 5.14 around 20 dB input SNR caused by the method utilizing the gain shape VQ cascade and the modified decision-directed SNR estimate is noted and further investigated. Hence, the averaged results are divided into four classes of noise types: **stationary** (F16, highway inside car), **transient fast** (jackhammer, wind, indoor soccer), **transient**

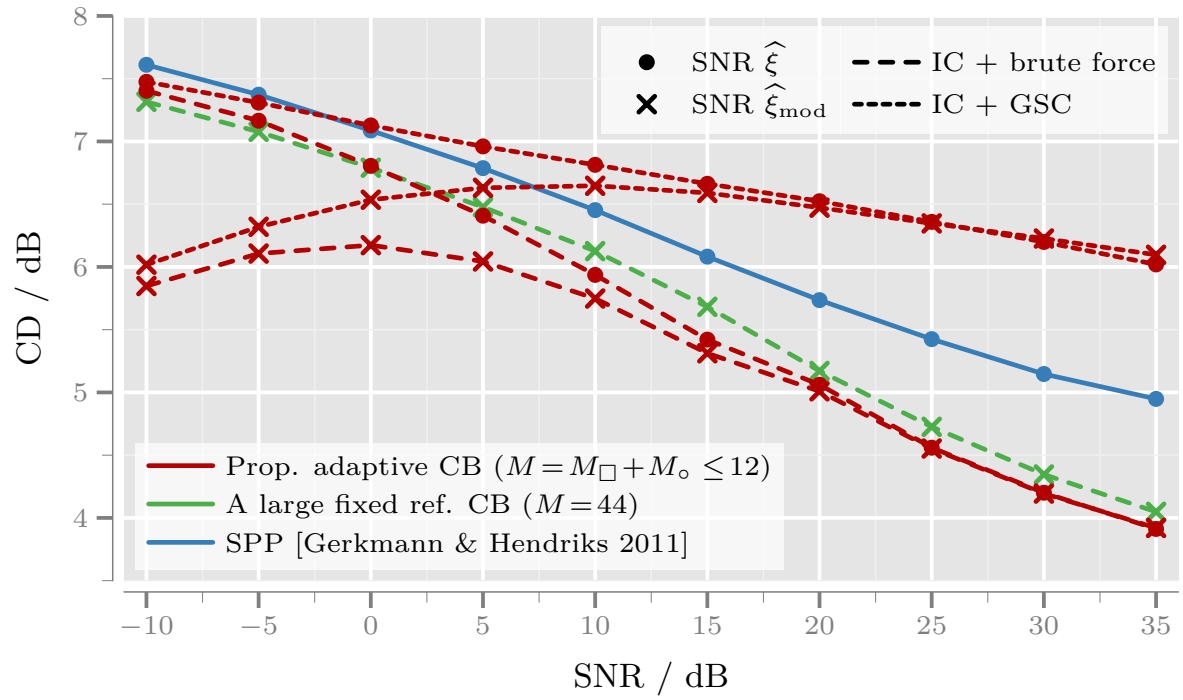



**Figure 5.15:** The *cepstral distance* (CD) is depicted over the input SNR (IC: *information combining* block, GSC: gain shape cascade).

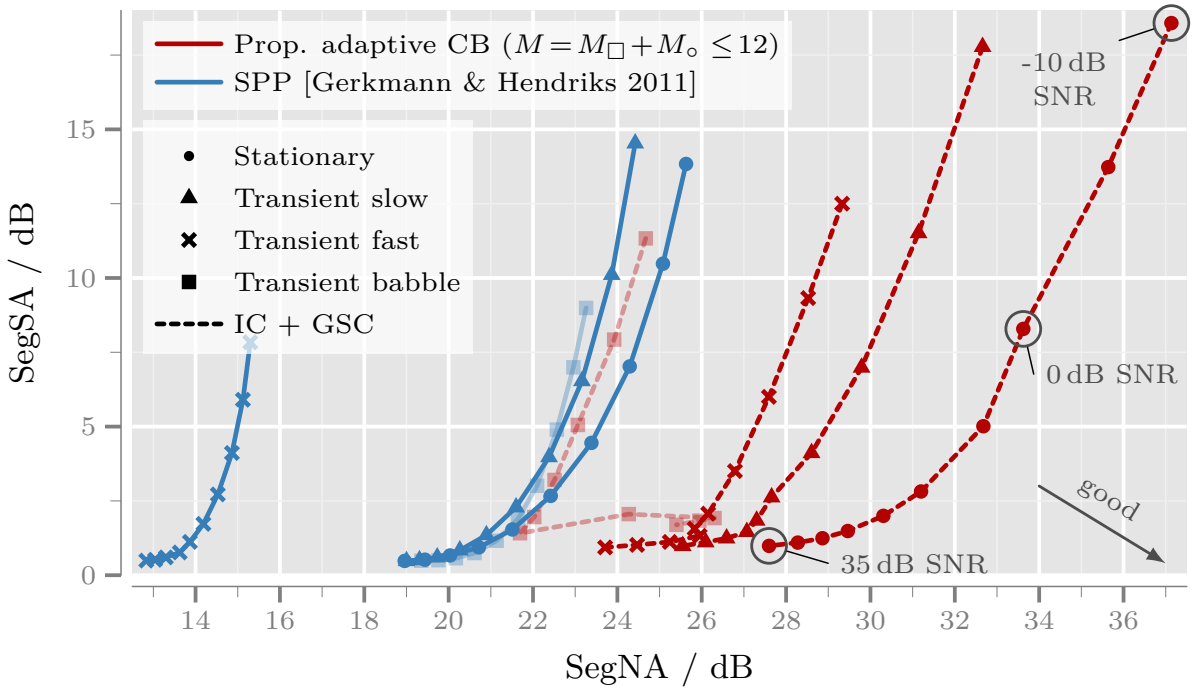



**Figure 5.16:** The *segmental speech attenuation* (SegSA) is depicted over the *segmental noise attenuation* (SegNA) with the input SNR as variable parameter for different noise types (IC: *information combining*, GSC: gain shape cascade).

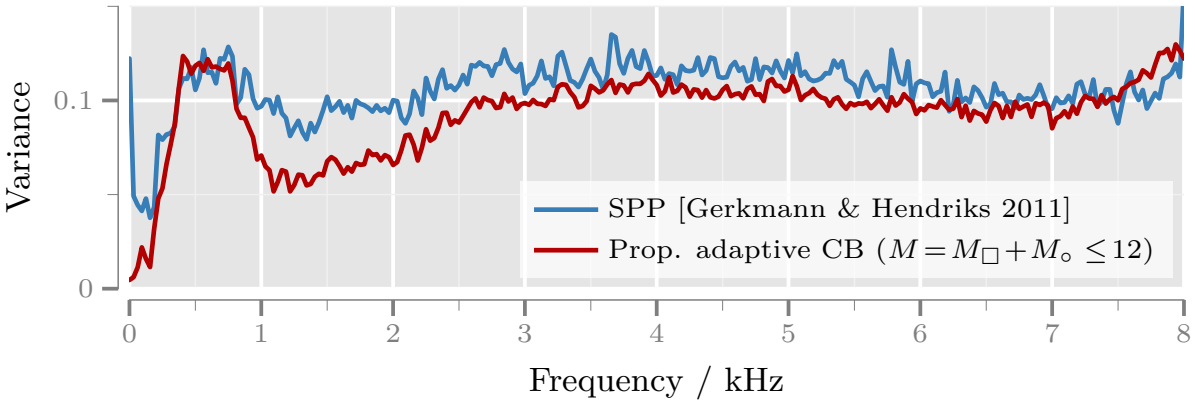



**Figure 5.17:** Variance of spectral weights $\mathcal{G}(\lambda, \mu)$ depicted over the frequency.

**slow** (outside traffic road, living-room, train station, forest, modulated Gaussian noise), and **transient babble** (pub noise, inside train). The results of each noise class are depicted in Fig. 5.16. For the sake of clarity, the results from the transient babble noise class are presented transparent. From Fig. 5.16 it is obvious that the observed unsteady course is caused by babble noise (--■--). Due to the inherent problem with speech-like noise types this behavior can be expected. Since speech seems to be present all the time, as wrongly indicated by the VAD, the noise codebook does not exhibit suitable noise knowledge. Hence, the noise estimate after codebook matching is underestimated. This is true for both codebook matching approaches, but more distinctive for the gain shape VQ cascade driven approach. Starting with 15 dB input SNR, babble noise is detected reliably as noise by the proposed VAD. Thus, the observed performance leap in the SegNA measure is the result at 20 dB input SNR.

The remaining noise type classes (--●--, --▲--, --✖--) describe a steady course over the input SNR. However, a slightly increased SegSA of approx. 1 dB is observed for very high input SNR and for very low input SNR the SegSA is also increased.

In addition, Fig. 5.16 allows to compare the proposed noise reduction system utilizing the full complexity reduction (——) with a conventional state-of-the-art system utilizing the *SPP* noise estimator (——). With the exception of babble noise, a tremendous performance gain is achieved by the proposed noise reduction system, especially for the noise type class "transient fast", e. g., with up to 12 dB SegNA improvement.

In order to analyze the occurrence of *musical tone* artifacts, a noisy input signal is generated consisting of ten different, six seconds long stationary and non-stationary noise types mixed with five male and female English speakers taken from the TIMIT database at 0 dB SNR. The variance of the spectral gains $\mathcal{G}(\lambda, \mu)$ is depicted in Fig. 5.17 over the frequency for the conventional state-of-the-art (——) system and the proposed system (——). Although the proposed system achieves best scores regarding noise attenuation (26.7 dB vs. 18.6 dB SegNA) and speech distortion (6.2 dB vs. 6.5 dB CD) in this example, the variance of the spectral weighting gains is decreased. This is an indicator for strongly reduced *musical*

*tones*, which is confirmed by informal listening test. In contrast to the conventional noise reduction, *musical tones* are almost removed.

## 5.9 Summary

A generic solution is formulated for the joint speech and noise estimation problem given the noisy observation. The solution considers several speech and noise estimates and provides optimal mixing coefficients with respect to minimized estimation error power regarding the noisy observation. At first, various estimates of the noisy observation are computed by permuting all different speech and noise estimates. Based on a distance measure between these estimates and the noisy observation optimal mixing coefficients for each frequency bin of the individual speech and noise estimates are determined. Applying the optimal mixing coefficients to the individual speech and noise estimates yields the final refined speech and noise estimate. This procedure is called *Information Combining*.

Although the proposed noise codebook online adaptation minimizes the probability of missing *a priori* knowledge on noise, it is not guaranteed that an appropriate codebook entry is available for each noisy observation $\mathcal{Y}$, e. g., due to changing noise while speech is present. In such cases, a second noise estimate $\widehat{\mathcal{N}}_{\mathrm{Stat}}$, e. g., provided by the newly proposed statistical noise estimator *Baseline Tracing*, is additionally considered. Utilizing the codebook driven speech estimate $\widehat{\mathcal{S}}_{\mathrm{CB}}$ and the two different noise estimates $\widehat{\mathcal{N}}_{\mathrm{CB}}$, $\widehat{\mathcal{N}}_{\mathrm{Stat}}$, two estimations $\widehat{\mathcal{Y}}_{1,2}(\lambda,\mu)$ for the noisy observation $\mathcal{Y}$ are computed. For both permutations $\widehat{\mathcal{Y}}_{1,2}(\lambda,\mu)$ a distance to the noisy observation is calculated. Based on these distances, optimal mixing coefficients for each frequency bin are calculated which minimize the estimation error power regarding the noisy observation. Subsequently, both estimates are merged resulting in the refined noise estimate $\widehat{\mathcal{N}}_{\mathrm{IC}}$. Given a second speech estimate, e. g., from the last enhanced frame of the speech enhancement system, the *Information Combining* is likewise extended to also provide a refined speech estimate $\widehat{\mathcal{S}}_{\mathrm{IC}}$. The evaluation verified a tremendous improvement of noise attenuation, while the speech distortion is reduced simultaneously.

With respect to feasible applications, e. g., mobile phones, a significant complexity reduction is necessary which is accomplished by replacing the brute force codebook matching. In a first step of the complexity reduction, the brute force search is only applied in phases of speech activity exploiting the information from VAD. During speech absence a gain shape VQ is utilized, i. e., the spectral shape is determined using the noise codebook in a first step and the calculation of the associated gain in a second step. With respect to gain shape VQ two scenarios exist which allow the substitution of the brute force codebook matching. Given a very high SNR, the brute force search can be replaced by gain shape VQ utilizing a speech codebook and setting the noise estimate to zero. In the opposed case, a noise codebook can be utilized and the speech estimate yields zero. With these considerations, two cascades of gain shape VQs are constructed. The first cascade consists of gain shape VQ utilizing a noise codebook which provides the estimate $\widehat{\mathcal{N}}_{\mathrm{GS1}}(\lambda,\mu)$.

Hereafter, a second VQ utilizing a speech codebook processes $\max(\mathcal{Y} - \widehat{\mathcal{N}}_{\mathrm{GS1}}, 0)$ and yields the speech estimate $\widehat{\mathcal{S}}_{\mathrm{GS1}}(\lambda, \mu)$. The second cascade is structured vice versa. While the first cascade provides a reliable noise estimate for low SNR, the second cascade provides a robust speech estimate for high SNR. Utilizing the *Information Combining* procedure again, the best of all estimates is combined yielding the final estimates of speech and noise. Hence, the brute force search is replaced by four VQs and subsequent *Information Combining*. The proposed complexity reduction decreases the number of distance calculations by a factor of approximately 60 from 12359 to 213. The proposed codebook matching exhibiting the full complexity reduction is characterized by almost the same performance regarding noise and speech attenuation, but at the expense of moderately increased speech distortion compared with the brute force search. Hence, *Information Combining* can be used for both, improving the estimation quality and reducing the complexity.

The proposed codebook based noise reduction system clearly outperforms conventional state-of-the art noise reduction systems. The evaluation of the new modified decision-directed *a priori* SNR estimate $\widehat{\xi}_{\mathrm{mod}}$, incorporating also the speech estimate, confirmed a superior performance. While the noise attenuation is improved, the speech distortion is reduced at the same time. A tremendous performance gain is achieved, especially for transient and fast types of noise, e. g., of up to 12 dB improved noise attenuation compared with state-of-the art systems. Although the proposed system achieves best scores regarding noise attenuation and speech distortion, the variance of the spectral weighting gains is decreased compared with conventional systems. This is an indicator for strongly reduced *musical tone* artifacts, which is confirmed by informal listening tests. *Musical tones* are almost removed.

# Chapter 6

# Real-Time Implementation

## 6.1 High Quality Video Conferencing

A multimodal signal processing concept is proposed [CoVR 2013; Schlien et al. 2013] which is suitable for flexible high-quality multi-point video conferencing. In contrast to other, commercially available high quality solutions, e. g., from Cisco, Tandberg, and Polycom, this system has been intentionally designed for off-the-shelf consumer electronics at low cost. The desired high-quality communication is achieved by a novel integration of dedicated algorithms for signal analysis and signal enhancement, combined with state-of-the-art coding and transmission techniques. The proposed multimodal signal processing concept enables a new audio-video scene composition as a key feature, where the most active participants are placed side by side in a virtual conference at the receiver (see Fig. 6.1). The gained information is further employed to control the media encoders for improved compression efficiency. The identification and extraction of the talkers – and their audio and video signals – represents the major challenge, especially with multiple participants at the clients.

The technical focus regarding video analysis is on face detection and tracking. For audio analysis, near field beamforming has been identified as the most important aspect. The results of the video analysis are input to the audio analysis. Besides the classical task of attenuating competing sound sources and background noise, the beamformer outputs are further used for speaker activity estimation. Metadata generated from these analyses is further exchanged and exploited in the network side processing and the receivers.

Concepts of joint processing or multimodal fusion for improved multimedia signal analysis have been a long-term research topic. In [Bub et al. 1995] an early scheme for visually guided beamforming has been proposed. A system with a video camera and two microphones has been discussed, e. g., in [Zhou et al. 2008]. Recently, [Minotto et al. 2014] used an eight-microphone-array and a video camera for multimodal voice activity detection and sound source localization. Related work on participant detection, localization and composition of audio-visual signals has been presented, e. g., in [Jansen et al. 2011; Q. Liu et al. 2014; Zhang et al. 2008]. A general survey on multimodal fusion can be found in [Atrey et al. 2010]. [Strobel et al. 2001] provide a good overview on joint audio-video object localization.

For evaluation and demonstration of the proposed concept, a real-time prototype


**(a)** Demonstrator **(b)** Scene composition

**Figure 6.1:** Real-time demonstrator setup and illustration of a scene composition of most active talkers in a video conference

was developed. This prototype enables experiments and evaluations of the proposed algorithms in real environments and conditions. The real-time prototype video conferencing system is *IP Multimedia Subsystem* (IMS) compliant. It consists of consumer electronics, namely stereo loudspeakers, eight microphones, a video camera, and two LCD-TVs. A single quadcore PC is able to perform all audio and video signal processing and all IMS-related services of the client. The setup of the real-time demonstrator is shown in Fig. 6.1a.

The first presentation of the demonstrator was on the *International Workshop on Acoustic Signal Enhancement* (IWAENC 2012) [Hamm et al. 2012]. Here, the individual audio and video parts were showcased. These were the beamformer, the audio rendering, the *artificial bandwidth extension* (BWE), and the person detection and tracking. Techniques for BWE [Jax & Vary 2003] extend the limited audio bandwidth of current narrowband telephone systems (0.3 – 3.4 kHz) towards the wideband frequency range (0.05 – 7 kHz). The BWE is applied to the narrowband signal, e. g., from telephone dial-in, to align the speech quality in terms of the acoustic bandwidth. This improves the speech intelligibility as well as the listening comfort in heterogeneous communication scenarios [Heese et al. 2012a]. For the demonstrations at the CeBIT 2013 and the *Workshop Audiosignal- und Sprachverarbeitung* (WASP) at the INFORMATIK 2013 [Schlien et al. 2013], the complete demonstrator was showcased. Live conferences between the demo location and two clients at RWTH Aachen University were established. Thereby, the final scene composition output as well as the effect of the BWE for telephone dial-in were successfully demonstrated in a real-life scenario.

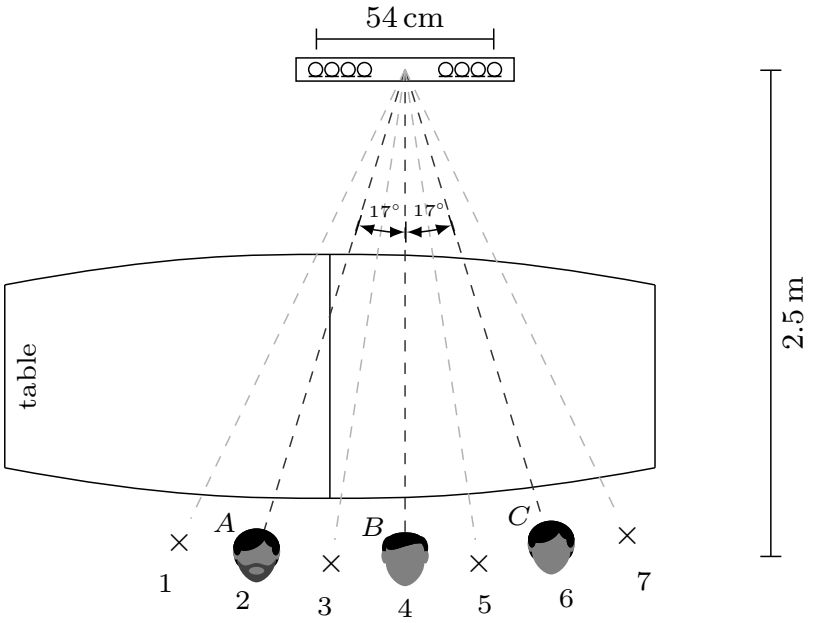



**Figure 6.2:** Experimental setup: video conferencing scenario

### 6.1.1 Evaluation of Speaker Activity Estimation

The identification and separation of the most active talkers is the task of the proposed multimodal signal analysis. The resulting activity indices jointly comprise the information from the audio and video analysis. For their evaluation an assessment of the spatial information from video as well as the activity estimation from audio is needed. This section focuses on the evaluation of the audio analysis and the impact of the spatial information provided by the video analysis. The stability of the tracking algorithm and the possible bitrate savings due to *region of interest* (ROI) encoding have been evaluated in [Hosten et al. 2013] and [Bulla et al. 2013].

A typical video conference scenario was arranged in a room, as depicted in Fig. 6.2. The room has a reverberation time of 0.32 s and the eight sensors of the microphone array have a spacing of 4 cm with a gap of 30 cm for the camera mounting in the center. The sampling frequency is set to $f_s = 48$ kHz. Audio and video signals of three participants were recorded with a duration of about 5 min comprising single- and multi-talk sequences in different variations. The three participants were placed in front of the microphone array at a distance of 2.5 m at $0°$, $-17°$, and $17°$ azimuth in the horizontal plane.

The performance of the proposed audio analysis, i. e., beamforming and soft voice activity detection (VAD), is evaluated in comparison to a conventional beamformer and two other soft VAD systems, respectively. All possible combinations of beamforming and soft VAD are applied to the test signals such that six different combinations are tested in total. The configurations of the two beamformers are:

- **Near Field Beamformer (NFB)** - as proposed, cf. Sec. 2.1
  The NFB is configured with $N = 6$ non-uniform sub-bands. The corresponding frequency range of each sub-band is given in Table F.1. The degree of the

FIR sub-band beamformer filters $\mathbf{h}_n^m$ was set to $L = 4$. In advance, seven individual filter sets for the near field beamformers were pre-computed for seven talker positions in front of the array with an azimuthal resolution of 9°. During runtime, the video analysis unit first detects the number and the positions of participants. With respect to these positions the most appropriate filter sets for the near field beamformers are chosen for each detected participant.

- **Delay-and-Sum Beamformer (DSB)** according to [Laakso et al. 1996]
  Usually, this conventional beamformer operates under the assumption that the target is far away from the microphone array. This allows the usage of simple geometric rules for the determination of the parameters. Because the far-field assumption does not hold in this application, the delay from each participant to the eight microphones was calculated using the correct distances on the basis of the provided angle from the video analysis assuming a distance of 2.5 m. This represents the best-case scenario for this beamformer.

For the soft VAD the following approaches are compared:

- **Activity Index (AI)** - as proposed, cf. Sec. F.1
  The activity index calculation was configured with an audio frame length of $T_{\mathrm{F}} = 30\,\mathrm{ms}$ and a maximum statistics buffer $\overline{V}_{\mathrm{Buffer}}(\lambda)$ which contains frames of the past 30 s. The initial energy normalization parameters were set to $L_N = 30$, $\gamma = 0.8$, and $\overline{V}_{\mathrm{min}} = 0.08$,
- **Ghosh** according to [Ghosh et al. 2011],
- **Sohn** according to [Sohn et al. 1999].

Both conventional soft VAD (Ghosh, Sohn) systems are parameterized as suggested in their original publications. The task of the audio analysis is to mark the phases of activity and inactivity individually for the three talkers, i. e., to perform a speaker activity estimation as a function of time *and* space.

Finally, an objective evaluation of the six combinations is performed which is based on a numerical comparison of the VAD $v_{\mathrm{bin,n}}(\lambda)$ with the ground truth VAD $v_{\mathrm{true,n}}(\lambda)$, where $n$ represents the participant index. Thus, for each soft VAD value a corresponding hard decision value $v_{\mathrm{bin,n}}(\lambda)$ is derived named VAD-AI, VAD-Ghosh, and VAD-Sohn, respectively. For a better comparability, the thresholds for the determination of these hard decision values were adjusted such that all systems yield detection rates in the same order of magnitude. The numerical evaluation is performed in terms of three VAD measures[1]:

- *Accuracy rate* $P_{\mathrm{a}}$: Percentage of signal frames with correct VAD estimation;
- *Detection rate* (or true positive rate) $P_{\mathrm{d}}$: Fraction of active speech frames that are detected correctly;
- *False alarm rate* (or false positive rate) $P_{\mathrm{f}}$: Fraction of speech frames without speech that are classified erroneously as speech.

All VAD measures for the six combinations are detailed in Appendix F.2.

[1] The objective scores are detailed in Appendix C.4

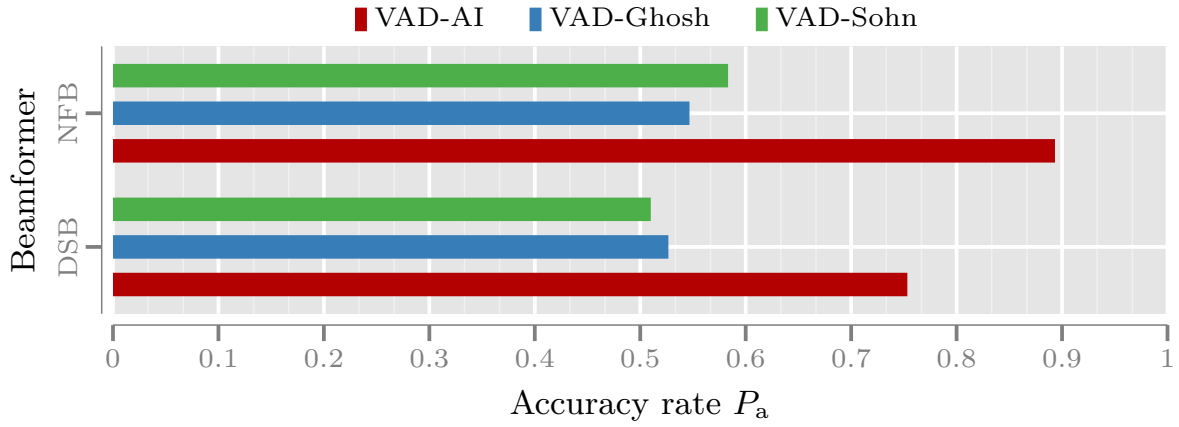



**Figure 6.3:** Accuracy rate $P_a$ for different beamformers and VAD algorithms. The results are evaluated and averaged over the occupied talker positions as indicated by the video analysis stage.

Fig. 6.3 summarizes the VAD accuracy rate $P_a$ by averaging over the three talker positions. It is obvious that the proposed near field beamformer (NFB) in combination with the VAD-AI (▮) achieves the best score with a $P_a$ of 0.88. With respect to the accuracy rate $P_a$ the remaining combinations of beamformers and VAD algorithms (▮, ▮) show a significant performance loss. Comparing the NFB and DSB, all VAD algorithms exhibit better scores when combined with the NFB.

Comparing the results of all talkers individually in Tab. F.2–F.4 the best rates are observed for the proposed system, i. e., detection rates $P_d$ in an order of 80 % and false alarm rates $P_f$ of 5-6 %. It is of special interest, that these results are consistent among the three speakers reflecting the robustness of the approach. As the hard decision thresholds were adjusted to yield similar detection rates, the low accuracy rates $P_a$ of the conventional VAD algorithms (Ghosh, Sohn) are due to extremely high false alarm rates $P_f$.

In a last experiment, the performance of the stand alone audio analysis is investigated, i. e., without location information from the video analysis stage. In this case, beamformers have to be run in parallel for all possible participant positions (i. e., positions 1 to 7 according to Fig. 6.2). For the evaluation, however, the focus is on the results for the most critical non-occupied positions 3 and 5. Here, cross-talk from neighboring occupied positions on both sides can occur.

The performance is quantified in terms of accuracy rate $P_a$ and false alarm rate $P_f$.[2] The detailed results can be found in Appendix F.2, Tab. F.5. The corresponding averaged accuracy rates $P_a$ are depicted in Fig. 6.4. Again the combination of NFB with VAD-AI (▮) leads to the best performance. It is notable that this result of $P_a \approx 0.87$ is almost equal to the result of Fig. 6.3 with $P_a \approx 0.88$. In contrast, the performance degradation of the five remaining combinations of beamformers and VAD algorithms is even more pronounced.

[2] Since there are no talkers on these positions, the detection rate $P_d$ can not be calculated ($N_d = 0$).

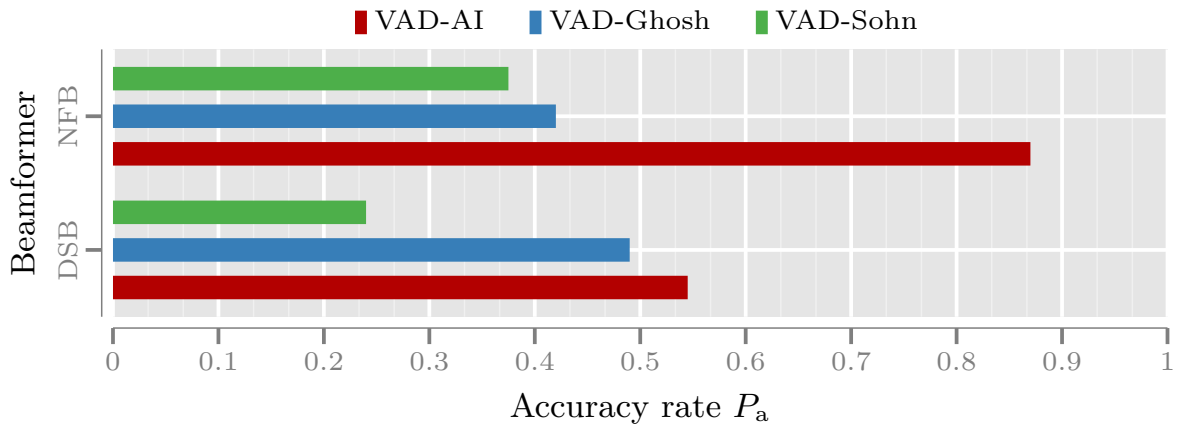



**Figure 6.4:** Accuracy rate $P_a$ for different beamformers and VAD algorithms. The results are evaluated and averaged over the most critical non-occupied talker positions 3 and 5. This experiment illustrates the case if no video information would be available.

This experiment shows the reliability and the stability of the audio analysis by itself and additionally allows to quantify the impact of the video analysis on performance and complexity. If the complementary information from the video analysis is used, the accuracy rate for the non-occupied positions increases to 100 %, i. e., 13 percentage points better than by single-modal audio analysis. In terms of computational complexity, a reduction of over 50 % for the beamformer unit can be achieved in this scenario as only three beamformers instead of seven are active in parallel. Since the talker positions are necessary anyway for ROI encoding and scene composition, no additional complexity is required for the video analysis stage.

If the accuracy rates over all positions 1 to 7 are averaged for the case when video information is available, the overall accuracy rate results in 95 %. At this accuracy, virtually no artifacts in the scene composition of the demonstrator system occur. In practice, all experiments with the real-time prototype using the NFB and the VAD-AI verified a stable and reliable operation of the audio-visual scene composition as indicated by the objective scores.

## 6.2 Real-Time Speech Enhancement for Mobile Phones

A real-time prototyping platform is proposed for rapid implementation, demonstration, and evaluation of speech enhancement algorithms. The focus is on telephone applications. The *FreeSWITCH – Open Source multi platform Software-PBX* [Minessale et al. 2013] is an appropriate software basis for this purpose and supports all eligible operating systems (Linux, OS X, Windows). It provides a powerful audio signal processing *application programming interface* (API) and implements necessary core features such as transcoding of audio codecs and media handling. Its small system requirements makes it attractive for *single-board computer*s (SBCs). Moreover, various communication technology back-ends are integrated, e. g., *voice*

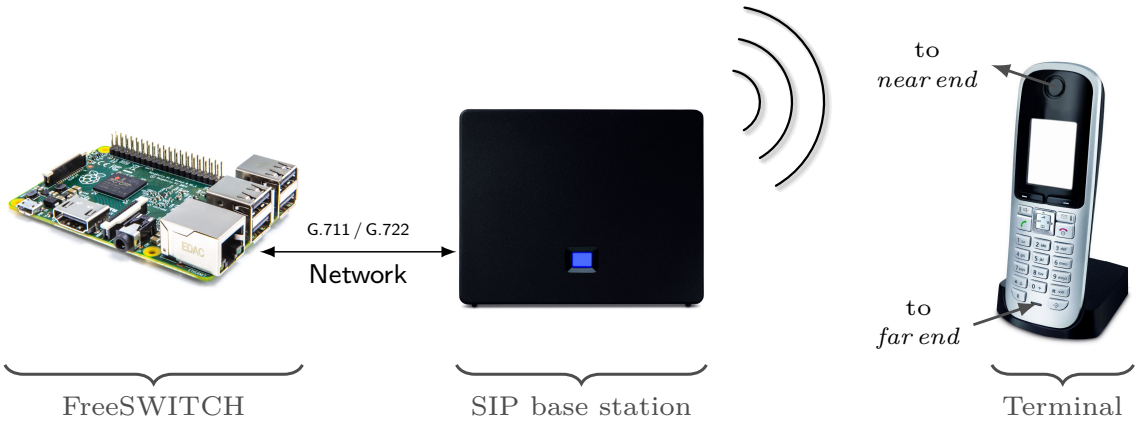



**Figure 6.5:** Real-time speech enhancement for telephone applications.

*over IP* (VoIP), *public switched telephone network* (PSTN), and *integrated services digital network* (ISDN).

Utilizing a SBC such as a Raspberry PI, a compact and comprehensive real-time demonstrator, as depicted in Fig. 6.5, is assembled at low cost. Due to its geometric size, this setup enables experiments under real conditions at various locations with minimum effort. The SBC hosts FreeSWITCH as well as a simple user interface using several LEDs. A standard CAT-iq capable SIP base station serves as gateway between the terminal and FreeSWITCH. The terminal is a standard wideband CAT-iq handset and supports the ITU-T G.711A (300 Hz – 3.4 kHz) and G.722 (50 Hz – 7 kHz) audio codecs. Furthermore, the terminal is able to remote control FreeSWITCH via *dual-tone multi-frequency signaling* (DTMF). Hence, interactive control of the speech enhancement algorithms is possible. Since the audio processing is organized as a processing chain, the interaction of different speech enhancement algorithms can be investigated.

Implemented applications are speech enhancement for the *near end* as well as the *far end*, e. g.,

- **Noise reduction** A noisy *near end* signal captured by the microphone of the handset, is enhanced for the *far end* by applying noise reduction, cf. Chap. 3, 4 and 5;

- ***Near-end listening enhancement* (NELE)** On the other hand, the intelligibility of a clean *far end* signal, perceived in strong *near end* environmental noise, is enhanced by a pre-processing of the *far end* signal, e. g., [Heese et al. 2014b; Sauert et al. 2014; Sauert & Vary 2010; Shankar Chanda & Park 2007];

- ***Artificial bandwidth extension* (BWE)** The limited audio bandwidth of narrowband telephone systems (300 Hz – 3.4 kHz) is extended towards the wideband frequency range (50 Hz – 7 kHz) [Jax & Vary 2003]. This improves the speech intelligibility [Heese et al. 2012a] as well as the listening comfort in heterogeneous communication scenarios.

The demonstrator was showcased on the *International Workshop on Acoustic Signal Enhancement* (IWAENC 2012) [Hamm et al. 2012] demonstrating the BWE, on the *ITG Fachtagung Sprachkommunikation* (ITG 2014) [Heese et al. 2014b] presenting NELE, and on the *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2014) [Sauert et al. 2014] performing NELE as well as statistical based noise reduction.

### 6.2.1 Codebook Based Noise Reduction

Considering the computational capabilities of current SBCs, a further complexity reduction of the proposed codebook based noise reduction is necessary, cf. Sec. 5.7 and Sec. 5.8.6. Since the speech codebook comprises spectral envelopes, the estimation performance of speech remains behind compared with the codebook driven noise estimation, which exhibits the full spectral resolution.

In a first step, the speech codebook size is investigated. Figure 6.6 depicts different *speech codebooks* comprising $L \in \{4, 5, 8\}$ entries. For five or more codebook entries, voiced and unvoiced sounds are modeled separately, e. g., entry number one models the *long-term speech spectrum average* (LTA) while entry number four characterizes unvoiced sounds. With respect to the mentioned estimation quality of speech, the codebook size is decreased from 128 entries to $L = 5$, which reduces the number of distance calculations according to Eq. (5.62) from 213 to $N_{\mathrm{D,GS}} = 28.5$ on average using $M = 12$ *noise codebook* entries.

With respect to the algorithmic complexity, the codebook driven VAD is identified as expensive. In each frame $\lambda$ the speech codebook is adapted to the current noisy pitch, employing a cepstral approach (cf. Sec. 4.4.2), which is computational expensive. Hence, the influence of pitch adaptation on the VAD performance is analyzed. For this purpose, the same benchmark as in Sec. 4.4.4 is carried out, except that the speech codebook size is set to $L = 5$ entries (same training sequence of 3073 s) and the speech codebook pitch adaptation shown in Fig. 4.6 is deactivated.

When applying a VAD, a compromise between detection-rate $P_{\mathrm{d}}$ and false-alarm-rate $P_{\mathrm{f}}$ has to be made by choosing an appropriate threshold. This compromise can be visualized, utilizing a *receiver operating characteristic* (ROC) curve as a function of varying thresholds[3]. A fixed but arbitrary threshold corresponds to a specific point on the ROC curve. The averaged results for the original and modified codebook based VAD are depicted in Fig. 6.7 separately for various *signal-to-noise ratio*s (SNRs). It is notable, that a significant difference between the codebook based approaches turns out only for low input SNR values below 5 dB. The effect is more pronounced for large thresholds resulting in small false-alarm rates $P_{\mathrm{f}}$. Since the pitch adaptation of the speech codebook has its major influence at very low SNR conditions and in particular for large thresholds, it is discarded in the following.

In order to analyze the impact of the modified VAD and the reduced number of speech codebook entries with respect to the noise reduction, a further benchmark

[3] For the sake of clarity, the thresholds $thr \in \{0, 1\}$ are discarded in the presented figure.

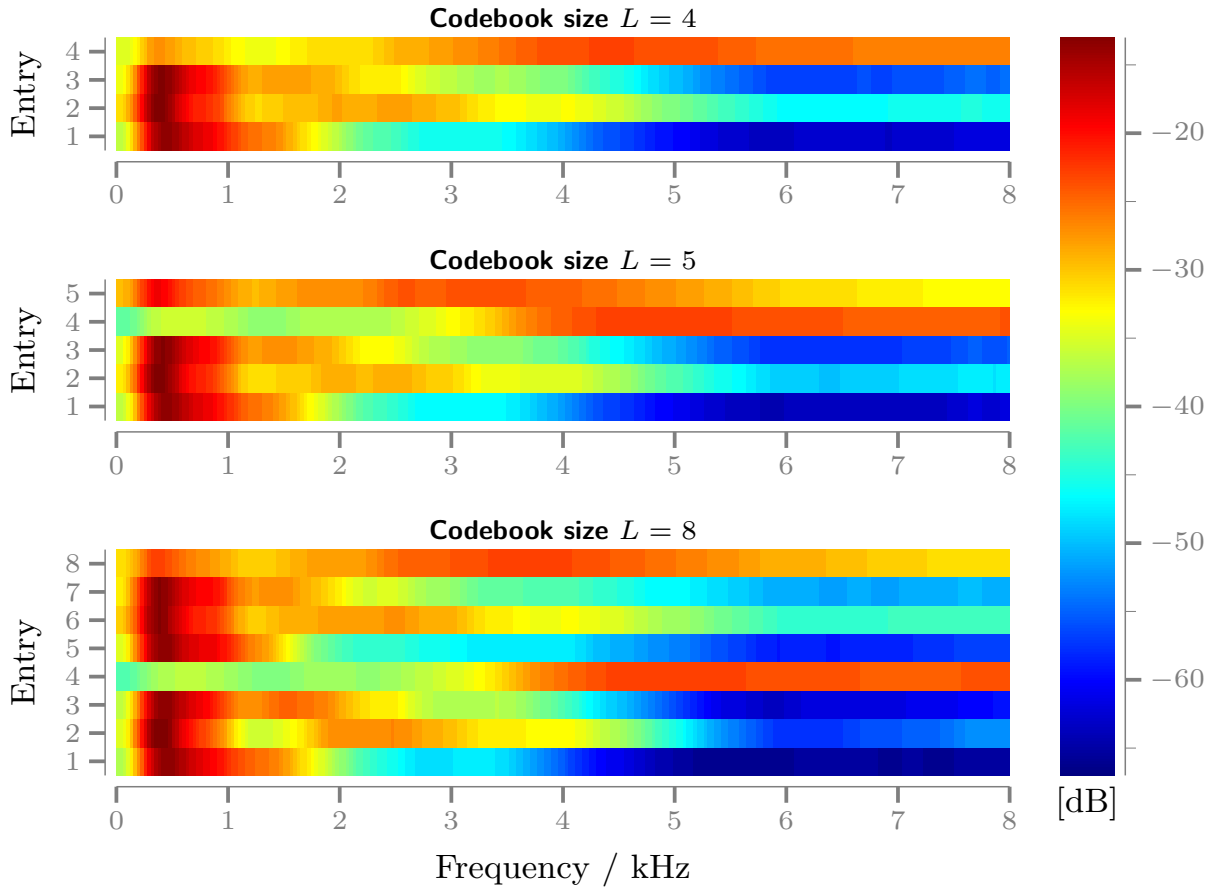



**Figure 6.6:** Different speech codebooks are depicted with various codebook sizes as spectrograms.

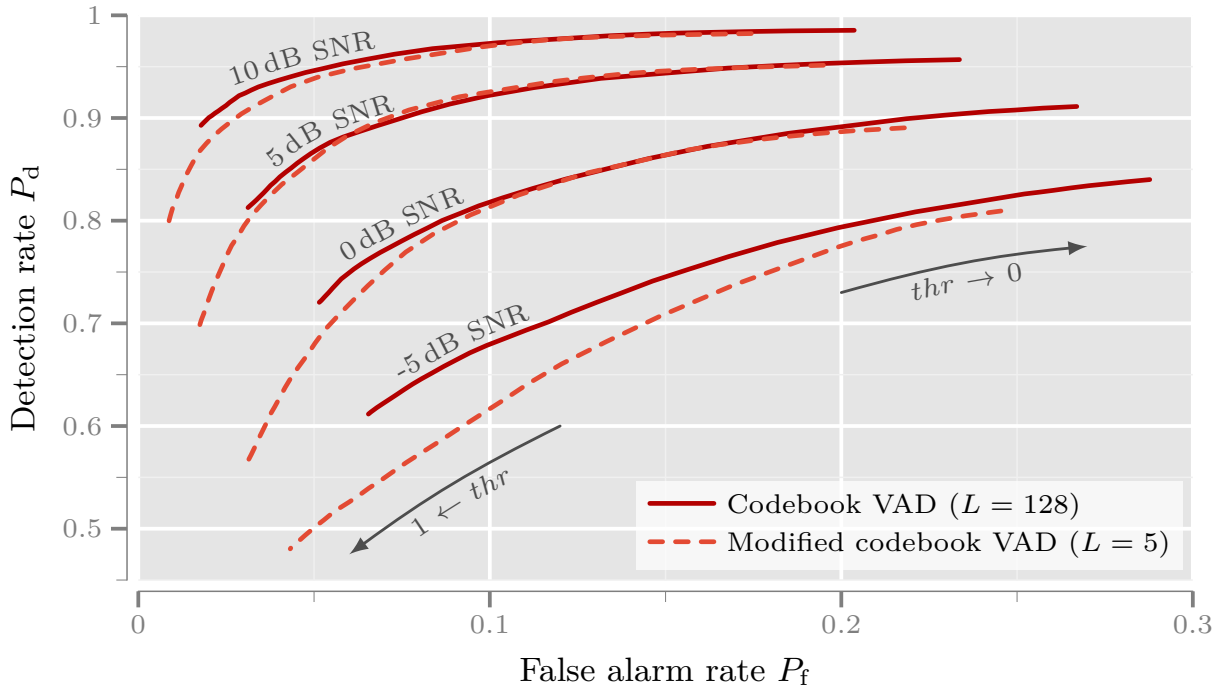



**Figure 6.7:** The ROC curves are depicted for a varying threshold $thr$ and for various input SNRs.

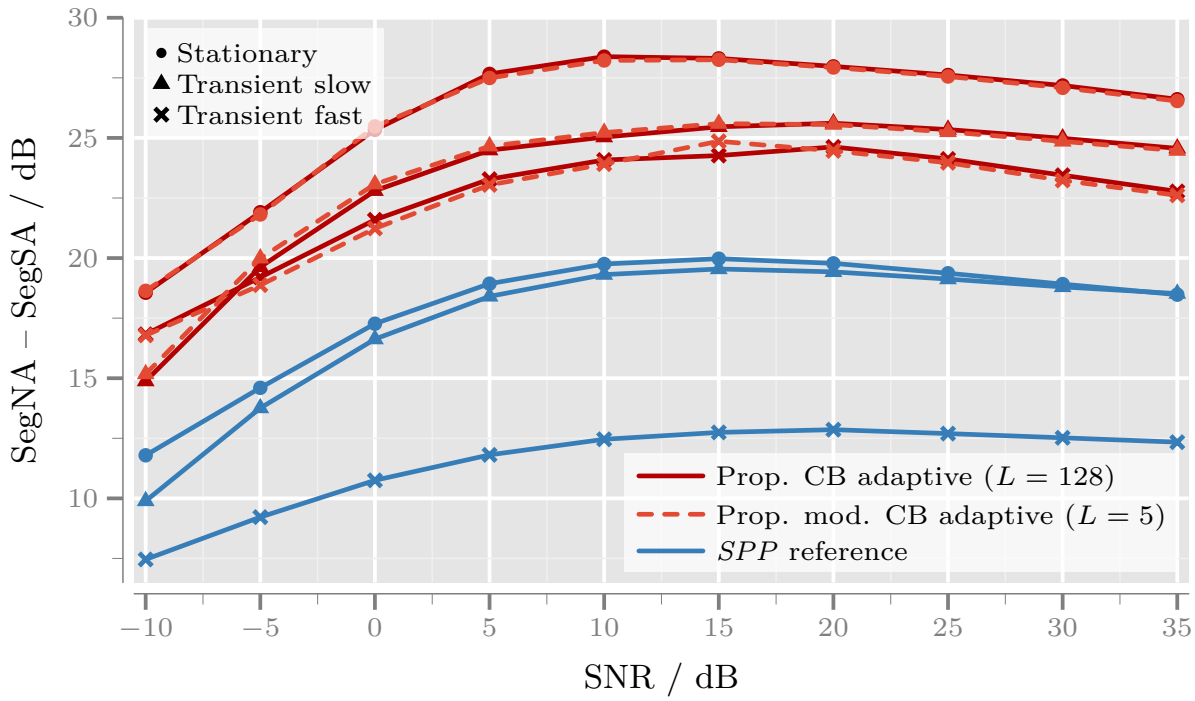



**Figure 6.8:** The difference between SegNA and SegSA is depicted over the input SNR for various types of noise.

is carried out. Thus, the associated benchmark from Sec. 5.8.6 is repeated using the modified VAD and $L = 5$ speech codebook entries and $M = 12$ noise codebook entries. Figure 6.8 presents the averaged results in terms of the difference between *segmental noise attenuation* (SegNA) and *segmental speech attenuation* (SegSA) depicted over the input SNR. The difference between SegNA and SegSA corresponds to the noise reduction performance. In addition, the results of the *SPP* based enhancement system (——), as described in Sec. 5.8.1, are depicted for reference.

It is obvious that the difference among the codebook based enhancement systems (——, - - -) regarding SegNA – SegSA is negligible, independent of the type of noise (•, ▲, ✖). Comparing the conventional state-of-the-art system utilizing the *SPP* noise estimator (——), a tremendous performance gain is achieved by the proposed modified noise reduction system (- - -), especially for the noise type class "transient fast", e. g., up to 11 dB SegNA – SegSA improvement.

In addition, informal listening tests confirmed a very similar performance among the codebook based noise reduction systems. However, the codebook based system comprising $L = 5$ speech codebook entries (- - -) produces artifacts very rarely. These are caused due to a slightly increased number of estimation outliers with respect to speech. Bounding the minimum spectral weighting gain to -15 dB, which is common in noise reduction, the performance is virtually unaffected. Hence, a speech codebook size of $L = 5$ is used, which reduces the number of distance calculations to only $N_{\mathrm{D,GSC}} = 28.5$ on average. This includes the distance calculations necessary for the VAD. Moreover, the reduced number of distance calculations allows the real-time processing of the codebook driven noise reduction on current SBCs.

# Chapter 7

# Summary

Single-microphone and multi-microphone speech enhancement techniques for mobile communication are investigated. First, the acoustic front-end of the digital processing chain is addressed by a novel filter design and optimization concept of a near field beamformer. This pre-processing stage guarantees an improved SNR for subsequent single-channel speech enhancement. Simplified assumptions regarding the statistical characteristics of noise signals typically limit the performance of state-of-the-art single-microphone systems and implicate unpleasant artifacts in terms of *musical tones*. In this thesis new methods and strategies of *Information Combining* have been developed to tackle in particular the problem of noise estimation in case of non-stationarity. The proposed single-microphone speech enhancement algorithms clearly outperform conventional systems with respect to high noise attenuation and low speech distortion. At the same time, *musical tone* artifacts are almost eliminated by the significantly improved speech and noise estimation accuracy. This is confirmed by numerous benchmarks with objective instrumental measures as well as real-time experiments with demonstrators.

## Near Field Beamforming

A novel concept for the filter design of a filter-and-sum beamformer based on numerical near field optimization is presented. The beamformer consists of a non uniform filterbank with FIR filters in the sub-bands The proposed design strategy combines the advantages of decoupled sub-band filters with a frequency resolution according to the human auditory system. The optimization scheme allows to closely approximate a predefined reception characteristic which can be freely chosen according to the application. The novel system provides a distinct spatial selectivity independently of the frequency. Hence, the beamformer achieves a substantial SNR already at the acoustic front-end. Switching between different reception characteristics, e. g., for speaker selection in a video conference scenario, can be easily achieved using several pre-computed filter sets.

## Single-Microphone Based Noise Suppression

### *Baseline Tracing*

A novel short-term noise *power spectral density* (PSD) estimator *Baseline Tracing* is presented. The basic idea consists of a constrained logarithmic magnitude tracing of

the noisy observation separately for each frequency bin. This constraint magnitude change causes a certain inertia of the noise estimate over time which corresponds to the temporal statistics of noise. The estimator can be explained in terms of delta modulation with an adaptive step size, operating in the slope overload mode. In the linear amplitude domain, the short-term noise PSD of the current frame is calculated by a simple scaling of the last noise estimate with a frequency and time dependent tracing factor. Stretching or compressing is decided according to the sign of the difference between the last short-term noise PSD estimate and the current noisy frame. Doing so, the estimator traces the noisy observation. Since speech onsets are assumed as sudden rises in the noisy observation, the tracing factor has to be selected to only follow the slow variations of the noise. A fixed as well as an adaptive tracing factor are introduced which take into account the long-term speech spectrum average and the frame SNR. Compared to state-of-the-art systems, the new *Baseline Tracing* algorithm with the fixed tracing factor performs similar with respect to the noise PSD error measure while performing superior utilizing the adaptive tracing factor. The noise reduction performance is characterized by a low speech distortion and simultaneously high noise attenuation. The proposed concept has extremely low computational complexity and memory footprint. With these characteristics it is especially well suited for applications where processing power and memory is limited.

### Exploiting Spectral Dependencies

An approach to wideband speech enhancement is proposed that exploits spectral dependencies between the low band (50 Hz – 4 kHz) and the high band (4 kHz – 7 kHz) of speech signals for improved noise reduction in the high band. While a conventional noise suppression takes place in the low band, a joint noise suppression approach is applied in the high band. Features from the enhanced low band signal are extracted and used to estimate sub-band energies of the high band using techniques known from artificial bandwidth extension. Compared to MFCC features, the utilized RASTA-PLP features are more robust against short-term noise variations and include furthermore a speaker normalization. The weighting gains determined from these energy estimates are adaptively combined with conventional gains, obtained in addition for the high band. This *combining* in the high band is possible employing a pre-trained look-up table which depends on the average low band SNR and the respective high band SNR. In order to increase the perceived speech quality if only a noisy low band signal has been received, a slightly modified version of the system can additionally be used to perform a joint noise reduction and artificial bandwidth extension.

### Codebook Based Speech and Noise Estimation

*A priori* knowledge about speech and noise allows to model and to cope with highly non-stationary noise environments. Starting point is a brute force codebook matching approach, which provides the upper performance bound and serves as reference codebook processing scheme. The basic concept is based on a superposition

of scaled speech and noise codebook entries. The speech codebook is pre-trained once in advance, while the noise codebook is additionally adapted to new noise types online. Training vectors for online noise codebook updates are identified using a novel *voice activity detector* (VAD) and a codebook mismatch measure.

The VAD is realized as part of the codebook matching, but utilizes only *a priori* knowledge of speech. A speech gain is provided in each frame which is a reliable speech indicator and may contain a noise floor, especially at low input SNR. By means of a baseline tracing algorithm, similar to noise reduction, the noise floor is removed and subsequently the gain is mapped to soft VAD values between zero and one. Instrumental measurements confirmed a consistent improvement compared to state-of-the-art systems, resulting in higher detection rates at significant lower false alarm rates, even for low input SNR and highly non-stationary noise.

### Information Combining

A generic solution is formulated for the joint speech and noise estimation problem given the noisy observation. The solution considers several speech and noise estimates and provides optimal mixing coefficients with respect to minimized estimation error power regarding the noisy observation. At first, various estimates of the noisy observation are computed by permuting all different speech and noise estimates. Based on a distance measure between these estimates and the noisy observation optimal mixing coefficients for each frequency bin of the individual speech and noise estimates are determined. Applying the optimal mixing coefficients to the individual speech and noise estimates yields the final refined speech and noise estimate. This procedure is called *Information Combining*. The estimation error power after *Information Combining* is less than the minimum of the error power of the individual estimates. Utilizing *Information Combining* two main restrictions of codebook based speech and noise estimation are tackled:

Although the noise codebook is updated online, it is not guaranteed that an appropriate codebook entry is available for each noisy observation. A noise codebook update is prevented, for example, if the ambient noise changes while speech is still present. In such cases, the noise estimation is restricted, but this impact is compensated utilizing the proposed *Information Combining*. The necessary second noise estimate is provided by a statistical noise estimator, e. g., the new proposed *Baseline Tracing*. Given a second speech estimate, e. g., from the last enhanced frame of the speech enhancement system, the *Information Combining* is capable to provide also a refined speech estimate. The evaluation verified a tremendous improvement of noise attenuation, while the speech distortion is reduced simultaneously. Hence, the proposed *Information Combining* is used to overcome missing *a priori* codebook knowledge.

Facing practical application scenarios the brute force codebook matching is too expensive and a substantial complexity reduction is necessary. With respect to the *Information Combining* procedure, the brute force codebook driven speech and noise estimates can be replaced by two cascades of gain shape *vector quantizer* (VQ) estimates. While the first cascade provides a reliable noise estimate for low

SNR, the second cascade provides a robust speech estimate for high SNR. Utilizing the *Information Combining* procedure again, the best of all estimates is combined yielding the final estimates of speech and noise. Hence, the brute force search is replaced by four gain shape VQs and subsequent *Information Combining*, reducing the number of distance calculations in each frame by a factor of 60.

The simulations confirm that, the proposed codebook based noise reduction system clearly outperforms conventional state-of-the art noise reduction systems. A tremendous performance gain is achieved, especially for transient and fast types of noise with up to 12 dB improved noise attenuation. Although the proposed system achieves best scores regarding noise attenuation and speech distortion, the variance of the spectral weighting gains is decreased compared to conventional systems. This is a strong indicator for significantly reduced *musical tone* artifacts. Informal listening tests confirmed that *musical tones* are almost removed by this technique.

## Real-Time Implementation

The practical capability of the proposed algorithms is demonstrated by two applications. The novel near field beamformer is embedded in a high quality video conferencing client. The identification and separation of the most active talkers is the target of the proposed multimodal signal analysis. Exploiting information from video and audio analysis, the most active speakers are determined as a function of time *and* space. On this basis, the most active participants are artificially placed side by side in a conference at the receiver. Due to the novel near field beamformer actual no artifacts in the artificial scene composition of the demonstrator occur.

With respect to single-microphone speech enhancement, the codebook driven enhancement system has been further investigated. Considering the computational capabilities of current *single-board computer*s (SBCs), a complexity reduction is carried out for both, the codebook matching as well as the VAD. It turns out, that already $L = 5$ carefully selected speech codebook entries are sufficient without affecting the overall performance. Utilizing a software based *private branch exchange* (PBX) a prove of concept is implemented on a lightweight embedded computing platform.

## Conclusion

The proposed *Information Combining* is a powerful method to merge the best of several speech and noise estimates. It is of special interest, that the estimation error power after *Information Combining* is less than the minimum of the error power of the individual estimates. In the context of codebook driven noise suppression, the proposed method is so efficient that the brute force search can be replaced by several gain shape VQs estimates without loosing notable performance. Moreover, missing *a priori* codebook knowledge is compensated incorporating a statistical fallback noise estimator. Hence, *Information Combining* can be used for both improving the estimation quality and reducing the complexity. The resulting estimation quality of speech and noise is such accurate, that the occurrence of undesired *musical tones* is almost avoided – a decisive step towards artifact-free speech enhancement.

# Appendix A

# Optimized filter coefficients

## A.1 Free Field

### A.1.1 Sub-band beamformer

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.55096 | 0.95313 | 0.97401 | 0.98146 | 0.99065 | 0.98817 | 0.98622 | 0.98715 |
| 2 | 0.82418 | 0.77801 | −0.22916 | −0.27949 | 0.91870 | 0.56402 | 0.91225 | 0.98034 |
| 3 | 0.89846 | −0.90478 | −0.96104 | −0.97088 | −0.97977 | −0.97411 | −0.93783 | 0.95025 |
| 4 | 0.71215 | −0.94958 | −0.97249 | −0.97588 | −0.98681 | −0.98492 | −0.97337 | 0.94182 |
| 5 | 0.04905 | −0.95201 | −0.96443 | −0.95723 | −0.98661 | −0.98280 | −0.96814 | 0.67551 |
| 6 | −0.86549 | −0.91395 | −0.30493 | 0.96073 | −0.97812 | −0.97785 | −0.96386 | 0.89431 |
| 7 | −0.79211 | 0.91849 | 0.97995 | 0.99018 | 0.78722 | −0.71738 | −0.72880 | −0.74877 |
| 8 | −0.18672 | 0.98019 | 0.99185 | 0.99526 | 0.98659 | 0.97245 | 0.91072 | −0.67535 |

**Table A.1:** Optimized filter coefficients of sub-band 1 (1 Hz – 268 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.41510 | −0.64213 | 0.70482 | −0.85706 | 0.97949 | 0.97739 | 0.97799 | 0.98783 |
| 2 | 0.04329 | −0.06134 | 0.91892 | −0.72908 | 0.33629 | −0.76612 | 0.46939 | 0.97850 |
| 3 | 0.14936 | 0.74476 | 0.88976 | 0.07453 | −0.96120 | −0.96928 | −0.95924 | 0.95238 |
| 4 | 0.06091 | −0.17004 | 0.90249 | −0.71018 | −0.96844 | −0.97641 | −0.97315 | 0.84845 |
| 5 | −0.06483 | −0.27719 | 0.59026 | −0.58652 | −0.94317 | −0.97184 | −0.97118 | −0.63597 |
| 6 | −0.03628 | −0.84770 | 0.57097 | −0.71253 | 0.93546 | −0.92366 | −0.95713 | −0.74877 |
| 7 | 0.28061 | −0.84274 | −0.79429 | 0.79044 | 0.98456 | 0.95764 | −0.10158 | −0.70714 |
| 8 | 0.66698 | −0.89254 | 0.32669 | 0.92647 | 0.99223 | 0.98590 | 0.96347 | 0.11570 |

**Table A.2:** Optimized filter coefficients of sub-band 2 (268 Hz – 839 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.16934 | 0.95933 | −0.24197 | −0.97999 | 0.17398 | 0.97410 | 0.85469 | 0.19047 |
| 2 | 0.13251 | −0.94855 | 0.88287 | 0.33649 | −0.93793 | 0.69086 | −0.74307 | 0.02136 |
| 3 | 0.14344 | −0.96963 | 0.43211 | 0.95602 | −0.85648 | −0.64071 | −0.30945 | 0.46151 |
| 4 | 0.36741 | −0.96720 | 0.90696 | −0.55528 | 0.37535 | −0.89908 | 0.34223 | 0.75848 |
| 5 | 0.19800 | −0.92705 | 0.74132 | −0.92679 | 0.86626 | −0.86340 | 0.00064 | 0.34685 |
| 6 | 0.12718 | 0.32359 | 0.90968 | −0.96756 | 0.32632 | −0.92725 | −0.79812 | −0.28475 |
| 7 | −0.22162 | 0.90298 | −0.48449 | 0.14034 | 0.00994 | −0.66801 | −0.71688 | −0.74175 |
| 8 | 0.03805 | −0.49440 | −0.61180 | 0.98136 | 0.90343 | 0.94759 | 0.91011 | 0.09941 |

**Table A.3:** Optimized filter coefficients of sub-band 3 (839 Hz – 1549 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.22750 | −0.18613 | 0.15414 | −0.38216 | −0.17940 | 0.52722 | 0.99139 | −0.17122 |
| 2 | −0.60693 | 0.30275 | −0.04624 | 0.82993 | 0.33335 | −0.96943 | 0.31047 | −0.38974 |
| 3 | 0.23268 | 0.05010 | −0.24007 | −0.30234 | −0.18941 | −0.94356 | −0.94837 | 0.83630 |
| 4 | 0.71505 | −0.21618 | −0.06626 | −0.86615 | −0.43333 | 0.19182 | 0.09029 | 0.87494 |
| 5 | −0.59813 | −0.45857 | 0.85817 | 0.68876 | 0.90434 | 0.65783 | 0.77719 | −0.75468 |
| 6 | −0.23858 | 0.30301 | −0.17516 | 0.27536 | 0.45845 | −0.93277 | −0.89761 | −0.50089 |
| 7 | 0.37863 | 0.50336 | −0.86079 | −0.46740 | −0.87402 | −0.45443 | −0.96228 | 0.78333 |
| 8 | −0.04781 | −0.53750 | 0.70673 | 0.06062 | 0.57535 | 0.99006 | 0.52855 | −0.30255 |

**Table A.4:** Optimized filter coefficients of sub-band 4 (1549 Hz – 2614 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.02943 | −0.08057 | 0.12956 | −0.11367 | 0.17923 | −0.27472 | 0.43424 | 0.73270 |
| 2 | −0.01953 | 0.06891 | −0.12670 | 0.15798 | −0.49291 | 0.10985 | −0.86247 | −0.43910 |
| 3 | −0.03017 | 0.03451 | −0.07490 | −0.00192 | 0.45217 | −0.03731 | −0.58577 | −0.37712 |
| 4 | 0.06483 | −0.12040 | 0.25986 | −0.22230 | 0.54781 | −0.22690 | 0.35583 | 0.56455 |
| 5 | 0.00290 | −0.01343 | −0.02651 | 0.13035 | −0.87764 | 0.64568 | −0.12909 | −0.12932 |
| 6 | −0.02983 | 0.03832 | −0.14253 | 0.03680 | 0.10643 | −0.03254 | −0.33007 | −0.08442 |
| 7 | 0.01079 | 0.01040 | 0.10111 | −0.06978 | 0.88298 | 0.18600 | −0.16915 | −0.08760 |
| 8 | 0.01489 | −0.05148 | 0.01889 | 0.00158 | −0.68755 | 0.30672 | 0.23880 | 0.04269 |

**Table A.5:** Optimized filter coefficients of sub-band 5 (2614 Hz – 4731 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.00574 | 0.00020 | 0.00307 | −0.01478 | −0.01924 | −0.05988 | −0.01492 | 0.39988 |
| 2 | −0.01152 | −0.01407 | 0.00497 | 0.02998 | 0.00708 | −0.03735 | −0.06875 | −0.48879 |
| 3 | 0.02491 | 0.02607 | −0.00599 | −0.05903 | 0.07766 | 0.01011 | −0.52104 | 0.33019 |
| 4 | −0.02408 | −0.05223 | 0.02301 | 0.06007 | −0.20990 | −0.22424 | −0.04987 | −0.26629 |
| 5 | 0.02490 | 0.05199 | −0.01970 | −0.05770 | 0.45996 | 0.42535 | −0.13117 | −0.00202 |
| 6 | −0.01140 | −0.05120 | 0.02401 | 0.03041 | −0.47523 | 0.02861 | 0.04660 | 0.04486 |
| 7 | 0.00583 | 0.02600 | −0.01203 | −0.01244 | 0.48048 | 0.31310 | 0.05593 | −0.05442 |
| 8 | 0.00035 | −0.01134 | 0.00570 | −0.00044 | −0.32945 | −0.08982 | −0.02700 | 0.00278 |

**Table A.6:** Optimized filter coefficients of sub-band 6 (4731 Hz – 12049 Hz)

### A.1.2 Full-band beamformer

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.07114 | 0.08709 | 0.09633 | 0.00189 | 0.02439 | 0.02870 | −0.02841 | 0.02239 |
| 2 | −0.14259 | −0.13731 | −0.31511 | 0.00477 | −0.13586 | −0.01264 | 0.04188 | −0.09627 |
| 3 | 0.10294 | −0.02893 | 0.35399 | −0.08881 | 0.26313 | 0.01171 | −0.11276 | 0.23645 |
| 4 | −0.04071 | 0.18617 | −0.26620 | 0.18509 | −0.26647 | 0.01308 | 0.14426 | −0.29070 |
| 5 | −0.15850 | −0.13198 | 0.11949 | −0.06987 | 0.03489 | 0.10417 | −0.21182 | 0.16350 |
| 6 | 0.22776 | 0.10295 | −0.03086 | −0.14443 | 0.19931 | −0.06697 | 0.08140 | 0.10562 |
| 7 | −0.22693 | −0.01725 | −0.14352 | 0.22281 | −0.33637 | −0.01542 | −0.05573 | −0.15016 |
| 8 | −0.07602 | 0.13150 | 0.11367 | 0.04684 | 0.40464 | 0.20658 | 0.02553 | −0.01017 |
| 9 | 0.27695 | 0.19307 | −0.02782 | −0.24612 | −0.50664 | 0.00226 | −0.21517 | 0.15765 |
| 10 | −0.17478 | −0.15521 | −0.23465 | 0.24704 | 0.25736 | −0.14822 | 0.05646 | 0.05032 |
| 11 | −0.10660 | −0.03741 | −0.19088 | −0.09286 | 0.34148 | 0.14630 | 0.03514 | −0.20626 |
| 12 | 0.04732 | 0.26890 | 0.22693 | 0.15783 | −0.70559 | 0.28692 | −0.11297 | 0.13077 |
| 13 | 0.15779 | −0.17038 | −0.03417 | 0.05641 | 0.23886 | −0.33055 | −0.22697 | 0.12992 |
| 14 | −0.26853 | 0.09891 | −0.36892 | −0.24271 | 0.36039 | 0.11821 | 0.13105 | −0.03884 |
| 15 | −0.11058 | −0.33622 | 0.03956 | 0.27293 | −0.13001 | 0.20169 | 0.11744 | −0.11779 |
| 16 | 0.15380 | 0.47596 | 0.27903 | 0.29828 | −0.61597 | −0.01738 | −0.41010 | 0.11662 |
| 17 | 0.03261 | 0.12745 | 0.28921 | −0.47747 | 0.62070 | −0.18012 | 0.08566 | 0.13936 |
| 18 | −0.33998 | −0.45806 | −0.96085 | 0.10487 | 0.22450 | 0.17929 | 0.05957 | −0.12795 |
| 19 | 0.12421 | 0.34335 | 0.53825 | 0.19959 | −0.61342 | 0.14576 | 0.01997 | 0.02016 |
| 20 | 0.05259 | 0.16416 | 0.32805 | 0.25343 | 0.04403 | −0.34609 | −0.27675 | 0.02015 |
| 21 | −0.14684 | −0.15170 | −0.43962 | −0.74714 | 0.51735 | 0.31674 | 0.07206 | 0.07732 |
| 22 | −0.17884 | −0.05999 | −0.10195 | 0.46129 | −0.21492 | −0.05072 | 0.06525 | 0.08307 |
| 23 | 0.27463 | 0.36049 | 0.23246 | 0.30053 | −0.25435 | −0.17904 | 0.00963 | −0.21249 |
| 24 | −0.17838 | −0.35165 | 0.06593 | −0.52715 | 0.16324 | 0.15735 | −0.29198 | 0.03816 |
| 25 | −0.30466 | 0.19168 | 0.19709 | 0.17942 | 0.20223 | 0.10700 | 0.26580 | 0.36063 |
| 26 | 0.51328 | 0.30026 | −0.46385 | 0.02337 | −0.14990 | −0.21520 | −0.20237 | −0.32870 |
| 27 | −0.51609 | −0.55543 | 0.28848 | 0.00660 | −0.08665 | −0.02601 | 0.08246 | −0.04273 |
| 28 | 0.07060 | 0.31306 | 0.19634 | −0.13852 | 0.00847 | 0.39711 | 0.00331 | 0.28108 |
| 29 | 0.05222 | 0.18153 | 0.10638 | 0.12000 | 0.21565 | −0.38240 | −0.22298 | −0.01403 |
| 30 | −0.07365 | 0.12908 | −0.23198 | −0.17577 | −0.11658 | 0.12652 | 0.12301 | −0.35409 |
| 31 | −0.02498 | −0.22959 | 0.06270 | −0.08679 | −0.28814 | 0.16540 | −0.00623 | 0.58806 |
| 32 | −0.03200 | 0.11154 | 0.09239 | 0.35299 | 0.39531 | −0.07707 | −0.00347 | −0.52084 |
| 33 | −0.09729 | 0.39019 | 0.13024 | −0.51276 | −0.16029 | 0.02570 | −0.36402 | 0.29806 |

| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 34 | 0.07975 | −0.11300 | −0.29604 | 0.14854 | −0.17850 | 0.08740 | 0.29717 | 0.09506 |
| 35 | 0.16184 | −0.03297 | −0.12553 | 0.09833 | 0.11316 | 0.15085 | −0.00296 | −0.26798 |
| 36 | −0.36035 | 0.10496 | 0.09135 | 0.08446 | 0.08578 | −0.19027 | −0.43140 | 0.19261 |
| 37 | 0.18286 | 0.13877 | 0.24811 | −0.39375 | −0.27993 | 0.27624 | 0.28181 | 0.05073 |
| 38 | 0.14298 | 0.05902 | −0.62067 | 0.31191 | 0.10679 | 0.04091 | −0.15775 | −0.05187 |
| 39 | −0.14837 | −0.24901 | 0.06011 | 0.14549 | 0.06177 | −0.05738 | −0.05364 | −0.00948 |
| 40 | −0.12265 | 0.11767 | 0.13854 | −0.29634 | −0.16578 | 0.04207 | −0.03765 | 0.07745 |
| 41 | 0.27018 | 0.17324 | 0.06151 | 0.12862 | −0.01377 | 0.27238 | −0.03023 | −0.06111 |
| 42 | −0.15926 | −0.14847 | −0.32444 | 0.25567 | 0.10368 | −0.20113 | −0.08089 | 0.17735 |
| 43 | −0.02650 | −0.00145 | −0.05649 | −0.26119 | −0.09827 | 0.07262 | 0.02996 | −0.23661 |
| 44 | 0.09616 | −0.11129 | 0.31829 | 0.07862 | −0.06023 | 0.17022 | 0.00084 | 0.15807 |
| 45 | −0.13720 | 0.36793 | −0.16663 | 0.16253 | 0.11725 | −0.13079 | −0.17175 | 0.05730 |
| 46 | 0.15632 | −0.34275 | −0.14746 | −0.11990 | −0.10969 | 0.08482 | 0.17584 | −0.15601 |
| 47 | −0.13379 | 0.12700 | 0.15515 | 0.03144 | 0.04388 | −0.00712 | −0.11806 | 0.11733 |
| 48 | 0.04912 | 0.01166 | −0.04966 | 0.03649 | −0.01433 | 0.01163 | 0.02065 | −0.03265 |

**Table A.7:** Optimized full-band filter coefficients

## A.2 Reverberant Room

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.99948 | −0.99447 | 0.99846 | 0.99955 | 0.99282 | 0.37780 | 0.30044 | −0.99972 |
| 2 | −0.99917 | −0.97780 | 0.99788 | 0.99946 | −0.02509 | −0.99309 | 0.98454 | −0.99968 |
| 3 | −0.99825 | 0.96024 | 0.99604 | 0.99929 | −0.98965 | −0.99538 | 0.99549 | −0.99961 |
| 4 | −0.50439 | 0.97300 | 0.83226 | 0.99896 | −0.99320 | −0.99534 | 0.99788 | −0.99946 |
| 5 | 0.99804 | 0.01759 | −0.99617 | 0.99794 | −0.99330 | −0.99272 | 0.99877 | −0.99900 |
| 6 | 0.99897 | −0.98678 | −0.99814 | −0.15093 | −0.99011 | 0.94023 | 0.99920 | −0.81912 |
| 7 | 0.99928 | −0.99446 | −0.99877 | −0.99794 | 0.08199 | 0.99595 | 0.99943 | 0.99911 |
| 8 | 0.99944 | −0.99673 | −0.99907 | −0.99896 | 0.99359 | 0.99819 | 0.99958 | 0.99958 |

**Table A.8:** Optimized filter coefficients of sub-band 1 (1 Hz – 268 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.71926 | −0.48637 | −0.99670 | 0.98006 | 0.99505 | −0.99653 | −0.99882 | 0.66781 |
| 2 | 0.96959 | 0.80318 | −0.99555 | −0.06960 | 0.99423 | −0.99346 | −0.99851 | 0.83343 |
| 3 | 0.96745 | 0.74212 | −0.99179 | −0.94972 | 0.99349 | −0.93087 | −0.99789 | 0.96119 |
| 4 | 0.70022 | 0.06959 | 0.02213 | −0.94596 | 0.99205 | 0.99040 | −0.99641 | 0.96985 |
| 5 | −0.97312 | −0.09498 | 0.99256 | −0.87271 | 0.98635 | 0.99444 | −0.99129 | 0.60781 |
| 6 | −0.98668 | 0.86139 | 0.99634 | 0.15725 | −0.94581 | 0.99489 | −0.52403 | −0.98877 |
| 7 | −0.98976 | 0.95937 | 0.99756 | −0.86858 | −0.99366 | 0.99276 | 0.97721 | −0.99609 |
| 8 | −0.98912 | 0.98193 | 0.99813 | −0.97641 | −0.99716 | −0.46079 | 0.88619 | −0.99804 |

**Table A.9:** Optimized filter coefficients of sub-band 2 (268 Hz – 839 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.96654 | 0.99688 | 0.99690 | −0.06746 | 0.99461 | −0.99274 | −0.99772 | 0.97905 |
| 2 | 0.26084 | 0.00165 | 0.94118 | −0.99321 | 0.99054 | −0.83507 | −0.99424 | 0.65517 |
| 3 | 0.87508 | −0.99270 | −0.98928 | −0.99146 | 0.95105 | 0.96827 | 0.05269 | −0.89135 |
| 4 | 0.79587 | −0.99180 | −0.98641 | 0.59280 | −0.97846 | 0.72550 | 0.98716 | −0.80780 |
| 5 | −0.07375 | −0.63699 | −0.36849 | 0.99134 | −0.98443 | −0.97900 | 0.98561 | 0.54200 |
| 6 | −0.60351 | 0.99387 | 0.97906 | 0.99264 | −0.33187 | −0.98534 | −0.01608 | 0.75816 |
| 7 | −0.50297 | 0.99678 | 0.74940 | −0.87677 | 0.99441 | 0.07934 | −0.98924 | 0.26644 |
| 8 | −0.61543 | 0.99738 | −0.99457 | −0.99721 | 0.99805 | 0.99543 | −0.99381 | −0.72760 |

**Table A.10:** Optimized filter coefficients of sub-band 3 (839 Hz – 1549 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.99888 | 0.99566 | 0.12940 | 0.55757 | 0.98984 | −0.99840 | 0.27608 | −0.05137 |
| 2 | −0.17041 | 0.04779 | −0.98830 | 0.65512 | 0.60206 | −0.99729 | 0.99201 | −0.80324 |
| 3 | 0.99725 | −0.99007 | −0.78251 | −0.69055 | −0.71427 | −0.99260 | 0.99066 | −0.59038 |
| 4 | 0.99665 | −0.98827 | 0.99390 | −0.82807 | 0.91535 | 0.88406 | 0.23134 | 0.46190 |
| 5 | −0.01613 | −0.16682 | 0.99669 | −0.75661 | 0.83123 | 0.97628 | −0.99100 | 0.66678 |
| 6 | −0.99637 | 0.98831 | 0.99647 | 0.43024 | −0.95707 | −0.11961 | −0.99271 | 0.53379 |
| 7 | −0.99609 | 0.99013 | −0.53478 | 0.78183 | −0.97554 | −0.97768 | −0.25621 | −0.25858 |
| 8 | 0.99662 | −0.88154 | −0.99838 | −0.45846 | 0.84962 | 0.67823 | 0.99708 | −0.46794 |

**Table A.11:** Optimized filter coefficients of sub-band 4 (1549 Hz – 2614 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.06206 | −0.90128 | −0.92650 | −0.68340 | −0.12689 | 0.27685 | −0.99737 | 0.44595 |
| 2 | 0.94055 | 0.96540 | 0.89525 | 0.99654 | 0.66967 | 0.53433 | −0.83029 | 0.97797 |
| 3 | −0.96877 | 0.92654 | 0.55340 | 0.27228 | −0.95850 | 0.08120 | 0.98684 | −0.96775 |
| 4 | −0.72231 | −0.72542 | −0.93887 | −0.99692 | −0.95720 | 0.57541 | 0.25564 | −0.97797 |
| 5 | 0.97648 | −0.80982 | −0.83429 | −0.99687 | 0.40895 | 0.74836 | −0.98162 | 0.85678 |
| 6 | 0.82205 | 0.39063 | 0.96683 | 0.80609 | 0.89284 | −0.51218 | −0.59309 | 0.97257 |
| 7 | −0.98853 | 0.34685 | 0.81806 | 0.99638 | −0.93634 | −0.34762 | 0.98807 | −0.42139 |
| 8 | 0.05032 | 0.13273 | −0.99559 | −0.89950 | −0.44582 | 0.98312 | −0.80257 | −0.30492 |

**Table A.12:** Optimized filter coefficients of sub-band 5 (2614 Hz – 4731 Hz)

| | Microphone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | −0.24818 | −0.15122 | 0.32139 | −0.07186 | −0.39530 | 0.55145 | −0.99959 | 0.19819 |
| 2 | 0.44246 | 0.42300 | −0.95083 | −0.10051 | 0.97126 | −0.04898 | 0.75347 | 0.44091 |
| 3 | −0.67355 | −0.96852 | 0.98746 | 0.73087 | −0.99801 | −0.25283 | −0.36988 | −0.98893 |
| 4 | 0.74090 | 0.99633 | −0.22434 | −0.99719 | 0.17210 | 0.99650 | −0.84638 | 0.99282 |
| 5 | −0.88435 | −0.33316 | −0.99626 | 0.28482 | 0.60361 | 0.71572 | 0.99747 | −0.56924 |
| 6 | 0.83249 | −0.76373 | 0.79479 | 0.75302 | −0.99726 | −0.78153 | −0.63883 | 0.11439 |
| 7 | −0.61275 | 0.99917 | −0.10899 | −0.99891 | 0.43522 | 0.99843 | 0.30493 | −0.03826 |
| 8 | 0.22406 | −0.44149 | −0.36195 | 0.41225 | −0.32549 | −0.08265 | 0.03311 | 0.05252 |

**Table A.13:** Optimized filter coefficients of sub-band 6 (4731 Hz – 12049 Hz)

## Appendix B

# Equivalent Variance of Recursive and Mean Average Smoothing

By assuming an uncorrelated signal $x(k)$ and equating the variance of the mean short-term expectation $\overline{\mathbb{E}}_K\{\cdot\}$ and the recursive short-term expectation $\tilde{\mathbb{E}}_\alpha\{\cdot\}$ estimators, the equivalent rectangular window length of the recursive short-term expectation estimator can be calculated. The corresponding block diagram is depicted in Fig. B.1, where $x(k)$ is the input signal and $h_K(k), h_\alpha(k)$ are the impulse responses of the short-term expectation operators. The averaged output signal $y(k)$ is given for the mean short-term expectation operator by $y_K(k)$ and for the recursive short-term expectation operator by $y_\alpha(k)$, respectively.

The impulse response of the mean short-term expectation $\overline{\mathbb{E}}_K\{\cdot\}$ operator is given by

$$h_K(k) = \begin{cases} \frac{1}{K} & \text{for } 0 \leq k < K \\ 0 & \text{else ,} \end{cases} \tag{B.1}$$

and the impulse response of the recursive short-term expectation operator $\tilde{\mathbb{E}}_\alpha\{\cdot\}$ is defined as

$$h_\alpha(k) = \begin{cases} (1-\alpha)\cdot\alpha^k & \text{for } k \geq 0,\ 0 < \alpha < 1 \\ 0 & \text{else .} \end{cases} \tag{B.2}$$

The parameters $K$ and $\alpha$ control the smoothing properties of the respective short-term expectation estimator. A relation between $K$ and $\alpha$ is derived in the following by equating the variance of both short-term expectation estimators.

In general, the variance of the output signal $y(k)$ is given by

$$\sigma_y^2 = \mathbb{E}\left\{y^2(k)\right\} - \left(\mathbb{E}\left\{y(k)\right\}\right)^2 , \tag{B.3}$$

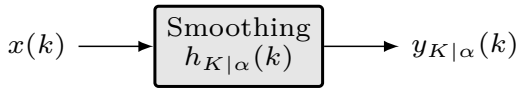



**Figure B.1:** Block diagram of smoothing operation

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator. Decomposing the output signal $y(k)$ into an alternating component $\tilde{y}(k)$ and a mean $\mathbb{E}\{y(k)\}$ yields,

$$y(k) = \tilde{y}(k) + \mathbb{E}\{y(k)\}, \tag{B.4}$$

and the expression $\mathbb{E}\left\{y^2(k)\right\}$ from Eq. (B.3) is now formulated in terms of Eq. (B.4) by

$$\begin{aligned}\mathbb{E}\left\{y^2(k)\right\} &= \mathbb{E}\left\{(\tilde{y}(k) + \mathbb{E}\{y(k)\})^2\right\} &\text{(B.5)}\\ &= \mathbb{E}\left\{\tilde{y}^2(k)\right\} + 2 \cdot \mathbb{E}\{\tilde{y}(k)\}\,\mathbb{E}\{y(k)\} + (\mathbb{E}\{y(k)\})^2 &\text{(B.6)}\\ &= \mathbb{E}\left\{\tilde{y}^2(k)\right\} + (\mathbb{E}\{y(k)\})^2 \ . &\text{(B.7)}\end{aligned}$$

Moreover, utilizing Eq. (B.7) in Eq. (B.3), the variance of $y(k)$ is given by,

$$\sigma_y^2 = \mathbb{E}\left\{\tilde{y}^2(k)\right\} \ . \tag{B.8}$$

In the following derivation the filter impulse response $h(k)$ represents either the short-term mean expectation or the recursive short-term expectation operator. Using the Wiener-Lee relation, the auto-correlation function of the alternating component $\tilde{y}(k)$ is given by

$$\varphi_{\tilde{y}\tilde{y}}(i) = \varphi_{\tilde{x}\tilde{x}}(i) * \varphi_{hh}(i), \tag{B.9}$$

where $*$ denotes the linear discrete convolution operator, $\varphi_{\tilde{x}\tilde{x}}$ represents the auto-correlation function of the alternating component $\tilde{x}(k)$ of the input signal and $\varphi_{hh}$ is the auto-correlation function regarding the filter impulse response $h(k)$. Assuming $\tilde{x}(k)$ as zero mean and white yields

$$\begin{aligned}\varphi_{\tilde{y}\tilde{y}}(i) &= \varphi_{\tilde{x}\tilde{x}}(0) \cdot \delta(i) * \varphi_{hh}(i) \ , &\text{(B.10)}\\ \varphi_{\tilde{y}\tilde{y}}(i) &= \varphi_{\tilde{x}\tilde{x}}(0) \cdot \varphi_{hh}(i) \ . &\text{(B.11)}\end{aligned}$$

Using the relation,

$$\sigma_y^2 = \mathbb{E}\left\{\tilde{y}^2(k)\right\} = \varphi_{\tilde{y}\tilde{y}}(0), \tag{B.12}$$

the variance of $y(k)$ is given by,

$$\begin{aligned}\sigma_y^2 &= \varphi_{\tilde{x}\tilde{x}}(0) \cdot \varphi_{hh}(0) &\text{(B.13)}\\ &= \mathbb{E}\left\{\tilde{x}^2(k)\right\} \sum_{j=0}^{\infty} h^2(j) &\text{(B.14)}\\ &= \sigma_{\tilde{x}}^2 \sum_{j=0}^{\infty} h^2(j) \ . &\text{(B.15)}\end{aligned}$$

By equating the variance of the output signals $y_K(k)$ and $y_\alpha(k)$ of both short-term expectation operators,

$$\sigma_{y_K}^2 = \sigma_{y_\alpha}^2 \ , \tag{B.16}$$

a relation between $K$ and $\alpha$ can be found. Utilizing Eq. (B.15), the variance of the mean short-term expectation operator $\overline{\mathbb{E}}_K\{\cdot\}$ is given by,

$$\sigma_{y_K}^2 = \sigma_{\tilde{x}}^2 \sum_{j=0}^{\infty} h_K^2(j) = \sigma_{\tilde{x}}^2 \sum_{j=0}^{K-1} \frac{1}{K^2} = \sigma_{\tilde{x}}^2 \cdot \frac{1}{K} \tag{B.17}$$

whereas the variance of the recursive short-term expectation operator $\tilde{\mathbb{E}}_\alpha\{\cdot\}$ yields

$$\sigma_{y_\alpha}^2 = \sigma_{\tilde{x}}^2 \sum_{j=0}^{\infty} h_\alpha^2(j) = \sigma_{\tilde{x}}^2 \sum_{j=0}^{\infty} (1-\alpha)^2 \cdot \alpha^{2\cdot j} \tag{B.18}$$

$$= \sigma_{\tilde{x}}^2 (1-\alpha)^2 \cdot \frac{1}{1-\alpha^2} = \sigma_{\tilde{x}}^2 \frac{1-\alpha}{1+\alpha}\ , \tag{B.19}$$

with

$$\sum_{j=0}^{\infty} \alpha^{2\cdot j} = \frac{1}{1-\alpha^2}\ . \tag{B.20}$$

Finally, the smoothing parameter $\alpha$ of the recursive short-term expectation estimator is given by,

$$\alpha = \frac{K-1}{K+1}, \tag{B.21}$$

in terms of the equivalent rectangular window length $K$ in samples and vice versa,

$$K = \frac{1+\alpha}{1-\alpha}. \tag{B.22}$$

# Appendix C

# Evaluation System for Speech Enhancement

The evaluation of speech enhancement algorithms is a difficult task since the speech quality is perceived subjectively. The aim of the evaluation is to quantify the subjectively perceived speech quality. So far, the best way to evaluate speech enhancement is probably to conduct a listening test. However, such tests are very time consuming and costly as a large number of participants is required to get statistically significant results.

On the other hand, so called instrumental measures also allow to assess the speech quality. Each of the instrumental measures aim to predict different aspects of the subjectively perceived speech quality, e. g., in terms of speech distortion and noise attenuation. The interpretation of several instrumental measurements allows a ranking of the investigated speech enhancement algorithms.

In this thesis the evaluation of the speech enhancement algorithms is based on the evaluation framework and instrumental measures proposed in [Gustafsson et al. 1996; Quackenbush et al. 1988]. In the following, a brief overview of the evaluation framework as well as the instrumental measures is given.

The framework is illustrated in Fig. C.1. The spectral weighting gains $\mathcal{G}(\lambda, \mu)$

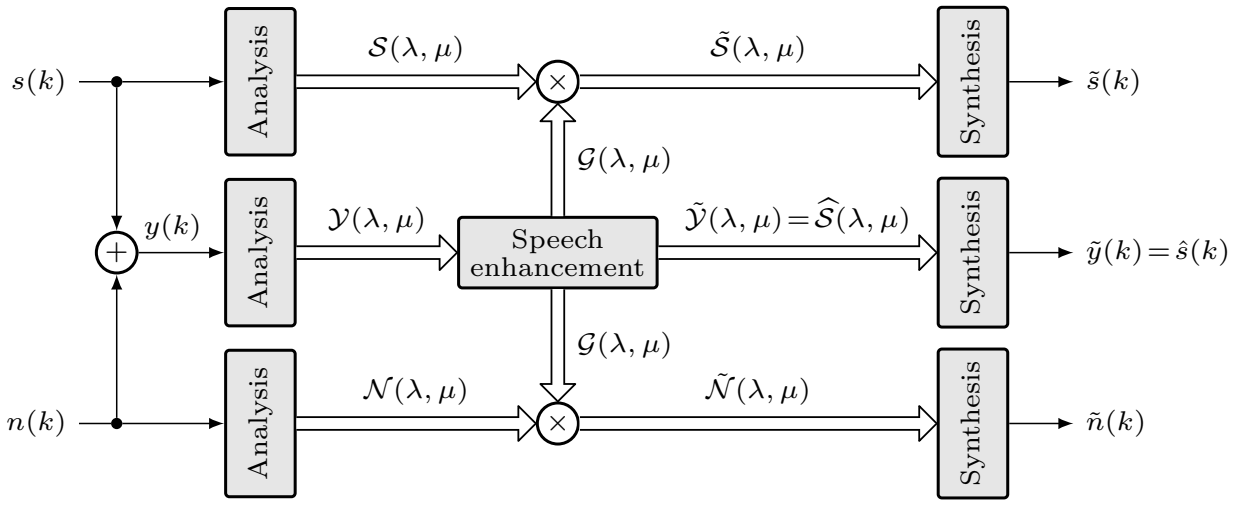



**Figure C.1:** Evaluation framework for speech enhancement

are determined in the *short-term Fourier domain* (STFD) based on the noisy input signal $y(k)$, which is the sum of the clean speech $s(k)$ and the noise signal $n(k)$. Besides, the noisy input signal $\mathcal{Y}(\lambda, \mu)$, the spectral weighting gain is applied to the clean speech component $\mathcal{S}(\lambda, \mu)$ as well as the noise component $\mathcal{N}(\lambda, \mu)$. Finally, the resulting filtered signals $\widehat{\mathcal{S}}(\lambda, \mu) = \tilde{\mathcal{Y}}(\lambda, \mu)$, $\tilde{\mathcal{S}}(\lambda, \mu)$, and $\tilde{\mathcal{N}}(\lambda, \mu)$ are transformed back into the time domain, where $\hat{s}(k)$ denotes the enhanced noisy output signal, $\tilde{s}(k)$ is the filtered speech component, and $\tilde{n}(k)$ is the filtered noise component, respectively. This allows to investigate the influence of the enhancement algorithms on the noisy input signal as well as on speech and noise separately.

# C.1 Input Signal-to-Noise Ratio

The noisy input signal $y(k)$ is generated from a clean speech signal $s(k)$ which is degraded by an additive noise component $n(k)$. In order to control the degradation the *signal-to-noise ratio* (SNR) of the input signal $y(k)$ can be adjusted.

For the adjustment of the SNR only signal samples with speech presence are considered. Note that the noise signal is assumed to be active all the time. The speech presence is determined by the objective measurement of the active speech level according to [ITU-T Recommendation P.56 1993]. Hence, the corresponding power of speech $P_s$ and noise $P_n$ are computed by

$$P_s = \frac{1}{\#\{\underline{M}_{\mathrm{S}}\}} \sum_{\kappa \in \underline{M}_{\mathrm{S}}} s(\kappa)^2 \,, \tag{C.1}$$

$$P_n = \frac{1}{\#\{\underline{M}_{\mathrm{S}}\}} \sum_{\kappa \in \underline{M}_{\mathrm{S}}} n(\kappa)^2 \,, \tag{C.2}$$

where $\underline{M}_{\mathrm{S}}$ is a vector which contains all signal samples with speech presence and $\#\{\underline{M}_{\mathrm{S}}\}$ is the number of elements of vector $\underline{M}_{\mathrm{S}}$. Given the desired SNR value $SNR_{dB}$ in dB, the scaling factor $a$ of the noise signal component is computed according to

$$a = \sqrt{\frac{P_s}{P_n \cdot 10^{SNR_{dB}/10}}} \tag{C.3}$$

and the noisy signal yields

$$y(k) = s(k) + a \cdot n(k) \ . \tag{C.4}$$

# C.2 Instrumental Measures for Speech Enhancement

## C.2.1 Segmental Speech and Noise Attenuation

The *segmental speech attenuation* (SegSA) and *segmental noise attenuation* (SegNA) are defined as the segmented power ratios between the original speech and noise

signals and their filtered versions, respectively. The measures SegSA and SegNA are given in dB and defined by

$$\mathrm{SegSA/dB} = \frac{1}{\#\{\underline{M}_\mathrm{S}\}} \sum_{\lambda \in \underline{M}_\mathrm{S}} \left( 10 \cdot \log_{10} \left( \frac{\sum_{\kappa=0}^{L_F-1} s(\kappa + \lambda \cdot L_F)^2}{\sum_{\kappa=0}^{L_F-1} \tilde{s}(\kappa + \lambda \cdot L_F)^2} \right) \right), \quad \text{(C.5)}$$

$$\mathrm{SegNA/dB} = \frac{1}{\#\{\underline{M}_\mathrm{N}\}} \sum_{\lambda \in \underline{M}_\mathrm{N}} \left( 10 \cdot \log_{10} \left( \frac{\sum_{\kappa=0}^{L_F-1} n(\kappa + \lambda \cdot L_F)^2}{\sum_{\kappa=0}^{L_F-1} \tilde{n}(\kappa + \lambda \cdot L_F)^2} \right) \right), \quad \text{(C.6)}$$

where $\underline{M}_\mathrm{S}$ denotes all frames with speech presence and $\underline{M}_\mathrm{N}$ is the set of frames to be evaluated in total. $\#\{\underline{M}_\mathrm{S}\}$ and $\#\{\underline{M}_\mathrm{N}\}$ denote the number of frames in each set $\underline{M}_\mathrm{S}$ and $\underline{M}_\mathrm{N}$, respectively. The frame size is represented by $L_F$.

Although the SegSA is not directly related to the manner of speech distortion, the difference between SegNA and SegSA indicates the effective noise reduction. For values greater than 0 dB the application of noise reduction appears reasonable.

### C.2.2 Segmental Speech Signal-to-Noise Ratio

The *segmental speech SNR* (SegSpSNR) is defined as the geometric mean of the SNR of short signal segments, where the difference between the original speech signal $s(k)$ and its filtered version $\tilde{s}(k)$ is considered as noise. The SegSpSNR is also given in dB and defined as

$$\mathrm{SegSpSNR}(\lambda) = 10 \cdot \log_{10} \left( \frac{\sum_{\kappa=0}^{L_F-1} s(\kappa + \lambda \cdot L_A)^2}{\sum_{\kappa=0}^{L_F-1} \left( s(\kappa + \lambda \cdot L_A) - \tilde{s}(\kappa + \lambda \cdot L_A) \right)^2} \right), \quad \text{(C.7)}$$

$$\mathrm{SegSpSNR/dB} = \frac{1}{\#\{\underline{M}_\mathrm{S}\}} \sum_{\lambda \in \underline{M}_\mathrm{S}} \mathrm{SegSpSNR}(\lambda)\,, \quad \text{(C.8)}$$

where $\underline{M}_\mathrm{S}$ is a vector which contains all frames with speech presence and $\#\{\underline{M}_\mathrm{S}\}$ is the number of elements of vector $\underline{M}_\mathrm{S}$. This measure is an indicator for speech distortion. Higher values of SegSpSNR result in a better performance. However, no information about possible noise reduction is provided.

### C.2.3 Cepstral Distance

The real cepstrum of a signal $s(k)$ is defined as the *inverse DFT* (IDFT) of the logarithm of the magnitude spectrum of the signal. For the signal frame $\lambda$ of $s(k)$

the cepstrum is calculated according to

$$\mathbb{C}_x(\lambda, q) = \mathrm{IDFT}\{\ln|\mathrm{DFT}\{s_\lambda(\kappa)\}|\}\ , \tag{C.9}$$

where $s_\lambda(\kappa)$ denotes the samples of signal frame $\lambda$, $\kappa = 0, \ldots, L_F$ is the sample index within the frame, and $q = 0, \ldots, N_{\mathrm{DFT}} - 1$ represents the cepstral bin index (quefrency).

The *cepstral distance* (CD) corresponds to the speech distortion and is defined as distance of the clean speech cepstrum $\mathbb{C}_s(\lambda, q)$ and the filtered clean speech cepstrum $\mathbb{C}_{\tilde{s}}(\lambda, q)$. In general, the magnitude spectrum $|\mathcal{S}(\lambda, \mu)|$ of $s(k)$ is fully described by $N_{\mathrm{DFT}}$ cepstral coefficients. However, the coarse structure of the spectrum is of interest which corresponds to the first cepstral coefficients. Hence, the cepstral distance is calculated for first $N_{CD} = \lceil 0.1 \cdot L_F \rceil$ cepstral coefficients according to

$$\mathrm{CD}(\lambda) = \frac{10}{\ln(10)}\sqrt{(\mathbb{C}_s(\lambda,0) - \mathbb{C}_{\tilde{s}}(\lambda,0))^2 + 2\sum_{q=1}^{N_{CD}}(\mathbb{C}_s(\lambda,q) - \mathbb{C}_{\tilde{s}}(\lambda,q))^2}, \tag{C.10}$$

$$\mathrm{CD/dB} = \frac{1}{\#\{\underline{M}_{\mathrm{S}}\}}\sum_{\lambda\in\underline{M}_{\mathrm{S}}}\mathrm{CD}(\lambda)\ , \tag{C.11}$$

where $\underline{M}_{\mathrm{S}}$ denotes all frames with speech presence and $\#\{\underline{M}_{\mathrm{S}}\}$ is the number of elements of vector $\underline{M}_{\mathrm{S}}$. Lower values of the CD indicate a better performance.

### C.2.4 PESQ

The *perceptual evaluation of speech quality* (PESQ) measure [Rix et al. 2001] aims to provide an objective measure of the perceived audio quality that predicts the results of a subjective listening test. PESQ compares the original clean speech signal $s(k)$ with the enhanced speech signal $\hat{s}(k) = \tilde{y}(k)$. The resulting PESQ values are related to the *mean-opinion score* (MOS) and range from one (bad) to 4.5 (no distortion).

## C.3 Instrumental Measures for Noise Estimation

The logarithmic error measures between the estimated $\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2$ and the true short-term noise *power spectral density* (PSD) $|\mathcal{N}(\lambda,\mu)|^2$ are defined as

$$\mathrm{Err}(\lambda,\mu) = \frac{|\mathcal{N}(\lambda,\mu)|^2}{\left|\widehat{\mathcal{N}}(\lambda,\mu)\right|^2}\ , \tag{C.12}$$

$$\mathrm{LogErr} = \frac{1}{\#\{\underline{M}_{\mathrm{N}}\}N_{\mathrm{DFT}}}\sum_{\lambda\in\underline{M}_{\mathrm{N}}}\sum_{\mu=0}^{N_{\mathrm{DFT}}-1}\left|10\log_{10}\mathrm{Err}(\lambda,\mu)\right|, \tag{C.13}$$

$$\mathrm{LogErr}_{\mathrm{Under}} = \frac{1}{\#\{\underline{M}_{\mathrm{N}}\} N_{\mathrm{DFT}}} \sum_{\lambda \in \underline{M}_{\mathrm{N}}} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left| \max(0, 10 \log_{10} \mathrm{Err}(\lambda, \mu)) \right| , \quad \text{(C.14)}$$

$$\mathrm{LogErr}^{\mathrm{Over}} = \frac{1}{\#\{\underline{M}_{\mathrm{N}}\} N_{\mathrm{DFT}}} \sum_{\lambda \in \underline{M}_{\mathrm{N}}} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left| \min(0, 10 \log_{10} \mathrm{Err}(\lambda, \mu)) \right| , \quad \text{(C.15)}$$

$$\mathrm{LogErr} = \mathrm{LogErr}^{\mathrm{Over}} + \mathrm{LogErr}_{\mathrm{Under}} , \quad \text{(C.16)}$$

where $\underline{M}_{\mathrm{N}}$ denotes all frames to be evaluated in total and $\#\{\underline{M}_{\mathrm{N}}\}$ is the number of elements of vector $\underline{M}_{\mathrm{N}}$. Lower values indicate a better performance. In applications such as speech enhancement an overestimation of the true noise power, as indicated by $\mathrm{LogErr}^{\mathrm{Over}}$, likely results in an attenuation of the speech and thus in speech distortions. On the other hand, a noise power underestimation, pointed out by the $\mathrm{LogErr}_{\mathrm{Under}}$ probably causes a lower noise attenuation.

## C.4 Instrumental Measures for VAD

The instrumental measures are based on the numerical comparison of $v_{\mathrm{bin}}(\lambda) \in \{0, 1\}$ from the *voice activity detector* (VAD) algorithm under test with the ground truth VAD $v_{\mathrm{true}}(\lambda) \in \{0, 1\}$. The true speech presence $v_{\mathrm{true}}(\lambda)$ is provided by the objective measurement of the active speech level according to [ITU-T Recommendation P.56 1993] which is computed from the clean speech signal $s(k)$. Based on $v_{\mathrm{bin}}(\lambda)$ and $v_{\mathrm{true}}(\lambda)$ three VAD measures are defined:

- *Accuracy rate* $P_{\mathrm{a}}$: Percentage of speech frames with correct VAD-estimation;
- *Detection rate* (or true positive rate) $P_{\mathrm{d}}$: Fraction of active speech frames that are detected correctly;
- *False alarm rate* (or false positive rate) $P_{\mathrm{f}}$: Fraction of speech frames without speech that are classified erroneously as speech.

To calculate these measures, three sets of frames are necessary. Here, $\underline{M}_{\mathrm{A}}$ denotes the set of all frames, $\underline{M}_{\mathrm{S}}$ is the set of frames *with* speech activity ($v_{\mathrm{true}}(\lambda) = 1$), and $\underline{M}_{\mathrm{F}}$ is the set of frames *without* speech activity ($v_{\mathrm{true}}(\lambda) = 0$). Let $\#\{\underline{M}_{\mathrm{A}}\}$, $\#\{\underline{M}_{\mathrm{S}}\}$, and $\#\{\underline{M}_{\mathrm{F}}\}$ denote the number of frames in each set, respectively. The objective VAD measures can now be formulated according to

$$P_{\mathrm{a}} = 1 - \frac{1}{\#\{\underline{M}_{\mathrm{A}}\}} \cdot \sum_{\lambda \in \underline{M}_{\mathrm{A}}} \left| v_{\mathrm{bin}}(\lambda) - v_{\mathrm{true}}(\lambda) \right| , \quad \text{(C.17)}$$

$$P_{\mathrm{d}} = \frac{1}{\#\{\underline{M}_{\mathrm{S}}\}} \cdot \sum_{\lambda \in \underline{M}_{\mathrm{S}}} v_{\mathrm{bin}}(\lambda) , \quad \text{(C.18)}$$

$$P_{\mathrm{f}} = \frac{1}{\#\{\underline{M}_{\mathrm{F}}\}} \cdot \sum_{\lambda \in \underline{M}_{\mathrm{F}}} v_{\mathrm{bin}}(\lambda) . \quad \text{(C.19)}$$

# Appendix D

# Independence Assumption of Speech and Noise

Most algorithms in speech enhancement are derived based on *short-term power spectrum* (STPS) quantities. Moreover, noise as well as speech is often estimated from the STPS of the noisy observation $|\mathcal{Y}(\lambda,\mu)|^2$. According to the additive signal model, the STPS $|\mathcal{Y}(\lambda,\mu)|^2$ of the noisy observation is given in terms of speech $\mathcal{S}(\lambda,\mu)$ and noise $\mathcal{N}(\lambda,\mu)$ by

$$
\begin{aligned}
|\mathcal{Y}(\lambda,\mu)|^2 &= |\mathcal{S}(\lambda,\mu)|^2 + |\mathcal{N}(\lambda,\mu)|^2 && \text{(D.1)}\\
&\quad + \mathcal{S}(\lambda,\mu)\mathcal{N}(\lambda,\mu)^* + \mathcal{N}(\lambda,\mu)\mathcal{S}(\lambda,\mu)^* && \text{(D.2)}\\
&= |\mathcal{S}(\lambda,\mu)|^2 + |\mathcal{N}(\lambda,\mu)|^2 && \text{(D.3)}\\
&\quad + \underbrace{2\,|\mathcal{S}(\lambda,\mu)|\,|\mathcal{N}(\lambda,\mu)|\cos(\vartheta_{\mathcal{S}}(\lambda,\mu)-\vartheta_{\mathcal{N}}(\lambda,\mu))}_{\text{cross-term}} && \text{(D.4)}
\end{aligned}
$$

where $\vartheta_{\mathcal{S}}(\lambda,\mu)$ and $\vartheta_{\mathcal{N}}(\lambda,\mu)$ denote the phase of speech and noise, respectively.

During the codebook matching procedure, as described in Sec. 4.1.2, the cross-term of $|\mathcal{Y}(\lambda,\mu)|^2$ is neglected. To verify the irrelevance of the cross-term, the independence assumption of speech and noise is investigated with respect to short-term signal frames $\lambda$ in the following.

At first, an error measure is determined. With $Err_{\mathrm{ct}}(\lambda,\mu)$ denoting the error power which is associated with the cross-term of $|\mathcal{Y}(\lambda,\mu)|^2$,

$$
Err_{\mathrm{ct}}(\lambda,\mu) = 2\,|\mathcal{S}(\lambda,\mu)|\,|\mathcal{N}(\lambda,\mu)|\cos(\vartheta_{\mathcal{S}}(\lambda,\mu)-\vartheta_{\mathcal{N}}(\lambda,\mu))\ , \tag{D.5}
$$

the relative error of the cross-term is defined for the current frame $\lambda$ according to

$$
RelErr_{\mathrm{ct}}(\lambda)/\mathrm{dB} = 10\cdot\log_{10}\left(\frac{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1}|Err_{\mathrm{ct}}(\lambda,\mu)|}{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1}|\mathcal{Y}(\lambda,\mu)|^2}\right)\ . \tag{D.6}
$$

In order to analyze the relative cross-term error $RelErr_{\mathrm{ct}}(\lambda)$ dependent on different speech and noise signals as well as the input SNR, a benchmark is performed. Therefore, noisy signals are generated from all permutations of the following parameters:

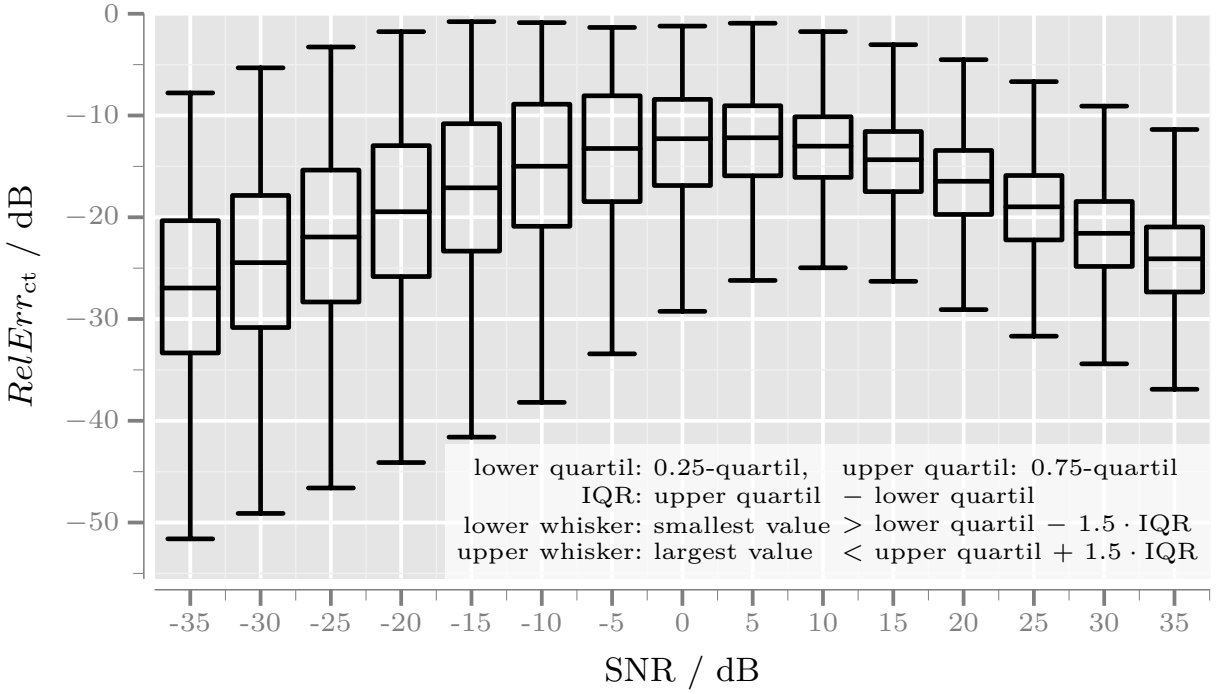



**Figure D.1:** Boxplot of the relative cross-term error $RelErr_{\mathrm{ct}}$ depicted over the input SNR.

- The input SNR ranges from -35 dB to 35 dB in 5 dB steps [1];
- 10 randomly chosen sentences belonging to 5 male and 5 female randomly chosen speakers from the TIMIT database [Garofolo & Consortium 1993] are selected and concatenated;
- The resulting speech sequences are mixed with 10 different types of noise (F16, midsize car, outside traffic, train station, inside train, living room, nature, pub noise, wind, jackhammer).

For the evaluation only signal frames with speech presence are considered. The speech presence is determined by the objective measurement of the active speech level according to [ITU-T Recommendation P.56 1993] from the clean speech component. From all resulting signal frames a boxplot over the input SNR is created and depicted in Fig. D.1.

As expected, the largest relative cross-term error $RelErr_{\mathrm{ct}}(\lambda)$ occurs at 0 dB input SNR with a median value of $RelErr_{\mathrm{ct}} = -13\,\mathrm{dB}$. Since the estimation error of codebook driven speech and noise estimation is in the range of several dB, the influence of the cross-term is considered as non-dominating error. In addition, experiments have confirmed that the resulting error with respect to the application of noise reduction is negligible.

[1] The mixing procedure is detailed in Appendix C.1. Note that for the calculation of the scaling factor to adjust the input SNR only speech and noise signal sections with speech presence are considered.

## Appendix E

# Optimization of $\sigma_n^2$ in the MMSE sense

The optimal gain $\sigma_n^2$ can be found by minimizing the distance between the noisy observation and its estimate. Hence, the optimization of $\sigma_n^2$ for a fixed but arbitrary combination of speech codebook entry $|\mathbb{S}_l(\mu)|^2$ and noise codebook entry $|\mathbb{N}_m(\mu)|^2$ is calculated in the *minimum mean-square error* (MMSE) sense for the current frame $\lambda$. Rewriting Eq. (4.9) yields

$$\left|\widehat{\mathcal{Y}}_{l,m,\sigma_n}(\mu)\right|^2 = \sigma_y^2 \left|\mathbb{S}_l(\mu)\right|^2 + \sigma_n^2 \left(|\mathbb{N}_m(\mu)|^2 - |\mathbb{S}_l(\mu)|^2\right) \tag{E.1}$$

$$= \sigma_y^2 \left|\mathbb{S}_l(\mu)\right|^2 + \sigma_n^2 \left|\mathbb{D}_{l,m}\right|^2 , \tag{E.2}$$

with $|\mathbb{D}_{l,m}|^2 = |\mathbb{N}_m(\mu)|^2 - |\mathbb{S}_l(\mu)|^2$. The estimation error $\operatorname{dist}\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{Y}}}$ in the *mean-square error* (MSE) sense between the noisy observation $\mathcal{Y}(\mu)$ and its estimate $\widehat{\mathcal{Y}}_{l,m,\sigma_n}(\mu)$ is given by

$$\operatorname{dist}\Big|_{\text{SE}}^{\mathcal{Y},\widehat{\mathcal{Y}}} = |\mathcal{Y}(\mu)|^2 - \left|\widehat{\mathcal{Y}}_{l,m,\sigma_n}(\mu)\right|^2 \tag{E.3}$$

$$= |\mathcal{Y}(\mu)|^2 - \sigma_y^2 \left|\mathbb{S}_l(\mu)\right|^2 - \sigma_n^2 \left|\mathbb{D}_{l,m}\right|^2 \tag{E.4}$$

$$\operatorname{dist}\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{Y}}} = \sum_{\mu=0}^{N_{\text{DFT}}-1} \left(\operatorname{dist}\Big|_{\text{SE}}^{\mathcal{Y},\widehat{\mathcal{Y}}}\right)^2 \stackrel{!}{=} \min . \tag{E.5}$$

Building the partial derivation of $\operatorname{dist}\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{Y}}}$ with respect to $\sigma_n^2$ and setting to zero yields the extremum of the distance given by

$$\frac{\partial}{\partial\sigma_n^2}\left(\operatorname{dist}\Big|_{\text{MSE}}^{\mathcal{Y},\widehat{\mathcal{Y}}}\right) = \sum_{\mu=0}^{N_{\text{DFT}}-1} 2\cdot\left(\operatorname{dist}\Big|_{\text{SE}}^{\mathcal{Y},\widehat{\mathcal{Y}}}\right)\frac{\partial\left(\operatorname{dist}\Big|_{\text{SE}}^{\mathcal{Y},\widehat{\mathcal{Y}}}\right)}{\partial\sigma_n^2} \stackrel{!}{=} 0 \tag{E.6}$$

$$= \sum_{\mu=0}^{N_{\text{DFT}}-1} 2\left(|\mathcal{Y}(\mu)|^2 - \sigma_y^2\left|\mathbb{S}_l(\mu)\right|^2 - \sigma_n^2\left|\mathbb{D}_{l,m}\right|^2\right)\left(-\left|\mathbb{D}_{l,m}\right|^2\right)$$

$$= 2\cdot\sigma_n^2\cdot\sum_{\mu=0}^{N_{\text{DFT}}-1}\left(\left|\mathbb{D}_{l,m}\right|^2\right)^2 + 2\cdot\sigma_y^2\cdot\sum_{\mu=0}^{N_{\text{DFT}}-1}\left|\mathbb{S}_l(\mu)\right|^2\left|\mathbb{D}_{l,m}\right|^2$$

$$+ 2 \cdot \sigma_y^2 \cdot \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \mathcal{Y}(\mu) \left|\mathbb{D}_{l,m}\right|^2$$

Hence, Eq. (E.6) can be transformed and $\sigma_n^2$ is expressed as:

$$\sigma_n^2 = \frac{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1} |\mathcal{Y}(\mu)|^2 \cdot |\mathbb{D}_{l,m}|^2 - \sigma_y^2 \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} |\mathbb{S}_l(\mu)|^2 \cdot |\mathbb{D}_{l,m}|^2}{\sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left(|\mathbb{D}_{l,m}|^2\right)^2} . \tag{E.7}$$

Since the second partial derivation of Eq. (E.6) with respect to $\sigma_n^2$ yields

$$\frac{\partial^2}{\partial^2 \sigma_n^2} \left( \mathrm{dist} \left|_{\mathrm{MSE}}^{\mathcal{Y},\widehat{\mathcal{Y}}} \right. \right) = 2 \cdot \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \left(|\mathbb{D}_{l,m}|^2\right)^2 , \tag{E.8}$$

and is greater than zero, the found extremum is in fact a minimum of $\mathrm{dist} \left|_{\mathrm{MSE}}^{\mathcal{Y},\widehat{\mathcal{Y}}} \right.$. Since in general Eq. (E.9) is *not* fulfilled,

$$\sigma_y^2 \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} |\mathbb{S}_l(\mu)|^2 \cdot |\mathbb{D}_{l,m}|^2 \leq \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} |\mathcal{Y}(\mu)|^2 \cdot |\mathbb{D}_{l,m}|^2 , \tag{E.9}$$

it is possible that $\sigma_n^2$ is negative, which violates the model assumption, i. e., $\sigma_n^2$ represents the short-term power of noise.

# Appendix F

# High Quality Video Conferencing

## F.1 Activity Index Calculation

The activity index is a soft quantification of the activity of each participant on a continuous scale between 0 (no activity) and 1 (high activity). It is based on the separated speech signals provided by the parallel beamformers.

For the activity index calculation, it is beneficial to use only frequency sub-bands which exhibit a reasonable SNR. Experimental studies have shown that noise, e.g., structure-borne sound, dominate the sound field especially in the first sub-band (1-268 Hz, cf., Table F.1). Hence, the lowest frequency band is discarded for the activity index calculation. Thus, the activity index calculation relies on the energy of the remaining frequency bands only (cf. Sec. 2, Sec. 2.3.1).

The determination of the activity index $v_{\mathrm{soft,n}}(\lambda)$ of participant $n$ is carried out on signal frame $\lambda$ of the corresponding beamformer output signal $\hat{s}_n(k)$. The typical audio frame length $T_\mathrm{F}$ ranges between 20 ms and 40 ms leading to a frame size of $N_\mathrm{F} = \lfloor f_\mathrm{s} \cdot T_\mathrm{F} \rfloor$samples. The short-term energy of the audio signal of participant $n$ is calculated by

$$V_n(\lambda) = \sum_{i=0}^{N_\mathrm{F}-1} \hat{s}_n^2(\lambda \cdot N_\mathrm{F} + i)\,. \tag{F.1}$$

Due to remaining noise and sudden outliers this energy fluctuates. Thus, recursive smoothing of the energy is applied according to

$$\overline{V}_n(\lambda) = \alpha_2 \cdot \overline{V}_n(\lambda - 1) + (1 - \alpha_2) \cdot V_n(\lambda)\,. \tag{F.2}$$

The smoothing factor $\alpha_2$ is chosen to be 0.98 ($\hat{=}$2 ms equivalent rectangular window

**Table F.1:** Filterbank sub-bands

| Band | Frequency range / Hz | Band | Frequency range / Hz |
|---|---|---|---|
| 1 | 1 - 268 | 4 | 1549 - 2614 |
| 2 | 268 - 839 | 5 | 2614 - 4731 |
| 3 | 839 - 1549 | 6 | 4731 - 12049 |

length at $f_\mathrm{s} = 48\,\mathrm{kHz}$) which results in a system that still adapts quickly to changes while the larger fluctuations are leveled out.

This smoothed energy could directly be used as an indicator of activity of participant $n$. Since the frame energy depends strongly on the recording level of the microphone array an additional step is necessary to map the frame energy into a target scale from 0 (no activity) to 1 (strong activity). It was observed that the smoothed energy measure provides values that increase steeply between situations with no activity and high level of activity. Both, the change in gradient and the mapping of the frame energy values can be achieved simultaneously by means of a sigmoid function. The activity index $v_{\mathrm{soft,n}}(\lambda)$ is calculated by

$$v_{\mathrm{soft,n}}(\lambda) = \frac{1}{1 + \mathrm{e}^{-\beta \cdot \left\{\overline{V}_n(\lambda) - \gamma\right\}}}\,, \tag{F.3}$$

and ensures that the values show a more smooth transition between the different activity levels of the participants.

The parameters $\beta$ and $\gamma$ of the sigmoid function depend strongly on the expected minimum and maximum smoothed frame energy. Since these quantities are related to the calibration of the microphones and background noise, they are not known *a priori* and adaptive adjustment of the parameters is required. Therefore, the maximum statistics of a sliding time window containing the smoothed energy frames of the last seconds (typically 30 – 180 s) are exploited.

Given an audio frame buffer $\overline{V}_{\mathrm{Buffer}}(\lambda)$ containing the energies of the past frames sorted in descending order, an estimate of the expected maximum frame speech energy is obtained by averaging the $L_N$ highest-energy frames according to

$$\overline{V}_{\mathrm{maxStat}} = \frac{1}{L_N} \sum_{i=0}^{L_N - 1} \overline{V}_{\mathrm{Buffer}}(i)\,. \tag{F.4}$$

The parameter $\gamma$, which defines the center of the sigmoid function, can now be calculated according to

$$\gamma = \max\left\{\frac{\overline{V}_{\mathrm{maxStat}}}{2}, \overline{V}_{\mathrm{min}}\right\}, \tag{F.5}$$

with $\overline{V}_{\mathrm{min}}$ serving as a lower bound to prevent underestimation for the expected frame speech energy, e. g., in the initialization phase. The gradient of the function is controlled by $\beta$ which can be determined using the inverse of (F.3) by

$$\beta = -\frac{\ln\left(\frac{1}{0.99} - 1\right)}{\gamma}\,. \tag{F.6}$$

With this choice of $\beta$ and $\gamma$ the activity index for a frame energy of $\overline{V}(\lambda) = \overline{V}_{\mathrm{maxStat}}$ results in $v_{\mathrm{soft}}(\lambda) = 0.99$. The parameters $\beta$ and $\gamma$ are updated according to this procedure in each frame.

## F.2 Detailed Objective VAD Measures

| Participant | Accuracy rate $P_a$ | | Detection rate $P_d$ | | False alarm rate $P_f$ | |
|---|---|---|---|---|---|---|
| # | NFB | DSB | NFB | DSB | NFB | DSB |
| A | **0.90** | 0.82 | **0.85** | 0.90 | **0.06** | 0.23 |
| B | **0.87** | 0.75 | **0.73** | 0.68 | **0.05** | 0.22 |
| C | **0.91** | 0.69 | **0.84** | 0.72 | **0.06** | 0.32 |

**Table F.2:** Detailed objective VAD measures for VAD-AI

| Participant | Accuracy rate $P_a$ | | Detection rate $P_d$ | | False alarm rate $P_f$ | |
|---|---|---|---|---|---|---|
| # | NFB | DSB | NFB | DSB | NFB | DSB |
| A | 0.58 | 0.53 | 0.73 | 0.57 | 0.54 | 0.50 |
| B | 0.50 | 0.48 | 0.50 | 0.46 | 0.50 | 0.51 |
| C | 0.56 | 0.57 | 0.80 | 0.64 | 0.65 | 0.46 |

**Table F.3:** Detailed objective VAD measures for VAD-Ghosh

| Participant | Accuracy rate $P_a$ | | Detection rate $P_d$ | | False alarm rate $P_f$ | |
|---|---|---|---|---|---|---|
| # | NFB | DSB | NFB | DSB | NFB | DSB |
| A | 0.64 | 0.56 | 0.83 | 0.87 | 0.50 | 0.67 |
| B | 0.56 | 0.49 | 0.77 | 0.86 | 0.55 | 0.70 |
| C | 0.55 | 0.48 | 0.78 | 0.86 | 0.56 | 0.69 |

**Table F.4:** Detailed objective VAD measures for VAD-Sohn

| VAD | Position | Accuracy rate $P_a$ | | False alarm rate $P_f$ | |
|---|---|---|---|---|---|
| | (cf. Fig. 6.2) | NFB | DSB | NFB | DSB |
| VAD-AI | 3 | **0.81** | 0.47 | **0.19** | 0.53 |
| | 5 | **0.93** | 0.62 | **0.07** | 0.38 |
| VAD-Ghosh | 3 | 0.42 | 0.47 | 0.58 | 0.53 |
| | 5 | 0.42 | 0.51 | 0.58 | 0.49 |
| VAD-Sohn | 3 | 0.39 | 0.24 | 0.61 | 0.76 |
| | 5 | 0.36 | 0.24 | 0.64 | 0.76 |

**Table F.5:** Detailed objective VAD measures for all VADs at positions between the talkers without video information

# Mathematical Notation & Abbreviations

## Mathematical Notation

In this thesis, the following conventions are used to denote quantities: vectors are underlined, e. g., $\underline{y}$, scalar values are not, e. g., $y$. The cardinality of a vector, i. e., the number of elements is indicated by the $\#\{\cdot\}$ operator, e. g., $\#\{\underline{y}\}$. Estimated or approximated variables are marked with a hat, e. g., $\hat{y}$, and averaged or smoothed values are denoted with a bar, e. g., $\bar{y}$.

Time-domain signals are written in lower-case letters, e. g., $y(k)$ with the sample index $k$. The complex-valued *discrete Fourier transform* (DFT) coefficients are labeled with the calligraphic upper-case letters, e. g., $\mathcal{Y}(\lambda, \mu)$ with DFT bin index $\mu \in \{0, 1, \ldots, N_{\mathrm{DFT}} - 1\}$, even DFT size $N_{\mathrm{DFT}}$, and frame index $\lambda$.

## Mathematical Operators

| | |
|---|---|
| $\approx$ | approximately equal to |
| $\widehat{=}$ | equivalent to (usually a unit conversion) |
| $\overset{!}{=}$ / $\overset{!}{\leq}$ | shall be equal to / shall be less than or equal to |
| $\wedge$ / $\vee$ | logical and / or |
| $\in$ | element of |
| $\forall$ | for all |
| $x^*$ | complex conjugate of $x$ |
| $\lvert x \rvert$ | absolute value of $x$ |
| $\lfloor x \rfloor$ | floor function, i. e., largest integer which is not greater than $x$ |
| $\lceil x \rceil$ | ceiling function, i. e., smallest integer which is not less than $x$ |
| $\mathbb{E}\{x(k)\}$ | expectation value of $x(k)$ |
| $\mathrm{Re}\{x\}$ | real part of $x$ |
| $\mathrm{Im}\{x\}$ | imaginary part of $x$ |
| $\exp\{x\}$ | exponential function $\mathrm{e}^x$ |
| $\log\{x\}$ | logarithm of $x$ to base 10 |
| $\max_x\{f(x)\}$ | maximum of $f(x)$ over $x$ |

| | |
|---|---|
| $\arg\max_x\{f(x)\}$ | argument $x$ of maximum of $f(x)$ over $x$ |
| $\underset{x}{\text{mean}}\{f(x)\}$ | average of $f(x)$ over all $x$ of a finite set |

## Principal Symbols

| | |
|---|---|
| $\alpha(\lambda)$ | time dependent scaling parameter of Baseline Tracing |
| $\alpha_{\Phi}$ | VAD based noise PSD smoothing factor |
| $\alpha_{\mathcal{G}}(\lambda,\mu)$ | parameter of spectral weighting gain |
| $\alpha_{\xi}$ | decision directed SNR smoothing factor |
| $\beta(\lambda,\mu)$ | tracing factor of Baseline Tracing |
| $\beta_{\mathcal{G}}(\lambda,\mu)$ | parameter of spectral weighting gain |
| $\delta(k)$ | unit impulse sequence |
| $\Delta(\lambda,\mu)$ | adaptive step-size parameter of Baseline Tracing |
| $\vartheta_{\mathcal{Y}}(\mu)$ | phase of noisy signal in the DFT domain |
| $\vartheta_{\mathcal{N}}(\mu)$ | phase of noise signal in the DFT domain |
| $\vartheta_{\mathcal{S}}(\mu)$ | phase of speech signal in the DFT domain |
| $\kappa$ | time index within a single signal frame $\lambda$ |
| $\lambda$ | frame index |
| $\mu$ | DFT bin index |
| $\phi(\mu)$ | speech dependent scaling parameter over the frequency of Baseline Tracing |
| $\Omega$ | normalized frequency |
| $\widehat{\mathcal{N}}(\lambda,\mu)$ | estimated DFT coefficients of noise signal |
| $\mathcal{C}$ | codebook containing codebook entry vectors |
| $D(\lambda)$ | parameter of Baseline Tracing |
| $d(\lambda)$ | parameter of Baseline Tracing |
| $\text{dist}(\mathcal{P},\widehat{\mathcal{P}})$ | distance between power spectra $\mathcal{P}(\mu)$ and $\widehat{\mathcal{P}}(\mu)$ |
| $\text{dist}\Big\vert_{\text{IS}}^{\mathcal{P},\widehat{\mathcal{P}}}(\lambda)$ | Itakura-Saito distance between power spectra $\mathcal{P}(\mu)$ and $\widehat{\mathcal{P}}(\mu)$ |
| $\text{dist}\Big\vert_{\text{MSE}}^{\mathcal{P},\widehat{\mathcal{P}}}(\lambda)$ | MSE between power spectra $\mathcal{P}(\mu)$ and $\widehat{\mathcal{P}}(\mu)$ |
| $\text{dist}\Big\vert_{\text{REL}}^{\mathcal{P},\widehat{\mathcal{P}}}(\lambda)$ | Relative energy distance between power spectra $\mathcal{P}(\mu)$ and $\widehat{\mathcal{P}}(\mu)$ |
| e | Euler's number |
| $E_n^{\mathcal{N}}(\lambda,\mu)$ | estimation error of the noise estimate |
| $E_s^{\mathcal{S}}(\lambda,\mu)$ | estimation error of the speech estimate |
| $E_i^{\mathcal{Y}}(\lambda,\mu)$ | estimation error of the noisy observation estimate |
| $f$ | continuous frequency |
| $f_{\text{P}}$ | Pitch frequency |

| | |
|---|---|
| $f_\mathrm{s}$ | sampling rate |
| $g_w$ | window normalization factor |
| $\mathcal{G}(\lambda, \mu)$ | spectral weighting gain |
| $g_\mathrm{min}$ | lower bound for spectral weighting gain |
| j | imaginary unit |
| $k$ | sample index |
| $L_A$ | frame advance in number of samples |
| $L_F$ | frame size in number of samples |
| $\mathrm{LTA}(f)$ | long-term speech spectrum average |
| $\mathrm{LTA}^{-1}(\mu)$ | inverse long-term speech spectrum average |
| $c(\lambda, \mu)$ | information combining coefficients |
| $n(k)$ | noise signal in the time domain |
| $N_\mathrm{DFT}$ | DFT size, i. e., number of DFT bins |
| $\widehat{\mathcal{N}}_\mathrm{CB}(\lambda, \mu)$ | DFT coefficients of codebook estimated noise signal |
| $\mathbb{N}_m(\mu)$ | noise codebook entry with entry index $m$ |
| $\sigma_n$ | gain factor of noise codebook entry |
| $\mathcal{N}(\lambda, \mu)$ | DFT coefficients of noise signal $n(k)$ |
| $\mathbb{N}$ | set of positive integers |
| $\mathbb{N}_0$ | set of non-negative integers |
| $p$ | noise estimate change in percent every 10 ms |
| $q$ | quefrency bin index |
| $\mathcal{S}(\mu)$ | DFT coefficients of speech signal $s(k)$ |
| $s(k)$ | speech signal in the time domain |
| $\widehat{\mathcal{S}}(\mu)$ | estimated DFT coefficients of speech signal |
| $\hat{s}(k)$ | enhanced speech signal in the time domain |
| $\widehat{\mathcal{S}}_\mathrm{CB}(\lambda, \mu)$ | DFT coefficients of codebook estimated speech signal |
| $\mathbb{S}_l(\mu)$ | speech codebook entry with entry index $l$ |
| $\sigma_s$ | gain factor of speech codebook entry |
| $\gamma(\lambda, \mu)$ | *a posteriori* SNR |
| $\xi(\lambda, \mu)$ | *a priori* SNR |
| $\mathrm{SNR}_\mathrm{CB}(\lambda, \mu)$ | codebook SNR |
| $\mathrm{SNR}_\mathrm{DD}(\lambda, \mu)$ | decision directed SNR |
| $\mathrm{SNR}_\mathrm{i}(\lambda, \mu)$ | instantaneous SNR |
| $t$ | continuous time |
| $T_A$ | frame shift in seconds |
| $T_F$ | frame length in seconds |
| $\mathcal{T}$ | training set for codebook creation |

| | |
|---|---|
| $w(k)$ | window function |
| $y(k)$ | noisy signal in the time domain |
| $\mathcal{Y}(\lambda, \mu)$ | DFT coefficients of noisy signal $y(k)$ |
| $\lvert\mathcal{Y}(\lambda, \mu)\rvert$ | magnitude of noisy signal $y(k)$ |
| $\widehat{\mathcal{Y}}(\lambda, \mu)$ | DFT coefficients of estimated noisy signal |
| $z$ | $z$-transform |
| $\mathbb{Z}$ | set of integers |
| $\widehat{\Phi}_{nn}(\lambda, \mu)$ | short-term estimate of PSD of noise |
| $\overline{\Phi}_{nn}(\lambda, \mu)$ | short-term PSD of noise |

## Acronyms

**API** application programming interface

**AR** auto-regressive

**BWE** artificial bandwidth extension

**CAT-iq** cordless advanced technology – internet and quatliy

**CD** cepstral distance

**DFT** discrete Fourier transform

**DSP** digital signal processor

**DTMF** dual-tone multi-frequency signaling

**FFT** fast Fourier transform

**FIR** finite impulse response

**GSC** generalized sidelobe canceller

**HMM** hidden markov model

**IDFT** inverse DFT

**IIR** infinite impulse response

**IMS** IP Multimedia Subsystem

**ISDN** integrated services digital network

**LCMV** linearly constrained minimum variance

**LPC** linear prediction coefficient

**LSF** line spectral frequencies

**LTA** long-term speech spectrum average

**MAP** maximum *a posteriori*

**MFCC** mel frequency cepstral coefficients

**ML** maximum likelihood

**MMSE** minimum mean-square error

**MOS** mean-opinion score

**MSE** mean-square error

**MVDR** minimum variance distortionless response

**MWF** multichannel Wiener filter

**NELE** near-end listening enhancement

**PBX** private branch exchange

**PDF** probability density function

**PESQ** perceptual evaluation of speech quality

**PSD** power spectral density

**PSTN** public switched telephone network

**QMF** quadrature mirror filter

**RASTA-PLP** relative spectral transform - perceptual linear prediction

**ROC** receiver operating characteristic

**ROI** region of interest

**SBC** single-board computer

**SegNA** segmental noise attenuation

**SegSA** segmental speech attenuation

**SegSpSNR** segmental speech SNR

**SIP** session initiation protocol

**SNR** signal-to-noise ratio

**SNR** *a priori* SNR

**SNR** *a posteriori* SNR

**SPP** speech presence probability

**STFD** short-term Fourier domain

**STPS** short-term power spectrum

**VAD** voice activity detector

**VoIP** voice over IP

**VQ** vector quantizer

**ZCR** zero-crossing rate

# Bibliography

Publications by the author are marked with an asterisk (∗).


**Atrey**, Pradeep K.; **Hossain**, M. Anwar; **Saddik**, Abdulmotaleb El; **Kankanhal**, Mohan S. (2010). "Multimodal fusion for multimedia analysis: a survey". In: *Multimedia Systems* 16.2, pp. 345–379. ISSN: 1432-1882. DOI: 10.1007/s00530-010-0182-0 (cit. on p. 135).

**Baasch**, Christin; **Rajan**, Vasudev Kandade; **Krini**, Mohamed; **Schmidt**, Gerhard (2014). "Low-Complexity Noise Power Spectral Density Estimation For Harsh Automobile Environments". In: *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 219–223 (cit. on pp. 4, 28, 39).

**Beaugeant**, Christophe; **Schönle**, Martin; **Varga**, Imre (2006). "Challenges of 16 kHz in Acoustic Pre- and Post-Processing for Terminals". In: *IEEE Communications Magazine* 44.5, pp. 98–104. ISSN: 0163-6804. DOI: 10.1109/MCOM.2006.1637953 (cit. on p. 54).

**Benesty**, Jacob; **Chen**, Jingdong; **Huang**, Yiteng; **Cohen**, Israel (2009). *Noise reduction in speech processing*. Vol. 2. Springer Science & Business Media (cit. on pp. 3, 19).

**Benesty**, Jacob; **Sondhi**, M. Mohan; **Huang**, Yiteng (2007). *Springer Handbook of Speech Processing*. Englisch. 2008th ed. Berlin ; London: Springer. ISBN: 978-3-540-49125-5 (cit. on pp. 3, 19, 23).

**Bertsekas**, Dimitri P. (1996). *Constrained optimization and Lagrange multiplier methods*. Academic press. ISBN: 1-886529-04-3 (cit. on p. 104).

**Boll**, Steven F. (1979). "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 27.2, pp. 113–120. ISSN: 0096-3518. DOI: 10.1109/TASSP.1979.1163209 (cit. on pp. 3, 19, 33, 65).

**Brandstein**, Michael; **Ward**, Darren B. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag (cit. on p. 7).

**Breithaupt**, Colin; **Gerkmann**, Timo; **Martin**, Rainer (2007). "Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement Without Musical Noise". In: *IEEE Signal Processing Letters* 14.12, pp. 1036–1039. ISSN: 1070-9908. DOI: 10.1109/LSP.2007.906208 (cit. on p. 3).

**Breithaupt**, Colin; **Krawczyk**, Martin; **Martin**, Rainer (2008). "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pp. 4037–4040. DOI: 10.1109/ICASSP.2008.4518540 (cit. on p. 3).

**Brennan**, Donald G. (2003). "Linear diversity combining techniques". In: *Proceedings of the IEEE* 91.2, pp. 331–356. ISSN: 0018-9219. DOI: 10.1109/JPROC.2002.808163 (cit. on p. 98).

**Bronstein**, Ilja N.; **Semendjajew**, Konstantin A.; **Musiol**, Gerhard; **Mühlig**, Heiner (1999). *Taschenbuch der Mathematik*. 4th ed. Verlag Harri Deutsch. ISBN: 3-8171-2014-1 (cit. on pp. 66, 104).

**Bub**, Udo; **Hunke**, Martin; **Waibel**, Alex (1995). "Knowing who to listen to in speech recognition: visually guided beamforming". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Detroit, MI, USA, pp. 848–851. DOI: 10.1109/ICASSP.1995.479827 (cit. on p. 135).

∗ **Bulla**, Christopher; **Feldmann**, Christian; **Schäfer**, Magnus; **Heese**, Florian; **Schlien**, Thomas; **Schink**, Martin (2013). "High Quality Video Conferencing: Region of Interest Encoding and Joint Video/Audio Analysis". In: *International Journal on Advances in Telecommunications* 6.3 & 4, pp. 153–163. ISSN: 1942-2601 (cit. on pp. 6, 137).

**Byrd**, Richard H.; **Gilbert**, Jean Charles; **Nocedal**, Jorge (2000). "A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming". In: *Mathematical Programming* 89.1, pp. 149–185 (cit. on p. 13).

**Chen**, R.F.; **Chan**, C.F.; **So**, H.C.; **Lee**, J.; **Leung**, C.Y. (2009). "Speech enhancement in car noise environment based on an analysis-synthesis approach using harmonic noise model". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pp. 4413–4416. DOI: 10.1109/ICASSP.2009.4960608 (cit. on pp. 4, 19).

**Cho**, Yong Duk; **Kondoz**, Ahmet (2001). "Analysis and improvement of a statistical model-based voice activity detector". In: *Signal Processing Letters, IEEE* 8.10, pp. 276–278 (cit. on p. 80).

**Cohen**, Israel (2003). "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging". In: *Speech and Audio Processing, IEEE Transactions on* 11.5, pp. 466–475 (cit. on pp. 4, 28).

**Cohen**, Israel; **Berdugo**, Baruch (2001). "Speech enhancement for non-stationary noise environments". In: *Signal Processing* 81.11, pp. 2403–2418. ISSN: 0165-1684. DOI: 10.1016/S0165-1684(01)00128-1 (cit. on p. 30).

— (2002). "Noise estimation by minima controlled recursive averaging for robust speech enhancement". In: *IEEE Signal Processing Letters* 9.1, pp. 12–15. ISSN: 1070-9908. DOI: 10.1109/97.988717 (cit. on p. 28).

**Cooley**, James W.; **Tukey**, John W. (1965). "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90, pp. 297–301. ISSN: 0025-5718. DOI: 10.2307/2003354 (cit. on p. 24).

**CoVR** (2013). *Connected Visual Reality (CoVR)*. URL: http://www.covr.rwth-aachen.de (cit. on p. 135).

**Cox**, John Charles (1984). "The minimum detectable delay of speech and music". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84*. Vol. 9, pp. 136–139. DOI: 10.1109/ICASSP.1984.1172559 (cit. on p. 21).

**Crochiere**, Ronald E. (1980). "A weighted overlap-add method of short-time Fourier analysis/Synthesis". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 28.1, pp. 99–102. ISSN: 0096-3518. DOI: 10.1109/TASSP.1980.1163353 (cit. on p. 24).

**De Haan**, Jan Mark; **Grbic**, Nedelko; **Claesson**, Ingvar; **Nordholm**, Sven (2001). "Design of oversampled uniform DFT filter banks with delay specification using quadratic optimization". In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. Vol. 6. IEEE, pp. 3633–3636 (cit. on p. 7).

**Deng**, Feng; **Bao**, Changchun (2016). "Speech enhancement based on AR model parameters estimation". In: *Speech Communication* 79, pp. 30–46. ISSN: 0167-6393. DOI: 10.1016/j.specom.2016.02.006 (cit. on p. 65).

**Doblinger**, Gerhard (1995). "Computationally Efficient Speech Enhancement By Spectral Minima Tracking In Subbands". In: *Proc. Eurospeech*, pp. 1513–1516 (cit. on pp. 4, 28, 36, 43).

**Doclo**, Simon; **Moonen**, Marc (2003). "Design of Far-field and Near-field Broadband Beamformers using Eigenfilters". In: *Signal Processing* 83.12, pp. 2641–2673 (cit. on p. 8).

**Dörbecker**, Matthias; **Ernst**, Stefan (1996). "Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation". In: *European Signal Processing Conference, 1996. EUSIPCO 1996. 8th*, pp. 1–4 (cit. on p. 4).

**Ephraim**, Yariv; **Malah**, David (1984). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 32.6, pp. 1109–1121 (cit. on pp. 3, 19, 31, 50, 52, 60, 65, 72, 111, 114).

— (1985). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 33.2, pp. 443–445. ISSN: 0096-3518. DOI: 10.1109/TASSP.1985.1164550 (cit. on pp. 3, 19, 52, 65).

**Erkelens**, Jan S.; **Hendriks**, Richard C.; **Heusdens**, Richard; **Jensen**, Jesper (2007). "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.6, pp. 1741–1752. ISSN: 1558-7916. DOI: 10.1109/TASL.2007.899233 (cit. on p. 3).

∗ **Esch**, Thomas; **Heese**, Florian; **Geiser**, Bernd; **Vary**, Peter (2010**a**). "Wideband Noise Suppression Supported by Artificial Bandwidth Extension Techniques". In: *ICASSP*. IEEE, pp. 4790–4793 (cit. on pp. 6, 54, 57, 60).

∗ **Esch**, Thomas; **Rüngeler**, Matthias; **Heese**, Florian; **Vary**, Peter (2010**b**). "A Modified Minimum Statistics Algorithm for Reducing Time Varying Harmonic Noise". In: *ITG-Fachtagung Sprachkommunikation*. Berlin, Germany: VDE Verlag GmbH. ISBN: 978-3-8007-3300-2 (cit. on p. 6).

∗ — (2010**c**). "Combined Reduction of Time Varying Harmonic and Stationary Noise Using Frequency Warping". In: *Conference Record of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*. Piscataway, NJ, USA: IEEE, pp. 533–537. ISBN: 978-1-4244-9720-1 (cit. on pp. 4, 6, 19).

∗ — (2012). "Estimation of Rapidly Time-Varying Harmonic Noise for Speech Enhancement". In: *IEEE Signal Processing Letters* 19.10, pp. 659–662. ISSN: 1070-9908. DOI: 10.1109/LSP.2012.2211011 (cit. on p. 6).

**ETSI EG 202 396-1** (2009). *Speech and multimedia Transmission Quality (STQ); Part 1: Background noise simulation technique and background noise database* (cit. on p. 88).

**ETSI Recommendation GSM 06.32** (1996). *GSM recommendations for VAD: GSM 06.32, GSM 06.42, GSM 06.82; Voice activity detection for full rate speech traffic channels* (cit. on pp. 79, 87, 90).

**Fisher**, Etan; **Rafaely**, Boaz (2011). "Near-Field Spherical Microphone Array Processing With Radial Filtering". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.2, pp. 256–265. ISSN: 1558-7916. DOI: 10.1109/TASL.2010.2047421 (cit. on p. 8).

**Garofolo**, John S.; **Consortium**, Linguistic Data (1993). *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium (cit. on pp. 78, 87, 88, 93, 110, 117, 166).

**Geiser**, Bernd; **Taddei**, Hervé; **Vary**, Peter (2007). "Artificial Bandwidth Extension without Side Information for ITU-T G.729.1". In: *Proceedings of European Conference on Speech Communication and Technology (INTERSPEECH)*. ISCA, pp. 2493–2496 (cit. on pp. 54, 57, 58).

**Gerkmann**, Timo; **Breithaupt**, Colin; **Martin**, Rainer (2008). "Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors". In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.5, pp. 910–919. ISSN: 1558-7916. DOI: 10.1109/TASL.2008.921764 (cit. on p. 30).

**Gerkmann**, Timo; **Hendriks**, Richard C. (2011). "Noise power estimation based on the probability of speech presence". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 145–148 (cit. on pp. 4, 28–30, 34, 36, 38, 43, 65, 67, 111, 113).

— (2012). "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20.4, pp. 1383–1393 (cit. on pp. 29, 30, 34, 52).

**Ghosh**, Prasanta Kumar; **Tsiartas**, Andreas; **Narayanan**, Shrikanth (2011). "Robust Voice Activity Detection Using Long-Term Signal Variability". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 19.3, pp. 600–613. ISSN: 1558-7916. DOI: 10.1109/TASL.2010.2052803 (cit. on pp. 80, 87, 89, 90, 92, 96, 138).

**Godsill**, Simon; **Buchner**, Herbert; **Skoglund**, Jan (2015). "Detection and suppression of keyboard transient noise in audio streams with auxiliary keybed microphone". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 379–383. DOI: 10.1109/ICASSP.2015.7177995 (cit. on pp. 4, 19).

**Gonzalez**, Sira; **Brookes**, Mike (2014). "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.2, pp. 518–530. ISSN: 2329-9290. DOI: 10.1109/TASLP.2013.2295918 (cit. on p. 66).

**Griffiths**, Lloyd J.; **Jim**, Charles W. (1982). "An alternative approach to linearly constrained adaptive beamforming". In: *IEEE Transactions on Antennas and Propagation* 30.1, pp. 27–34. ISSN: 0018-926X. DOI: 10.1109/TAP.1982.1142739 (cit. on p. 7).

**Gustafsson**, Stefan; **Martin**, Rainer; **Vary**, Peter (1996). "On the Optimization of Speech Enhancement Systems Using Instrumental Measures". In: *Proceedings of Workshop on Quality Assessment in Speech, Audio and Image Communication*. ITG / EURASIP, pp. 36–40 (cit. on p. 159).

∗ **Hadad**, Elior; **Heese**, Florian; **Gannot**, Sharon; **Vary**, Peter (2014). "Multichannel Audio Database in Various Acoustic Environments". In: *IWAENC*. EURECOM, Sophia Antipolis, France (cit. on pp. 6, 17).

∗ **Hamm**, Laurits; **Engelbert**, Tobias; **Lausuch**, Jose; **Nicolas**, Arturo Martin de; **Kandasamy**, Ramsundar; **Schink**, Martin; **Feldmann**, Christian; **Bulla**, Christopher; **Schäfer**, Magnus; **Heese**, Florian; **Schlien**, Thomas; **Antweiler**, Christiane (2012). "Connected Visual Reality – High Quality Audio Visual Communication in Heterogeneous Networks". In: *International Workshop on Acoustic Signal Enhancement (IWAENC)*. RWTH Aachen University (cit. on pp. 136, 142).

**Hansen**, John H.L. (1991). "Speech enhancement employing adaptive boundary detection and morphological based spectral constraints". In: *, 1991 International Conference on Acoustics, Speech, and Signal Processing, 1991. ICASSP-91*, 901–904 vol.2. DOI: 10.1109/ICASSP.1991.150485 (cit. on p. 33).

**Hänsler**, Eberhard; **Schmidt**, Gerhard (2006). *Topics in Acoustic Echo and Noise Control, Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise and Speech Processing*. ISBN: 3-540-33212-X (cit. on pp. 3, 19).

— (2008). *Speech and Audio Processing in Adverse Environments*. Signals and Communication Technology. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-70601-4 (cit. on pp. 3, 19).

**Hao**, Yue; **Bao**, Changchun (2015). "An improved dictionary learning method for speech enhancement". In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 144–147. DOI: 10.1109/APSIPA.2015.7415490 (cit. on p. 65).

**Harris**, Fredric J. (2004). *Multirate signal processing for communication systems*. Prentice Hall PTR (cit. on pp. 13, 15).

**Haykin**, Simon; **Litva**, John; **Shepherd**, Terence J. (1993). *Radar Array Processing* (cit. on p. 7).

**Haykin**, Simon; **Liu**, KJ Ray (2010). *Handbook on array processing and sensor networks*. Vol. 63. John Wiley & Sons (cit. on p. 7).

∗ **Heese**, Florian; **Esch**, Thomas; **Geiser**, Bernd; **Vary**, Peter (2010). "Noise Reduction for Wideband Speech Exploiting Spectral Dependencies Based on Conditional Estimation". In: *ITG-Fachtagung Sprachkommunikation*. Berlin: VDE Verlag GmbH. ISBN: 978-3-8007-3300-2 (cit. on pp. 6, 54).

∗ **Heese**, Florian; **Esch**, Thomas; **Vary**, Peter (2011). "Dual Channel Reduction of Rapidly Varying Harmonic and Random Noise Using a Spot Microphone". In: *Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany* (cit. on p. 6).

∗ **Heese**, Florian; **Geiser**, Bernd; **Vary**, Peter (2012**a**). "Intelligibility Assessment of a System for Artifical Bandwidth Extension of Telephone Speech". In: *DAGA*. DEGA, pp. 905–906 (cit. on pp. 6, 54, 136, 141).

∗ **Heese**, Florian; **Nelke**, Christoph Matthias; **Niermann**, Markus; **Vary**, Peter (2014**a**). "Selflearning Codebook Speech Enhancement". In: *ITG Fachtagung Sprachkommunikation*. VDE Verlag GmbH (cit. on p. 6).

∗ **Heese**, Florian; **Niermann**, Markus; **Vary**, Peter (2014**b**). "Real-Time Near-End Listening Enhancement for Mobile Phones". In: *ITG Fachtagung Sprachkommunikation*. Show and Tell Demonstration. VDE Verlag GmbH (cit. on pp. 141, 142).

∗ — (2015). "Speech-Codebook Based Soft Voice Activity Detection". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, QLD: IEEE, pp. 4335–4339. DOI: 10.1109/ICASSP.2015.7178789 (cit. on pp. 6, 80).

∗ **Heese**, Florian; **Schäfer**, Magnus; **Vary**, Peter; **Hadad**, Elior; **Markovich Golan**, Shmulik; **Gannot**, Sharon (2012**b**). "Comparison of supervised and semi-supervised beamformers using real audio recordings". In: *2012 IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, pp. 1–5. DOI: 10.1109/EEEI.2012.6376965 (cit. on p. 6).

∗ **Heese**, Florian; **Schäfer**, Magnus; **Wernerus**, Jona; **Vary**, Peter (2013). "Numerical Near Field Optimization of a Non-Uniform Sub-band Filter-and-Sum Beamformer". In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC: IEEE (cit. on pp. 6, 8).

∗ **Heese**, Florian; **Vary**, Peter (2015). "Noise PSD Estimation By Logarithmic Baseline Tracing". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, QLD: IEEE. DOI: 10.1109/ICASSP.2015.7178803 (cit. on pp. 6, 33, 65).

**Hendriks**, Richard C.; **Heusdens**, Richard; **Jensen**, Jesper (2010). "MMSE based noise PSD tracking with low complexity". In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4266–4269 (cit. on pp. 4, 28, 29, 65, 67).

**Hermansky**, Hynek (1990). "Perceptual linear predictive (PLP) analysis of speech". In: *The Journal of the Acoustical Society of America* 87.4, pp. 1738–1752. ISSN: 0001-4966. DOI: 10.1121/1.399423 (cit. on p. 57).

**Hermansky**, Hynek; **Morgan**, Nelson (1994). "RASTA processing of speech". In: *IEEE Transactions on Speech and Audio Processing* 2.4, pp. 578–589. ISSN: 1063-6676. DOI: 10.1109/89.326616 (cit. on p. 57).

**Hosten**, Peter; **Steiger**, Andreas; **Feldmann**, Christian; **Bulla**, Christopher (2013). "Performance evaluation of object representations in mean shift tracking". In: *Proc. Int. Conf. Advances Multimedia*. Venice, Italy, pp. 1–6 (cit. on p. 137).

**Huber**, Johannes B.; **Huettinger**, Simon (2003). "Information processing and combining in channel coding". In: *Proc. 3rd Int. Symp. Turbo Codes*, pp. 1–5 (cit. on pp. 2, 98).

**Itakura**, Fumitada (1975). "Line spectrum representation of linear predictor coefficients of speech signals". In: *The Journal of the Acoustical Society of America* 57.S1, S35–S35. ISSN: 0001-4966. DOI: 10.1121/1.1995189 (cit. on p. 73).

**Itakura**, Fumitada; **Saito**, Shuzo (1968). "Analysis synthesis telephony based on the maximum likelihood method". In: *Proceedings of the 6th International Congress on Acoustics*. Vol. 17. pp. C17–C20, pp. C17–C20 (cit. on p. 72).

**ITU-T Recommendation P.50** (1999). *Artificial voices (ITU-T Recommendation P.50)* (cit. on pp. 41, 53).

**ITU-T Recommendation P.56** (1993). *Telephone Transmission Quality Objective Measuring Apparatus: Objective Measurement of Active Speech Level*. Series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks; Methods for Objective and Subjective Assessment of Quality (cit. on pp. 74, 88, 95, 160, 163, 166).

**Jansen**, Jack; **Cesar**, Pablo; **Bulterman**, Dick C. A.; **Stevens**, Tim; **Kegel**, Ian; **Issing**, Jochen (2011). "Enabling Composition-Based Video-Conferencing for the Home". In: *Multimedia, IEEE Transactions on* 13.5, pp. 869–881. DOI: 10.1109/TMM.2011.2159369 (cit. on p. 135).

**Jax**, Peter; **Vary**, Peter (2003). "On artificial bandwidth extension of telephone speech". In: *Signal Processing, IEEE* 83.8 (cit. on pp. 136, 141).

— (2004). "Feature selection for improved bandwidth extension of speech signals". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*. Vol. 1, I–697–700 vol.1. DOI: 10.1109/ICASSP.2004.1326081 (cit. on p. 57).

— (2006). "Bandwidth extension of speech signals: a catalyst for the introduction of wideband speech coding?" In: *IEEE Communications Magazine* 44.5, pp. 106–111. ISSN: 0163-6804. DOI: 10.1109/MCOM.2006.1637954 (cit. on p. 54).

**Jayant**, Nuggehally S.; **Noll**, Peter (1984). "Digital Coding of Waveforms, Principles and Applications to Speech and Video". In: Englewood Cliffs NJ, USA: Prentice-Hall, p. 688 (cit. on pp. 38, 39).

**Jeub**, Marco; **Nelke**, Christoph Matthias; **Krüger**, Hauke; **Beaugeant**, Christophe; **Vary**, Peter (2011). "Robust dual-channel noise power spectral density estimation". In: *Signal Processing Conference, 2011 19th European*, pp. 2304–2308 (cit. on p. 4).

**Kennedy**, Rodney A.; **Abhayapala**, Thushara; **Ward**, Darren B.; **Williamson**, Robert C. (1996). "Nearfield broadband frequency invariant beamforming". In: *, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings*. Vol. 2, 905–908 vol. 2. DOI: 10.1109/ICASSP.1996.543268 (cit. on p. 8).

**Kitawaki**, Nobuhiko; **Itoh**, Kenzo (1991). "Pure delay effects on speech quality in telecommunications". In: *IEEE Journal on Selected Areas in Communications* 9.4, pp. 586–593. ISSN: 0733-8716. DOI: 10.1109/49.81952 (cit. on p. 21).

**Kleijn**, W. Bastian; **Paliwal**, Kuldip K. (1995). *Speech Coding and Synthesis*. New York, NY, USA: Elsevier Science Inc. ISBN: 0-444-82169-4 (cit. on p. 73).

**Laakso**, Timo I.; **Välimäki**, Vesa; **Karjalainen**, Matti; **Laine**, Unto K. (1996). "Splitting the unit delay - Tools for fractional delay filter design". In: *IEEE Signal Processing Magazine* 13.1, pp. 30–60. ISSN: 1053-5888. DOI: 10.1109/79.482137 (cit. on p. 138).

**Land**, Ingmar; **Huber**, Johannes B. (2006). *Information combining*. Vol. 1. Now Publishers Inc (cit. on pp. 2, 98).

**Land**, Ingmar; **Huettinger**, Simon; **Hoeher**, Peter A.; **Huber**, Johannes B. (2005). "Bounds on information combining". In: *IEEE Transactions on Information Theory* 51.2, pp. 612–619. ISSN: 0018-9448. DOI: 10.1109/TIT.2004.840883 (cit. on pp. 2, 98).

**Lim**, Jae S.; **Oppenheim**, Alan V. (1979). "Enhancement and bandwidth compression of noisy speech". In: *Proceedings of the IEEE* 67.12, pp. 1586–1604. ISSN: 0018-9219. DOI: 10.1109/PROC.1979.11540 (cit. on pp. 3, 32, 52, 60).

**Linde**, Yoseph; **Buzo**, Andres; **Gray**, Robert M. (1980). "An Algorithm for Vector Quantizer Design". In: *Communications, IEEE Transactions on* 28.1, pp. 84–95. ISSN: 0090-6778. DOI: 10.1109/TCOM.1980.1094577 (cit. on pp. 58, 75).

**Liu**, Qingju; **Aubrey**, Andrew J.; **Wang**, Wenwu (2014). "Interference Reduction in Reverberant Speech Separation With Visual Voice Activity Detection". In: *Multimedia, IEEE Transactions on* 16.6, pp. 1610–1623. DOI: 10.1109/TMM.2014.2322824 (cit. on p. 135).

**Loizou**, Philipos C. (2013). *Speech enhancement: theory and practice*. CRC press (cit. on pp. 3, 23).

**Löllmann**, Heinrich W. (2011). "Allpass-Based Analysis-Synthesis Filter-Banks: Design and Application". In: Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN) 30. Ed. by Peter Vary (cit. on pp. 9, 14, 25).

**Löllmann**, Heinrich W.; **Hildenbrand**, Matthias; **Geiser**, Bernd; **Vary**, Peter (2009). "IIR QMF-Bank Design for Speech and Audio Subband Coding". In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 269–272. ISBN: 978-1-4244-3679-8 (cit. on p. 54).

**Lorenzelli**, Fluvio; **Wang**, Arthur; **Yao**, Kung (1996). "Broadband array processing using subband techniques". In: *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 5. IEEE, pp. 2876–2879 (cit. on p. 7).

**Lotter**, Thomas; **Vary**, Peter (2005). "Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model". In: *EURASIP Journal on Applied Signal Processing* 2005.7, pp. 1110–1126 (cit. on pp. 3, 52, 65).

**Marin-Hurtado**, Jorge I.; **Anderson**, David V. (2011). "FFT-Based Block Processing in Speech Enhancement: Potential Artifacts and Solutions". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.8, pp. 2527–2537. ISSN: 1558-7916. DOI: 10.1109/TASL.2011.2150215 (cit. on p. 23).

**Markovich Golan**, Shmulik; **Gannot**, Sharon; **Cohen**, Israel (2009). "Multichannel Eigenspace Beamforming in a Reverberant Environment with Multiple Interfering Speech Signals". In: *alp* 17.6, pp. 1071–1086 (cit. on p. 7).

**Martin**, Rainer (1994). "Spectral Subtraction Based on Minimum Statistics". In: *Proceedings of European Signal Processing Conference (EUSIPCO)*. Edinburgh, Scotland, Great Britain, pp. 1182–1185 (cit. on p. 28).

**Martin**, Rainer (2001). "Noise power spectral density estimation based on optimal smoothing and minimum statistics". In: *Speech and Audio Processing, IEEE Transactions on* 9.5, pp. 504–512 (cit. on pp. 4, 28, 29, 34, 37, 38, 65).

— (2005). "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors". In: *IEEE Transactions on Speech and Audio Processing* 13.5, pp. 845–856. ISSN: 1063-6676. DOI: 10.1109/TSA.2005.851927 (cit. on p. 3).

— (2006). "Bias compensation methods for minimum statistics noise power spectral density estimation". In: *Signal Processing*. Applied Speech and Audio Processing 86.6, pp. 1215–1229. ISSN: 0165-1684 (cit. on pp. 4, 28, 34, 36–38, 43, 60, 67, 111, 117).

**Martin**, Rainer; **Heute**, Ulrich; **Antweiler**, Christiane (2008). *Advances in Digital Speech Transmission*. Wiley. ISBN: 978-0-470-72717-1 (cit. on p. 76).

**McAulay**, Robert; **Malpass**, Marilyn (1980). "Speech enhancement using a soft-decision noise suppression filter". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 28.2, pp. 137–145. ISSN: 0096-3518. DOI: 10.1109/TASSP.1980.1163394 (cit. on pp. 3, 28, 31, 79).

**Minessale**, Anthony; **Collins**, Michael S.; **Schreiber**, Darren; **Chandler**, Raymond (2013). *FreeSWITCH 1.2*. Packt Publishing. ISBN: 978-1-78216-100-4 (cit. on p. 140).

**Minotto**, Vicente P.; **Jung**, Claudio R.; **Lee**, Bowon (2014). "Simultaneous-Speaker Voice Activity Detection and Localization Using Mid-Fusion of SVM and HMMs". In: *Multimedia, IEEE Transactions on* 16.4, pp. 1032–1044. DOI: 10.1109/TMM.2014.2305632 (cit. on p. 135).

**Murthi**, Manohar N.; **Rao**, Bhaskar D. (2000). "All-pole modeling of speech based on the minimum variance distortionless response spectrum". In: *IEEE Transactions on Speech and Audio Processing* 8.3, pp. 221–239. ISSN: 1063-6676. DOI: 10.1109/89.841206 (cit. on p. 73).

**Nelke**, Christoph Matthias; **Beaugeant**, Christophe; **Vary**, Peter (2013). "Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7279–7283. DOI: 10.1109/ICASSP.2013.6639076 (cit. on p. 4).

**Nelke**, Christoph Matthias; **Naylor**, Patrick A.; **Vary**, Peter (2015). "Corpus Based Reconstruction of Speech Degraded by Wind Noise". In: *Proceedings of European Signal Processing Conference (EUSIPCO)*. Nice, France: EURASIP, pp. 869–873. ISBN: 978-0-9928626-4-0 (cit. on p. 4).

**Nelke**, Christoph Matthias; **Vary**, Peter (2015). "Wind Noise Short Term Power Spectrum Estimation Using Pitch Adaptive Inverse Binary Masks". In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, pp. 5068–5072. DOI: 10.1109/ICASSP.2015.7178936 (cit. on p. 4).

∗ **Niermann**, Markus; **Heese**, Florian; **Vary**, Peter (2015). "Intelligibility Enhancement For Hands-Free Mobile Communication". In: *Proceedings of German Annual Conference on Acoustics (DAGA)*. DEGA, pp. 384–387 (cit. on p. 6).

**Nordholm**, Sven; **Low**, Siow Yong; **Claesson**, Ingvar; **Yiu**, Ka Fai Cedric (2008). "Non-uniform Optimal Subband Beamforming: An Evaluation on Real Acoustic Measurements". In: *Image and Signal Processing, 2008. CISP'08. Congress on*. Vol. 5. IEEE, pp. 747–751 (cit. on p. 7).

**NTT-Corporation** (1994). *Multi-lingual speech database for telephonometry* (cit. on pp. 39, 53, 60).

**Oppenheim**, Alan V.; **Schafer**, Ronald W.; **Buck**, John R, et al. (1989). *Discrete-time signal processing*. Vol. 2. Prentice-hall Englewood Cliffs (cit. on p. 22).

**Paliwal**, Kuldip K.; **Lyons**, James G.; **Wójcicki**, Kamil K. (2010). "Preference for 20-40 ms window duration in speech analysis". In: *2010 4th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–4. DOI: 10.1109/ICSPCS.2010.5709770 (cit. on p. 22).

**Papoulis**, Athanasios; **Pillai**, S. Unnikrishna (2002). *Probability, Random Variables, and Stochastic Processes, Auflage: 4*. ISBN: 978-0-07-112256-6 (cit. on p. 25).

**Proakis**, John G.; **Salehi**, Masoud (2001). *Communication Systems Engineering*. English. 2 edition. Upper Saddle River, N.J: Prentice Hall. ISBN: 978-0-13-061793-4 (cit. on p. 39).

**Quackenbush**, Schuyler R.; **Barnwell**, Thomas Pinkney; **Clements**, Mark A. (1988). *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ (cit. on p. 159).

**Rabiner**, Lawrence R.; **Schafer**, Ronald W. (1978). *Digital processing of speech signals*. Prentice Hall (cit. on pp. 21, 57).

**Rix**, Antony W.; **Beerends**, John G.; **Hollier**, Michael P.; **Hekstra**, Andries P. (2001). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. Vol. 2, pp. 749–752 (cit. on p. 162).

**Rosca**, Justinian; **Balan**, Radu; **Fan**, Ning Ping; **Beaugeant**, Christophe; **Gilg**, Virginie (2002). "Multichannel voice detection in adverse environments". In: *Proceedings of EUSIPCO* (cit. on p. 80).

**Rosenkranz**, Tobias (2010). "Noise codebook adaptation for codebook-based noise reduction". In: *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv* (cit. on pp. 65–67, 76).

— (2012). *Codebook-Based Speech Enhancement - Robust and Efficient Approaches*. München: Dr. Hut. ISBN: 978-3-8439-0363-9 (cit. on p. 76).

**Rosenkranz**, Tobias; **Puder**, Henning (2012**a**). "Improved Gain Estimation for Codebook-Based Speech Enhancement". In: *Speech Communication; 10. ITG Symposium; Proceedings of*, pp. 1–4 (cit. on pp. 65–67).

— (2012**b**). "Improving robustness of codebook-based noise estimation approaches with delta codebooks". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20.4, pp. 1177–1188 (cit. on p. 67).

**Ryan**, James G.; **Goubran**, Rafik A. (2000). "Array optimization applied in the near field of a microphone array". In: *IEEE Transactions on Speech and Audio Processing* 8.2, pp. 173–176. ISSN: 1063-6676. DOI: 10.1109/89.824702 (cit. on p. 8).

∗ **Sauert**, Bastian; **Heese**, Florian; **Vary**, Peter (2014). "Real-Time Near-End Listening Enhancement for Mobile Phones". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Show and Tell Demonstration. IEEE (cit. on pp. 141, 142).

**Sauert**, Bastian; **Vary**, Peter (2010). "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement". In: *ITG-Fachtagung Sprachkommunikation*. Berlin, Germany: VDE Verlag GmbH. ISBN: 978-3-8007-3300-2 (cit. on p. 141).

∗ **Schäfer**, Magnus; **Heese**, Florian; **Wernerus**, Jona; **Vary**, Peter (2012). "Numerical Near Field Optimization of Weighted Delay-and-Sum Microphone Arrays". In: *IWAENC*. IWAENC (cit. on pp. 6, 8, 13–15).

∗ **Schlien**, Thomas; **Heese**, Florian; **Schäfer**, Magnus; **Antweiler**, Christiane; **Vary**, Peter (2013). "Audiosignalverarbeitung für Videokonferenzsysteme". German. In: *WASP*. Vol. Vol. P-220. Lecture Notes in Informatics (LNI) - Proceedings. Gesellschaft für Informatik, pp. 2987–3001. ISBN: 978-3-88579-614-5 (cit. on pp. 6, 135, 136).

**Schroeder**, Manfred R. (1965). *Apparatus for suppressing noise and distortion in communication signals*. US Patent 3,180,936. Google Patents (cit. on p. 3).

**Schuster**, Arthur (1898). "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena". en. In: *Terrestrial Magnetism* 3.1, pp. 13–41. ISSN: 0272-7528. DOI: 10.1029/TM003i001p00013 (cit. on p. 34).

**Shahnaz**, C.; **Zhu**, W. P.; **Ahmad**, M. O. (2005). "Robust Pitch Estimation At Very Low SNR Exploiting Time and Frequency Domain Cues". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*. Vol. 1, pp. 389–392. DOI: 10.1109/ICASSP.2005.1415132 (cit. on p. 66).

**Shankar Chanda**, Pinaki; **Park**, Sungjin (2007). "Speech intelligibility enhancement using tunable equalization filter". In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 4. IEEE, pp. IV–613 (cit. on p. 141).

**Sigg**, Christian D.; **Dikk**, Tomas; **Buhmann**, Joachim M. (2012). "Speech Enhancement Using Generative Dictionary Learning". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.6, pp. 1698–1712. ISSN: 1558-7916. DOI: 10.1109/TASL.2012.2187194 (cit. on p. 65).

**Sohn**, Jongseo; **Kim**, Nam Soo; **Sung**, Wonyong (1999). "A statistical model-based voice activity detection". In: *Signal Processing Letters, IEEE* 6.1, pp. 1–3 (cit. on pp. 80, 87, 90, 92, 96, 138).

**Soong**, Frank K.; **Juang**, Biing-Hwang (1984). "Line spectrum pair (LSP) and speech data compression". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.* Vol. 9, pp. 37–40. DOI: 10.1109/ICASSP.1984.1172448 (cit. on p. 73).

**Sreenivas**, Thippur; **Kirnapure**, Pradeep (1996). "Codebook constrained Wiener filtering for speech enhancement". In: *IEEE Transactions on Speech and Audio Processing* 4.5, pp. 383–389. ISSN: 1063-6676. DOI: 10.1109/89.536932 (cit. on pp. 65, 66).

**Srinivasan**, Sriram; **Samuelsson**, Jonas; **Kleijn**, W. Bastian (2006). "Codebook driven short-term predictor parameter estimation for speech enhancement". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.1, pp. 163–176. ISSN: 1558-7916. DOI: 10.1109/TSA.2005.854113 (cit. on pp. 65–67).

— (2007). "Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.2, pp. 441–452. ISSN: 1558-7916. DOI: 10.1109/TASL.2006.881696 (cit. on pp. 65–67).

**Strobel**, Norbert; **Spors**, Sascha; **Rabenstein**, Rudolf (2001). "Joint Audio-Video Object Localization and Tracking". In: *Signal Processing Magazine, IEEE* 18.1, pp. 22–31 (cit. on p. 135).

**Taghia**, Jalal; **Taghia**, Jalil; **Mohammadiha**, Nasser; **Sang**, Jinqiu; **Bouse**, Vaclav; **Martin**, Rainer (2011). "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643. DOI: 10.1109/ICASSP.2011.5947389 (cit. on p. 28).

**Taghizadeh**, Mohammad J.; **Garner**, Philip N.; **Bourlard**, Hervé; **Abutalebi**, Hamid R.; **Asaei**, Afsaneh (2011). "An integrated framework for multi-channel multi-source localization and voice activity detection". In: *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 92–97 (cit. on p. 80).

**Talmon**, Ronen; **Cohen**, Israel; **Gannot**, Sharon (2013). "Single-Channel Transient Interference Suppression With Diffusion Maps". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.1, pp. 132–144. ISSN: 1558-7916. DOI: 10.1109/TASL.2012.2215593 (cit. on pp. 4, 19).

**Tan**, Lee Ngee; **Borgstrom**, Bengt J.; **Alwan**, Abeer (2010). "Voice activity detection using harmonic frequency components in likelihood ratio test". In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4466–4469. DOI: 10.1109/ICASSP.2010.5495611 (cit. on pp. 80, 87, 90, 92, 96).

**Vähätalo**, Antti; **Johansson**, Ingemar (1999). "Voice activity detection for GSM adaptive multi-rate codec". In: *IEEE Workshop on Speech Coding Proceedings*, pp. 55–57 (cit. on p. 80).

**Van Compernolle**, Dirk (1989). "Noise adaptation in a hidden Markov model speech recognition system". In: *Computer Speech & Language* 3.2, pp. 151–167 (cit. on pp. 28, 79).

**Van Veen**, Barry D.; **Buckley**, Kevin M. (1988). "Beamforming: A versatile approach to spatial filtering". In: *IEEE ASSP Magazine* 5.2, pp. 4–24 (cit. on p. 7).

**Varga**, Andrew; **Steeneken**, Herman J. M.; **Jones**, D (1992). "The noisex-92 study on the effect of additive noise on automatic speech recognition system". In: *Reports of NATO Research Study Group (RSG. 10)* (cit. on pp. 39, 44, 53, 61).

**Vary**, Peter (1985). "Noise suppression by spectral magnitude estimation —mechanism and theoretical limits—". In: *Signal Processing* 8.4, pp. 387–400. ISSN: 0165-1684. DOI: 10.1016/0165-1684(85)90002-7 (cit. on pp. 3, 32).

**Vary**, Peter; **Heute**, Ulrich; **Hess**, Wolfgang (1998). *Digitale Sprachsignalverarbeitung*. German. 1st ed. Stuttgart, Germany: Teubner Verlag (cit. on p. 3).

**Vary**, Peter; **Martin**, Rainer (2006). *Digital Speech Transmission - Enhancement, Coding & Error Concealment*. Chichester, UK: John Wiley & Sons, Ltd. ISBN: 978-0-471-56018-0 (cit. on pp. 3, 19, 22, 28, 75, 76, 79, 87).

**Vaseghi**, Saeed V. (1996). *Advanced Signal Processing and Digital Noise Reduction*. Deutsch. 1996th ed. Chichester ; New York: Vieweg+Teubner Verlag. ISBN: 978-3-519-06451-0 (cit. on p. 32).

**Veth**, Johan de; **Boves**, Louis (1998). "Channel normalization techniques for automatic speech recognition over the telephone". In: *Speech Communication* 25.1–3, pp. 149–164. ISSN: 0167-6393. DOI: 10.1016/S0167-6393(98)00034-X (cit. on p. 67).

**Wang**, David L.; **Lim**, Jae S. (1982). "The unimportance of phase in speech enhancement". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 30.4, pp. 679–681. ISSN: 0096-3518. DOI: 10.1109/TASSP.1982.1163920 (cit. on p. 32).

**Ward**, Darren B.; **Kennedy**, Rodney A.; **Williamson**, Robert C. (1995). "Theory and Design of Broadband Sensor Arrays with Frequency Invariant Far-field Beam Patterns". In: *The Journal of the Acoustical Society of America* 97, p. 1023 (cit. on p. 8).

**Wei**, Bo; **Gibson**, Jerry D. (2000). "Comparison of distance measures in discrete spectral modeling". In: *Proceedings of IEEE Digital Signal Processing Workshop 2000* (cit. on p. 72).

**Westphal**, Martin (1997). "The Use Of Cepstral Means In Conversational Speech Recognition". In: *In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech*, pp. 1143–1146 (cit. on p. 67).

**Wiener**, Norbert (1949). *Extrapolation, interpolation, and smoothing of stationary time series*. Vol. 2. MIT press Cambridge, MA (cit. on p. 3).

**Yoshioka**, Takuya; **Nakatani**, Tomohiro; **Okuno**, Hiroshi G. (2010). "Noisy speech enhancement based on prior knowledge about spectral envelope and harmonic structure". In: *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 4270–4273. DOI: 10.1109/ICASSP.2010.5495681 (cit. on p. 66).

**Zhang**, Cha; **Yin**, Pei; **Rui**, Yong; **Cutler**, R.; **Viola**, P.; **Sun**, Xinding; **Pinto**, N.; **Zhang**, Zhengyou (2008). "Boosting-Based Multimodal Speaker Detection for Distributed Meeting Videos". In: *Multimedia, IEEE Transactions on* 10.8. bibtex: ZhYiRuCu08, pp. 1541–1552. ISSN: 1520-9210. DOI: 10.1109/TMM.2008.2007344 (cit. on p. 135).

**Zhao**, Yong; **Liu**, Wei; **Langley**, Richard J. (2011). "Subband design of fixed wideband beamformers based on the least squares approach". In: *Signal Processing* 91.4, pp. 1060–1065 (cit. on p. 7).

**Zhou**, Huiyu; **Taj**, Murtaza; **Cavallaro**, Andrea (2008). "Target Detection and Tracking With Heterogeneous Sensors". In: *IEEE J. Select. Topics Signal Process.* 2.4, pp. 503–513. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2008.2001429 (cit. on p. 135).

**Zwicker**, Eberhard (1982). *Psychoakustik*. 1st ed. Springer Berlin Heidelberg. ISBN: 3-540-11401-7 (cit. on p. 55).

**Zwicker**, Eberhard; **Fastl**, Hugo (1990). *Psychoacoustics: Facts and Models*. Englisch. Berlin ; New York: Springer. ISBN: 978-3-540-52600-1 (cit. on p. 21).