Wind Noise Reduction
– Signal Processing Concepts –

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

Christoph Matthias Nelke

aus Aachen

Berichter: Universitätsprofessor Dr.-Ing. Peter Vary
Reader Dr. Patrick Naylor

Tag der mündlichen Prüfung: 10. Mai 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 43)
ISSN 1437-6768
ISBN 978-3-958896-107-7

Wissenschaftsverlag Mainz
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34
Fax: 02 41 / 87 55 77
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.

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 creating such an inspiring and pleasant working environment. I am also indebted to Dr. Patrick Naylor for being my co-supervisor and for the fruitful discussions we had during my research visit at the Imperial College London.

I would like to take the opportunity to thank all my current and former colleagues as well as the permanent staff at the institute their scientific support and the enjoyable working environment. I want to thank Dr.-Ing. Christiane Antweiler, Dr.-Ing. Hauke Krüger, Dipl.-Ing. Florian Heese, Dr.-Ing. Marco Jeub, Dr.-Ing. Bastian Sauert, Dipl.-Ing. Thomas Schlien, Dr.-Ing. Matthias Rüngeler, Dr.-Ing. Magnus Schäfer, Dipl.-Ing. Daniel Haupt, M.Sc. Stefan Liebich, M.Sc. Markus Niermann, M.Sc. Tim Schmitz, and Max Mascheraux for the intense scientific and technical discussions and support and proof-reading this work. Of course, I would like to express my gratitude to the students who supported this work, in particular, Dipl.-Ing. Christian Herglotz, M.Sc. Niklas Nawroth, M.Sc. Maximilian Kentgens, Dipl.-Ing. Christopher Bulla, and M.Sc. Tobias Moll and the Speech and Audio Processing Group at the Imperial College London for making my visit so pleasant.

This work was accompanied by projects with Intel Mobile Communication in Sophia Antipolis, France and Allentown, Pennsylvania, USA. Special thanks to Dr. Christophe Beaugeant and Dr. Navin Chatlani for the good and friendly collaboration.

Finally, I want to thank my friends and family, especially my parents Kerstin and Peter, and my sister Annika for their support over the years. Dear Eva thank you for the encouragement and the support, especially during the last months of this work.

Aachen, May 2016 *Christoph Nelke*

Abstract

With the technological progress, devices, such as mobile phones, tablet computers or hearing aids, can be used in a large variety of every-day situations for mobile communication. Acoustic background noise signals, which are picked up with the desired speech signal, can impair the signal quality and the intelligibility of a conversation. A special noise type is generated outdoors, if the microphone is exposed to a wind stream resulting in strong-rumbling noise, which is highly non-stationary. As a result, conventional approaches for noise reduction fail in the case of noise induced by wind turbulences.

This thesis is focused on the development of signal processing concepts, which reduce the undesired effects of wind noise. The key contributions are:

- Signal analysis of wind noise
- Digital signal model for wind noise generation
- Signal processing algorithms for detection and reduction of wind noise signals.

All these topics are considered with the focus on the development of algorithms for single and dual microphone systems.

The analysis of recorded wind signals is the first step and gives valuable information for the estimation and reduction of wind noise. Furthermore it leads to a signal model for the generation of reproducible artificial wind noise signals.

For the enhancement of the disturbed speech, an estimate of the underlying wind noise signal is required. In contrast to state-of-the-art noise estimation algorithms, the spectral shape and energy distribution is exploited for the distinction between speech and wind noise components leading to a novel estimation scheme of the wind noise short-term power spectrum. Considering a system with two microphone inputs, the complex coherence function of the two recorded signals is exploited for wind noise estimation. In addition to commonly used noise reduction schemes by spectral weighting, an innovative concept for speech enhancement is developed by using techniques known from artificial bandwidth extension. Highly disturbed speech parts are replaced by corresponding parts from an artificial speech signal.

Objective measures indicate a significant increase of both the signal-to-noise ratio and the speech intelligibility. Besides, two application examples show that the proposed methods are very efficient and robust in realistic scenarios.

Contents

Chapter 1

Introduction

Today, a world without mobile communication is inconceivable: everyone is reachable almost everywhere. With a nearly complete network coverage of mobile telephony services, it is possible to make phone calls in almost any environment. This provides many advantages but also leads to technical challenges to guarantee high speech quality for all use cases. 25 years ago, when mostly traditional fixed-line networks were in use, phone calls usually took place inside rooms, where only low acoustic disturbance could be expected. With the possibility of making a phone call outdoors, many noise sources picked up by the microphone impair the signal quality. These noise signals can severely degrade both the speech quality and intelligibility at the far-end side.

Besides mobile telephony, the appearance of noise can be an even more significant problem for hearing-impaired listeners using a hearing aid, which also amplifies the noise signals. This is not only annoying but can also make a conversation impossible, due to a reduced intelligibility.

The perceived noise can be caused by many acoustic scenarios, such as the sound inside a car, or close to a highly frequented street, or the voices of surrounding talkers in a crowd. A special noise type occurs in windy environments, when the air stream generates a highly non-stationary disturbance in the recorded signal. While for the reduction of stationary background noise signals many well established methods can be found in the literature, the suppression of fast varying wind noise signals is still an open issue. This thesis provides a first complete overview on the analysis, generation and reduction of wind noise from a digital signal processing perspective.

1.1 Relation to Prior Work

In the past decades, many approaches have been published dealing with the problem of reducing the undesired noise components within a speech signal. Early concepts for noise reduction can be found in [Wie57], [LO79], [Bol79] or [MM80]. They all rely on a spectral subtraction or a Wiener filter solution. The required estimate of the noise power spectral density (PSD) is given either by known statistics about the noise signal or by averaging the signal power in segments of speech absence, e.g., in speech pauses or at the beginning of the recorded signal. These techniques assume stationary noise signal characteristics and mostly rely on a voice activity detector (VAD).

Since a VAD is erroneous in the presence of noise and the scenario of stationary noise signals is not always given, the techniques of noise PSD estimation were refined. The first algorithms apply a minima tracking in each frequency band independently. The most prominent methods in this field are *Minimum Statistics* by Martin [Mar01] and the minima controlled recursive averaging by Cohen [Coh03]. Further improvements in terms of estimating time-varying noise were developed by Hendriks and Gerkmann [HHJ10], [GH11], and Heese [HV15].

All the aforementioned methods have been developed for the estimation and reduction of background noise in general and show reasonable results in cases for stationary or only slowly varying noise signals with a signal-to-noise ratio above 0 dB. Because these methods can not guarantee a sufficient noise estimation for all scenarios, several algorithms can be found dealing with special noise types as:

- keyboard noise [SSA07], [GBS15],
- harmonic car engine noise [CCS+09] [ERHV10],
- multi-talker babble noise [ML13],
- car horn noise [CBK15].

With the increasing computational power of digital signal processors (DSP), more and more of these specialized algorithms can be integrated into communication devices.

Wind noise reduction belongs into this class of algorithms dealing with high non-stationary noise signals. If a mobile communication device is used outdoors in a windy environment, the air stream of the wind meets an obstacle, e.g., the housing of a mobile phone, and turbulences are generated. The turbulent air flow close to a microphone leads to annoying, low frequency, rumbling artifacts in the recorded signal. In many applications the small dimensions and design constraints of the devices do not allow the usage of a wind shield. Thus, it is necessary to reduce the wind noise by means of digital signal processing. Due to its high level of non-stationarity, conventional noise estimation methods fail at this point. Although wind noise is a common problem outdoors, only a few contributions can be found, the most important examples are [KMT+06], [Kat07], [HWB+12], [Elk07], and [FB10]. Kuroiwa et al. proposed in [KMT+06] to store wind-templates and estimate the rough spectral shape by a comparison of the stored samples with the observed noisy signal. A simple high-pass filter approach based on a wind detection was presented by Kates in [Kat07]. Hofmann et al. developed an algorithm, which identifies wind presence by applying techniques from image processing on the observed noisy spectrum for the detection of connected areas [HWB+12]. Dual microphone concepts were derived by Elko in [Elk07] as well as Franz and Bitzer in [FB10], where both algorithms exploit the low spatial correlation between wind noise signals recorded at different microphone positions.

1.2 Structure of this Thesis

In Chapter 2 the general problem of noise reduction is depicted. The underlying signal model is presented for the single and multi microphone case. The structure of noise reduction realized in the discrete Fourier transform (DFT) domain is explained, introducing the overlap-add framework for speech enhancement. The procedure of background noise PSD estimation is exemplary demonstrated by the speech presence probability (SPP) based method [GH11]. Furthermore, the most common approaches of noise suppression are presented, which are variants of spectral subtraction by means of a spectral weighting and the Wiener filter realized in the frequency domain. These state-of-the-art techniques are the starting point for the following research on wind noise suppression.

Chapter 3 deals with the analysis of wind noise signals, which is fundamental for the subsequent estimation and reduction concepts. In a first step, the generation of the acoustic signal is described, which becomes audible in the presence of wind close to the microphone. Then the characteristics in a digital signal representation are considered in the time- and frequency-domain. Based on the derived specific features, several approaches for the detection of wind noise in a noisy speech signal are discussed and compared. These detection methods are a key element of the wind noise reduction systems in this work. Finally, a signal model for the generation of wind noise is proposed including an auto-regressive (AR) process for the spectral characteristics and a Markov-chain for the temporal characteristics. This model plays an important role during the development and the reproducible evaluation of wind noise reduction algorithms within this thesis.

The wind noise reduction task is addressed in Chapter 4, which is the main part of this work. Novel solutions for the suppression of wind noise and the enhancement of the desired speech signal are presented. For a single microphone system, two state-of-the-art methods ([KMT^{+}06], [HWB^{+}12]) are considered as reference for the wind noise estimation. Since these algorithms can not always guarantee a good wind noise suppression, new concepts for wind noise estimation are developed. The innovative approach of the two proposed techniques is that they exploit the different spectral energy distribution of speech and wind noise.

The system of wind noise reduction is also extended to configurations with two microphones. Here, a solution is developed, with a coherence based wind noise estimator. Especially, the use of the phase of the complex coherence leads to good noise reduction performance. All algorithms are evaluated in competitive studies with real wind noise recordings using different spectral gain calculation realizations.

A further priority is the development of a new concept for speech enhancement, which is in general independent of the microphone configuration. In contrast to the conventional spectral weighting, the alternative approach reconstructs highly disturbed parts of the speech with an artificial signal, applying the source-filter model for speech production.

In Chapter 5 two application examples for speech enhancement in a mobile phone are presented. In addition to the more theoretic algorithmic concepts, also typical problems that arises from practice have to be considered for a balanced system design. The issue of combined wind noise and background noise reduction for the application of a single microphone system is discussed and a solution for a suppression of both disturbances is proposed. As the task of speech enhancement is always accompanied with the aspect of background noise estimation, solutions are developed for a dual microphone mobile phone. Here, the focus was to bypass the limitation of coherence based estimators for diffuse background noise in realistic environments.

Parts of the results of this thesis have been pre-published in the following references: [JSK^{+}10, JNK^{+}11, HJN^{+}11, JNBV11, NNJ^{+}12, JHN^{+}12, JNH^{+}13, NBV13, HNNV14, NCBV14, NV14a, NV14b, NCBV15, NV15, NNV15, NJV16].

These references are marked by an underlined label, i.e., [____], throughout the thesis.

Chapter 2

Noise Reduction Techniques

Many approaches for enhancing a speech signal, which is degraded by noise, can be found in the literature of the last three decades. Different realizations were proposed depending on the available number of microphones, the noise type, the application, and the source signal. Further variations are possible regarding the internal structure of the algorithm. Throughout this thesis, all considerations target at a real-time processing of the recorded signals, e.g., hearing aid application or in mobile phones as exemplified in Figure 2.1. With this constraint, only causal modifications of the signals are possible, i.e., signal properties at the current point in time and from the past can be taken into account but no information from future segments is available. Besides, the signal is processed in short time segments, since most of the considered signals are only stationary within this short duration (see, e.g., [VM06]). This short-term stationarity is necessary, because the noisy input signal is modified in a constant manner during one segment, e.g., by filtering with a fixed but arbitrary set of coefficients.

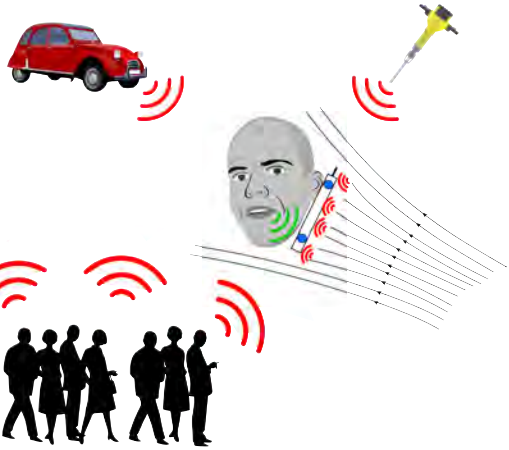


Figure 2.1: Wind and background noise scenario for a mobile phone with two microphones.

In this chapter, the structure of a noise reduction system is described. The aim is to highlight the aspects, which are important for suppressing background noise in a conventional structure. These are namely the analysis-synthesis framework (Section 2.2), the estimation procedure of the noise power spectral density (PSD) (Section 2.3.1) and the signal-to-noise-ratio (SNR) (Section 2.3.2), and the calculation of the spectral gain function (Section 2.3.3). The last section of this chapter gives some insights in the performance of conventional background noise reduction approaches in the case of wind noise.

2.1 Problem Statement

The general problem of recorded signals in the presence of noise is depicted in Figure 2.1 for the scenario of a mobile phone equipped with two microphones (marked in blue). The microphones of the device pick up not only the desired speech signal $s(k)$ (green) but also a superposition with different noise signals $n_j(k)$ generated somewhere in the surrounding (red). The signal is digitized and fed to a digital signal processor (DSP), where it is possible to apply modifications. The recorded noisy signals of the two microphones are given by

$$x(k) = h_1(k) * s(k) + \sum_j n_{1,j}(k) \tag{2.1}$$

$$y(k) = h_2(k) * s(k) + \sum_j n_{2,j}(k), \tag{2.2}$$

where k is the discrete time index and the index j represents the noise sources. The convolution operation $*$ models the impulse responses $h_{1,2}(k)$ from the speech source to the microphones. Their influence is mainly the reverberation due to the room acoustics, which can also impair the speech quality [JSK$^+$10]. However, this problem is not a focus of this work, corresponding approaches can be found, e.g., in [Jeu12] or [NG10]. The general aim is to obtain a good estimate $\hat{s}(k)$ of the clean speech signal and to transmit an enhanced signal to the far-end speaker.

In this work different representations are used for the description of signals in the frequency-domain. Considering an analog signal $x(t)$ over the continuous time t, the Fourier transform (FT) reads

$$X(f) = \int\limits_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt. \tag{2.3}$$

For a digital signal $x(k)$ either the Fourier transform of discrete signals (FTDS) with the continuous *normalized radiant frequency* $\Omega = 2\pi f / f_\mathrm{s}$

$$X(\Omega) = \sum_{k=-\infty}^{\infty} x(k) e^{-j\Omega k} \tag{2.4}$$

or the discrete Fourier transform (DFT) over an finite number of M signal samples

$$X(\lambda,\, \mu) = \sum_{\kappa=0}^{M-1} x_\lambda(\kappa) e^{-j\frac{2\pi\mu\kappa}{M}} \text{ , with } \mu = 0, \ldots, M-1. \tag{2.5}$$

with the discrete frequency bin μ and the sample position κ in the frame λ. The relevant representation is apparent from the context of the used signals in this work.

2.2 Speech Processing System

All approaches considered in this work can be described by the structure shown in Figure 2.2. Single or multi microphone inputs (using K microphones) are considered, which cover many applications. The latest generation of mobile phones are equipped with two or more microphones and hearing aids exploit the advantage of using more than one microphone, too. As initially mentioned, short-term processing is applied by segmenting the input signals into frames, which may overlap. Typical values for speech processing are a frame size of 10-30 ms and an overlap of half the frame-size (see, e.g., [Loi13]). If not otherwise stated a frame size of 20 ms is applied in this work.

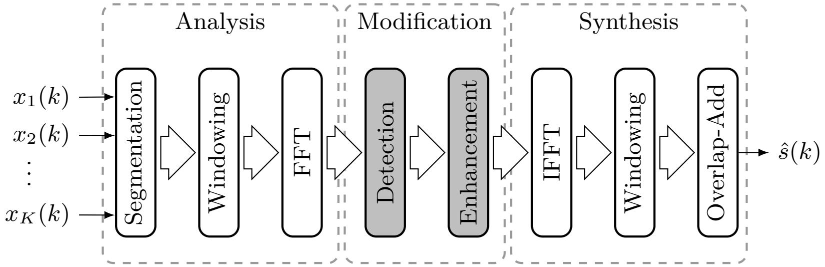


Figure 2.2: Structure of speech processing systems for noise reduction.

After the segmentation, the frames of length L_F samples are multiplied with a window function, in order to counteract the *spectral leakage effect* [VM06]. Frequently used window functions are, e.g., the Hann window, the Hamming window or the Blackmann window [OSB$^+$89]. Because the window function is applied twice (in the analysis as well as in the synthesis stage), a square-root Hann window of length L_F is used in this work as

$$w(k) = \sqrt{\frac{1}{2}\left(1 - \cos\left(\frac{2\pi k}{L_\mathrm{F}-1}\right)\right)}, \tag{2.6}$$

with

$$k = 0 \ldots L_\mathrm{F} - 1, \tag{2.7}$$

which multiplies after analysis and synthesis to a conventional Hann window. Applying the window twice in the analysis and synthesis, lowers also the negative effects of changing spectral modifications by interpolation of the overlapping parts of successive frames [MHA11]. These changing modifications are necessary for the noise reduction task for non-stationary signals. Besides, the prerequisite is fulfilled that with an overlap of half of the frame-size the windows of successive frames add up to one. This behavior is depicted in Figure 2.3 by the dashed line, where the frame index is denoted by λ.

After windowing, the frames are transformed into the frequency-domain by a discrete Fourier transform (DFT)[1] of size M. The corresponding short-term Fourier spectrum of a signal $x(k)$ in frame λ is given by

$$X(\lambda,\mu) = \mathrm{DFT}_M\{x_\lambda(\kappa)\} = \mathrm{DFT}_M\{w(\kappa)\cdot x(\lambda\cdot L_\mathrm{F}/2+\kappa)\}, \qquad (2.8)$$
$$\text{with } \kappa = 0,\dots,L_\mathrm{F}-1 \text{ and } \mu = 0,\dots,M-1,$$

where μ is the discrete frequency bin and κ is the sample position within one signal frame. The subscript M indicates the length of the DFT, where zero-padding of $M - L_\mathrm{F}$ samples is applied if $M > L_\mathrm{F}$.

The noise reduction is applied in the frequency-domain and can roughly be separated into the two stages of detection and enhancement. The detection may comprise the identification of noise and speech in the input signal and also the measurement of the degree of degradation, e.g., given by the spectral SNR. Based on the results of the detection stage the enhancement is applied. Different realizations will be considered for these stages and will be discussed in Chapter 4. After these modifications, the signal frames must be reconstructed resulting in a time-domain signal $\hat{s}(k)$. This is realized by first applying an inverse fast Fourier transform (IFFT) and again a windowing. In a last step, the time-domain frames are added

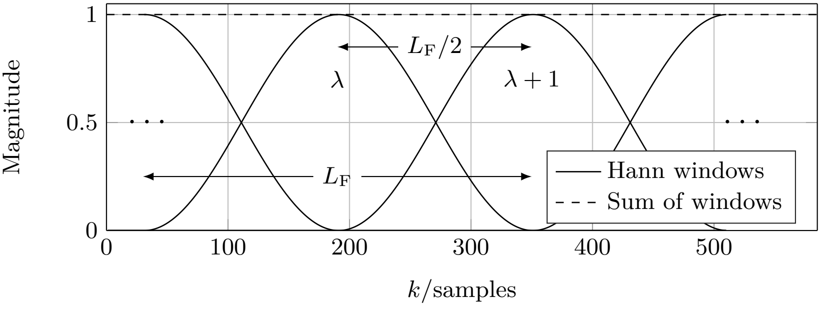


Figure 2.3: Sequence of Hann window functions with $L_\mathrm{F} = 320$ samples and an overlap of $L_\mathrm{F}/2$ for two frames λ and $\lambda + 1$.

[1] The fast Fourier transform (FFT) is used throughout this work as an efficient implementation of the DFT.

with the same overlap as in the analysis stage. The IFFT, the second windowing procedure, and the overlap-add step are widely known as the synthesis stage of the described structure.

Both the analysis and the synthesis stage are not subject of this work and are mostly used in the implementation described above. Different implementations can be found for the analysis-synthesis framework, e.g., by a filter-bank structure (see [Löl11] and references therein). The focus of this work are the two highlighted modification blocks in Figure 2.2, i.e., the detection of wind noise and the enhancement of the degraded speech signal.

2.3 Conventional Noise Reduction

Most state-of-the-art noise reduction systems for background noise reduction are realized in a framework as described in the previous section. A scalable solution for one or two microphone input signals is shown in Figure 2.4. A common way to suppress noise is given by first estimating[2] the short-term PSD of the noise $\widehat{\Phi}_{nn}(\lambda,\, \mu)$ and subsequently applying a spectral weighting. Usually, the weighting gains are computed based on the noise PSD estimate and optionally an estimate of the current $\mathrm{SNR}(\lambda, \mu)$ given by the *a priori* SNR $\widehat{\xi}(\lambda, \mu)$ or the *a posteriori*

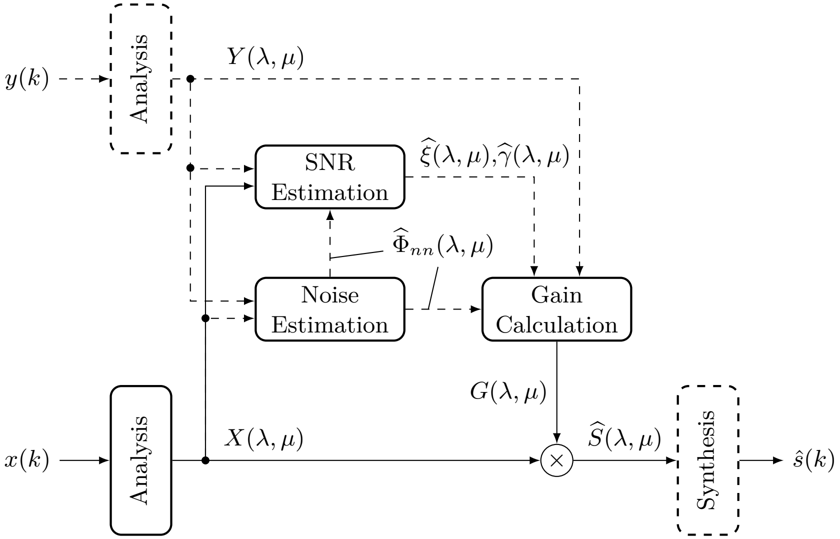


Figure 2.4: Scalable noise reduction system working in the short-term Fourier-domain (dashed lines correspond to the optional second microphone signal).

[2] In this thesis, the $\widehat{\ }$ symbol depicts the estimate of a signal or parameter.

SNR $\widehat{\gamma}(\lambda, \mu)$. Multiplying the noisy input spectrum $X(\lambda, \mu)$ with the spectral gain $G(\lambda, \mu)$ results into an estimate $\widehat{S}(\lambda, \mu)$ of the clean speech spectrum. The synthesis stage produces the corresponding time-domain representation $\hat{s}(k)$ as output of the noise reduction system. If a second microphone signal $y(k)$ is available, the SNR and noise estimation as well as the spectral gain calculation can exploit information from this signal. For both microphone signals, it is assumed that the desired speech signal $S(\lambda,\ \mu)$ and the noise signals $N_{1,2}(\lambda,\ \mu)$ superpose to the input signals, as defined in Equations 2.1 and 2.2. Then the following short-term frequency-domain model is used

$$X(\lambda,\ \mu) = S_1(\lambda,\ \mu) + N_1(\lambda,\ \mu), \tag{2.9}$$

$$Y(\lambda,\ \mu) = S_2(\lambda,\ \mu) + N_2(\lambda,\ \mu), \tag{2.10}$$

where the spectra $S_1(\lambda,\ \mu)$ and $S_2(\lambda,\ \mu)$ are the short-term frequency-domain representations of the filtered speech components

$$s_1(k) = h_1(k) * s(k), \tag{2.11}$$

$$s_2(k) = h_2(k) * s(k). \tag{2.12}$$

2.3.1 Noise PSD Estimation

Several algorithms were proposed in the past for the estimation of the noise PSD in speech signals. Usually, they are based on the assumption that the desired speech signal and the unwanted noise signal can be separated by their temporal statistics. A simple way to estimate the noise PSD is given by a voice activity detector (VAD). The noise PSD can be updated in speech pauses using a first-order recursive smoothing with $0 < \alpha < 1$,

$$\widehat{\Phi}_{nn}(\lambda, \mu) = \alpha \cdot \widehat{\Phi}_{nn}(\lambda - 1, \mu) + (1 - \alpha) \cdot |X(\lambda, \mu)|^2 \tag{2.13}$$

assuming that the input $X(\lambda,\ \mu)$ only contains noise, and kept constant during speech activity ($\alpha = 1$) [VM06]. In the last years more sophisticated approaches were proposed. Most prominent examples are *Minimum Statistics* by Martin [Mar01], the MMSE Noise PSD Tracker by Hendriks e.a. [HHJ10] and the approach based on the speech presence probability (SPP) proposed by Gerkmann and Hendriks [GH11]. Investigating the capability of estimating the PSD of time-varying noise signals, the SPP based method showed the highest accuracy (see results in [GH11]). Because wind noise is characterized by a high level of non-stationarity, this method will used in the following as state-of-the-art method for conventional background noise estimation.

SPP Based Noise Estimation

The aforementioned VAD yields a hard decision for a given signal segment, if speech is present or not. In contrast to that, the speech presence probability (SPP) measure is a time and frequency dependent value between zero and one for the speech activity. For a Gaussian distribution of the real and imaginary parts of speech and noise spectral coefficients, a mathematical expression can be derived for the SPP. Using Bayes' theorem, the probability p of speech presence $\mathcal{H}_1$ [CB01], given a noisy spectrum observation $X(\lambda, \mu)$ and a noise PSD estimate $\widehat{\Phi}_{nn}(\lambda, \mu)$ can be expressed as[3]

$$p(\mathcal{H}_1|X(\lambda,\mu)) = \left(1 + (1+\xi_{\text{opt}}) \exp\left(-\frac{|X(\lambda,\mu)|^2}{\widehat{\Phi}_{nn}(\lambda,\mu)} \frac{\xi_{\text{opt}}}{\xi_{\text{opt}}+1}\right)\right)^{-1}. \tag{2.14}$$

Furthermore, it was assumed in Equation 2.14 that the absence of speech $\mathcal{H}_0$ and the presence of speech $\mathcal{H}_1$ are equally probable, i.e.,

$$p(\mathcal{H}_0) = p(\mathcal{H}_1) = 0.5. \tag{2.15}$$

A post-processing of $p(\mathcal{H}_1|X(\lambda, \mu))$ is applied to avoid a stagnation at high values close to one in terms of a recursive smoothing and an upper limit of the smoothed SPP. In [GH11] it was proposed that the SPP measure can be used as a soft VAD to control the update of the noise periodogram estimate as follows

$$|\widehat{N}(\lambda,\, \mu)|^2 = p(\mathcal{H}_0|X(\lambda,\, \mu)) \cdot |X(\lambda,\, \mu)|^2 + p(\mathcal{H}_1|X(\lambda,\, \mu)) \cdot \widehat{\Phi}_{nn}(\lambda,\, \mu) \tag{2.16}$$

with the probability of speech absence

$$p(\mathcal{H}_0|X(\lambda,\, \mu)) = 1 - p(\mathcal{H}_1|X(\lambda,\, \mu)). \tag{2.17}$$

It must be noted, that the noise PSD estimate from the previous frame $\widehat{\Phi}_{nn}(\lambda-1, \mu)$ is used in Equation 2.14 to compute the SPP value. Finally, recursive smoothing of the periodogram results in the short-term estimate of the noise PSD

$$\widehat{\Phi}_{nn}(\lambda,\, \mu) = 0.8 \cdot \widehat{\Phi}_{nn}(\lambda - 1, \mu) + 0.2 \cdot |\hat{N}(\lambda,\, \mu)|^2. \tag{2.18}$$

Simulations carried out in [GH11] showed that this approach for noise PSD estimation is capable of tracking noise signals even in the case of at least slowly time-varying noise.

2.3.2 Signal-to-Noise-Ratio Estimation

Many algorithms for the gain calculation require an estimate of the signal-to-noise-ratio. Namely these are the *a priori* SNR ξ and the *a posteriori* SNR γ and their estimates are defined by [MM80]

$$\widehat{\xi}(\lambda,\, \mu) = \frac{\widehat{\Phi}_s(\lambda,\, \mu)}{\widehat{\Phi}_n(\lambda,\, \mu)} = \frac{\widehat{\mathrm{E}}\{|S(\lambda,\, \mu)|^2\}}{\widehat{\mathrm{E}}\{|N(\lambda,\, \mu)|^2\}} \tag{2.19}$$

[3] According to [GH11], the fixed optimal *a priori* SNR ξ_{opt} should be chosen to $10\log_{10}(\xi_{\text{opt}}) = 15\,\text{dB}$, if the true *a priori* SNR lies between $-\infty$ and 20 dB.

and

$$\widehat{\gamma}(\lambda,\, \mu) = \frac{|X(\lambda,\, \mu)|^2}{\widehat{\Phi}_n(\lambda,\, \mu)} = \frac{|X(\lambda,\, \mu)|^2}{\widehat{\mathrm{E}}\{|N(\lambda,\, \mu)|^2\}}, \tag{2.20}$$

where $\widehat{\mathrm{E}}\{\;\}$ represents the short-term average of its argument. For a given estimate of the noise PSD estimate $\widehat{\Phi}_{nn}$, the *a posteriori* SNR is easily measurable and the *a priori* SNR can be expressed as

$$\widehat{\xi}(\lambda,\, \mu) = \frac{\widehat{\Phi}_s(\lambda,\, \mu)}{\widehat{\Phi}_{nn}(\lambda,\, \mu)} = \frac{|X(\lambda,\, \mu)|^2}{\widehat{\Phi}_{nn}(\lambda,\, \mu)} - 1 = \widehat{\gamma}(\lambda,\, \mu) - 1. \tag{2.21}$$

Again it is assumed that speech and noise are uncorrelated leading to the cross PSD estimates

$$\widehat{\Phi}_{sn}(\lambda,\, \mu) = \widehat{\Phi}_{ns}(\lambda,\, \mu) = 0. \tag{2.22}$$

2.3.3 Spectral Gain Calculation

For the suppression of the unwanted noise in the input signal, the noisy spectrum $X(\lambda,\, \mu)$ is multiplied with the spectral gain $G(\lambda,\, \mu)$ (see Figure 2.4). The resulting estimate of the clean speech DFT coefficients are given by

$$\widehat{S}(\lambda,\, \mu) = G(\lambda,\, \mu) \cdot X(\lambda,\, \mu) = G(\lambda,\, \mu) \cdot R(\lambda,\, \mu) e^{\mathrm{j}\eta(\lambda,\, \mu)}, \tag{2.23}$$

where $R(\lambda, \mu)$ and $\eta(\lambda, \mu)$ are the magnitude and phase of the noisy signal $X(\lambda,\, \mu)$. Usually, the spectral gain $G(\lambda,\, \mu)$ is real-valued in the range between zero and one. Consequently, only the magnitudes of the noisy DFT coefficients are modified. The influence of the phase can be neglected in most of the cases because the human auditory system is rather insensitive w.r.t. phase distortions [WL82].

A widely used rule for the calculation of the spectral gains is represented by the Wiener filter $G_{\mathrm{W}}(\lambda,\, \mu)$ [LO79], which minimizes the mean square error

$$\widehat{\mathrm{E}}\{|S(\lambda,\, \mu) - \widehat{S}(\lambda,\, \mu)|^2\} = \widehat{\mathrm{E}}\{|S(\lambda,\, \mu) - G_{\mathrm{W}}(\lambda,\, \mu)(S(\lambda,\, \mu) + N(\lambda,\, \mu))|^2\} \tag{2.24}$$

between the clean speech and its estimate independently for each frequency bin μ. By partial derivation to the real and imaginary part of $G_{\mathrm{W}}(\lambda,\, \mu)$ it can be shown that

$$\mathrm{Im}\{G_{\mathrm{W}}(\lambda,\, \mu)\} = 0 \tag{2.25}$$

and

$$\mathrm{Re}\{G_{\mathrm{W}}(\lambda,\, \mu)\} = \frac{\widehat{\mathrm{E}}\{|S(\lambda,\, \mu)|^2\}}{\widehat{\mathrm{E}}\{|S(\lambda,\, \mu)|^2\} + \widehat{\mathrm{E}}\{|N(\lambda,\, \mu)|^2\}} = \frac{\widehat{\Phi}_s(\lambda,\, \mu)}{\widehat{\Phi}_s(\lambda,\, \mu) + \widehat{\Phi}_n(\lambda,\, \mu)} \tag{2.26}$$

or expressed by the *a priori* SNR $\widehat{\xi}(\lambda,\, \mu)$ estimate as

$$G_{\mathrm{W}}(\lambda,\, \mu) = \frac{\widehat{\xi}(\lambda,\, \mu)}{\widehat{\xi}(\lambda,\, \mu) + 1}. \tag{2.27}$$

A further approach which is often used is represented by Boll's idea of spectral subtraction [Bol79], which tries to reconstruct the speech spectrum by subtracting an estimate of the noise magnitude from the noisy speech spectrum as

$$|\widehat{S}(\lambda,\,\mu)| = |X(\lambda,\,\mu)| - \widehat{\mathrm{E}}\{|N(\lambda,\,\mu)|\}. \tag{2.28}$$

This leads to the gain computation rule

$$G(\lambda,\,\mu) = 1 - \frac{\widehat{\mathrm{E}}\{|N(\lambda,\,\mu)|\}}{|X(\lambda,\,\mu)|}. \tag{2.29}$$

A generalized version of the initial function by Boll was proposed by Hansen in [Han91] incorporating the two parameters α_S and β_S, and using the noise estimate $\widehat{N}(\lambda,\,\mu)$

$$G_\mathrm{S}(\lambda,\,\mu) = \sqrt{\left[1 - \left(\frac{|\widehat{N}(\lambda,\,\mu)|^2}{|X(\lambda,\,\mu)|^2}\right)^{\beta_\mathrm{S}}\right]^{\alpha_\mathrm{S}}}. \tag{2.30}$$

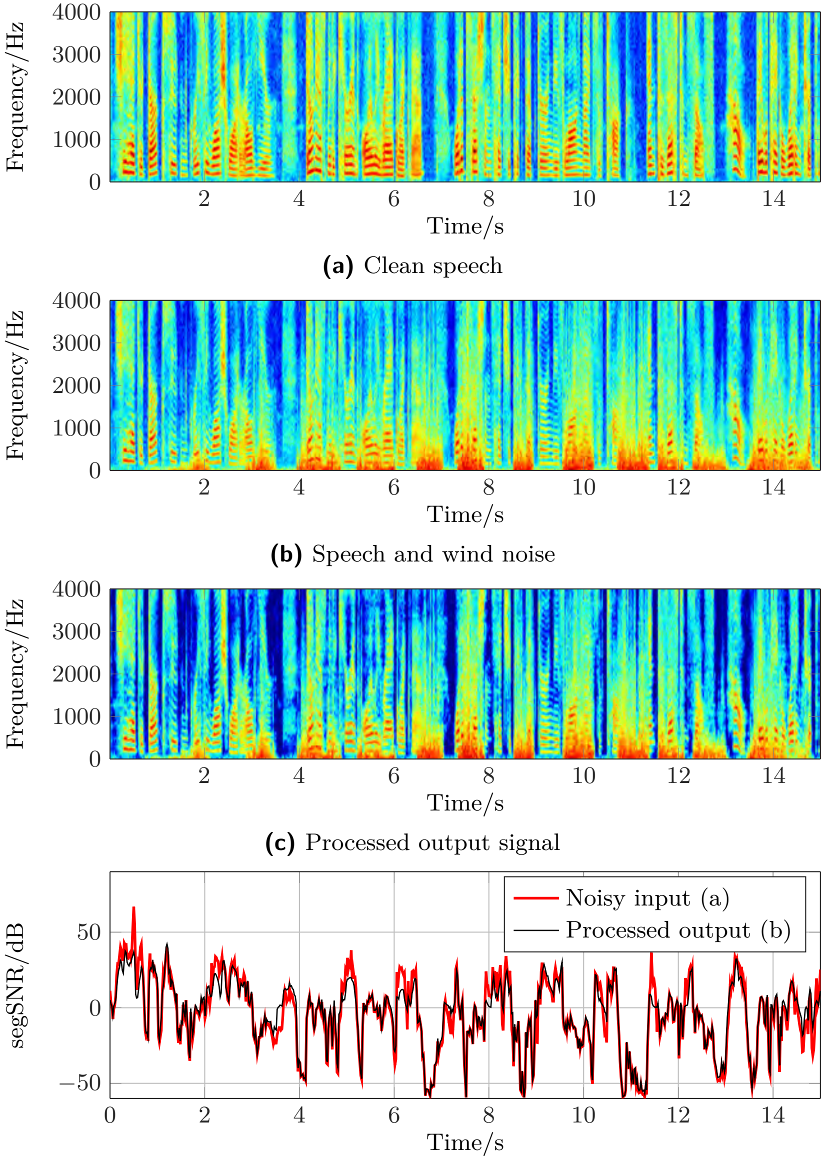
Different parameter settings provoke different realizations of the spectral subtraction gain. E.g., $\alpha_\mathrm{S} = 2$ and $\beta_\mathrm{S} = 0.5$ yields the magnitude subtraction proposed by Boll, power subtraction is given for $\alpha_\mathrm{S} = \beta_\mathrm{S} = 1$, and $\alpha_\mathrm{S} = 2$ and $\beta_\mathrm{S} = 1$ leads to the Wiener filter (c.f., Equation 2.26).

2.4 Conventional Noise Reduction Applied to Wind Noise Signals

In this section, an experiment is carried out by applying a conventional background noise reduction technique to a speech signal disturbed by wind noise. Here, the SPP based method [GH11] estimates the noise PSD, and the spectral gain is computed using the general spectral subtraction method as defined in Equation 2.30 with $\alpha_\mathrm{S} = 0.5$ and $\beta_\mathrm{S} = 2$.

Figure 2.5 shows different signals of the noise reduction task. In Figure 2.5a the spectrogram of the desired clean speech signal is represented, which is not known in a real scenario. The noisy input and output signals of the system are depicted by the spectrograms in Figures 2.5b and 2.5c, respectively. The low-frequency wind gusts are still clearly visible in the output spectrogram, e.g., at $t = 10\,\mathrm{s}$. A more precise insight on the performance of the noise reduction is given by the segmental SNR (segSNR) and is presented in Figure 2.5d. This widely used measure for the speech quality computes the SNR in each frame [QB88], where a high value indicates a good signal quality. Usually, the averaged value for a signal is computed in order to rate the performance of the noise reduction system under test by a single score.[4]

[4] Further information on the evaluation of noise reduction systems using instrumental measures is given in Appendix A.1.



(a) Clean speech

(b) Speech and wind noise

(c) Processed output signal

(d) Segmental SNR of (a) and (b)

Figure 2.5: Wind noise reduction using SPP based noise estimation [GH11] and spectral subtraction [Han91].

Here, the time-dependent values are presented by the red curve (input signal) and the black curve (output signal) in each frame. It can be seen, that over the whole signal length no or only a marginal improvement is visible. This holds for segments containing speech and noise (e.g., $t = 4 \ldots 5\,\mathrm{s}$) as well as segments with pure wind noise (e.g., around $t = 11\,\mathrm{s}$).

This experiment illustrates that conventional noise reduction systems fail in the case of wind noise. The poor results motivate the development of algorithms especially designed for the estimation and reduction of wind noise.

Chapter 3

Signal Analysis

All investigations in this work are aimed towards the enhancement of speech signals disturbed by wind noise. The first step towards this goal is an analysis of the disturbance. Therefore, it is necessary to investigate the recorded signals and derive characteristic properties to distinguish between speech and wind noise.

In general, noise reduction concerns the problem of suppressing sound sources, which are not the desired speech signal. Here, often the term background noise is used, which implies, that the source of the desired speech signal is closer to the used microphones than the unwanted noise sources as depicted in Figure 2.1. In contrast to that, wind noise is locally generated by an air stream around the device, which picks up the sound. In some publications wind noise is named “sensor artifacts” (e.g., in [SF12]), because it can not be related to a real sound source. In order to distinguish between wind noise and noise signals generated by sound sources in the ambience the term background noise will be used in the following for the latter.

Many publications are dealing with aero-acoustics, which describes the sound generation by air flows. Most of these investigations are carried out in the field of aerospace and automotive engineering (e.g. [Geo89], [Cro07], [MM09]). These studies consider artificially generated wind during the flight with an airplane or the car while driving. In contrast to that, this thesis takes into account the wind stream, which arises naturally in an outdoor environment caused by meteorologic phenomena. The main difference between these two scenarios is the range of the expected wind speed. While in the case of a driving car or an airplane wind speeds between 10 up to 300 m/s are considered, typically the wind speed takes values between between 0 and 20 m/s in an outdoor scenario.

Since this work deals with the processing of a digitized signal, only a short introduction in the generation of wind noise is given in Section 3.1. The used measurement setup is presented in Section 3.2. For the detection and reduction of the recorded noise it is more important to investigate the statistics of the recorded signals. This is carried out in Section 3.3 and emphasizes the difference between wind noise signals and background noise signals. The impact of wind noise on the speech quality in a communication system is evaluated in Section 3.4. Different approaches for a wind detection in short signal segments are presented in Section 3.5. Based on the signal statistics a model is derived in Section 3.6, for the simulation of the influence of wind noise in a recorded signal the generation of a reproducible artificial wind noise signal.

3.1 Wind Noise Generation

As aforementioned, wind noise in an outdoor environment is considered, where the flow velocity usually exhibits frequent changes. The variations of the velocity often described as wind gusts are provoked by large structures or natural objects such as buildings, cars, or trees in the vicinity. These obstacles in the air flow generate turbulences on a large scale, which are noticed as gusts. Due to the chaotic behavior of these turbulences, an exact information on the wind speed as well as the wind direction is not available.

A closer investigation is necessary to understand the acoustics, which are responsible for the generated sound. Figure 3.1 illustrates the scenario of the example of a mobile phone. Even if the wind direction and speed of the wind are known, the mobile phone or the head of the talker influence the air stream locally by a great amount. This effect can be transferred to any device equipped with microphones without a wind shield such as hearing aids, headsets or laptops. Consequently, the wind direction and speed close to the microphone can not be predicted and are assumed to be random variables.

Many publications are dealing with the sound generated aerodynamically. They all have in common, that turbulences in the air stream are responsible for the sound. Lighthill presented a general theory for the generation of the sound, where he explained the mechanics of the conversion from kinetic energy in an air stream to acoustic energy ([Lig52], [Lig54]). Furthermore, Lighthill mentioned, that "frequencies in the flow are identical with those of the sound produced", which leads to a high correlation between the wind speed and the measurable acoustic signal. The air flow around a solid surface is depicted in Figure 3.2 for two different wind speeds.

Because of friction losses the velocity of the flow is decreasing from the free-field velocity u_∞ towards the surface of the object. For a low free-field velocity a laminar flow profile is generated, which is shown in Figure 3.2a. The stream

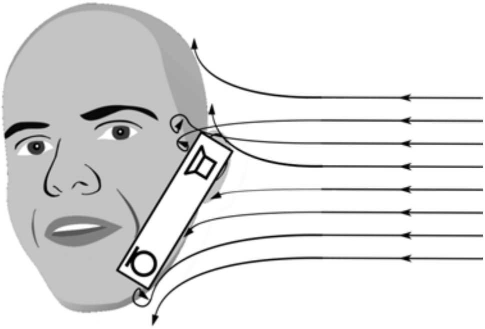
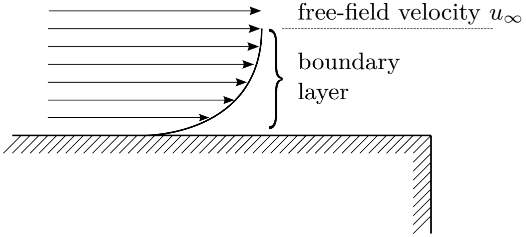
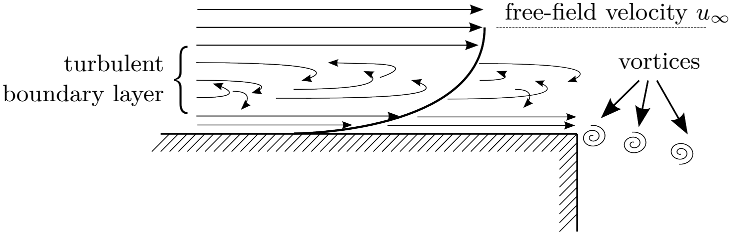


Figure 3.1: Wind stream around head and mobile phone.



(a) Laminar flow



(b) Turbulent flow

Figure 3.2: Airflow around a solid object with increasing free-field velocity from 3.2a to 3.2b.

consists of parallel layers with different velocities and the range of the flow stream, where the velocity is less than 99% of the free-field velocity is defined as boundary layer. As the wind speed U increases, the air stream will develop into a turbulent flow (Figure 3.2b). The threshold between a laminar and a turbulent stream is determined by the Reynolds number (e.g. [MM09])

$$R_e = \frac{\rho U D_\mathrm{c}}{\nu}, \tag{3.1}$$

as a function of the wind speed, where ρ and ν are the density and the viscosity of air, respectively. D_c is called the characteristic linear dimension and describes the size and geometric shape of the object in the air stream. In addition to the turbulent layer, vortices are shed at edges of the object. Bradley et al. focused on the investigation of effects of wind on hand-held communication devices [BWHB03]. They stated that the acoustic signal generation in a turbulent air flow can be decomposed into two main components.

- *Trailing edge vortex shedding*: On trailing edges in the air flow vortices are periodically generated. Depending on the velocity and the geometry of the surface, the periodical vortices lead to a tone at a defined frequency. Considering a constant air flow this will lead to a measurable peak in the spectrum [BWHB03], which is well below 50 Hz for normal outdoor wind conditions and dimensions of mobile communication devices.

- *Boundary layer turbulences*: As depicted in Figure 3.2b, turbulences occur within the boundary layer. They generate sound with a broader spectrum with main energy at lower frequencies.

Because in outdoor environments the wind is not a constant air stream, the vortex shedding frequency varies permanently and will not result in an isolated spectral peak, as shown in [BWHB03]. For realistic scenarios, the boundary layer turbulences are the main origin for the audible wind noise.

A mathematical description of the measured spectra of wind noise was developed by Strasberg ([Str88]). He stated that the logarithmic spectrum level $L_{\log}$ of the wind noise signal may be written as

$$L_{\log}(f) = 67 + 63\log_{10}(U) - 33\log_{10}(f) - 23\log_{10}(D_{\mathrm{c}}), \tag{3.2}$$

with the frequency f. The loudness level $L_{\log}$ is computed to a reference sound pressure of $20\,\mu$Pa. Transforming Equation (3.2) into a linear representation the sound pressure spectrum is given by

$$P(f) = \frac{20\,\mu\mathrm{Pa} \cdot 10^{3.35} \cdot U^{3.15}}{f^{1.65} \cdot D_{\mathrm{c}}^{1.15}}. \tag{3.3}$$

The relation shown in Equations 3.2 and 3.3 were derived empirically from several measurements, so an exact prediction of the relation between the sound and the wind speed or frequency is not possible. However, two important relations are given by Equation (3.3). The sound pressure rises with increasing wind speed ($P(f) \sim U^{3.15}$) and the sound pressure decreases with increasing frequency ($P(f) \sim 1/f^{1.65}$). Especially, the latter dependency is significant to explain the low-frequency energy distribution of wind noise, which will examined more detailed in Section 3.3.3.

3.2 Wind Noise Measurements

For the investigations in this work, several measurements were carried out, where mainly two scenarios were considered. For the investigation of wind noise under realistic conditions, outdoor recordings are the most appropriate way to obtain relevant wind noise data. The drawback of these measurements is, that it is hardly possible to avoid additional background noises such as movement in trees, passing cars or other noises generated by the wind in the surrounding of the recording set-up. For a precise analysis of a signal, it is required, that the considered signal is stored separately. Therefore, additional measurements under laboratory conditions can be helpful using an artificially generated air stream. This set-up can be realized in an audio lab, which provides a low-reverberant room with a reverberation time $T_{60}<$ 100 ms and an acoustic decoupling from other background noises. Here, a compressed air connection generates an adjustable air stream without further background noise sources.

Measurements using an artificial head to simulate the near-end speaker were carried out considering both the hand-held position (HHP) and the hands-free position (HFP) according to the European Telecommunications Standards Institute (ETSI) standard ETSI EG 201 377-2 ([ETS04]). More details and audio samples can be found in [NV14b].

3.3 Signal Statistics

The methods presented in this work all aim to reduce the effect of wind noise in recorded signals, which are available as digitized data. Acoustic countermeasures such as wind shields or wind insensitive microphone positions are not considered and their operating principle is only shortly explained in Section 4.1. For the reduction of wind noise by means of digital signal processing it is necessary to examine the statistics and spectral characteristics of wind noise in the recorded signal. The aim of this analysis is to identify characteristics, which provide a differentiation between the desired speech signal and the unwanted wind noise. First, a short description of the sound of wind noise is given in Section 3.3.1. In Section 3.3.2 and Section 3.3.3 the temporal and spectral features of wind noise are analyzed. For devices equipped with more than one microphone, the spatial characteristics of the recorded signals are of interest, which are investigated in Section 3.3.4. For the reduction methods presented in Chapter 4, it is assumed that the noisy input signal is a linear combination of the speech signal and wind noise. This is however not true in all cases. Therefore possible non-linear effects are discussed in Section 3.3.5.

3.3.1 Acoustics of Wind Noise

Wind noise generates a distinct sound in a recorded signal, which is normally immediately recognized by a listener. It is characterized by a low-frequency rumbling sound, which is closely related to the wind conditions of the near-end speaker. Figure 3.3 shows a sample of a typical wind noise recording taken outdoors. The spectrogram is given at the top and the corresponding time-domain signal is plotted at the bottom[1].

The spectrogram view clearly exhibits the low-frequency characteristic of wind noise with a spectrum, which exceeds the frequency range greater than 1 kHz only in segments with high wind noise levels, e.g., around $t = 2\,\mathrm{s}$. But even in these parts of the signal the main energy is located at lower frequencies. The fluctuations in the rumbling sound can also be seen in Figure 3.3 in both the spectrogram and the time-domain representation. Fast fluctuations in the noise signal are not only more annoying than a constant noise floor, but also reduce the intelligibility of speech, see, e.g., [FP90], [RV05] or [BG09]. The authors of these publications compared the recognition rate of speech in presence of constant noise and fluctuating noise signals. They found out, that for equal SNR values fluctuating noise signals always lead to significantly lower intelligibility results.

[1] Unless otherwise noted, all signals throughout this thesis are sampled with a sampling frequency of $f_\mathrm{s} = 16\,\mathrm{kHz}$

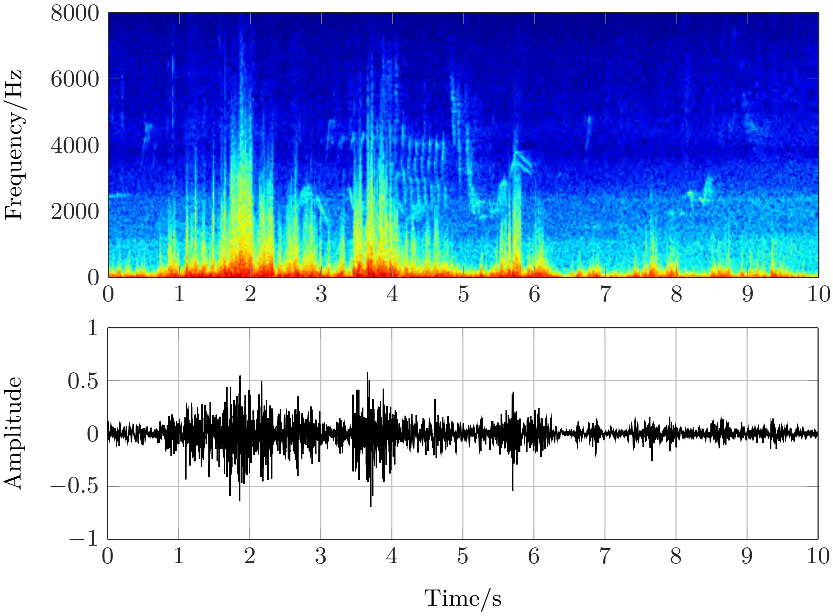


Figure 3.3: Typical wind noise sample from an outdoor recording [NV14b].

3.3.2 Temporal Characteristics

For the estimation and reduction of background noise in a speech signal, usually the temporal statistics are exploited as described in Chapter 2. This section investigates the characteristics of wind noise in a time-domain representation and compare them with speech signals and other noise signals. Since realistic scenarios are of interest for the reduction of wind noise, outdoor recordings are considered.

To reflect the temporal properties, in Figure 3.4 the progress of the frame energy $E_{\mathrm{ST}}(\lambda)$ of different signals is depicted[2], which is given for a signal $x(k)$ as

$$E_{\mathrm{ST}}(\lambda) = \sum_{k=\lambda \cdot L_{\mathrm{F}}+1}^{\lambda \cdot (L_{\mathrm{F}}+1)} x^2(k), \tag{3.4}$$

where L_{F} is the frame length of 320 samples ($\widehat{=}$ 20 ms) in which the signal is assumed to be stationary.

From the ETSI background noise database [ETS09], three typical background noise types *Inside Train Noise1*, *Work Noise Jackhammer* and *Pub Noise* are chosen for the investigations. In Figure 3.4, five seconds of the wind noise from the signal

[2] The frame energy is depicted in the unit $\mathrm{dB}_{\mathrm{FS}}$ referring to full-scale signal, i.e., the maximum scale is $x_{\max} = \pm 1$.

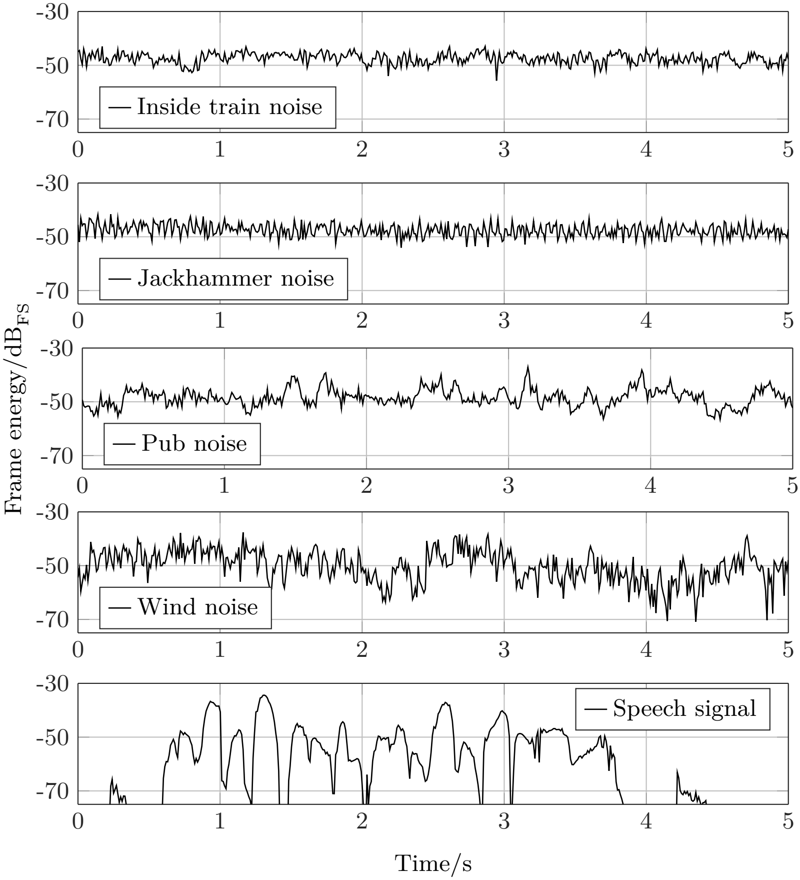


Figure 3.4: Frame energy of different noise signals and a speech signal.

given in Figure 3.3 is also shown and the bottom plot shows a sentence of female speech taken from the TIMIT database [LKS89]. As explained in Chapter 2 the degree of stationarity is deciding for the success of conventional noise reduction techniques. The temporal progress of the energy of the noise signals in Figure 3.4 shows an increasing degree of non-stationarity from *Inside Train Noise1* to *Pub Noise* and even more variations over time for the wind noise signal. For the speech signal, the frame energy suddenly rises after speech pauses and decreases in the same way at the end of speech activity. This behavior and the assumed constant noise level usually suffices to separate speech and noise signals. In the case of wind noise the sudden changes of the signal level during a wind gust does not fulfill this assumption. To quantify the degree of non-stationarity the short-term variance

$\sigma^2_{E,\mathrm{ST}}(\lambda)$ of the frame energy

$$\sigma^2_{E,\mathrm{ST}}(\lambda) = \frac{1}{L} \sum_{l=\lambda-(L-1)/2}^{\lambda+(L-1)/2} \left(E_{\mathrm{ST}}(l) - \overline{E_{\mathrm{ST}}}(\lambda)\right)^2, \tag{3.5}$$

is computed, where L is the number of consecutive frames considered for the computation and $\overline{}$ depicts the mean value over L frames. For investigating the stationarity, the variance over a duration of 100 ms ($L = 5$ frames of 20 ms) is taken into account. In Table 3.1 the averaged values of the variance $\overline{\sigma^2}_{E,\mathrm{ST}}$ over signals of 20 seconds are depicted.

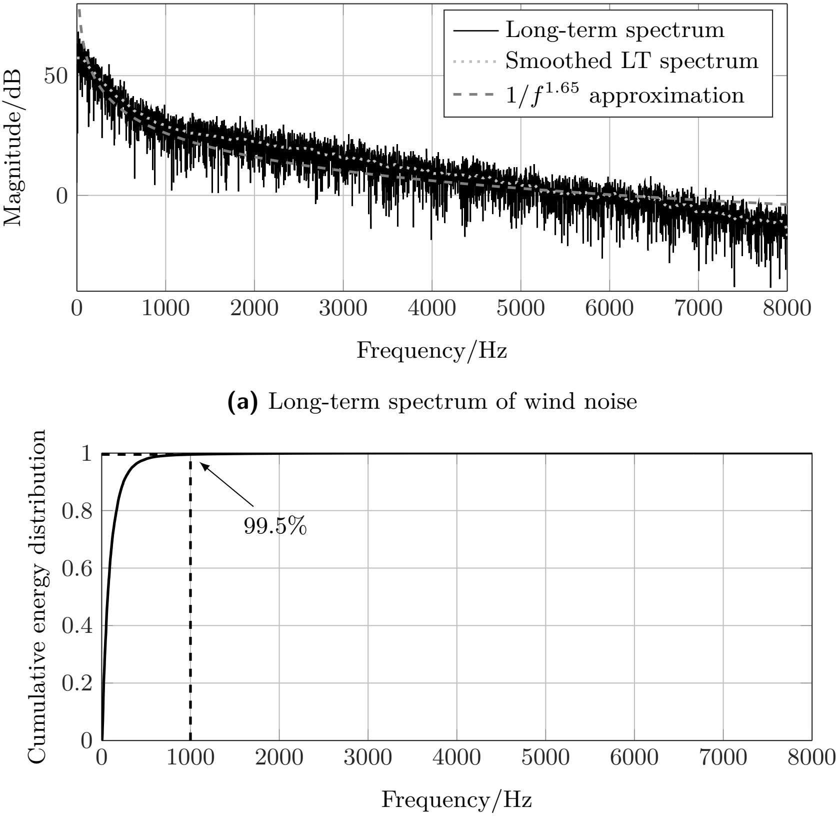
	Train noise	Jackhammer noise	Pub noise	Wind noise
$\overline{\sigma^2}_{E,\mathrm{ST}}/\mathrm{dB_{FS}}$	2.09	2.55	3.79	12.23

Table 3.1: Variance of short-term energy for different noise types.

It can be seen that the jack-hammer and the pub noise show a slightly higher variation than the train noise. But in contrast to the three background noise types, the variance of wind noise is significantly higher with a value over 12 dB$_{\mathrm{FS}}$. Besides the described fast variation of the wind signal level, also the signal energy varies over longer time intervals of several seconds (see Figure 3.3). In realistic scenarios, there are also periods of still air, which might occur between two wind gusts. These silent parts of the wind noise signal can further increase the variance, but are not taken into account for a better comparability with the other noise types. The temporal characteristics of wind noise illustrated in this section differ significantly from noise signals usually considered in typical speech enhancement problems. Especially, the high short-term variance is responsible for the low performance of conventional noise reduction schemes and motivates the development of techniques designed for wind noise reduction.

3.3.3 Spectral Characteristics

As for the temporal analysis, the investigation of the spectral properties of a signal in the discrete Fourier transform (DFT) domain can be carried out in a short-term (ST) and long-term (LT) consideration. Firstly, a general representation of the LT spectrum is given in Figure 3.5. For the depicted curves 60 seconds of wind noise from [NV14b] are taken. The LT spectrum in Figure 3.5a is given by the solid black line, and the dotted gray curve illustrates the general characteristic given by a smoothing over frequencies. Furthermore, the dashed gray line shows the decay related to $1/f^{1.65}$ as defined in Equation 3.3. It can be seen, that this mathematical definition does not perfectly fit the LT spectrum, but gives a good approximation of the rough spectral distribution of wind noise. As mentioned in Section 3.1, this description was derived from measurements with several microphones and might



(a) Long-term spectrum of wind noise

(b) Cumulative energy distribution of wind noise

Figure 3.5: Spectral energy distribution of wind noise.

be adapted to one certain microphone type. An easy way to adjust the spectral decay of the approximation can be realized by choosing different exponents ν of the frequency f as

$$\widetilde{N}(f) = \frac{1}{f^{\nu}} \ , \text{ with } \nu > 0. \tag{3.6}$$

A different representation of the spectral energy is presented in Figure 3.5b. The cumulative energy distribution beginning from low frequencies shows that most of the energy (99.5%) is below 1 kHz. This is important with regard to which parts of the speech is distorted. Speech only partly covers this frequency range. Mostly voiced speech segments are present in these frequencies (0-3000 Hz), while unvoiced speech can be expected at higher frequencies. A more detailed investigation on the influence on speech signals follows in Section 3.4.

The short-term (ST) spectral characteristics are shown in Figure 3.6 using three segments of 20 ms from the wind signal depicted in Figure 3.3. The segments are chosen from parts representing different wind levels of the signal. In addition two variants of the approximations from Equation 3.6 ($\nu = 1.65$ and $\nu = 1$) are given. Differences between spectral shapes of the wind noise segments are visible, which do not strictly follow the relation of the $1/f^{1.65}$ shape. But with the introduction of the parameter ν the magnitude of the ST wind spectrum can adopted for a better approximation.

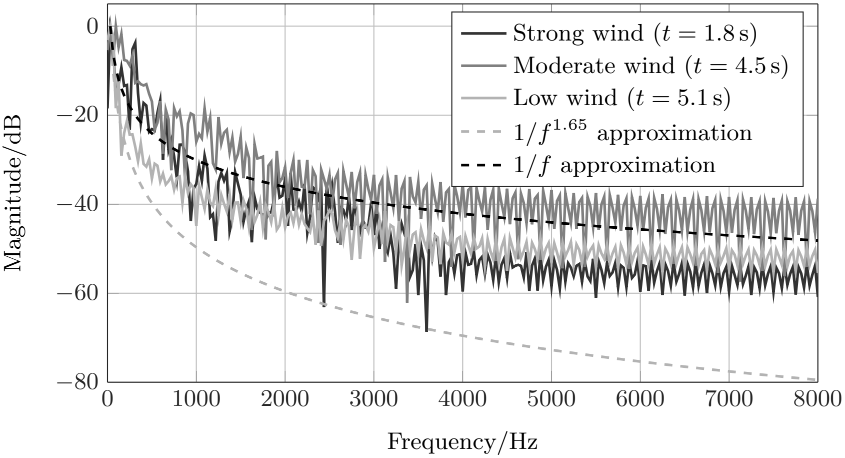


Figure 3.6: Short-term spectra of wind noise segments of different wind intensity. The corresponding temporal positions of the signal depicted in Figure 3.3 are given in brackets.

3.3.4 Multi Microphone Properties

In the current generation of smartphones, commonly the devices are equipped with more than one microphone. Many mobile phones have a primary microphone at the bottom of the device and at least one additional microphone at the top and/or the back of the housing. The additional microphone signals are usually exploited for background noise estimation and reduction. Hearing aids might also use two microphones at each device to apply a spatial filtering to the captured signals. The main difference between the two applications is the distance d_m between the two microphones. For mobile phones, a distance of 10 cm is quite common, whereas the microphones of hearing aids are closely spaced with a distance of about 1 cm. Figure 3.7 shows the general set-up of a dual microphone system recording a sound signal arriving from the angle θ.

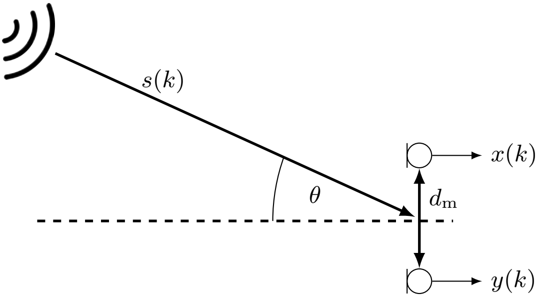


Figure 3.7: Dual microphone setup.

For the processing of multi microphone signals, often the spatial correlation is exploited to distinguish between different acoustic scenarios or sound fields. Considering the time-domain representation the cross-correlation of the signals can be investigated. A more useful analysis provides a frequency dependent correlation measure given by the coherence function between two signals $x(k)$ and $y(k)$ with limited energy (e.g., signals segments)

$$\Gamma_{xy}(\Omega) = \frac{\Phi_{xy}(\Omega)}{\sqrt{\Phi_{xx}(\Omega) \cdot \Phi_{yy}(\Omega)}}, \tag{3.7}$$

with the auto- and cross-PSDs $\Phi_{xx}(\Omega)$, $\Phi_{yy}(\Omega)$ and $\Phi_{xy}(\Omega)$ of the microphone signals $x(k)$ and $y(k)$. In general, the coherence function is complex-valued with a magnitude less than or equal to one. Often the so called magnitude squared coherence (MSC)

$$\mathcal{C}_{xy}(\Omega) = \frac{|\Phi_{xy}(\Omega)|^2}{\Phi_{xx}(\Omega) \cdot \Phi_{yy}(\Omega)} \tag{3.8}$$

is used instead, yielding real values between zero and one, where a high correlation leads to values close to one.

Different sound fields can be distinguished by their coherence properties. There is a variety of different coherence models, which can be mathematically derived for several acoustic scenarios (see, e.g., [Bit02]). Here, the most prominent three coherence models are relevant and will be explained in the following.

Coherent Sound Field

In a scenario depicted in Figure 3.7, a coherent sound field is generated by a single sound source. The corresponding complex coherence is given by

$$\Gamma_{xy}^{\mathrm{Coh}}(\Omega) = \cos(\Omega f_\mathrm{s} d_\mathrm{m} \cos(\theta)/c) - j \sin(\Omega f_\mathrm{s} d_\mathrm{m} \cos(\theta)/c), \tag{3.9}$$

where c is the speed of sound[3] [Kut09]. The MSC for this sound field is $\mathcal{C}_{xy}(\Omega) = 1$ for all frequencies and independent of the angle of arrival θ. Extensions to this model for more than one sound source can be found in [Bit02]. Results from experiments with two microphone setups are shown in Figure 3.8a, where for both configurations (2 and 10 cm microphone distance) the expected high value of the MSC can be measured over the complete frequency range.

Diffuse Sound Field

A more complex scenario is described by the so-called diffuse sound field. In that case, the sound is generated by numerous independent sound sources equally distributed around the microphone array. For this sound field, the frequency dependent real-valued coherence function becomes [Kut09]

$$\Gamma_{xy}^{\mathrm{Dif}}(\Omega) = \mathrm{sinc}(\Omega f_{\mathrm{s}} d_{\mathrm{m}}/c). \tag{3.10}$$

Dependent on the microphone distance the lower frequencies show a higher coherence, while for higher frequencies the coherence decreases. This characteristic is depicted by the dashed curves in Figure 3.8b. Many background noise situations reflects a diffuse sound field, where the noise sources are distributed around the microphones, e.g., babble noise from a crowd or street noise from many cars in the background. Therefore often a diffuse noise field is assumed, when a dual microphone set-up is examined.

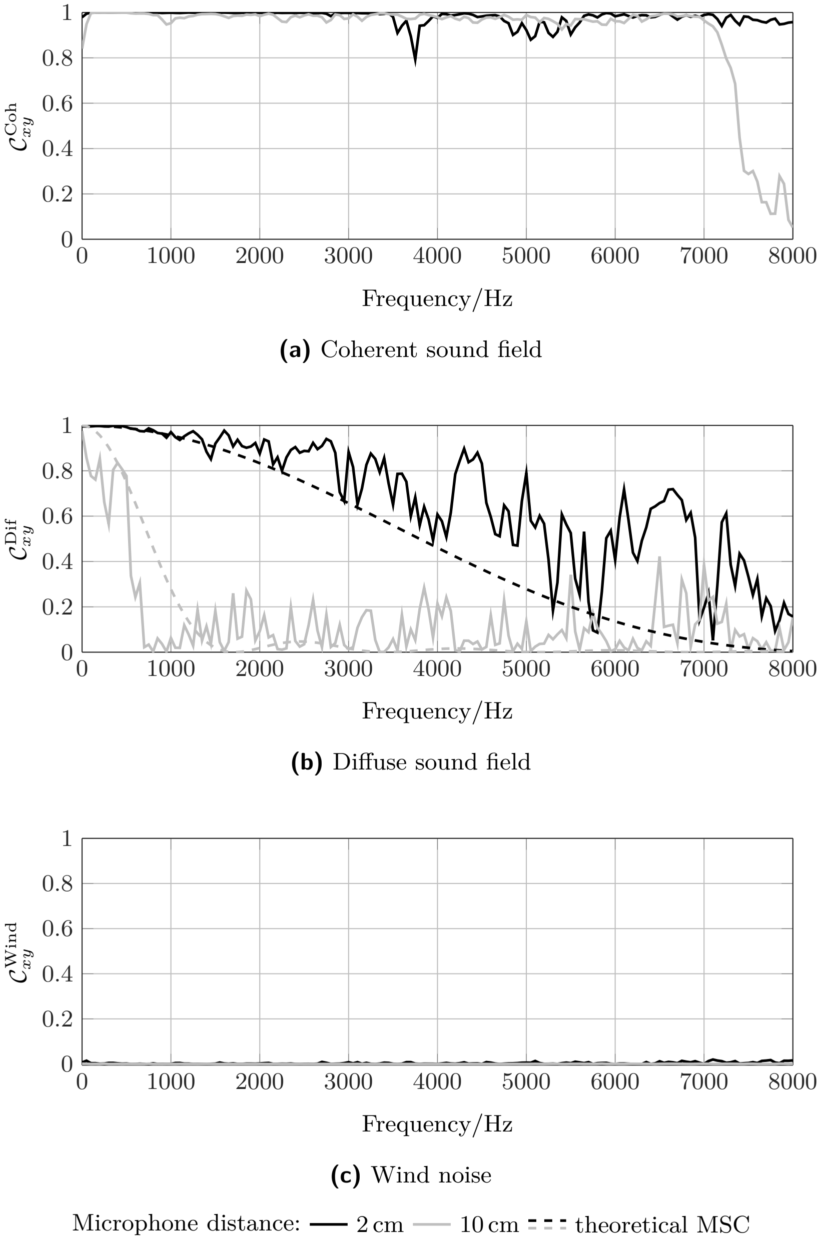
Again measurements were carried out, while a diffuse sound field was generated according to ETSI standard 202 396-1 [ETS09]. The measured MSC curves are shown by the solid curves in Figure 3.8b. Especially for the microphone distance of 10 cm depicted by the gray line, the low MSC values for frequencies higher than 1000 Hz are clearly visible while for the smaller distance of 2 cm (black lines) the MSC descends only slowly over frequency.

Incoherent Sound Field

As mentioned at the beginning of this chapter, the generation of wind noise differs significantly from other sound signals. Because the turbulences in the boundary layer are responsible for the generated noise signals, the sound sources are located in the direct proximity to the microphones themselves. Thus, the sound generation mechanisms can be seen as independent acoustic sources close to the microphone positions, which leads to a low spatial correlation for recorded wind noise signals in multi microphone scenarios. In literature different mathematical expressions can be found to describe the coherence in a boundary layer turbulence field. The authors in [Cor64] and [Elk07] assume that the coherence can be formulated as an exponential decay

$$\Gamma_{xy}^{\mathrm{Wind}}(\Omega) = \exp\left(-\frac{\alpha_{\mathrm{D}} \Omega f_{\mathrm{s}} d_{\mathrm{m}}}{0.8U}\right) \tag{3.11}$$

[3] $c = 343\,\mathrm{m/s}$ is considered in this thesis, which is given at an air temperature of 20° C.



(a) Coherent sound field

(b) Diffuse sound field

(c) Wind noise

Figure 3.8: Magnitude squared coherence (MSC) of different sound fields displayed by measured values (solid) and theoretical curves (dashed).

over frequency and microphone distance d_m with an empirically determined decay constant α_D. The relation in Equation 3.11 would introduce some high coherent parts at lower frequencies ($\Gamma_{xy}^{\mathrm{Wind}}(0) = 1$). A coherence function

$$\Gamma_{xy}^{\mathrm{Wind}}(\Omega) = 0 \tag{3.12}$$

over the complete frequency range is assumed in [SF12], which implies that the wind noise components in each microphone signal are completely uncorrelated. Measurements support the latter assumption as depicted by the curves in Figure 3.8c for both microphone distances. Thus, in the following a zero coherence property is assumed for wind noise signals.

3.3.5 Non-linear Effects

Usually, noise in a speech signal is described as an additive component, which presents the noisy microphone signal as a linear combination of the clean speech signal and the pure noise signal. However, wind noise exhibits partially very high signal levels, which might lead to non-linear effects. Consequently, two types of non-linear effects are worth to be investigated in more detail.

High levels of the input signal can lead to amplitudes in the captured signal, which are higher than the dynamic range of components of the recording device. This might be the microphone itself or limits of the signal amplifier and/or the analog-digital converter. Such a violation is called clipping and results in samples in the recorded signal, which are limited to the maximum signal level. An example is given in Figure 3.9 showing the spectrogram and the time-domain representation of a recorded speech signal in a windy situation[4]. Both the low frequency wind noise and the harmonic structure of the speech signal are clearly visible in the spectrogram. The samples, which are clipped, are marked with red rectangles in the lower plot of Figure 3.9. As a result, in the spectrogram the clipped areas of the noise reveals also high frequency components, which can be clearly seen in speech pauses around $t = 0.3\,\mathrm{s}$ or $t = 2.8\,\mathrm{s}$ (marked by the arrows below the spectrogram in Figure 3.9). Short segments with clipped samples can be seen as nearly ideal Dirac impulses, as they reach the maximum amplitude for a short durationa of only a few samples. In the short-time frequency-domain, a single Dirac impulse results in the broad spectral representation, which is visible in the depicted spectrogram. Besides the clipped segments, the higher frequencies seem to be unaffected by the wind noise. Several approaches can be found for the de-clipping of audio signals (see, e.g., [AEJ+12] and references therein). But all these algorithms presume that high-level segments of desired signal are responsible for the limitation of the recorded signal. In the described case in Figure 3.9 the clipping is caused by the noise signal. Thus, a restoration of the clipped signal parts is not desired, because this mainly restores the wind noise portions in the signal. A better treatment would be an attenuation or suppression of the signal segments clipped by wind noise.

[4] The noisy speech signal was directly recorded, using a loudspeaker-microphone setup in an outdoor environment.

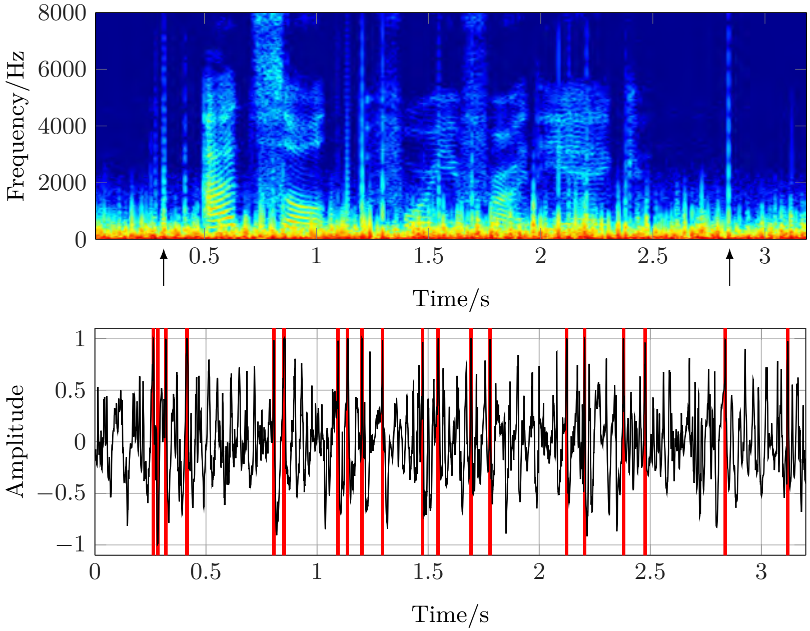


Figure 3.9: Noisy speech clipped due to high wind levels.

Apart from clipping, the high pressure level by the wind noise might lead to a displaced operating point of the recording hardware due to a mechanical offset introduced by the wind stream. This may result into different impacts, e.g., an extreme excursion of the microphone membrane and also saturation effects from the amplifiers. Both incidents might lead to a non-linear behavior of the recording chain. Non-linear distortions of audio hardware can be determined by the total harmonic distortion (THD), which is given by the power response $P(f)$ of the test device to a sine wave at frequency f. Any non-linear behavior will generate additional signal components at multiples of the excitation frequency. Therefore, the THD in the discrete frequency representation

$$\mathrm{THD}(\mu) = \sqrt{\frac{P(2 \cdot \mu) + P(3 \cdot \mu) + \ldots P(N \cdot \mu)}{P(\mu)}} \tag{3.13}$$

can be inspected to investigate any non-linear behavior. Usually, a THD up to 0.5% to 1 % is tolerated for high quality audio recordings. In this work, only the influence of wind noise on the used microphone is examined. The power response $P(\mu)$ is given by the squared magnitude of the discrete spectrum of the measured signal. In the experiments the used microphones (Sennheiser ME 2) are exposed to a great amount of wind while simultaneously sine signals are played by a loudspeaker as excitation signals for the THD measurement. It is ensured for all measurements that no signal parts are clipped, because only the influence of high wind noise levels to the microphone characteristics is investigated. As stated in Section 3.3.3 most of the wind noise energy is located well below 1 kHz. Here, an additional safety bandwidth of 1 kHz is taken in order to not influence the THD measurements beginning from 2 kHz. It turned out that taking the first 5 harmonics of the excitation sine signal are sufficient to measure the THD. With these two aforementioned constraints and a sampling frequency of 48 kHz, only a small frequency range between 2 kHz and 3.5 kHz can be investigated, which is depicted in Figure 3.10.

For the investigations two different wind speeds were considered (dotted and solid curves) and a reference measurement with no wind is also depicted in Figure 3.10 by the dashed line. A small increase of the THD can be seen for the measurements with wind noise. But the absolute THD value is still quite low (<0.3 %), which indicates that the non-linear steady state distortions induced by wind noise are not crucial and will be neglected in the following. A more crucial problem might be the aforementioned clipping in segments with high wind levels.

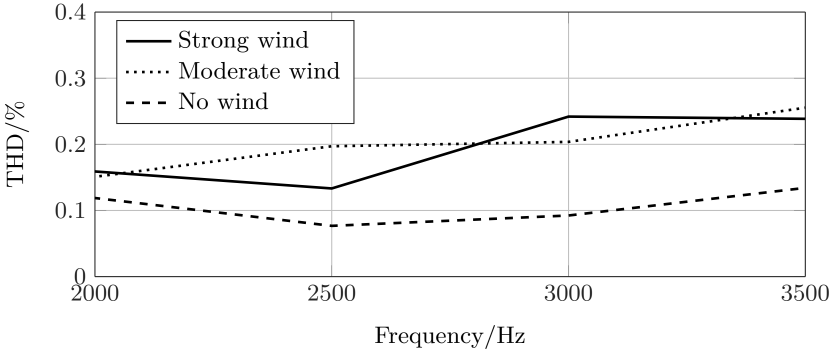


Figure 3.10: Total harmonic distortion (THD) at different wind speeds.

3.4 Influence on Speech Communication Systems

This section investigates the effect of wind noise on the quality and intelligibility of a speech signal. Therefore speech was recorded with an artificial head simulating the near-end speaker. The speech levels were chosen to 89.3 $\mathrm{dB_{SPL}}$ at the mouth reference point and to 65.3 $\mathrm{dB_{SPL}}$ at the hand-held position (HHP) and the hands-free position (HFP), respectively, as defined in [ETS04]. The HHP represents the normal position of mobile phone during a telephone conversation close to the head (c.f. Figure 3.1). Using the phone in the speakerphone mode, the HFP defines a position of the phone 50 cm in front of the head of the speaker. Speech samples of female and male speakers from [Kab02] were randomly taken. The degree of degradation was measured in terms of the speech quality by the perceptual evaluation of speech quality (PESQ) value, see [RBHH01], [IT01], [IT07] and the intelligibility given by the short-time objective intelligibility (STOI) [THHJ10].

The PESQ value in the used implementation ranges from 1 (poor quality) to 4.5 (no degradation) and the intelligibility coefficient estimated by STOI ranges from 0 to 1, where 1 indicates a perfect intelligibility. Besides, the global SNR was calculated over the whole signal length. For the two positions three scenarios were investigated: a constant slow wind stream (≈ 5 m/s), a constant fast wind stream (≈ 10 m/s) and a varying wind stream with wind speeds up to 10 m/s. The latter condition reflects a realistic scenario in which gusts of the wind leads to fast changes of the wind speed. The evaluation of all scenarios is given in Table 3.2.

		SNR/dB	PESQ	STOI
slow wind	HHP	6.08	1.38	0.93
(≈ 5 m/s)	HFP	-9.19	1.04	0.79
fast wind	HHP	-5.41	1.09	0.87
(≈ 10 m/s)	HFP	-20.68	1.02	0.7
wind gusts	HHP	-2.95	1.09	0.78
(up to 10 m/s)	HFP	-18.22	1.06	0.52

Table 3.2: Quality measures from noisy speech in hand-held position (HHP) and hands-free position (HFP).

Clearly negative SNR values can be seen in almost all cases, except the slow wind case in HHP. This extreme annoying noise impairs the speech quality as seen by the very low PESQ values. Furthermore, the wind has influence on the speech intelligibility given by the decreased STOI measures. This is especially true for the last considered scenario, the varying wind stream which reflects the most realistic condition. Here even higher SNR values in the wind gust scenario show a lower speech intelligibility for both the HHP and the HFP.

The results of the presented investigations shows that wind noise can be a severe problem for many communication devices in terms of the perceived speech quality and the intelligibility. Hence, it is necessary to develop algorithms for the detection and reduction of wind noise.

3.5 Wind Noise Detection

In this section a frame-wise detection of wind noise is considered, which can be realized either in the time- or frequency-domain. Several algorithms for wind noise detection can be found in the field of signal processing for hearing aids (see, e.g., [Kat08] for an overview). A good detection of wind noise is the first step towards a suppression of distortion in the captured signals. Furthermore, a detection method for wind noise is very helpful for outdoor recordings and videos, where a degradation of the recorded signal by wind might not be noticed during the recording process. In this case a warning for the user could be displayed to indicate the presence of wind noise. In the following, the most promising approaches for the detection of wind noise in a single microphone signal are presented and compared in terms of their accuracy, which were also presented in [NJV16].

3.5.1 Time Domain Approaches

Methods for wind noise detection in the time-domain use the input signal $x_\lambda(\kappa)$, where $\kappa = 0 \dots L_\mathrm{F} - 1$ states the sample position within the frame λ. The frames are available as 20 ms segments with an overlap of half frame-size and windowed with a square-root Hann window, which is the standard configuration of the analysis block in the considered noise reduction system.

3.5.1.1 Zero Crossing Rate

The zero crossing rate (ZCR) is defined as the number of sign-changes of a given signal within a fix duration, i.e., the rate at which the signal changes from positive to negative magnitudes or back and is defined as

$$\mathrm{ZCR}(\lambda) = \frac{1}{L_\mathrm{F} - 1} \sum_{\kappa=1}^{L_\mathrm{F}-1} \mathbb{I}\{x_\lambda(\kappa) \cdot x_\lambda(\kappa - 1) < 0\} \ \in [0, 1] \tag{3.14}$$

where L_F is the frame-size and the indicator function $\mathbb{I}\{A\}$ is 1 if its argument A is true and 0 otherwise. The ZCR is dependent on the frequency components and is a well known feature in the field of voice activity detectors (VAD). Low frequency signals result in slow changes of the time signal and thus a low number of sign-changes is generated resulting in a ZCR close to zero. Higher frequencies in the considered signal will produce more sign-changes, which leads to ZCR-values closer to one. Because each signal can be seen as a sum of sine waves representing the different frequency components, the frequency component with the highest amplitude will mainly affect the ZCR. To detect wind segments, it is proposed in [NLZIT10] to measure the ZCR in each signal frame, as the high amplitudes at low frequencies will also generate a low ZCR. For the wind noise detection, it is preferable to have a soft decision in terms of an indicator in the range between zero and one for the two conditions *wind inactive* and *wind active*, respectively. Thus the wind noise indicator based on the ZCR is simply defined as

$$\mathcal{I}_{ZCR}(\lambda) = 1 - \mathrm{ZCR}(\lambda). \tag{3.15}$$

3.5.1.2 Short-Term Mean

A further result of the low frequency characteristic of wind noise can be investigated by the normalized short-term mean (NSTM) of the signal. Usually, the digital representation of an acoustic signal can be assumed to be zero-mean (see, e.g., [WMG79], [Mar05]). Besides, almost every recording equipment shows a certain high-pass characteristic, e.g., with a cut-off frequency at 5-10 Hz. This is necessary to remove the direct component (DC) in the complete signal, which impairs the further processing of the signal such as the quantization. The zero-mean property is only valid in a long-term sense, while shorter signal segments can show a DC depending on its frequency components. The DC or mean value of short segments can be used to detect low frequency parts in a signal and is here defined in a normalized way as

$$\mathcal{I}_{\mathrm{NSTM}}(\lambda) = \left| \frac{\sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} x_\lambda(\kappa)}{\sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} |x_\lambda(\kappa)|} \right| . \tag{3.16}$$

Because the sign of the DC provides no information, the absolute value is taken and the normalization with the sum of the absolute values of $x_\lambda(\kappa)$ leads to values close to zero for high frequency components. For a signal containing only a DC, the two sums in Equation 3.16 will have the same amplitude and thus the NSTM will be one. An analysis of the NSTM is carried out to investigate the influence of different frequency components in a considered signal. It is assumed that the a signal can be decomposed into its frequency components each represented by a sine or cosine wave according to:

$$x_\lambda(\kappa) = \sum_{\mu=1}^{N} a_\lambda(\mu) \cos(2\pi \cdot (f_\mu / f_{\mathrm{s}} \kappa + \phi_\lambda(\mu))), \tag{3.17}$$

which can be seen as discrete cosine transformation (DCT) (see [ANR74]) of a signal. The index μ describes the discrete frequency f_μ of each cosine component, which is weighted by $a_\lambda(\mu) \geq 0$ and delayed by the phase term $\phi_\lambda(\mu)$. Equation 3.16 can now be rewritten to

$$\mathcal{I}_{\mathrm{NSTM}}(\lambda) = \left| \frac{\sum\limits_{\mu=1}^{N} a_\lambda(\mu) \sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} \cos(2\pi \cdot (f_\mu / f_{\mathrm{s}} \kappa + \phi_\lambda(\mu)))}{\sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} |x_\lambda(\kappa)|} \right| \tag{3.18}$$

$$= \frac{1}{\sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} |x_\lambda(\kappa)|} \cdot \left| \sum_{\mu=1}^{N} a_\lambda(\mu) \widetilde{\mathcal{I}}_{\mathrm{NSTM},\mu} \right| , \tag{3.19}$$

where

$$\widetilde{\mathcal{I}}_{\mathrm{NSTM},\mu} = \sum_{\kappa=0}^{L_{\mathrm{F}}-1} \cos(2\pi \cdot (f_\mu / f_{\mathrm{s}} \kappa + \phi_\lambda(\mu))) \tag{3.20}$$

is the non-normalized NSTM of one cosine component at frequency f_μ. From Equation 3.19 follows that $\mathcal{I}_{\mathrm{NSTM}}$ is the weighted sum of the NSTM of each frequency component μ. An experiment for each cosine component is carried out, where the normalized NSTM $\mathcal{I}_{\mathrm{NSTM},\mu}$ for each frequency f_μ is calculated separately. For the simulation the usual frame-size of 20 ms is used. It is assumed that in natural signals the phase of each frequency component is randomly distributed, thus every possible value for ϕ_μ is considered[5] and the values were averaged afterwards.

The resulting curve is plotted in Figure 3.11 for the frequency range of f_μ between 0 and 2000 Hz. It is obvious that a for

$$f_\mu = m \cdot 1/L_{\mathrm{F}} \cdot f_{\mathrm{s}} = m \cdot 50\,\mathrm{Hz}, \quad m \in \mathbb{N}^+ \tag{3.21}$$

the NSTM takes the value 0, because in these cases one or multiples of the cosine period length are equal to the frame size L_{F} and the resulting sum in Equation 3.20 over a whole cosine period is zero. Besides, the general behavior of the depicted curve shows, as expected, high NSTM-values for low frequencies and vice versa. As shown in Equation 3.19 the complete NSTM based wind indicator $\mathcal{I}_{\mathrm{NSTM}}$ is represented by the weighted sum of its frequency components given by $\widetilde{\mathcal{I}}_{\mathrm{NSTM},\mu}$. Of special interest are frequencies below 50 Hz (marked by the gray dashed line in Figure 3.11), where a great amount of the energy of wind noise is located. The higher frequencies between 100-2000 Hz, where the main speech energy is distributed (see, e.g., [BDT+94]), show a clearly lower NSTM value.

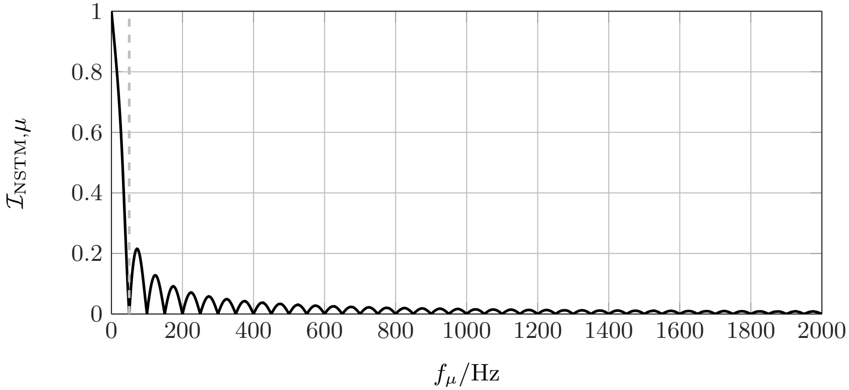


Figure 3.11: Theoretical NSTM values of single cosine components from Equation 3.20 for 20 ms frames. The dashed line represents the frequency with a period length equal to the considered frame-size.

[5] Usually, the phase term ϕ_μ is a continuous variable. For the experiment only discrete values are considered in the range $\phi_\mu = 0, \tau_{\mathrm{s}}, 2 \cdot \tau_{\mathrm{s}} \ldots, \lfloor 1/f_\mu \rfloor$. This reflects a cyclic sample-wise shift ($\tau_{\mathrm{s}} = 1/f_{\mathrm{s}}$) of each frame over the whole period of the cosine at the considered frequency f_μ.

A further experiment with real wind noise signals and speech signals is carried out to confirm the considerations previously made. Both, a clean speech signal and a pure wind noise signal are segmented and windowed as described in Section 2.2. The NSTM is calculated for each frame according to Equation 3.16 and the experiment is repeated for different frame sizes between 5 and 100 ms.

The results are given in Figure 3.12, where the averaged values of all frames are represented by the black and gray curve for wind noise and speech, respectively. It can be seen that the zero-mean property is valid for speech for frame sizes greater than 20 ms and a clear distinction between speech and wind noise is possible for all considered frame sizes. Thus, the NSTM can be used to detect wind noise without a great influence of speech signals, which might be active at the same time.

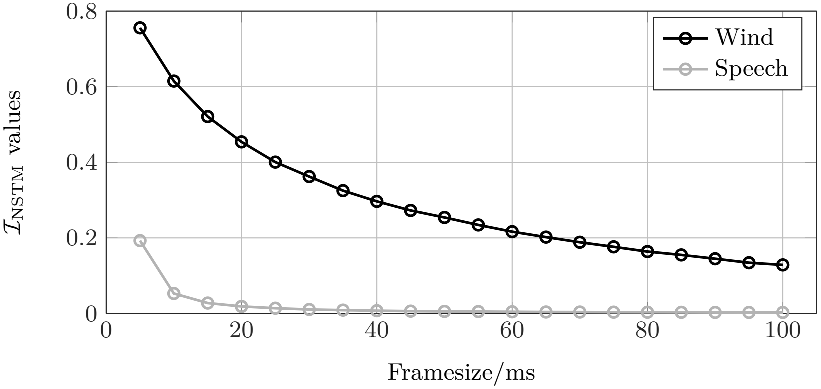


Figure 3.12: Measurement of NSTM for speech and wind noise signals.

3.5.2 Frequency Domain Approaches

For a wind noise detection in the frequency-domain the DFT representation of the input signal spectrum as $X(\lambda,\, \mu)$ with frame-index λ and discrete frequency bin μ is considered. In this section, the vector notation of $X(\lambda,\, \mu)$ will be used as

$$\mathbf{X}(\lambda) = [|X(\lambda, 0)|, |X(\lambda, 2)|, \dots, |X(\lambda, M/2)|]^T, \tag{3.22}$$

containing the magnitudes of the complex DFT coefficients for each frequency bin. As the DFT provides a symmetric spectrum, only the first $M/2 + 1$ has to be observed. All presented methods have in common that they exploit the decreasing spectral characteristic of wind noise over frequency.

3.5.2.1 Negative Slope Fit

One detector presented in [NLZIT10] is based on the idea that the magnitude of the spectrum of wind noise can be roughly approximated by a linear decay over the frequency, which can be expressed as

$$\widehat{\mathbf{X}}(\lambda) = a_1 \cdot \boldsymbol{\mu} + a_0 \tag{3.23}$$

with the frequency vector

$$\boldsymbol{\mu} = [0, 1, \ldots, M/2]^T \tag{3.24}$$

The parameters a_0 and a_1 control the DC and the slope of the approximation and will be denoted by

$$\boldsymbol{a} = [a_0, a_1]^T. \tag{3.25}$$

Combining the frequency vector with a vector $\mathbf{1} = [1, 1, \ldots, 1]^T$ containing $M/2{+}1$ ones as a $2{\times}(M/2{+}1)$, matrix

$$\mathbf{M} = [\mathbf{1}, \boldsymbol{\mu}] \tag{3.26}$$

Equation 3.23 can be written as

$$\widehat{\mathbf{X}}(\lambda) = \mathbf{M} \cdot \boldsymbol{a}. \tag{3.27}$$

Because for wind noise a negative slope is expected, the approach is named negative slope fit (NSF). A least square analysis can be applied to compute the optimal parameters for a given spectrum $\widehat{\mathbf{X}}(\lambda)$ minimizing the squared error

$$e(\lambda) = ||\mathbf{X}(\lambda) - \widehat{\mathbf{X}}(\lambda)||^2 \overset{!}{=} \min. \tag{3.28}$$

Setting the derivation with respect to the parameter vector $\boldsymbol{a}$ to zero leads to the optimal solution

$$\boldsymbol{a}_{\mathrm{opt}}(\lambda) = (\mathbf{M}^T\mathbf{M})^{-1} \cdot \mathbf{M}^T \cdot \mathbf{X}(\lambda). \tag{3.29}$$

According to [NLZIT10], two conditions must be fulfilled to classify the current frame as wind noise. Firstly, the slope of the approximated spectrum must be negative ($a_1 < 0$) and secondly the squared error $e(\lambda)$ must be smaller than a certain threshold. Normalizing the error to the energy of the observed spectrum the two conditions can be combined to the wind indicator

$$\mathcal{I}_{\mathrm{NSF}}(\lambda) = \begin{cases} 1 - \dfrac{e(\lambda)}{||\mathbf{X}(\lambda)||^2} & \text{, for } a_1 < 0, \\ 0 & \text{, otherwise.} \end{cases} \tag{3.30}$$

in the range between zero and one. A closer investigation of this algorithm has shown that an increased performance can be achieved by applying the indicator only on a limited frequency range between 0 and 1000 Hz, where most wind energy is expected.

3.5.2.2 Signal Sub-band Centroids

In [NCBV14] and [NV15] a method is proposed that investigates the energy distribution of a given spectrum. There are many ways to describe the energy distribution, e.g., by the spectral envelope or spectral flatness measures. A feature known from automatic speech recognition (ASR) systems are the so-called sub-band signal centroids (SSC) (see, e.g., [Pal98]). They depict the center-of-gravity in a given sub-band range from f_1 to f_2 and are defined for a signal x by

$$\Xi_{f_1,f_2} = \frac{\int\limits_{f_1}^{f_2} \Phi_{xx}(f) \cdot f df}{\int\limits_{f_1}^{f_2} \Phi_{xx}(f) df} \tag{3.31}$$

For a theoretical investigation of the sub-band signal centroid (SSC), this continuous frequency-domain representation is considered. It is assumed that the wind noise magnitude spectrum can be approximated by an $1/f$ slope, which yields in the wind noise PSD approximation

$$\Phi_{nn}(f) \approx \frac{\beta}{f^2}. \tag{3.32}$$

The parameter β scales the total signal energy of the wind noise PSD. Inserting Equation 3.32 in Equation 3.31, β cancels out and the integrals can be solved, giving the following expression for the definition of the wind SSC

$$\Xi_{f_1,f_2,\text{wind}} = f_1 \cdot f_2 \cdot \left(\frac{\ln(f_2) - \ln(f_1)}{f_2 - f_1} \right) \tag{3.33}$$

as a function of the frequency limits f_1 and f_2. An interesting feature is that $\Xi_{f_1,f_2,\text{wind}}$ tends towards zero, if $f_1 \to 0$, i.e., the considered sub-band begins at $f = 0\,\text{Hz}$.

Similar to the $1/f$-approximation of the wind noise a description of the speech spectrum is required to investigate the behavior of the SSC for speech signals. Here, the so-called long-term average speech spectrum (LTASS) is used as it is defined in the ITU-T P.50 standard for the generation of an artificial voice signal [IP99]. The LTASS $\Upsilon(f)$ is a mathematical description of the spectral characteristic of speech and defines the logarithmic spectrum density in dB relative to 1 pW/m^2 [IP99] as

$$\Upsilon_{\log}(f) = -376.44 + 465.44 \cdot \log_{10}(f) - 157.75 \cdot \log_{10}(f)^2 + 16.71 \cdot \log_{10}(f)^3 \tag{3.34}$$

and is depicted in Figure 3.13.

Although, the LTASS also exhibits a low-frequency characteristic, where most of the energy is located between 200 and 500 Hz, the spectral energy distribution measured by the SSCs will depict a clear distinction between speech and noise. An important adjustment for the SSC determination is the choice of the sub-band range $f_1 \dots f_2$ (or $\mu_1 \dots \mu_2$ in the discrete case, respectively).

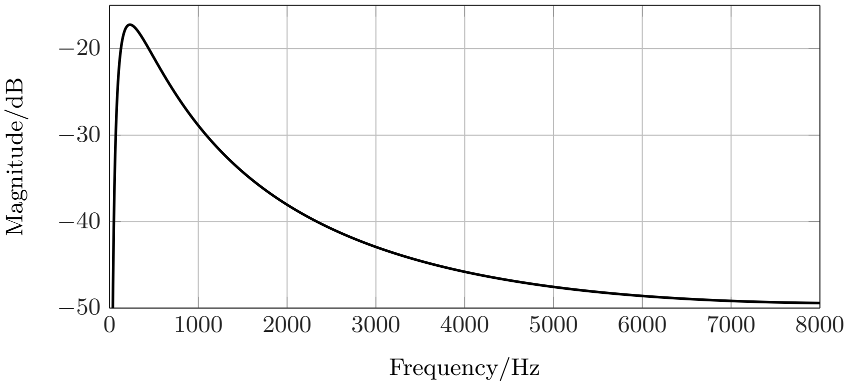


Figure 3.13: Long-term average speech spectrum according to [IP99].

Using the definition of the wind noise SSC from Equation 3.33 and the measured SSC of the LTASS representing the speech, different parametrization of f_1 and f_2 are compared in Figure 3.14. While on the x-axis different f_1-values are considered, each color of the depicted curves represents one choice of f_2. The curves only show values for $f_1 < f_2$, because this condition must be fulfilled for the computation of the SSCs. For most of the displayed values of f_1, no distinct difference can be observed between the dashed lines representing the speech SSC and the solid lines representing the noise SSC. But as expected, if the lower frequency limit f_1 tends towards 0 Hz the wind noise SSC also converges towards zero, while the speech SSCs takes a value of approximately 500 Hz as shown in the magnified view in Figure 3.14. As a result, $f_1 = 0$ Hz is a good choice for the SSC computation while different f_2 values only show a minor influence.

For the implementation in a digital signal processing system, the discrete frequency-domain representation from Equation 3.35 is used beginning at low frequencies ($\mu_1 = 0$) up to the discrete frequency bin μ_2 corresponding to f_2.

$$\Xi_{\mu_1,\mu_2}(\lambda) = \frac{f_\mathrm{s}}{M} \frac{\sum\limits_{\mu=\mu_1}^{\mu_2} \widehat{\Phi}_{xx}(\lambda,\, \mu) \cdot \mu}{\sum\limits_{\mu=\mu_1}^{\mu_2} \widehat{\Phi}_{xx}(\lambda,\, \mu)}, \tag{3.35}$$

The factor f_s/M causes a conversion of the SSC from the discrete frequency-domain to a representation in Hz. The power spectral density (PSD) of a signal is defined as long-term expectation over all frames λ

$$\Phi_{xx}(\mu) = \underset{\lambda}{\mathrm{E}}\{|X(\lambda,\, \mu)|^2\}. \tag{3.36}$$

As for real-time applications it is not possible to compute the expectation over the whole signal length (i.e., all frames), an alternative approach for the estimation

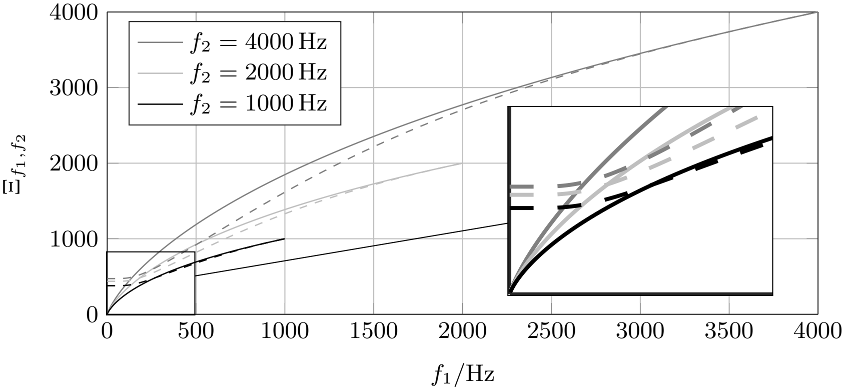


Figure 3.14: Signal centroids for theoretical energy distribution of speech (dashed lines) and wind noise (solid lines).

of the time-varying PSD in Equation 3.35 is given by the recursive smoothing approach

$$\widehat{\Phi}_{xx}(\lambda,\,\mu) = \alpha \cdot \widehat{\Phi}_{xx}(\lambda-1,\mu) + (1-\alpha)\cdot |X(\lambda,\,\mu)|^2, \tag{3.37}$$

where the smoothing constant α must be chosen in the range between 0 and 1 and controls the adaptation speed of the estimate.

A study of measured wind noise and speech SSCs for the frequency range up to $f_2 = 4000\,\mathrm{Hz}$ is given in Figure 3.15 using speech data from the TIMIT database [LKS89]. Here, 6 minutes of voiced speech segments are taken into account. Unvoiced speech segments are omitted, because they show only low energy in the considered frequency range, where wind noise is active (see, e.g., Figure 3.9) and would be treated as a speech pause for SSC computation in the described frequency range. To investigate the influence of wind, 6 minutes of recorded wind noise from [NV14b] are analyzed. Both signals are segmented into frames of 20 ms and the signal centroids are computed for every frame resulting in the depicted distributions. The wind noise as well as the speech SSCs show slightly higher values than the theoretically determined curves in Figure 3.14. These deviations can be explained by the non-continuous frequency resolution, which is necessary for the behavior derived from Equation 3.33. Nonetheless, a clear difference is visible between the speech and wind noise SSCs showing only a small overlap. Again, a wind indicator is desired, which takes only values in the range between 0 and 1. Setting $f_1 = 0\,\mathrm{Hz}$ leads to SSC values close to zero for wind noise, whereas speech will generate higher values with a theoretical maximum of f_2. The SSC-based wind indicator is finally defined as

$$\mathcal{I}_{\mathrm{SSC}}(\lambda) = \frac{f_2 - \Xi_{\mu_1,\mu_2}(\lambda)}{f_2} \quad \in [0,1], \tag{3.38}$$

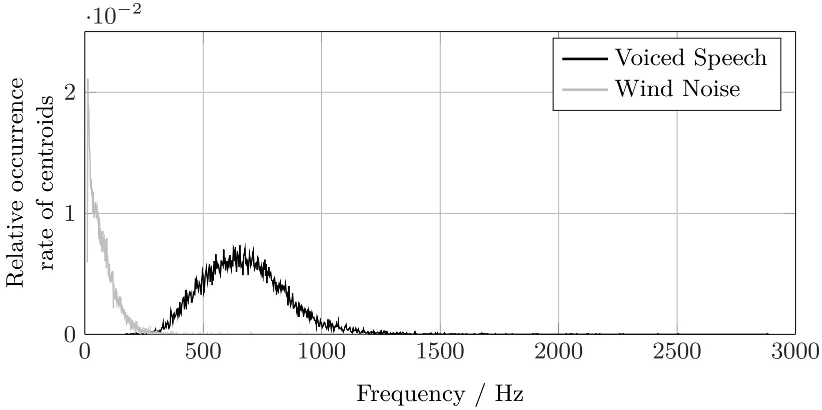


Figure 3.15: Distribution of speech and wind noise centroids.

using the discrete frequency computation of Equation 3.35.

3.5.2.3 Template Spectrum Combination

A different approach for the detection of wind noise is derived from a concept for noise estimation using codebooks with pre-trained speech and noise entries (see, e.g., [HNNV14]). The basic idea is that the noisy spectral magnitude $|X(\lambda,\,\mu)|$ can be decomposed into the speech template $|\widetilde{S}_i(\mu)|$ with the index i from a speech codebook and a noise template $|\widetilde{N}_j(\mu)|$ with index j from a noise codebook. Then, the template spectrum combination (TSC) of the noisy magnitude spectrum is approximated by

$$|\widehat{X}(\lambda,\,\mu)| = \sigma_{\mathrm{TSC}}(\lambda) \cdot |\widetilde{S}_i(\mu)| + (1 - \sigma_{\mathrm{TSC}}(\lambda)) \cdot |\widetilde{N}_j(\mu)|. \qquad (3.39)$$

Because all signals in Equation 3.39 tagged with the $\widetilde{\ }$-operator are normalized to a frame-energy of 1, the codebook weight $\sigma_{\mathrm{TSC}}(\lambda)$ takes values between 0 and 1. An extensive search is applied using all combination of codebook entries $\widetilde{S}_i(\mu)$ and $\widetilde{N}_j(\mu)$ and discrete values for the codebook weight σ_{TSC} for an estimation of the noise spectrum in [SSK07] or [HNNV14]. Here, a simplified procedure is applied to detect wind noise by using only a single representative for the speech and wind noise component. For the speech component $\widetilde{S}(\mu)$ the previously introduced long-term average speech spectrum (LTASS) of Equation 3.34 is used in a linear description, while the $1/f$-approximation represents the wind noise component $\widetilde{N}(\mu)$. As in Equation 3.22, a vector notation $\mathbf{X}(\lambda)$, $\widehat{\mathbf{X}}(\lambda)$, $\widetilde{\mathbf{S}}(\lambda)$, $\widetilde{\mathbf{N}}(\lambda)$ is employed to describe the magnitudes of the DFT coefficients in each frame λ. By minimizing the mean square error between a given input signal $\mathbf{X}(\lambda)$ and the estimate $\widehat{\mathbf{X}}(\lambda)$ defined in

Equation 3.39

$$||\mathbf{X}(\lambda)-\widehat{\mathbf{X}}(\lambda)||^2 = ||\mathbf{X}(\lambda)-\sigma_{\mathrm{TSC}}(\lambda)\cdot\widetilde{\mathbf{S}}(\lambda)-(1-\sigma_{\mathrm{TSC}}(\lambda))\cdot\widetilde{\mathbf{N}}(\lambda)||^2 \overset{!}{=} \min \tag{3.40}$$

an optimal template weight $\sigma_{\mathrm{TSC,opt}}$ can be derived by taking the derivative with respect to σ_{TSC} and setting the result to zero yielding in

$$\sigma_{\mathrm{TSC,opt}} = \frac{\widetilde{\mathbf{N}}^T\widetilde{\mathbf{N}} - \widetilde{\mathbf{S}}^T\widetilde{\mathbf{N}} + \mathbf{X}^T \cdot (\widetilde{\mathbf{S}} - \widetilde{\mathbf{N}})}{||\widetilde{\mathbf{S}} - \widetilde{\mathbf{N}}||^2}, \tag{3.41}$$

where the frame index λ is omitted for the sake of clarity. Since all quantities in Equation 3.39 are normalized to a frame-energy of 1, the template gain $\sigma_{\mathrm{TSC,opt}}$ indicates the amount of the speech component and $1-\sigma_{\mathrm{TSC,opt}}$ the amount of the wind noise component. Thus, the template weight can be used as wind detector according to:

$$\mathcal{I}_{\mathrm{TSC}}(\lambda) = 1 - \sigma_{\mathrm{TSC,opt}}(\lambda). \tag{3.42}$$

3.5.3 Performance of Single Microphone Wind Detection

For evaluation, noisy speech signals are first manually labeled to determine performance of the detection methods. Two sets of signal frames are defined as $\mathcal{M}_\mathrm{s}$ for frames containing only clean speech and $\mathcal{M}_\mathrm{w}$ including all frames with wind noise activity. These sets are displayed exemplary in Figure 3.16. Here, the speech and wind signals are depicted separately to clarify the beginnings and endings of the respective activity. In the evaluation process, only the superposition of both signals is considered.

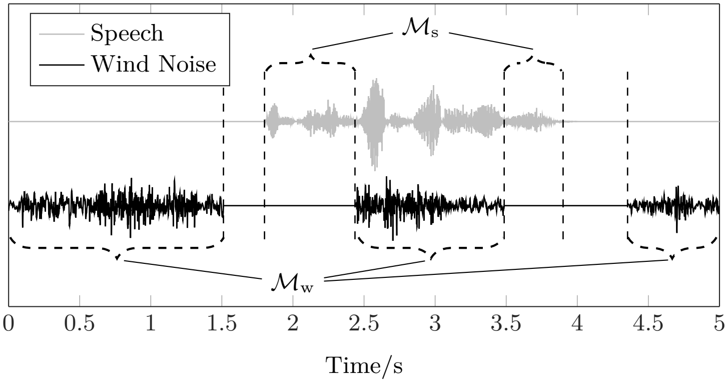


Figure 3.16: Example of speech and wind noise signals for the definition of the sets $\mathcal{M}_\mathrm{s}$ and $\mathcal{M}_\mathrm{w}$.

All described algorithms for the detection of wind noise are compared in the following by means of two measures. Firstly, the accuracy of the wind noise detection is measured by the wind detection rate

$$\mathcal{P}_{\mathrm{w}}(\zeta) = \frac{\#\{\mathcal{I}(\lambda) > \zeta\}}{\#\{\mathcal{M}_{\mathrm{w}}\}}, \quad \lambda \in \mathcal{M}_{\mathrm{w}} \tag{3.43}$$

where $\#\{\cdot\}$ denotes the cardinality, i.e. for the numerator in Equation 3.43 the number of elements in the considered set of frames in which the wind indicator $\mathcal{I}(\lambda)$ is greater than a threshold ζ. In a similar way the speech misdetection rate is defined by

$$\mathcal{P}_{\bar{\mathrm{s}}}(\zeta) = \frac{\#\{\mathcal{I}(\lambda) > \zeta\}}{\#\{\mathcal{M}_{\mathrm{s}}\}}, \quad \lambda \in \mathcal{M}_{\mathrm{s}}, \tag{3.44}$$

and counts the amount of clean speech, which is erroneously detected as wind noise. Both measures describes important performance properties of the wind detection. On the one hand, a high detection rate of wind noise is desired for a sufficient removal of the distortion in a subsequent step. But on the other hand, no clean speech segments should be detected as wind, which results in a low speech misdetection rate.

Both rates defined in Equations 3.43 and 3.44 are dependent on a threshold ζ, which is applied to the wind indicator. Since all wind detection methods result in an indicator between 0 (no wind) and 1 (wind active), a good comparison between the algorithms is given by passing through values between 0 and 1 and measuring the resulting detection rates. Taking both the speech misdetection rate and the wind detection rate at different thresholds into account, the so-called receiver operating characteristic (ROC) can be generated as depicted in Figure 3.17.

An evaluation was carried out taking randomly chosen speech sentences from the TSP database [Kab02]. The clean speech is mixed with wind noise segments from [NV14b] with duration between 0.3 and 3 s. The corresponding noisy speech signal is segmented into frames of 20 ms with an overlap of 10 ms, where in 70% of the frames wind is active and in 50% of the frames speech is active. Both speech and wind are active in about 30% of the frames. The speech and wind activity is manually labeled based on the clean speech and the pure wind noise signals to determine the sets $\mathcal{M}_{\mathrm{s}}$ and $\mathcal{M}_{\mathrm{w}}$, which are required for Equations 3.43 and 3.44. The global signal-to-noise-ratio (SNR) of the signal was -5 dB, which reflects a realistic situation (c.f., [NV14b]).

For each algorithm, a curve displays different operating points, which belong to certain values of the threshold ζ applied to the corresponding wind indicator. A good detection results in a high $\mathcal{P}_{\mathrm{w}}$ value and a low $\mathcal{P}_{\bar{\mathrm{s}}}$ value, as indicated by the arrows. The upper right corner of Figure 3.17 represents thresholds close to zero, while the lower left corner depicts thresholds close to one. Because some of the above mentioned approaches only take discrete values, e.g., a discrete frequency bin or a discrete number of zero-crossings, some of the curves show partially large gaps between the working points. The ROC can be roughly separated into two parts:

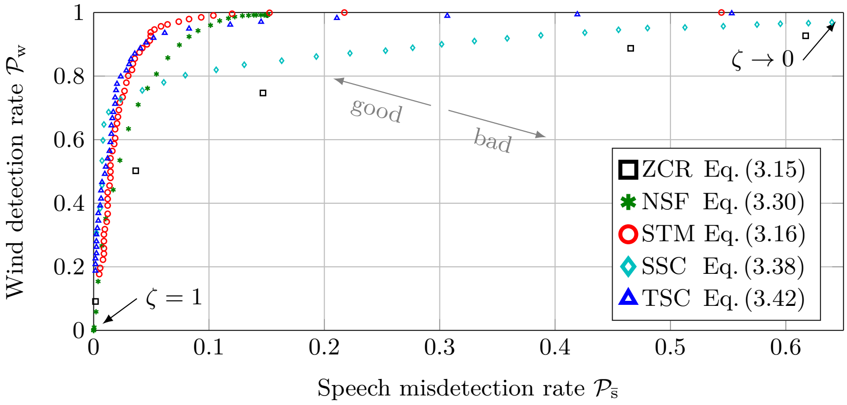


Figure 3.17: Receiver operating characteristic of single microphone various wind noise detection methods.

- The fast ascending section, where all algorithms show a low misdetection rate. Here, the centroid based method (SSC) and the template spectrum combination (TSC) show the best results.

- A section, where the detection rate rises slowly, but the misdetection increases. In this range, the detector resulting from the normalized short-term mean (NSTM) and the TSC method gives the best results.

The remaining two methods, zero crossing rate (ZCR) and negative slope fit (NSF), give only very inaccurate findings for all operating points.

In conclusion, the NSTM and the TSC methods presents the best trade-off between a low misdetection rate of speech and a high wind noise detection rate. If a really low misdetection rate is required the SSC concept outperforms the two aforementioned methods in some regions.

3.5.4 Dual Microphone Wind Noise Detection

Considering a system with two microphone signals as depicted in Figure 3.7, the correlation between the signals can be exploited for the detection of wind. The acoustic generation process of wind noise is given by turbulences, which are close to the microphones and can be seen as a vast number of independent sound sources for each microphone (cf. Section 3.3.4). Thus, a low correlation is assumed for wind noise. A speech signal is usually represented by a point source (neglecting reverberation effects), resulting in a high correlation.

3.5.4.1 Average Short-Term Coherence

Exploiting the correlation properties of speech and wind noise, the magnitude squared coherence (MSC) $\mathcal{C}(\lambda, \mu)$, introduced in Section 3.3.4, is applied and is defined as frame and frequency dependent quantity

$$\mathcal{C}(\lambda,\, \mu) = \frac{|\widehat{\Phi}_{xy}(\lambda,\, \mu)|^2}{\widehat{\Phi}_{xx}(\lambda,\, \mu)\widehat{\Phi}_{yy}(\lambda,\, \mu)}. \tag{3.45}$$

The required short-term estimates of the auto- and cross-PSDs $\widehat{\Phi}_{xx}(\lambda,\, \mu)$, $\widehat{\Phi}_{yy}(\lambda,\, \mu)$ and $\widehat{\Phi}_{xy}(\lambda,\, \mu)$ are computed via recursive smoothing as

$$\widehat{\Phi}_{xx}(\lambda,\, \mu) = \alpha \cdot \widehat{\Phi}_{xx}(\lambda - 1, \mu) + (1 - \alpha) \cdot X(\lambda,\, \mu) \cdot X^*(\lambda,\, \mu), \tag{3.46}$$

and

$$\widehat{\Phi}_{xy}(\lambda,\, \mu) = \alpha \cdot \widehat{\Phi}_{xy}(\lambda - 1, \mu) + (1 - \alpha) \cdot X(\lambda,\, \mu) \cdot Y^*(\lambda,\, \mu), \tag{3.47}$$

where $\{^*\}$ denotes the complex conjugate and $\alpha = 0.5$ is chosen. As depicted in Figures 3.8a and 3.8c speech shows a value close to one, while wind noise takes values close to zero. For the wind detection, only a single score in each frame is desired. Hence, the MSC $\mathcal{C}(\lambda,\, \mu)$ can be averaged over a specific frequency range to lower the variance. Using a frequency range in which mainly wind is assumed to be active, e.g., 0-500 Hz, the mean MSC value in this range $\bar{\mathcal{C}}_{0-500\,\mathrm{Hz}}$ is used as wind noise indicator

$$\mathcal{I}_{\mathrm{MSC}}(\lambda) = 1 - \bar{\mathcal{C}}_{0-500\,\mathrm{Hz}}(\lambda) = 1 - \frac{\sum\limits_{\mu=1}^{\mu=\mu_{500}} \mathcal{C}(\lambda,\, \mu)}{\mu_{500}}, \tag{3.48}$$

where μ_{500} is the frequency bin corresponding to a frequency of 500 Hz.

3.5.5 Performance of Dual Microphone Wind Detection

To evaluate the performance of the dual microphone wind noise detection, the experiment explained in Section 3.5.3 is carried out using a dual microphone recording of wind noise from [NV14b]. Again the wind detection rate $\mathcal{P}_{\mathrm{w}}$ and speech misdetection $\mathcal{P}_{\bar{\mathrm{s}}}$ rate are taken into account. For decreasing microphone distance the MSC of wind noise might show higher values, as indicated by Equation 3.11. Thus a smaller microphone distances exhibit the crucial scenarios for the dual microphone wind noise detection task. Therefore, a small microphone distance of 2 cm is considered here as test case. For the coverage and length of the speech and wind signal, the same parameters were chosen as for the single microphone case in Section 3.5. The results are again presented as ROC curves in Figure 3.18.

The dual microphone MSC based approach presented in Equation 3.48 is compared with the three single microphone methods, which gave the best results. It can be seen that the MSC method shows similar results with the best single microphone algorithm in each working point but does not lead to an increased performance. Only if a really high detection rate is required ($\mathcal{P}_\mathrm{w}$>0.99), the MSC method yields a slightly better performance. The expected gain of exploiting the low spectral correlation of two microphone signals is compensated by an effect from the calculation of the coherence. For the recursive smoothing defined in Equation 3.47, the fast changing characteristics of the wind noise signals are spread over time and thus a fast tracking of wind noise onsets and offsets in not guaranteed.

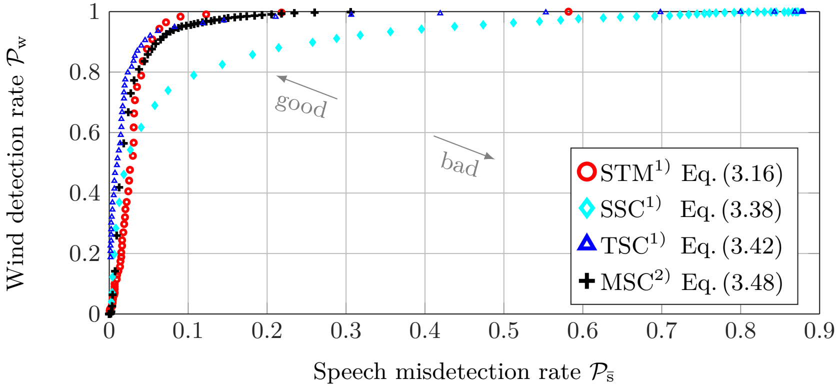


Figure 3.18: Receiver operating characteristic of wind noise detectors: $^{1)}$single microphone, $^{2)}$dual microphone.

3.6 Model for Wind Noise Generation

For the development and evaluation of algorithms that suppress wind noise, audio data of the noise signal is required. Because reproducible measurements of wind noise are difficult and costly, an approach is presented for simulating wind noise signals under precisely defined conditions. Considering simulations of windmill-powered plants or hazard assessment of wind sensitive structures, models were proposed, which predict time series of the wind speed. The derived models are based on measurement data and presented, e.g., in [JL86] and [SS01]. They consider a coarse temporal resolution of hourly wind speed data. In the case of audio signal processing, a considerably finer temporal resolution of the wind noise model is required, which will be derived in this section.

Based on the investigations from the previous sections, a model is proposed, which generates an artificial wind noise signal with pre-defined features [NV14b]. It should be mentioned that the derived model does not reflect the physics of wind noise generation. Primarily, the aim is to provide signals with similar statistics and spectral characteristics as recorded wind noise. A block diagram of this model is depicted in Figure 3.19 and can be divided into three stages:

1. generation of an excitation signal $e(k)$,
2. weighting with a time-dependent gain $g(k)$ yielding the weighted excitation $\tilde{e}(k)$,
3. filtering with $A(z)$, which adapts the spectral shape of the synthesized wind noise signal $n_{\mathrm{syn}}(k)$.

The explanation of all three stages follows in Sections 3.6.1 - 3.6.3.

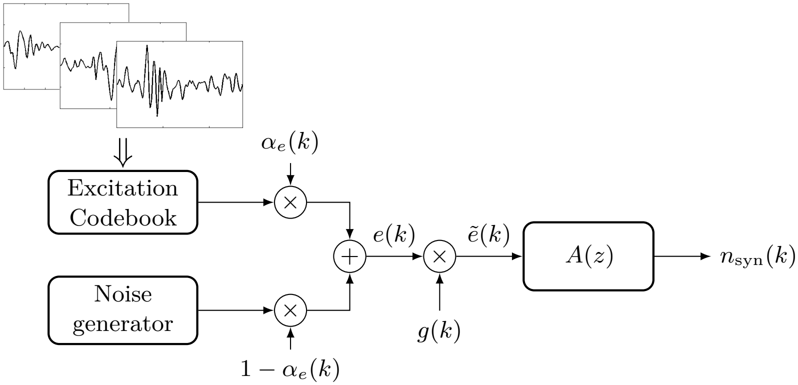


Figure 3.19: Wind noise model proposed in [NV14b].

3.6.1 Modeling the Temporal Characteristics

Regarding the acoustic signal, which is generated by wind in a device equipped with one microphone, a two-sided consideration of the temporal characteristics is necessary. In a long-term sense of several seconds, the noise is determined by the current wind speed in close proximity to the device. Due to shadowing effects, the local wind speed is not always equal to the global wind speed in a free-field scenario but both wind speeds are usually highly correlated. A closer look provides the short-term behavior of the wind noise signal considered in 20 ms frames, where the sound is dominated by the turbulences in the air stream. The turbulences can be close to the microphone, resulting in the low-frequency rumbling sound or in further distance yielding in more constant noise level.

Both aspects are illustrated in Figure 3.20 in terms of the (short-term) frame energy and the (long-term) smoothed version of the frame energy of a measured wind noise signal. The gain $g(k)$ in the proposed model shown in Figure 3.19 controls the temporal characteristics of the generated wind noise signal with respect to both the long-term and the short-term behavior. In [NV14b], it was proposed to determine one gain, which models both the long-term and the short-term. An

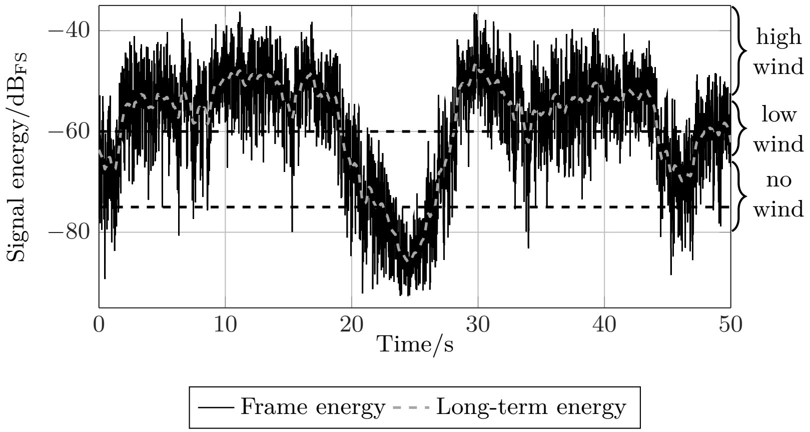


Figure 3.20: Classification of wind noise signal.

advanced way in modeling both characteristics is given by a decomposition of the gain into a product

$$g(k) = g_{\mathrm{ST}}(k) \cdot g_{\mathrm{LT}}(k) \tag{3.49}$$

of a short-term gain $g_{\mathrm{ST}}(k)$, which is combined with the long-term gain $g_{\mathrm{LT}}(k)$ by multiplication.

Long-term Gain

The long-term energy is determined by the current wind speed generating the acoustic signal. Usually, the wind speed is rising during a wind gust continuously to a high level and is then falling again. A wind gust may last a time span below one second, but usually takes several seconds. The long-term behavior is exemplified by the smoothed frame energy shown by the dashed gray line in Figure 3.20. It is calculated by a recursive smoothing of the frame energy with a smoothing constant of $\alpha = 0.99$.

The temporal progress of the LT energy of measured wind noise in Figure 3.20 can roughly be divided into three classes. In the first case the measured noise results from flow sound not generated in the vicinity of the microphone (*low wind*). When a wind gust arises, the sound level suddenly rises due to turbulences close to the microphone position (*high wind*). A third case is given in the absence of wind (*no wind*). The three classes can be seen as three discrete states of a Markov model reflecting different wind conditions. Similar models were derived for the long-term behavior of the wind speed in [JL86] or [SS01]. The 3-state model depicted in Figure 3.21 is used in the following to model the long-term temporal characteristics of wind noise.

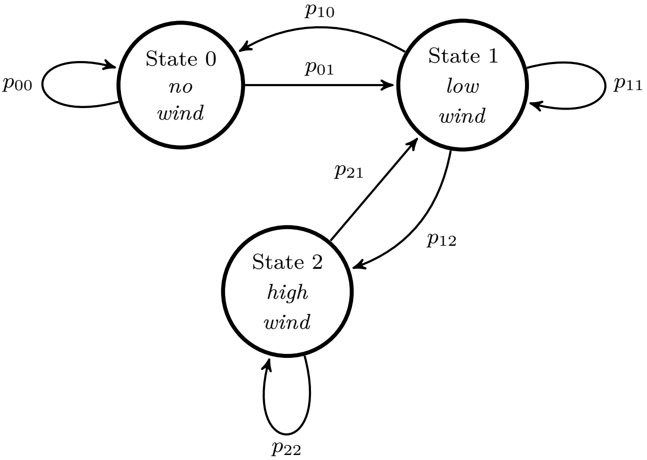


Figure 3.21: 3-state Markov model.

The transition probabilities of the model are given by p_{ij} from state i to state j. The probabilities p_{ii} determine the duration and occurrence rate of the corresponding wind condition in state i. For the provided model it was assumed that state 1 (*low wind*) is always the transition between *no wind* and *high wind*. Therefore, the transition probabilities p_{02} and p_{20} were set to zero and are not depicted in Figure 3.21. The remaining transition probabilities can be trained by wind noise measurements. This is done by first labeling ranges of *no*, *low* and *high wind* in a given signal as exemplified in Figure 3.20 and compute the corresponding probabilities afterwards. The thresholds defining the ranges of the wind noise activity must be chosen manually and are -60 $\mathrm{dB_{FS}}$ and -75 $\mathrm{dB_{FS}}$ for the considered wind recordings as depicted in Figure 3.20.

The gain $g_{\mathrm{LT}}(k)$ for the long-term behavior is then calculated by using the trained Markov model, which produces a sequence of states $s_{\mathrm{seq}}(\lambda)$ to control the wind noise activity in each frame λ. Based on this sequence a gain value is assigned to each state s_i, which is previously determined by the mean values gained from the corresponding states of the wind noise measurements. The resulting values only consist of three discrete values (s_0, s_1, s_2). In order to smooth the sudden changes of the gain values, the gain sequence is calculated by convolution with a Hann window $h_{\mathrm{smooth}}(\lambda)$ creating the frame dependent long-term gain

$$g_{\mathrm{LT}}(\lambda) = \sum_{\kappa=0}^{M} h_{\mathrm{smooth}}(\kappa) \cdot s_{\mathrm{seq}}(\lambda - (\kappa - M/2)), \tag{3.50}$$

where the length M of $h_{\mathrm{smooth}}(\kappa)$ corresponds to 0.5 seconds. The values of the gain sequence $s_{\mathrm{seq}}(\lambda)$ must reflect the average energy relation between the different states defined for the Markov model. Therefore, the for the three states the values $s_0 = 0$, $s_1 = 0.1 \mathrel{\widehat{=}} -20\,\mathrm{dB}$, and $s_2 = 1 \mathrel{\widehat{=}} 0\,\mathrm{dB}$. This also reflects the different levels of the long-term energy depicted in the wind noise segment in Figure 3.20. The sample-wise long-term gain $g_{LT}(k)$ takes the constant value during each frame, which is determined by Equation 3.50.

Short-term Gain

While the long-term gain primarily controls the presence and absence of wind noise, the instantaneous signal level is simulated by the short-term gain $g_{\mathrm{ST}}(k)$. As explained in Section 3.3.2 the short-term energy $E_{\mathrm{ST}}(\lambda)$ of one frame shows high variation over time, which is characteristic for wind noise. This behavior is modeled by the short-term gain $g_{\mathrm{ST}}(k)$.

First, the statistics of $E_{\mathrm{ST}}(\lambda)$ are analyzed. For the long-term measurement of wind speed, statistical models were derived in [SL00] and [LL00], which assume that the wind speed data can be approximated by a Weibull distribution [Wei51]. The corresponding probability density function (PDF) of the wind speed U is expressed as

$$p_{\mathrm{W}}(U) = \begin{cases} \left(\frac{\kappa_{\mathrm{W}}}{\lambda_{\mathrm{W}}}\right)\left(\frac{U}{\lambda_{\mathrm{W}}}\right)^{\kappa_{\mathrm{W}}-1} \exp\left[-\left(\frac{U}{\lambda_{\mathrm{W}}}\right)^{\kappa_{\mathrm{W}}}\right] & \text{, if } U \geq 0 \\ 0 & \text{, else} \end{cases} \tag{3.51}$$

with the shape parameter κ_{W} and the scale parameter λ_{W}. A *maximum likelihood estimation* of the two parameters is given using the following equations [SL00]:

$$\kappa_{\mathrm{W}}(m+1) = \left(\frac{\sum_{i=1}^{N}(U_i)^{\kappa_{\mathrm{W}}(m)}\log(U_i)}{\sum_{i=1}^{N}(U_i)^{\kappa_{\mathrm{W}}(m)}} - \frac{\sum_{i=1}^{N}\log(U_i)}{N}\right)^{-1}, \tag{3.52}$$

$$\lambda_{\mathrm{W}} = \left(\frac{1}{N}\sum_{i=1}^{N}(U_i)^{\kappa_{\mathrm{W}}}\right)^{1/\kappa_{\mathrm{W}}}, \tag{3.53}$$

where U_i is the observed wind speed in time step i of N non-zero data points. Equation 3.52 must be solved iteratively and $\kappa_{\mathrm{W}} = 2$ is proposed in [SL00] as a suitable initialization for the first iteration $m = 0$. Thereafter, Equation 3.53 can be solved explicitly by inserting the found κ_{W}. All the aforementioned models are based on n long-term wind observations such as hourly averaged measurements (e.g., 72 measurements in [SL00]). For the proposed approach clearly, a shorter time duration is of interest such as the frame energy $E_{\mathrm{ST}}(\lambda)$ of 20 ms segments as investigated in Equation 3.4 in Section 3.3.2 as the set of N data points.

As mentioned in the beginning of this chapter, the acoustic sound levels generated by wind are related to its speed. For the purpose of modeling the short-term characteristics, the distribution of the short-term gain g_{ST} is of interest, which is related to the frame energy $E_{\mathrm{ST}}(\lambda)$ of the excitation signal $\tilde{e}(k)$ as

$$g_{\mathrm{ST}}(\lambda) = \sqrt{\frac{E_{\mathrm{ST}}(\lambda)}{\sum\limits_{k=0}^{L_{\mathrm{F}}-1} e^2(k)}}, \tag{3.54}$$

where L_{F} is the frame length. With the assumption of an energy normalized excitation signal $e(k)$, the significant relation is given by $g_{\mathrm{ST}} \sim \sqrt{E_{\mathrm{ST}}}$. Due to the known quadratic relation between wind speed and energy, i.e., $E_{\mathrm{ST}} \sim U^2$, it can be concluded, that the the short-term gain g_{ST} is linearly depending on the wind speed U and $\sqrt{E_{\mathrm{ST}}}$. A histogram of measured $\sqrt{E_{\mathrm{ST}}}$ values is given in Figure 3.22.

For the detected distribution, signal segments with no signal energy, i.e., in wind pauses, are excluded. These conditions are modeled by the long-term gain regarding the *no wind* case. Additionally, the PDF of a Weibull distribution is displayed by the dashed black curve. The parameters λ_{W} and κ_{W} were computed using the calculation instructions from Equations 3.52 and 3.53. Comparing the histogram data and the Weibull distribution, it is evident, that the PDF provides a sufficient approximation of the wind noise energy even on shorter time scale than in [SL00] and [LL00].

For the generation of the short-term gain g_{ST} in each frame the so-called inverse transform technique is applied, which adapts a uniform distributed variable to a given PDF, if the inverse of the cumulative distribution function (CDF) exists, see [Dev86]. The CDF of the Weibull distribution reads

$$P_{\mathrm{W}}(U) = \begin{cases} 1 - e^{-(U/\lambda_{\mathrm{W}})^{\kappa_{\mathrm{W}}}}, \text{ if } U \geq 0, \\ 0 \text{ , else,} \end{cases} \tag{3.55}$$

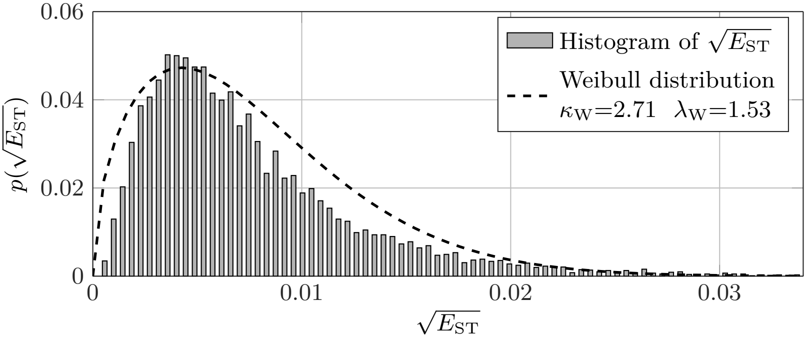


Figure 3.22: Distribution of wind noise energy $\sqrt{E_{\mathrm{ST}}}$.

which is invertible and can be applied to produce a random variable with a Weibull distribution.

An example of the temporal progress of simulated gains is depicted in Figure 3.23. The black curve represents the long-term behavior while the modulated version of the short-term gain $g_{\text{LT}}(k) \cdot g_{\text{ST}}(k)$ is shown by the thinner gray curve. The parameters from Figure 3.22 for the Weibull distribution are applied and the used transition probabilities of the Markov model are given later in the evaluation of the in Section 3.6.4.

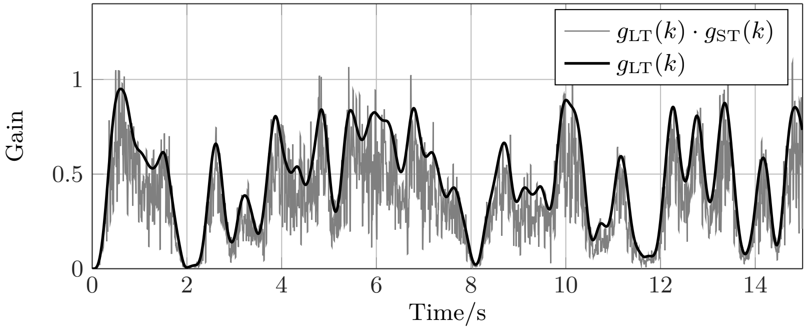
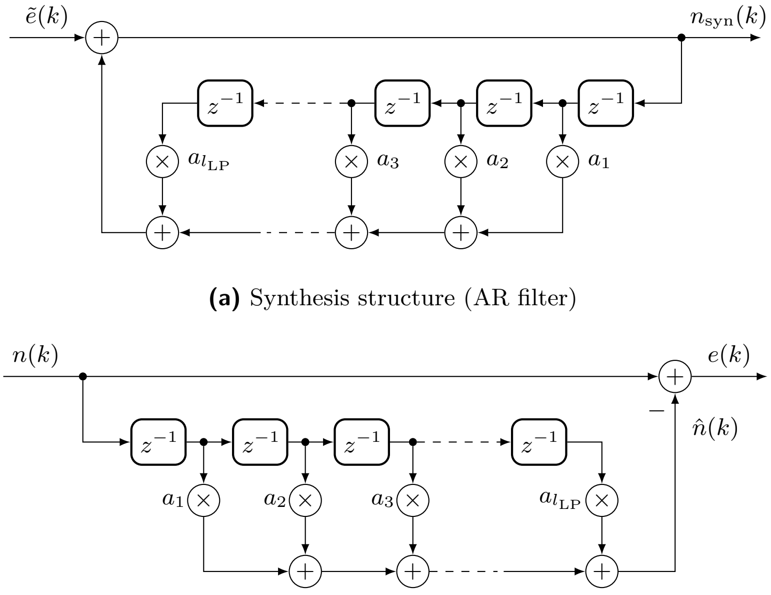


Figure 3.23: Simulated long-term and short-term gain.

3.6.2 Modelling the Spectral Characteristics

A common description for correlated time series, such as digitized audio signals, is given by an auto-regressive (AR) process (see, e.g., [Dur60]). For the proposed realization of the wind noise, an AR model is applied in terms of the all-pole filter $A(z)$. This filter controls the spectral envelope of the generated noise signal $n_{\text{syn}}(k)$. The basic structure of an AR process of order l_{LP} is shown in Figure 3.24a, where the excitation signal $\tilde{e}(k)$ is recursively filtered by the coefficients of $a_1 \ldots a_n$. The value of the coefficients defines the spectral behaviour of the synthesized noise signal $n_{\text{syn}}(k)$. E.g., in the case of linear predictive coding (LPC) the coefficients determine the position and shape of the formants of a coded speech signal (see, e.g., [VM06]).

In general, there are multiple approaches to estimate the coefficients a_i of the filter $A(z)$. All these methods are based on the analysis structure given in Figure 3.24b. The optimal coefficients a_i are chosen, such that the power of the error signal $e(k)$ between the given signal $n(k)$ and the estimated version $\hat{n}(k)$ is minimized. In speech coding usually a block-wise adaptation of usual frame-sizes between 10 and 30 ms is applied using the *auto-correlation method* or the *covariance method* [VM06]. Because of the fast changing signal characteristics of wind noise, an estimation method with a finer temporal resolution is chosen here, which is



(a) Synthesis structure (AR filter)

(b) Analysis structure

Figure 3.24: Filter structures for linear predictive coding.

given by a sequential adaptation using the *normalized least-mean-square* (NLMS) algorithm [Hay96]. With the notation of the signal vector

$$\mathbf{n}(k-1) = [n(k-1), n(k-2), \ldots, n(k-l_{\mathrm{LP}})]^T \tag{3.56}$$

and the coefficient vector

$$\mathbf{a}(k) = [a_1(k), a_2(k), \ldots, a_{l_{\mathrm{LP}}}(k)]^T, \tag{3.57}$$

this method provides a sample-wise calculation of a_i using the following update rule:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + 2 \cdot \vartheta \frac{e(k)\mathbf{n}(k-1)}{||\mathbf{n}(k-1)||^2}, \tag{3.58}$$

where the error signal is calculated as

$$e(k) = n(k) - \mathbf{a}^T(k)\mathbf{n}(k-1). \tag{3.59}$$

The adaptation speed is controlled by the step-size constant ϑ, which must be limited to the range

$$0 < \vartheta < 1 \tag{3.60}$$

for stability reasons.

The NLMS algorithm is applied to estimate the coefficients describing the spectral shape of wind noise using wind noise recordings as input signal $n(k)$. To prevent a wrong adaptation in periods without any wind noise in the recordings, Equation 3.58 is modified to

$$\mathbf{a}(k+1) = \mathbf{a}(k) + 2 \cdot \vartheta \frac{e(k)\mathbf{n}(k-1)}{||\mathbf{n}(k-1)||^2 + \varepsilon}, \tag{3.61}$$

where ε avoids a division by zero in case of absence of wind in the considered signal samples. As investigated in Section 3.3.3, wind noise is identified as a low frequency signal with a distinct spectral shape, which is similar to a $1/f$-slope (see Equation 3.3). If the spectral characteristics of the simulated wind noise are determined by an AR filter, two steps are necessary:

- choice of sufficiently high order l_{LP} of $A(z)$ and
- determination of the values of the coefficients a_i.

A measure for the quality of the analysis structure in Figure 3.24b is given by the prediction gain, which is determined by the relation between the input signal power and the error signal power

$$G_{\mathrm{P}} = 10 \log_{10} \frac{\mathrm{E}\{n^2(k)\}}{\mathrm{E}\{e^2(k)\}}. \tag{3.62}$$

The higher the prediction gain the better the AR filter approximates the input signal $n(k)$. Results from experiments with different orders of the filter in the analysis structure are presented in Figure 3.25 for both, the speech signals and the recorded wind noise signals. Most striking is the extensively higher gain for wind noise signals. This property can be explained by the high energy at low frequencies for wind, which might lead to a distinct DC in short signal segments. Such DC can

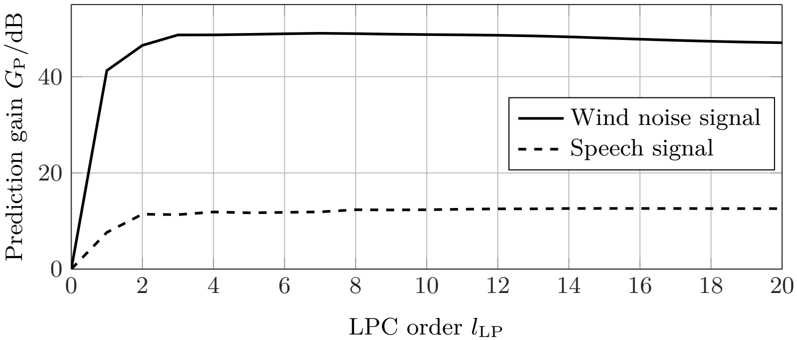


Figure 3.25: Prediction gain of LPC analysis.

already be removed by LP filter of order 1, which is evident by the prediction gain of over 40 dB for wind noise. This feature is implicitly exploited for the NSTM wind noise detector.

Furthermore, a saturation of the gain is reached for filter orders $l_{\mathrm{LP}} > 4$ for both signals with the considered sampling frequency of 16 kHz. Hence, no improvement in approximating the spectral shape of wind noise with the filter $A(z)$ can be expected by choosing a higher prediction order than 5. In the case of wind noise for prediction orders greater than 10, the prediction gain even slightly decreases again. It is assumed that fast changes in the wind noise contradict with a longer constraint length of the analysis filter. Thus, a higher order leads to erroneous prediction and a lower prediction gain.

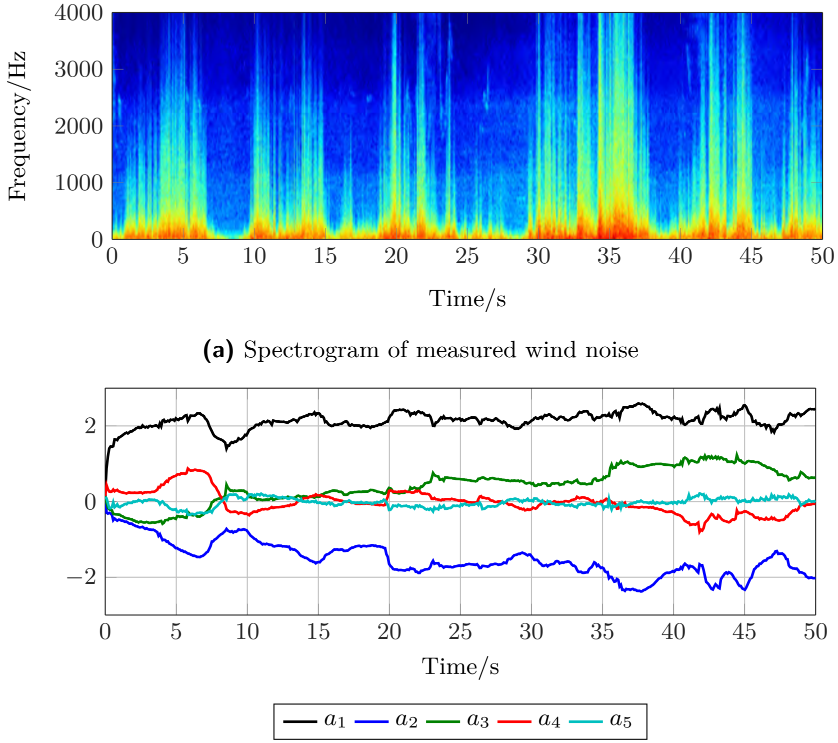
After setting the prediction order to 5 the values of the coefficients a_i have to be determined. The results of a sequential LP analysis of 50 seconds of recorded wind noise with $\vartheta = 0.1$ and $\varepsilon = 10^{-5}$ are shown in Figure 3.26. After a short settling process in the beginning the coefficients show only small variations over time. Besides periods without or with low wind noise ($t = 8\,\mathrm{s}$), the coefficients take almost constant values for long periods with wind activity. Considering the curves depicting the coefficients of highest order a_4 and a_5, it is evident that they only take small values close to zero. This observation supports the assumption that a low model order of 5 is sufficiently high as a representation.

The most simple way to realize $A(z)$ for the proposed model, is to use a fixed set of prototypical coefficients, which results in a constant shape of the spectral envelope, which could be measured in Figure 3.26. This concept is applied in the following and small variations of the spectral characteristics, as usually observed in wind noise signals, can also be generated by the excitation signal, which will be described in the following section in more detail.

3.6.3 Excitation Generation

The linear prediction (LP) coefficients determine the filter $A(z)$ in Figure 3.19 and the gain $g(k)$ controls the energy of the synthesized signal over time. In this way $a_1 ... a_5$ define the spectral shape of the produced signal $n_{\mathrm{syn}}(k)$. An easy way to produce the synthetic wind noise would be to use a white noise process as excitation $e(k)$. After filtering with the AR filter and weighting with the gain function the resulting signal has the same spectral and temporal characteristics as measured wind noise. But the synthetic noise does not reflect the characteristic listening impression of a real wind noise signal, especially in the *high wind* segments. For an entirely theoretical examination this property would not constitute a problem as long as the statistical characteristics are modeled, e.g., to test the performance of a speech enhancement system by objective quality measures.

As the generated signal should also be usable for human listeners, e.g., in a listening test, a natural sound of the synthetic wind signal is desired. For the proposed system the natural sound is realized by choosing excitation sequences from real recordings. These are also approximately spectrally flat and thus do not



(a) Spectrogram of measured wind noise

(b) LP coefficients

Figure 3.26: Sequentially estimated predictor coefficients of wind noise using an AR process of 5th order.

influence the spectral characteristics of the generated signal. For the proposed model, the error signal $e(k)$, which emerges during the sequential estimation process of the LP coefficients (see Figure 3.24b) is segmented and stored in a codebook as depicted in Figure 3.19. From this pre-trained codebook sequences are randomly chosen.

While for the *high wind* case the aforementioned excitation signal leads to a realistic sound, the *low wind* case is characterized by a rather noise-like signal as it is given by a spectral shaping of a white noise signal. This behavior is controlled by the parameter $\alpha_e(k)$ dependent on the current state of the Markov model in Section 3.6.1. In the *high wind* case a value close to one is favorable (e.g., $\alpha_e(k) = 0.9$) while in the *low wind* case a lower value should be chosen (e.g., $\alpha_e(k) = 0.1$). By this process for the excitation signal generation a more natural sound is produced with a very similar listening impression of the synthetic wind noise as recorded wind noise signals.

3.6.4 Validation of the Model

In this section an investigation of the simulated wind noise signal of the proposed model is carried out. Therefore, the temporal and spectral characteristics are compared to the results investigated in Sections 3.3.2 and 3.3.3. The model is implemented as proposed in Sections 3.6.1-3.6.3 using a model order of 5 for the AR filter and a fixed set of coefficients. The coefficients set was chosen to the values given in Table 3.3. These values are determined by averaging[6] the estimated coefficients in signal segments with wind activity.

a_1	a_2	a_3	a_4	a_5
2.24	-1.81	0.72	-0.131	-0.03

Table 3.3: Fixed LP coefficients for the wind noise synthesis.

The codebook is derived by taking segments of 5-10 ms from the error signal. As proposed in [NV14b] a codebook size of 140 sequences is sufficient to generate a wind noise signal with a natural sound. The transition probabilities of the Markov model p_{ij} are trained from manually labeled wind noise signals in terms of their long term energy. The values applied in the considered implementation are presented in Table 3.4.

			$i \rightarrow$	
	p_{ij}	0	1	2
j $\downarrow$	0	0.99991	$8.0037 \cdot 10^{-5}$	0
	1	$3.36740 \cdot 10^{-5}$	0.99974	$2.26097 \cdot 10^{-4}$
	2	0	$2.08928 \cdot 10^{-4}$	0.99979

Table 3.4: Transition probabilities p_{ij} between the states of the Markov model.

An example of a synthesized wind noise signal is presented in the spectrogram in Figure 3.27. Comparing it with measured wind noise (e.g., Figure 3.26a), similar characteristics are clearly visible. This applies for the non-stationary behavior as well as the low-frequency nature of the signal.

An experiment is carried out measuring the short-term variance $\overline{\sigma^2}_{E,\mathrm{ST}}$ as proposed in Section 3.3.2. The model is used to create 200 synthetic wind noise signals with a length of 50 seconds. For each sample signal the $\overline{\sigma^2}_{E,\mathrm{ST}}$ is measured in segments with wind activity as proposed in Equation 3.5. The measured $\overline{\sigma^2}_{E,\mathrm{ST}}$

[6] Since the averaging process of LPC coefficients can lead to unstable filter structures, the averaging is carried out in the line spectral frequency (LSF) domain.

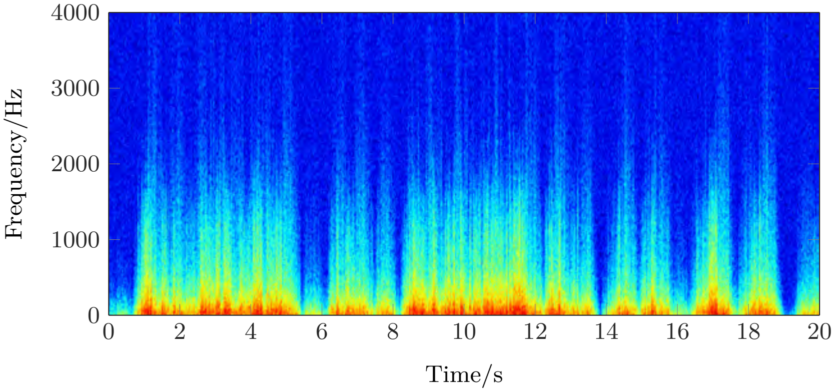


Figure 3.27: Spectrogram of synthesized wind noise.

values range approximately between 10 and 14 dB as shown in Figure 3.28. The mean $\overline{\sigma^2}_{E,\mathrm{ST}}$ of the synthetic wind noise is 12.41 dB, which is consistent with the mean $\overline{\sigma^2}_{E,\mathrm{ST}}$ of 12.23 dB measured for real wind noise signals. This measure indicates that the short-term variations over time of the simulated noise is similar to the measured wind noise.

A closer look at the spectral energy distribution is given by Figure 3.29, where the long-term spectrum of both measured and simulated wind noise is depicted. The black curves correspond to the simulated wind noise while the gray curves show the measured wind noise. Besides the spectra given by the solid lines, the smoothed spectra are also depicted by the dashed lines. These two curves bear a high amount

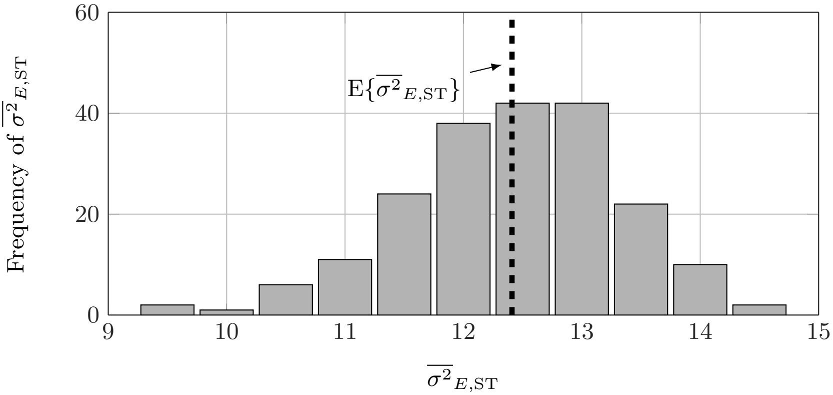


Figure 3.28: Distribution of the short-term variance $\overline{\sigma^2}_{E,\mathrm{ST}}$.

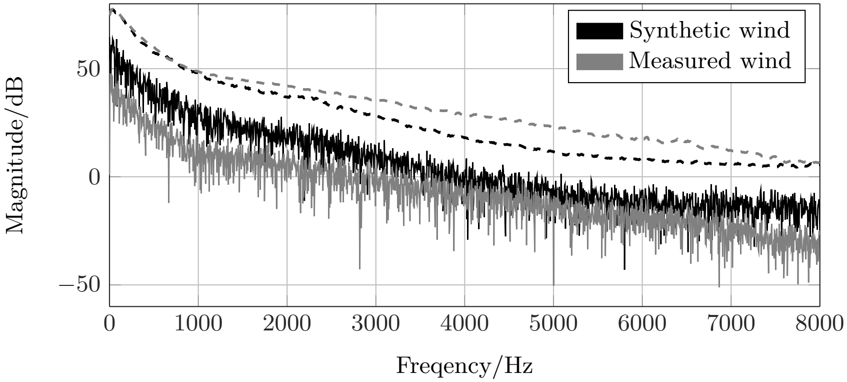


Figure 3.29: Spectral energy distribution of measured and simulated wind noise (dashed curves show smoothed progress of solid curves and are shifted on the y-axis for a better clarity).

of resemblance. This is especially true for frequencies below 1000 Hz, where most of the noise energy is distributed (see Figure 3.5b). For a better comparability, the curves shown in the figure are shifted with respect to their magnitude on the y-axis.

The investigation of the proposed wind noise model showed that the temporal and spectral characteristics of measured wind noise can be well approximated by a synthetic signal. The main parameters of the model determine the distribution of the short-term energy, the transition probabilities of the states of the Markov model and the coefficients controlling the spectral shape. These quantities are trained using recordings and can be adopted. For different use-cases, it can be useful to adjust the model to other applications by re-training the parameters based on different recordings. From informal listening tests the synthetic wind noise signal manifests a natural sound similar to wind noise recordings. This is achieved by applying excitation segments from the LPC analysis of real wind noise signals.

3.7 Conclusions

This chapter introduces the special characteristics of wind noise signals. The target is to point out significant differences between wind noise and other background noise types, which are usually assumed in the context of speech enhancement. First, the single microphone statistics in time- and frequency-domain representations of the signal are investigated. It turns out that the low-frequency shape of the spectrum can be roughly described by an $1/f$-decay over frequency f or more precisely by $1/f^{\nu}$ with the shape parameter ν. The temporal progress of the signal energy shows a considerably higher variation than other background noise types.

The next sections deals with detection of short segments in speech signal, which are degraded by wind noise. Different methods from literature are investigated and novel algorithms are proposed for both the single and dual microphone case. Two newly developed methods achieves similar high detection rates with ensuring a low false alarm rate for speech signals. These are the approaches, which exploit the normalized short-term mean (NSTM), and the technique based on separation of the noisy spectrum by a speech and wind template spectrum TSC, where the NSTM method is distinguished by its simplicity.

Based on the results from the wind noise analysis, a model is proposed for the generation of a synthetic wind noise signal. The temporal properties are separated into a long-term and a short-term gain controlling the energy of the generated signal. For the long-term gain, a Markov model with three states is applied. This long-term gain is mainly responsible for the absence or presence of wind noise. The typical fast variations of the signal are generated by the short-term gain. It has been shown that a random process following a Weibull distribution yields in a good emulation of the temporal progress of the signal. An auto-regressive filter is used to adjust the spectral energy distribution of the wind noise signal. For the proposed model a fixed choice of linear prediction coefficients shows a sufficient approximation of the distinct spectral shape. All parameters can be adapted to fit the synthesized signal to a given application, e.g., a certain microphone type or recording device. This model presents a valuable tool in the development process of wind noise reduction systems, as it provides precisely defined and repeatable test conditions.

Chapter 4

Wind Noise Reduction

This chapter deals with the reduction of wind noise in a captured speech signal. As discussed in Section 2.4, algorithms for background noise reduction, which are stated in this work as "conventional" methods, can not provide a sufficient wind noise reduction, if the speech signal is recorded in the presence of wind. Because wind noise is a severe problem, when mobile phones, microphones or hearing aids are used outdoors, special algorithms must be developed to combat the annoying disturbance of wind noise in the recorded signal.

As presented in Chapter 2, all considered methods for real-time noise suppression can be described by an analysis-synthesis structure. The modification and thus the enhancement of the speech signal is realized in the short-term discrete Fourier transform (DFT) domain. The most crucial part of the wind noise reduction is the detection step in Figure 2.2, which is usually realized as estimation of the current noise spectrum or short-term power spectral density (PSD). Several algorithms were presented in the past to estimate the background noise. Most prominent are the Minimum Statistics approach by Martin [Mar01], the minimum mean square error (MMSE) based noise PSD tracker by Hendriks et al. [HHJ10] and the SPP based noise estimator proposed by Gerkman and Hendriks [GH11]. In the last years particular algorithms were developed for the estimation of wind. In single microphone systems [KMT$^+$06], [HWB$^+$12] will be set as state-of-the-art methods. Considering devices equipped with more than one microphone, solutions can be found in [Elk07] and [FB10]. For both microphone configurations, more advanced methods are derived and will be presented in this thesis. The relevant publications can be found in [NCBV14, NV14a, NV15].

A widely used approach applies a time-varying spectral gain to the input spectrum to reduce the noise. Early solutions were proposed by Lim and Oppenheimer in terms of the well-known Wiener filter [LO79] and by Boll in terms of the spectral subtraction [Bol79] as explained in Section 2.3.3. Several publications can be found, which take into account *a priori* knowledge about speech and noise statistics. Exploiting the spectral statistics within a single frame, assumptions about the distribution of noise and speech discrete Fourier transform (DFT) coefficients can be made (see, e.g., [Lot04] or [Mar05]). A further improvement can be made by exploiting the temporal correlation between successive frame as it was shown by Esch in [Esc12]. Because these modifications to the spectral gain computation are based on stationary statistics of general background noise signals, it is not reasonable to apply them for the reduction of non-stationary wind noise. In this thesis well approved gain calculation rules given by the Wiener filter and the spectral subtraction and modified versions of these are applied for wind noise reduction.

Furthermore, an innovative approach for speech enhancement is presented in this work. Instead of a spectral weighting of the input signal, the clean speech is estimated using a model for synthesizing speech components. A widely known model for the process of speech generation is the so-called source-filter model ([RS78], [VM06]). The basic idea is to divide a speech signal into an excitation signal and a digital filter simulating the influence of the vocal tract. Many applications of this model can be found in the context of speech enhancement and most of the current and past speech codecs are based on this model (see, e.g., [Chu04], [VM06]). It will be shown that especially in the case of wind noise this model can be helpful for improving the processed speech [NNJ+12], [NNV15].

The remainder of this chapter is organized as follows. First, a short overview over acoustical countermeasures against the formation of wind noise before degrading the recorded signal will be given in Section 4.1. Because the focus of this work is the enhancement of speech signals by means of digital signal processing, this section gives only a brief insight in the mechanisms of wind shields. Considering a digital representation of the input signals, the following sections deal with the estimation and reduction of wind noise in a noisy speech signal. In Section 4.2, procedures for the estimation of the wind noise short-term power spectrum (STPS) using a single microphone are presented. A review on existing methods is given followed by the presentation of two advanced new concepts. Based on the wind noise estimate the subsequent spectral weighting is explained in Section 4.3. A dual microphone configuration is considered in Section 4.4 for the estimation and reduction of wind noise. In Section 4.5 the new concept for wind noise reduction incorporating a speech synthesis module is presented. Finally, conclusions are drawn at the end of this chapter.

4.1 Acoustical Countermeasures

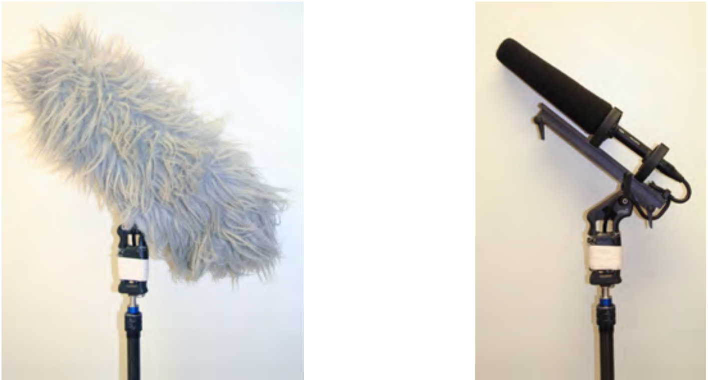
Besides techniques introduced in this thesis, which try to reduce the effect of wind noise by means of signal processing, many acoustical countermeasures exist to overcome the problem of wind for outdoor recordings. Mostly, this is realized by windscreens, where two types exist (see, e.g., [Wut92]):

(a) basket-style windscreens,

(b) foam windscreens.

The two concepts are shown in Figure 4.1 and their basic goal is to prevent the full velocity and the produced turbulences of the wind stream to reach the microphone. For both constructions the shape should be aerodynamically, because the windscreen itself should not introduce any further turbulences.

The concept of the basket-style version is, that an open frame is mounted around the microphone, which is covered with one or more layers of cotton, fine-mesh



(a) Basket-style windscreen [Wik06a] **(b)** Foam windscreen [Wik06b]

Figure 4.1: Different designs of windscreens.

fabric or fur (see Figure 4.1a). This frame encloses a volume of air around the microphone, which should be effected by the wind only to a little amount. The trapped air volume inside the basket-style windscreen can influence the frequency response especially at higher frequencies, where standing waves might affect the transfer behavior and directivity of the enclosed microphone.

The second type, represented by solid foam wind screens are much cheaper to produce and more robust (see Figure 4.1b). The use of porous material reduces the wind speed and also the generation of turbulences around the microphone. Since they have no frame to cause reflections, scattering or diffraction have, these windshields have only minor effects on the recorded sound field. Their main drawback is that they act as an acoustic low-pass filter, which can be easily compensated.

Measurements by Wuttke in [Wut92] showed that both types of windscreens can attenuate the wind induced noise up to almost 40 dB depending on the microphone type and wind condition. In general, the performance of the windscreens scales with their size. Thus, in many mobile applications the use of windscreens is not feasible or will lead only to an negligible amount of wind noise reduction.

4.2 Wind Noise Estimation

The input of the noise reduction system is a segmented version of the noisy input signal $x(k)$ (c.f. Figure 2.4) given in either a time-domain representation $x_\lambda(k)$ or a short-term spectral representation $X(\lambda, \mu)$. The crucial aspect of every noise reduction is the detection of the portion of speech and noise in each frame λ. If the noise is not detected or underestimated, annoying residual noise will appear in the output signal, while a false positive detection or an overestimation of the noise might result in an unwanted degradation of the speech signal. For conventional background noise signals, a rather stationary noise floor is assumed in the input signal. Thus, an explicit noise detection is not necessary and the noise can be separated from the speech by a temporal analysis in the short-term Fourier domain. This can be carried out by taking the minimum over a certain search window [Mar01] or by updating the noise estimate only when the probability of speech presence is assumed to be low [GH11]. For non-steady noise signals as wind noise, a frame-wise detection and estimation of noise activity is required for an efficient reduction. While the detection was discussed in Section 3.5, this section describes the process of wind noise estimation.

4.2.1 Review on Single Microphone Wind Noise Estimation

Figure 4.2 depicts the long-term power spectral density (PSD) of speech and wind noise computed by averaging the STPS of a complete signal. All depicted values are normalized with respect to a maximum value of each curve at 0 dB. The speech is separated into voiced (red) and unvoiced (blue) segments and the spectra are calculated from 60 seconds of randomly chosen speakers from the TIMIT database [LKS89], while the wind noise spectrum is gained from 60 seconds of wind recordings from [NV14b]. The main spectral overlap is given for voiced speech and wind noise and thus the main task is the enhancement of the frequency range in which both voiced speech and wind noise are active. The distorted segments with no spectral overlap of wind noise and unvoiced speech or no speech activity can be enhanced by a simple high-pass filter. A further positive effect is that unvoiced speech and high-pass filtered wind noise both have similar acoustic properties, which leads to a lower perceptual distortion. Thus, the main problem of wind noise reduction can be specified to the enhancement of voiced speech components.

The reduction of noise in a speech signal, as represented by the general structure in Figure 2.4, requires an estimate of the noise spectrum or noise short-term PSD.

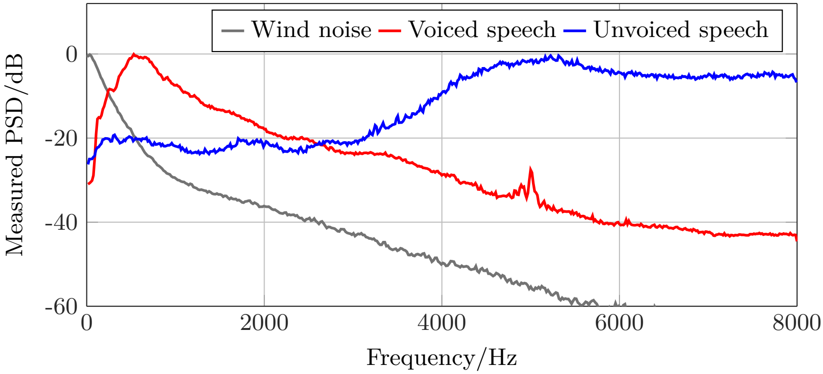


Figure 4.2: Power spectral density of wind noise and speech signals.

The latter is usually computed by the expectation of the signal spectrum over a certain time period, e.g., by a first-order recursive smoothing (c.f., Equation 3.37). For stationary signals this can efficiently reduce estimation errors but for fast varying signals, such as wind noise, any procedure of averaging or smoothing must be applied carefully, because this can reduce the accuracy to a great amount. Therefore, in the case of wind signals the quantity, which is required for the reduction is called short-term power spectrum (STPS)[1] of the wind noise and will be denoted by $|\mathcal{N}(\lambda, \mu)|^2$ or $|\widehat{\mathcal{N}}(\lambda, \mu)|^2$ for its estimate, respectively.

In the past, only a few proposals can be found for the estimation and reduction of wind noise. In the following, two methods from literature for the estimation of wind noise STPS will be introduced, which operate on a single microphone input signal and represent the most promising approaches.

4.2.1.1 Morphological Approach for Wind Noise Estimation

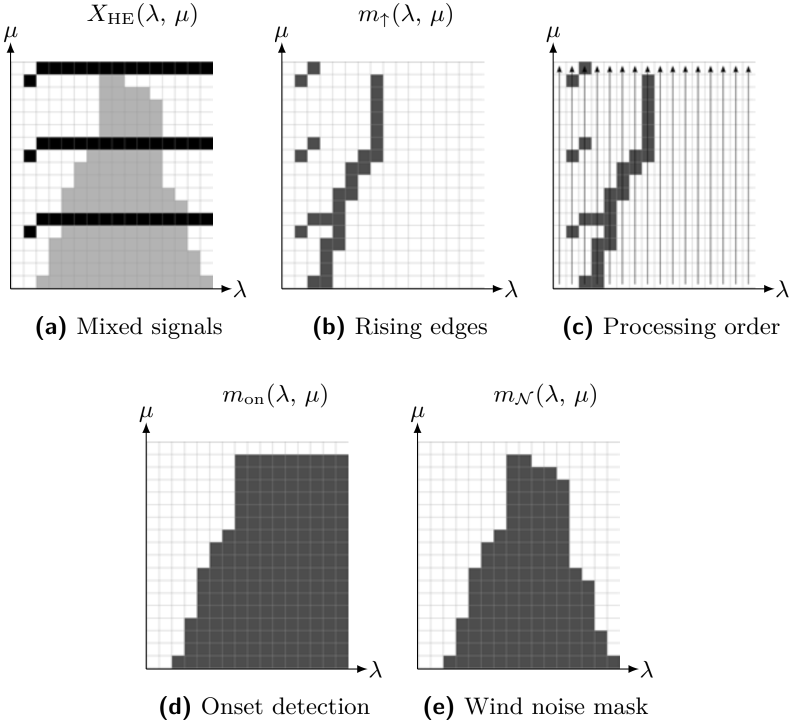
The approach in [HWB$^+$12] by Hofmann et al. regards the spectrogram of the noisy signal in the time-frequency plane as a two dimensional image. Considering a spectrogram, as presented in Figure 2.5b, parts affected by wind noise can be seen as connected areas in the time-frequency plane, while voiced speech shows the typical harmonic structure with high amplitudes at the fundamental frequency and its multiples. A separation of the connected areas is obtained by so-called morphological operations. These operations are usually applied in image processing tasks to detect connected areas (see, e.g., [FP03]). In the following, this algorithm will be denoted as morphological approach (MORPH).

The aim of the wind noise estimator of Hofmann is to determine areas in the time-frequency plane as a mask, which labels the appearance of wind buffets. A similar concept is known from many blind source separation algorithms, see, e.g., [YR04]. First, the high-energy components $X_{\mathrm{HE}}(\lambda, \mu)$ of the signal are exposed by comparing each frequency bin to a certain threshold or to a background noise estimate. The latter option is applied in the case that additional stationary noise sources also exist in the recorded signal. The steps for this procedure of computing the wind noise mask from the high-energy components are exemplary pictured in Figure 4.3 in the time-frequency plane. In Figure 4.3a a noisy voiced speech segment is given as input signal, where black and gray areas denote speech and wind noise, respectively. The first stage of the processing is given by a derivative $m'(\lambda, \mu)$ of the high-energy components $X_{\mathrm{HE}}(\lambda, \mu)$ with respect to the time, realized by the difference between successive frames as

$$\frac{\partial}{\partial t}|X_{\mathrm{HE}}(\lambda, \mu)| \approx |X_{\mathrm{HE}}(\lambda, \mu)| - |X_{\mathrm{HE}}(\lambda - 1, \mu)| = m'(\lambda, \mu). \tag{4.1}$$

The high-energy components are computed in [HWB$^+$12] by comparing the estimate

[1] The computation of the STPS should be normalized to the frame-size for a correct physical definition but will be omitted as it is usually done in literature. As the STPSs are always used in relation to each other (e.g., SNR) the dependency on the frame-size will be canceled out.


$X_{\mathrm{HE}}(\lambda,\,\mu)$
$m_{\uparrow}(\lambda,\,\mu)$
$m_{\mathrm{on}}(\lambda,\,\mu)$
$m_{\mathcal{N}}(\lambda,\,\mu)$
μ
λ

(a) Mixed signals **(b)** Rising edges **(c)** Processing order

(d) Onset detection **(e)** Wind noise mask

Figure 4.3: Steps towards the computation of the wind noise mask $m_{\mathcal{N}}(\lambda,\,\mu)$.

of a conventional noise estimator for constant noise to the noisy wind noise signal. By this procedure only speech and wind noise is assumed to stand out yielding $X_{\mathrm{HE}}(\lambda,\,\mu)$. From Equation 4.1 rising edges $m_{\uparrow}(\lambda,\,\mu)$ in the input signal can be detected by comparing the result with a threshold θ_{on}

$$m_{\uparrow}(\lambda,\,\mu) = \begin{cases} 1, & \text{if } m'(\lambda,\,\mu) > \theta_{\mathrm{on}} \\ 0, & \text{else.} \end{cases} \tag{4.2}$$

resulting in the labeled areas in Figure 4.3b. In the next step, a processing as shown in Figure 4.3c along the frequency axis is applied to find the onsets of the wind noise signal. An onset is defined by the two-dimensional non-linear recursive filter as

$$\begin{aligned} m_{\mathrm{on}}(\lambda,\,\mu) :=\, & \underbrace{\left(m_{\uparrow}(\lambda,\,\mu) \wedge m_{\mathrm{on}}(\lambda, \mu-1)\right)}_{\text{spectral connection}} \vee \underbrace{\left(m_{\uparrow}(\lambda,\,\mu) \wedge m_{\mathrm{on}}(\lambda-1, \mu)\right)}_{\text{temporal connection}} \vee \\ & \underbrace{\left(m_{\uparrow}(\lambda,\,\mu) \wedge \mu \le \mu_{\mathrm{low,max}}\right)}_{\text{lowest-frequency edges}}, \end{aligned} \tag{4.3}$$

where $\wedge$ and $\vee$ are logical conjunction and disjunction, respectively. The detection of the wind noise area starts from the low frequency bins below $\mu_{\text{low,max}}$, where only wind noise is assumed to be active. From this anchor, spectral and temporal connections are identified. By this processing isolated active frequency bins remaining from the harmonic pitch structure (e.g., in the upper left corner of Figures 4.3b and 4.3c) are removed resulting in the area, which is displayed in Figure 4.3d. A comparison of the observed signal energy of the unterminated wind mask $m_{\text{on}}(\lambda, \mu)$ in these bins with a heuristically chosen threshold identifies the complete shape of the wind noise $m_{\mathcal{N}}(\lambda, \mu)$ as depicted in Figure 4.3e. Applying the mask to the noisy input spectrum

$$|\widehat{\mathcal{N}}_{\text{MORPH}}(\lambda, \mu)|^2 = m_{\mathcal{N}}(\lambda, \mu) \cdot |X(\lambda, \mu)|^2, \quad \text{with } m_{\mathcal{N}}(\lambda, \mu) \in \{0, 1\}. \tag{4.4}$$

results into the wind noise STPS estimate. In [HWB+12] post processing is applied to remove isolated spectral notches by smoothing of the estimated mask in Equation 4.4 over frequency. This approach nicely estimates wind noise but has the drawback that low-frequency parts of the speech signal might also be included in the wind mask and thus be labeled as noise. More details on the implementation and choice of parameters can be found in [HWB+12].

4.2.1.2 Wind Noise Estimation Using Noise Templates

The idea of Kuroiwa et al. is based on a decomposition of the spectral shape of wind noise into its rough spectral structure, i.e., the spectral envelope, and the spectral fine structure [KMT+06]. This separation is realized in the cepstral domain, whereas the real cepstrum is defined as the inverse Fourier transform of the logarithmic spectrum

$$c_\lambda(q) = \frac{1}{N} \sum_{\mu=0}^{N-1} \log_{10}(|X(\lambda, \mu)|) e^{j\frac{2\pi\mu q}{M}} \;\; , \; q = 0, 1, \ldots, M-1 \tag{4.5}$$

with the cepstral coefficients $c_\lambda(q)$. This representation can be used to decompose a signal into “slow frequency” variations also referred to as the spectral envelope and the spectral fine structure represented by the lower and higher cepstral coefficients, respectively (see, e.g., [GM10]). The method presented by Kuroiwa is shown in a simplified version in Figure 4.4. In the sequel, this second reference method is denoted by cepstral wind reference (CWR) approach.

After a cepstral analysis of each noisy input frame, the cepstral coefficients are split up into the higher coefficients $c_\lambda(q_{\text{th}} + 1) \ldots c_\lambda(M - 1)$ and the lower coefficients $c_\lambda(0) \ldots c_\lambda(q_{\text{th}})$. While the higher coefficients are kept untouched, the lower coefficients are processed, which are mainly responsible for the spectral energy distribution and thus the accuracy of the wind noise estimate. The computation of the lower coefficients is carried out by using reference envelopes of wind noise. These references are trained in a separate step before the wind noise reduction is applied, using the lower cepstral coefficients of pure wind noise recordings. A

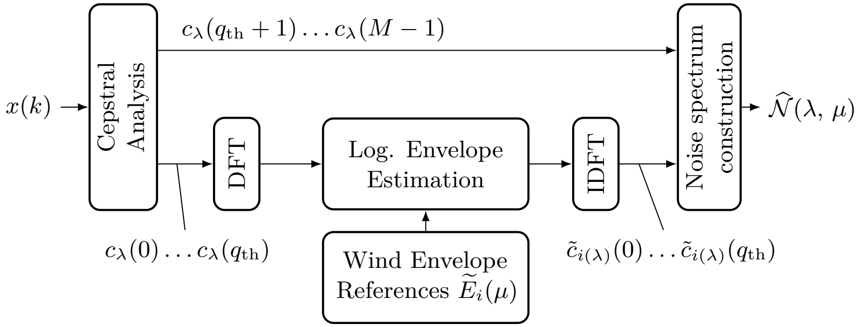

$c_\lambda(q_{\text{th}}+1)\dots c_\lambda(M-1)$
$x(k)$
Cepstral Analysis
DFT
Log. Envelope Estimation
IDFT
Noise spectrum construction
$\widehat{\mathcal{N}}(\lambda,\,\mu)$
$c_\lambda(0)\dots c_\lambda(q_{\text{th}})$
Wind Envelope References $\widetilde{E}_i(\mu)$
$\tilde{c}_{i(\lambda)}(0)\dots\tilde{c}_{i(\lambda)}(q_{\text{th}})$

Figure 4.4: Template based wind noise estimation.

subsequent vector quantization of the coefficients guarantees a limited number of references representing different wind noise conditions. In [KMT$^+$06] the LBG algorithm [LBG80] was proposed for the vector quantization and is also used in the investigated implementation.

During the noise estimation process, the lower cepstral coefficients $c(1)\dots c(q_{\text{th}})$ of the observed signal are transformed back into the DFT domain again yielding the logarithmic spectral envelope

$$\log_{10}|E(\lambda,\,\mu)| = \text{DFT}\{c(0),..,c(q_{\text{th}})\}. \tag{4.6}$$

From the stored reference envelopes $\widetilde{E}_i(\mu)$ the optimal candidate i_{opt} is taken, which minimizes squared error as

$$i_{\text{opt}}(\lambda) = \underset{i}{\text{argmin}}\left\{\sum_{\mu=0}^{\mu_{\text{th}}}(\log_{10}|E(\lambda,\,\mu)| - \log_{10}|\widetilde{E}_i(\mu)|)^2\right\} \tag{4.7}$$

in a lower frequency range limited by μ_{th} (e.g., up to 100 Hz as proposed in [KMT$^+$06]). To avoid the influence of the signal energy in the candidate search in Equation 4.7, both the reference envelopes and the observed envelope's energy are normalized to a constant value, e.g., one.

The cepstral coefficients $\tilde{c}_{i,\text{opt}}(q)$ corresponding to the optimal envelope $\widetilde{E}_{i,\text{opt}}(\mu)$ now replace the cepstral coefficients of the observed spectrum as depicted in the lower branch in Figure 4.4. Combining them with the higher cepstral bins from the input spectrum the complete spectrum is constructed. After the inverse cepstral transformation of the combined coefficients and an energy correction of the resulting spectrum an estimate $|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2$ of the wind noise STPS in the current frame is given.

4.2.2 Centroid Based Wind Noise Estimation

The new concepts proposed in the following are based on a classification of the current signal frame and subsequent estimation of the wind noise STPS. The classification aims to detect segments in the recorded signal, which contain pure wind noise, clean speech or a mixture of speech and and wind. In contrast to the pure detection described in Section 3.5, the classification must also give evidence about speech activity. As shown in Figure 4.2, wind noise mainly affects voiced speech. Thus, the classification aims to distinguish between voiced speech and wind noise.

The feature used for the classification is given by the sub-band signal centroid (SSC) $\Xi(\lambda)$, which were already defined in Section 3.5.2.2 for the wind detection as

$$\Xi_{\mu_1,\mu_2}(\lambda) = \frac{f_\mathrm{s}}{M} \frac{\sum_{\mu=\mu_1}^{\mu_2} \widehat{\Phi}_{xx}(\lambda,\,\mu)\cdot\mu}{\sum_{\mu=\mu_1}^{\mu_2} \widehat{\Phi}_{xx}(\lambda,\,\mu)} \tag{4.8}$$

and reflects the energy distribution of an observed short-term PSD $\widehat{\Phi}_{xx}(\lambda,\,\mu)$. Assuming that speech and wind noise signals are uncorrelated, the PSD of the noisy signal is given by the sum of the speech short-term PSD $\widehat{\Phi}_{ss}(\lambda,\,\mu)$ and wind short-term PSD $\widehat{\Phi}_{nn}(\lambda,\,\mu)$ as

$$\widehat{\Phi}_{xx}(\lambda,\,\mu) = \widehat{\Phi}_{ss}(\lambda,\,\mu) + \widehat{\Phi}_{nn}(\lambda,\,\mu). \tag{4.9}$$

With the definition of the short-term *a posteriori* SNR

$$\widehat{\gamma}(\lambda,\,\mu) = \frac{\widehat{\Phi}_{xx}(\lambda,\,\mu)}{\widehat{\Phi}_{nn}(\lambda,\,\mu)} \tag{4.10}$$

the definition of the SSC of the input signal $x(k)$ in Equation 4.8 can be rewritten as

$$\begin{aligned}
\Xi(\lambda) \quad &= \quad \frac{f_\mathrm{s}}{M}\Bigg(\frac{\sum\widehat{\Phi}_{ss}(\lambda,\,\mu)}{\sum\widehat{\Phi}_{ss}(\lambda,\,\mu)+\widehat{\Phi}_{nn}(\lambda,\,\mu)}\frac{\sum\widehat{\Phi}_{ss}(\lambda,\,\mu)\cdot\mu}{\sum\widehat{\Phi}_{ss}(\lambda,\,\mu)} \\
&\qquad + \frac{\sum\widehat{\Phi}_{nn}(\lambda,\,\mu)}{\sum\widehat{\Phi}_{ss}(\lambda,\,\mu)+\widehat{\Phi}_{nn}(\lambda,\,\mu)}\frac{\sum\widehat{\Phi}_{nn}(\lambda,\,\mu)\cdot\mu}{\sum\widehat{\Phi}_{nn}(\lambda,\,\mu)}\Bigg) \\
&= \quad \Xi_s\cdot\left(1-\frac{1}{\overline{\gamma}(\lambda)}\right)+\Xi_n\cdot\frac{1}{\overline{\gamma}(\lambda)},
\end{aligned} \tag{4.11}$$

where the indices of the sums in Equation 4.11 are defined over the frequency range between μ_1 and μ_2 as in Equation 4.8, but are omitted here for the sake of clarity. The centroids of clean speech and pure wind noise are presented by Ξ_s and Ξ_n, and $\overline{\gamma}(\lambda)$ is the mean short-term *a posteriori* SNR in one frame, i.e. averaged over the frequency range $\mu_1 \ldots \mu_2$. Now, a prediction of the SNR can be made from the measured SSC value, if Ξ_s and Ξ_n are known by rearranging Equation 4.11 to

$$\overline{\gamma}(\lambda) = \frac{\Xi_s - \Xi_n}{\Xi_s - \Xi(\lambda)}. \tag{4.12}$$

To illustrate this relation, an experiment is carried out measuring the SNR and the corresponding centroid frequency $\Xi(\lambda)$, in each frame for a speech signal disturbed by wind noise using the frequency range from 0 to 3000 Hz for μ_1 and μ_2. For the speech data, 3 minutes of randomly chosen speaker from the TIMIT database [LKS89] are taken and 3 minutes from the database in [NV14b] represents the wind noise. The measured data is sorted according to the SNR values and the resulting values averaged for all frames are depicted in Figure 4.5 by the black solid curve. Furthermore, it is assumed that the centroid frequency of wind noise and speech are $\Xi_s = 100\,\text{Hz}$ and $\Xi_n = 700\,\text{Hz}$, respectively. These prior conditions are taken, as the measurements in Figure 3.15 indicate these as realistic values. Inserting the aforementioned centroids Ξ_s and Ξ_n into Equation 4.11, the dashed gray curve follows from the considered SNR range. For both curves in Figure 3.15 the *a priori* SNR

$$\overline{\xi}(\lambda) = \overline{\gamma}(\lambda) - 1 \tag{4.13}$$

is considered for reasons of clarity. It can be seen, that there are no big deviations between the measured and theoretical relation for the signal centroids and the *a priori* SNR. From the SNR-dependent behaviour in Figure 4.5, three classes can be defined:

- **A** pure wind noise ($\Xi < 200\,\text{Hz}$)
- **B** both voiced speech and wind noise are active ($200\,\text{Hz} < \Xi < 550\,\text{Hz}$)
- **C** clean voiced speech ($\Xi > 550\,\text{Hz}$).

This classification will be used for the following two wind noise estimation approaches as shown in the decision diagram in Figure 4.6. In a first step a binary decision for each frame is made, if wind is active. The normalized short-term mean

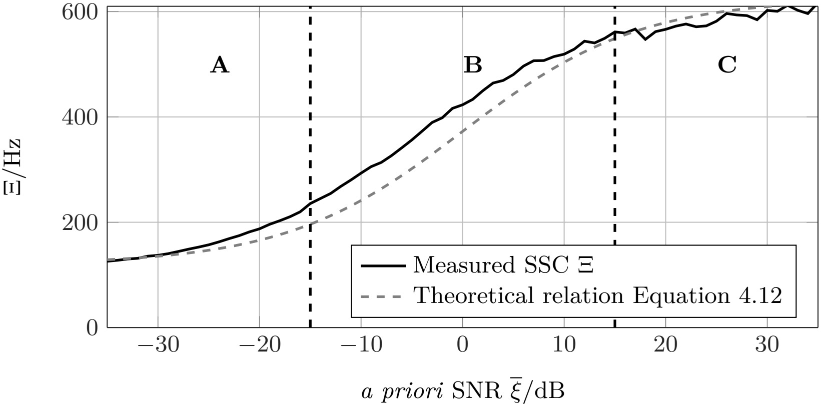


Figure 4.5: Signal centroid of voiced speech disturbed by wind noise.

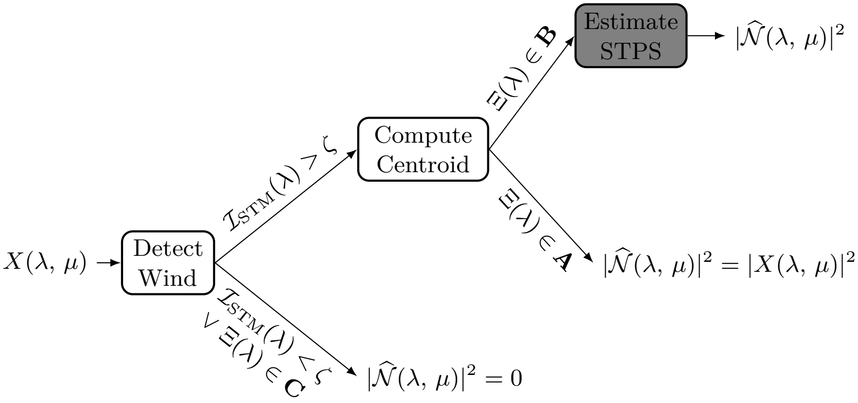

$X(\lambda,\,\mu)$
Detect Wind
$\mathcal{I}_{\mathrm{STM}}(\lambda) > \zeta$
Compute Centroid
$\Xi(\lambda) \in \mathbf{B}$
Estimate STPS
$|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2$
$\Xi(\lambda) \in \mathbf{A}$
$|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2 = |X(\lambda,\,\mu)|^2$
$\mathcal{I}_{\mathrm{STM}}(\lambda) < \zeta$
$\vee\ \Xi(\lambda) \in \mathbf{C}$
$|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2 = 0$

Figure 4.6: Decision diagram for wind noise estimation.

(NSTM) approach proposed in Section 3.5 showed the best detection rate (c.f., Figure 3.17) giving the wind indicator $\mathcal{I}_{\mathrm{NSTM}}(\lambda)$. The binary decision for wind activity is given by a comparison of the indicator with a threshold ζ. In the case of no wind activity ($\mathcal{I}_{\mathrm{NSTM}}(\lambda) < \zeta$) the noise estimate can be set to

$$|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2 = 0, \text{ if } \mathcal{I}_{\mathrm{NSTM}}(\lambda) < \zeta \vee\ \Xi(\lambda) \in \mathbf{C}. \tag{4.14}$$

In addition, the SSC in the current frame is checked for clean speech activity ($\Xi(\lambda) \in \mathbf{C}$) to ensure that wind noise reduction systems leave these parts untouched. Based on the measured centroid a further classification is possible between frames containing noisy speech or pure wind noise. In the case of pure wind noise ($\mathbf{A}$), the wind noise STPS estimation $|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2$ can easily be realized by taking the input spectrum

$$|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2 = |X(\lambda,\,\mu)|^2, \text{ if } \Xi(\lambda) \in \mathbf{A}. \tag{4.15}$$

With a given $\Xi(\lambda)$ value in each frame, an estimate of the SNR condition is possible as shown in Equation 4.12. But so far no frequency dependent estimate of the wind noise is realized for condition $\mathbf{B}$, which is required for the subsequent speech enhancement. For the most challenging case $\mathbf{B}$, two strategies will be presented, which exploit the distinguishable structures of speech and wind to estimate the wind noise STPS, when both speech and wind are present.

4.2.2.1 Minima Fitting Approach For Wind Noise Estimation

Because only voiced speech is expected in the lower frequency range, where the wind noise signal is active, the harmonic structure of this speech segments can be exploited. This means that the speech energy is located at the fundamental frequency and multiples of it. In between, i.e., at local minima of the magnitude spectrum, the noise spectrum is assumed to be detectable.

The short-term spectral characteristics of voiced speech and wind noise are demonstrated in Figure 4.7, where the noisy speech spectrum and the underlying wind noise are depicted by the black and gray curves, respectively. Furthermore two local minima $X_1(\lambda, \mu)$, $X_2(\lambda, \mu)$ of the noisy speech spectrum are marked by the black circles for frequencies above 100 Hz, where voiced speech is expected (see, e.g., [VM06]).

The task of estimating the wind noise STPS during voiced speech activity can be realized by exploiting the local minima. Those points of the noisy spectrum can be used to fit an approximation of the wind noise spectrum. Different concepts for the approximation can be applied. If all local minima are taken into account a least square regression is possible. Since in the higher frequency range only a negligible amount of wind noise is expected, this approximation will overemphasize the high frequencies.

Based on the spectral shapes of voiced speech and wind noise, the method presented in [NCBV14] and [NCBV15] approximates the wind noise spectrum as an $1/f^{\nu}$-decay over the frequency, which was introduced as the distinct spectral shape of wind in Section 3.3.3. The expression

$$\widetilde{N}_{1/f}(\lambda, \mu) = \beta(\lambda) \cdot \mu^{-\nu(\lambda)} \tag{4.16}$$

is introduced to describe the spectral shape of the wind noise signal in each frame.

The parameter $\beta(\lambda)$ and $\nu(\lambda)$ control the noise power and the spectral slope of the approximation, respectively. Both parameters has to be determined in every frame. The curve-fitting concept is illustrated in Figure 4.7, where the dashed gray curve represents the $1/f^{\nu}$ decay for this example. The noisy speech and the wind

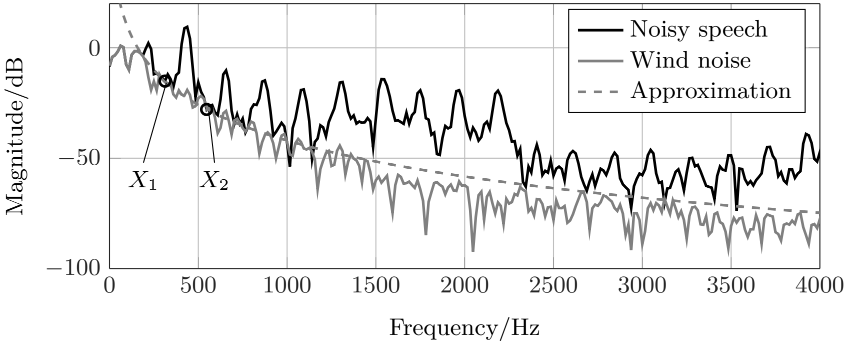


Figure 4.7: Wind noise estimation by approximation of local minima (method Min-Fit [NCBV14]).

noise spectrum are shown by the solid black and gray curves, respectively, and the used data points $X_1(\lambda,\, \mu)$, $X_2(\lambda,\, \mu)$ are marked by the black circles.

For the calculation of the $\beta(\lambda)$ and $\nu(\lambda)$ the amplitudes of at least two measured points $X_1(\lambda,\, \mu) = X(\lambda, \mu_1)$, $X_2(\lambda,\, \mu) = X(\lambda, \mu_2)$ from the observed spectrum $X(\lambda,\, \mu)$ are necessary leading to

$$\nu(\lambda) = \frac{\log\left(\left|\frac{X(\lambda,\mu_1)}{X(\lambda,\mu_2)}\right|\right)}{\log\left(\frac{\mu_2}{\mu_1}\right)} \tag{4.17}$$

and

$$\beta(\lambda) = {\mu_1}^{\nu(\lambda)} \cdot |X(\lambda, \mu_1)|. \tag{4.18}$$

In the example in Figure 4.7 the first two local minima at μ_1 and μ_2 above 100 Hz are taken, which are identified by a simple comparison with their neighbouring values. To ensure that the considered local minima are between the multiples of the fundamental frequency they must fulfill the two conditions:

1. The distance between two adjacent minima must be at least 50 Hz, as this is lowest expected fundamental frequency and thus smallest distance between two valleys of the harmonic structure of speech.

2. The negative peaks representing the local minima must show a negative peak prominence[2] of at last 1 dB to ensure that the considered minima correspond to a harmonic valley.

Alternatively, one or more measurement points, which are not local minima, can be chosen directly from the spectrum below 100 Hz, where no speech energy is expected. In this range many microphones show a certain high-pass characteristic, which influences the approximation and thus the noise estimate. Therefore, the frequency range above 100 Hz is used in the following. Considering the observations from Section 3.3.3 the parameter $\nu(\lambda)$ controlling the spectral slope is bounded to the range between 1 and 2. This lowers the amount of over and under estimation of the wind noise spectrum, which might lead to severe artifacts during the speech enhancement process. As seen in Figure 4.7, the noise approximation can exceed the current noisy signal frame for low frequencies ($<$ 200 Hz), therefore the estimate defined by Equation 4.16 is limited by the noisy signal $X(\lambda,\, \mu)$ in the current frame as

$$|\widehat{\mathcal{N}}_{1/f}(\lambda,\, \mu)|^2 = \min\left\{\widetilde{N}^2_{1/f}(\lambda,\, \mu), |X(\lambda,\, \mu)|^2\right\}. \tag{4.19}$$

Experiments using a curve fitting approach using more than two local minima does not improve the approximation accuracy. This arises from the fact that the minima

[2] The prominence of a peak is defined by the minimum amount that the signal must descend on either side of the peak before either climbing back to a level higher than the peak or reaching an endpoint of the signal.

at higher frequencies are often not only related to the wind noise spectrum but to other portions in the captured signals, e.g. sensor noise or further background noise.

The only necessary steps for this wind noise estimation are the determination of the local minima and the computation of the parameter β and ν for the spectrum approximation, which makes this algorithm to a solution featured by a low computational complexity.

4.2.2.2 Pitch Adaptive Wind Noise Estimation

The technique proposed in [NV15] for the estimation of the wind noise STPS is proposed, which takes into account a parameter describing the harmonic structure of voiced speech signals given by the fundamental frequency f_0. The fundamental frequency is the inverse of the pitch cycle, which determines the periodicity of the speech signal.

There exists a variety of algorithms, for estimating the fundamental frequency f_0 or its discrete representation μ_0 in short segments of a speech signal (see, e.g., [Hes83] for an overview). They can be roughly divided into methods working in the time-domain and methods working in a transform domain, mostly the DFT domain. It turned out that frequency-domain approaches showed the most robust results towards wind noise, because only a narrow spectral region of voiced speech is influenced by the wind signal. For the proposed system the harmonic product spectrum (HPS) was chosen as pitch estimator ([Nol70]):

$$\hat{\mu}_0(\lambda) = \arg\max_{\mu} \left\{ \frac{\prod_{l=1}^{M_H} |X(\lambda, l \cdot \mu)|}{\prod_{l=1}^{M_H} |X(\lambda, l \cdot \mu + \lceil \mu/2 \rfloor)|} \right\}, \tag{4.20}$$

where $\lceil x \rfloor$ denotes the closest natural number to x and M_H is the number of considered harmonics. In [ISM08] Equation 4.20 was used to compute the pitch frequency of band-limited speech, where the frequencies below 300 Hz are completely missing. It turned out that in the case of wind noise, where mainly the lower frequencies are corrupted, the HPS also gives quite good results for the pitch estimation. It must be mentioned, that pitch estimation in general requires larger frame-sizes than the 20 ms usually applied in this work to detect also low fundamental frequencies to 50 Hz. Therefore, the HPS method is carried out on frames of 50 ms length.

The idea of the method presented in [NV15] is to use the knowledge of the energy distribution of the speech spectrum for a given fundamental frequency. By eliminating the harmonic speech components in the noisy spectrum, i.e., by setting the corresponding frequency bins to zero, the remaining parts of the spectrum are assumed to belong to the wind noise spectrum. This is realized by using a so-called pitch adaptive inverse binary mask (P-IBM).

Binary masks are usually used to separate speech and noise by multiplying a spectral gain

$$G_{\mathrm{BM}}(\lambda,\,\mu) = \begin{cases} 1, \text{if } |S(\lambda,\,\mu)|^2 > \mathrm{LC}(\mu), \\ 0, \text{otherwise} \end{cases} \tag{4.21}$$

to the noisy spectrum $X(\lambda,\,\mu)$. The resulting output signal only contains parts, where the speech power $|S(\lambda,\,\mu)|^2$ is higher than a certain local criterion $\mathrm{LC}(\mu)$. This criterion is usually a threshold, which might depend on the local SNR. Applying an ideal binary mask can improve the intelligibility or the performance of an automatic speech recognition system (see, e.g., [GB14] and references therein). Normally, binary masks completely cancel out parts of the undesired noise signal. This leads to a sufficient but also aggressive noise suppression, which may introduce unwanted artifacts to the output signal. Furthermore, due to the binary gain of the mask based processing it follows, that the noise is not reduced in time-frequency units, where both speech and noise are active. This residual noise also results in annoying effects in the output signal.

Here, the aim is to cancel out the harmonic components of voiced speech segments in the time-frequency plane by applying the P-IBM to the noisy signal. For this purpose, the binary mask is defined as follows

$$G_{\text{P-IBM}}(\lambda,\,\mu) = \begin{cases} 0, \text{if } \mu \in \mathbb{M}_0(\lambda) \\ 1, \text{else,} \end{cases} \tag{4.22}$$

with the set $\mathbb{M}_0(\lambda)$ of frequency bins belonging to speech activity

$$\mathbb{M}_0(\lambda) = \{\kappa \cdot \hat{\mu}_0(\lambda) - \mu_\Delta, \ldots, \kappa \cdot \hat{\mu}_0(\lambda), \ldots \kappa \cdot \hat{\mu}_0(\lambda) + \mu_\Delta\}\,,\ \forall \kappa \in \mathbb{N} \tag{4.23}$$

and $\hat{\mu}_0(\lambda)$ depicts the discrete frequency bin corresponding to the fundamental frequency estimate. The parameter μ_Δ determines a frequency range around the pitch bin to ensure the cancellation of the speech signal by the P-IBM. The concept is displayed in Figure 4.8. The harmonic structure of the speech in the noisy signal is clearly visible by the peaks at multiples of f_0 in the black curve, which is removed by the binary mask $G_{\text{P-IBM}}(f)$ shown as the dashed gray curve. An estimate of the speech-free amplitude spectrum is then given by

$$A_{\mathrm{noSp}}(\lambda,\,\mu) = G_{\text{P-IBM}}(\lambda,\,\mu) \cdot |X(\lambda,\,\mu)| \tag{4.24}$$

in which the speech components are set to zero.

An important parameter for the determination of the binary mask in Equations 4.22-4.23 is the width of zero-segments μ_Δ. On the one hand, if μ_Δ is too small residual parts of the speech spectrum will be identified as noise resulting into unwanted attenuation of the desired speech signal in the further steps of the noise reduction process. On the other hand, too wide zero-segments leads to smaller remaining parts of the spectrum and thus a less accurate STPS estimate of the wind noise.

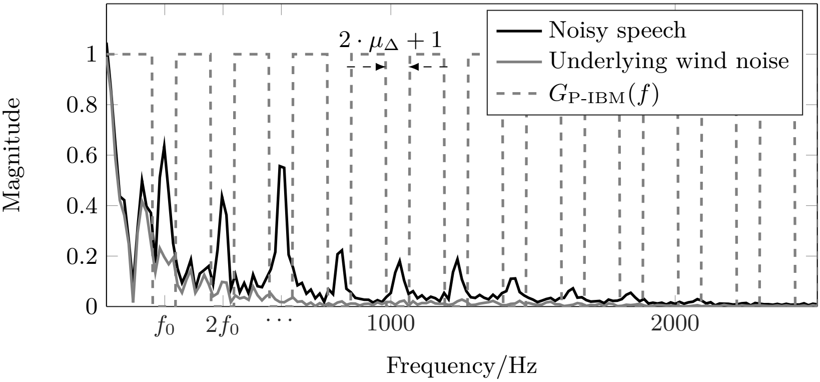


Figure 4.8: Pitch adaptive masking with $\mu_\Delta \mathrel{\widehat{=}} 50\,\text{Hz}$, $M = 512$, $f_\text{s} = 16\,\text{kHz}$ (method P-IBM [NV15]).

Due to the segmentation and windowing of the signal in the noise reduction framework (see Figure 2.2) the considered spectrum $X(\lambda,\,\mu)$ has a limited frequency resolution and the so-called leakage effect causes a spreading of the spectrum (see [OSB+89]). Because of the latter effect discrete frequency components, such as the harmonic structure of voiced speech, are spread to a broader range. Mathematically this can be described by a convolution of the spectrum with the spectrum of the window function resulting from the multiplication of the window function in the time-domain. The spectrum of the used square-root Hann window is depicted in Figure 4.9 for the considered frame-size of 20 ms. The dashed line marks the point, where the spectrum decreases by 10 dB. The 10 dB decrease from a single harmonic of the speech spectrum turned out to be a good trade-off between a low leakage effect of the speech harmonics and a broad width of the zero-segments. Therefore, this frequency range of approximately 50 Hz is used to define the width μ_Δ of the zero-segments in the binary mask definition in Equation 4.23.

Since only wind noise is assumed to occupy the non-zero parts of $A_\text{noSp}(\lambda,\,\mu)$, this spectrum is taken as the starting point for the STPS estimation. The parts around multiples of f_0, which are set to zero, are linearly interpolated using the known adjacent non-zero frequency bins at $\mu = \kappa \cdot \mu_0 \pm (\mu_\Delta + 1)$, resulting into the noise STPS estimate

$$|\widehat{\mathcal{N}}_{\text{P-IBM}}(\lambda,\,\mu)|^2 = \begin{cases} A^2_\text{noSp}(\lambda,\,\mu)\,, \text{if } \mu \in \mathbb{M} \\ A^2_\text{lin}(\lambda,\,\mu)\,, \text{else}, \end{cases} \tag{4.25}$$

where $A_\text{lin}(\lambda,\,\mu)$ is the interpolated wind noise spectrum.

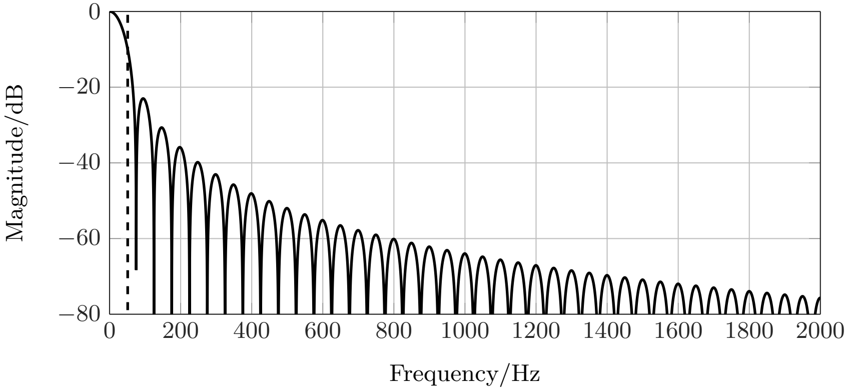


Figure 4.9: Spectrum of a 20 ms square-root Hann window.

Limitation of Wind Noise Over-estimation

Because wind noise only shows very low energy at higher frequencies, the aforementioned method can over-estimate the wind spectrum in this range, if the binary gain does not cancel out the complete speech spectrum. To prevent an over-estimation, a reliability check is performed exploiting the curve-fitting concept already used in Section 4.2.2.1. It was shown that the spectrum of wind noise can be approximated by an $1/f$ slope over the frequencies f. Therefore, the noise STPS estimate is limited at higher frequencies ($\mu > \mu_{\text{low}}$) by an $1/f^2$ curve starting from the averaged power $\sigma^2_{N,\text{low}}(\lambda)$ in the lower band ($\mu < \mu_{\text{low}}$) of the noise estimate from Equation 4.25

$$|\widehat{\mathcal{N}}_{\text{P-IBM}}(\lambda,\, \mu)|^2 = \min\left\{|\widehat{\mathcal{N}}_{\text{P-IBM}}(\lambda,\, \mu)|^2, \sigma^2_{N,\text{low}}(\lambda)/\mu^2\right\} \text{for } \mu > \mu_{\text{low}}. \quad (4.26)$$

The frequency limit corresponding to μ_{low} is set to the limit 2000 Hz for the reliability check. Below this frequency, most of the wind noise energy is located (see Figure 4.2) and thus this range covers the most relevant part of the wind noise spectrum.

4.2.3 Effects of Recursive Smoothing

Many noise estimators apply a first-order recursive smoothing to either the input signals as in [Mar01] or to both the input and to the estimated noise PSD as in [HHJ10] or [GH11]. The aim is to reduce the variance and improve the accuracy of the estimate. For background noise, which is assumed to be stationary or only slowly varying over time, this procedure might be helpful to reduce the impact of outliers in the estimation process. For highly non-stationary noise signals, such as wind noise, this smoothing conflicts with a sufficient high tracking speed of the noise estimate.

An experiment is carried out in the following, which shows the influence of recursive smoothing on the noise estimate accuracy. The noise signal in each frame is assumed to be known as $N(\lambda, \mu)$ and the smoothed version yielding to the short-term noise PSD estimate is given by

$$\widehat{\Phi}_{nn}(\lambda, \mu) = \alpha \cdot \widehat{\Phi}_{nn}(\lambda - 1, \mu) + (1 - \alpha) \cdot |N(\lambda, \mu)|^2. \tag{4.27}$$

The smoothing constant α determines the trade-off between good variance reduction ($\alpha \to 1$) and a high tracking speed ($\alpha \to 0$). The accuracy of the noise estimate in each frame is essential for the performance of the complete noise reduction system. A measure, which is often used to quantify the accuracy, is the logarithmic error[3] $e_{\log}$ between the noise PSD estimate and the real noise, where a lower error indicates a more accurate estimate (see Equation A.4). Usually, the noise PSD estimate is compared to a smoothed version of the real noise, e.g., with a smoothing constant $\alpha = 0.8$ ([HHJ10], [GH11], [TTM$^+$11]).

To obtain information about the influence of the recursive smoothing, the error between the short-term PSD estimate $\widehat{\Phi}_{nn}(\lambda, \mu)$ obtained by smoothing described in Equation 4.27 and the true squared magnitude of the noise $|N(\lambda, \mu)|^2$ is investigated.

Three typical background noise signals (car, babble and jackhammer) from [ETS09] and a recorded wind signal from [NV14b] are considered. The results are shown in Figure 4.10 using smoothing constants $0 < \alpha < 0.995$. It can be seen, that for the three background noise signals the smoothing only introduces a moderate error up to a maximum below 5 dB even in the case of smoothing factor close to one and the rather non-stationary noise types babble and jackhammer noise. For the dashed gray curve, presenting the logarithmic error of wind noise, the values

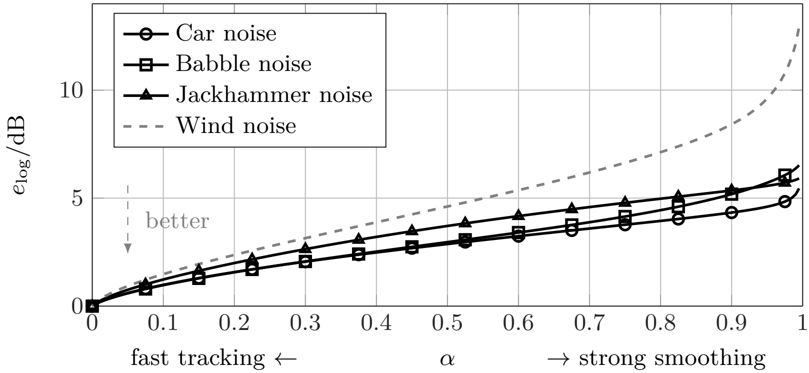


Figure 4.10: Logarithmic error $e_{\log}$ between squared magnitude $|N(\lambda, \mu)|^2$ of noise spectrum and its short-term PSD estimate $\widehat{\Phi}_{nn}(\lambda, \mu)$.

[3] Appendix A.2 gives a detailed description of the computation of the logarithmic error.

are significantly higher. Especially for $\alpha > 0.9$, which is a commonly chosen value for conventional noise estimation algorithms, the error increases dramatically. This simulation shows that only a light smoothing of the wind noise should be applied, which in turn provides a smaller variance reduction of the estimate. Thus, in the following the STPS estimate of the noise signal in each frame is directly employed for the speech enhancement process, i.e., $\alpha = 0$.

4.2.4 Evaluation of Wind Noise Estimation

In the considered speech enhancement system as shown in Figure 2.4 different components influence the performance of the noise reduction. A crucial role is played by the accuracy of the estimated wind noise STPS on which the subsequent spectral weighting is computed. The presented algorithms in Section 4.2.1.1-4.2.2.2 will be compared in terms to their accuracy for noisy speech signals with different input SNR scenarios. Again, the logarithmic error $e_{\log}$ is used as the quality measure. Speech sentences from male and female talkers are randomly taken from the TIMIT database [LKS89] and mixed with wind noise recordings from [NV14b] corresponding to SNR scenarios from -15 to 15 dB. Realistic wind noise conditions are mostly in the SNR range between -5 and 5 dB. The length of the signals are 60 seconds, where 3 different shifts of the noise signal are considered resulting in signals with a length of 3 minutes for each SNR scenario.

The results of the logarithmic error are shown in Figure 4.11 comparing the different wind noise estimation algorithms. The cepstral reference based method CWR from Section 4.2.1.2 shows the largest error in the lower SNR range. Only for high SNR values (> 10 dB) this method outperforms the other considered approaches. Considering the proposed methods, the Min-Fit approach from Section 4.2.2.1 show similar results as the morphological algorithm (MORPH) from Section

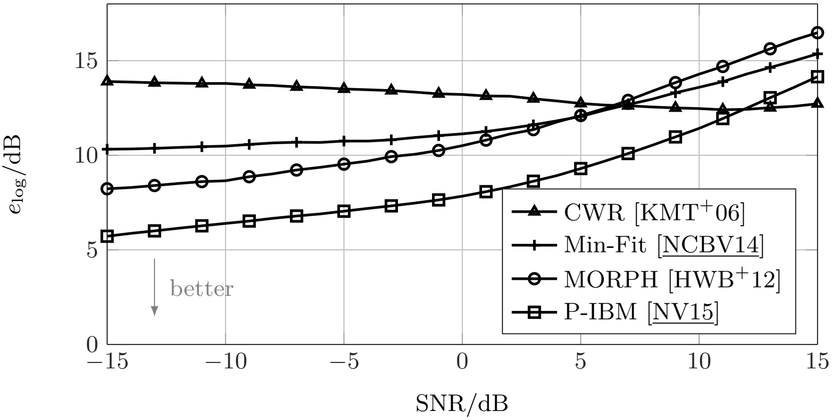


Figure 4.11: Estimation accuracy in terms of the logarithmic error.

4.2.1.1, while the pitch adaptive inverse binary mask (P-IBM) (Section 4.2.2.2) method outperforms the three other approaches. Especially in the SNR range -5 to 5 dB, which reflects realistic scenarios, the noise estimation by the P-IBM procedure shows the lowest error and thus the highest accuracy.

4.3 Wind Noise Reduction Based on Spectral Filtering

In the previous section only the estimation of the wind noise was discussed and evaluated. For the application in a communication application obviously the performance of the complete system as depicted in Figure 2.4 in terms of noise reduction or speech enhancement is deciding. Thus, the second crucial part of the speech enhancement, the computation of the spectral gain, is investigated and evaluated in this section.

In the past, a variety of rules for the gain calculation were developed (see, e.g., [Esc12] and references therein). Because the focus of this work is the detection and estimation of wind noise, only the most common algorithms already introduced in Section 2.3.3 are considered here, which is the Wiener filter

$$G_{\mathrm{W}}(\lambda,\, \mu) = \frac{\widehat{\xi}(\lambda,\, \mu)}{\widehat{\xi}(\lambda,\, \mu) + 1} \tag{4.28}$$

and can also described in form of the generalized spectral subtraction

$$G_{\mathrm{S}}(\lambda,\, \mu) = \sqrt{\left[1 - \left(\frac{|\widehat{\mathcal{N}}(\lambda,\, \mu)|^2}{|X(\lambda,\, \mu)|^2}\right)^{\beta_{\mathrm{S}}}\right]^{\alpha_{\mathrm{S}}}} \tag{4.29}$$

with the parameter $\alpha_{\mathrm{S}} = 2$ and $\beta_{\mathrm{S}} = 1$. Since in a real scenario the required SNRs $\xi(\lambda, \mu)$ or noise power spectra $|\mathcal{N}(\lambda,\, \mu)|^2$ are only available as estimates, usually estimation errors arise for all noise estimation methods as shown in Section 4.2.4. Especially, short segments in which the noise estimate is inaccurate can lead to severe artifacts in the output signal. An underestimation of the noise leads to short residual noise segments also known as “musical tones” while an overestimation might lead to an undesired cancellation of parts of the speech signal. To overcome these problems different strategies were proposed in the past. Two of those will be investigated and also taken into the evaluation process at the end of this section.

4.3.1 Decision Directed SNR Estimation

Ephraim and Malah proposed the so-called “decision-directed” approach (DDA) presented in [EM84] to update the *a priori* SNR $\widehat{\xi}(\lambda,\, \mu)$ with the smoothing constant α_ξ

$$\widehat{\xi}(\lambda,\, \mu) = \alpha_\xi \cdot \frac{|\widehat{S}(\lambda - 1, \mu)|^2}{|\widehat{\mathcal{N}}(\lambda - 1, \mu)|^2} + (1 - \alpha_\xi) \cdot \max\{\widehat{\gamma}(\lambda,\, \mu) - 1, 0\}, \tag{4.30}$$

where α_ξ is typically in the range $0.9 < \alpha_\xi < 0.99$ and $\widehat{S}(\lambda - 1, \mu)$ is the spectrum of the enhanced previous frame. Usually, this procedure contributes to higher subjective quality of the enhanced speech, especially to attenuate "musical tones". In terms of transient or fast changing signal characteristics the "decision-directed" approach might lead to a reduced performance due the smoothing over consecutive signal frames. In [EM84] a high smoothing constant of $\alpha_\xi = 0.98$ is proposed for the SNR computation to reduce variations of the spectral gains. For the considered wind noise reduction system the *a posteriori* SNR estimate $\widehat{\gamma}(\lambda,\, \mu)$ in Equation 4.30 is computed from the STPS in each frame as

$$\widehat{\gamma}(\lambda,\, \mu) = \frac{|X(\lambda,\, \mu)|^2}{|\widehat{\mathcal{N}}(\lambda,\, \mu)^2} \tag{4.31}$$

for the Wiener filter using the presented wind noise estimators for the STPS estimation.

4.3.2 Spectral Subtraction with Recursive Gain Curves

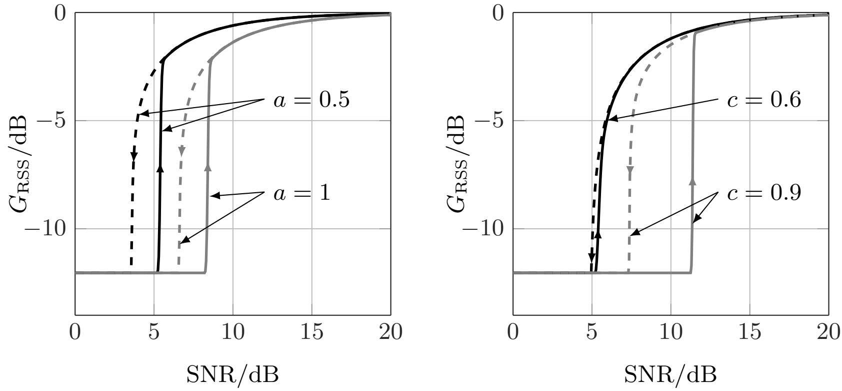
A different approach for the gain calculation was proposed by Linhard and Haulick in [LH99] using also the spectral subtraction method as shown in the general form in Equation 4.29. Their gain calculation rule was also proposed by Hofmann e.a. for wind noise suppression in [HWB$^+$12].

The motivation was to avoid single outliers during the gain calculation process, which result from estimation errors of the wind noise STPS. Therefore, a recursive calculation rule was proposed using the gain function from the previous frame $G_{\mathrm{RSS}}(\lambda - 1,\, \mu)$ for the computation of the gain in the current frame $G_{\mathrm{RSS}}(\lambda,\, \mu)$. In the Wiener filter realization of Equation 4.29 ($\alpha_{\mathrm{S}} = 2,\ \beta_{\mathrm{S}} = 1$), the recursive computation rule is given by

$$G_{\mathrm{RSS}}(\lambda,\, \mu) = \max\left\{1 - \frac{a}{\widehat{\gamma}(\lambda,\, \mu)((1-c) + c(G_{\mathrm{RSS}}(\lambda-1,\, \mu) - G_{\min}))}, G_{\min}\right\}. \tag{4.32}$$

The important part of the gain calculation is the weighting of the *a posteriori* SNR $\widehat{\gamma}(\lambda,\, \mu) = |X(\lambda,\, \mu)|^2/|\widehat{\mathcal{N}}(\lambda,\, \mu)|^2$ with a factor depending on the previous gain $G_{\mathrm{RSS}}(\lambda - 1,\, \mu)$. This relation introduces a hysteresis into the gain rule leading to different curves for increasing and decreasing SNR values. The position and width of the hysteresis range is controlled by the parameters a and c, respectively. Exemplary curves are shown in Figure 4.12. The solid and dashed lines present the progress for rising and decreasing SNR values, respectively.

In Fig 4.12a the position of the hysteresis range is shifted by the choice of a. A greater value of a results into an earlier decrease of the gain for higher SNR values, which leads to a more aggressive noise suppression. The effect of the parameter c is depicted in Figure 4.12b controlling the width of the hysteresis range. The aim of the hysteresis during the gain calculation is, that the gain function remains



(a) Varying hysteresis position
$c = 0.8$, $G_{\min} = 0.25$

(b) Varying hysteresis width
$a = 1$, $G_{\min} = 0.25$

Figure 4.12: Recursive spectral subtraction gain curves for increasing (—) and decreasing (- - -) SNR values.

longer in the state of a strong noise reduction (for increasing low SNR values) or low noise reduction (for decreasing high SNR values). Using this technique should reduce the effect of single outliers during the noise estimation procedure and the associated artifacts in the output signal such as musical tones.

4.3.3 Evaluation of the Wind Noise Reduction Performance

In this section an evaluation of the complete noise reduction system containing the single microphone wind noise estimators introduced in Section 4.2.1 and using the previously presented gain calculation rules. Different measures were proposed in the past to rate the quality of a speech enhancement system. Many can be related to the desired signal-to-interference ratios (see, e.g., [QB88]). The segmental attenuation of both, the desired speech signal (speech attenuation (SA)), and the noise signal (noise attenuation (NA)) are calculated. These are a commonly used methods for the evaluation of the performance of noise reduction systems (see, e.g., [Esc12], [Jeu12]). As a low SA and at the same time a high NA is desired, the difference NA-SA is an indicator for an improvement of the processed signal and will be used in the following to compare the investigated algorithms.[4]

The improvement in terms of the SNR or NA-SA is highly correlated with the subjective listening impression of the quality of speech signals but gives no information about the intelligibility of the speech signal. There are many discussions, whether a single microphone approach can enhance the intelligibility of speech in

[4] The computation of the SA and NA measure is explained in Appendix A.1.

general (see, e.g., [HL07]). The outcome is that for most algorithms and noise scenarios the intelligibility can not be increased and is not related to the experienced speech quality. However, in some cases an improvement of the intelligibility is measurable also by listening tests, especially for noise signals, which seems to be sparse with regard to their spectral energy distribution such as low frequency car noise [HL07]. If no improvements of the intelligibility are measurable, the applied signal processing should at least not decrease the intelligibility.

A measure for the intelligibility of noisy speech is the speech intelligibility index (SII) standardized by the American National Standards Institute (ANSI) in [ANS97]. The calculation is based on the speech level distortion in different subbands considering psycho-acoustic effects such as masking, perception thresholds and a non-uniform frequency resolution. The SII is used as a second quality measure for the evaluation of the algorithms always comparing the processed signals with the noisy input signals. The SII takes values between zero and one values higher than 0.75 indicates a good communication system while values below 0.45 correspond to a poor system.

For the evaluation, clean speech randomly chosen from the TIMIT database [LKS89] is mixed with wind noise recordings from [NV14b] according to different SNR scenarios between 15 and 15 dB. Again, the SNR range of -5 . . . 5 dB depicts the most realistic conditions for outdoor recordings but also for very low SNR ranges and almost clean speech scenarios the performance of the algorithms is of interest. Therefore, the above mentioned SNR range is considered.

Figure 4.13 shows the results in terms of the NA-SA measure for the four wind noise estimation approaches presented in Section 4.2.1. Over the complete SNR range all algorithms provide a positive NA-SA value, which demonstrates an improvement. Up to 5 dB, the approach using the pitch adaptive inverse binary mask (P-IBM) indicates the highest quality enhancement with NA-SA values of

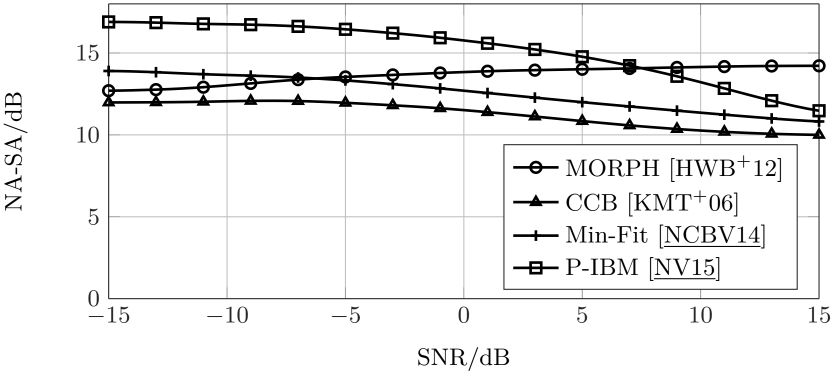


Figure 4.13: Noise attenuation - speech attenuation (NA-SA) using different noise estimators and general Wiener filter rule.

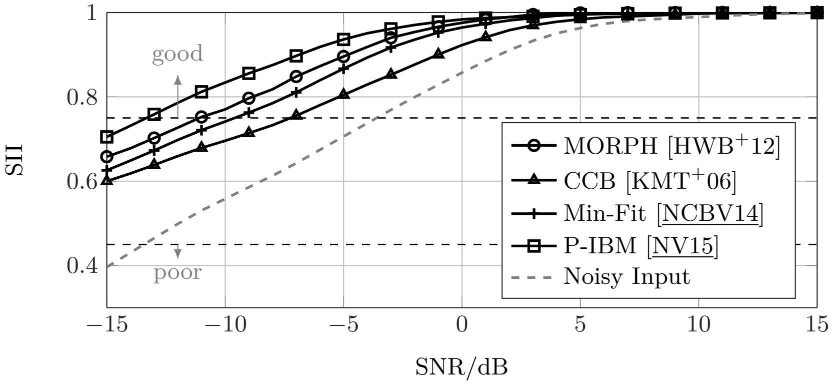


Figure 4.14: Speech intelligibility using different noise estimators and general Wiener filter rule.

over 15 dB. For higher SNR values the morphological approach (MORPH) gives a slightly better results, where in total this algorithm shows a relative constant improvement of approximately 14 dB. The minima-fitting (Min-Fit) method and the method based on the cepstral codebooks (CCB) show the smallest improvement, which are, however, not much lower than the other two noise estimators.

Considering the SII, the results of the experiments using the different wind noise estimators are shown in Figure 4.14 together with the SII of the noisy input signal represented by the dashed gray line. As for the NA-SA measure, the SII investigations confirm an improvement for all algorithms. For the complete SNR range a fixed ranking can be observed. Again the P-IBM approach shows the highest values followed by the morphological method, the Min-Fit method and the cepstral codebook algorithm. All algorithm achieve an SII value representing a good intelligibility for input SNR values greater than -7 dB, for the P-IBM method even for SNR values greater than -13 dB.

The small divergence between the ranking of the algorithms with respect to the considered measures can be explained by the fact that the audible speech quality is not always correlated with the intelligibility. A rather aggressive noise reduction can lead to lower noise attenuation - speech attenuation (NA-SA) values because of the introduced speech attenuation, but might be beneficial for the speech intelligibility. In conclusion, all noise estimators showed an improvement in terms of the quality and the intelligibility. For the most realistic wind noise scenarios the P-IBM method results in the highest improvements.

In the second part of this evaluation section, the three different approaches for gain computation are compared:

1. the recursive spectral subtraction (RSS) computation rule described in (Section 4.3.2) with the parameter $a = 0.3$ and $c = 0.75$,

2. Wiener filter using the decision directed approach (DDA) for SNR estimation (Section 4.3.1),

3. the original spectral subtraction from Equation 4.29.

The calculated spectral gains of the three algorithm are limited to the minimum gain $G_{\min} \mathrel{\widehat{=}} -40\,\mathrm{dB}$.

As only the influence of the gain computation is of interest, the best wind wind noise estimator from the previous results, the P-IBM method, is applied. The results are shown in Figures 4.15 for the NA-SA and 4.16 for the SII and

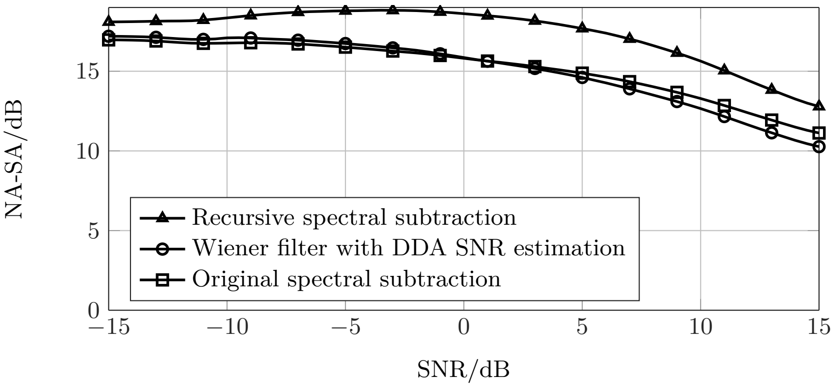


Figure 4.15: Noise attenuation - speech attenuation (NA-SA) of different gain computation rules using the P-IBM wind noise estimator.

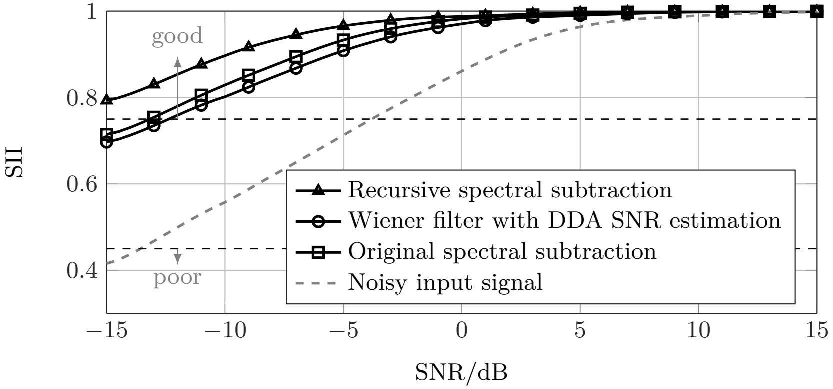


Figure 4.16: Speech intelligibility of different gain computation rules using the P-IBM wind noise estimator.

the same wind noise conditions as used before. It can be seen, that the recursive approach results in small but consistent improvements for both measures compared to the original spectral subtraction. In contrast to that, the DDA method for SNR estimation tends to slightly decrease the performance of the noise reduction system. This is a result of the smoothing of the SNR over time which lower the effect of outliers in the noise estimation procedure but leads to an inaccurate tracking speed of the fast variations of the wind noise as it was shown in Section 4.2.3.

The results of the simulations show that the combination of the pitch adaptive wind noise estimator P-IBM and the recursive spectral subtraction approach for the spectral gain calculation RSS achieves the highest noise reduction and intelligibility improvements.

4.4 Dual Microphone Wind Noise Reduction

State-of-the-art smartphones and digital hearing aids use two or more microphones for the signal acquisition and use characteristics from both recorded signals also for noise reduction purposes. For a directional processing, beamforming techniques may be used, e.g., [BW01] or [BCH08]. Because the speech signal and the wind noise signal can not be separated due to their directional properties, dual channel wind noise reduction algorithms usually exploit the correlation or more specific the differing coherence properties of speech and wind noise. The methods proposed in the past are all based on directly computing a spectral gain for the removal of wind noise without the intermediate step of a wind noise estimation. Two methods from literature will be introduced in Section 4.4.1 and 4.4.2. A novel coherence based method to estimate the wind noise STPS will be discussed in Section 4.4.3 ([NV14a]).

The angle of arrival of the desired speech signal is often determined by a specific scenario, e.g., for a mobile phone or a hearing aid in constant orientation to the speaker. Besides, methods for estimating the direction of arrival (DOA) can be applied. DOA estimation is a well studied field and an overview for applications in mobile phones can be found in [Nel09], where the cross-correlation based method by Knapp and Carter [KC76] showed the highest robustness towards noise. Further approaches can be found in [RFB81] proposing a *least-mean-square* (LMS) algorithm or [Ben00] using an adaptive eigenvalue decomposition (AED) for DOA estimation. Dependent on the DOA the microphone signals delay will be compensated. This procedure is usually carried out in a pre-processing step before the noise reduction (e.g., by a fractional delay filter [LVKL96]). For all considered approaches in this section, the DOA of the speech signal is assumed to be known and the resulting delay between the signals is compensated and is not scope of this work.

4.4.1 Coherence Weighting

Franz and Bitzer proposed a multi-microphone algorithm for wind noise reduction in [FB10]. The approach consists of two stages of which the first performs a

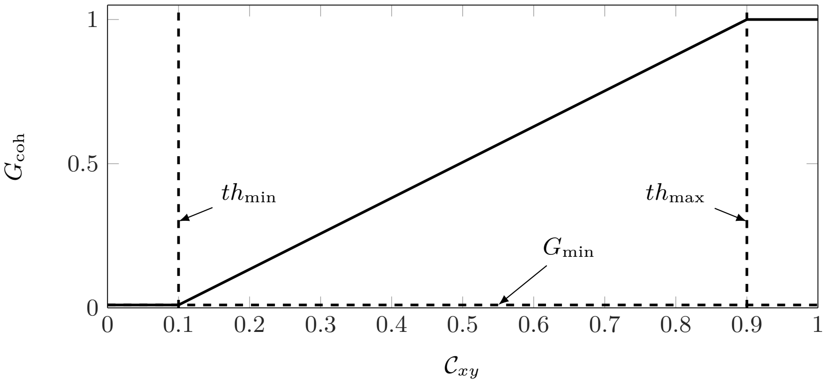


Figure 4.17: Spectral weighting based on the magnitude squared coherence $\mathcal{C}_{xy}(\lambda,\,\mu)$ as proposed in [FB10].

wind noise reduction in general. The second stage is especially designed for the application of binaural hearing aids and replaces disturbed signal parts from one monaural signal by the corresponding clean parts from the other monaural signal. Because the required shadowing from the wind for at least one microphone is usually not given, only the first stage is considered here. This stage directly uses the magnitude squared coherence (MSC) $\mathcal{C}_{xy}(\lambda,\,\mu)$ as defined in Equation 3.8 for the two microphone signals x and y. The dual microphone wind noise suppression gain is then defined as

$$G_{\mathrm{coh}}(\lambda,\,\mu) = \max\left\{\min\left\{\frac{(\mathcal{C}_{xy}(\lambda,\,\mu) - th_{\mathrm{min}})\cdot(1-G_{\mathrm{min}})}{th_{\mathrm{max}} - th_{\mathrm{min}}} + G_{\mathrm{min}}, 1\right\}, G_{\mathrm{min}}\right\}. \tag{4.33}$$

The parameters th_{max}, th_{min} and G_{min} limit the gain function as depicted in Figure 4.17. The definition for this suppression gain is motivated by the described coherence properties of speech and wind noise as shown in Section 3.3.4. In the case of wind noise the MSC $\mathcal{C}_{xy}(\lambda,\,\mu)$ tends to zero, which leads to a suppression in these frequency bins. A speech signal produces a high coherence and generates gain values close to one. The thresholds th_{min} and th_{max} allow a headroom for some fluctuations around $\mathcal{C}_{xy}(\lambda,\,\mu) = 0$ for pure wind noise and $\mathcal{C}_{xy}(\lambda,\,\mu) = 1$ for clean speech. Otherwise, variations of $\mathcal{C}_{xy}(\lambda,\,\mu)$ in these ranges will lead to some unwanted artifacts in the filtered output signal.

4.4.2 Differential Array Wind Noise Suppression

A further dual microphone method for wind noise suppression is presented by Elko in [Elk07]. Again, it is proposed to apply a spectral weighting gain, which is directly calculated from the input signals. The basic idea of this algorithm can be derived from an observation made with so-called differential arrays. They achieve a directional filtering by using the difference of two microphone signals, where the directionality can be modified by delaying and weighting of the signals [HB04]. This approach works efficiently for small microphone distances ($d_\mathrm{m} < 10\,\mathrm{cm}$) but shows a high sensitivity to uncorrelated noise in the microphone signals (see Chapter 4 in [BW01]) such as sensor self-noise or wind noise. This sensitivity is usually not desired, because instead of a noise attenuation an amplification of the uncorrelated wind noise is performed. The principle of the differential array can be used vice versa to the original approach for the detection and reduction of wind noise. Therefore, the sum and difference of the microphone short-term PSD estimates are considered as

$$\begin{aligned} \widehat{\Phi}_\mathrm{sum}(\lambda,\,\mu) &= \alpha\cdot\widehat{\Phi}_\mathrm{sum}(\lambda-1,\,\mu) + (1-\alpha)\cdot|X(\lambda,\,\mu)+Y(\lambda,\,\mu)|^2 \quad (4.34)\\ \widehat{\Phi}_\mathrm{diff}(\lambda,\,\mu) &= \alpha\cdot\widehat{\Phi}_\mathrm{diff}(\lambda-1,\,\mu) + (1-\alpha)\cdot|X(\lambda,\,\mu)-Y(\lambda,\,\mu)|^2 \quad (4.35)\end{aligned}$$

defining the power ratio

$$PR(\lambda,\,\mu) = \frac{\widehat{\Phi}_\mathrm{diff}(\lambda,\,\mu)}{\widehat{\Phi}_\mathrm{sum}(\lambda,\,\mu)}. \tag{4.36}$$

According to [Elk07], the sum and difference PSDs from Equations 4.34 and 4.35 can be expressed in terms of the coherent speech short-term PSD $\widehat{\Phi}_{ss}(\lambda,\,\mu)$ and the wind noise short-term PSD $\widehat{\Phi}_{nn}(\lambda,\,\mu)$ as

$$\begin{aligned} \widehat{\Phi}_\mathrm{sum}(\lambda,\,\mu) = {} & 4\cdot\widehat{\Phi}_{ss}(\lambda,\,\mu) + 4\cdot\widehat{\Phi}_{nn}(\lambda,\,\mu)\cdot\mathcal{C}_\mathrm{W}(\mu)\\ & + 2\cdot\widehat{\Phi}_{nn}(\lambda,\,\mu)\cdot(1-\mathcal{C}_\mathrm{W}(\mu)) + \widehat{\Phi}_{\mathrm{mic}_x}(\lambda,\,\mu) + \widehat{\Phi}_{\mathrm{mic}_y}(\lambda,\,\mu) \quad (4.37)\end{aligned}$$

$$\begin{aligned} \widehat{\Phi}_\mathrm{diff}(\lambda,\,\mu) = {} & 4\cdot\widehat{\Phi}_{ss}(\lambda,\,\mu)\cdot\sin^2\left(\frac{\pi\tilde{d}_\mathrm{m}\mu f_\mathrm{s}}{cM}\right)\\ & + 4\cdot\widehat{\Phi}_{nn}(\lambda,\,\mu)\cdot\mathcal{C}_\mathrm{W}(\lambda,\,\mu)\sin^2\left(\frac{\pi\tilde{d}_\mathrm{m}\mu f_\mathrm{s}}{UM}\right)\\ & + 2\cdot\widehat{\Phi}_{nn}(\lambda,\,\mu)\cdot(1-\mathcal{C}_\mathrm{W}(\mu)) + \widehat{\Phi}_{\mathrm{mic}_x}(\lambda,\,\mu)+\widehat{\Phi}_{\mathrm{mic}_y}(\lambda,\,\mu) \quad (4.38)\end{aligned}$$

with the coherence function $\mathcal{C}_\mathrm{W}(\mu)$ of wind noise. The self-noise of the two microphone signals is expressed by the PSDs $\widehat{\Phi}_{\mathrm{mic}_{x|y}}(\lambda,\,\mu)$. Neglecting the self-noise and assuming a zero coherence of wind noise $\mathcal{C}_\mathrm{W}(\lambda,\,\mu) = 0$ (cf. Equation 3.12), the power ratio in Equation 4.36 in the case of pure wind noise ($\widehat{\Phi}_{ss}(\lambda,\,\mu) = 0$) turns to

$$PR_n(\mu) = 1 \tag{4.39}$$

and in the case of a clean coherent speech signal ($\widehat{\Phi}_{nn}(\lambda,\,\mu) = 0$) to

$$PR_s(\mu) = \sin^2\left(\frac{\pi\,\tilde{d}_\mathrm{m}\,\mu f_\mathrm{s}}{cM}\right), \tag{4.40}$$

which is only dependent on the effective microphone distance $\tilde{d}_{\mathrm{m}}$. This distance is defined by the angle θ between the microphone axis and the incident direction of the speech signal

$$\tilde{d}_{\mathrm{m}} = \cos(\theta) \cdot d_{\mathrm{m}}. \tag{4.41}$$

With the aforementioned assumption of delay compensated speech signals, i.e., $\theta = 90°$, follows $\tilde{d}_{\mathrm{m}} = d_{\mathrm{m}}$. The resulting power ratios for wind noise and coherent speech signals are shown in Figure 4.18 by the dashed and solid curves, respectively, where for the speech signal three different effective microphone distances are considered. It can be seen that the distinction between speech and wind noise improves with smaller microphone distances. The suppression gain $G_{\mathrm{diff}}(\lambda,\, \mu)$ to reduce the wind noise in speech signal is stated in [Elk07] as the ratio between the theoretical power ratio for speech in Equation 4.40 and the measured power ratio in the current frame $PR(\lambda,\, \mu)$ from Equation 4.36

$$G_{\mathrm{diff}}(\lambda,\, \mu) = \frac{PR_s(\mu)}{PR(\lambda,\, \mu)}. \tag{4.42}$$

The idea is to reduce the noisy input signal by the ratio between the measured power ratio $PR(\lambda,\, \mu)$ and the predicted power ratio $PR_s(\mu)$ for a clean speech signal. From Equation 4.40 and the curves in Figure 4.18, it can be seen that the separation between speech and wind noise works better the smaller the microphone distance is. But even for a relative big microphone distance of 10 cm, a reasonable separation in the relevant frequency range below 1500 Hz is ensured.

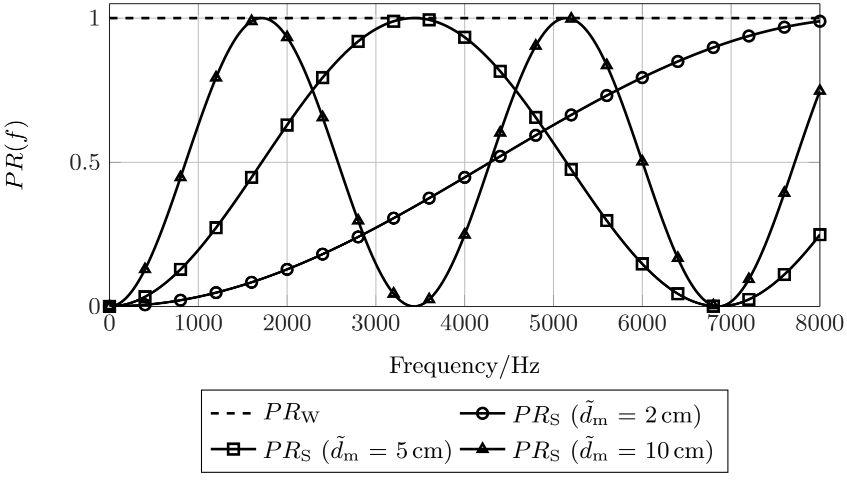


Figure 4.18: Power ratios for wind (PR_{W}, Equation 4.39) and speech (PR_s, Equation 4.40) for different microphone distances.

4.4.3 Coherence Based Wind Noise Estimation

In contrast to the two aforementioned methods, which directly compute a suppression gain, the algorithm proposed in [NV14a] first performs a noise estimation and then applies a noise reduction based on a spectral weighting. This separation of wind noise estimation and reduction can be useful as the choice of the subsequent gain calculation gives an additional degree of freedom for the design of the speech enhancement system. Furthermore, the noise estimate can be combined with other disturbance estimates, e.g., background noise, acoustic echo or reverberation.

For the noise estimation also the low coherence of wind noise and the high coherence of speech is considered. In [DE96], Dörbecker proposed a noise estimator for a dual microphone system expecting uncorrelated, i.e., incoherent background noise signals. The dual microphone signal model in DFT domain is given by[5]

$$\begin{aligned} X(\lambda,\,\mu) &= S(\lambda,\,\mu)\cdot H_1(\lambda,\,\mu) + N_1(\lambda,\,\mu) \qquad (4.43)\\ Y(\lambda,\,\mu) &= S(\lambda,\,\mu)\cdot H_2(\lambda,\,\mu) + N_2(\lambda,\,\mu). \qquad (4.44)\end{aligned}$$

Equal noise power levels

$$\widehat{\Phi}_{n_1n_1}(\lambda,\,\mu) \approx \widehat{\Phi}_{n_2n_2}(\lambda,\,\mu) \approx \widehat{\Phi}_{nn}(\lambda,\,\mu), \qquad (4.45)$$

and similar transfer functions $H_{1|2}(\lambda,\,\mu)$ of the desired speech signal

$$|H_1(\lambda,\,\mu)| \approx |H_2(\lambda,\,\mu)| \approx |H(\lambda,\,\mu)|, \qquad (4.46)$$

are assumed in both microphones. Then, the magnitude squared cross power spectrum can be expressed for uncorrelated noise signals $N_1(\lambda,\,\mu)$ and $N_2(\lambda,\,\mu)$ as

$$|\widehat{\Phi}_{xy}(\lambda,\,\mu)|^2 = \widehat{\Phi}_{ss}(\lambda,\,\mu)^2 \cdot |H(\lambda,\,\mu)|^4 \qquad (4.47)$$

and the product of the power spectra of each microphone signal can be written as

$$\widehat{\Phi}_{xx}(\lambda,\,\mu)\cdot\widehat{\Phi}_{yy}(\lambda,\,\mu) = \left(\widehat{\Phi}_{nn}(\lambda,\,\mu) + |H(\lambda,\,\mu)|^2\cdot\widehat{\Phi}_{ss}(\lambda,\,\mu)\right)^2. \qquad (4.48)$$

Taking the square root of Equations 4.47 and 4.48, they can be combined and rearranged for an estimate of the noise PSD

$$\widehat{\Phi}_{nn,\mathrm{Coh}}(\lambda,\,\mu) = \sqrt{\widehat{\Phi}_{xx}(\lambda,\,\mu)\cdot\widehat{\Phi}_{yy}(\lambda,\,\mu)} - |\widehat{\Phi}_{xy}(\lambda,\,\mu)|, \qquad (4.49)$$

where the short-term estimates of the PSDs are defined by the recursive smoothing approach as

$$\widehat{\Phi}_{xy}(\lambda,\,\mu) = \alpha\cdot\widehat{\Phi}_{xy}(\lambda-1,\mu) + (1-\alpha)\cdot X(\lambda,\,\mu)\cdot Y^*(\lambda,\,\mu). \qquad (4.50)$$

The noise estimate from Equation 4.49 can be used for the subsequent speech enhancement, e.g., based on a spectral weighting. However, it has the drawback that

[5] The influence of the impulse responses $h_1(k)$ and $h_2(k)$ (see Equations 2.1 and 2.2) is modeled by the transfer functions $H_1(\lambda,\,\mu)$ and $H_2(\lambda,\,\mu)$.

due to the smoothing process for the computation of the PSDs, the aforementioned problem of an slow adaptation occurs as demonstrated in Section 4.2.3. In [DE96] a smoothing constant close to one ($\alpha = 0.96$) is proposed to reduce the variance of the estimated PSDs, which is sufficient to follow the characteristics of general background noise types, but may introduce high estimation errors in the case of highly non-stationary wind noise (see Figure 4.10).

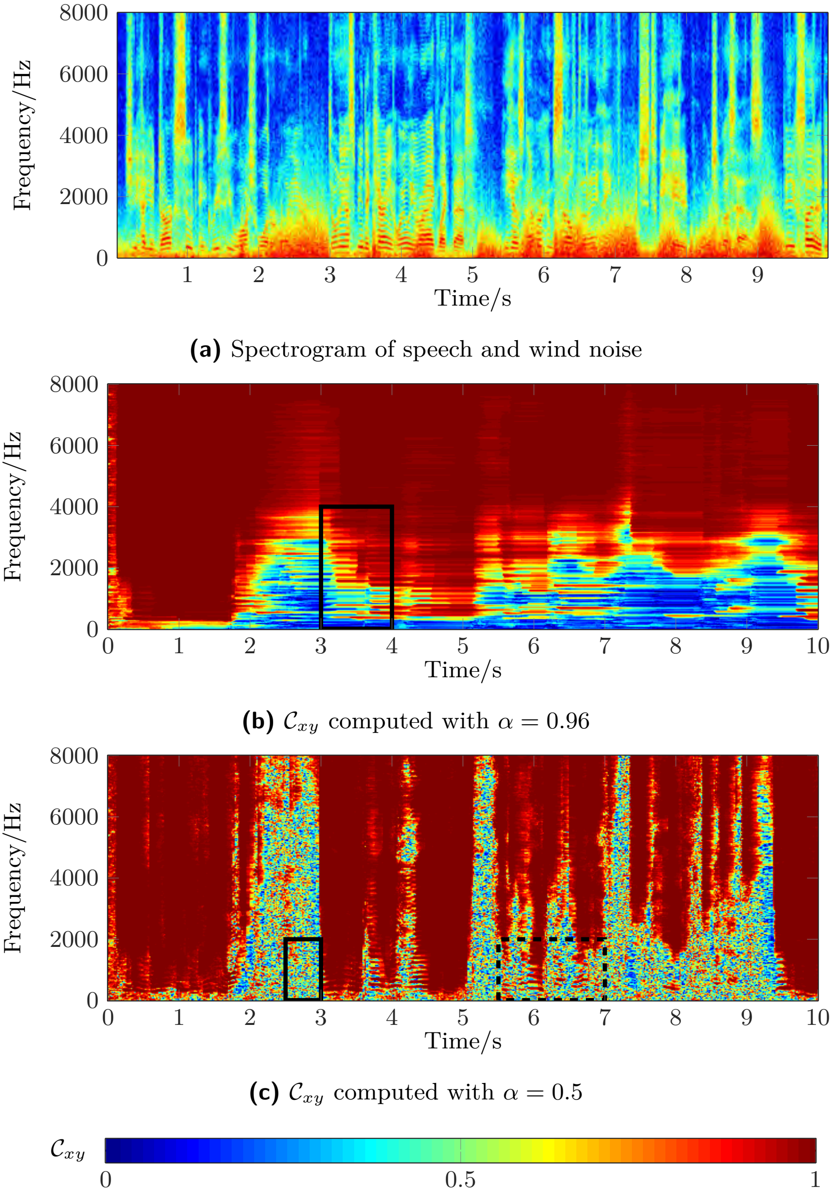
The effect of the smoothing constant on the dual microphone signals is investigated and shown in Figure 4.19. A sequence of 10 seconds of speech is mixed with wind noise signals taken from [NV14b], where the recordings were carried out with a dual microphone mock-up phone with a microphone distance of 10 cm.

In 4.19a the spectrogram of the noisy speech of one microphone signal is presented while 4.19b and 4.19c shows the MSC values over time and frequency using smoothing constants of $\alpha = 0.96$ and $\alpha = 0.5$, respectively. For the illustration of the coherence, the red areas represent parts with high coherence close to one while the blue areas depict incoherent segments. As the MSC $\mathcal{C}_{xy}$ is the normalized version of the cross-PDS of the two signals $x(k)$ and $y(k)$ (see Equation 3.8), the performance of the noise estimate in Equation 4.49 can be predicted from the accuracy of the MSC calculation. In Figure 3.8 it was shown, that the scenarios of clean speech and pure wind noise are characterized by $\mathcal{C}(\lambda, \mu) = 1$ and $\mathcal{C}(\lambda, \mu) = 0$, respectively. Consequently, an overestimation, e.g., $\alpha = 0.5$, of the MSC and thus of the cross-PSD leads to an underestimation of the wind noise in Equation 4.49. In the same way an underestimation of the MSC, e.g., $\alpha = 0.96$, leads to a too high wind noise estimate. Thus, the two effects displayed in Figure 4.19 have a great influence on the accuracy of the wind noise estimate.

The trade-off between variance reduction and estimation accuracy in terms of the tracking speed is clearly visible. On the one hand the choice of $\alpha = 0.96$ in Figure 4.19b results in a good estimation of the true values of the MSC, e.g., in the case of wind noise as the blue areas indicating the expected low coherence in the low-frequency range ($f < 1000\,\mathrm{Hz}$). The high smoothing constant cause a smearing of the MSC values over time, which is clearly visible at $t = 3\,\mathrm{s}$ (black box), where a speech segment begins with only low wind noise energy but blue areas still indicate a low coherence. On the other hand, the coherence in Figure 4.19c computed with a low smoothing constant ($\alpha = 0.5$) shows a direct adaptation at this speech onset with a high coherence. Here, even the harmonic structure of voiced speech segments is visible during wind noise highlighted by the dashed black box. The drawback of the high variance in the estimate becomes apparent in segments, where only wind is active, e.g., in the solid black box before $t = 3\,\mathrm{s}$.

The slow adaptation of the coherence for $\alpha = 0.96$ might be negligible in the case of stationary or only slowly varying noise signals, which were assumed in the original approach [DE96], but deteriorate the performance of noise estimators for non-stationary noise such as wind noise. Therefore, in the following two strategies are proposed concerning this problem leading to an improved wind noise estimation.

Both approaches are further developments of the original approach from [DE96] by exploiting not only the magnitude of the coherence function as before, but also



(a) Spectrogram of speech and wind noise

(b) $\mathcal{C}_{xy}$ computed with $\alpha = 0.96$

(c) $\mathcal{C}_{xy}$ computed with $\alpha = 0.5$

Figure 4.19: Short-term coherence for different smoothing constants.

the phase of the complex coherence. As evident in Equation 3.7 the phase of the complex coherence $\Gamma_{xy}(\lambda, \mu)$ is only affected by the cross-PSD $\widehat{\Phi}_{xy}(\lambda, \mu)$, because the auto PSDs are always real-valued. Choosing the smoothing constant $\alpha = 0$ for the calculated cross PSD leads to the phase in each frame

$$\varphi_\Gamma(\lambda, \mu) = \angle\{\widehat{\Phi}_{xy}(\lambda, \mu)\} = \angle\{X(\lambda, \mu)\} - \angle\{Y(\lambda, \mu)\}, \tag{4.51}$$

which is determined as phase difference between the two input signals $X(\lambda, \mu)$ and $Y(\lambda, \mu)$. For a coherent signal the phase difference is only dependent on the DOA of this signal. A not compensated delay τ between the signals will generate a linear phase function

$$\varphi_\Gamma(\lambda, \mu) = \frac{2\pi\mu\tau f_\mathrm{s}}{M}. \tag{4.52}$$

The measured phase of the coherence of the same signals as in Figure 4.19 is represented in Figure 4.20 in a time-frequency representation for compensated DOA, i.e., $\tau = 0$. The zero phase of the speech signal is clearly visible by the green areas in the undisturbed segments, while parts of the signal in which wind is dominant the phase takes random values in the interval $-\pi \ldots \pi$.

As mentioned before the DOA is assumed to be known and the corresponding delay is compensated ($\tau = 0$). For a mixed signal containing similar speech and noise levels in each microphone signal

$$|S_1(\lambda, \mu)| \approx |S_2(\lambda, \mu)| = S(\lambda, \mu) \tag{4.53}$$

$$|N_1(\lambda, \mu)| \approx |N_2(\lambda, \mu)| = N(\lambda, \mu) \tag{4.54}$$

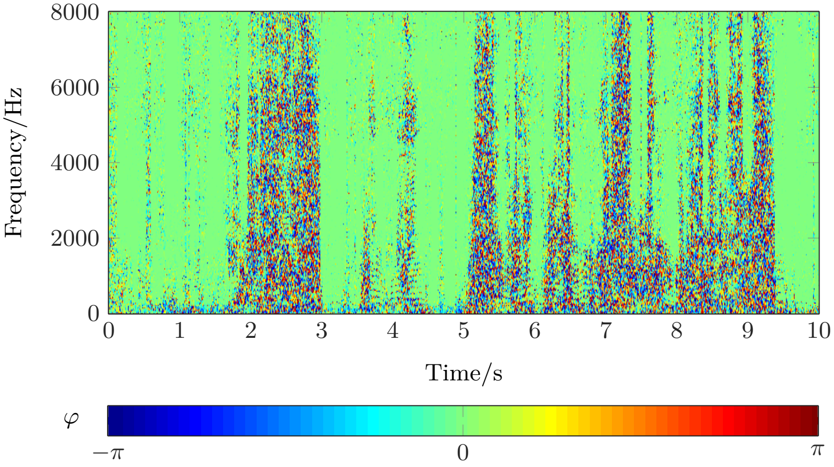


Figure 4.20: Coherence phase $\angle\{X(\lambda, \mu)\cdot Y^*(\lambda, \mu)\}$ of speech and wind noise.

the coherence phase can be expressed as[6]

$$\varphi_\Gamma = \arctan\left(\frac{|S||N|(\sin(\varphi_{s1}-\varphi_{n2})+\sin(\varphi_{n1}-\varphi_{s2}))+|N|^2\sin(\varphi_{n1}-\varphi_{n2})}{|S|^2+|S||N|(\cos(\varphi_{s1}-\varphi_{n2})+\cos(\varphi_{n1}-\varphi_{s2}))+|N|^2\cos(\varphi_{n1}-\varphi_{n2})}\right). \tag{4.55}$$

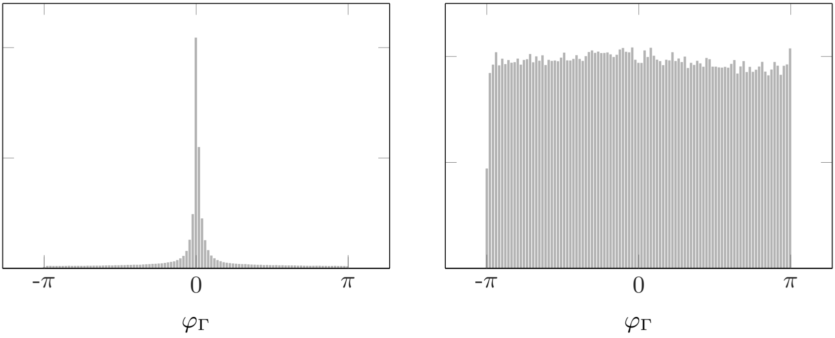
For the sake of brevity the frequency and time indices are omitted in this equation. A direct relation between the SNR ($|S|^2/|N|^2$) and the phase φ_Γ is not possible, since the phases of speech signals $\varphi_{s1|2}$ and noise signals $\varphi_{n1|2}$ are randomly distributed and unknown. However, it can be seen that in the case of pure wind noise ($S = 0$) φ_Γ takes the value of the difference of noise phases

$$\varphi_{\Gamma,\text{wind}} = \varphi_{n1}-\varphi_{n2} \tag{4.56}$$

and in the case of clean speech ($N = 0$)

$$\varphi_{\Gamma,\text{speech}} = 0. \tag{4.57}$$

The measured distribution of the phase in the case of wind noise and clean coherent speech is shown in Figure 4.21. As expected, the zero phase behaviour of the speech is apparent and result in a peak at $\varphi_\Gamma = 0$. For wind noise a uniform distribution of the phase between $-\pi$ and π is given. This property is exploited in the following by two proposals for advanced wind noise estimation using dual microphone signals.



(a) Speech

(b) Wind noise

Figure 4.21: Phase distribution of wind noise and speech signals.

4.4.3.1 Decision Directed Wind Noise Estimation

As shown in Equation 4.55, a single phase value of φ_Γ of the coherence in a one time-frequency bin can not be mapped to the degree of distortion. Therefore, the

[6] The derivation of Equation 4.55 can be found in Appendix B.

distribution of the phase within one signal frame is investigated in [NV14a] to further develop the approach by Dörbecker [DE96]. As shown in Figure 4.21 the distribution of the noise phase follows a uniform distribution, which is in general characterized by a variance of $A^2/3$ for a range of values between $-A \dots A$. For the variance of the phase normalized by $\pi^2/3$

$$\sigma_\varphi^2(\lambda) = \frac{3}{\pi^2} \sum_{\mu=1}^{\mu_\mathrm{c}} \frac{\varphi_\Gamma(\lambda,\, \mu)^2}{\mu_\mathrm{c}+1} \tag{4.58}$$

follows that $\sigma_\varphi^2(\lambda)$ takes a value of one for a uniform distribution between $-\pi \dots \pi$ in the case of wind noise. For the zero-phase segments of the clean speech short-term cross PSD $\widehat{\Phi}_{xy}(\lambda,\, \mu)$, values close to zero are expected. The frequency limit for the variance computation defined by μ_c should be chosen to a range in which both wind noise and speech are active, e.g., to $0 \dots 4000\,\mathrm{Hz}$. The variance of the phase information represents a wind and speech indicator and can be used to update the noise estimate similar as proposed by Ephraim and Malah for the decision directed signal-to-noise-ratio (SNR) estimation scheme [EM84]. Here, the phase variance is applied as parameter defining the cross-fade factor between the noise estimate $\hat{\Phi}_{nn}(\lambda,\, \mu)$ given in Equation 4.49 and the input signal $X(\lambda,\, \mu)$ as

$$|\widehat{\mathcal{N}}_\mathrm{DDWE}(\lambda,\, \mu) = (1-\sigma_\varphi^2(\lambda)) \cdot \hat{\Phi}_{nn,\mathrm{Coh}}(\lambda,\, \mu) + \sigma_\varphi^2(\lambda)|X(\lambda,\, \mu)|^2. \tag{4.59}$$

Here, the smoothing constant for the computation of the PSDs is chosen to $\alpha = 0.5$ to allow a fast adaptation to changes in the wind noise characteristic. The cross-fade presented in Equation 4.59 circumvent the issue of overestimating the coherence in noise only segments ($\sigma_\varphi^2(\lambda) \to 1$) as visualized by the red areas in Figure 4.19c by taking directly the input spectrum as noise estimate. In this way the problem of underestimating the noise signal in speech pauses resulting from the aforementioned overestimation of the coherence is bypassed.

4.4.3.2 Adaptive Smoothing Factor for Improved Coherence Estimation

The second proposed advance is a modified calculation of the cross and auto PSDs in Equation 4.50, which are required for the coherence estimation. As discussed above the coherence estimation heavily depends on the choice of the smoothing factor α. For an exact coherence value in segments containing only wind noise, α should be close to one but a smaller α ensures a fast adaptation, e.g., at the beginning of speech activity. This trade-off is bypassed by an adaptive smoothing factor based on the phase variance $\sigma_\varphi^2(\lambda)$ calculated in each frame as an indicator for the predominant signal component (speech or wind noise). The adaptive smoothing factor is determined by a sigmoid characteristic as

$$\alpha_\mathrm{ad}(\lambda) = \frac{1}{1-(1-\sigma_\varphi^2(\lambda))^2} \tag{4.60}$$

and the relation is shown in Figure 4.22. A similar relation was previously proposed in [Mar01] for an optimal smoothing parameter in dependency of the *a posteriori* SNR.

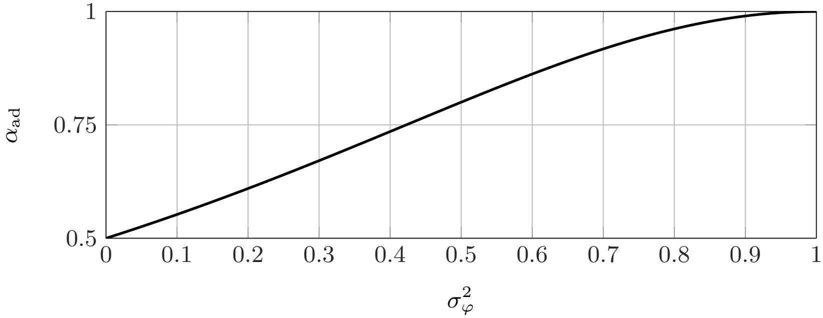


Figure 4.22: Mapping between phase variance and adaptive smoothing factor.

The adaptive smoothing factor is automatically limited to the range $0.5 \ldots 1$ and guarantees a fast adaptation to coherent speech parts ($\sigma_\varphi^2(\lambda) \to 0$) and a low variance during wind activity ($\sigma_\varphi^2(\lambda) \to 1$). The resulting coherence computed with the adaptive smoothing factor is shown in Figure 4.23. An improvement compared to both coherence plots with constant smoothing factors in Figure 4.19 is clearly visible. A fast adaptation at speech onsets is given, e.g., at $t = 3\,\mathrm{s}$ given by the sharp red edge of the red area. At the same time low coherence values indicated by the blue areas arises at segments with pure wind noise, e.g., at $t = 5 \ldots 5.5\,\mathrm{s}$. The adaptive smoothing parameter can now be used for the computation of the cross- and auto-PSDs required for the noise estimate in Equation 4.49.

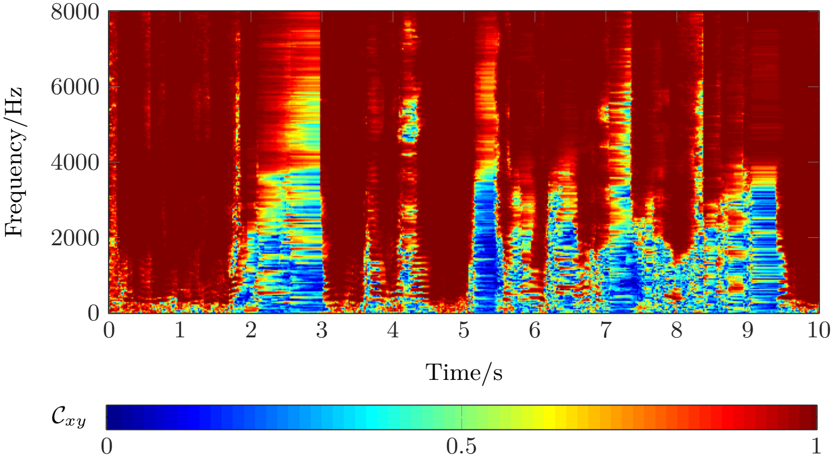


Figure 4.23: Coherence computation with adaptive smoothing factor α_{ad}.

4.4.3.3 Estimation Accuracy of Dual Microphone Wind Noise Estimation

In this section the discussed variants of coherence based wind noise estimation schemes are evaluated and compared by means of their accuracy measured by the logarithmic error $e_{\log}$ (see Equation A.4). For this evaluation dual microphone noise recordings from [NV14b] are mixed with speech recordings made with a dual microphone mock-up phone with a microphone distance of 10 cm. Figure 4.24 presents the results for the original approach by Dörbecker and the two advancements described in the previous section.

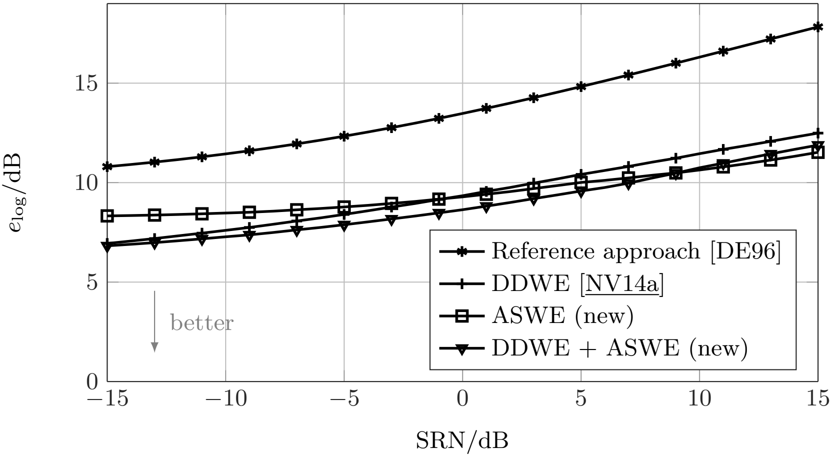


Figure 4.24: Wind noise estimation accuracy of dual microphone approaches.

The decision directed wind noise estimation (DDWE) is defined by Equation 4.59 and for the adaptive smoothing approach adaptive smoothing wind noise estimation (ASWE) the noise estimate is computed as proposed by Dörbecker but with adaptive smoothing constants. Besides, the combination of both advancements (DDWE + ASWE) is also taken into account during the evaluation procedure. The advancements of the original approach yield in a significant improvement for all considered input SNRs indicated by a decrease of the logarithmic error between 5 and 7 dB. If only one modification is considered, the phase-based cross-fading shows a better performance. The combination of both concepts generates only a marginal lower logarithmic error than one of the modifications. The small improvement of the combination arises from the fact that both methods uses the additional information gained from the phase in a similar way for the update of noise estimate.

4.4.4 Evaluation of Dual Microphone Wind Noise Reduction

As already introduced for the single microphone solutions the performance of all considered dual microphone wind noise reduction concepts is compared by

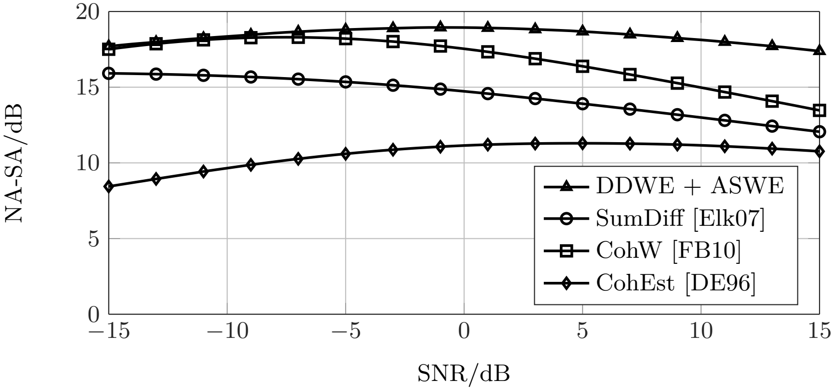


Figure 4.25: Noise reduction performance of dual microphone systems.

the NA-SA measure and the SII. The coherence based weighting (CohW) from Section 4.4.1, the differential array approach (SumDiff) from Section 4.4.2, and the original coherence based noise estimator by Dörbecker (CohEst) are evaluated. The proposed wind noise estimator exploiting the phase information is used in the realization, where the combination of the adaptive smoothing and the noise cross-fade (DDWE + ASWE) is taken into account. The methods providing a wind noise estimate are applied with the modified spectral subtraction gain rule explained in Section 4.3.2, as this method showed the highest improvements in Section 4.3.3. The results are depicted in in Figures 4.25 and 4.26.

For the NA-SA values, the combined phase based wind estimation (DDWE +

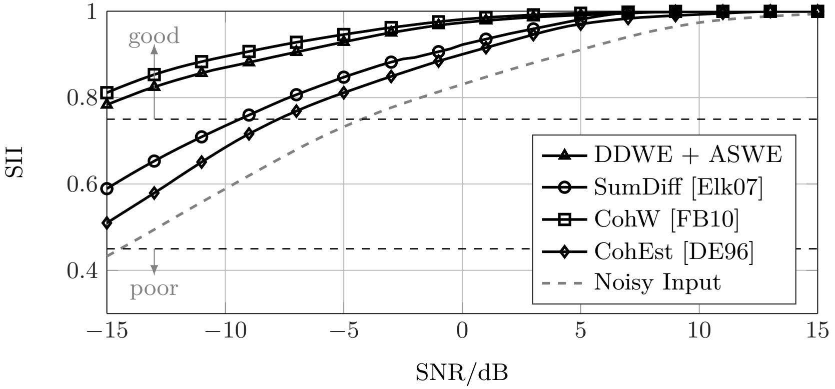


Figure 4.26: Intelligibility performance of dual microphone systems.

ASWE) scheme achieves the highest performance. The SumDiff and CohW methods also show high NA-SA values for all considered SNR scenarios. As expected, the original coherence based approach (CohEst) results in the lowest performance. The predicted speech intelligibility in terms of SII indicates similar high improvements for the proposed wind estimator (DDWE + ASWE) and the coherence based spectral weighting (CohW), where the CohW method shows the highest SII values for low SNRs. The original approach (CohEst) again only leads to a small SII gain.

In conclusion, all algorithms achieve an improvement in terms of the depicted measures. The proposed method for wind noise estimation clearly outperforms the original coherence based approach for noise estimation. Here, the improvements can be realized by exploiting the phase information of the two microphone signals. The new method can also shows a better performance then the two methods from literature for dual microphone wind noise reduction, if both measures the SII and NA-SA are both of interest.

4.5 Wind Noise Reduction via Partial Speech Synthesis

So far, the conventional realization for a noise reduction system by means of a spectral weighting as introduced in Figure 2.4 is considered in this chapter. In this section a new alternative approach is introduced to enhance a distorted speech signal as shown in Figure 4.27.

The analysis and synthesis of the framework is again implemented as an overlap-add structure by first segmenting and windowing the time-domain signal and transforming it into the DFT domain. Subsequently, two steps are proposed to enhance the noisy input spectrum $X(\lambda, \mu)$:

1. wind noise reduction (WNR) stage yielding the spectrum $\widetilde{X}(\lambda, \mu)$,
2. speech synthesis stage generating a synthetic speech spectrum $\widetilde{S}(\lambda, \mu)$.

Both signals are combined, leading to an estimate $\widehat{S}(\lambda, \mu)$ of the clean speech signal. The motivation of this alternative design is given by the fact that even the best

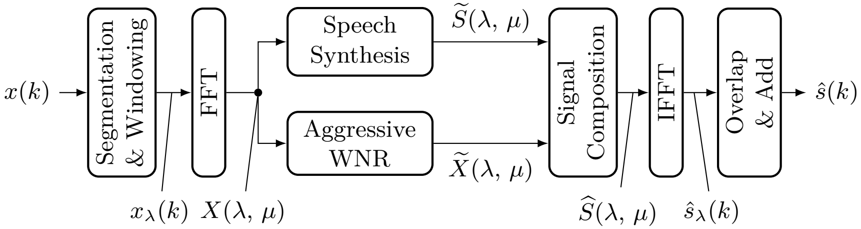


Figure 4.27: Alternative speech enhancement system.

candidates of the wind noise reduction systems presented in Secs. 4.2-4.4 tend to introduce a high-pass effect to the filtered speech. This is due to the extreme low SNR conditions at low frequencies. To overcome this issue, the synthesis stage is incorporated into the process of speech enhancement. The signals $x_\lambda(k)$ and $\hat{s}_\lambda(k)$ denote the segmented time-domain signals in the current frame λ of the input signal and the enhanced output signal, respectively.

An initial version of this algorithm is proposed in [NNJ+12], which applies a technique similar to an artificial band width extension (ABWE) to the noisy speech signal. This system is further developed in [NNV15] incorporating knowledge about the speech signal characteristics in terms of pre-trained codebooks. Both methods will be presented in the following Secs. 4.5.1 and 4.5.2.

4.5.1 Reconstruction Based on Partial Synthesis

The basic concept proposed in [NNJ+12] is to consider the distorted lower frequency parts of a speech corrupted by wind noise as missing parts of the speech resulting into a band-limited signal. The problem of enhancing band-limited speech is a well studied objective in the case of speech coding. Heterogeneous communication networks do not allow a transmission of the full frequency range, even though parts of the network are capable to transmit the considered speech with a wide frequency range. This problem is solved by the so-called artificial band width extension (ABWE), where the missing parts of the signal are reconstructed using *a priori* knowledge and statistical models for speech signals (see, e.g., [Jax02], [Gei12]).

The system, which is designed to reconstruct the missing or highly disturbed

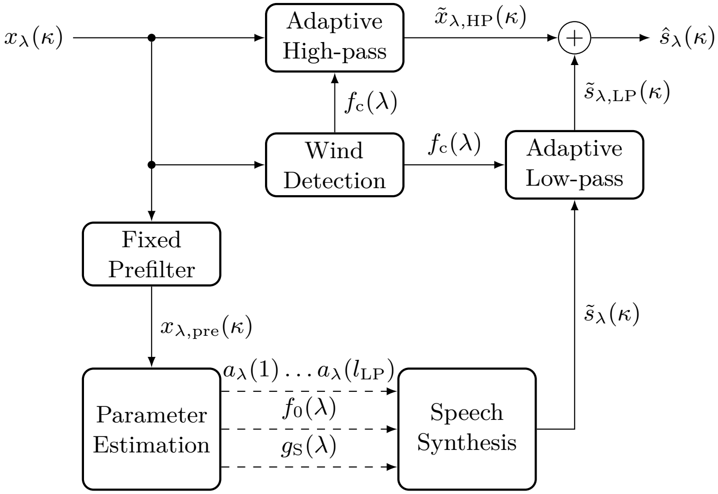

$x_\lambda(\kappa)$
Adaptive High-pass
$\tilde{x}_{\lambda,\mathrm{HP}}(\kappa)$
+
$\hat{s}_\lambda(\kappa)$
$f_\mathrm{c}(\lambda)$
$\tilde{s}_{\lambda,\mathrm{LP}}(\kappa)$
Wind Detection
$f_\mathrm{c}(\lambda)$
Adaptive Low-pass
Fixed Prefilter
$x_{\lambda,\mathrm{pre}}(\kappa)$
$\tilde{s}_\lambda(\kappa)$
$a_\lambda(1)\dots a_\lambda(l_\mathrm{LP})$
Parameter Estimation
$f_0(\lambda)$
$g_\mathrm{S}(\lambda)$
Speech Synthesis

Figure 4.28: Wind noise reduction using partial speech synthesis (PSYN).

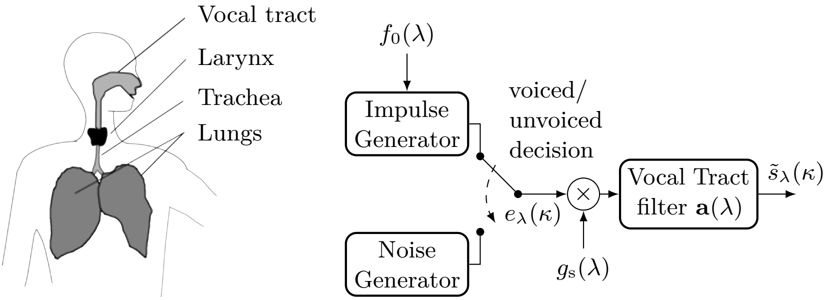
parts of the speech signal, is shown in Figure 4.28. For the sake of clarity the analysis and synthesis parts of the framework shown in Figure 4.27 are omitted.

Speech Synthesis

The core part of the system is the speech synthesis block to generate a synthetic noise-free speech signal. Here, the frequently used source-filter model is applied [VM06]. This model is derived from the process of speech generation in the human body and is depicted in Figure 4.29.

The most important organs of speech production are highlighted in Figure 4.29a. The airflow produced by the lungs is modulated by the larynx, where the vocal chords generate the so-called excitation signal. This is either a periodic signal or a noise-like signal. The vocal tract, consisting of the mouth, nose and throat, acts as an acoustic resonator and performs a filtering, i.e., a spectral shaping of the excitation signal. The filtered signal is then radiated via the lips and the nostrils. The periodic parts of the excitation signal are voiced speech segments resulting in vowels while the noise-like excitation leads to unvoiced speech such as fricatives.

Although there are several more categories of speech, e.g., plosive or mixed segments, the partitioning into voiced and unvoiced speech leads to the widely used source-filter model for speech production in Figure 4.29b. The equivalent to the excitation is represented by either an impulse generator or a noise generator for voiced or unvoiced sounds, respectively. The time lag between the impulses for voiced segments is determined by the pitch period T_0 or the fundamental frequency $f_0 = 1/T_0$ and the noise-like signal can be given by, e.g., a white noise signal. As discussed earlier and demonstrated in Section 4.2.1, wind noise mainly shows a spectral overlap with voiced speech. Therefore, unvoiced speech can be separated by



(a) Organs of speech production

(b) Source-filter model producing a synthetic speech signal $\tilde{s}_\lambda(\kappa)$

Figure 4.29: Generation of voice in human body and digital source-filter model.

a simple high-pass filter as realized in the upper branch signal $\tilde{x}_\lambda(k)$ in the proposed system in Figure 4.28. Consequently, the source-filter model is only employed to produce voiced speech segments. The influence of the vocal tract is simulated by the filter with the time-varying coefficient vector

$$\mathbf{a}(\lambda) = [a_\lambda(1), \ldots, a_\lambda(l_{\mathrm{LP}})] \tag{4.61}$$

of order l_{LP}.

For the generation of the artificial speech, several steps are necessary. First, the excitation impulse train in the current voiced speech frame of length L_{F} is defined by

$$e_\lambda(\kappa) = \sum_{i=0}^{M_0-1} \delta(\kappa - i \cdot N_0), \ \kappa = 0, \ldots, L_{\mathrm{F}}, \tag{4.62}$$

with the discrete equivalent of the pitch period

$$N_0 = \lceil T_0 \cdot f_{\mathrm{s}} \rfloor = \lceil f_{\mathrm{s}}/f_0 \rfloor, \tag{4.63}$$

and

$$M_0 = \lfloor L_{\mathrm{F}}/N_0 \rfloor \tag{4.64}$$

is the number of pitch cycles in one signal frame. The index κ represents the sample position within the current frame λ. The signal power is controlled by the time varying gain $g_{\mathrm{s}}(\lambda)$ resulting into the weighted excitation signal

$$\tilde{e}_\lambda(\kappa) = g_{\mathrm{s}}(\lambda) \cdot e_\lambda(\kappa). \tag{4.65}$$

A digital filter models the effect of the vocal tract on the excitation signal. It is realized by the linear predictive coding (LPC) synthesis filter as an all-pole filter with the coefficients $a_\lambda(i)$. The output of the source-filter model is the synthetic speech signal

$$\tilde{s}_\lambda(\kappa) = \tilde{e}_\lambda(\kappa) + \sum_{i=1}^{l_{\mathrm{LP}}} \tilde{s}_\lambda(\kappa - i) \cdot a_\lambda(i), \tag{4.66}$$

where l_{LP} is the linear prediction (LP) order. For the considered application in the system presented in Figure 4.28, a frame-wise processing is necessary, therefore all quantities of the model for the speech synthesis are dependent on the frame index λ and must be estimated each frame.

Parameter Estimation

In the proposed system in Figure 4.28, all parameters for the speech synthesis are estimated by first applying a fixed pre-filter, which reduces the influence of the wind noise on the speech signal features. A high-pass filter with a cut-off frequency of 200 Hz and a high slope steepness to ensure that the low frequency effects of the wind noise are strongly reduced. In the considered implementation, this is achieved by a high-order finite impulse response (FIR) filter $h_{\mathrm{pre}}(k)$ of 160 taps in the case of $f_{\mathrm{s}} = 16\,\mathrm{kHz}$.

The all-pole filter for the vocal tract filter are represented by LPC coefficients. For the estimation of the predictor coefficients $a_\lambda(1)\dots a_\lambda(l_{\mathrm{LP}})$ an efficient algorithm is the Levinson-Durbin recursion ([Lev47], [Dur60]), which is applied on the pre-filtered noisy input signal

$$x_{\lambda,\mathrm{pre}}(k) = h_{\mathrm{pre}}(k) * x_\lambda(k). \tag{4.67}$$

The order of the vocal tract filter represented by $a_\lambda(1)\dots a_\lambda(l_{\mathrm{LP}})$ was found to be sufficiently high for $l_{\mathrm{LP}} = 20$ ([VM06]).

There exists a large number of methods for the estimation of the fundamental frequency or the pitch period of speech signals. Thorough investigations have shown that algorithms working in the frequency-domain yield most robust results in case of wind noise disturbance. Here the harmonic product spectrum (HPS) method is applied for pitch estimation, which was introduced in Section 4.2.2.2 and defined in Equation 4.20.

The gain $g_{\mathrm{s}}(\lambda)$ controlling the power of the synthetic speech segments is computed comparing the excitation signal $e_{\lambda,\mathrm{pre}}(\kappa)$ of the noisy, pre-filtered signal and the excitation $e_\lambda(\kappa)$ produced by the pulse train as described in Equation 4.62. Ideally, the power of the reconstructed excitation signal should be equal to the power of the excitation signal of the clean speech signal. The sum of the squared residual signal $e_{\lambda,\mathrm{pre}}(\kappa)$ is directly accessible from Levinson-Durbin recursion as the prediction error for the computed LPC coefficients. Then the gain can be calculated as

$$g_{\mathrm{s}}(\lambda) = \sqrt{\frac{\sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} e^2_{\lambda,\mathrm{pre}}(\kappa)}{\sum\limits_{\kappa=0}^{L_{\mathrm{F}}-1} e^2_{\lambda}(\kappa)}}. \tag{4.68}$$

Speech Composition

The combination of the two signal branches depicted in Figure 4.27 is realized by two contrary filters (low-pass and high-pass) with the cut-off frequency $f_{\mathrm{c}}(\lambda)$. Through the upper branch only the noise-free parts $\tilde{x}_\lambda(k)$ of the signal pass by applying a high-pass filter. The remaining components of the system reconstruct the missing speech signal parts. The cut-off frequency $f_{\mathrm{c}}(\lambda)$ defines the amount of reconstructed speech in the output signal and is controlled by the wind detection. The power ratio between between a low-frequency range and the frequency range up to $f_\mathrm{s}/2$ is used as

$$f_\mathrm{c}(\lambda) = f_\mathrm{max} \cdot \frac{\sum\limits_{\mu=0}^{\mu_\mathrm{hi}} |X(\lambda,\, \mu)|^2}{\sum\limits_{\mu=0}^{M/2-1} |X(\lambda,\, \mu)|^2} \tag{4.69}$$

to determine the cut-off frequency. The parameter for the limit of the low-frequency range μ_hi is chosen to 100 Hz as in this range no speech activity is expected and only wind noise will cover this part of the spectrum. The parameter f_max controls the maximum range of the reconstructed speech in the output signal. In [NNJ+12], $f_\mathrm{max} = 1500\,\mathrm{Hz}$ was found to give a good trade-off between wind noise suppression and artifacts of the synthetic speech in the output of the system. A higher value will result in a more aggressive wind noise reduction but will also introduce a wider range of artificial speech, which leads to a "robotic sound" of the processed signal. The final output signal of the proposed system is then given by the sum of the low-pass filtered synthetic speech $\tilde{s}_{\lambda,\mathrm{LP}}(\kappa)$ for the reconstruction of the noisy parts and the noise-free speech parts $\tilde{x}_{\lambda,\mathrm{HP}}(\kappa)$ gained from the high-pass filter

$$\hat{s}_\lambda(\kappa) = \tilde{x}_{\lambda,\mathrm{HP}}(\kappa) + \tilde{s}_{\lambda,\mathrm{LP}}(\kappa). \tag{4.70}$$

4.5.2 Corpus-based Wind Noise Reduction

In the system proposed in the previous section many components and parameters are chosen heuristically by extensive subjective investigations yielding an enhanced output signal as it will be shown in Section 4.5.4. However, an advancement is proposed in [NNV15] using also the new concept of reconstructing the missing or highly noisy parts of the speech signal by a synthetic speech signal. The main difference is to incorporate pre-trained information gained from a clean speech corpus into the wind noise reduction task. This system will be denoted as corpus-based wind noise reduction corpus-based wind noise reduction (CORP).

The system is presented in Figure 4.30, again omitting the analysis and synthesis parts of the framework and also the FFT/IFFT stages. The main parts are the signal combination, realized here as a binary spectral gain function $G_\mathrm{bin}(\lambda,\, \mu)$ and the speech synthesis stage. The latter exploits not only information from the current input signal as the pitch frequency $f_0(\lambda)$, but also pre-trained information gained from a clean speech corpus, which is applied during the speech synthesis process. Therefore, the term *corpus based speech synthesis* is used. As a post-processing step, a residual noise reduction is applied. For the calculations of the binary gain $G_\mathrm{bin}(\lambda,\, \mu)$, the speech synthesis, and the residual noise reduction gain $G_\mathrm{W}(\lambda,\, \mu)$ a wind noise STPS estimate $|\widehat{\mathcal{N}}(\lambda,\, \mu)|^2$ is required. Here, the P-IBM method is applied, which was presented in Section 4.2.2.2 and turned out to give the best results (see Figures 4.11, 4.13 and 4.14).

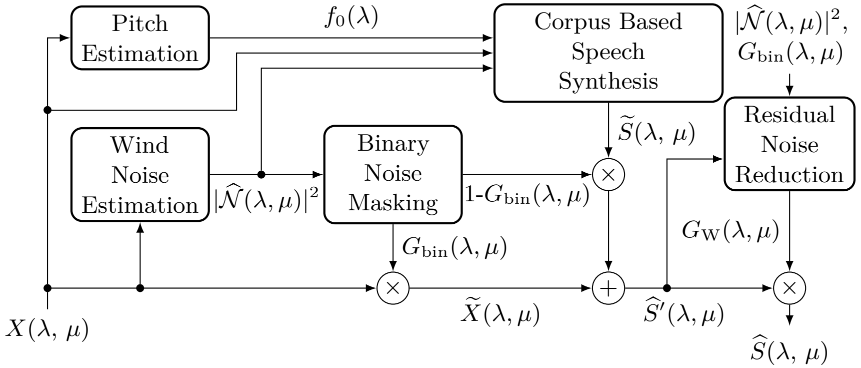

Pitch Estimation
$f_0(\lambda)$
Corpus Based Speech Synthesis
$|\widehat{\mathcal{N}}(\lambda,\mu)|^2$, $G_{\mathrm{bin}}(\lambda,\mu)$
Residual Noise Reduction
Wind Noise Estimation
$|\widehat{\mathcal{N}}(\lambda,\mu)|^2$
Binary Noise Masking
1-$G_{\mathrm{bin}}(\lambda,\mu)$
$\widetilde{S}(\lambda,\,\mu)$
$G_{\mathrm{bin}}(\lambda,\mu)$
$G_{\mathrm{W}}(\lambda,\mu)$
$X(\lambda,\,\mu)$
$\widetilde{X}(\lambda,\mu)$
$\widehat{S}'(\lambda,\mu)$
$\widehat{S}(\lambda,\,\mu)$

Figure 4.30: System for corpus based speech enhancement CORP.

Signal Composition by Binary Mask

The aim of this stage of the algorithm is to compose the signal $\widehat{S}'(\lambda,\,\mu)$ of parts of the masked input signal $\widetilde{X}(\lambda,\,\mu)$ and parts of the synthetic speech signal $\widetilde{S}(\lambda,\,\mu)$, which is denoted by the signal composition block in Figure 4.27. The frequency dependent composition is realized by the binary mask $G_{\mathrm{bin}}(\lambda,\,\mu)$ applied to the noisy input $X(\lambda,\,\mu)$ and inverted mask $(1 - G_{\mathrm{bin}}(\lambda,\,\mu))$ to the synthetic speech. The aim is to cancel out highly impaired parts in the input signal and replace them with $\widetilde{S}(\lambda,\,\mu)$. As explained in Section 4.2.2.2, a binary mask is commonly determined by comparing a local criterion $\mathrm{LC}(\lambda,\,\mu)$ for each time-frequency bin, e.g., the SNR, to a frequency dependent threshold $th(\mu)$

$$G_{\mathrm{bin}}(\lambda,\,\mu) = \begin{cases} 0, & \text{if, } \mathrm{LC}(X(\lambda,\,\mu), |\widehat{\mathcal{N}}(\lambda,\,\mu)|^2) < th(\mu) \\ 1, & \text{otherwise.} \end{cases} \tag{4.71}$$

In the proposed system the speech presence probability (SPP) is used as local criterion as clean speech indicator. It is defined according to [CB01] as

$$\begin{aligned} \mathrm{LC}(\lambda,\mu) &= \mathrm{LC}(X(\lambda,\,\mu), |\widehat{\mathcal{N}}(\lambda,\,\mu)|^2) \\ &= \left(1+(1+\xi_{\mathrm{opt}})\exp\left(-\frac{|X(\lambda,\mu)|^2}{|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2}\cdot\frac{\xi_{\mathrm{opt}}}{\xi_{\mathrm{opt}}+1}\right)\right)^{-1}, \end{aligned} \tag{4.72}$$

where the constant parameter ξ_{opt} is the optimal *a priori* SNR ($\widehat{=}$ 15 dB as proposed in [GH11]). The SPP has values between 0 and 1 for each frequency bin and is compared to the frequency dependent threshold as indicated by (4.71): $th(\mu) = 0.95$ for $0 \le f \le 500\,\mathrm{Hz}$ and $th(\mu) = 0.75$ for $f > 500\,\mathrm{Hz}$. Thus, the lower frequencies are more likely set to zero, where most of the wind energy is assumed. The noise STPS $|\widehat{\mathcal{N}}(\lambda,\,\mu)|^2$ is estimated by the pitch adaptive method

[NV15] presented in Section 4.2.2.2, which showed the best performance for the single microphone wind noise reduction schemes. In this chapter a setup using only one microphone is considered. In the case of dual microphone configurations, the coherence based method [NV14a] derived in Section 4.4.3 can be applied for the wind noise estimation. The binary gain is multiplied with the noisy input signal $X(\lambda,\mu)$ yielding the masked signal $\widetilde{X}(\lambda,\mu)$.

Speech Synthesis

The corpus based speech synthesis is depicted more detailed in Figure 4.31. The input values are the noisy input $X(\lambda,\ \mu)$ the wind noise STPS estimate $|\widehat{\mathcal{N}}(\lambda,\ \mu)|^2$ and the fundamental frequency $f_0(\lambda)$. The goal is to produce a voiced speech signal applying the source-filter model already shown in Figure 4.29b. The corresponding components, i.e., excitation generation, vocal tract filter, and the gain $g_\mathrm{s}(\lambda)$, can be found in the bottom branch of Figure 4.31. The synthetic speech signal is again given by filtering the excitation $e_\lambda(\kappa)$ with the vocal filter $\tilde{\mathbf{a}}(\lambda) = [\tilde{a}_\lambda(1), \ldots, \tilde{a}_\lambda(l_\mathrm{LP})]$ (Equation 4.66). In contrast to the previously described system, for the generation of the excitation signal $e_\lambda(\kappa)$, a pitch cycle extracted from clean speech is taken as template pitch cycle (TPC), which is shown in Figure 4.32. The length of this pitch cycle is inversely proportional to its fundamental frequency $f_{0,\mathrm{TPC}}$. To adjust the excitation signal the TPC is time-warped by the ratio

$$R(\lambda) = \frac{f_{0,\mathrm{TPC}}}{f_0(\lambda)} \tag{4.73}$$

by re-sampling[7] of the TPC with $R(\lambda)$. Different speakers were tested (male and female) for the TPC with only marginal differences, therefore only one TPC is

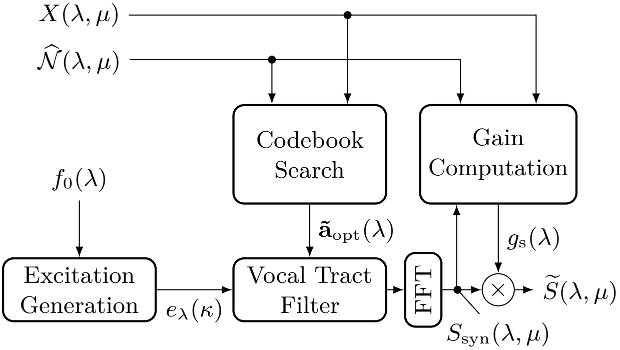


Figure 4.31: Corpus based speech synthesis component from Figure 4.30.

[7] For the re-sampling process the resample function of *Matlab* was used.

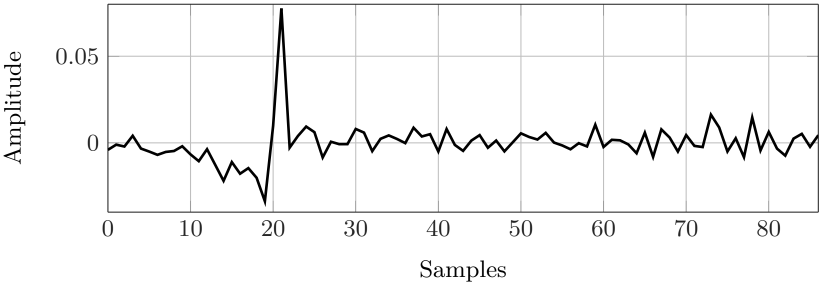


Figure 4.32: Template pitch cycle (TPC) used for the excitation signal generation.

applied, which is taken from a male speaker from the training data set of the TIMIT database [LKS89].

The generation of the excitation signal is depicted in Figure 4.33 using repeated re-sampled TPCs. The process is exemplified by three frames, where two issues have to be covered during the generation.

1. Continuous transition between consecutive frames:
   To avoid discontinuities between consecutive frames, only the second half of the L_F samples of each frame are updated, as shown by the gray highlighted segments in Figure 4.33. By this procedure the overlapping parts of the frames are identical for the assumed overlap of half frame-size in the used framework.

2. Pitch synchronicity:
   If each updated part starts with the beginning of the (time-warped) TPC, the generated excitation signal will not result into a pitch synchronous signal because the last pitch cycle is not necessarily attached in its full length, i.e. until its last sample. E.g., in Figure 4.33, the first excitation update in frame $\lambda - 1$ ends a few samples after the third TPC starts. Therefore, the missing fraction of the last used TPC of the previous frame is used as as starting point for the current frame λ. This is realized by a circular shift of the TPC by κ samples.

The amount of samples for the required shift of the TPC in the current frame is defined by

$$\delta(\lambda) = \frac{L_\mathrm{F}/2}{N_0(\lambda-1)} - \lfloor \frac{L_\mathrm{F}/2}{N_0(\lambda-1)} \rfloor \tag{4.74}$$

and the operation of the circular shift of $x(k)$ by δ samples is described by

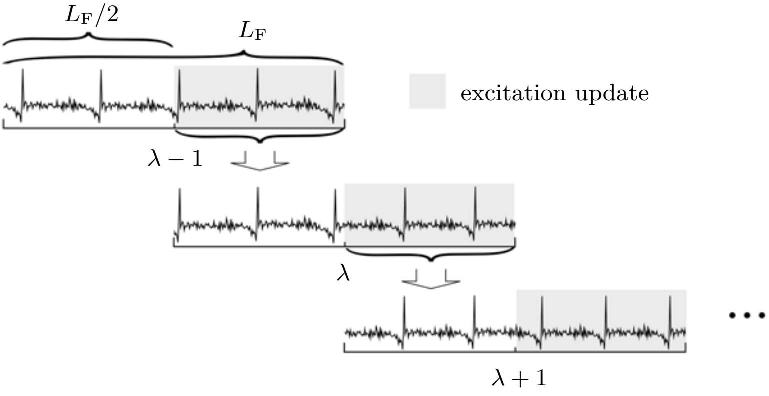


Figure 4.33: Excitation signal generation in three consecutive frames.

$\mathrm{CS}(x(k),\delta)$. Then the excitation signal in a frame λ is then determined as

$$\tilde{e}_\lambda(\kappa) = \begin{cases} \tilde{e}_{\lambda-1}(k+L_\mathrm{F}/2), \text{ if } \kappa < L_\mathrm{F}/2 \\ \mathrm{CS}(\,\overline{\mathrm{TPC}}_{f_0(\lambda)}(\kappa),\delta), \text{ else.} \end{cases} \tag{4.75}$$

The periodically repeated TPC, which is re-sampled to the current fundamental frequency $f_0(\lambda)$ is denoted by $\overline{\mathrm{TPC}}_{f_0(\lambda)}(\kappa)$.

The vocal tract filter in Figure 4.31 is obtained by means of a codebook in which representations $\tilde{\mathbf{a}}_i$ of the filter coefficients gained from clean speech are stored. The vector

$$\tilde{\mathbf{p}}_i = [\tilde{p}_{1,i} \ldots \tilde{p}_{K_\mathrm{CB},i}]^T \tag{4.76}$$

is the i-th entry and contains K_CB features describing the spectral envelope. Additionally, the associated LPC coefficients $\tilde{\mathbf{a}}_i$ are stored in the codebook with the aim to find the optimal coefficient vector $\tilde{\mathbf{a}}_\mathrm{opt}(\lambda)$ in each frame by comparing the currently observed feature vector $\mathbf{p}(\lambda)$ with the stored vectors $\tilde{\mathbf{p}}_i$. This concept was already deployed for the purpose of background noise estimation methods proposed in [Ros10] or [HNNV14]. The features describing the spectral envelope of the noise and speech signals are given by cepstral coefficients [Ros10] or the DFT representation [HNNV14]. In contrast to these methods, the codebook is used for the speech synthesis in the proposed system.

The codebook is derived from the training data set of the TIMIT database [LKS89] using only voiced speech segments, because speech pauses and unvoiced segments are not generated by the speech synthesis and should therefore not be represented by the codebook. For the codebook generation, the voiced speech is segmented in the same way as for the noise reduction process, i.e., into frames of

20 ms with an overlap of half frame-size. To reduce the number of entries in the codebook the k-means algorithm is employed as vector quantizer[8] [MRG85].

Different parameters can be taken to describe the vocal tract filter in each frame. Taking the aforementioned cepstral coefficients or directly the spectral amplitudes in each frequency bin is possible. For speech coding applications also the LPC coefficients or the line spectral frequency (LSF) are used [Ita75]. The latter are known to be robust to quantization effects. This is an important issue as the size of the codebooks, i.e., the number of entries, is limited in order to comply certain complexity aspects. Four different features are considered for the estimation of the vocal tract filter parameters:

1. Linear predictive coding (LPC) coefficients: Coefficients of the auto-regressive (AR) filter representing the vocal tract by means of an infinite impulse response (IIR) filter.

2. Mel-frequency cepstral coefficients (MFCC): Cepstral coefficients using a non-uniform frequency resolution, which is adopted to the frequency resolution of human auditory system [RJ93]. This representation is widely used for speech and music recognition tasks (see, e.g., [DM80]).

3. Line spectral frequencies (LSF): The representation proposed by Itakura [Ita75] contains exactly the same information as the LPC coefficients by computing the roots of the palindromic and antipalindromic polynoms of the LPC polynom. Broadly speaking, they represent the positions of the poles and zeros of the spectral envelope.

4. Spectral envelope (SPENV): As a description for the vocal tract filter also the complete spectral envelope can be taken into account, which is given by the DFT representation of the LPC filter. In contrast to the other three features, the spectral envelope is not a compact representation since all frequency bins must be stored into the codebook.

During the estimation process the trained codebook entries are compared to the features calculated from the input signal. To reduce the effect of the wind noise on the codebook search, a spectral subtraction is applied using the wind noise STPS estimate. The considered parameter from the input signal is computed in the current frame resulting in a de-noised parameter vector $\mathbf{p}(\lambda)$. The optimal codebook entry is given by minimizing the mean square error (MSE) between the feature vector $\mathbf{p}(\lambda)$ of the de-noised input signal and each codebook entry $\tilde{\mathbf{p}}_i$

$$i_{\text{opt}}(\lambda) = \arg\min_i \{||\tilde{\mathbf{p}}_i - \mathbf{p}(\lambda)||^2\}. \tag{4.77}$$

Figure 4.34 compares the performance of the four features for the description of the vocal tract filter. The performance is measured by comparing the squared magnitude $|\widetilde{A}(\lambda, \mu)|^2$ of the spectral envelope of the estimated vocal tract by

[8] The implementation in *voicebox* by Brookes was used for the vector quantization [B$^+$11].

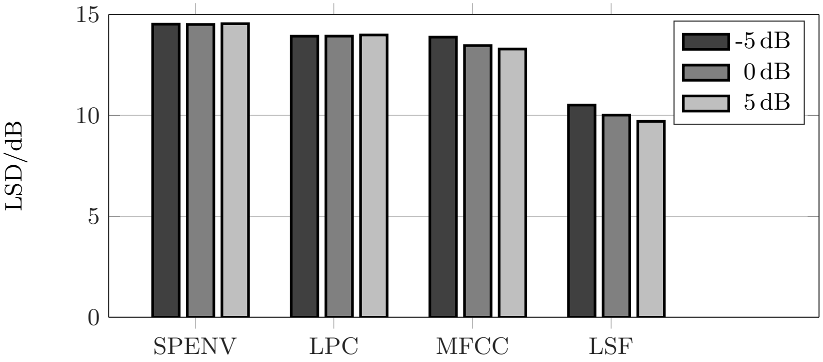


Figure 4.34: Effective logarithmic spectral distortion of estimated envelopes for three different input SNRs and codebooks with 512 entries.

a considered feature with the squared magnitude of the true spectral envelope $|A(\lambda,\, \mu)|^2$ of the clean speech signal. The error is computed by the logarithmic spectral distortion (LSD) between the two power spectra as

$$\mathrm{LSD}_{\mathrm{dB}} = \frac{1}{\mathcal{K}} \sum_{\lambda=0}^{\mathcal{K}} \sqrt{\sum_{\mu \in \tilde{\mu}} \left(10 \log_{10} \frac{|A(\lambda,\, \mu)|^2}{|\widetilde{A}(\lambda,\, \mu)|^2} \right)^2} \qquad (4.78)$$

where only the frequency bins $\tilde{\mu}$ are taken into account, which needs to be replaced, i.e., where the binary mask of Equation 4.71 is zero.

The results in Figure 4.34 are gained for three SNR scenarios and for the four features stored in a codebook of 512 entries gained from three minutes of voiced speech randomly taken from the training set of the TIMIT database. For all SNR conditions, the LSF representation offers the lowest distortion, which shows that they are the most robust towards the degradation of the input signal but also to the applied vector quantizer during the codebook generation. These results support the knowledge from speech coding that LSFs are a good choice for a quantized representation of the vocal tract filter coefficients and they will be used in the following.

A second experiment is carried out to investigate the influence of the training data. During the codebook generation, two parameters can be adjusted, the size, i.e., the number of codebook entries and the duration of the training sequence. The impact of both parameters is shown in a two-dimensional representation in Figure 4.35 using LSFs as feature vector again in terms of the LSD. The duration of the training sequence is given on the x-axis while the y-axis depicts the number of codebook entries and the gray scale reflects the LSD value.

Besides for very small codebook sizes of 16 or 32 entries and short durations the computed LSD values are not varying to a great amount. Furthermore, the

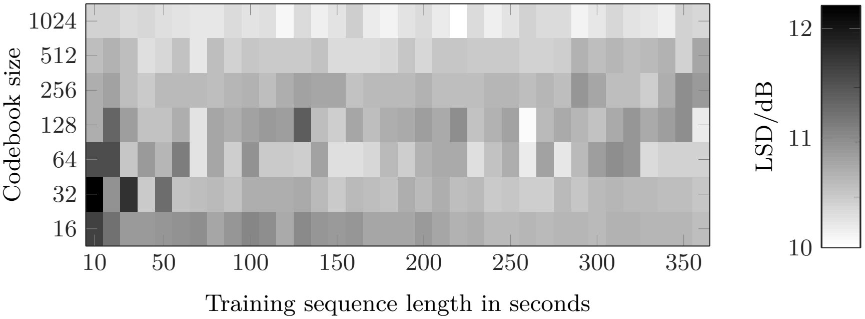


Figure 4.35: LSD for different variants of the codebook generation using LSF vocal tract filter representation.

length of the training data seems not to influence the result if the codebook size is sufficient high (e.g., 512 entries or more). In the following, a codebook of 512 entries gained from 3 minutes of training data is used, as larger codebooks or more training data does not indicate significant improvements.

The last missing parameter for the speech synthesis part in Figure 4.31 is the gain $g_s(\lambda)$, which is multiplied with the DFT representation $S_{\text{syn}}(\lambda, \mu)$ of the generated speech signal. Because the spectral distribution of the synthetic speech signal is already defined by the excitation signal and the vocal tract filter, only a global gain is required controlling the power of each frame. In the ideal case, $\widetilde{S}(\lambda, \mu)$ has the same power as the unknown clean speech signal frame $S(\lambda, \mu)$. To adjust the power the gain computation is realized as follows

$$g_s(\lambda) = \sqrt{\frac{\sum\limits_{\mu} \left[|X(\lambda, \mu)|^2 - |\widehat{\mathcal{N}}(\lambda, \mu)|^2 \right]}{\sum\limits_{\mu} |S_{\text{syn}}(\lambda, \mu)|^2}}, \tag{4.79}$$

which can be seen as a spectral subtraction of the noise estimate $|\widehat{\mathcal{N}}(\lambda, \mu)|^2$ with respect to a whole signal frame. After the multiplication with the gain factor the artificial speech signal $\widetilde{S}(\lambda, \mu)$ can be used for the signal composition as explained before.

Residual Noise Reduction

So far, the proposed system in Figure 4.30 only applies a binary processing either to reconstruct the signal ($G_{\text{bin}}(\lambda, \mu) = 0$) or to keep the noisy input signal ($G_{\text{bin}}(\lambda, \mu) = 1$). A high amount of wind noise suppression can be achieved by tuning the threshold for the binary gain computation in Equation 4.72 to a more aggressive setting, i.e., to set gains to zero for lower SPP values. However, this introduces a higher fraction of artificial speech in the output signal on the expense of an unnatural sound. A better solution is given by applying the binary decision as described before in order to reconstruct only the highly noisy parts of the signal. The remaining noise is then removed by a conventional noise reduction as proposed in Section 4.3. This means that the noise estimate $|\widehat{\mathcal{N}}(\lambda,\, \mu)|^2$ is used along with the modified spectral subtraction of Section 4.3.2, which is applied to the unmasked, i.e, non-reconstructed frequency bins.

4.5.3 On the Phase Reconstruction

All conventional noise reduction methods, which apply a spectral gain only enhance the magnitude of the noisy input spectra

$$\widehat{S}(\lambda,\, \mu) = G(\lambda,\, \mu) \cdot X(\lambda,\, \mu) = |\widehat{S}(\lambda,\, \mu)| \cdot e^{\mathrm{j}\eta(\lambda,\, \mu)} \tag{4.80}$$

keeping the noisy phase $\eta(\lambda,\, \mu)$ of the complex spectrum $X(\lambda,\, \mu)$. In this section a discussion is carried out about the phase of the synthesized speech spectrum applied in the aforementioned concepts for wind noise reduction. Several publications can be found on the topic of phase processing in the terms of speech enhancement. The experiments reported by Wang and Lim [WL82] showed that the phase only has an influence on the processed speech at very low SNRs (-25 dB) for long frame-sizes of 400 ms. In other cases the incorporation of the clean phase does not result in any improvement. Ephraim and Malah derived that the MMSE estimate of the complex spectrum of the clean speech leads to the known Wiener solution keeping the noisy phase [EM84]. The calculations made by Vary in [Var85] predicts that phase deviations are only perceived for SNRs lower 6 dB.

In the last years several approaches were presented, which address speech enhancement processing also incorporating phase modifications of the noisy signals (see, e.g., [KG12], [GKR12], [MS14]). In total, the improvement is limited and only a combined processing of phase and magnitude of the spectral coefficients indicates an improvement ([MS14]). All methods require an estimate of the fundamental frequency to apply pitch synchronous adaptation of the analysis-synthesis framework.

The proposed generation of synthetic speech explained in Equation 4.75 and Figure 4.33 can be seen as a synchronization of the generated excitation signal to the fixed analysis-synthesis framework. Thus, the phase of the generated speech signal is of great importance for the pitch synchronicity. Keeping the noisy phase introduces discontinuities in the overlapping parts of the frames, which results in severe artifacts and the periodicity of the initially voiced segments is destroyed. From the listening impression, segments generated with a noisy phase sounds similar as unvoiced speech, which is of course not desired. After these considerations, the synthetic speech signal is applied for both magnitude and phase reconstruction in the proposed concepts presented in Sections 4.5.1 and 4.5.2.

4.5.4 Performance Results

The two proposed systems including the speech synthesis into the noise reduction process:

- partial speech synthesis (PSYN) (Section 4.5.1)
- corpus-based wind noise reduction (CORP) (Section 4.5.2)

are evaluated with wind recordings and compared to three methods using only a conventional spectral weighting:

- the SPP based algorithm from [GH11], which can be seen as the state-of-the-art approach for background noise estimation,
- the morphological technique (MORPH) [HWB+12] (see Section 4.2.2.2)„
- the masked based approach (P-IBM) [NV15] (see Section 4.2.1.1).

Both algorithms for wind noise estimation MORPH and P-IBM gave sufficiently accurate wind noise estimates. These methods for noise estimation are used in combination with modified spectral subtraction (see Section 4.3.2).

Because of the non-linear processing, which is introduced by the speech synthesis in the two alternative approaches, the quality measures used before (NA-SA and SII) can not be calculated, since they require the filtered clean speech signal and filtered pure noise signal. These two signals are not given in the new concepts for wind noise reduction where parts of the input signal are replaced by a synthetic speech signal. Thus, two other measures are used for the evaluation of the algorithms:

1. Perceptual evaluation of speech quality (PESQ): A measure standardized by the International Telecommunication Union (ITU) [IT01] to predict the perceptual rating of human listeners on a mean opinion score (MOS) between 0.5 (bad) and 4.5 (no distortions) as proposed by [RBHH01]. Here, the wideband extension ([IT07]) is applied for the considered audio signals with a sampling frequency of 16 kHz.
2. Segmental SNR (segSNR): A widely used measure, which computes a segmental, i.e., frame-wise ratio between the clean speech signal and the error between the clean and processed speech [QB88]. The averaged values of frames, where both speech and wind noise are active results in the considered measure.[9] A higher value indicates an improvement.

The experiment is carried out with 270 s speech data randomly chosen from the test set of the TIMIT database. Wind noise segments from real recordings [NV14b] were added with lengths between 0.3 and 3 s. The level of the wind noise is adjusted to a realistic scenario resulting in mostly negative SNR values in frames, where both speech and wind are active. For the shown PESQ results the percentage

[9] More details on the computation of segmental SNR (segSNR) are given in Appendix A.1.

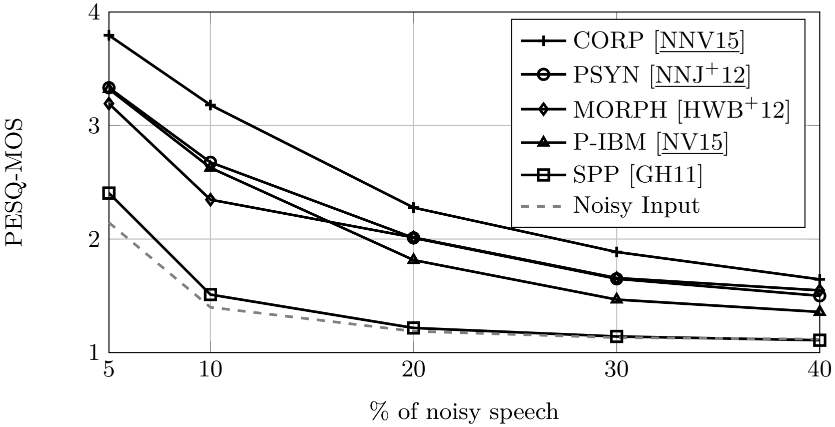


Figure 4.36: PESQ-MOS results for different degrees of degradation.

of the length of voice activity, which is corrupted by wind noise is given (shown on the x-axis of Figure 4.36). Because the PESQ measure shows saturation effects for low SNR values (<-5 dB) and high SNR values (>5 dB), the amount of noise can be adjusted with a finer resolution by the percentage of noisy speech, i.e., the temporal overlap of speech segments and noise segments.

The results in terms of the PESQ values in Figure 4.36 show that all considered algorithms yield an enhancement of the perceptual evaluation of speech quality (PESQ) value of the noisy speech, as depicted by the dashed gray reference line. As expected, the SPP method, which is designed for background noise tracking, is not capable to follow the non-stationary characteristics of wind noise. Thus only marginal improvements can be seen. The PSYN concept and the two conventional approaches based on noise estimation and spectral weighting (MORPH and P-IBM) show similar results for all degrees of degradation. The best performance for all scenarios is achieved by the corpus based method (CORP) with PESQ improvements up to 2 MOS values. Investigations using CORP method without the spectral weighting applied as post-filter (see Figure 4.30) show only marginal lower results.

The second measure, the segmental SNR, is depicted in Figure 4.37 averaged for all noise scenarios. Again, all methods show an improvement compared to the SNR value of the noisy input, which is represented by the dashed gray line. The corpus based speech synthesis method shows the best performance with a gain over 16 dB compared to the noisy input signal. Besides, the insufficient noise reduction performance of the SPP for conventional background noise estimation is demonstrated by only a low improvement of about 4 dB segSNR.

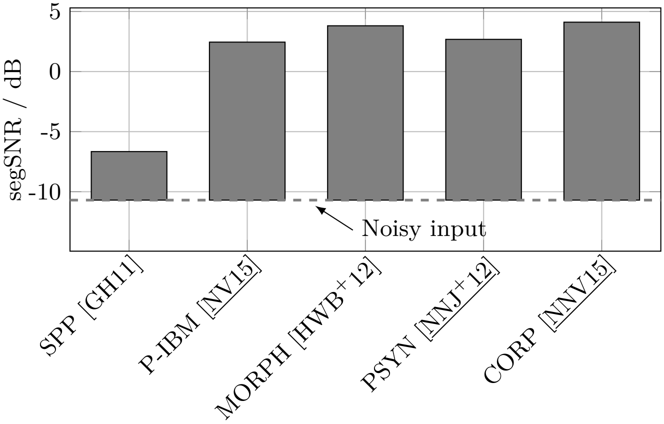


Figure 4.37: segSNR Results, - - - represents measures of noisy input signals.

4.6 Conclusions

In this chapter different concepts for the enhancement of speech degraded by wind noise are presented. Systems using a single microphone or dual microphone configurations are investigated. As the special characteristic of wind noise makes it necessary to develop algorithms especially designed for the statistical properties of wind noise, new concepts for both configurations are developed.

First, a single microphone noise reduction system based on spectral weighting is considered. For the required wind noise STPS estimate, two new noise estimation schemes are proposed exploiting the spectral energy distribution of wind and speech. Since the first step of the estimation is an wind detection, the NSTM method from Section 3.5.1.2 is used in the schemes, which showed the highest accuracy. The sub-band signal centroid played an important role for the classification of noisy signal, i.e., if speech, wind, or both signals are active. A subsequent exploration of the spectral shapes of speech and wind noise leads to two novel algorithms to estimate the STPS of wind noise minima fitting approach (Min-Fit) and the pitch Adaptive binary mask (P-IBM). Where the Min-Fit algorithm features a low complexity, P-IBM leads to a more accurate noise estimate, indicated by a low logarithmic error of the STPS estimate (3 to 8 dB lower than all considered methods for all relevant scenarios). Combined with the recursive spectral subtraction gain computation, a high wind noise reduction is achieved, where the pitch adaptive approach P-IBM also clearly outperforms previously presented wind noise reduction systems in terms of the NA-SA measure and the SII.

For applications using two microphones, the coherence properties of speech and wind noise can be taken into account for the noise reduction. A wind noise STPS estimator is proposed in Section 4.4.3, which solves the problem of fast changes of the noise level by a decision directed scheme for the noise estimate and an adaptive update scheme for the coherence computation (DDWE + ASWE). The key point is to incorporate phase information of the complex coherence function. The new method (DDWE + ASWE) shows better performance than state-of-the art methods for dual microphone wind noise reduction for different conditions. A further advantage of the proposed method is that the noise estimation is carried out separately. This can be useful, if the signal is processed by additional enhancement steps.

All methods for speech enhancement based on a spectral weighting, have the drawback that they introduce undesired attenuation of the speech signal in parts with a very low local SNR. Because of the high signal levels of wind noise at low frequencies, this leads to an high-pass effect on the output signal. This problem is circumvented by an innovative approach for speech enhancement, which reconstructs parts of the speech. Two concepts using the source-filter model of speech production are presented, where the use of information stored in pre-trained codebooks is the key to ensure a high speech quality. These methods have a higher computational complexity compared to the approaches applying only a spectral weighting, but the evaluation under realistic conditions showed a great performance gain in terms of the PESQ measure and the segmental SNR.

In summary, for a single microphone system and a noise reduction by spectral weighting the combination of the P-IBM wind noise estimation and the recursive spectral subtraction method should be chosen. If a low-complexity solution is is required, the minima fitting approach can also be taken into account for the noise estimation. Using two microphones, the new proposed coherence based wind noise estimation exploiting the phase information shows the best results. In cases where the complexity is not a crucial point, the concept applying a partial speech synthesis can further improve the speech enhancement performance.

Chapter 5

Application to Mobile Phones

In this chapter the application of speech enhancement algorithms in mobile phones is considered. Often, the applied methods make assumptions about the acoustic environment, e.g., in terms of a certain signal model and the resulting properties and statistics. In real environments, derivations from these assumptions might lead to a limited performance and call for modifications of the proposed systems. Here, two applications are considered within this chapter, which deals with typical problems that arise from practice.

While most smart-phones are equipped with at least two microphones there exist a great number of so-called feature phones with limited functionality and only a single microphone. For single-microphone systems, the simultaneous occurrence of wind noise and background noise is a common scenario. As a speech enhancement system must be robust to this condition, the first application in this chapter is the combined reduction of background noise and wind noise. Different approaches will be discussed to ensure a high suppression of all noise signals.

The second application deals with conventional background noise reduction for mobile phones using two microphones. Here, two use cases are considered with different acoustic characteristics, the normal hand-held position (HHP) and the hands-free position (HFP). Because solutions for the HHP were already presented in detail in the work of Jeub in [Jeu12], this work will focus on noise reduction for the HFP case. In these conditions, usually the coherence models of speech and noise can be exploited. However, these models and the coherence properties of real signals lead to limitations of the noise reduction system. Thus, in the second part of this chapter solutions are presented to circumvent this limitations. Several solutions for wind noise reduction using two microphones were already introduced in Section 4.4. A further combination of the proposed advanced background noise reduction with dual microphone wind noise could be possible but is not considered here.

5.1 Combined Wind and Background Noise Reduction

As mobile phones can be used in many situations, usually, there is not only a single disturbance but a mixture of different noise signals impairing the speech quality. In addition to wind noise further noise sources might occur, e.g., traffic noise from a street near by, or inside-car noise if the phone call is taking place inside a convertible car. This section discusses different possible options to combine a general background noise reduction with a wind noise reduction. Here, the single microphone setting of the overlap-add structure depicted in Figure 2.4 is considered with a noise estimation stage and subsequent spectral weighting to enhance the desired speech signal.

5.1.1 Concept for Combined Noise Reduction

Different configurations are conceivable, which incorporate both background noise and wind noise reduction. E.g., both estimates of background noise and wind noise could be carried out independently and this leads to a parallel processing for both noise types. This is however not favorable for the following reason. All considered wind noise estimators rely on the assumption, that the input signal only contains clean speech and wind noise. They exploit spectral properties of the clean speech and pure wind noise in order to achieve a processing, which is not based on the temporal characteristics of speech and noise. The presence of further signal portions such as additional background noise will influence the wind detection and estimation. Therefore, a serial processing is applied, where first the conventional background noise is reduced and then the wind noise reduction is carried out. The underlying model of the noisy input signal is given by

$$x(k) = s(k) + n_{\mathrm{b}}(k) + n_{\mathrm{w}}(k), \tag{5.1}$$

or in the short-term discrete Fourier transform (DFT) domain

$$X(\lambda,\, \mu) = S(\lambda,\, \mu) + N_{\mathrm{b}}(\lambda,\, \mu) + N_{\mathrm{w}}(\lambda,\, \mu), \tag{5.2}$$

where the subscripts identifies background noise (b) and wind noise (w) portions in the noisy signal.

The used structure for the combined noise reduction is presented in Figure 5.1, where two setups can be chosen by position of switch A.

- Switch A in position ①: wind noise detection and wind noise estimation based only on the modified spectrum $\widehat{S}'(\lambda,\, \mu)$.
- Switch A in position ②: wind noise detection based on input spectrum $X(\lambda,\, \mu)$ and wind noise estimation based on the modified spectrum $\widehat{S}'(\lambda,\, \mu)$.

The first stage applies a conventional background noise reduction using the speech presence probability (SPP) method of Gerkmann and Hendriks [GH11] for the noise PSD estimation $\widehat{\Phi}_{nn,\mathrm{b}}(\lambda,\, \mu)$ (see Section 2.3.1), which is known to give reasonable results for many background noise types. Applying a spectral gain $G_1(\lambda,\, \mu)$, this results in the first enhanced signal $\widehat{S}'(\lambda,\, \mu)$. The second stage for the wind noise reduction is realized by the pitch adaptive inverse binary mask (P-IBM) method proposed in Section 4.2.2.2 for the estimation of the wind short-term power spectrum (STPS) $|\widehat{N}_{\mathrm{w}}(\lambda,\, \mu)|^2$ and the calculation of a gain $G_2(\lambda,\, \mu)$.

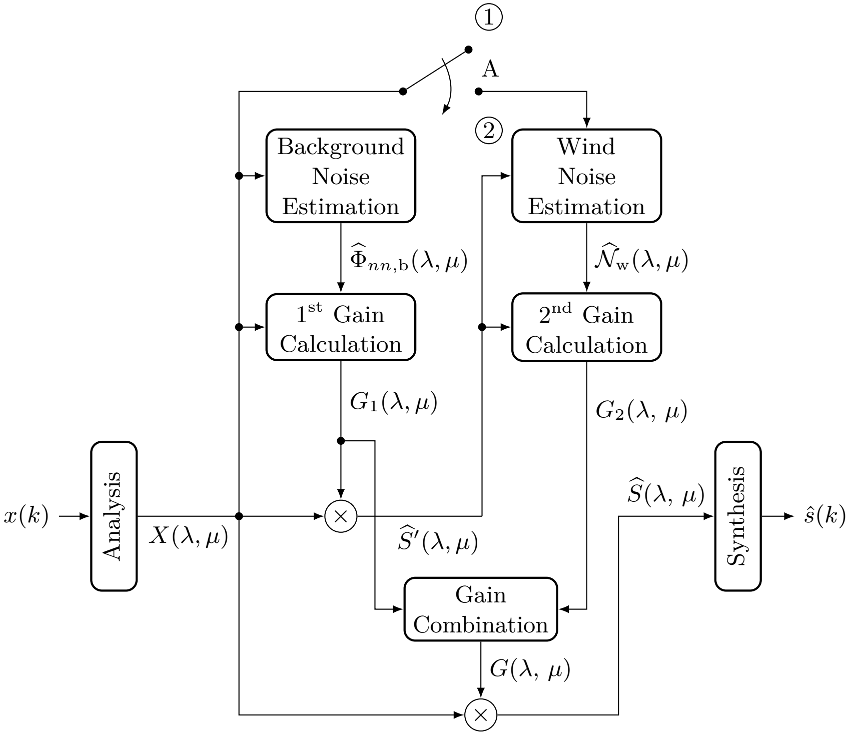


Figure 5.1: System for combined background noise and wind noise reduction.

The normalized short-term mean (NSTM) used for the wind detection (see Section 3.5.1.2), is not influenced by any zero-mean signal (e.g., additional background noise). But, the processing by the background noise reduction in the first stage can remove or reduce the short-term offset caused by the wind noise. In this case the important feature for the wind detection is removed. This leads to undetected parts of wind noise activity in the observed signal and thus remaining wind noise components. Therefore, the unfiltered input is used for the wind noise detection, if switch A is in position②. For the remaining processing steps of the pitch adaptive inverse binary mask (P-IBM) algorithm for wind noise estimation, the pre-filtered signal $\widehat{S}'(\lambda,\,\mu)$ is applied as explained in Section 4.2.2.2.

The partial speech synthesis concept presented in Section 4.5 is based on the assumption that the occurring noise is sparse with respect of its energy distribution in the time-frequency domain. This is fulfilled for wind noise but usually not for background noise in general, which can cover a larger range in both time and frequency. Thus this concept of speech enhancement is not applied for the combined noise reduction in this section. Besides, for the application in mobile phones, computational complexity is always a constraint for signal processing algorithms. This also pleads for the a noise reduction via spectral weighting, which is characterized by a lower complexity.

Using the background noise PSD estimate $\widehat{\Phi}_{nn,\mathrm{b}}(\lambda,\,\mu)$ and the wind noise STPS estimate $\widehat{\mathcal{N}}_{\mathrm{w}}(\lambda,\,\mu)$ two spectral gains $G_1(\lambda,\,\mu)$ and $G_2(\lambda,\,\mu)$ are computed. For the background noise reduction the Wiener rule using the decision directed approach (DDA) for SNR estimation [EM84] is applied, while the gain of the second stage is calculated by the recursive spectral subtraction rule (see Section 4.3.2). As depicted in Figure 5.1, both gains are combined to the gain $G(\lambda,\,\mu)$, which is finally multiplied with the noisy spectrum $X(\lambda,\,\mu)$ for the desired noise suppression. Different gain combinations are possible and will be discussed.

A serial processing of the two noise reduction stages leads to a concept where both gains are multiplied successively to the noisy spectrum. Then the combined gain reads

$$G(\lambda,\,\mu) = G_1(\lambda,\,\mu) \cdot G_2(\lambda,\,\mu) \tag{5.3}$$

and an aggressive noise reduction is realized because a multiplication of two gains in the range between one and zero will always lead to a combined gain smaller than both gains $G_1(\lambda,\,\mu)$ and $G_2(\lambda,\,\mu)$.

A further quite aggressive approach is to use the minimum of both gains

$$G(\lambda,\,\mu) = \min\{G_1(\lambda,\,\mu), G_2(\lambda,\,\mu)\}, \tag{5.4}$$

which limits the combined gain at least to the smaller of the to gains.

To realize a more moderate combined gain it also possible to average both gains $G_1(\lambda,\,\mu)$ and $G_2(\lambda,\,\mu)$. Here, the arithmetic mean

$$G(\lambda,\,\mu) = \frac{G_1(\lambda,\,\mu) + G_2(\lambda,\,\mu)}{2} \tag{5.5}$$

and the geometric mean

$$G(\lambda,\,\mu) = \sqrt{G_1(\lambda,\,\mu) \cdot G_2(\lambda,\,\mu)} \tag{5.6}$$

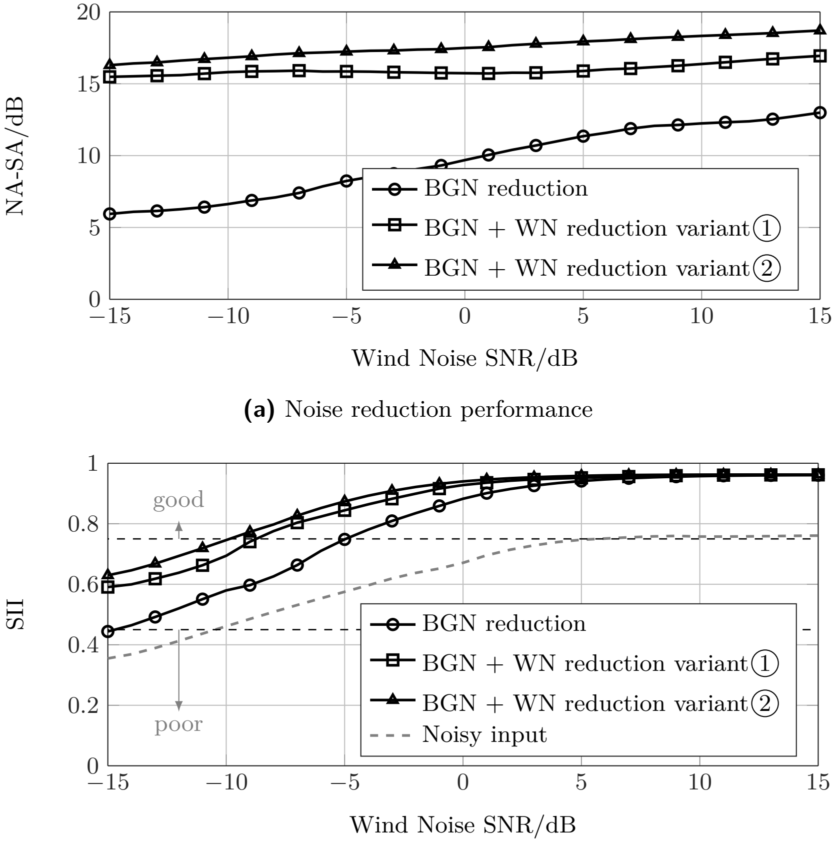
are considered. An analysis of the performance of the different combinations is given in the following section. For all proposed setups in Equations 5.3 to 5.6, the combined gain $G(\lambda,\,\mu)$ is limited to -40 dB.

5.1.2 Results

For the evaluation, noisy speech signals are generated containing both wind and background noise as depicted in Equation 5.1. To reflect different scenarios, both noise signals are scaled to different SNR values. The background noise signals are taken from the ETSI database ([ETS09]) using one of three typical noise types for an outdoor environment (*Fullsize Car1 130Kmh, Outside Traffic Road, Work Noise Jackhammer*).

The proposed scheme for combined wind and background noise reduction is evaluated using the noise attenuation - speech attenuation (NA-SA) metrics for the noise reduction performance and the speech intelligibility index (SII) measures to predict the intelligibility enhancement. The following presented measures are averaged over the three considered background noise types.

In a first investigation the two variants controlled by the position of switch A (①,②) in Figure 5.1 are compared using the gain combination by multiplication (Equation 5.3). In order to investigate different background noise and wind noise scenarios, two experiments are carried out. Firstly, in Figure 5.2 the speech-to-

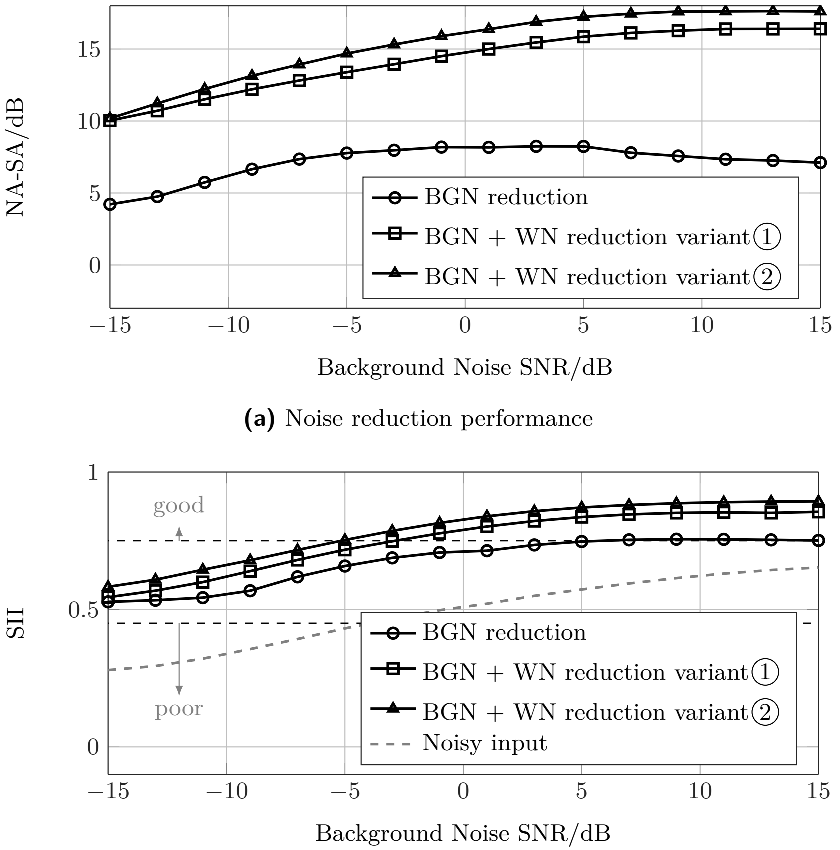


(a) Noise reduction performance

(b) Intelligibility enhancement

Figure 5.2: Results different wind noise (WN) SNR and a fixed background noise (BGN) SNR of 5 dB.

wind-noise ratio is varied between -15 and 15 dB using a fixed background noise SNR of 5 dB for the simulations. Secondly, the speech-to-background-noise ratio takes values between -15 and 15 dB while a wind noise SNR condition of -5 dB is considered (see Figure 5.3).



(a) Noise reduction performance

(b) Intelligibility enhancement

Figure 5.3: Results for different background noise (BGN) SNR and a fixed wind noise (WN) SNR of -5 dB.

The motivation for both fixed SNR values in the experiments is that realistic conditions of the considered noise type should be investigated. For the noise attenuation (NA) required for the NA-SA values, the reduction of the complete noise (wind noise + background noise) is taken into account. The evaluation is carried out to compare three configurations for the aforementioned SNR scenarios:

1. background noise (BGN) reduction,
2. serial processing of background noise and wind noise (WN) reduction without any exchange of information (BGN + WN) (switch A open),
3. the modified combination proposed in Figure 5.1 (switch A closed).

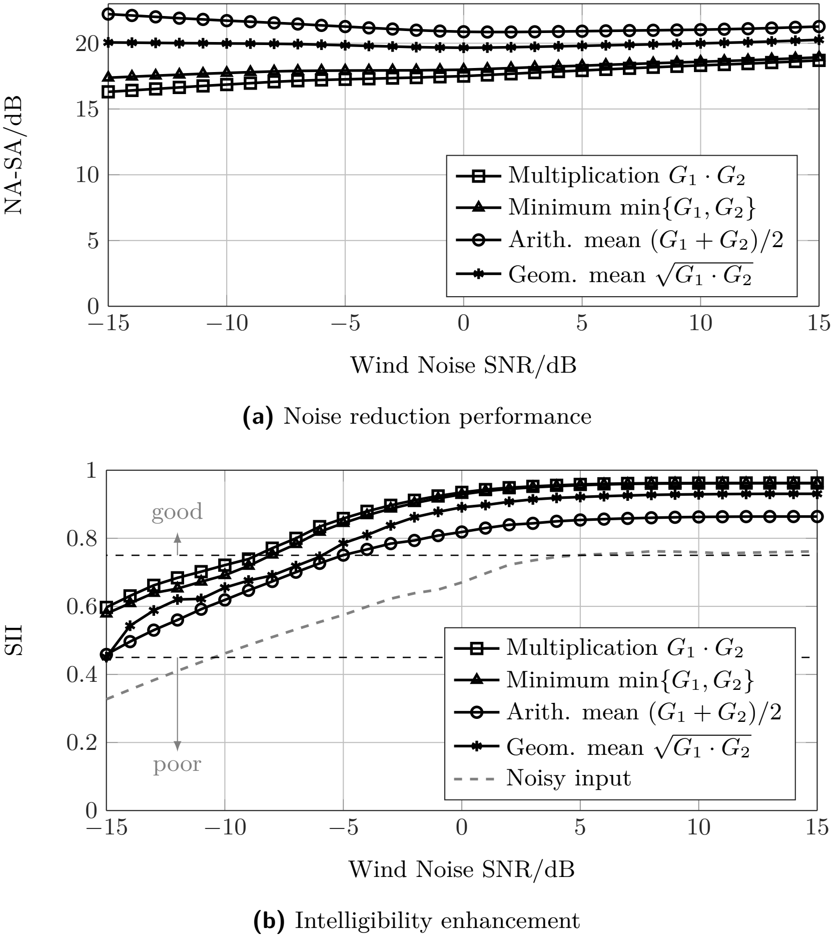
For all considered wind noise SNR conditions in Figure 5.2, the modified combination results in the highest performance for both the noise reduction and speech intelligibility enhancement. As expected, the BGN reduction alone only shows limited improvements resulting in lower values compared to a combined approach, especially for the NA-SA measure in Figure 5.2a. It can be seen, that both configurations for the combined reduction achieve high noise reduction and a great enhancement of the speech intelligibility. In some cases the conventional noise reduction also removes parts of the wind noise signal, which are necessary for the detection and the associated wind noise estimation. Thus, the modified combination, where the unfiltered input is used for the detection stage of the proposed estimation concept, results in a higher performance, due to a better detection of the wind noise signal.

A similar behaviour is observed for both measures regarding a variation of the background noise level in Figure 5.3. Here, it is also noticeable, that the difference between the three considered methods is decreasing for lower SNR values. This is due to the fact that in these conditions the background noise is dominant and thus the background noise reduction dominates the quality of the complete noise reduction system.

The second aspect consider in this evaluation is the gain combination of the two gains calculated for the background noise reduction $G_1(\lambda, \mu)$ and the wind noise reduction $G_2(\lambda, \mu)$. The proposed approaches in Equations 5.3-5.6 are compared using the same SNR scenarios as for the previous investigations and the NA-SA measure and the SII values. The results are shown in Figure 5.4 for varying wind noise SNRs and in Figure 5.5 for varying background noise SNRs. For both SNR scenarios two issues stand out.

As expected, the multiplication and the minimum of the two gains results in a quite aggressive noise reduction and the high speech attenuation leads to degraded NA-SA values. This can be seen in Figures 5.4a and 5.5a where the multiplication and the minimum leads to the lowest measures in all cases. The averaging of the two gains results to a better noise reduction performance ensuring a high NA-SA value of 20 to 22 dB for the arithmetic mean combination representing the best performance.

A different performance can be seen from the intelligibility enhancement depicted in Figures 5.4b and 5.5b. All proposed methods yields an improvement compared to the SII of the noisy input presented by the dashed gray curve. As already explained in previous parts of this work an aggressive noise reduction might not improve the subjective auditory impression but achieves an enhanced intelligibility. The aggressive methods showing the lowest NA-SA measures provide the highest SII improvements and vice versa for the moderate methods which average the two



(a) Noise reduction performance

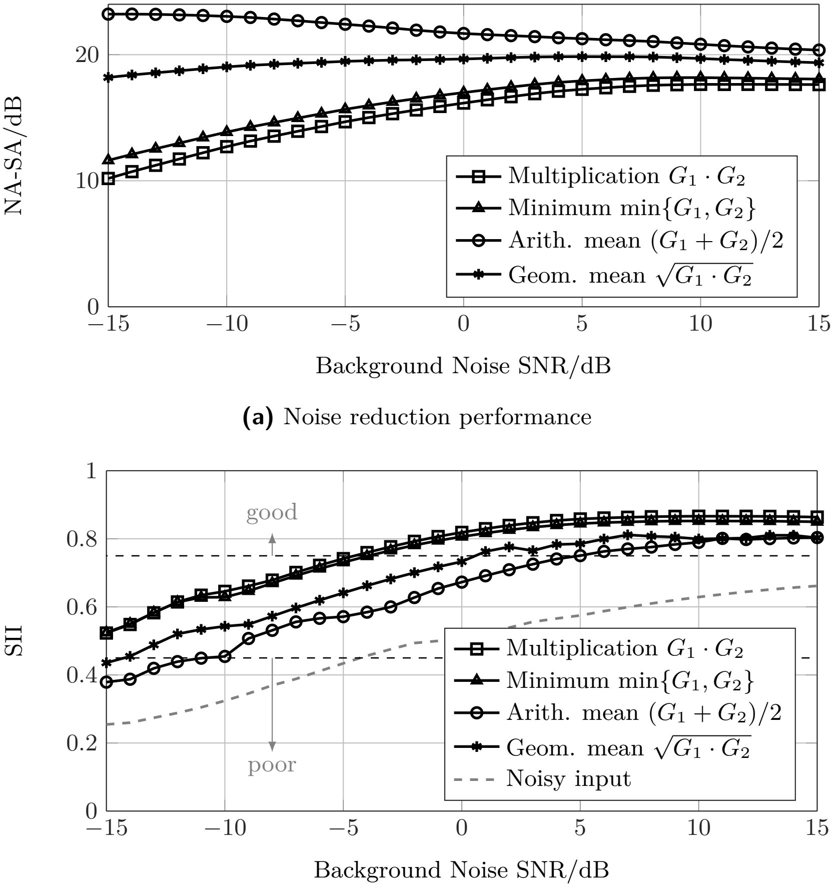
(b) Intelligibility enhancement

Figure 5.4: Results for different gain combinations for different wind noise SNR and a fixed background noise SNR of 5 dB.

spectral gains. This leads to a difference of 0.1 for the SII between the multiplication method and the arithmetic mean combination.

The auditory impression of the output signals supports this results. For the multiplication of the two gains, parts of speech are clearly degraded and in some cases, where both wind noise and background noise are active in the lower frequency range (e.g., for car noise), speech is partially completely attenuated but not necessarily unintelligible.

The results presented in this section support the proposed combined noise



(a) Noise reduction performance

(b) Intelligibility enhancement

Figure 5.5: Results performance for different gain combinations for different background noise SNR and a fixed wind noise SNR of -5 dB.

reduction concept, where first the background noise is estimated and reduced and subsequently the wind noise is considered. For the wind noise detection based on the NSTM the noisy input signal should be used, as the processing for background noise reduction decreases the detection accuracy. The choice of the gain combination of the background noise reduction gain and the wind noise reduction gain depends on the application. If a high noise reduction performance is desired, the arithmetic mean of the gains leads to best results. In contrast to that the aggressive approach of the gain multiplication achieves the highest intelligibility improvements. A good trade-off is given by the geometric mean of the two gains with a high NA-SA measure also considerably good SII improvements. Here, also a good listening impression is provided where not too high speech attenuation is introduced.

5.2 Dual Microphone Noise Reduction

Figure 5.6 depicts the considered dual microphone arrangement for the mobile phone application. This configuration can be found in many currently available smart-phones. The setup allows a microphone distance of approximately 10 cm. While the primary microphone is always at the bottom of the device, the reference microphone can be placed at the top or the back of the phone. The signals of both microphones can be exploited for the reduction of background noise for the two scenarios explained in the following.

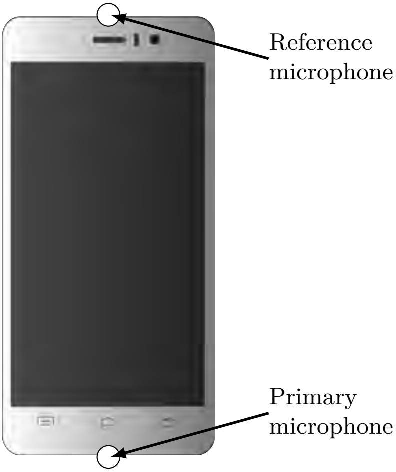


Figure 5.6: Dual microphone configuration for mobile phone.

5.2.1 Hand-held Telephony

In the hand-held position, the primary microphone is close to the the mouth to ensure a high level of the desired speech signal. At the reference microphone, clearly lower speech levels can be observed. In contrast to this, the noise signal levels in both microphones are very similar, if a homogeneous noise field is given. In [JHN+12] the level differences of the two microphone signals were exploited yielding a frequency dependent voice activity detector (VAD). Based on the VAD, the noise power spectral density (PSD) estimate can be calculated by taking either the signal from the primary microphone (speech absence) or from the reference microphone (speech presence). This procedure is applied along with a modified Wiener filter for the noise reduction, which also incorporates the power level differences of the microphone signals. A more detailed presentation of this method and evaluation results can be found in [JHN+12], [HJN+11] and [Jeu12].

5.2.2 Hands-free Telephony

Besides the previously described case of hand-held position, mobile phones can operate in the hands-free mode. This can be useful, when more than one person participates at the near-end side, for phone calls during a car drive, or for making video calls. Since the mobile device is not in a steady position as in the hand-held case, no assumptions about the power levels of speech and noise can be made at the two microphone positions. In most of the cases the power level differences are very similar for both speech and background noise. Hence, other characteristics must be taken into account for the differentiation between speech and noise. The primary and reference microphone are treated of equal value in the following.

For many situations, the sound field of the undesired background noise can be assumed as a diffuse noise field as explained in Section 3.3.4. Then, the spatial correlation between signals can be exploited in terms of the frequency dependent short-term coherence function

$$\Gamma_{xy}(\lambda,\,\mu) = \frac{\widehat{\Phi}_{xy}(\lambda,\,\mu)}{\sqrt{\widehat{\Phi}_{xx}(\lambda,\,\mu)\cdot\widehat{\Phi}_{yy}(\lambda,\,\mu)}}. \tag{5.7}$$

The short-term estimates of the auto and cross PSDs ($\widehat{\Phi}_{xx}(\lambda,\,\mu)$, $\widehat{\Phi}_{yy}(\lambda,\,\mu)$, $\widehat{\Phi}_{xy}(\lambda,\,\mu)$) are computed by the first order smoothing defined in Equations 3.46 and 3.47.

For an ideal diffuse noise field, $\Gamma_{xy}(\lambda,\,\mu)$ can be modeled by the sinc function (see Equation 3.10). The speech is often assumed to be coherent ($\Gamma_{ss}(\lambda,\,\mu) = 1$). However, these conditions are not exactly fulfilled in many real environments, i.e., $\Gamma_{ss}(\lambda,\,\mu) \neq 1$. One constraint is that the microphones do not show an omnidirectional characteristic due to the mounting into the mobile phone. This effect as well as reflections and reverberation have an impact on the coherence properties of the speech signals [BW01], [Jeu12]. Additionally, the assumption of an ideal diffuse noise field is mostly not fulfilled, because of some coherent noise sources in the background. These coherent portions result in an increase of the noise coherence function. The deviations of measured coherence functions from the theoretical curves are shown in Figure 3.8a and 3.8b for speech and noise, respectively. A further drawback of the coherence properties even under ideal conditions, is that both speech and noise exhibit high coherence values at low frequencies. Thus the separation is more difficult in this frequency range.

The proposed noise estimation tackling these problems is realized in two steps and is depicted in Figure 5.7 ([NBV13]). The advantages of a single and dual microphone processing are combined. The first stage is the single microphone speech presence probability (SPP) based noise estimation method [GH11] as introduced

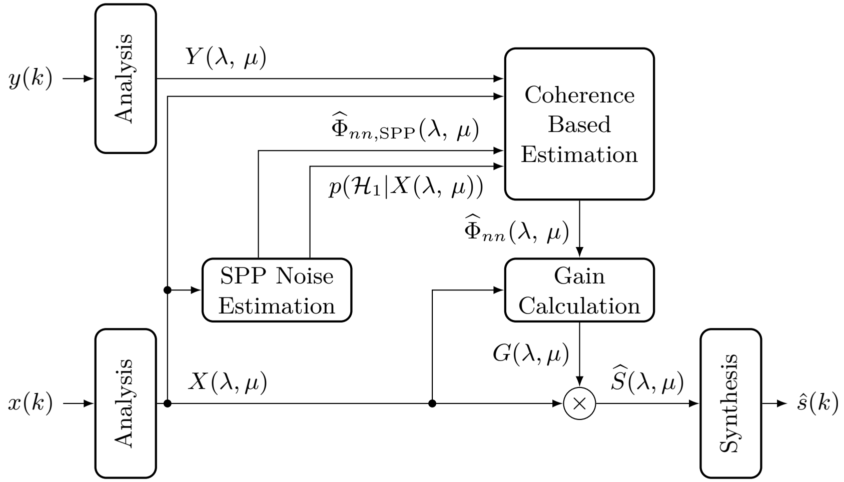

$y(k)$
Analysis
$Y(\lambda,\,\mu)$
Coherence Based Estimation
$\widehat{\Phi}_{nn,\mathrm{SPP}}(\lambda,\,\mu)$
$p(\mathcal{H}_1|X(\lambda,\,\mu))$
$\widehat{\Phi}_{nn}(\lambda,\,\mu)$
SPP Noise Estimation
Gain Calculation
$G(\lambda,\mu)$
$x(k)$
Analysis
$X(\lambda,\mu)$
$\widehat{S}(\lambda,\mu)$
Synthesis
$\hat{s}(k)$

Figure 5.7: Dual microphone system for background noise reduction.

in Section 2.3.1. The resulting estimate of the noise PSD $\widehat{\Phi}_{nn,\mathrm{SPP}}(\lambda,\,\mu)$ according to Equation 2.18 is calculated using the signal of the first microphone $X(\lambda,\,\mu)$. Besides, the SPP $p(\mathcal{H}_1|X(\lambda,\,\mu))$ is computed in each time-frequency bin (see Equation 2.14). Both quantities are used in the second stage, which also incorporates the coherence properties of the two microphone signals $X(\lambda,\,\mu)$ and $Y(\lambda,\,\mu)$ for the noise PSD estimation. The coherence based component of the proposed system also incorporates an update of the speech coherence function $\Gamma_{ss}(\lambda,\,\mu)$ and the noise coherence function $\Gamma_{nn}(\lambda,\,\mu)$, which might vary over time. The noise PSD estimate is then used for the SNR estimation and subsequent spectral gain computation as depicted in Figure 5.7.

Coherence Based Noise Estimation

The coherence based noise estimation can be seen as a generalized version of the method by Dörbecker in [DE96] already mentioned in Section 4.4.3. A first adaptation to diffuse noise fields was proposed in [JNK+11] and further developed in [NBV13] in order to circumvent limitations, which arises in practice.

We assume that speech and noise signals are uncorrelated. Then, the auto- and cross PSDs of the input signals are given by

$$\widehat{\Phi}_{xx}(\lambda,\,\mu) \quad = \quad \widehat{\Phi}_{s_1s_1}(\lambda,\,\mu) + \widehat{\Phi}_{n_1n_1}(\lambda,\,\mu) \tag{5.8}$$

$$\widehat{\Phi}_{yy}(\lambda,\,\mu) \quad = \quad \widehat{\Phi}_{s_2s_2}(\lambda,\,\mu) + \widehat{\Phi}_{n_2n_2}(\lambda,\,\mu) \tag{5.9}$$

$$\widehat{\Phi}_{xy}(\lambda,\,\mu) \quad = \quad \widehat{\Phi}_{s_1s_2}(\lambda,\,\mu) + \widehat{\Phi}_{n_1n_2}(\lambda,\,\mu). \tag{5.10}$$

Furthermore, we assume a homogeneous speech and noise field in both microphone signals of the system, i.e.

$$\widehat{\Phi}_{s_1 s_1}(\lambda,\,\mu) \;=\; \widehat{\Phi}_{s_2 s_2}(\lambda,\,\mu) \;=\; \widehat{\Phi}_{ss}(\lambda,\,\mu) \tag{5.11}$$

$$\widehat{\Phi}_{n_1 n_1}(\lambda,\,\mu) \;=\; \widehat{\Phi}_{n_2 n_2}(\lambda,\,\mu) \;=\; \widehat{\Phi}_{nn}(\lambda,\,\mu). \tag{5.12}$$

In [JNK$^+$11], we assumed ideal coherent speech ($\Gamma_{ss}(\lambda,\,\mu) = 1$). This is, however, not always fulfilled in real situation as it was shown in Section 3.3.4. In the following we neglect this assumption and thus, the cross PSD in (5.10) can be rewritten with (5.7) and (5.11, 5.12) as

$$\widehat{\Phi}_{xy}(\lambda,\,\mu) = \Gamma_{ss}(\lambda,\,\mu) \cdot \widehat{\Phi}_{ss}(\lambda,\,\mu) + \Gamma_{nn}(\lambda,\,\mu) \cdot \widehat{\Phi}_{nn}(\lambda,\,\mu), \tag{5.13}$$

where $\Gamma_{ss}(\lambda,\,\mu)$ and $\Gamma_{nn}(\lambda,\,\mu)$ are the coherence functions of the speech and noise signals[1], respectively. Inserting Equations (5.11) and (5.12) in Equations (5.8) and (5.9) and using the geometric mean of the two auto PSDs leads to

$$\sqrt{\widehat{\Phi}_{xx}(\lambda,\,\mu) \cdot \widehat{\Phi}_{yy}(\lambda,\,\mu)} = \widehat{\Phi}_{ss}(\lambda,\,\mu) + \widehat{\Phi}_{nn}(\lambda,\,\mu). \tag{5.14}$$

Resolving Equation (5.13) into

$$\widehat{\Phi}_{ss}(\lambda,\,\mu) = \frac{\widehat{\Phi}_{xy}(\lambda,\,\mu) - \Gamma_{nn}(\lambda,\,\mu) \cdot \widehat{\Phi}_{nn}(\lambda,\,\mu)}{\Gamma_{ss}(\lambda,\,\mu)} \tag{5.15}$$

and inserting in Equation 5.14 results in

$$\Phi_{nn}^{'}(\lambda,\,\mu) = \frac{\sqrt{\widehat{\Phi}_{xx}(\lambda,\,\mu) \cdot \widehat{\Phi}_{yy}(\lambda,\,\mu)} - \frac{\widehat{\Phi}_{xy}(\lambda,\,\mu)}{\Gamma_{ss}(\lambda,\,\mu)}}{1 - \frac{\Gamma_{nn}(\lambda,\,\mu)}{\Gamma_{ss}(\lambda,\,\mu)}}. \tag{5.16}$$

In periods, where speech is not predominant (i.e., in speech pauses), it turned out that a weighted average with the noisy input signal (e.g., from the first microphone) is more accurate than the estimate from (5.16). Therefore, the final noise PSD estimate of the coherence based stage is given by

$$\widehat{\Phi}_{nn,\mathrm{coh}}(\lambda,\,\mu) = \rho_{coh}(\lambda,\,\mu) \cdot \Phi_{nn}^{'}(\lambda,\,\mu) + (1 - \rho_{coh}(\lambda,\,\mu)) \cdot |X(\lambda,\,\mu)|^2. \tag{5.17}$$

The weighting factor

$$\rho_{\mathrm{coh}}(\lambda,\,\mu) = \frac{\widehat{\Gamma}_{xy}(\lambda,\,\mu) - \Gamma_{nn}(\lambda,\,\mu)}{\Gamma_{ss}(\lambda,\,\mu) - \Gamma_{nn}(\lambda,\,\mu)} \tag{5.18}$$

is a function of the measured short-term coherence $\widehat{\Gamma}_{xy}(\lambda,\,\mu)$ in the current signal frame. It is calculated as defined in Equation 5.7, where the required PSDs are given by the first-order recursive smoothing given in Equation 3.47. The parameter $\rho_{\mathrm{coh}}(\lambda,\,\mu)$ can be interpreted as a dual microphone soft voice indicator, similar to the single channel SPP value $p(\mathcal{H}_1|X(\lambda,\,\mu))$.

[1] The coherence functions for the speech and noise signals are defined in the same way as described in Equation 5.7, but using the cross- and auto-PSDs of speech or noise only.

Coherence Function Update

The coherence based noise estimate given in Equations 5.16 to 5.18 requires the coherence functions of the speech signals $\Gamma_{ss}(\lambda, \mu)$ and noise signals $\Gamma_{nn}(\lambda, \mu)$. This can be constant functions as described in [JNK+11]. In practice, $\Gamma_{ss}(\lambda, \mu)$ and $\Gamma_{nn}(\lambda, \mu)$ are not known and might also change over time. Therefore, we propose to update $\Gamma_{ss}(\lambda, \mu)$ by recursive smoothing with α_Γ in periods, where speech is predominant and $\Gamma_{nn}(\lambda, \mu)$ in periods, where speech is absent. The speech presence probability $p(\mathcal{H}_1|X(\lambda, \mu))$ from Equation 2.18 determines these periods by applying a simple threshold. The update rule is based on the short-term coherence $\widehat{\Gamma}_{xy}(\lambda, \mu)$ and reads for the noise coherence function

$$\widehat{\Gamma}_{nn}(\lambda, \mu) = \alpha_\Gamma \cdot \widehat{\Gamma}_{nn}(\lambda-1, \mu) + (1-\alpha_\Gamma)\cdot \widehat{\Gamma}_{xy}(\lambda, \mu), \;\; \forall\, \mu \in \{p(\mathcal{H}_1|X(\lambda, \mu)) < 0.1\}. \tag{5.19}$$

This rule uses the speech pauses to update the noise coherence function $\widehat{\Gamma}_{nn}(\lambda, \mu)$ in time-frequency bins with a low SPP.

The same rule can not be applied directly for the update of the speech coherence function because a high SPP value $p(\mathcal{H}_1|X(\lambda, \mu))$ does not necessarily indicate a noise-free speech segment. Hence, the influence of the noise must be taken into account. Using Equations 5.8 - 5.10 and assuming again that noise and speech signals are uncorrelated, the coherence function of Equation 5.7 can be expressed as

$$\begin{aligned}\Gamma_{xy}(\lambda, \mu) &= \frac{\widehat{\Phi}_{s_1s_2}(\lambda, \mu) + \widehat{\Phi}_{n_1n_2}(\lambda, \mu)}{\sqrt{\widehat{\Phi}_{xx}(\lambda, \mu)\widehat{\Phi}_{yy}(\lambda, \mu)}} = \frac{\widehat{\Phi}_{s_1s_2}(\lambda, \mu) + \widehat{\Phi}_{n_1n_2}(\lambda, \mu)}{\widehat{\Phi}_{ss}(\lambda, \mu) + \widehat{\Phi}_{nn}(\lambda, \mu)} \\ &= \frac{\widehat{\Phi}_{s_1s_2}(\lambda, \mu)}{\widehat{\Phi}_{ss}(\lambda, \mu)}\left(1+\frac{\widehat{\Phi}_{nn}(\lambda, \mu)}{\widehat{\Phi}_{ss}(\lambda, \mu)}\right)^{-1} + \frac{\widehat{\Phi}_{n_1n_2}(\lambda, \mu)}{\widehat{\Phi}_{nn}(\lambda, \mu)}\left(1+\frac{\widehat{\Phi}_{ss}(\lambda, \mu)}{\widehat{\Phi}_{nn}(\lambda, \mu)}\right)^{-1}.\end{aligned} \tag{5.20}$$

With the definition of the *a posteriori* SNR

$$\gamma(\lambda, \mu) = \frac{\widehat{\Phi}_{xx}(\lambda, \mu)}{\widehat{\Phi}_{nn}(\lambda, \mu)} = \frac{\widehat{\Phi}_{ss}(\lambda, \mu) + \widehat{\Phi}_{nn}(\lambda, \mu)}{\widehat{\Phi}_{nn}(\lambda, \mu)} \tag{5.21}$$

and inserting the coherence function for speech $\Gamma_{ss}(\lambda, \mu)$ and noise $\Gamma_{nn}(\lambda, \mu)$ in Equation 5.20 the coherence can be rewritten as

$$\Gamma_{xy}(\lambda, \mu) = \Gamma_{ss}(\lambda, \mu)\frac{\gamma(\lambda, \mu) - 1}{\gamma(\lambda, \mu)} + \Gamma_{nn}(\lambda, \mu)\frac{1}{\gamma(\lambda, \mu)}. \tag{5.22}$$

For the *a posteriori* SNR, the noise PSD estimate from the previous frame and the smoothed noisy input are used to compute $\widehat{\Phi}_{nn}(\lambda, \mu)$ and $\widehat{\Phi}_{xx}(\lambda, \mu)$. Now Equation 5.22 can be rearranged and finally leads to the corrected speech coherence function

$$\Gamma_{ss,\mathrm{cor}}(\lambda,\,\mu) = \Gamma_{xy}(\lambda,\,\mu)\frac{\gamma(\lambda,\,\mu)}{\gamma(\lambda,\,\mu)-1} - \Gamma_{nn}(\lambda,\,\mu)\frac{1}{\gamma(\lambda,\,\mu)-1}. \tag{5.23}$$

As we now consider the influence of the noise signals, the update of the speech coherence function can be carried out similarly to Equation 5.19 during periods, where speech is active, i.e., with a high SPP value. The computation rule is then given by the following expression

$$\widehat{\Gamma}_{ss}(\lambda,\,\mu) = \alpha_{\Gamma}\cdot\widehat{\Gamma}_{ss}(\lambda-1,\,\mu)+(1-\alpha_{\Gamma})\cdot\Gamma_{ss,\mathrm{cor}}(\lambda,\,\mu),\;\; \forall\mu \in \{p(\mathcal{H}_1|X(\lambda,\,\mu)) > 0.9\}. \tag{5.24}$$

The smoothing constants in Equation 5.19 and Equation 5.24 are chosen to $\alpha_{\Gamma} = 0.95$ and the coherence functions are initialized as $\widehat{\Gamma}_{ss}(0,\mu) = 1$ for the speech and $\widehat{\Gamma}_{nn}(0,\mu)$ for an ideal diffuse noise field as expressed in Equation 3.10.

The second issue mentioned at the beginning of this section is the similar coherence characteristic of speech and noise for low frequencies. This leads to an inaccurate distinction between speech and noise signals. To circumvent this problem the SPP noise estimate $\widehat{\Phi}_{n,\mathrm{SPP}}(\lambda,\,\mu)$ is incorporated in the problematic frequency range. Then the final noise PSD estimate of the complete system is given by combining the estimates from Equations 2.18 and 5.17 and reads

$$\widehat{\Phi}_{nn}(\lambda,\,\mu) = \begin{cases} \widehat{\Phi}_{nn,\mathrm{SPP}}(\lambda,\,\mu), & \text{if } \mu < \mu_{\mathrm{s}} \\ \widehat{\Phi}_{nn,\mathrm{coh}}(\lambda,\,\mu), & \text{else,} \end{cases} \tag{5.25}$$

where μ_s represents the split-frequency between the single microphone and dual microphone noise estimate. Here, we propose to use the frequency, where the MSC of the ideal diffuse coherence in (3.10) takes the value 0.5. All parameters for the SPP based components of the system are chosen as proposed in [GH11].

5.2.3 Evaluation

As in Chapter 4, the estimation accuracy as well as the noise reduction performance is rated using the logarithmic error $e_{\log}$ of the noise PSD estimate (see Equation A.4) and the speech quality measures NA-SA and SII. For realistic signal generation, a mock-up phone is used, which is equipped with two microphones with a distance of 10 cm. The speech signal of the hands-free scenario is produced by an artificial head including a mouth simulator (HEAD acoustics HMS II.3), where the mock-up phone is situated 50 cm in front of the head according the ETSI EG 201 377-2 standard [ETS04]. The diffuse noise field is generated by four loudspeakers in the **ind** audio laboratory[2]. This is carried out by the procedure defined in the ETSI standard EG 202 396-1 [ETS09] using the four noise signals from the provided

[2] The audio laboratory is a measurement room with low reverberation ($T_{60} < 100\,\mathrm{ms}$) and good isolation from surrounding signals.

database (*pub noise, work noise jackhammer, outside traffic crossroads, fullsize car1 130Kmh*) and two artificial noise signals (constant and modulated white noise). All results shown in the following are averaged over all noise types. The evaluation is carried out, comparing the single channel SPP based method, [GH11] the original coherence based approach (CohB) presented in [JNK$^+$11] assuming constant coherence properties, and the proposed advanced method [NBV13].

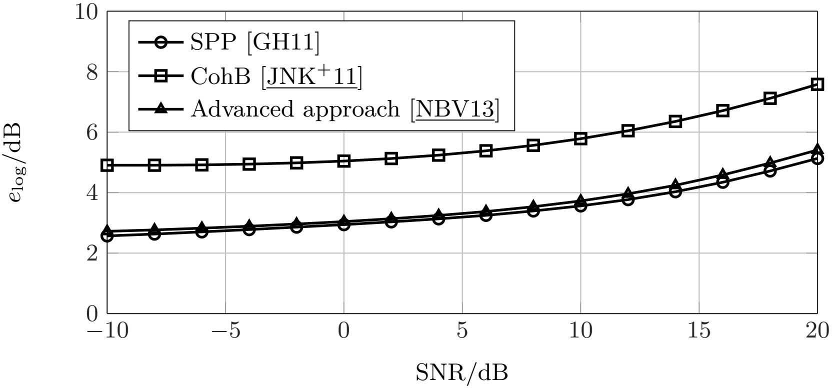
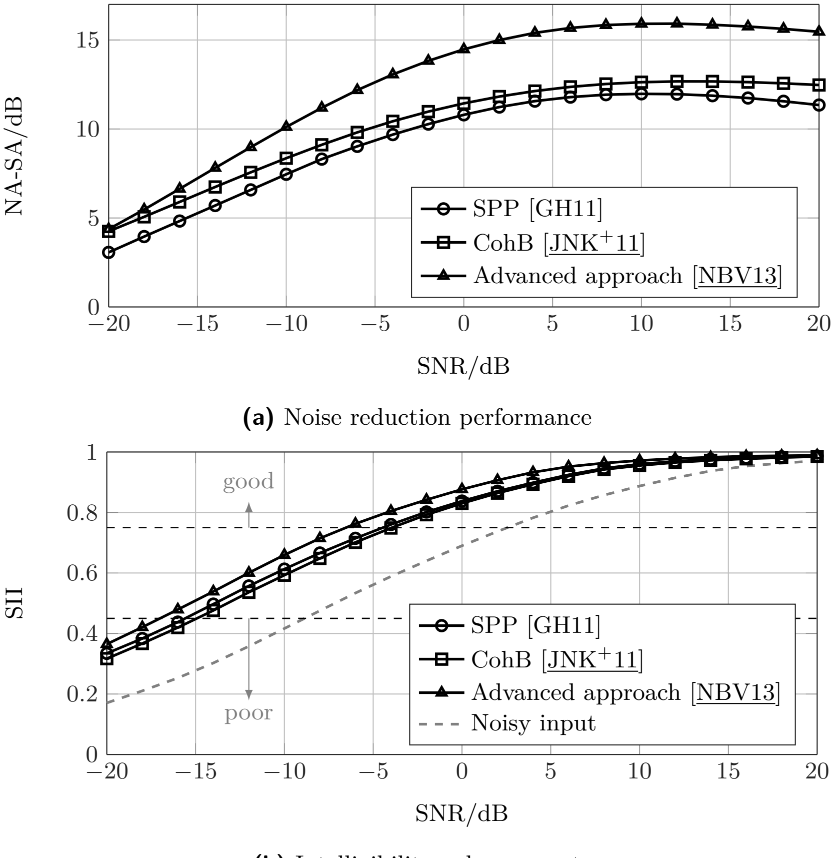


Figure 5.8: Estimation accuracy in terms of the logarithmic error $e_{\log}$.

The logarithmic error in Figure 5.8 is depicted for different input SNRs between -10 and 20 dB. For all cases both the SPP and the advanced method show the best results with approximately 2 dB lower error. This results seem not to indicate any advantages from the use of dual microphone characteristics, but considering the noise reduction performance presented in Figure 5.9, the advanced approach shows the highest improvement. In contrast to the estimation accuracy the SPP based method results in lower values in terms of the NA-SA measure as shown in Figure 5.9a. This is due to the property, that the SPP noise tracker applies a rather aggressive noise reduction, i.e., a high noise reduction is applied at the price of undesired speech attenuation.

The intelligibility enhancement presented in Figure 5.9b indicates an improvement for all algorithms compared to the noisy input signals marked by the dashed gray line. Again, the advanced method achieves the highest SII value ensuring to avoid “poor” intelligibility conditions for SNRs greater than 16 dB.



(a) Noise reduction performance

(b) Intelligibility enhancement

Figure 5.9: Evaluation of speech enhancement performance.

5.3 Conclusions

In this chapter two realistic scenarios are discussed in which the speech signal captured by a mobile phone is degraded by different noise types. For both scenarios, solutions are proposed to bypass problems usually occurring in realistic environments. These are, that not always ideal conditions can be assumed as the coherence properties or appearance of only clean speech and wind noise without any further background noise.

First, the scenario is investigated, where not only wind noise but also background noise is present. This is an important issue, because the complete speech enhancement system must be robust to scenarios with additional noise sources. The proposed scheme applies a conventional noise reduction followed by the wind noise reduction. The evaluation with different noise signals showed that an efficient noise reduction can be achieved. A further improvement can be reached, if the wind noise detection exploits properties gained from the unfiltered input signal. Different combinations of the background noise and wind noise suppression are investigated. Here the geometric mean of the spectral gains for two reduction stages leads to a high noise reduction of up to 22, dB and at the same time an enhanced intelligibility.

In the second part of this chapter, dual microphone solutions are presented to combat background noise for the application of mobile phones in hand-held and hands-free position. A short description of the principle is given, which exploits the power level differences of speech and noise for the detection and estimation of the noise PSD. A more detailed solution is presented in the case of the hands-free scenario. Here, the coherence properties of the speech and the noise field are considered. A system is proposed that solves two problems of a coherence based processing:

1. non-ideal coherence properties,
2. high correlation of low-frequency diffuse noise.

This is realized by a combination of a single microphone system exploiting the temporal characteristics in terms of the SPP with the coherence based processing using both microphone signals. Here, a clear improvement of the noise reduction performance is measurable using real recordings captured by a dual microphone mock-up phone.

Chapter 6

Summary

So far conventional approaches for speech enhancement are not capable to reduce wind noise. Hence, special algorithms are developed and presented. Different prerequisites are considered, driven by the number of microphones or the application of the used system. The *temporal*, *spectral*, and in the dual microphone case the *spatial* properties are investigated for the detection, estimation and reduction of wind noise. All proposed algorithms are evaluated with real recordings and compared to state-of-the-art wind noise reduction methods. It turns out that the proposed techniques clearly outperform the previous methods with respect to the increase in signal-to-noise ratio and speech intelligibility. This was proven by numerous benchmarks with standard objective and perceptual measures for speech quality assessment.

Signal Analysis and Modelling

After a short introduction in the principles of noise estimation and speech enhancement, the first focus of the thesis was the investigation of wind noise from a digital signal processing perspective. In a thorough analysis the statistics of the recorded digital representation of wind noise and its distinct characteristics were presented in detail. Different properties in the time domain, the discrete Fourier transform (DFT) domain or regarding the spatial correlation of wind noise signals captured by two microphones were explored, always with regard to detect wind noise in a recorded signal in short segments. Based on the analysis of wind noise, different approaches for the detection were presented and compared in terms of their accuracy and robustness towards the presence of speech signals. In the time domain, the normalized short-term mean (NSTM) approach, which exploits the offset introduced by the wind noise, showed the best performance. Similar results were achieved by a method in the frequency domain that decomposes the noisy speech signal into a speech template spectrum and a wind noise template spectrum. For systems with two microphones the averaged short-term coherence is applied as wind detector. Contrary to the expectation that the use of two microphone signals leads to an improved detection the results indicates only comparable detection rates. This is due to the computation of the coherence, which always includes an averaging process over time and leads to a decreased adaptation speed to the fast changing signal characteristics of wind noise.

Using the knowledge of the statistics of wind noise signals, a model was derived for the generation of artificial wind noise as digital signal. So far only models for the prediction of the long-term behavior of wind are known. The proposed model generates a signal with clearly higher temporal resolution and played a significant role for the development and testing chain of speech enhancement algorithms. The spectral characteristics are reproduced by an auto-regressive (AR) filter with prototypical coefficients. The non-stationary temporal behavior is simulated by a time-varying gain. It was shown that with an appropriate parameterization, the short-term energy can be modeled by a Weibull distribution. For the long-term behavior a Markov model has been applied for the representation of the different wind intensities. Comparative analyses showed such a high similarity between real wind noise recordings and the generated wind noise that the time- and cost-consuming recordings could be reduced to a minimum.

Wind Noise Estimation and Reduction

The main part of this thesis dealt with the estimation of the short-term power spectrum (STPS) of wind noise and the enhancement of the degraded speech signal. As all concepts for speech enhancement of a noisy signal require an estimate of the underlying noise, methods were developed, which can precisely determine the wind noise spectrum in a given signal containing both speech and wind. The spectral shapes of speech and wind noise were exploited for a distinction. The experimental comparison with other state-of-the-art wind noise estimators showed that the new methods lower the logarithmic error in order of 5 dB in all relevant wind noise conditions. The wind noise estimators were also compared as part of commonly used overlap-add structure with a spectral weighting gain for noise suppression. Here again, the proposed algorithm achieved the best performance considering both the noise reduction and the intelligibility enhancement. An improvement in terms of the difference between noise and speech attenuation (NA-SA) of over 15 dB can be achieved in all relevant cases.

Many present-day mobile devices are equipped with two microphones. Therefore, a new approach was derived for the estimation of wind noise using the short-term coherence. To solve the problems introduced by the non-stationary behavior of the wind noise, besides the magnitude, also the phase of the complex valued coherence has been used for the wind noise estimation. A comparison with other dual microphone wind noise reduction methods demonstrated similar intelligibility enhancement results, but an improved noise reduction performance.

All wind noise reduction concepts applying a spectral weighting gain suffer from a strong speech attenuation in the highly degraded parts at lower frequencies. Therefore, an innovative concept for speech enhancement was introduced. The basic idea is to partially reconstruct the degraded speech spectrum by parts of an artificially generated speech spectrum. By means of techniques known from the artificial bandwidth extension and pre-trained speech codebooks, the widely used source-filter model for speech production has been incorporated in the speech enhancement process. Perceptual measures and an evaluation in terms of the segmental SNR proved that the new concept can mitigate to a large extend the effects introduced by the conventional spectral weighting. This system is a completely new approach for speech enhancement and can be extended to combat a wider range of noise types. The final evaluation of all algorithms was performed with real wind noise recordings to prove their efficiency under realistic conditions.

Noise Reduction for Mobile Phones

It is of special interest that the speech enhancement techniques also hold in the context of realistic acoustic, i.e., non-ideal, situations. Exemplary, two concrete scenarios were discussed dealing with problems in realistic environments for a mobile phone application.

First, the integration of a wind noise reduction component into a conventional noise reduction system is investigated. In this context, the operation order was discussed with the consensus that first the background noise reduction should be applied followed by the wind noise reduction. The evaluation with speech signals degraded by both background and wind noise manifests this structure. A modification of this serial processing could even further improve the performance.

The second scenario considers dual microphone mobile phones for the use in a diffuse background noise field. In case of a hand-held telephony the power level differences of speech and noise can be exploited for the estimation of the noise power spectral density (PSD) and the subsequent background noise reduction. For the hands-free condition, a coherence based method was adopted to solve two problems of realistic recordings: (i) non-ideal coherence properties of the signals and (ii) high-coherent parts of diffuse noise for lower frequencies.

In conclusion, it can be stated that the algorithms proposed in this thesis can efficiently reduce the effects of wind noise and background noise in speech signals. Especially, the wind noise reduction techniques improves intelligibility in terms of an speech intelligibility index (SII) score indicating a poor intelligibility (SII < 0.45) to a range of good intelligibility (SII > 0.8). In addition, informal listening tests confirm a high quality of the processed speech signals. With these results, a high signal quality in many mobile communication devices can be ensured even under severe outdoor conditions. This thesis is the first, which addresses the complete problem of detecting and reducing wind noise from a signal processing perspective. The results provide valuable concepts for many applications, such as mobile mobile phones, outdoor microphones or hearing aids. All considered algorithms besides the partial speech synthesis are characterized by a low computational complexity, which is comparable to conventional noise reduction methods.

Appendix A

Evaluation Environment

A.1 Evaluation of Speech Enhancement

All considered speech enhancement algorithms in this thesis are realized in an overlap-add structure. The complete set-up for algorithms applying a spectral gain $G(\lambda,\, \mu)$ for noise reduction is depicted in Figure A.1. For the evaluation of the methods not only the mixed noisy signal $x(k)$ is used, but also the clean speech signal $s(k)$ and the pure noise signal $n(k)$, which are also available in the simulation environment. The same analysis procedure in terms of segmentation, windowing and fast Fourier transform (FFT) is applied to all input signals yielding the frequency domain representations $S(\lambda,\, \mu)$, $N(\lambda,\, \mu)$, and $X(\lambda,\, \mu)$. For systems

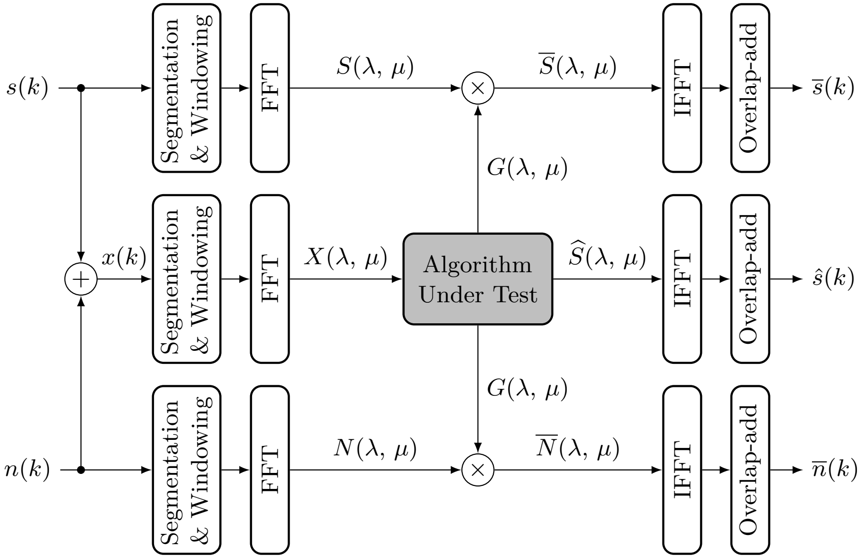


Figure A.1: Evaluation structure of speech enhancement algorithms

applying a spectral weighting, the gain function $G(\lambda, \mu)$ can also be multiplied with the clean speech spectrum $S(\lambda, \mu)$ and the pure noise spectrum $N(\lambda, \mu)$ where $G(\lambda, \mu)$ is only calculated based on the information given by $X(\lambda, \mu)$. After the noise reduction stage the three input signals results into an enhanced noisy signal $\widehat{S}(\lambda, \mu)$, the filtered clean speech $\overline{S}(\lambda, \mu)$ and the filtered noise signal $\overline{N}(\lambda, \mu)$ and their time-domain representations $\hat{s}(k)$, $\overline{s}(k)$ and $\overline{n}(k)$.

Different quality measures can be computed from a comparison the input and output signals of the presented evaluation structure. In this thesis, the noise attenuation - speech attenuation (NA-SA), speech intelligibility index (SII), perceptual evaluation of speech quality (PESQ) and segmental SNR (segSNR) are used.

Segmental Speech and Noise Attenuation

Comparing the clean speech $s(k)$ with the filtered speech $\overline{s}(k)$ and the input noise $n(k)$ with the filtered noise $\overline{n}(k)$, the segmental attenuation of the speech and noise signals due to the applied noise reduction can be calculated as

$$\mathrm{SA/dB} \;=\; \frac{1}{\#\{\mathcal{K}_\mathrm{s}\}} \sum_{l\in\mathcal{K}_\mathrm{s}} \left(10 \cdot \log_{10}\left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} s^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \overline{s}^2(k + l \cdot L_\mathrm{F})}\right)\right) \tag{A.1}$$

$$\mathrm{NA/dB} \;=\; \frac{1}{\#\{\mathcal{K}_\mathrm{t}\}} \sum_{l\in\mathcal{K}_\mathrm{t}} \left(10 \cdot \log_{10}\left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} n^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} \overline{n}^2(k + l \cdot L_\mathrm{F})}\right)\right) \tag{A.2}$$

For the speech attenuation only the set $\mathcal{K}_\mathrm{s}$ of frames with speech activity is considered, while for the noise attenuation the complete set off all signal frames $\mathcal{K}_\mathrm{t}$ is used.

Regarding the NA and SA measures separately, no direct proposition about the speech quality can be made. But difference between NA and SA indicates the effective noise reduction performance and predicts an enhancement for values greater 0 dB.

Segmental Signal-to-Noise Ratio

A further measure for the rating of the signal quality is the segmental signal-to-noise ratio. It is defined by the mean of all segments K_s with voice activity as follows

$$\mathrm{segSNR/dB} = \frac{1}{\#\{\mathcal{K}_\mathrm{s}\}} \sum_{l\in\mathcal{K}_\mathrm{s}} \left(10 \cdot \log_{10}\left(\frac{\sum_{k=0}^{L_\mathrm{F}-1} s^2(k + l \cdot L_\mathrm{F})}{\sum_{k=0}^{L_\mathrm{F}-1} (s(k + l \cdot L_\mathrm{F}) - \hat{s}(k + l \cdot L_\mathrm{F}))^2}\right)\right). \tag{A.3}$$

A.2 Evaluation of Noise Estimation Accuracy

The accuracy of the noise is often determined by means of the error between the noise estimate by the considered algorithm and a known reference noise signal (see, e.g., [TTM+11], [GH11]). Using the evaluation structure of Figure A.1, the true noise signal is given and can be used as reference. The logarithmic error is defined as

$$e_{\log}/\mathrm{dB} = \frac{1}{ML} \sum_{\lambda=0}^{L-1} \sum_{\mu=0}^{M-1} \left| 10 \log_{10} \left(\frac{\mathcal{N}_{\mathrm{ref}}(\lambda,\, \mu)}{\widehat{\mathcal{N}}(\lambda,\, \mu)} \right) \right| . \tag{A.4}$$

In conventional noise reduction systems often the noise PSD is estimated by a first-order recursive smoothing approach. As discussed throughout this work, smoothing of the wind noise estimate can lower the accuracy. Hence, the short-term power spectrum (STPS) of the noise signal

$$|\mathcal{N}_{\mathrm{ref}}(\lambda,\, \mu)|^2 = |N(\lambda,\, \mu)|^2 \tag{A.5}$$

is used as noise reference.

Appendix B

Derivation of Coherence Phase

The phase of the complex coherence function will be derived in the following as function of the magnitude and phase of the complex speech and noise spectra from a dual microphone configuration. It is assumed that both speech and noise show similar levels in both microphone signals. Then, the spectra of the two signals in both microphones 1|2 read

$$S_{1|2}(\lambda,\,\mu) = |S(\lambda,\,\mu)| \cdot e^{j\varphi_{s_{1|2}}(\lambda,\,\mu)}, \tag{B.1}$$

$$N_{1|2}(\lambda,\,\mu) = |N(\lambda,\,\mu)| \cdot e^{j\varphi_{n_{1|2}}(\lambda,\,\mu)}. \tag{B.2}$$

The noisy input signals of the two microphone are given by

$$X(\lambda,\,\mu) = |S(\lambda,\,\mu)| \cdot e^{j\varphi_{s_1}(\lambda,\,\mu)} + |N(\lambda,\,\mu)| \cdot e^{j\varphi_{n_1}(\lambda,\,\mu)}, \tag{B.3}$$

$$Y(\lambda,\,\mu) = |S(\lambda,\,\mu)| \cdot e^{j\varphi_{s_2}(\lambda,\,\mu)} + |N(\lambda,\,\mu)| \cdot e^{j\varphi_{n_2}(\lambda,\,\mu)}. \tag{B.4}$$

Regarding the short-term complex coherence function defined by

$$\widehat{\Gamma}(\lambda,\,\mu) = \frac{\widehat{\Phi}_{xy}(\lambda,\,\mu)}{\widehat{\Phi}_{xx}(\lambda,\,\mu) \cdot \widehat{\Phi}_{yy}(\lambda,\,\mu)}, \tag{B.5}$$

the auto PSDs $\widehat{\Phi}_{xx}(\lambda,\,\mu)$ and $\widehat{\Phi}_{yy}(\lambda,\,\mu)$ are real-valued, and only the cross-PSD $\widehat{\Phi}_{xy}(\lambda,\,\mu)$ is complex-valued. Hence, the phase of Equation B.5 is determined by $\widehat{\Phi}_{xy}(\lambda,\,\mu)$. In Section 4.4.3, the phase of the coherence is exploited to achieve a sufficient adaptation speed to the fast changing wind noise characteristics. Therefore, the smoothing constant α for the PSD calculation (see Equation 4.50) is set to zero. For the computation of the magnitude squared coherence (MSC) $\mathcal{C}_{xy}$, this choice of α is not recommended, because the required PSDs must be calculated as expectation over a certain time-span (see, [Car87]) and $\alpha = 0$ will lead to $\mathcal{C}_{xy} = 1$ for all signal types. But for the phase of the coherence or the cross-PSD, a characteristic information is given by this instantaneous calculation in each frame and will be shown in the following. That means, for the choice $\alpha = 0$ the cross-PSD computation reads

$$\widehat{\Phi}_{xy}(\lambda,\,\mu) = X(\lambda,\,\mu) \cdot Y^{*}(\lambda,\,\mu). \tag{B.6}$$

For the sake of clarity, the time and frequency indices λ and μ are omitted in the following equations. Inserting Equations B.3 and B.4 into Equation B.6, the cross-PSD reads

$$\begin{aligned} X \cdot Y^* \quad = \quad & |S|^2 \cdot \cos(\varphi_{s_1} - \varphi_{s_2}) + |N|^2 \cdot \cos(\varphi_{n_1} - \varphi_{n_2}) \\ & + |S \cdot N| \cdot (\cos(\varphi_{s_1} - \varphi_{n_2}) + \cos(\varphi_{n_1} - \varphi_{s_2})) \\ & + j \cdot [|S|^2 \cdot \sin(\varphi_{s_1} - \varphi_{s_2}) + |N|^2 \cdot \sin(\varphi_{n_1} - \varphi_{n_2}) \\ & + |S \cdot N| \cdot (\sin(\varphi_{s_1} - \varphi_{n_2}) + \sin(\varphi_{n_1} - \varphi_{s_2}))]. \end{aligned} \tag{B.7}$$

With the assumption of delay-compensated speech signals, i.e., $\varphi_{s_1} = \varphi_{s_2}$, the phase of the coherence or cross PSD is given by

$$\begin{aligned} \varphi_\Gamma = \angle\{\widehat{\Phi}_{xy}(\lambda,\, \mu)\} = \ \angle\{X \cdot Y^*\} = \arctan\left(\frac{\mathrm{Im}\{X \cdot Y^*\}}{\mathrm{Re}\{X \cdot Y^*\}}\right) \\ = \arctan\left(\frac{|N|^2 \sin(\varphi_{n_1} - \varphi_{n_2}) + |S||N|(\sin(\varphi_{s_1} - \varphi_{n_2}) + \sin(\varphi_{n_1} - \varphi_{s_2}))}{|S|^2 + |N|^2 \cos(\varphi_{n_1} - \varphi_{n_2}) + |S||N|(\cos(\varphi_{s_1} - \varphi_{n_2}) + \cos(\varphi_{n_1} - \varphi_{s_2}))}\right) \end{aligned} \tag{B.8}$$

as it is used in Equation 4.55. This relation between the distribution of the phase φ_Γ and the amplitudes of speech $|S|$ and noise $|N|$ can now be exploited for the detection as it is shown in Section 4.4.3.

Mathematical Notation & Abbreviations

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
$\vert x\vert$	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
$\lceil x \rfloor$	rounding function, i. e., closest integer to x
$\mathrm{E}\{x(k)\}$	expectation value of $x(k)$
$\widehat{\mathrm{E}}\{x(k)\}$	short-term expectation value of $x(k)$
$\mathrm{Re}\{x\}$	real part of x
$\mathrm{Im}\{x\}$	imaginary part of x
$\max\limits_{x}\{f(x)\}$	maximum of $f(x)$ over x
$\arg\max\limits_{x}\{f(x)\}$	argument x of maximum of $f(x)$ over x
$\overline{x}$	average of x
$\#\{X\}$	cardinality of X, i.e., number of elements in X
$\vert\vert\mathbf{x}\vert\vert$	norm, i.e., Euclidean distance of the vector $\mathbf{x}$
$\mathbf{x}^T$	transpose of the vector $\mathbf{x}$

Non-Mathematical Operators

$\hat{x}$	estimate of the signal or parameter x

$\widetilde{x}$ signals or parameters, which are not the direct result of observed signals, e.g., synthetic signals or pre-trained information in codebooks

Principal Symbols

α smoothing constant for recursive PSD calculation

$\alpha_{\mathbf{D}}$ decay constant of wind coherence model

α_e mixing parameter of wind noise excitation

α_ξ smoothing constant for “decision-directed” *a priori* SNR estimation

$\alpha_{\mathbf{S}}$ spectral subtraction parameter

$\beta_{\mathbf{S}}$ spectral subtraction parameter

γ *a posteriori* SNR

θ angle of arrival of signal

η phase complex DFT coefficients of noisy speech

κ sample index in the current frame

$\kappa_{\mathbf{W}}$ shape parameter of Weibull distribution

λ frame index

$\lambda_{\mathbf{W}}$ scale parameter of Weibull distribution

μ frequency bin

ν viscosity of air

Ξ sub-band signal centroid (frequency range of sub-band may be given as subscript)

ξ *a priori* SNR

$\xi_{\mathbf{opt}}$ optimal *a priori* SNR used in [GH11]

φ_Γ phase of complex coherence

Φ power spectral density

$\widehat{\Phi}$ short-term estimate of power spectral density

$\widehat{\Phi}_{nn}$ noise power spectral density estimate

Φ_{xx} (auto) power spectral density of a signal $x(k)$

Φ_{xy} cross power spectral density of the signals $x(k)$ and $y(k)$

ρ density of air

σ_{φ}^{2} phase variance

$\sigma_{E,\mathbf{ST}}^{2}$ variance of short-term frame energy over i frames

$\overline{\sigma^{2}}_{E,\mathbf{ST}}$ mean short-term variance of frame energy

$\sigma_{\mathbf{TSC}}$ weight for codebook decomposition

Υ long-term average speech spectrum

ω angular frequency ($\omega \equiv 2\pi f$)

Ω normalized angular frequency ($\Omega \equiv 2\pi f/f_{\mathrm{s}}$)

c speed of sound (343 m/s)

$\mathcal{C}$ magnitude squared coherence

$\mathbf{dB_{FS}}$ decible relative to full scale $[-1\ldots1]$

$D_{\mathbf{c}}$ characteristic dimension

$d_{\mathbf{m}}$ microphone distance

$\tilde{d}_{\mathbf{m}}$ effective microphone distance depending on θ

$e_{\mathbf{log}}$ logarithmic error

$E_{\mathbf{ST}}$ short-term energy of one signal frame

f continuous (analog) frequency

$f_{\mathbf{s}}$ sampling frequency

$g_{\mathbf{LT}}$ long-term gain

$G_{\mathbf{P}}$ prediction gain

$G_{\mathbf{S}}$ spectral subtraction filter gain in generalized form

$g_{\mathbf{ST}}$ short-term gain

$G_{\mathbf{W}}$ Wiener filter gain

$\mathcal{H}_0$ speech absence

$\mathcal{H}_1$ speech presence

$\mathcal{I}$ wind indicator, if not otherwise stated in the range between 0 and 1

k sample index

$K_{\mathbf{CB}}$ length of one codebook vector

$L_{\mathbf{F}}$ frame size

$L_{\mathbf{log}}$ logarithmic spectrum level

$l_{\mathbf{LP}}$ order of LP filter

M length of the DFT

$\mathcal{N}$ short-term spectrum of wind noise

$n_{\mathbf{syn}}$ synthetic wind noise signal

P sound pressure spectrum

$\mathcal{P}_{\bar{\mathbf{s}}}$ speech misdetection rate

$\mathcal{P}_{\mathbf{w}}$ wind detection rate

$p_{\mathbf{W}}$ PDF of a Weibull distribution

R magnitude complex DFT coefficients of noisy speech

R_e Reynolds number

$s_{\mathbf{seq}}$ sequence of discrete states

$\widetilde{S}$ synthetic speech signal

t continuous time variable

T_{60} reverberation time

U wind speed

$\tilde{U}$ normalized wind speed

u_∞ free-field velocity

Acronyms

ABWE artificial band width extension

AED adaptive eigenvalue decomposition

ANSI American National Standards Institute

AR auto-regressive

ASR automatic speech recognition

ASWE adaptive smoothing wind noise estimation

CDF cumulative distribution function

DC direct component

DCT discrete cosine transformation

DDA decision directed approach

DFT discrete Fourier transform

DOA direction of arrival

DSP digital signal processor

ETSI European Telecommunications Standards Institute

FFT fast Fourier transform

FIR finite impulse response

HFP hands-free position

HHP hand-held position

HPS harmonic product spectrum

IFFT inverse fast Fourier transform

IIR infinite impulse response

ITU International Telecommunication Union

LMS *least-mean-square*

LP linear prediction

LPC linear predictive coding

LSD logarithmic spectral distortion

LSF line spectral frequency

LT long-term

LTASS long-term average speech spectrum

MFCC mel-frequency cepstral coefficients

MMSE minimum mean square error

MORPH morphological approach

MOS mean opinion score

MSC magnitude squared coherence

MSE mean square error

NA noise attenuation

NLMS *normalized least-mean-square*

NSF negative slope fit

NSTM normalized short-term mean

PDF probability density function

PESQ perceptual evaluation of speech quality

P-IBM pitch adaptive inverse binary mask

PSD power spectral density

PSYN partial speech synthesis

ROC receiver operating characteristic

RSS recursive spectral subtraction

SA speech attenuation

segSNR segmental SNR

segSSNR segmental speech-signal-to-noise-ratio

SII speech intelligibility index

SNR signal-to-noise-ratio

SPP speech presence probability

SSC sub-band signal centroid

ST short-term

STOI short-time objective intelligibility

STPS short-term power spectrum

THD total harmonic distortion

TPC template pitch cycle

TSC template spectrum combination

VAD voice activity detector

WNR wind noise reduction

ZCR zero crossing rate

Bibliography

[AEJ+12] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. "Audio Inpainting". *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, March 2012.

[ANR74] N. Ahmed, T. Natarajan, and K. Rao. "Discrete Cosine Transform". *IEEE Trans. on Computers*, vol. 23, no. 1, pp. 90–93, January 1974.

[ANS97] ANSI S3.5-1997. "Methods for the Calculation of the Speech Intelligibility Index", 1997.

[B+11] M. Brookes et al. "Voicebox: Speech Processing Toolbox for MATLAB". *Imperial College, London, United Kingdom, Software available from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, March 2011.

[BCH08] J. Benesty, J. Chen, and Y. Huang. *Microphone array signal processing*, vol. 1. Springer Science & Business Media, 2008.

[BDT+94] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Hetu, J. Kei, C. Lui, et al. "An International Comparison of Long-term Average Speech Spectra". *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2108–2120, October 1994.

[Ben00] J. Benesty. "Adaptive Eigenvalue Decomposition Algorithm for Passive Acoustic Source Localization". *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, January 2000.

[BG09] J. Bernstein and K. Grant. "Auditory and Auditory-visual Intelligibility of Speech in Fluctuating Maskers for Normal-hearing and Hearing-impaired Listeners". *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3358–3372, May 2009.

[Bit02] J. Bitzer. *Mehrkanalige Geräuschunterdrückungssysteme - Eine vergleichende Analyse.* PhD thesis, Universität Bremen, September 2002.

[Bol79] S. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, April 1979.

[BW01] M. Brandstein and D. Ward. *Microphone Arrays - Signal Processing Techniques and Applications.* Springer Verlag, 2001.

[BWHB03] S. Bradley, T. Wu, S. Hünerbein, and J. Backman. “The Mechanisms Creating Wind Noise in Microphones”. *Audio Engineering Society, 114th Convention*, Amsterdam, The Netherlands, March 2003.

[Car87] G. Carter. “Coherence and Time Delay Estimation”. *Proceedings of the IEEE*, vol. 75, no. 2, pp. 236–255, February 1987.

[CB01] I. Cohen and B. Berdugo. “Speech Enhancement for Non-stationary Noise Environments”. *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, June 2001.

[CBK15] N. Chatlani, C. Beaugeant, and P. Kroon. “Low Complexity Single Microphone Tonal Noise Reduction in Vehicular Traffic Environments”. *Proc. of European Signal Processing Conf. (EUSIPCO)*, Nice, France, September 2015.

[CCS$^+$09] R. Chen, C. Chan, H. So, J. Lee, and C. Leung. “Speech Enhancement in Car Noise Envoronment Based on an Analysis-synthesis Approach Using Harmonic Noise Model”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Taipei, Taiwan, April 2009.

[Chu04] W. C. Chu. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders.* John Wiley & Sons, 2004.

[Coh03] I. Cohen. “Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging”. *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 5, pp. 466–475, September 2003.

[Cor64] G. M. Corcos. “The Structure of the Turbulent Pressure Field in Boundary-layer Flows”. *Journal of Fluid Mechanics*, vol. 18, pp. 353–378, February 1964.

[Cro07] M. J. Crocker. *Handbook of Noise and Vibration Control.* John Wiley & Sons, 2007.

[DE96] M. Dörbecker and S. Ernst. “Combination of Two-Channel Spectral Subtraction and Adaptive Wiener Post-Filtering for Noise Reduction and Dereverberation”. *Proc. of European Signal Processing Conf. (EUSIPCO)*, Trieste, Italy, September 1996.

[Dev86] L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, 1986.

[DM80] S. B. Davis and P. Mermelstein. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, August 1980.

[Dur60] J. Durbin. “The Fitting of Time-Series Models”. *Revue de l’Institut International de Statistique / Review of the International Statistical Institute*, vol. 28, no. 3, pp. 233–244, 1960.

[Elk07] G. Elko. “Reducing Noise in Audio Systems”, Patent US7171008, 2007.

[EM84] Y. Ephraim and D. Malah. “Speech Enhancement Using a Minimum-mean Square Error Short-time Spectral Amplitude Estimator”. *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[ERHV10] T. Esch, M. Rüngeler, F. Heese, and P. Vary. “A Modified Minimum Statistics Algorithm for Reducing Time Varying Harmonic Noise”. *ITG-Fachtagung Sprachkommunikation*. VDE Verlag GmbH, October 2010.

[Esc12] T. Esch. *Model-Based Speech Enhancement Exploiting Temporal and Spectral Dependencies*. Dissertation, IND, RWTH Aachen, April 2012.

[ETS04] ETSI EG 201 377-2. “Speech Processing, Transmission and Quality Aspects (STQ); Specification and Measurement of Speech Transmission Quality; Part 2: Mouth-to-ear Speech Transmission Quality Including Terminals”, April 2004.

[ETS09] ETSI EG 202 396-1. “Speech and Multimedia Transmission Quality (STQ); Part 1: Background Noise Simulation Technique and Background Noise Database”, March 2009.

[FB10] S. Franz and J. Bitzer. “Multi-Channel Algorithms for Wind Noise Reduction and Signal Compensation in Binaural Hearing Aids”. *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, August 2010.

[FP90] J. M. Festen and R. Plomp. “Effects of Fluctuating Noise and Interfering Speech on the Speech-reception Threshold for Impaired and Normal Hearing”. *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, October 1990.

[FP03] D. A. Forsyth and J. Ponce. *Computer Vision - A Modern Approach*. Prentice-Hall, 2003.

[GB14] S. Gonzalez and M. Brookes. “Mask-based Enhancement for Very Low Quality Speech”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Florence, Italy, May 2014.

[GBS15] S. Godsill, H. Buchner, and J. Skoglund. “Detection and Suppression of Keyboard Transient Noise in Audio Streams with Auxiliary Keybed Microphone”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Brisbane, Australia, April 2015.

[Gei12] B. Geiser. *High-Definition Telephony over Heterogeneous Networks.* Dissertation, IND, RWTH Aachen, June 2012.

[Geo89] A. George. "Automobile Aeroacoustics". *American Institute of Aeronautics and Astronautics (AIAA) Journal*, vol. 1067, 1989.

[GH11] T. Gerkmann and R. Hendriks. "Noise Power Estimation Based on the Probability of Speech Presence". *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2011.

[GKR12] T. Gerkmann, M. Krawczyk, and R. Rehr. "Phase Estimation in Speech Enhancement; Unimportant, Important, or Impossible?". *Electrical Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pp. 1–5, November 2012.

[GM10] T. Gerkmann and R. Martin. "Cepstral Smoothing with Reduced Computational Complexity". *Proc. ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010.

[Han91] J. Hansen. "Speech Enhancement Employing Adaptive Boundary Detection and Morphological Based Spectral Constraints". *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Toronto, Canada, April 1991.

[Hay96] Haykin. *Adaptive Filter Theory.* Prentice Hall, 1996.

[HB04] Y. Huang and J. Benesty. *Audio Signal Processing for Next-Generation Multimedia Communication Systems.* Springer Science & Business Media, 2004.

[Hes83] W. Hess. *Pitch Determinaton of Speech Signals.* Springer Verlag, 1983.

[HHJ10] R. Hendriks, R. Heusdens, and J. Jensen. "MMSE Based Noise PSD Tracking with Low Complexity". *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas, USA, March 2010.

[HJN$^+$11] C. Herglotz, M. Jeub, C. Nelke, C. Beaugeant, and P. Vary. "Evaluation of Single- and Dual-Channel Noise Power Spectral Density Estimation Algorithms for Mobile Phones". *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Aachen, Germany, September 2011. ITG, DEGA.

[HL07] Y. Hu and P. C. Loizou. "A Comparative Intelligibility Study of Single-microphone Noise Reduction Algorithms". *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, September 2007.

[HNNV14] F. Heese, C. Nelke, M. Niermann, and P. Vary. "Selflearning Codebook Speech Enhancement". *ITG-Fachtagung Sprachkommunikation*, Erlangen, Germany, September 2014. VDE Verlag GmbH.

[HV15] F. Heese and P. Vary. “Noise PSD Estimation by Logarithmic Baseline Tracing”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Brisbane, Australia, April 2015. IEEE.

[HWB⁺12] C. Hofmann, T. Wolff, M. Buck, T. Haulick, and W. Kellermann. “A Morphological Approach to Single-Channel Wind-Noise Suppression”. *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, September 2012.

[IP99] ITU-P. “P.50: Artificial Voices”, September 1999.

[ISM08] B. Iser, G. Schmidt, and W. Minker. *Bandwidth Extension of Speech Signals*, vol. 13. Springer, 2008.

[IT01] ITU-T. “ITU-T Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs”, February 2001.

[IT07] ITU-T. “ITU-T Rec. P.862.2: Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs”, November 2007.

[Ita75] F. Itakura. “Line Spectrum Representation of Linear Predictor Coefficients of Speech Signals”. *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[Jax02] P. Jax. *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds.* Dissertation, IND, RWTH Aachen, October 2002.

[Jeu12] M. Jeub. *Joint Dereverberation and Noise Reduction for Binaural Hearing Aids and Mobile Phones.* PhD Thesis, IND, RWTH Aachen, August 2012.

[JHN⁺12] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary. “Noise Reduction for Dual-Microphone Mobile Phones Exploiting Power Level Differences”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Kyoto, Japan, March 2012.

[JL86] D. Jones and M. Lorenz. “An Application of a Markov Chain Noise Model to Wind Generator Simulation”. *Mathematics and Computers in Simulation*, vol. 28, no. 5, pp. 391–402, October 1986.

[JNBV11] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary. “Blind Estimation of the Coherent-to-Diffuse Energy Ratio From Noisy Speech Signals”. *Proc. of European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, August 2011.

[JNH⁺13] M. Jeub, C. Nelke, C. Herglotz, P. Vary, and C. Beaugeant. “Noise Reduction for Dual-Microphone Communication Devices”, Patent US 2013/0054231, 2013.

[JNK$^+$11] M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary. "Robust Dual-Channel Noise Power Spectral Density Estimation". *Proc. of European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, August 2011.

[JSK$^+$10] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary. "Do We Need Dereverberation for Hand-Held Telephony?". *International Congress on Acoustics (ICA)*, Sydney, Australia, August 2010. Australian Acoustical Society.

[Kab02] P. Kabal. "TSP Speech Database". Technical report, McGill University, Montreal, Canada, September 2002.

[Kat07] J. Kates. "Hearing Aid with Suppression of Wind Noise", Patent 2007/0030989, 2007.

[Kat08] J. M. Kates. *Digital Hearing Aids.* Plural Publishing, Inc, 2008.

[KC76] C. Knapp and G. Carter. "The Generalized Correlation Method for Estimation of Time Delay". *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, August 1976.

[KG12] M. Krawczyk and T. Gerkmann. "STFT Phase Improvement for Single Channel Speech Enhancement". *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, September 2012.

[KMT$^+$06] S. Kuroiwa, Y. Mori, S. Tsuge, M. Takashina, and F. Ren. "Wind Noise Reduction Method for Speech Recording Using Multiple Noise Templates and Observed Spectrum Fine Structure". *Intern. Conf. on Communication Technology*, Guilin, China, November 2006.

[Kut09] H. Kuttruff. *Room Acoustics.* Taylor & Francis, London, 2009.

[LBG80] Y. Linde, A. Buzo, and R. Gray. "An Algorithm for Vector Quantizer Design". *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, January 1980.

[Lev47] N. Levinson. "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction". *Journal of Mathematical Physics*, vol. 25, no. 4, pp. 261–278, January 1947.

[LH99] K. Linhard and T. Haulick. "Noise Subtraction with Parametric Recursive Gain Curves". *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, Budapest, Hungary, September 1999.

[Lig52] M. J. Lighthill. "On Sound Generated Aerodynamically. I General Theory". *Proceedings of the Royal Society*, vol. 211, no. 1107, pp. 564–587, March 1952.

[Lig54] M. J. Lighthill. "On Sound Generated Aerodynamically. II. Turbulence as a Source of Sound". *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 222, no. 1148, pp. 1–32, February 1954.

[LKS89] L. Lamel, R. Kassel, and S. Seneff. "Speech Database Development: Design and Analysis of the Acoustic-phonetic Corpus". *Speech Input/Output Assessment and Speech Databases*, 1989.

[LL00] I. Y. Lun and J. C. Lam. "A Study of Weibull Parameters Using Long-term Wind Observations". *Renewable Energy - An International Journal*, vol. 20, no. 2, pp. 145–153, June 2000.

[LO79] J. Lim and A. Oppenheim. "Enhancement and Bandwidth Compression of Noisy Speech". *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

[Loi13] P. Loizou. *Speech Enhancement: Theory and Practice.* CRC Press, 2013.

[Löl11] H. W. Löllmann. *Allpass-Based Analysis-Synthesis Filter-Banks: Design and Application.* Dissertation, IND, RWTH Aachen, 2011.

[Lot04] Lotter. *Single and Multimicrophone Speech Enhancement for Hearig Aids.* Dissertation, IND, RWTH Aachen, 2004.

[LVKL96] T. Laakso, V. Valimaki, M. Karjalainen, and U. Laine. "Splitting the Unit Delay [FIR/all-pass Filter Design]". *Signal Processing Magazine, IEEE*, vol. 13, no. 1, pp. 30–60, January 1996.

[Mar01] R. Martin. "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics". *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, July 2001.

[Mar05] R. Martin. "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors". *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 845–856, September 2005.

[MHA11] J. Marin-Hurtado and D. Anderson. "FFT-Based Block Processing in Speech Enhancement: Potential Artifacts and Solutions". *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 8, pp. 2527–2537, November 2011.

[ML13] N. Mohammadiha and A. Leijon. "Nonnegative HMM for Babble Noise Derived From Speech HMM: Application to Speech Enhancement". *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 998–1011, May 2013.

[MM80] R. McAulay and M. Malpass. "Speech Enhancement Using a Soft-decision Noise Suppression Filter". *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, April 1980.

[MM09] G. Müller and M. Möser. *Handbook of Engineering Acoustics.* Springer, 2009.

[MRG85] J. Makhoul, S. Roucos, and H. Gish. “Vector Quantization in Speech Coding”. *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1551–1588, November 1985.

[MS14] P. Mowlaee and R. Saeidi. “Time-frequency Constraints for Phase Estimation in Single-channel Speech Enhancement”. *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 337–341, Aachen, Germany, September 2014.

[NBV13] C. Nelke, C. Beaugeant, and P. Vary. “Dual Microphone Noise PSD Estimation for Mobile Phones in Hands-Free Position Exploiting the Coherence and Speech Presence Probability”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013.

[NCBV14] C. Nelke, N. Chatlani, C. Beaugeant, and P. Vary. “Single Microphone Wind Noise PSD Estimation Using Signal Centroids”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Florence, Italy, May 2014.

[NCBV15] C. Nelke, N. Chatlani, C. Beaugeant, and P. Vary. “Audio processing devices and audio processing methods”, Patent WO 2015/061116 A8, 2015.

[Nel09] C. M. Nelke. “Mehrkanalige Störgeräuschreduktion für Mobiltelefone”. Diploma thesis, IND, RWTH Aachen, Templergraben 55, 52056 Aachen, January 2009.

[NG10] P. A. Naylor and N. D. Gaubitch. *Speech dereverberation.* Springer Science & Business Media, 2010.

[NJV16] C. M. Nelke, P. Jax, and P. Vary. “Wind Noise Detection: Signal Processing Concepts for Speech Communication”. *Proc. of German Annual Conference on Acoustics (DAGA).* Deutsche Gesellschaft für Akustik (DEGA), March 2016.

[NLZIT10] E. Nemer, W. LeBlanc, M. Zad-Issa, and J. Thyssen. “Single-Microphone Wind Noise Suppression”, Patent 2010/00209, 2010.

[NNJ$^+$12] C. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary. “Single Microphone Wind Noise Reduction Using Techniques of Artificial Bandwidth Extension”. *Proc. of European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, August 2012.

[NNV15] C. Nelke, P. Naylor, and P. Vary. “Corpus Based Reconstruction of Speech Degraded by Wind Noise”. *Proc. of European Signal Processing Conf. (EUSIPCO)*, Nice, France, 2015.

[Nol70] A. Noll. “Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate”. *Proc. of the Symposium on Computer Processing in Communications*, vol. 14, pp. 779–797, 1970.

[NV14a] C. Nelke and P. Vary. “Dual Microphone Wind Noise Reduction by Exploiting the Complex Coherence”. *ITG-Fachtagung Sprachkommunikation*, Erlangen, Germany, September 2014.

[NV14b] C. Nelke and P. Vary. “Measurement, Analysis and Simulation of Wind Noise Signals for Mobile Communication Devices”. *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC), Download audio samples from: http://www.ind.rwth-aachen.de/en/research/tools-downloads/wind-noise-database/*, Sophia-Antipolis, France, September 2014.

[NV15] C. Nelke and P. Vary. “Wind Noise Short Term Power Spectrum Estimation Using Pitch Adaptive Inverse Binary Masks”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Brisbane, Australia, April 2015.

[OSB+89] A. Oppenheim, R. Schafer, J. Buck, et al. *Discrete-time Signal Processing*, vol. 2. Prentice-hall Englewood Cliffs, 1989.

[Pal98] K. Paliwal. “Spectral Subband Centroid Features for Speech Recognition”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Seattle, USA, May 1998.

[QB88] S. Quackenbush and T. Barnwell. *Objective Measures of Speech Quality*. Prentice-Hall, Inc., 1988.

[RBHH01] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. “Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs”. *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Salt Lake City, Utah, USA, May 2001.

[RFB81] F. Reed, P. Feintuch, and N. Bershad. “Time Delay Estimation Using the LMS Adaptive Filter–Static Behavior”. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 3, pp. 561–571, June 1981.

[RJ93] L. Rabiner and B.-H. Juang. “Fundamentals of Speech Recognition”. 1993.

[Ros10] T. Rosenkranz. “Cobebuch-basierte Geräuscheuschreduktion mit cepstraler Modellierung”. *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010.

[RS78] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[RV05] K. Rhebergen and N. Versfeld. "A Speech Intelligibility Index-based Approach to Predict the Speech Reception Threshold for Sentences in Fluctuating Noise for Normal-hearing Listeners". *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, April 2005.

[SF12] R. Scharrer and J. Fels. "Fuzzy Sound Field Classification in Devices with Multiple Acoustic Sensors". *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, September 2012.

[SL00] J. Seguro and T. Lambert. "Modern Estimation of the Parameters of the Weibull Wind Speed Distribution for Wind Energy Analysis". *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 85, no. 1, pp. 75–84, March 2000.

[SS01] A. Sahin and Z. Sen. "First-order Markov chain Approach to Wind Speed Modelling". *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 89, no. 3–4, pp. 263–269, March 2001.

[SSA07] A. Subramanya, M. Seltzer, and A. Acero. "Automatic Removal of Typed Keystrokes from Speech Signals". *Signal Processing Letters, IEEE*, vol. 14, no. 5, pp. 363–366, May 2007.

[SSK07] S. Srinivasan, J. Samuelsson, and W. Kleijn. "Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments". *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 441–452, February 2007.

[Str88] M. Strasberg. "Dimensional Analysis of Windscreen Noise". *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 544–548, February 1988.

[THHJ10] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen. "A Short-time Objective Intelligibility Measure for Time-frequency Weighted Noisy Speech". *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas, USA, March 2010.

[TTM$^+$11] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin. "An Evaluation of Noise Power Spectral Density Estimation Algorithms in Adverse Acoustic Environments". *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011.

[Var85] P. Vary. "Noise Suppression by Spectral Magnitude Estimation – Mechanism and Theoretical Limits". *Signal processing*, vol. 8, no. 4, pp. 387–400, July 1985.

[VM06] P. Vary and R. Martin. *Digital Speech Transmission. Enhancement, Coding and Error Concealment.* Wiley-VCH Verlag, 2006.

[Wei51] W. Weibull. “A Statistical Distribution Function of Wide Applicability”. *Journal of applied mechanics*, vol. 23, pp. 981–997, September 1951.

[Wie57] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications.* John Wiley & Sons, 1957.

[Wik06a] Wikipedia The Free Encyclopedia. “Microphone Boom with Large Fur Windshield”. Online picture by Galak76, November 2006. https://de.wikipedia.org/wiki/Datei:Mic_boom_with _windshield.jpg.

[Wik06b] Wikipedia The Free Encyclopedia. “Microphone Boom with Light Foam Windshield”. Online picture by Galak76, November 2006. https://de.wikipedia.org/wiki/Datei:Mic_boom_with_light _foam_windshield.jpg.

[WL82] D. Wang and J. Lim. “The Unimportance of Phase in Speech Enhancement”. *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, August 1982.

[WMG79] D. Wong, J. Markel, and J. Gray, A. “Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform”. *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, August 1979.

[Wut92] J. Wuttke. “Microphones and Wind”. *Journal of the Audio Engineering Society*, vol. 40, no. 10, pp. 809–817, October 1992.

[YR04] O. Yilmaz and S. Rickard. “Blind Separation of Speech Mixtures via Time-frequency Masking”. *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.