Post-Filter Optimization for Multichannel Automotive Speech Enhancement

Von der Fakultät für Elektrotechnik, Informationstechnik, Physik der Technischen Universität Carolo-Wilhelmina zu Braunschweig zur Erlangung der Würde eines Doktor-Ingenieurs (Dr.-Ing.) genehmigte

Dissertation

vorgelegt von

Huajun Yu, M.Sc.
aus Shanghai, China
Huajun Yu

Post-Filter Optimization for Multichannel Automotive Speech Enhancement
This dissertation was written during my time as a research assistant at the Institut für Nachrichtentechnik (IfN) of Technische Universität Braunschweig. The research results of this work would not have been possible without the support of several people. It is my pleasure to take this opportunity to thank them all.

First of all, I would like to express my deepest and sincere gratitude to my supervisor Prof. Dr.-Ing. Tim Fingscheidt for his continuous active support throughout my time under his supervision. He has definitely contributed the most to this thesis by providing me valuable suggestions, intensive discussions and precious feedbacks.

I am grateful to Prof. Dr.-Ing. Gerhard Schmidt for being the co-supervisor of this thesis and showing a lot of interest in my research work. I would also like to thank Prof. Dr.-Ing. Jörg Schöbel for being the chair of the examination board.

Furthermore, I would like to thank my colleagues for providing me a very enjoyable working environment. In particular, I am thankful to Dipl.-Ing. Patrick Bauer, Dipl.-Ing. Balázs Fodor, Dipl.-Ing. Florian Pflug, Marc-André Jung, M.Sc., Dr.-Ing. Volker Märgner, Dipl.-Inform. Werner Pantke, Dipl.-Ing. Simon Receveur, Dipl.-Ing. Martin Jacob, Dipl.-Inform. Frithjof Hummes and Dr.-Ing. Suhadi for their valuable discussions and proof-readings of parts of this thesis. I am also indebted to Andreas Gudat and Uwe Hellrung for supporting me with hardware construction issues and my student Lars Tebelmann for his contribution.

Finally, I would like to give my special thanks to my parents Ningsheng Yu and Xiaomei Fu, my wife Qing Yang, and my children Leyang Yu and Qingying Yu for their consistent love, understanding, patience, and moral support, especially during the phase of writing this dissertation.

Shanghai, November 2013

Huajun Yu
Abstract

In an automotive environment, quality of speech communication using a hands-free equipment is often deteriorated by interfering car noise. In order to preserve the speech signal without car noise, a multichannel speech enhancement system including a beamformer and a post-filter can be applied. Since employing a beamformer alone is insufficient to substantially reducing the level of car noise, a post-filter has to be applied to provide further noise reduction, especially at low frequencies. In this thesis, two novel post-filter designs along with their optimization for different driving conditions are presented. The first post-filter design utilizes an adaptive smoothing factor for the power spectral density estimation as well as a hybrid noise coherence function. The hybrid noise coherence function is a mixture of the diffuse and the measured noise coherence functions for a specific driving condition. The second post-filter design applies a new multichannel decision-directed a priori SNR estimator based on both temporal and spatial smoothing. For different driving conditions, both post-filters are instrumentally optimized: For the first post-filter, the optimal adaptive smoothing factor and the optimal hybrid noise coherence function are obtained. For the second post-filter, the weighting factors of the temporal and spatial smoothing parts are optimized. Compared to state-of-the-art post-filters, both post-filter designs employing the optimized parameters improve the overall noise reduction performance significantly for different driving conditions.

Generally, manually finding the optimal parameterization of a noise reduction algorithm is a time-consuming task. In this thesis, the two new post-filter designs are thus instrumentally optimized by using a figure of merit (FoM). We define the FoM as an entity, which comprises three independent instrumental measures for the speech component quality, the level of noise attenuation, and the amount of musical tones. Particularly, a new weighted log kurtosis ratio measure is proposed for instrumental musical tones assessment in a black-box test manner, which does not mandate any knowledge of internal variables of the noise reduction algorithm under test and can be applied to a wide range of noise reduction algorithms. Subjective listening tests reveal that the weighted log kurtosis ratio measurements can provide a high correlation to the perceived amount of musical tones. In addition, a single-channel application example of jointly optimizing the smoothing factor and the a priori SNR floor of the decision-directed a priori SNR estimation is shown using an FoM. For some noise reduction algorithms, yet unknown optimal values of the parameters of interest are identified by applying the FoM-based instrumental optimization method and subjectively verified.
Kurzfassung


Contents

Acknowledgments i

Abstract iii

Kurzfassung v

1 Introduction 1
   1.1 Speech Acquisition In a Car ............................................. 1
   1.2 Outline of the Thesis ..................................................... 4

2 Basic Principles of Beamforming and Post-Filtering 7
   2.1 Spatial Characteristics of Noise Fields ............................... 7
   2.2 Multichannel MMSE Estimation .......................................... 13
       2.2.1 General Approach .................................................... 13
       2.2.2 Decomposition ....................................................... 15
   2.3 Beamforming ............................................................... 19
       2.3.1 Microphone Array Design and Performance Measures ............... 19
       2.3.2 Linearly Constrained Minimum Variance (LCMV) Beamforming .. 23
       2.3.3 Minimum Variance Distortionless Response (MVDR) Beamforming 23
       2.3.4 Generalized Sidelobe Canceller (GSC) Beamforming ............ 28
       2.3.5 Other Beamforming Techniques ..................................... 30
   2.4 Post-Filtering ............................................................. 30
       2.4.1 Designs for an Incoherent Noise Field ............................ 31
       2.4.2 General Designs ...................................................... 33
   2.5 Summary ................................................................. 36

3 Instrumental Evaluation Framework 37
   3.1 Processing Framework ..................................................... 37
       3.1.1 Single-Channel Setup ............................................... 38
       3.1.2 Multichannel Setup .................................................. 39
   3.2 Speech-Related Instrumental Quality Measures ....................... 40
       3.2.1 Perceptual Evaluation of Speech Quality (PESQ) ................ 41
       3.2.2 Segmental Speech-to-Speech Distortion Ratio (SSDR) .......... 42
       3.2.3 Speech Mean Opinion Score (S-MOS) ............................. 43
   3.3 Noise-Related Instrumental Quality Measures ....................... 45
       3.3.1 Noise Mean Opinion Score (N-MOS) .............................. 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.2</td>
<td>Signal-to-Noise Ratio Improvement</td>
<td>46</td>
</tr>
<tr>
<td>3.3.3</td>
<td>White-Box Log Kurtosis Ratio</td>
<td>47</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Black-Box Log Kurtosis Ratio</td>
<td>49</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Black-Box Weighted Log Kurtosis Ratio</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>Instrumental Optimization Using a Figure of Merit (FoM)</td>
<td>65</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Independency of Instrumental Quality Measures</td>
<td>65</td>
</tr>
<tr>
<td>3.4.2</td>
<td>FoM-Based Instrumental Optimization Method</td>
<td>66</td>
</tr>
<tr>
<td>3.5</td>
<td>FoM-Based Single-Channel Optimization of an A Priori SNR Estimation</td>
<td>68</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Approach</td>
<td>68</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Performance Analysis</td>
<td>69</td>
</tr>
<tr>
<td>3.6</td>
<td>Summary</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>Multichannel In-Car Speech, Noise, and CAN Bus Database</td>
<td>79</td>
</tr>
<tr>
<td>4.1</td>
<td>Application Scenario</td>
<td>79</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Collection</td>
<td>81</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Speech Signals</td>
<td>81</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Noise and CAN Bus Signals</td>
<td>82</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Post-Processing</td>
<td>83</td>
</tr>
<tr>
<td>4.3</td>
<td>Applied Data Set</td>
<td>84</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
<td>84</td>
</tr>
<tr>
<td>5</td>
<td>Post-Filter Design with Modified PSD Estimation and Hybrid Noise</td>
<td>85</td>
</tr>
<tr>
<td>5.1</td>
<td>Baselines</td>
<td>85</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Approaches</td>
<td>85</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Performance Analysis</td>
<td>86</td>
</tr>
<tr>
<td>5.2</td>
<td>PSD Estimation with an Adaptive Smoothing Factor</td>
<td>88</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Approach</td>
<td>88</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Performance Analysis</td>
<td>89</td>
</tr>
<tr>
<td>5.3</td>
<td>Hybrid Noise Coherence Functions</td>
<td>91</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Noise Coherence Function Measurement</td>
<td>91</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Approach</td>
<td>94</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Performance Analysis</td>
<td>96</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>Post-Filter Design with Multichannel Decision-Directed Approach</td>
<td>103</td>
</tr>
<tr>
<td>6.1</td>
<td>Single-Channel Baseline Approach by Temporal Smoothing</td>
<td>103</td>
</tr>
<tr>
<td>6.2</td>
<td>Multichannel Approaches</td>
<td>105</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Baseline Approach by Spatial Smoothing</td>
<td>105</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Approach by Temporal and Spatial Smoothing</td>
<td>106</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Performance Analysis</td>
<td>107</td>
</tr>
<tr>
<td>6.3</td>
<td>Summary</td>
<td>111</td>
</tr>
<tr>
<td>7</td>
<td>Conclusions</td>
<td>113</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Current in-car infotainment systems include several features like mobile communication, navigation, and multimedia entertainment. For all these features, an efficient and safe interaction between the driver and the infotainment system is required. In general, speech communication can be treated as the most natural and efficient way of human-to-human as well as human-to-machine communication. Thanks to the hands-free equipment, an easy-to-use and safe voice-controlled infotainment system can be realized in a car. However, due to interfering signals, especially car noise, a speech enhancement system has to be applied to the hands-free equipment for acquiring clean speech without car noise. Recently, multichannel speech enhancement systems consisting of beamforming and post-filtering have been shown to be capable of providing noise attenuation with good speech quality. However, in an automotive environment with different driving conditions, state-of-the-art approaches to the post-filter estimation just utilize the same \textit{a priori} knowledge throughout. This thesis deals with multichannel automotive speech enhancement by focusing on new post-filter designs, which can be optimized for different driving conditions.

1.1 Speech Acquisition In a Car

In the context of acquiring clean speech in a car, an in-car integrated hands-free equipment is much more favored by drivers than using a separate mobile headset owing to comfortability and safety (Oh et al., 1992). One of the key functionalities of the hands-free equipment is speech enhancement\(^1\), since interfering background noise in a car can dramatically degrade the perceived speech quality. Noise reduction algorithms shall, therefore, provide sufficient noise attenuation and keep the speech signal undistorted at the same time. In the literature, noise reduction algorithms can typically be categorized into single-channel and multichannel approaches.

Employing a single-channel noise reduction algorithm in a car, one microphone is typically being positioned in the rear mirror, in the sunlight visor, or in the overhead light module. These positions are typically preferred, since they provide a relatively short

\(^1\)In the following, the terms speech enhancement and noise reduction are used interchangeably.
distance to the driver and have a direct acoustical path between the microphone and the

Single-channel noise reduction has been intensively researched for several decades. It
usually applies a weighting rule in the frequency domain to extract the clean speech from
the noisy speech signal. Starting with pioneer works using a spectral subtraction approach
(Boll, 1979, Lim and Oppenheim, 1979), state-of-the-art weighting rules are generally es-
estimated by employing different statistical models of the clean speech and noise probability
density functions (PDFs), as well as different cost functions (Ephraim and Malah, 1984,
addition, Fingscheidt et al. (2008) proposed a data-driven approach, which can be utilized
to train the weighting rule without using any statistical models of the clean speech and
noise PDFs.

In the field of multichannel noise reduction, a microphone array consisting of two or
more omnidirectional microphones has gained a lot of research interest in recent decades
with early works of Flanagan et al. (1985) for the pioneer design of microphone array
in large rooms, Cox et al. (1987) for analyzing adaptive beamforming algorithms, and
Kellermann (1991) for implementing a digital microphone array instead of early analog
concepts at that time. Based on the microphone array, a beamforming algorithm can
be realized exploiting not only temporal and spectral information, but also spatial in-
formation of the underlying noise field. Utilizing the additional spatial information, the
beamformer can be spatially directed to the source of the clean speech and thus cancels
noise signals from all other directions. Beamforming algorithms are mostly realized in the
frequency domain. For automotive applications, microphone arrays are commonly posi-
tioned in the rear mirror (Wolff and Buck, 2010), in the sunlight visor (Dahl and Claesson,
1999, Zhang and Hansen, 2003), or in the overhead light module (Buck et al., 2009). In
this thesis, a new head unit-integrated microphone array is proposed, which provides
an integrated solution without extra cabling. In an automotive environment, car noise
power spectral density (PSD) has a low-pass characteristic, which means that most of
its energies are concentrated at low frequencies. However, beamforming algorithms alone
yield insufficient noise attenuation at low frequencies (Simmer et al., 2001). In order to
solve this problem, a post-filter was firstly proposed by Zelinski (1988), which can be
applied after the beamformer showing an improved noise attenuation performance. Fur-
thermore, it has been shown that such a beamformer and post-filter structure is equiva-
lent to the multichannel minimum mean square error (MMSE) estimator (Simmer et al.,
2001). Zelinski’s post-filter was further improved by Simmer and Wasiljeff (1992) exhib-
ting a less distorted speech signal. Both Zelinski’s and Simmer’s post-filters were specif-
ically designed for an incoherent noise field. However, Meyer and Simmer (1997) have
shown that the automotive noise field can be better approximated by a diffuse noise field.
Therefore, a more generalized post-filter design based on the noise coherence function
was proposed by McCowan and Bourlard (2003), who applied the diffuse noise coherence
function to their post-filter estimation. McCowan’s post-filter was further slightly modi-
ﬁed by Lefkimmiatis and Maragos (2006). Applying the diffuse noise coherence function,
a significant improvement in terms of noise attenuation can be achieved by McCowan’s
and Lefkimmiatis’s post-filters compared to Zelinski’s and Simmer’s post-filters. However, McCowan’s and Lefkimmiatis’s post-filters still suffer from noise distortion in terms of musical tones.

Concerning the post-filter estimation, the clean speech and noise PSD estimations and the noise coherence function can be regarded as the two most important components. Current approaches employ a fixed smoothing factor to the recursive clean speech and noise PSD estimations (Zelinski, 1988, Simmer and Wasiljeff, 1992, McCowan and Bourlard, 2003, Lefkimmiatis and Maragos, 2006), which faces a compromise between speech distortion and noise distortion in terms of musical tones. For the noise coherence function, the diffuse assumption can be applied as a good model for car noise (Meyer and Simmer, 1997). However, it is unknown to what extent it is actually valid for different driving conditions like different driving speeds, air-conditioning operation levels, or windows being open or closed. Such car-specific driving information can be provided by using controller area network bus (CAN bus) data of modern cars in real-time. In this thesis, a modified PSD estimation utilizing an adaptive smoothing factor and a new hybrid noise coherence function are proposed to formulate a new post-filter design. The hybrid noise coherence function comprises the diffuse and the measured noise coherence functions for a specific driving condition. In addition, another new post-filter design is also presented in this thesis, which applies a novel multichannel decision-directed (DD) *a priori* SNR estimator exploiting not only temporal smoothing as proposed by Ephraim and Malah (1984) in their single-channel DD approach, but also spatial smoothing utilizing the multichannel inputs.

For each single-channel or multichannel noise reduction algorithm, an instrumental quality assessment is chosen to evaluate whether the targeted quality requirements have been achieved. Instrumental evaluation of speech enhancement algorithms mainly includes speech- and noise-related instrumental quality measures. Until now, several quality aspects in both fields can be well evaluated in an instrumental manner (see ITU-T Rec. P.1100 and ITU-T Rec. P.1110, ITU-T, 2008, 2009). However, in the context of noise distortion, especially for musical tones, in spite of several works on the analysis of musical tones (Cappé, 1994, Breithaupt and Martin, 2011), there are less publications on instrumentally assessing musical tones. Recently, an instrumental musical tones measure exploiting the log kurtosis ratio was presented by Uemura et al. (2008, 2009) and Inoue et al. (2011) showing a high correlation to the perceived amount of musical tones. Nevertheless, their instrumental musical tones measure belongs to a *white-box* measure mandating knowledge of internal variables of the noise reduction algorithm under test. Furthermore, it can only be applied to a specific noise reduction algorithm, namely the spectral subtraction approach. To cope with this problem, a *black-box* instrumental musical tones measure based on the weighted log kurtosis ratio is proposed in this thesis, which does not require any knowledge of internal variables of the noise reduction algorithm under test and can be applied to a wide range of noise reduction algorithms.

When developing noise reduction algorithms, an overall quality trade-off usually exists between the speech quality, noise attenuation, and noise distortion in terms of musical tones of the enhanced speech signal. In order to achieve the best overall noise reduction
performance of a noise reduction algorithm, it is typically necessary to manually tune different parameters to their optimal values. Such an optimization task, however, turns out to be as a time-consuming task. In this thesis, an instrumental entity defined as a figure of merit (FoM), which consists of three independent instrumental quality measures typically applied to evaluate noise reduction algorithms, is proposed. The optimal parameters of the noise reduction algorithm can then be automatically identified by using the FoM. Utilizing the FoM-based instrumental optimization method, the two newly proposed post-filter designs can thus be optimized for different driving conditions. The optimized post-filter parameters hold the potential of being saved and later being selected by the CAN bus data for the corresponding driving condition in a real-time manner.

1.2 Outline of the Thesis

This thesis can be divided into two main parts: The first part concentrates on a new black-box instrumental musical tones measure. Based on this instrumental musical tones measure and along with two other instrumental quality measures for the speech component quality and the level of noise attenuation, an FoM-based instrumental optimization method is elaborated. The second part addresses the optimization of two novel post-filter designs. The post-filter optimization for different driving conditions is carried out by using the FoM. The remainder of this thesis is structured as follows:

Chapter 2 begins with a theoretical description of spatial characteristics for different noise fields using noise coherence functions. The decomposition of the multichannel MMSE estimator into a beamformer and a post-filter follows then subsequently. Based on a clear classification by using design features, state-of-the-art beamforming and post-filtering algorithms are briefly outlined.

Chapter 3 shows a unified instrumental quality assessment framework in its single-channel and multichannel setups. Based on this framework, several state-of-the-art speech- and noise-related instrumental quality measures are introduced. Furthermore, a new black-box instrumental musical tones measure using the weighted log kurtosis ratio, in comparison to the established white-box one, is presented and analyzed. The FoM-based instrumental optimization method is then defined by means of three independent instrumental quality measures for the speech component quality, the level of noise attenuation, and the amount of musical tones. Furthermore, an application example of jointly optimizing two key parameters of the single-channel decision-directed a priori SNR estimation by using an FoM is shown for four noise reduction algorithms.

Chapter 4 is dedicated to the description of a multichannel in-car speech, noise, and CAN bus database, which was acquired for this thesis. Furthermore, the applied data set selected from the multichannel database, which is employed for the post-filter optimization in Chapters 5 and 6, is also described.

Chapter 5 presents a new post-filter design using a modified PSD estimation and a hybrid noise coherence function. Firstly, the performance of four state-of-the-art post-
filters are evaluated in order to choose the appropriate baselines for Chapters 5 and 6. An adaptive smoothing factor for the PSD estimation is then described. The noise coherence function measurement applying the multichannel in-car speech and noise database provides a new insight into the general diffuse noise field assumption for car noise. Based on this observation, a new hybrid noise coherence function consisting of the diffuse and the measured noise coherence functions is proposed. Furthermore, for different driving conditions, the FoM-based instrumental optimization of the new post-filter design employing the adaptive smoothing factor for the PSD estimation and the hybrid noise coherence function is shown in detail.

Chapter 6 deals with another new post-filter design employing a multichannel decision-directed (DD) a priori SNR estimator. By exploiting the multichannel inputs, a multichannel a priori SNR estimation based on spatial smoothing is formulated. Having this multichannel a priori SNR estimation and the single-channel DD a priori SNR estimation using temporal smoothing at our disposal, a new multichannel DD a priori SNR estimation based on both temporal and spatial smoothing is then proposed. Applying the multichannel DD a priori SNR estimate to design a new post-filter, the weighting factors of the multichannel DD approach are then shown to be optimized by using the FoM-based instrumental optimization method.

Chapter 7 provides concluding remarks of this thesis.
1. Introduction
Chapter 2

Basic Principles of Beamforming and Post-Filtering

For multichannel noise reduction, there are several algorithms of beamforming and post-filtering in the literature. In order to give a structured overview of these algorithms, spatio-temporal characteristics of four common noise fields are described in Section 2.1, which are essential to develop different beamformers and post-filters. In Section 2.2, the multichannel MMSE estimation is derived along with its decomposition into a beamformer and a post-filter. Consequently, based on a classification using distinguishing design features, different beamforming and post-filtering algorithms are briefly described in Sections 2.3 and 2.4, respectively.

2.1 Spatial Characteristics of Noise Fields

According to the theory of room acoustics, the sound pressure of a wave propagating in an anechoic free wave field can be described by $p(p, t)$ (unit [Pa]) with $p$ being the position vector in the Euclidean space and $t$ being the continuous time, respectively (Kuttruff, 2007). The propagation of a sound wave, i.e., an acoustic signal, can be defined in a homogeneous, lossless and dispersion-free medium using the wave equation (Kuttruff, 2007)

$$\nabla^2 p(p, t) = \frac{1}{c^2} \frac{\partial^2 p(p, t)}{\partial t^2}, \quad (2.1.1)$$

where $\nabla^2$ is the Laplace operator producing the divergence of the gradient of a function, $\partial$ is the partial differentiation operator, and $c$ is the speed of sound, respectively. It can be seen in (2.1.1) that the sound wave described by $p(p, t)$ belongs to a so-called

$^2$Homogeneity means that the speed of propagation keeps constant. Lossless assures that the attenuation of the wave amplitude does not depend on the medium. A medium is called dispersion-free, if the medium is linear and the phase velocity of a propagating wave in such a medium is independent of its frequencies.

$^3$Under a normal room temperature of 20°C, we obtain $c = 344 \text{ m/s}$. This value is applied throughout this thesis.
spatio-temporal signal. In order to solve (2.1.1) w.r.t. \( p(p, t) \), a distinction between a plane wave and a spherical wave has to be made.

Assuming a plane wave, the term \( p(p, t) \) in the wave equation (2.1.1) can then be solved as (Kuttruff, 2007)

\[
p(p, t) = P \exp \left( j(\omega_w t - k^T p) \right),
\]

where \( P \) [Pa] represents the amplitude of the plane wave and \((\cdot)^T\) denotes the transpose operator, respectively. The term \( \omega_w = 2\pi f_w \) is the angular frequency of the wave with \( f_w \) being the wave frequency. The wavenumber vector \( \mathbf{k} \) is defined as (Kuttruff, 2007)

\[
\mathbf{k} = (2\pi/\lambda_w)\mathbf{u},
\]

with \( \lambda_w \) being the wavelength and \( \mathbf{u} \) being the unit vector (\(|\mathbf{u}| = 1\) with \(|\cdot|\) being the magnitude operator). The wavenumber vector \( \mathbf{k} \) thus governs the direction of the plane wave propagation, which is defined by the direction of \( \mathbf{u} \). Furthermore, applying \( \lambda_w = c/f_w \) and \( \omega_w = 2\pi f_w \) in (2.1.3), the magnitude of the wavenumber vector \( |\mathbf{k}| \) can be derived as \(|\mathbf{k}| = \omega_w/c\). The solution of (2.1.2) is called a plane wave, since all points with the same value of \( p(p, t) \) are in the same plane defined by the scalar product \( k^T p \) (Kuttruff, 2007). The plane wave model is assumed when the distance \( |p| \) of the point \( p \) to the sound source is much bigger than the wavelength \( \lambda_w \). A wave field fulfilling this condition is defined as far field (Kuttruff, 2007). In the automotive environment, the distance between the driver’s mouth and the hands-free equipment usually satisfies the far field assumption. Therefore, throughout this thesis, the far-field model is applied.

Under the assumption of a spherical wave, \( p(p, t) \) in the wave equation (2.1.1) can be solved as (Kuttruff, 2007)

\[
p(p, t) = \frac{P}{|p|} \exp \left( j(\omega_w t - |\mathbf{k}||p|) \right),
\]

where \(|p|\) represents the distance of the point \( p \) to the source, which is located in the center of the sphere. A spherical wave propagates then in the radial direction from the source. Compared to the plane wave solution given in (2.1.2), the amplitude of a spherical wave \( P/|p| \) depends on its distance \(|p|\) from the source and decreases with increasing distance. The spherical wave model is only applied when the point of observation is located quite close to the source, which is referred to as near field (Kuttruff, 2007).

In a real noise field, e.g., a car noise environment, which has more than one noise source and is usually not anechoic but having direct and reflected sound waves, the sound wave propagation can be deterministically modeled by a room impulse response. The room impulse response is a function, which depends on the position \( p \) and the time instant \( t \). However, a real noise field is usually too complex to be modeled analytically using the room impulse response. Therefore, a statistical description of the noise field by using a spatio-temporal cross-correlation function between two different random points inside the noise field is often more preferred (Cron and Sherman, 1962). The sound wave \( p(p, n) \) [Pa] at a certain point \( p \) in the noise field can thus be defined as a random process
with \( n \) being the discrete time index using a sampling frequency \( f_s \). The spatio-temporal relationship between two points \( i \) and \( j \) in space and time can be described through a cross-correlation function as (Eckart, 1953)

\[
\varphi(p_i, p_j, n_i, n_j) = \mathbb{E}\{p(p_i, n_i) \cdot p(p_j, n_j)\},
\]

where \( \mathbb{E}\{\cdot\} \) is the expectation operator. It is commonly assumed that the noise field is spatially homogeneous and temporally stationary meaning that the spatio-temporal correlation of two points inside the noise field depends only on their relative difference \( d = p_j - p_i \) and \( n = n_j - n_i \), respectively, not on their absolute positions and time instances (Eckart, 1953). Consequently, (2.1.5) can then be written as

\[
\varphi(p_i, p_j, n_i, n_j) = \varphi_{i,j}(d, n) = \mathbb{E}\{p(p_i, n_i) \cdot p(p_i + d, n_i + n)\}.
\]

The noise field can further be assumed to be isotropic, when physical properties of the noise field are identical regardless of directions (Eckart, 1953). Based on the property of isotropy, the cross-correlation function \( \varphi_{i,j}(d, n) \) defined in (2.1.6) depends only on the Euclidean distance \( d = |d| \) as

\[
\varphi_{i,j}(d, n) = \mathbb{E}\{p(p_i, n_i) \cdot p(p_i + d, n_i + n)\}.
\]

In addition, the auto-correlation of one point in time can be defined as

\[
\varphi_{i,i}(n) = \mathbb{E}\{p(p_i, n_i) \cdot p(p_i + n)\}.
\]

Applying the discrete Fourier transform (DFT), the auto-power spectral density (PSD) \( \Phi_{i,i}(k) \) and the cross-PSD \( \Phi_{i,j}(d, k) \) can then be calculated as (Oppenheim et al., 1999)

\[
\Phi_{i,i}(k) = \sum_{n=0}^{K-1} \varphi_{i,i}(n)e^{-j\Omega n}, \quad \Phi_{i,j}(d, k) = \sum_{n=0}^{K-1} \varphi_{i,j}(d, n)e^{-j\Omega n},
\]

where \( \Omega = 2\pi \frac{k}{K} \) is the normalized angular frequency, \( k = 0, 1, ..., K - 1 \) is the frequency bin, and \( K \) is the DFT length, respectively. Based on \( \Phi_{i,i}(k) \) and \( \Phi_{i,j}(d, k) \), we can now define the complex-valued noise coherence function as (Trees, 2002)

\[
\Gamma_{i,j}(d, k) = \frac{\Phi_{i,j}(d, k)}{\sqrt{\Phi_{i,i}(k)\Phi_{j,j}(k)}} \in \mathbb{C},
\]

which describes the spatial and spectral relationship between two signals received by two different points inside an isotropic noise field. The term \( \mathbb{C} \) represents the set of complex numbers. In the following, we simply write \( \Phi_{i,j}(k) = \Phi_{i,j}(d, k) \) and \( \Gamma_{i,j}(k) = \Gamma_{i,j}(d, k) \). Alternatively, the noise coherence function \( \Gamma_{i,j}(k) \) can be also interpreted as the normalized cross-PSD through \( \Phi_{i,i}(k) \) and \( \Phi_{j,j}(k) \). As a result, \( \Gamma_{i,j}(k) \) owns the property of \(-1 \leq |\Gamma_{i,j}(k)| \leq 1\). Regarding a multichannel speech enhancement algorithm utilizing a microphone array with \( M \) microphones, the noise coherence matrix \( \Gamma_{NN}(k) \) can then be defined as

\[
\Gamma_{NN}(k) = \begin{pmatrix}
\Gamma_{1,1}(k) & \Gamma_{1,2}(k) & \cdots & \Gamma_{1,M}(k) \\
\Gamma_{2,1}(k) & \Gamma_{2,2}(k) & \cdots & \Gamma_{2,M}(k) \\
\vdots & \vdots & \ddots & \vdots \\
\Gamma_{M,1}(k) & \Gamma_{M,2}(k) & \cdots & \Gamma_{M,M}(k)
\end{pmatrix}.
\]
In this thesis, the far-field model is applied. In the following, four commonly applied noise fields are described utilizing the noise coherence function $\Gamma_{i,j}(k)$ (2.1.10).

**Incoherent Noise Field**

An incoherent noise field exists only when the noise signals captured by different microphones in a noise field are uncorrelated, which results into (Bitzer, 2001)

$$\Gamma_{i,j}^{\text{inc}}(k) = 0, \quad \forall i \neq j.$$  \hfill (2.1.12)

Applying (2.1.12) to the general noise coherence matrix $\mathbf{\Gamma}_{NN}(k)$ in (2.1.11), the incoherent noise coherence matrix $\mathbf{\Gamma}_{NN}^{\text{inc}}(k)$ turns out to be an identity matrix

$$\mathbf{\Gamma}_{NN}^{\text{inc}}(k) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{I}.$$  \hfill (2.1.13)

For microphone arrays with a small distance of adjacent microphones, an incoherent noise field can hardly be assumed. There always exists some extent of correlation between the microphone signals. However, the incoherent noise field can still be useful for modeling the electric self-noise of microphones with different manufacturing tolerances (Bitzer and Simmer, 2001).

**Coherent Noise Field**

In a coherent noise field, the noise signal propagates in a free wave field without any reflection. Assuming a point-shaped source, the plane wave of the source reaches two microphones $i$ and $j$ with a certain angle of arrival $\theta_o$. This scenario is depicted in Fig. 2.1. The time delay between two microphones $i$ and $j$ can thus be calculated as

$$\tau_{i,j} = \frac{d_{i,j} \cos \theta_o}{c},$$  \hfill (2.1.14)

with $d_{i,j}$ being the distance between the two microphones. Using $\tau_{i,j}$, the according phase shift can then be computed as (Proakis and Dimitris, 2007)

$$\Delta \Omega = \Omega \cdot f_s \cdot \tau_{i,j} = 2\pi f_s \cdot \tau_{i,j} \cdot \frac{k}{K} = \frac{2\pi k \cdot f_s \cdot d_{i,j} \cos \theta_o}{c \cdot K}.$$  \hfill (2.1.15)

Consequently, the cross-PSD $\Phi_{i,j}(k)$ can be derived from the source auto-PSD $\Phi_o(k)$ using the phase shift $\Delta \Omega$ as (Bitzer, 2001)

$$\Phi_{i,j}(k) = \Phi_o(k) \exp \left( -j \frac{2\pi k f_s d_{i,j} \cos \theta_o}{c K} \right).$$  \hfill (2.1.16)

Assuming the noise field to be homogeneous, we can thus obtain $\Phi_{i,i}(k) = \Phi_{j,j}(k) = \Phi_o(k)$ (Habets et al., 2008). Inserting $\Phi_{i,j}(k)$ into (2.1.10) and let it be normalized by $\Phi_{i,i}(k)$ and $\Phi_{j,j}(k)$, we obtain the coherence function (Bitzer, 2001)

$$\Gamma_{i,j}^{\text{coh}}(k) = \exp \left( -j \frac{2\pi k f_s d_{i,j} \cos \theta_o}{c K} \right) \in \mathbb{C}.$$  \hfill (2.1.17)
2.1. Spatial Characteristics of Noise Fields

It turns out that $\Gamma_{i,j}^{\text{coh}}(k)$ of the coherent noise field actually exhibits the phase shift of microphones in the frequency domain and accordingly the time delay in the time domain. In practice, the coherent noise field can be applied to direct noise signal without any reflection, e.g., in an echo-free environment or an open air environment. The noise signals of different microphones are thus highly correlated in a coherent noise field.

Until now, a coherent noise field with just one point-shaped source has been explained. For more complicated scenarios with more than one point-shaped sources, it is possible to apply the superposition of coherent noise fields (Habets and Gannot, 2007). In the following, the superposition of coherent noise fields being generated by uncorrelated point-shaped sources will be explained in a spherical (3-dimensional) geometry and a cylindrical (2-dimensional) geometry.

Spherically Isotropic Noise Field

We assume that there is an infinite number of uncorrelated point-shaped sources being uniformly distributed on a spherical surface, which is shown in Fig. 2.2. Under the far-field assumption, the radius of the sphere $r$ owns the property of $r \gg d_{i,j}$. Accordingly, an integration of $\Gamma_{i,j}^{\text{coh}}(k)$ (2.1.17) of the infinite superposed coherence noise fields over $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$ yields (Cook et al., 1955)

$$
\Gamma_{i,j}^{\text{dif}}(k) = \frac{1}{4\pi} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} \exp\left(-j \frac{2\pi k f_s d_{i,j} \cos \theta}{cK}\right) \sin \theta \, d\phi \, d\theta
$$

$$
= \frac{\sin\left(\frac{2\pi k f_s d_{i,j}}{cK}\right)}{2\pi k f_s d_{i,j}}
$$

$$
= \text{sinc}\left(\frac{2\pi k f_s d_{i,j}}{cK}\right) \in \mathbb{R},
$$

which is the coherence function of a spherically isotropic noise field. The spherically isotropic noise field is also called as the diffuse noise field (Cook et al., 1955). The co-

Figure 2.1: Two microphones in a coherent noise field shown in a Cartesian coordinate system
herence function of a diffuse noise field thus owns the property of a sinc function with \( \text{sinc}(x) = \frac{\sin(x)}{x}, \ x \in \mathbb{R} \). Compared to the coherence function of a coherent noise field, the diffuse coherence function is naturally real-valued and it is controlled by the distance between two microphones. In practice, the diffuse noise field can be treated as a good approximation model for car noise and office noise (Meyer and Simmer, 1997, McCowan and Bourlard, 2003). If the windows are closed, and air-condition is switched off, we can easily image an infinite number of (uncorrelated) noise sources distributed over the entire car interior and windows.

\[ \text{Cylindrically Isotropically Noise Field} \]

Like the spherically isotropic noise field, a cylindrically isotropic noise field can be modeled with an infinite number of uncorrelated point-shaped sources. However, the point-shaped sources are uniformly distributed on a ring with the radius \( r \), which is depicted in Fig. 2.3. Please note that the contribution of all rings of the cylinder along the z axis is the same (Habets and Gannot, 2007). Considering the far field assumption with \( r \gg d_{i,j} \), the coherence function of a cylindrically isotropic noise field can then be computed by the integration of \( \Gamma^{\text{coh}}_{i,j}(k) \) (2.1.17) over \( \phi \in [0, 2\pi] \) as (Jacobsen, 1962)

\[
\Gamma^{\text{cyl}}_{i,j}(k) = \frac{1}{2\pi} \int_{\phi=0}^{2\pi} \exp \left( -j \frac{2\pi k f s d_{i,j} \cos \phi}{cK} \right) d\phi = J_0 \left( \frac{2\pi k f s d_{i,j}}{cK} \right) \in \mathbb{R} ,
\]

where \( J_0(\cdot) \) is the zero\(^{\text{th}}\)-order Bessel function of the first kind. The cylindrically isotropic noise field provides also a real-valued noise coherence function. It can be applied to cocktail-party noise, where several speakers are talking at the same time within a room having good absorption in the floor and ceiling (Dörbecker, 1998).
2.2 Multichannel MMSE Estimation

For multichannel automotive speech enhancement, an optimal multichannel filter in the sense of a minimum mean square error (MMSE) criterion has attracted great interest in the recent years. In the literature, the multichannel Wiener filter can be derived according to the multichannel MMSE estimation. In the following, the multichannel Wiener filter will be derived in the frequency domain. Furthermore, the decomposition of the multichannel Wiener filter into a beamformer and a post-filter is presented.

![Multichannel Wiener Filter Diagram](image_url)

**Figure 2.4:** Block diagram of the multichannel Wiener filter

2.2.1 General Approach

Let's regard a multichannel system consisting of $M$ input microphone signals as shown in Fig. 2.4, where each microphone signal can be expressed as $y_i(n) = s_i'(n) + n_i'(n)$ with $i = 1, \ldots, M$. The term $s_i'(n)$ is the sampled delayed clean speech signal $s(t - \tau_i)$ and $n_i'(n)$ is the additive noise component, respectively. The speech signal attenuation at each microphone is ignored under the far field assumption. In general, the multichannel
Wiener filter is implemented in the frequency domain: Before applying DFT, the noisy microphone signal $y'_i(n)$ is firstly segmented into frames with a frame length $L$ using an analysis window with the same length $L$ and a frame shift of $\Delta L$. After the processing in the frequency domain, the enhanced speech frame is then transformed back into the time domain with an inverse DFT (IDFT), a synthesis window, and the overlap-and-add (OLA) operation with the previously enhanced speech frame. Applying such a framework, the vector of microphone signals can then be formulated with the frame index $\ell$ and the frequency bin $k$ as

$$Y'(\ell, k) = S'(\ell, k) + N'(\ell, k) = S(\ell, k) \cdot D(k) + N'(\ell, k). \tag{2.2.1}$$

with $Y'(\ell, k) = [Y'_1(\ell, k), Y'_2(\ell, k), \ldots, Y'_M(\ell, k)]^T$, $S'(\ell, k) = S(\ell, k) \cdot D(k)$, and noise $N'(\ell, k) = [N'_1(\ell, k), N'_2(\ell, k), \ldots, N'_M(\ell, k)]^T$, respectively. The term $D(k)$ is the propagation vector representing the delays of the desired source signal $S(\ell, k)$ based on a reference microphone depending on the microphone array geometry as shown in Fig. 2.5

$$D(k) = \left[\exp\left(-j\frac{2\pi f_s \tau_1}{K}\right), \exp\left(-j\frac{2\pi f_s \tau_2}{K}\right), \ldots, \exp\left(-j\frac{2\pi f_s \tau_M}{K}\right)\right]^T. \tag{2.2.2}$$

Following the signal model given in (2.2.1), the enhanced speech signal of the multichannel Wiener filter can be written as (Simmer et al., 2001)

$$\hat{S}(\ell, k) = W^H(\ell, k) \cdot Y'(\ell, k), \tag{2.2.3}$$

where $W(\ell, k)$ represents the filter coefficients vector, and $(\cdot)^H$ denotes the Hermitian operator, respectively. Defining the error signal as $E(\ell, k) = S(\ell, k) - \hat{S}(\ell, k)$, the error

\[\text{To ensure quasi-stationarity of the speech signal, the frame length is usually chosen within a length of about 30 ms (Allen, 1977)}\]
signal PSD can be computed as

\[
\Phi_{EE}(\ell, k) = E\{E(\ell, k) \cdot E^*(\ell, k)\} \tag{2.2.4}
\]

\[
= E\{[S(\ell, k) - \hat{S}(\ell, k)] \cdot [S(\ell, k) - \hat{S}(\ell, k)]^*\}
\]

\[
= E\{[S(\ell, k) - \mathbf{W}^H(\ell, k) \cdot \mathbf{Y}'(\ell, k)] \cdot [S(\ell, k) - \mathbf{W}^H(\ell, k) \mathbf{Y}'(\ell, k)]^*\},
\]

where \((\cdot)^*\) denotes the complex conjugate operator. Based on the definition of the speech signal PSD \(\Phi_{SS}(\ell, k)\), the cross-PSD vector \(\Phi_{Y'S}(\ell, k)\), and the auto-PSD matrix \(\Phi_{Y'Y'}(\ell, k)\)

\[
\Phi_{SS}(\ell, k) = E\{S(\ell, k) \cdot S^*(\ell, k)\} \tag{2.2.5}
\]

\[
\Phi_{Y'S}(\ell, k) = E\{Y'(\ell, k) \cdot S^*(\ell, k)\} \tag{2.2.6}
\]

\[
\Phi_{Y'Y'}(\ell, k) = E\{Y'(\ell, k) \cdot Y'^H(\ell, k)\} \tag{2.2.7}
\]

(2.2.4) can be further formulated as

\[
\Phi_{EE}(\ell, k) = E\{\Phi_{SS}(\ell, k) - \mathbf{W}^H(\ell, k) \Phi_{Y'S}(\ell, k) - \Phi_{Y'S}^H(\ell, k) \mathbf{W}(\ell, k)
\]

\[
+ \mathbf{W}^H(\ell, k) \Phi_{Y'Y'}(\ell, k) \mathbf{W}(\ell, k)\} \tag{2.2.8}
\]

According to the MMSE criterion, the optimal filter can be achieved by minimizing the sum of \(\Phi_{EE}(\ell, k)\) for all frequency bins. However, since \(\Phi_{EE}(\ell, k)\) is real-valued and non-negative for each frequency bin \(k\) within each frame, the optimal filter vector \(\mathbf{W}_{\text{opt}}(\ell, k)\) can be achieved by minimizing \(\Phi_{EE}(\ell, k)\) given in (2.2.8). Since \(\Phi_{EE}(\ell, k)\) is a quadratic function of \(\mathbf{W}(\ell, k)\), its global minimum can be obtained by setting the gradient of \(\Phi_{EE}(\ell, k)\) w.r.t. \(\mathbf{W}(\ell, k)\) to zero (Haykin, 2001)

\[
\nabla_{\mathbf{W}}(\Phi_{EE}(\ell, k)) = 2\frac{\partial \Phi_{EE}(\ell, k)}{\partial \mathbf{W}^*(\ell, k)} = -2\Phi_{Y'S}(\ell, k) + 2\Phi_{Y'Y'}(\ell, k) \mathbf{W}(\ell, k) \overset{!}{=} \mathbf{0}, \tag{2.2.9}
\]

where \(\nabla\) is the gradient operator. The solution of \(\Phi_{Y'Y'}(\ell, k) \mathbf{W}_{\text{opt}}(\ell, k) = \Phi_{Y'S}(\ell, k)\) is then the Wiener-Hopf equation being expressed in a multichannel style with matrix notations (Simmer et al., 2001). Assuming that \(\Phi_{Y'Y'}(\ell, k)\) is positive definite, the optimal multichannel Wiener filter can be computed as (Simmer et al., 2001)

\[
\mathbf{W}_{\text{opt}}(\ell, k) = \Phi_{Y'Y'}^{-1}(\ell, k) \Phi_{Y'S}(\ell, k) \tag{2.2.10}
\]

2.2.2 Decomposition

The general multichannel Wiener filter expressed in (2.2.10) is derived with no explicit assumptions. However, if \(s(n)\) and \(n'_i(n)\) are further assumed to be statistically independent, \(\Phi_{Y'S}(\ell, k)\) can then be derived by inserting (2.2.1) into (2.2.6) as

\[
\Phi_{Y'S}(\ell, k) = E\{[S(\ell, k) \mathbf{D}(k) + \mathbf{N}'(\ell, k)] \cdot S^*(\ell, k)\}
\]

\[
= \Phi_{SS}(\ell, k) \mathbf{D}(k) + \underbrace{\Phi_{N'S}(\ell, k)}_{=0}
\]

\[
= \Phi_{SS}(\ell, k) \mathbf{D}(k). \tag{2.2.11}
\]
In a similar way, inserting (2.2.11) into (2.2.10), the auto-correlation matrix $\Phi_{NN'}(\ell, k)$ can be simplified into

$$
\Phi_{NN'}(\ell, k) = E\{[S(\ell, k)D(k) + N'(\ell, k)]\cdot[S(\ell, k)D(k) + N'(\ell, k)]^H\} \\
= \Phi_{SS}(\ell, k)D(k)D^H(k) + \Phi_{NS}(\ell, k)D^H(k) + D(k)\Phi_{SN'}(\ell, k) + \Phi_{NN'}(\ell, k) \\
= \Phi_{SS}(\ell, k)D(k)D^H(k) + \Phi_{NN'}(\ell, k).
$$

(2.2.12)

Inserting (2.2.11) and (2.2.12) into (2.2.10), the multichannel Wiener filter can now be rewritten as

$$
W_{opt}(\ell, k) = \left[\Phi_{SS}(\ell, k)D(k)D^H(k) + \Phi_{NN'}(\ell, k)\right]^{-1}\Phi_{SS}(\ell, k)D(k).
$$

(2.2.13)

Applying the Sherman-Morrison-Woodbury formula (Haykin, 2001) and assuming $\Phi_{NN'}$ to be positive definite, (2.2.13) can thus be derived as (Simmer et al., 2001)

$$
W_{opt}(\ell, k) = \frac{\Phi_{NN'}^{-1}(\ell, k)D(k)}{D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k)} \cdot \frac{\Phi_{SS}(\ell, k)}{\Phi_{SS}(\ell, k) + [D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k)]^{-1}},
$$

(2.2.14)

which can be seen as a decomposition into two parts. The first part represents a minimum variance distortionless response (MVDR) beamformer, which will be explained in detail in Section 2.3.3. Applying firstly a beamformer to the clean speech signal only, the PSD of the clean speech signal after the beamformer can be computed as

$$
\Phi_{SS_{BF}} = E\{[W^H(\ell, k)S(\ell, k)D(k)]\cdot[W^H(\ell, k)S(\ell, k)D(k)]^*\} \\
= \Phi_{SS}(\ell, k)W^H(\ell, k)D(k)W(\ell, k) \\
= \Phi_{SS}(\ell, k)|W^H(\ell, k)D(k)|^2.
$$

(2.2.15)

Utilizing $W_{MVDR}(\ell, k)$ for an MVDR beamformer as given in (2.2.14), we achieve

$$
\Phi_{SS_{BF}} = \Phi_{SS}(\ell, k)\left|\frac{D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k)}{D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k)}\right|^2 = \Phi_{SS}(\ell, k).
$$

(2.2.16)

Obviously, the PSD of the clean speech signal remains unchanged by being filtered through the MVDR beamformer. The PSD of the noise signal after the beamformer can be calculated in the same way as

$$
\Phi_{NN_{BF}}(\ell, k) = E\{[W^H(\ell, k)N(\ell, k)][W^H(\ell, k)N(\ell, k)]^*\} \\
= N^H(\ell, k)W(\ell, k) \\
= W^H(\ell, k)\Phi_{NN'}(\ell, k)W(\ell, k).
$$

(2.2.17)

Utilizing $W(\ell) = W_{MVDR}(\ell, k)$ as given in (2.2.14), we get

$$
\Phi_{NN_{BF}}(\ell, k) = \frac{D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k)}{(D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k))^2} = [D^H(k)\Phi_{NN'}^{-1}(\ell, k)D(k)]^{-1}.
$$

(2.2.18)
Inserting (2.2.16) and (2.2.18) into (2.2.14) yields

\[ W_{opt}(\ell,k) = \frac{\Phi_{NN}^{-1}(\ell,k)D(k)}{D^H(k)\Phi_{NN}^{-1}(\ell,k)D(k)} \cdot \left[ \frac{\Phi_{SS}(\ell,k)}{\Phi_{SS}(\ell,k) + \Phi_{NB-NF}(\ell,k)} \right] . \]  

(2.2.19)

Comparing (2.2.19) to (2.2.14), it can be seen that the second part in (2.2.14) is actually a single-channel Wiener filter under the assumption of the clean speech signal and the noise signal being statistically independent (Hänsler and Schmidt, 2004). Applying this decomposition, the optimal multichannel Wiener filter can be compactly formulated as

\[ W_{opt}(\ell,k) = W_{MVDR}(\ell,k) \cdot H_{PF}(\ell,k) , \]  

(2.2.20)

where \( H_{PF}(\ell,k) \) denotes the single-channel Wiener post-filter after the MVDR beamformer. The multichannel Wiener filter thus comprises an MVDR beamformer and a post-filter. However, there are different beamformer and post-filter design approaches. A classification based on their design features is shown in Fig. 2.6. In Sections 2.3 and 2.4, all beamformer and post-filter design approaches listed in Fig. 2.6 are briefly discussed.
Figure 2.6: Topology of different beamformers and post-filters, with references to equations
2.3 Beamforming

In the following, several microphone array layouts are shortly introduced with a focus on the one-dimensional linear equidistant microphone array, which is applied in this thesis. In addition, a brief overview of the instrumental performance measures for beamforming algorithms are presented. Subsequently, different beamforming algorithms are described along with some discussions on their advantages and disadvantages for automotive noise reduction algorithms.

2.3.1 Microphone Array Design and Performance Measures

For each beamforming algorithm, an array system is needed, in our case, it implies a microphone array with several omnidirectional microphones. The design aim of a microphone array is to process multichannel microphone signals employing a beamforming algorithm, so that the clean speech signal can be acquired without distortion and noise signals coming from all other directions are attenuated. A typical microphone array can be realized in a one-dimensional linear equidistant layout, which is depicted in Fig. 2.7 using four microphones with \( d \) being the distance of adjacent microphones. Generally, the source signal can be located anywhere with an angle of arrival \( \theta_o \) towards the microphone array. In the literature, two directions of the source signal towards the microphone array are defined as follows. The first one is the endfire direction, where the source signal is located on the same line of the microphone array axis, while the second one is the broadside direction, where the source signal is located vertically to the microphone array axis (Trees, 2002). Having multiple inputs, beamforming algorithms based on microphone arrays exploit not only temporal and spectral information, but also spatial information. However, similar to temporal aliasing caused by choosing the sampling frequency \( f_s \) to be less than twice of the signal’s upper cutoff frequency, spatial aliasing of array processing can result into ambiguous arriving directions of the source signal. To avoid spatial aliasing, the condition (Trees, 2002)

\[
\frac{d}{\lambda} \leq \frac{1}{1 + |\cos \theta_o|} \quad (2.3.1)
\]

has to be satisfied. Generally, microphone arrays shall be steered towards all possible angles of arrival. In addition, microphone signals are naturally broadband signals with their

![Figure 2.7: Layout of a one-dimensional linear equidistant microphone array using four microphones](image)
wavelength $\lambda$ satisfying $\lambda = \frac{c}{f}$. With these constraints, (2.3.1) can then be formulated as

$$d \leq \frac{\lambda_{\text{min}}}{2} = \frac{c}{2f_{\text{max}}}.$$  (2.3.2)

Let $f_{\text{max}} = \frac{f_s}{2}$ be the signal’s upper cutoff frequency, we obtain

$$d \leq \frac{c}{f_s},$$  (2.3.3)

which shall be met by a microphone array using the sampling frequency $f_s$ to avoid spatial aliasing. More particular, we can say that we obtain $d \leq \frac{c}{f_s}$ for an endfire array (or a general array), and $d \leq \frac{2c}{f_s}$ for a broadside array. For a microphone array applied to wideband signals with $f_s = 16$ kHz, the distance of adjacent microphones $d$ shall then be chosen equal to or less than 2.15 cm. However, on the other hand, beamforming algorithms generally exhibit very poor directivity for low frequencies (up to 1000 Hz) within such a small distance. Therefore, in practice, a compromise has to be taken for choosing the distance of adjacent microphones (Dörbecker, 1998). Compared to a linear equidistant microphone array, an inequidistant nested microphone array layout can also be applied, which includes several nested subarrays being responsible for different frequency regions (see Fischer and Kammeyer, 1997, Zheng et al., 2005). However, more microphones are required for a nested microphone array, which makes it nearly infeasible for a hands-free equipment in a car with its commonly limited space.

In addition, there are microphone array designs in a two-dimensional layout, e.g., a square layout (see Zelinski, 1988, Simmer et al., 1993, Hoshuyama et al., 1997), or a circular layout (see Tamai et al., 2003, Tashev and Malvar, 2005, Meyer and Elko, 2008, Habets and Benesty, 2011). Microphone arrays can also be designed in three-dimensional layouts, as in early works by Brandstein et al. (1997) and Flanagan et al. (1993). A common three-dimensional microphone layout is the spherical microphone array layout (Rafaely, 2005). Furthermore, Abhayapala and Gupta (2009) proposed a hybrid geometry to replace the classical spherical layout. Due to the limited space of a hands-free equipment for noise reduction in an automotive environment, two-dimensional, especially three-dimensional microphone arrays with more microphones than a linear one-dimensional microphone array are practically not feasible. Therefore, in this thesis, we apply a head unit-integrated one-dimensional linear equidistant microphone array using four microphones, which was prepublished in (Yu and Fingscheidt, 2009). The distance of adjacent microphones is 3.6 cm and the angle of arrival is 85.25°. The construction details of the applied microphone array can be found in Chapter 4.

Several instrumental measures can be applied to evaluate microphone arrays based on beamforming algorithms. In the following, four instrumental measures are briefly introduced.

**Array Gain**

Array gain is applied to measure the achieved signal-to-noise ratio (SNR) improvement between the output of a beamformer and one of the microphone inputs. According to
Bitzer and Simmer (2001), the array gain can be defined as
\[
AG(\ell, k) = \frac{\text{SNR}_{\text{BF}}(\ell, k)}{\text{SNR}_y(\ell, k)},
\]
where \(\text{SNR}_{\text{BF}}(\ell, k)\) is the SNR of the beamformer output signal and \(\text{SNR}_y(\ell, k)\) is the SNR of the \(i\)th microphone input. Under the assumption of stationary signals and a homogeneous noise field with \(\Phi_{N_i'N_j'}(\ell, k) = \Phi_{N_iN_j}(\ell, k)\) and \(\Phi_{S_i'S_j'}(\ell, k) = \Phi_{SS}(\ell, k)\), the microphone input SNR can be expressed as
\[
\text{SNR}_y(\ell, k) = \frac{\Phi_{SS}(\ell, k)}{\Phi_{NN}(\ell, k)}. \tag{2.3.5}
\]

The beamformer output SNR is defined by the clean speech PSD \(\Phi_{S_{\text{BF}}}S_{\text{BF}}(\ell, k)\) and the noise PSD \(\Phi_{N_{\text{BF}}N_{\text{BF}}}(\ell, k)\) after the beamformer as
\[
\text{SNR}_{\text{BF}}(\ell, k) = \frac{\Phi_{S_{\text{BF}}}S_{\text{BF}}(\ell, k)}{\Phi_{N_{\text{BF}}N_{\text{BF}}}(\ell, k)}. \tag{2.3.6}
\]

Applying (2.2.15) and (2.2.17) for \(\Phi_{S_{\text{BF}}}S_{\text{BF}}(\ell, k)\) and \(\Phi_{N_{\text{BF}}N_{\text{BF}}}(\ell, k)\), respectively, we get
\[
\text{SNR}_{\text{BF}}(\ell, k) = \frac{\Phi_{SS}(\ell, k)|W^H(\ell, k)D(\ell, k)|^2}{W^H(\ell, k)\Gamma_{NN}(\ell, k)W(\ell, k)}. \tag{2.3.7}
\]

Using a homogeneous noise field, we achieve further (Bitzer, 2001)
\[
\Phi_{NN}(\ell, k) = \Phi_{NN}(\ell, k)\Gamma_{NN}(\ell, k),
\]
where \(\Gamma_{NN}(\ell, k)\) is the noise coherence matrix with
\[
\Gamma_{i,j}(\ell, k) = \frac{\Phi_{N_i'N_j'}(\ell, k)}{\sqrt{\Phi_{N_i'N_i}(\ell, k)\Phi_{N_j'N_j}(\ell, k)}} \in \mathbb{C}. \tag{2.3.9}
\]

Inserting (2.3.8) into (2.3.7), we can finally compute the SNR of the beamformer output as
\[
\text{SNR}_{\text{BF}}(\ell, k) = \frac{\Phi_{S_{\text{BF}}}S_{\text{BF}}(\ell, k)}{\Phi_{N_{\text{BF}}N_{\text{BF}}}(\ell, k)} = \frac{\Phi_{SS}(\ell, k)|W^H(\ell, k)D(\ell, k)|^2}{\Phi_{NN}(\ell, k)W^H(\ell, k)\Gamma_{NN}(\ell, k)W(\ell, k)}. \tag{2.3.10}
\]

Inserting (2.3.10) and (2.3.5) into (2.3.4), we can then derive the array gain as
\[
AG(\ell, k) = \frac{|W^H(\ell, k)D(\ell, k)|^2}{W^H(\ell, k)\Gamma_{NN}(k)W(k)}, \tag{2.3.11}
\]
which depending only on the beamformer coefficients and the coherence matrix of a certain noise field. Please note, the array gain defined in (2.3.11) is the general expression depending on the frame index \(\ell\) and frequency bin \(k\). In this thesis, we focus on a fixed beamformer with \(W(k)\) and apply the noise coherence matrix \(\Gamma_{NN}(k)\) as described in Section 2.1\(^5\). Therefore, the following definition of array gain depending only on the frequency bin \(k\) is applied in this thesis as
\[
AG(k) = \frac{|W^H(k)D(k)|^2}{W^H(k)\Gamma_{NN}(k)W(k)}. \tag{2.3.12}
\]
\(^5\)The term \(\Gamma_{NN}(k)\) defined in Section 2.1 corresponds here to the term \(\Gamma_{N_i'N_j'}(\ell, k)\) in (2.3.11) without dependency on the frame index \(\ell\).
2. Basic Principles of Beamforming and Post-Filtering

White Noise Gain

Applying an incoherent noise field with its coherence matrix $\Gamma_{\text{inc}}^{\text{NN}}(k) = I$ to the array gain $AG(k)$ (2.3.12), the so-called white noise gain (WNG) can be obtained in a logarithmic scale as (Bitzer and Simmer, 2001)

$$WNG(k) = 10 \log \left( \frac{|W^H(k)D(k)|^2}{W^H(k)W(k)} \right).$$  \hspace{1cm} (2.3.13)

The white noise gain $WNG(k)$ is applied to measure the attenuation performance of a beamformer for the uncorrelated white noise.

Directivity Index

One important characteristic of a beamformer is its performance for attenuating diffuse noise. Such a performance can be measured by the directivity index (DI), which is derived by using the diffuse noise coherence matrix $\Gamma_{\text{diff}}^{\text{NN}}(k)$ in the array gain $AG(k)$ (2.3.12) in a logarithmic scale as (Bitzer and Simmer, 2001)

$$\text{DI}(k) = 10 \log \left( \frac{|W^H(k)D(k)|^2}{W^H(k)\Gamma_{\text{diff}}^{\text{NN}}(k)W(k)} \right).$$  \hspace{1cm} (2.3.14)

Beam Pattern

Unlike the array gain, which focuses only on the directivity response of a beamformer for the arriving direction of the desired clean speech signal, beam pattern (BP) measures the directivity response of a beamformer for a wavefront coming from a certain arriving direction defined by $(\phi, \theta)$. However, in this thesis we deal with the one-dimensional linear microphone array exhibiting a rotational symmetry w.r.t. $\phi$. Therefore, the coherence function of the wavefront depends only on $\theta$. The coherence matrix for such a wavefront is given with elements $\Gamma_{\text{wav}}^{\text{NN}}(k, \theta, i, j) = \exp \left( -j \frac{2\pi k f_s d_{i,j} \cos \theta}{c K} \right)$. Using the coherence matrix $\Gamma_{\text{wav}}^{\text{NN}}(k, \theta)$ of a wavefront coming from a certain arriving direction defined by $\theta$ in the array gain $AG(k)$ (2.3.12), the beam pattern (BP) can be expressed in a logarithmic scale as (Bitzer and Simmer, 2001)

$$\text{BP}(k, \theta) = -10 \log \left( \frac{|W^H(k)D(k)|^2}{W^H(k)\Gamma_{\text{wav}}^{\text{NN}}(k, \theta)W(k)} \right).$$  \hspace{1cm} (2.3.15)

It can be seen that the beam pattern depends not only on $k$, but also on $\theta$. 
2.3.2 Linearly Constrained Minimum Variance (LCMV) Beamforming

Among different beamforming techniques, the linearly constrained minimum variance (LCMV) beamforming has been applied in most cases. By specifying the LCMV beamforming, several linear constraints have to be fulfilled, while the beamformer output variance is set to be minimized at the same time. According to that, a set of linear constraints can be defined as an $M \times N_c$ constraint matrix $C$ with $N_c < M$, which satisfies (Trees, 2002)

$$C^H(\ell,k)W(\ell,k) = \Xi(\ell,k), \quad (2.3.16)$$

where each value of the $(N_c \times 1)$-dimensional vector $\Xi(\ell,k)$ combined with the corresponding linearly independent column of $C(\ell,k)$ represents one of the $N_c$ linearly independent constraints. Satisfying the linear constraints as given in (2.3.16), the LCMV beamformer can be derived by minimizing the beamformer output variance as (Trees, 2002)

$$W_{LCMV}(\ell,k) = \arg \min_{W(\ell,k)} W^H(\ell,k) \Phi_{YY'}(\ell,k) W(\ell,k) \left|_{C^H(\ell,k)W(\ell,k) = \Xi(\ell,k)} \right.. \quad (2.3.17)$$

To solve (2.3.17), a Lagrange multiplier can be applied, which leads to the final solution of the LCMV beamformer as (Trees, 2002)

$$W_{LCMV}(\ell,k) = \frac{\Phi_{YY'}^{-1}(\ell,k)C(\ell,k)\Xi(\ell,k)}{C^H(\ell,k)\Phi_{YY'}^{-1}(\ell,k)C(\ell,k)} \quad \cdot (2.3.18)$$

In the literature, the LCMV beamformer is often called Frost beamformer, since the result given in (2.3.18) was also derived by Frost (1972) by utilizing the constrained least mean square (LMS) algorithm.

Keeping the minimum variance requirement in (2.3.17) unchanged and selecting different suitable constraints for $C(\ell,k)$ and $\Xi(\ell,k)$ in (2.3.16), a number of beamformers can be obtained: E.g., the distortionless response constraint (Bitzer and Simmer, 2001), the eigenvector constraint (Er and Cantoni, 1985, van Veen, 1991), and the directional constraint (Steele, 1983, Yansouni and Inkol, 1983). Among these constraints, the most commonly employed one is the distortionless response constraint, which is used to derive the MVDR beamformer.

2.3.3 Minimum Variance Distortionless Response (MVDR) Beamforming

As a special case to LCMV beamforming, the minimum variance distortionless response (MVDR) beamforming is derived by replacing $C(\ell,k) = D(k)$ and $\Xi(\ell,k) = 1$ in (2.3.16) as (Bitzer and Simmer, 2001)

$$D^H(k)W_{MVDR}(\ell,k) = 1 \quad . \quad (2.3.19)$$
Using the propagation vector $\mathbf{D}(k)$ defined in (2.2.2) describing the delays of the clean speech signal for each microphone, the clean speech signal is then being kept undistorted by (2.3.19) meaning a distortionless response (DR).

In order to achieve minimum variance (MV) under the constraint of distortionless response for the clean speech signal, only the variance of noise signals shall be minimized, while the variance of the clean speech signal shall be kept unchanged. Therefore, instead of the variance of the noisy speech signal $\Phi_{Y'Y'}(\ell,k)$, only the variance of noise signals $\Phi_{N'N'}(\ell,k)$ needs to be minimized. Applying the constraint of distortionless response provided in (2.3.19) to (2.3.17) and minimizing the variance of noise signals $\Phi_{N'N'}(\ell,k)$, the MVDR beamformer can be derived as

$$W_{\text{MVDR}}(\ell,k) = \arg\min_{W(\ell,k)} \mathbf{W}^H(\ell,k)\Phi_{N'N'}(\ell,k)\mathbf{W}(\ell,k) \bigg|_{\mathbf{D}^H(k)\mathbf{W}(\ell,k) = 1}. \quad (2.3.20)$$

Utilizing a Lagrange multiplier, the final solution of the MVDR beamformer can then be formulated as (Bitzer and Simmer, 2001)

$$W_{\text{MVDR}}(\ell,k) = \frac{\Phi_{N'N'}^{-1}(\ell,k)\mathbf{D}(k)}{\mathbf{D}^H(k)\Phi_{N'N'}^{-1}(\ell,k)\mathbf{D}(k)}. \quad (2.3.21)$$

Assuming a homogeneous noise field with $\Phi_{N'N'}(k) = \Phi_{NN}(k) \cdot \Gamma_{NN}(k)$ and $\Gamma_{NN}(k)$ as being described in Section 2.1, the MVDR beamformer can be reformulated with the noise coherence matrix of a homogeneous noise field as

$$W_{\text{MVDR}}(k) = \frac{\Gamma_{NN}^{-1}(k)\mathbf{D}(k)}{\mathbf{D}^H(k)\Gamma_{NN}^{-1}(k)\mathbf{D}(k)}. \quad (2.3.22)$$

In designing an MVDR beamformer, two well defined noise fields are usually applied, which result into two well-known beamforming techniques referred to as the delay-and-sum beamforming and the superdirective beamforming.

**Delay-and-Sum Beamforming**

The MVDR beamformer for an incoherent noise field leads to the so-called delay-and-sum (DS) beamformer. The DS beamformer can be calculated by applying the incoherent noise coherence matrix $\Gamma_{NN}^{\text{inc}}(k) = \mathbf{I}$ in (2.3.22) as (Bitzer and Simmer, 2001)

$$W_{\text{DS}}(k) = \frac{\Gamma_{NN}^{-1}(k)\mathbf{D}(k)}{\mathbf{D}^H(k)\Gamma_{NN}^{-1}(k)\mathbf{D}(k)}. \quad (2.3.23)$$

This beamforming approach is defined as the delay-and-sum beamforming, since utilizing the DS beamformer, delays of microphone signals $Y'(\ell,k) = S(\ell,k) \cdot \mathbf{D}(k) + N'(\ell,k)$ are
2.3.3. Minimum Variance Distortionless Response (MVDR) Beamforming

compensated, which makes an in-phase summation of the clean speech signal:

\[
\hat{S}_{DS}(\ell, k) = W^H_{DS}(k) \cdot Y'(\ell, k)
\]

\[
= \frac{1}{M} D^H(k) \cdot [S(\ell, k) \cdot D(k) + N'(\ell, k)]
\]

\[
= \frac{1}{M} D^H(k) D(k) S(\ell, k) + \frac{1}{M} D^H(k) N'(\ell, k)
\]

\[
= S(\ell, k) + \frac{1}{M} \sum_{i=1}^{M} N_i(\ell, k).
\]  

(2.3.24)

The term \( \frac{1}{M} \sum_{i=1}^{M} N_i(\ell, k) \) in (2.3.24) represents the not-in-phase summation of the noise signal \( N(\ell, k) \) after the delay compensation, which leads to noise reduction. Note that the DS beamformer is implemented simply by delay and sum: \( \hat{S}_{DS}(\ell, k) = \frac{1}{M} D^H(k) \cdot Y'(\ell, k) \).

Superdirective Beamforming

In an automotive environment, car noise can be well approximated by a diffuse noise field (Meyer and Simmer, 1997). Applying the diffuse noise coherence matrix \( \Gamma_{NN}^{\text{diff}}(k) \) in (2.3.22), the superdirective (SD) beamforming can be formulated as (Bitzer and Simmer, 2001)

\[
W_{SD}(k) = \frac{\Gamma_{NN}^{\text{diff}}-1(k) D(k)}{D^H(k) \Gamma_{NN}^{\text{diff}}-1(k) D(k)}.
\]  

(2.3.25)

Note that the SD beamformer in practice is implemented as filter and sum: \( \hat{S}_{SD}(\ell, k) = W^H_{SD}(k) \cdot Y'(\ell, k) \).

Comparison between DS and SD Beamforming

Knowing the definition of the DS beamforming and the SD beamforming, it is worth to compare their performance using the instrumental measures introduced in Section 2.3.1. In Fig. 2.8(a), the directivity index for the DS beamforming and the SD beamforming realized by the applied head unit-integrated microphone array using 4 microphones is shown. It can be observed that the SD beamforming exhibits a much higher directivity than the DS beamforming at low frequencies. The DS beamforming yields close to 0 dB in the low-frequency region. This phenomenon is defined as superdirectivity in the literature (Bitzer and Simmer, 2001). At high frequencies, the SD beamforming provides about the same directivity index as the DS beamforming, since the diffuse noise field used to derive the SD beamformer becomes nearly uncorrelated for high frequencies as the incoherent noise field used for deriving the DS beamformer. However, when we look at the white noise gain (WNG) of both beamformers in Fig. 2.8(b), it can be seen that at low frequencies, the SD beamformer delivers a tremendous negative white noise gain in a logarithmic scale meaning an amplification of the uncorrelated white noise. In contrary to that, the DS beamformer shows a robust attenuation of white noise of 10 \( \log(M) = 6.02 \) dB with \( M = 4 \). This self-amplification of white noise makes an ideal SD beamforming unattractive for a
real practical design, since uncorrelated white noise nearly always exists as electric self-noise of microphones (Bitzer and Simmer, 2001). To solve this problem, the constrained superdirective beamforming technique can be applied.

Constrained Superdirective Beamforming

The problem of high WNG at low frequencies for the superdirective beamforming originates from the applied diffuse noise coherence function. A large value up to one is reached by the diffuse noise coherence function $\Gamma_{i,j}^{\text{dif}}(k)$ (2.1.18) at low frequencies. This leads to a large value up to one for off-diagonal elements of the diffuse noise coherence matrix at low frequencies, which results into the self-amplification of the uncorrelated white noise. In order to get rid of this problem, Gilbert and Morgan (1955) proposed an approach to add a very small constant to the diagonal entries of the diffuse noise coherence matrix, so that a constrained superdirective beamforming with less WNG can be achieved. However, this method makes the modified noise coherence matrix not really a coherence matrix, since diagonal entries may become larger than one. Therefore, a mathematically equivalent approach was provided by Bitzer and Simmer (2001), where each off-diagonal entry of the diffuse noise coherence matrix is scaled by a factor $(1 + \varrho)^{-1}$ as

$$\Gamma_{i,j}^{\text{dif}}(k) = \frac{\Gamma_{i,j}^{\text{dif}}(k)}{1 + \varrho}, \quad \forall i \neq j. \tag{2.3.26}$$

Employing the modified diffuse noise coherence matrix $\Gamma_{NN}^{\text{dif}}(k)$ with its off-diagonal entries being expressed by $\Gamma_{i,j}^{\text{dif}}(k)$ in (2.3.25), the constrained superdirective beamformer (CS) can then be formulated as

$$W_{CS}(k) = \frac{\Gamma_{NN}^{\text{dif}}(k)^{-1}D(k)}{D^H(k)\Gamma_{NN}^{\text{dif}}(k)^{-1}(k)D(k)}. \tag{2.3.27}$$
2.3.3. Minimum Variance Distortionless Response (MVDR) Beamforming

According to (2.3.26), we can observe that setting $\varrho$ to 0 results into an ideal SD beamforming without any constraints. For $\varrho \to +\infty$ we obtain $\Gamma_{i,j}^{\text{dif}}(k) \to 0$ yielding a DS beamforming. Therefore, the factor $\varrho$ provides a possibility to design the CS beamforming with acceptable WNG. In Figs. 2.9(a) and 2.9(b), the DI and WNG of the CS beamforming with three typical $\varrho$ values are plotted in comparison to the DS beamforming, respectively. In this thesis, the CS beamforming is computed with $\varrho = 0.01$ showing a good trade-off between the DI and the WNG performances. Figs. 2.10(a) and 2.10(b) depict the beam pattern of the DS beamforming and the CS beamforming using $\varrho = 0.01$, respectively. For both beamformers, a good directivity can only be achieved for $f > 1000$ Hz. However,
the energies of car noise mostly concentrate at low frequencies, therefore, a post-filter has to be applied after the MVDR beamforming to achieve further noise reduction. Different post-filter designs are presented in Section 2.4.

### 2.3.4 Generalized Sidelobe Canceller (GSC) Beamforming

Until now, the LCMV beamforming along with its special case of an MVDR beamforming have been derived. Griffiths and Jim (1982) proposed an alternative approach called the generalized sidelobe canceller (GSC) beamforming, which actually decomposes the LCMV beamforming into two orthogonal subspaces: The linear constraints (LC) subspace and the minimum variance (MV) subspace. The equivalence of the LCMV beamforming to the GSC beamforming was shown by Breed and Strauss (2002). Applying the two orthogonal subspaces, the optimal LCMV beamforming coefficients $W_{\text{LCMV}}(\ell,k)$ can then be divided into (Trees, 2002)

$$ W_{\text{GSC}}(\ell,k) = W_{\text{LC}}(\ell,k) - W_{\text{MV}}(\ell,k) . \quad (2.3.28) $$

The term $W_{\text{LC}}(\ell,k)$ represents the projection of $W_{\text{LCMV}}(\ell,k)$ onto the LC subspace, while the term $W_{\text{MV}}(\ell,k)$ stands for the projection of $W_{\text{LCMV}}(\ell,k)$ onto the MV subspace. The LC subspace is defined by the constraint matrix $C(\ell,k)$ (cf. (2.3.16)), while the MV subspace is defined by an $M \times (M - N_c)$ blocking matrix $B(\ell,k)$. To maintain orthogonality, $C(\ell,k)$ and $B(\ell,k)$ have to fulfill

$$ C^H(\ell,k)B(\ell,k) = 0 , \quad (2.3.29) $$

with 0 being an $N_c \times (M - N_c)$ zero matrix. Let us first consider the derivation of the optimal filter coefficients $W_{\text{LC}}(\ell,k)$ in (2.3.28). The projection can be achieved by using the subspace projection matrix $C_p(\ell,k)$ defined as (Trees, 2002)

$$ C_p(\ell,k) = C(\ell,k) \left[ C^H(\ell,k)C(\ell,k) \right]^{-1} C^H(\ell,k) . \quad (2.3.30) $$

Applying $C_p(\ell,k)$ and $W_{\text{LCMV}}(\ell,k)$ as derived in (2.3.18), $W_{\text{LC}}(\ell,k)$ can then be calculated as (Trees, 2002)

$$ W_{\text{LC}}(\ell,k) = C_p^H(\ell,k)W_{\text{LCMV}}(\ell,k) $$

$$ = C(\ell,k) \left[ C^H(\ell,k)C(\ell,k) \right]^{-1} \Xi(\ell,k) . \quad (2.3.31) $$

It can be seen that the filter coefficients $W_{\text{LC}}(\ell,k)$ depend only on $C(\ell,k)$ and $\Xi(\ell,k)$. In the case of employing only the distortionless response of an MVDR beamformer with $C(\ell,k) = D(k)$ and $\Xi(\ell,k) = 1$, (2.3.31) turns exactly to a delay-and-sum beamformer with (cf. (2.3.23))

$$ W_{\text{LC}}(k) = W_{\text{DS}}(k) = \frac{1}{M} D(k) . \quad (2.3.32) $$

In a second step, $W_{\text{MV}}(\ell,k)$ can be transformed into a cascaded form as (Trees, 2002)

$$ W_{\text{MV}}(\ell,k) = B^H(\ell,k)W_{\text{INT}}(\ell,k) , \quad (2.3.33) $$

where $W_{\text{INT}}(\ell,k)$ represents the filter coefficients of an interference cancellation module.
2.3.4. Generalized Sidelobe Canceller (GSC) Beamforming

Due to the orthogonality between the LC and the MV subspaces, the output of the blocking matrix $\mathbf{B}^H(\ell,k) \cdot \mathbf{Y}'(\ell,k)$ shall contain no signal of the LC subspace. This property gives the name *blocking matrix* to $\mathbf{B}(\ell,k)$ (Trees, 2002), since it shall reject all signals belonging to the LC subspace. In the case of $\mathbf{C}(\ell,k) = \mathbf{D}(k)$, $\mathbf{B}(\ell,k)$ shall block the clean speech signal in the LC subspace and leave only noise signals to pass by. There are different approaches to design the blocking matrix for providing noise-only signals, e.g., the non-adaptive approach in Griffiths and Jim (1982), which just subtracts the time-aligned microphone signals to get noise signals only, and the adaptive approach in Hoshuyama et al. (1999), which eliminates the clean speech signal by applying an adaptive LMS algorithm. The remaining noise signals in $\mathbf{B}^H(\ell,k) \cdot \mathbf{Y}'(\ell,k)$ shall then be used by the interference cancellation module. Recalling the minimum variance criterion (cf. (2.3.17)), the filter coefficients $\mathbf{W}_{\text{INT}}(\ell,k)$ can be calculated by minimizing the GSC beamforming output variance as

$$\mathbf{W}_{\text{INT}}(\ell,k) = \arg\min_{\mathbf{W}(\ell,k)} \left[ \mathbf{W}_{\text{GSC}}^H(\ell,k) \Phi_{\mathbf{Y}'\mathbf{Y}'}(\ell,k) \mathbf{W}_{\text{GSC}}(\ell,k) \right]$$

$$= \arg\min_{\mathbf{W}(\ell,k)} \left[ \mathbf{W}_{\text{LC}}(\ell,k) - \mathbf{B}^H(\ell,k) \mathbf{W}(\ell,k) \right]^H \Phi_{\mathbf{Y}'\mathbf{Y}'}(\ell,k) \left[ \mathbf{W}_{\text{LC}}(\ell,k) - \mathbf{B}^H(\ell,k) \mathbf{W}(\ell,k) \right].$$

(2.3.34)

The minimization can be solved by setting the gradient of (2.3.34) w.r.t. $\mathbf{W}(\ell,k)$ to zero. Thus, $\mathbf{W}_{\text{INT}}(\ell,k)$ can be derived as (Trees, 2002)

$$\mathbf{W}_{\text{INT}}(\ell,k) = \left[ \mathbf{B}^H(\ell,k) \Phi_{\mathbf{Y}'\mathbf{Y}'}(\ell,k) \mathbf{B}(\ell,k) \right]^{-1} \mathbf{B}^H(\ell,k) \Phi_{\mathbf{Y}'\mathbf{Y}'}(\ell,k) \mathbf{W}_{\text{LC}}(\ell,k).$$

(2.3.35)

The solution for $\mathbf{W}_{\text{INT}}(\ell,k)$ can also be adaptively computed by employing, e.g., the normalized least mean square (NLMS) algorithm (Hänsler and Schmidt, 2004).
The block diagram of the GSC beamforming is depicted in Fig. 2.11: In the upper reference path, a fixed beamforming module, e.g., a DS beamformer, can be applied to yield only the clean speech signal by fulfilling the linear constraint in the LC subspace. In the lower sidelobe canceller path, the blocking matrix and the interference cancellation module provide an adaptive filter structure utilizing the fixed beamforming output, i.e., the clean speech signal, as the reference signal and noisy microphone signals as input signals. The blocking matrix shall be designed in such a manner, that no parts of the clean speech signal can pass through it. However, this is usually not the case. The leakage of the clean speech signal into the interference cancellation module results into speech attenuation and distortion (Widrow et al., 1982). This problem was just partly reduced by applying a voice activity detector (VAD) to the blocking matrix and the interference cancellation module, so that the adaptation is only performed during speech pauses (Hoshuyama et al., 1999). It has been shown in Bitzer et al. (1998, 1999) that the GSC beamforming can provide more noise attenuation for coherent noise than for diffuse noise.

### 2.3.5 Other Beamforming Techniques

Besides the commonly applied LCMV beamforming with its special case MVDR beamforming and its alternative realization GSC beamforming, there are also some other beamforming techniques. Motivated by the human auditory system, a soft-constraint balancing the trade-off between noise reduction and speech distortion was introduced by Kaneda (1988) in the adaptive microphone array system for noise reduction (AMNOR) beamforming. Furthermore, a generalized singular value decomposition (GSVD) beamforming was proposed by Doclo and Moonen (2002), which is a subspace-based beamforming algorithm relying on no a priori knowledge of the source signal location and the microphone array layout. A useful analysis between the GSVD beamforming and the GSC beamforming can be found in Spriet et al. (2001). There are also beamforming algorithms utilizing native features of the speech signal, e.g., employing the dual excitation speech model (Brandstein, 1998), and applying the linear predictive coding (LPC) model (Brandstein, 1999).

### 2.4 Post-Filtering

As explained in Section 2.2.2, the multichannel Wiener filter can be decomposed into an MVDR beamformer (see Section 2.3.3) and a single-channel Wiener filter. Since employing a beamformer alone does not yield sufficient noise attenuation for car noise at low frequencies, the single-channel Wiener post-filter can be applied after the MVDR beamformer as the optimal post-filter (cf. (2.2.19))

\[
H_{PF}(\ell, k) = \frac{\Phi_{SS}(\ell, k)}{\Phi_{SS}(\ell, k) + \Phi_{NF,NPF}(\ell, k)},
\]

(2.4.1)
2.4.1 Designs for an Incoherent Noise Field

In the early publications (see Zelinski, 1988, Simmer and Wasiljeff, 1992), post-filters were designed based on the assumption of uncorrelated noise signals, which implies actually an incoherent noise field. The difference between Zelinski’s post-filter (Zelinski, 1988) and Simmer’s post-filter (Simmer and Wasiljeff, 1992) lies in the inputs to estimate the noise PSD for the post-filter. Such a difference makes Zelinski’s post-filter theoretically sub-optimal and Simmer’s post-filter theoretically optimal according to the post-filter definition in (2.4.1).

Zelinski’s Post-Filter

As shown in Fig. 2.12, the delay-compensated microphone signals are basis of the post-filter coefficient estimation (Zelinski, 1988)

\[
Y(\ell, k) = \text{diag}\{D^*(k)\} \cdot [S'(\ell, k) + N'(\ell, k)] \\
= 1 \cdot S(\ell, k) + \text{diag}\{D^*(k)\} \cdot N'(\ell, k) \\
= 1 \cdot S(\ell, k) + N(\ell, k),
\]

(2.4.2)

Figure 2.12: Block diagram of the multichannel Wiener filter being decomposed into a beamforming and post-filtering system (cf. Fig. 2.4)

which provides further noise attenuation. The block diagram of such a decomposition structure is shown in Fig. 2.12. In the following, two categories of post-filter designs are addressed, which include a specific design for uncorrelated noise signals in Section 2.4.1 and a generalized design based on the noise coherence function in Section 2.4.2.
with $\text{diag}\{\cdot\}$ being the diag operator and $\mathbf{1}$ denoting an all-ones vector. Zelinski took several assumptions in his post-filter design (Zelinski, 1988): (1) Noises among different microphones are uncorrelated resulting in an incoherent noise field; (2) speech and noise signals are uncorrelated; (3) the noise power spectrum is the same for all microphones, $\Phi_{N_iN_j}(\ell,k) = \Phi_{NN}(\ell,k)$, which implies actually a homogeneous noise field. Under these assumptions the auto-PSD $\Phi_{YY}(\ell,k)$ and cross-PSD $\Phi_{Y_jY_i}(\ell,k)$ of the delay-compensated microphone signals can be formulated as

$$\Phi_{YY}(\ell,k) = \Phi_{SS}(\ell,k) + \Phi_{NN}(\ell,k) \in \mathbb{R}, \quad (2.4.3)$$

$$\Phi_{Y_jY_i}(\ell,k) = \Phi_{SS}(\ell,k) \in \mathbb{R}, \quad (2.4.4)$$

Utilizing the first-order recursive updating approach proposed by Allen et al. (1977), the auto- and cross-PSD can thus be recursively estimated as

$$\hat{\Phi}_{YY}(\ell,k) = \alpha \hat{\Phi}_{YY}(\ell-1,k) + (1 - \alpha)Y^*_i(\ell,k)Y_i(\ell,k) \in \mathbb{R}, \quad (2.4.5)$$

$$\hat{\Phi}_{Y_jY_i}(\ell,k) = \alpha \hat{\Phi}_{Y_jY_i}(\ell-1,k) + (1 - \alpha)Y^*_i(\ell,k)Y_j(\ell,k) \in \mathbb{C}, \quad (2.4.6)$$

where $\alpha \in [0,1]$ is a fixed smoothing factor. The value of $\alpha$ is typically being chosen close to one (Allen et al., 1977). The value of $\alpha$ as applied in this thesis is provided later in Section 5.1.1. In the Zelinski post-filter estimation, a noise overestimation was taken by replacing $\Phi_{NN}(\ell,k)$ with $\Phi_{NN}(\ell,k)$ in (2.4.1) as

$$H_{PF}(\ell,k) = \frac{\Phi_{SS}(\ell,k)}{\Phi_{SS}(\ell,k) + \Phi_{NN}(\ell,k)}, \quad (2.4.7)$$

which neglects the noise attenuation effect of the beamformer. Please note that such a noise overestimation makes Zelinski’s post-filter theoretically sub-optimal. However, due to the deficiency of the MVDR beamformer at low frequencies (Bitzer and Simmer, 2001), the noise attenuation effect of the beamformer itself can be ignored for car noise. Inserting (2.4.3) and (2.4.4) into (2.4.7) and employing estimation with variance reduction over all possible microphone combinations, Zelinski’s post-filter can thus be formulated as (Zelinski, 1988)

$$H_{ZE}(\ell,k) = \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \text{Re}\left\{ \hat{\Phi}_{Y_jY_i}(\ell,k) \right\}}{\frac{1}{M} \sum_{i=1}^{M} \hat{\Phi}_{Y_jY_i}(\ell,k)}, \quad (2.4.8)$$

with $\text{Re}\{\cdot\}$ being the real operator, used to force the numerator to be real-valued. Please note, since it is not advisable to have $H_{ZE}(\ell,k)$ being negative, which may occur during the estimation, $H_{ZE}(\ell,k)$ shall be forced to be non-negative by

$$H_{ZE}(\ell,k) = \max\left\{ \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \text{Re}\left\{ \hat{\Phi}_{Y_jY_i}(\ell,k) \right\}}{\frac{1}{M} \sum_{i=1}^{M} \hat{\Phi}_{Y_jY_i}(\ell,k)}, H_{\min} \right\}, \quad (2.4.9)$$

with $\text{max}\{\cdot\}$ being the maximum operator and $H_{\min}$ being some lower floor. The value of $H_{\min}$ as applied in this thesis is provided later in Section 5.1.1.
2.4.2 General Designs

It can be seen that the most important assumption taken by Zelinski is the uncorrelated noise signals, which therefore simplifies the estimation without employing the noise coherence function. It will be exhibited later that by considering the noise coherence function, Zelinski’s post-filter can be regarded as a special case of a general post-filter design using the incoherent noise coherence function.

Simmer’s Post-Filter

Focusing on the noise overestimation by Zelinski’s post-filter, Simmer and Wasiljeff (1992) proposed another post-filter design by slightly modifying Zelinski’s approach in (2.4.9). Having the same assumptions as for Zelinski’s post-filter, Simmer and Wasiljeff (1992) applied the beamformer output to estimate the denominator of the post-filter. Under the assumption of uncorrelated speech and noise signals, the auto-PSD of the beamformer output \( \hat{S}_{\text{BF}}(\ell,k) \) can be expressed as (Simmer and Wasiljeff, 1992)

\[
\Phi_{\hat{S}_{\text{BF}} \hat{S}_{\text{BF}}}(\ell,k) = \Phi_{SS}(\ell,k) + \Phi_{NN_{\text{BF}}}(\ell,k) \in \mathbb{R},
\]

which is exactly the denominator of the optimal post-filter as defined in (2.4.1). Therefore, using \( \Phi_{\hat{S}_{\text{BF}} \hat{S}_{\text{BF}}}(\ell,k) \) instead of \( \frac{1}{M} \sum_{i=1}^{M} \Phi_{Y_i Y_i}(\ell,k) \) in (2.4.9), Simmer’s post-filter can be calculated as (Simmer and Wasiljeff, 1992)

\[
H_{SI}(\ell,k) = \max \left\{ \frac{\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \text{Re} \left\{ \hat{\Phi}_{Y_i Y_j}(\ell,k) \right\}}{\Phi_{\hat{S}_{\text{BF}} \hat{S}_{\text{BF}}}(\ell,k)}, H_{\text{min}} \right\},
\]

(2.4.11)

where \( \Phi_{\hat{S}_{\text{BF}} \hat{S}_{\text{BF}}}(\ell,k) \) can be recursively estimated by employing \( \hat{S}_{\text{BF}}(\ell,k) \) in (2.4.5). Comparing \( H_{SI}(\ell,k) \) (2.4.11) to \( H_{ZE}(\ell,k) \) (2.4.9), the only difference to be found is the denominator. Employing the auto-PSD of the beamformer output in the denominator, Simmer’s post-filter provides thus the theoretically optimal estimation of the Wiener post-filter for uncorrelated noise signals.

2.4.2 General Designs

The post-filter designs addressed so far only deal with uncorrelated noise signals in an incoherent noise field. However, in practice, an incoherent noise field hardly occurs and it is more interesting to have post-filters, which can be applied to different noise fields, like the diffuse noise field in a car. In the following, two generalized post-filter designs proposed by McCowan and Bourlard (2003) and Lefkimmiatis and Maragos (2006) are described, who applied the diffuse noise coherence function to their generalized post-filter designs.

McCowan’s Post-Filter

Applying the noise coherence function, McCowan and Bourlard (2003) extended Zelinski’s post-filter. However, unlike Zelinski, McCowan and Bourlard (2003) did not assume
noises among different microphones to be uncorrelated. With the *a priori* noise coherence function \( \Gamma_{i,j}(k) \) of the microphone pair \( i \) and \( j \) (see (2.1.10)), the auto-PSD \( \Phi_{Y_iY_i}(\ell,k) \), \( \Phi_{Y_jY_j}(\ell,k) \), and the cross-PSD \( \Phi_{Y_iY_j}(\ell,k) \) can be formulated as (McCowan and Bourlard, 2003)

\[
\Phi_{Y_iY_i}(\ell,k) = \Phi_{SS}(\ell,k) + \Phi_{NN}(\ell,k) \in \mathbb{R} ,
\]

(2.4.12)

\[
\Phi_{Y_jY_j}(\ell,k) = \Phi_{SS}(\ell,k) + \Phi_{NN}(\ell,k) \in \mathbb{R} ,
\]

(2.4.13)

\[
\Phi_{Y_iY_j}(\ell,k) = \Phi_{SS}(\ell,k) + \Gamma_{i,j}(k) \cdot \Phi_{NN}(\ell,k) \in \mathbb{C} .
\]

(2.4.14)

Summation of (2.4.12) and (2.4.13) leads to

\[
\Phi_{NN}(\ell,k) = \frac{1}{2} \left[ \Phi_{Y_iY_i}(\ell,k) + \Phi_{Y_jY_j}(\ell,k) \right] - \Phi_{SS}(\ell,k) .
\]

(2.4.15)

Inserting (2.4.15) into (2.4.14), \( \hat{\Phi}_{SS}^{(ij)}(\ell,k) \) can then be calculated straightforwardly as:

\[
\hat{\Phi}_{SS}^{(ij)}(\ell,k) = \max \left\{ \frac{\text{Re}\left\{ \hat{\Phi}_{Y_iY_i}(\ell,k) \right\} - \frac{1}{2} \text{Re}\left\{ \Gamma_{i,j}(k) \right\} \left[ \hat{\Phi}_{Y_iY_i}(\ell,k) + \hat{\Phi}_{Y_jY_j}(\ell,k) \right]}{1 - \text{Re}\left\{ \Gamma_{i,j}(k) \right\}}, 0 \right\} ,
\]

(2.4.16)

where \( \hat{\Phi}_{Y_iY_i}(\ell,k) \) and \( \hat{\Phi}_{Y_jY_j}(\ell,k) \) are estimated by using (2.4.5) and (2.4.6), respectively, and \( \text{Re}\{\cdot\} \) is used to force \( \hat{\Phi}_{SS}^{(ij)}(\ell,k) \) to be real-valued. However, the calculation in (2.4.16) may result in a negative value of \( \hat{\Phi}_{SS}^{(ij)}(\ell,k) \), which is theoretically not allowed for an auto-PSD. Please note that although this consideration was not mentioned in McCowan and Bourlard (2003), we consider it to enforce \( \hat{\Phi}_{SS}^{(ij)}(\ell,k) \) to be non-negative, therefore improving the performance. Please note that the real-valued diffuse noise coherence function \( \Gamma^{(ij)}_{\text{diff}}(k) \) (see (2.1.18)) is applied in (2.4.16) by McCowan and Bourlard (2003).

The same estimation step of Zelinski’s post-filter aiming at the variance reduction by averaging over all possible microphone combinations was also taken to estimate \( \hat{\Phi}_{SS}(\ell,k) \). The term \( \hat{\Phi}_{SS}(\ell,k) \) can thus be estimated as (McCowan and Bourlard, 2003)

\[
\hat{\Phi}_{SS}(\ell,k) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\Phi}_{SS}^{(ij)}(\ell,k) .
\]

(2.4.17)

Applying the same denominator as in Zelinski’s post-filter (2.4.9), McCowan’s post-filter is thus formulated by using \( \hat{\Phi}_{SS}(\ell,k) \) (2.4.17) for the numerator of the sub-optimal post-filter estimation (2.4.7) as (McCowan and Bourlard, 2003)

\[
H_{MC}(\ell,k) = \max \left\{ \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\Phi}_{SS}^{(ij)}(\ell,k), H_{\text{min}} \right\} .
\]

(2.4.18)

Therefore, McCowan’s post-filter is also theoretically sub-optimal compared to the optimal Wiener post-filter definition in (2.4.1) by neglecting the noise attenuation of the beamformer. However, applying McCowan’s post-filter, \( \hat{\Phi}_{SS}(\ell,k) \) can now be estimated.
for different noise fields by employing the noise coherence function. Actually, inserting the incoherent noise coherence function $\Gamma_{inc}^{i,j}(k) = 0$ into (2.4.16), $\hat{\Phi}_{SS}^{i,j}(\ell, k)$ will then reduce to $\text{Re}\left\{\hat{\phi}_{Y_iY_j}(\ell, k)\right\}$, which makes $H_{MC}(\ell, k)$ (2.4.18) and $H_{ZE}(\ell, k)$ (2.4.9) matching exactly. Therefore, the noise coherence function is the key component to the generalized post-filter estimation, and shall thus be chosen carefully to match the underlying noise field. In McCowan and Bourlard (2003), a significant noise attenuation improvement can be achieved by McCowan’s post-filter utilizing a diffuse noise coherence function compared to Zelinski’s post-filter for some real noise environments. The generalized post-filter estimation presented in this section thus provides a possibility to further improve the post-filter performance by employing the suitable noise coherence function.

Lefkimmiatis’s Post-Filter

In analogy to Simmer’s extension to Zelinski’s post-filter, Lefkimmiatis and Maragos (2006) have also extended McCowan’s post-filter in the same manner by imposing no noise overestimation. Taking the same assumption as McCowan’s post-filter, not only $\hat{\Phi}_{SS}^{i,j}(\ell, k)$, but also $\hat{\Phi}_{NN}^{i,j}(\ell, k)$ is solved from the equation group (2.4.12), (2.4.13), and (2.4.14) as (Lefkimmiatis and Maragos, 2006)

$$\hat{\Phi}_{NN}^{i,j}(\ell, k) = \max\left\{\frac{1}{2}\left[\hat{\Phi}_{Y_iY_i}(\ell, k) + \hat{\Phi}_{Y_jY_j}(\ell, k)\right] - \text{Re}\left\{\hat{\Phi}_{Y_iY_j}(\ell, k)\right\}, 0\right\}. \quad (2.4.19)$$

In the same manner, the real-valued diffuse noise coherence function $\Gamma_{dif}^{i,j}(k)$ is also applied by Lefkimmiatis and Maragos (2006) in (2.4.19). In analogy to averaging over all possible microphone combinations for estimation of $\hat{\Phi}_{SS}(\ell, k)$ in (2.4.17), $\hat{\Phi}_{NN}(\ell, k)$ can be estimated as (Lefkimmiatis and Maragos, 2006)

$$\hat{\Phi}_{NN}(\ell, k) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\Phi}_{NN}^{i,j}(\ell, k). \quad (2.4.20)$$

The auto-PSD of the noise signal after the beamformer can then be calculated according to (2.2.17) as (Lefkimmiatis and Maragos, 2006)

$$\hat{\Phi}_{NBF,NBF}(\ell, k) = \hat{\Phi}_{NN}(\ell, k)W^H_{MVDR}(\ell, k)\Gamma_{NN}(k)W_{MVDR}(\ell, k)$$

$$= \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\Phi}_{NN}^{i,j}(\ell, k)\left[W^H_{MVDR}(\ell, k)\Gamma_{NN}(k)W_{MVDR}(\ell, k)\right]. \quad (2.4.21)$$

Once $\hat{\Phi}_{NBF,NBF}(\ell, k)$ has been computed, the denominator of the optimal Wiener post-filter $\Phi_{SS}(\ell, k) + \Phi_{NBF,NBF}(\ell, k)$ in (2.4.1) can then be calculated by applying $\hat{\Phi}_{SS}(\ell, k)$ from (2.4.17). Lefkimmiatis’s post-filter can thus be formulated as (Lefkimmiatis and Maragos,
2. Basic Principles of Beamforming and Post-Filtering

\[ H_{LE}(\ell, k) = \max \left\{ \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\Phi}_{SS}^{(i,j)}(\ell, k) \right\}, \quad (2.4.22) \]

where \( \hat{\Phi}_{NBF-NBF}(\ell, k) \) and \( \hat{\Phi}_{SS}^{(i,j)}(\ell, k) \) are estimated according to (2.4.21) and (2.4.16), respectively.

Comparing \( H_{LE}(\ell, k) \) (2.4.22) to \( H_{MC}(\ell, k) \) (2.4.18), the only difference is the denominator, where Lefkimmiatis’s post-filter applies the estimated auto-PSD of the noise signal after the beamformer. By considering the noise attenuation effect of the beamformer, Lefkimmiatis’s post-filter thus provides a theoretically optimal post-filter estimation. On the other hand, Lefkimmiatis’s post-filter has also extended Simmer’s post-filter by applying the diffuse noise coherence function.

2.5 Summary

In this chapter, the basic principles of beamforming and post-filtering have been briefly addressed. For a better understanding of multichannel speech enhancement applying beamforming and post-filtering, spatial characteristics of four noise fields have been described by utilizing the noise coherence function.

The multichannel MMSE estimation is then explained in the discrete Fourier transform domain. Furthermore, it has been shown that the multichannel MMSE estimation can be decomposed into an MVDR beamformer and a Wiener post-filter. Different microphone array designs have been presented with different design layouts along with the beamforming instrumental performance measures. With a focus on LCMV beamforming, to which the MVDR beamforming is a special case and the GSC beamforming is an alternative realization, all three beamforming techniques have been explained. Furthermore, some other beamforming techniques existing in the literature have also been shortly introduced. For post-filtering techniques, four state-of-the-art post-filter designs have been derived, which can be categorized into special designs for the incoherent noise field and general designs based on the noise coherence function. Inside each category, both theoretically sub-optimal and optimal post-filter designs have been pointed out.
Chapter 3

Instrumental Evaluation Framework

In this chapter, we address instrumental quality assessment of noise reduction systems, which is preferred in practice compared to subjective quality assessment due to its low costs and efforts. Section 3.1 describes a unified instrumental evaluation framework in its single-channel and multichannel setups. Based on this framework, several state-of-the-art speech-related instrumental quality measures are described in Section 3.2. Noise-related instrumental quality measures are then introduced in Section 3.3. Particularly, in Section 3.3.5, a new weighted log kurtosis ratio measure is specified for instrumentally measuring noise distortion in a black-box manner, especially suited to report on musical tones. In Section 3.4, a new instrumental optimization method defined as a Figure of Merit (FoM) is presented. We define the FoM as an entity of three independent instrumental quality measures, which can be applied to instrumentally optimize different parameters of speech enhancement systems. To illustrate such a capability, Section 3.5 shows the FoM-based instrumental optimization of two key parameters of the a priori SNR estimation for single-channel noise reduction algorithms.

3.1 Processing Framework

Instrumental quality assessment of a noise reduction algorithm concentrates mainly on two aspects: The speech- and the noise-related qualities of the enhanced speech signal. Therefore, it is advantageous to isolate the speech and the noise components from the noisy input speech signal and the enhanced output speech signal. In science, the noisy speech signal $y(n) = s(n) + n(n)$ is conventionally constructed by digitally adding the clean speech signal $s(n)$ and the background noise signal $n(n)$, which makes the separation trivial at the input side. In this thesis, a DFT-based noise reduction scheme employing a spectral weighting rule is utilized. The spectral gains adapted on the basis of the noisy speech signal are logged and then separately applied to the clean speech signal and the noise signal, respectively, which yields the filtered clean speech signal and the filtered noise signal at the output side. Throughout the thesis, the applied parameters including the DFT length, the type of the analysis and the synthesis windows, and the frame shift for narrowband and wideband signals are listed in Table 3.1. In this thesis,
narrowband signals are obtained by applying the modified intermediate reference system (MIRS) weighting filter according to ITU-T Rec. P.830 (ITU-T, 1996b) to achieve the bandwidth of 300…3400 Hz, while wideband signals are pre-processed by applying the P.341 weighting filter according to ITU-T Rec. P.341 (ITU-T, 1995) to achieve the bandwidth of 50…7000 Hz. In the following, the unified instrumental evaluation framework is explained in its single-channel and multichannel setups.

### 3.1.1 Single-Channel Setup

The block diagram of the single-channel setup is depicted in Fig. 3.1. Throughout the thesis, the following pre-processing is applied to $s(n)$ and $n(n)$: The clean speech signal $s(n)$ is pre-normalized to –26 dBov (decibel overload) according to ITU-T Rec. P.56 (ITU-T, 1993). The noise signal $n(n)$ is scaled to different levels according to the same ITU-T Rec. P.56, so that different input SNR ($\text{SNR}_{\text{in}}$) conditions can be generated.

![Figure 3.1: Block diagram of the instrumental evaluation framework for a single-channel noise reduction algorithm](image)

After the analysis window and DFT, we get the noisy signal $Y(\ell,k)$, the clean speech signal $S(\ell,k)$, and the noise signal $N(\ell,k)$ separately. The spectral gain $G(\ell,k)$ of the
weighting rule is then adaptively computed based on the noisy speech signal \( Y(\ell, k) \). Applying \( G(\ell, k) \), the enhanced speech signal \( \hat{S}(\ell, k) \) can be calculated as

\[
\hat{S}(\ell, k) = G(\ell, k) \cdot Y(\ell, k) .
\] (3.1.1)

In the meantime, the spectral gain \( G(\ell, k) \) is logged and applied separately to \( S(\ell, k) \) and \( N(\ell, k) \)

\[
\tilde{S}(\ell, k) = G(\ell, k) \cdot S(\ell, k) ,
\] (3.1.2)
\[
\tilde{N}(\ell, k) = G(\ell, k) \cdot N(\ell, k) ,
\] (3.1.3)

where \( \tilde{S}(\ell, k) \) is the filtered speech signal and \( \tilde{N}(\ell, k) \) is the filtered noise signal, respectively. Due to the linear properties of DFT (Proakis and Dimitris, 2007), we get the following relationship

\[
Y(\ell, k) = S(\ell, k) + N(\ell, k) .
\] (3.1.4)

Multiplying both sides of (3.1.4) with \( G(\ell, k) \) and inserting (3.1.1), (3.1.2), and (3.1.3) into (3.1.4), we achieve

\[
G(\ell, k) \cdot Y(\ell, k) = G(\ell, k) \cdot S(\ell, k) + G(\ell, k) \cdot N(\ell, k) ,
\]
\[
\hat{S}(\ell, k) = \tilde{S}(\ell, k) + \tilde{N}(\ell, k) .
\] (3.1.5)

It can be seen that the enhanced speech signal \( \hat{S}(\ell, k) \) can thus be separated into the filtered speech signal \( \tilde{S}(\ell, k) \) and the filtered noise signal \( \tilde{N}(\ell, k) \). Finally, by applying IDFT, the synthesis window, and the overlap-and-add method, the enhanced speech signal \( \hat{s}(n) = \tilde{s}(n) + \tilde{n}(n) \) can be obtained in the time domain. Having \( s(n) \) and \( n(n) \) on the input side as well as \( \tilde{s}(n) \) and \( \tilde{n}(n) \) on the output side, different speech- and noise-related instrumental quality measures can be employed.

### 3.1.2 Multichannel Setup

We now extend the single-channel instrumental evaluation setup to the multichannel instrumental evaluation setup (Yu and Fingscheidt, 2009), which is shown in Fig. 3.2. Like the single-channel setup, the noisy signal \( y_i'(n) = s_i'(n) + n_i'(n) \) with \( s_i'(n) \) being the clean speech signal, \( n_i'(n) \) being the noise signal in channel \( i = 1, ..., M \), respectively, are the multichannel inputs to the speech enhancement system comprising a fixed beamformer and an adaptive post-filter. Applying ITU-T Rec. P.56 (ITU-T, 1993), the clean speech signal \( s_i'(n) \) is normalized to \(-26\) dBov and the noise signal \( n_i'(n) \) is scaled to different levels, so that different SNR\(\text{in}\) conditions can be generated for the beamformer and post-filter system.

The beamformer and post-filter system shown in Fig. 3.2 operates in the frequency domain with the beamformer coefficients vector \( W(k) \) and the post-filter coefficients \( H_{PF}(\ell, k) \). The noisy signal \( Y'(\ell, k) \) with \( Y'(\ell, k) = S'(\ell, k) + N'(\ell, k) \) is employed for the adaptation of \( H_{PF}(\ell, k) \). The adapted post-filter coefficients and the beamformer coefficients are then logged during the operation and applied to \( S'(\ell, k) \) and \( N'(\ell, k) \) separately.
Due to the linearity of DFT, we can thus achieve the relationship
\[
\hat{S}(\ell, k) = W^H(k) \cdot Y'(\ell, k) \cdot H_{PF}(\ell, k) = W^H(k) \cdot [S'(\ell, k) + N'(\ell, k)] \cdot H_{PF}(\ell, k) = W^H(k) \cdot S'(\ell, k) \cdot H_{PF}(\ell, k) + W^H(k) \cdot N'(\ell, k) \cdot H_{PF}(\ell, k) = \hat{S}(\ell, k) + \hat{N}(\ell, k),
\]
where \(\hat{S}(\ell, k)\) is the filtered speech signal and \(\hat{N}(\ell, k)\) is the filtered noise signal at the output of the beamformer and post-filter system. This results into the separation of the enhanced speech signal \(\hat{s}(n)\) into the filtered speech signal \(\hat{s}(n)\) and the filtered noise signal \(\hat{n}(n)\) as \(\hat{s}(n) = \hat{s}(n) + \hat{n}(n)\) in the time domain. However, speech- and noise-related instrumental quality measures usually require a single-channel clean speech signal \(s(n)\) and a single-channel noise signal \(n(n)\) as reference signals at the input side. Therefore, within the multichannel setup, the best clean speech signal from \(s'_i(n)\) with its corresponding noise signal \(n'_i(n)\) are then chosen as the single-channel reference signals.

### 3.2 Speech-Related Instrumental Quality Measures

In this section, several state-of-the-art speech-related instrumental quality measures based on the noisy speech signal \(y(n)\), the clean speech signal \(s(n)\), the enhanced speech signal \(\hat{s}(n)\), and the filtered clean speech signal \(\tilde{s}(n)\) are described. The main focus lies on the speech component quality measure utilizing the perceptual evaluation of speech quality (PESQ) in Section 3.2.1, while other speech-related instrumental quality measures are introduced in Section 3.2.1.
3.2.1 Perceptual Evaluation of Speech Quality (PESQ)

In the area of speech communication, speech quality is commonly assessed as the mean opinion score (MOS) defined by ITU-T Rec. P.10 (ITU-T, 2006) and Rec. P.800.1 (ITU-T, 1996a): The mean of opinion scores, i.e., of the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material. For a subjective listening test, such an MOS is defined as the MOS listening quality subjective (MOSLQS) using a scale of five speech quality categories: 1 (bad), 2 (poor), 3 (fair), 4 (good), and 5 (excellent). MOSLQS can be applied, e.g., to an absolute category rating (ACR) test, so that each listener can judge each sample by giving a suitable MOSLQS score.

However, performing a subjective listening test commonly costs time and efforts. Therefore, an instrumental measure is desired to yield MOS scores correlating well to the MOSLQS scores. An instrumental speech quality measure defined as the perceptual evaluation of speech quality (PESQ) has been defined in ITU-T Rec. P.862 (ITU-T, 2001) as a replacement to the previously specified perceptual speech quality measure (PSQM) in ITU-T Rec. P.861 (ITU-T, 1998). PESQ requires a reference signal (e.g., the clean speech signal) and a degraded signal (e.g., a processed speech signal after a speech codec). PESQ starts by pre-processing through time alignment and level alignment to the standard telephony listening level. The pre-processed signals are then separately transformed via a perceptual model, which is carried out with an auditory transformation similar to the human auditory system. The auditory transformation includes time-to-frequency transformation, Bark transformation, and the loudness transformation in a Sone scale employing Zwicker’s law (Zwicker and Fastl, 1999). After the auditory transformation, both signals are thus being transformed into their internal signal representations, which are similar to their psychoacoustical representations. Consequently, the differences between these two signals are computed as disturbances. The disturbances along with the internal representation of the two signals are fed into a cognitive model, which includes several non-linear computations, e.g., the time-frequency masking of inaudible disturbances, weighing of the additional frequency components introduced by signal processing, and non-linear averaging of disturbances over time and frequency. All computed values after the cognitive model are linearly combined to yield the raw MOS score (MOSraw) ranging from ~0.5 to 4.5. By employing the perceptual and cognitive models, PESQ can thus predict the perceptual speech quality in a reliable manner. To enable a direct comparison to MOSLQS with high correlation, the MOS listening quality objective (MOSLQO) for narrowband signals is computed by applying a mapping function (ITU Rec. P.862.1, ITU-T, 2003) as

\[
\text{MOSLQO} = 0.999 + \frac{4}{1 + \exp(-1.4945 \cdot \text{MOS}_{\text{raw}} + 4.6607)},
\]

which makes MOSLQO ∈ [1.02, 4.56] for narrowband signals. PESQ was originally designed to instrumentally measure narrowband speech quality. A wideband extension of PESQ was provided for instrumentally assessing wideband speech (50...7000 Hz) quality.
in ITU Rec. P.862.2 (ITU-T, 2005b), which already includes a respective mapping function

\[ \text{MOS}_{\text{LQO}} = 0.999 + \frac{4}{1 + \exp(-1.3669 \cdot \text{MOS}_{\text{raw}} + 3.8224)} \]  

(3.2.2)

Applying (3.2.2), we achieve \( \text{MOS}_{\text{LQO}} \in [1.04, 4.64] \) for wideband signals.

Conventionally, PESQ is employed to measure the coded speech quality by employing a clean speech signal as the reference and a coded speech signal without background noise as the degraded signal. PESQ has not been validated in its original design in ITU-T Rec. P.862 (ITU-T, 2001) for noise reduction algorithms. Later, it has been stated in ITU-T Rec. P.862.3 (ITU-T, 2007), that PESQ is not suitable for measuring the speech quality by using the noisy speech signal \( y(n) \) and the enhanced speech signal \( \hat{s}(n) \). The problem is that the perceptual and cognitive models of PESQ were designed in the first place without considering the background noise. Therefore, in this thesis, the speech component quality of a noise reduction algorithm is measured by means of PESQ as

\[ \text{MOS}_{\text{LQO}}^{\hat{s}} = \frac{\text{PESQ}(s(n), \hat{s}(n))}{\text{PESQ}(s(n), \hat{s}(n))} \]  

(3.2.3)

where the clean speech signal \( s(n) \) is used as the reference and the filtered clean speech signal \( \hat{s}(n) \) is the degraded signal. Utilizing \( s(n) \) and \( \hat{s}(n) \) as the inputs to the PESQ computation, \( \text{MOS}_{\text{LQO}}^{\hat{s}} \) indicates then the speech component quality without any background noise influences. The filtered clean speech signal can be interpreted as a processed speech component after a noise reduction algorithm similar to a coded speech signal after a speech codec without any background noise. The degree of underlying speech distortion, i.e., the speech component quality, can thus be well measured by using \( \text{MOS}_{\text{LQO}}^{\hat{s}} \) without constraints. Accordingly, a high \( \text{MOS}_{\text{LQO}}^{\hat{s}} \) score indicates a good speech component quality, while a low \( \text{MOS}_{\text{LQO}}^{\hat{s}} \) score indicates a bad speech component quality.

### 3.2.2 Segmental Speech-to-Speech Distortion Ratio (SSDR)

Gustafsson (1999) proposed the segmental speech-to-speech distortion ratio (SSDR), which measures the speech component quality by calculating the segmental difference between the clean speech signal \( s(n) \) and the filtered clean speech signal \( \hat{s}(n) \) in the time domain. The SSDR is thus computed for each segment \( \lambda_s \) as (Fingscheidt and Suhadi, 2006)

\[ \text{SSDR}(\lambda_s) = 10 \log \left[ \frac{\sum_{n=0}^{N-1} s^2(n + \lambda_s N)}{\sum_{n=0}^{N-1} (\hat{s}'(n + \lambda_s N) - s(n + \lambda_s N))^2} \right] \]  

(3.2.4)

where \( \hat{s}'(n) \) is the delay-compensated filtered speech signal and \( N \) is the segment length, respectively. Time alignment between \( s(n) \) and \( \hat{s}(n) \) is very important, otherwise a meaningful direct comparison of each time index cannot be performed. During the measurement, the analysis window length \( L \) usually equals to the segment length \( N \) with \( L = N \), which means that the segment index \( \lambda_s \) equals to the frame index \( \ell \). The segmental
SSDR can then be computed as the mean value of $\text{SSDR}(\lambda_s)$ over selected segments as (Fingscheidt and Suhadi, 2006)

$$\text{SSDR}_{\text{seg}} = \frac{1}{C(\Lambda_{H_1})} \sum_{\lambda_s \in \Lambda_{H_1}} \min\{\text{SSDR}(\lambda_s), \text{SSDR}_{\text{max}}\}, \quad (3.2.5)$$

where $C(\Lambda_{H_1})$ is the number of the set $\Lambda_{H_1}$ including the speech-active segments only and having $\text{SSDR}(\lambda_s) > -10$ dB (Suhadi, 2012). The detection of active speech is achieved by comparing the short-term mean amplitude of each segment of speech to a threshold computed by the active speech level computation according to the ITU-T Rec. P.56 (ITU-T, 1993). A segment is then categorized as speech-active if its short-term mean amplitude is greater than the computed threshold. Furthermore, $\text{SSDR}_{\text{seg}}$ shall be limited to a maximum value of $\text{SSDR}_{\text{max}} = 30$ dB (Suhadi, 2012).

According to the segmental SSDR computation using (3.2.4) and (3.2.5), a high segmental SSDR value represents a small difference between the clean speech signal and the filtered clean speech signal. Therefore, a high segmental SSDR indicates a good preservation of the speech component, while a low segmental SSDR indicates a bad preservation of the speech component.

### 3.2.3 Speech Mean Opinion Score (S-MOS)

In the ETSI EG Recommendation 202.396-3 (ETSI, 2008b), a speech mean opinion score (S-MOS) is proposed to measure the speech quality. Similar to the speech component quality measure $\text{MOS}_{\text{LQO}}$ employing PESQ, S-MOS utilizes also a psychoacoustical model originally referred to as relative approach proposed by Genuit (1996). According to Genuit (1996), human hearing is sensitive to sounds with fast varying temporal and spectral patterns, which can be considered as distinct patterns. Generally, a human does not need any reference signal to judge such distinct patterns. Therefore, the relative approach assumes that human hearing keeps on building adaptive references employing the previous heard sounds to judge the new coming sounds for distinct patterns. Motivated by this assumption, the relative approach thus needs no reference signal and uses only the signal under test.

In order to detect the distinct patterns in time and frequency, the signal under test is firstly transformed into the frequency domain by a filter bank analysis, typically with 1/12 octave according to ETSI (2008b) or by using the ear-related filter bank analysis as part of the hearing model proposed by Sottek and Genuit (2005). In the temporal-spectral domain, the signal is separately processed to build two adaptive reference signals for detecting spectral tonal components and temporal transient structures, respectively. For the detection of spectral tonal components, a regression over time for each frequency band is applied to estimate the human expectation of each frequency band in the next frame. Subsequently, within each frame a smoothing over frequency is carried out. A non-linear transformation according to Sottek’s hearing model (Sottek and Genuit, 2005) is applied afterwards. The output signal $R_{tc}(\ell, k)$ is then considered as the adaptive
reference signal for detecting spectral tonal components. For the detection of temporal transient structures, a smoothing over frequency is firstly performed within each frame. A regression over time for each frequency band is performed afterwards. Furthermore, a non-linear transformation according to Sottek’s hearing model is also performed. The output signal \( R_{ts}(\ell, k) \) is then treated as the adaptive reference signal for detecting temporal transient structures. Both adaptive reference signals can be interpreted as a running mean value of previous signal frames, which can be treated as the human reference signal for comparing the coming signal frame. Furthermore, the signal under test is also non-linearly transformed using Sottek’s hearing model as \( R(\ell, k) \). Sequentially, the differences \( |R(\ell, k) - R_{ts}(\ell, k)| \) and \( |R(\ell, k) - R_{tc}(\ell, k)| \) are separately computed. Finally, the result of the relative approach for the signal under test \( RA(\ell, k) \) is computed by combining these two differences with two adjustable weights, so that the weights of spectral tonal components and temporal transient structures can be differently set. For calculating S-MOS, only the temporal transient structures are considered in ETSI (2008b).

However, the relative approach can only measure how close a signal is to the human expectation of this signal. The human expectation of such a signal can yet be either good or bad. Therefore, using the relative approach alone is not sufficient to measure the speech quality. It needs a reference signal, which can be expected as a good signal having expectation of this signal. The human expectation of such a signal can yet be either good or bad. Therefore, using the relative approach alone is not sufficient to measure the speech quality. It needs a reference signal, which can be expected as a good signal having

\[
\Delta RA_{p,ref}(\ell, k) = RA_p(\ell, k) - RA_{ref}(\ell, k),
\]

which indicates how close the processed signal is to the reference signal. Therefore, \( \Delta RA_{p,ref}(\ell, k) \) can be applied to measure the speech quality of the processed signal. Applying S-MOS based on the relative approach to measure the speech quality, three signals are employed (ETSI, 2008b): The input noisy signal \( y(n) \), the enhanced speech signal \( \hat{s}(n) \), and the clean speech signal \( s(n) \). The \( RA(\ell, k) \) values are then separately computed for \( y(n) \), \( \hat{s}(n) \), and \( s(n) \) as \( RA_y(\ell, k), RA_{\hat{s}}(\ell, k), \) and \( RA_s(\ell, k) \), respectively. Subsequently, two RA differences are computed as (ETSI, 2008b)

\[
\Delta RA_{\hat{s},s}(\ell, k) = RA_{\hat{s}}(\ell, k) - RA_s(\ell, k), \quad \text{for } \ell \in \Lambda_{H_1}
\]

\[
\Delta RA_{\hat{s},y}(\ell, k) = RA_{\hat{s}}(\ell, k) - RA_y(\ell, k), \quad \text{for } \ell \in \Lambda_{H_1}
\]

where \( \Lambda_{H_1} \) is the set of speech-active frames, which are extracted by using a mask based on the clean speech signal \( s(n) \). Based on \( \Delta RA_{\hat{s},s}(\ell, k) \), the mean \( \mu_{\Delta RA_{\hat{s},s}} \) and the variance \( \sigma^2_{\Delta RA_{\hat{s},s}} \) are computed over time and frequency. In the same manner, the mean \( \mu_{\Delta RA_{\hat{s},y}} \) and the variance \( \sigma^2_{\Delta RA_{\hat{s},y}} \) are calculated for \( \Delta RA_{\hat{s},y}(\ell, k) \) over time and frequency. The mean \( \mu_s \) is computed for \( RA_s(\ell, k) \) employing only speech-active frames. Furthermore, a separate measure NA for the noise attenuation level is computed based on the three input signals. For details of NA, readers are referred to ETSI (2008b). The S-MOS score can then be computed by applying a linear quadratic regression using the six values \( (\mu_{\Delta RA_{\hat{s},s}}, \mu_{\Delta RA_{\hat{s},y}}, \mu_s, \sigma^2_{\Delta RA_{\hat{s},s}}, \sigma^2_{\Delta RA_{\hat{s},y}}, \text{NA}) \) as (ETSI, 2008b)

\[
S\text{-MOS}_i = c_0^{(i)} + \sum_{p=1}^{6} c_{1,p}^{(i)} \cdot V_p + \sum_{p=1}^{6} c_{2,p}^{(i)} \cdot V_p^2, \quad i = 1, 2, 3.
\]
where \( V_p \in \{ \mu_{\Delta R A_{s,y}}, \mu_{\Delta R A_{s,s}}, \mu_{\Delta s}, \sigma^2_{\Delta R A_{s,y}}, \sigma^2_{\Delta R A_{s,s}}, \text{NA} \} \), \( \sigma^{(i)}_{1,p} \), are the first order regression coefficients for \( V_p \) within the coefficient set \( i \), and \( \sigma^{(i)}_{2,p} \) are the second order regression coefficients for \( V^2_p \) within the coefficient set \( i \). According to ETSI (2008b), three coefficient sets are obtained for different noise qualities (including noise distortion and absolute noise level). The values of the three different coefficient sets can be found in ETSI (2008b). In order to obtain S-MOS, a coefficient set has to be firstly chosen based on the noise quality. For each coefficient set, the coefficients are computed in the way that the S-MOS score shows a good correlation to the subjective MOS score within the corresponding noise quality (ETSI, 2008b). Therefore, a high S-MOS score indicates a good speech quality, while a low S-MOS score indicates a low speech quality. Please note that unlike MOS\(_{\text{LQO}}\) in (3.2.3) depending only on the speech component quality, S-MOS measurements depend also on the level of noise attenuation and the noise quality. Since the computation of S-MOS in (3.2.9) employs the noise attenuation measurement as one of the parameters. Furthermore, different coefficient sets shall be applied to different noise qualities. Therefore, S-MOS depends not only on the speech quality, but also on the level of noise attenuation as well as the noise quality.

3.3 Noise-Related Instrumental Quality Measures

In this section, noise-related instrumental quality measures are addressed. The noise mean opinion score (N-MOS) employing the relative approach is described in Section 3.3.1. N-MOS, however, evaluates noise attenuation and noise distortion in a single measure. In contrast to that, the signal-to-noise ratio improvement for evaluating noise attenuation only and the new black-box weighted log kurtosis ratio for evaluating noise distortion (especially for musical tones) only are presented in Section 3.3.2 and Section 3.3.5, respectively.

3.3.1 Noise Mean Opinion Score (N-MOS)

Similar to the S-MOS, noise mean opinion score (N-MOS) was proposed in ETSI EG Recommendation 202.396-3 (ETSI, 2008b) for instrumentally measuring noise quality by employing the relative approach (Genuit, 1996). Applying the relative approach, two values \( RA_y(\ell,k) \), \( RA_s(\ell,k) \) are separately computed for \( y(n) \) and \( s(n) \), respectively (ETSI, 2008b). Unlike the S-MOS computation, the clean speech signal \( s(n) \) is employed now to extract speech pause frames (containing only noise parts) for computing the difference between \( RA_y(\ell,k) \) and \( RA_s(\ell,k) \) as (ETSI, 2008b)

\[
\Delta RA_{s,y}(\ell,k) = RA_s(\ell,k) - RA_y(\ell,k), \quad \ell \in \Lambda_{H_0},
\]

(3.3.1)

where \( \Lambda_{H_0} \) is the set of speech pause frames. Based on \( \Delta RA_{s,y}(\ell,k) \), the mean \( \mu_{\Delta RA_{s,y}} \) and the variance \( \sigma^2_{\Delta RA_{s,y}} \) are computed over time and frequency. Similarly, the mean \( \mu_{RA_s} \) and the variance \( \sigma^2_{RA_s} \) are calculated for \( RA_s(\ell,k) \) over time and frequency. The variance \( \sigma^2_{RA_y} \) is calculated for \( RA_y(\ell,k) \) over time and frequency. Furthermore, the residual noise
level NL is computed through speech pause frames of the enhanced speech signal \( \hat{s}(n) \). For computing the final N-MOS score, all six values \( \mu_{\Delta RA_{\hat{s},y}}, \mu_{RA_{\hat{s}}}, \sigma^2_{\Delta RA_{\hat{s},y}}, \sigma^2_{RA_{\hat{s}}}, \sigma^2_{RA_y}, NL \) are applied to a linear quadratic regression as (ETSI, 2008b)

\[
N\text{-MOS} = b_0 + \sum_{p=1}^{6} b_{1,p} \cdot U_p + \sum_{p=1}^{6} b_{2,p} \cdot U_p^2 ,
\]

where \( U_p \in \{ \mu_{RA_{\hat{s}}}, \mu_{RA_{\hat{s},y}}, \sigma^2_{RA_{\hat{s}}}, \sigma^2_{RA_{\hat{s},y}}, \sigma^2_{RA_y}, NL \} \), \( b_{1,p} \) are the first order regression coefficients for \( U_p \), and \( b_{2,p} \) are the second order regression coefficients for \( U_p^2 \). Unlike S-MOS computation in (3.2.9), only one coefficient set is trained, which leads the instrumental N-MOS to show a high correlation to the subjective MOS score. The values of the coefficients can be found in ETSI (2008b). Therefore, a high N-MOS score indicates a high noise quality, while a low N-MOS score indicates a low noise quality. Please note that according to ETSI (2008b), noise quality includes both the level of noise attenuation and noise distortion (i.e., the amount of musical tones). Therefore, N-MOS was designed to measure the noise quality including the level of noise attenuation and the amount of musical tones. The compact N-MOS measurement may become ambiguous in practice, e.g., a certain N-MOS score can imply a low level of noise attenuation, a large amount of musical tones, or both of them simultaneously. Therefore, it is desirable to have separate instrumental measures for the level of noise attenuation and the amount of musical tones.

### 3.3.2 Signal-to-Noise Ratio Improvement

The signal-to-noise ratio improvement (\( \Delta SNR \)) is designed to instrumentally measure the effective noise attenuation level. If both noise and speech components are attenuated, this measure shall only capture the effective relative improvement of the speech level against the noise level. Applying the clean speech signal \( s(n) \) and the noise signal \( n(n) \), the input SNR (\( SNR_{in} \)) can be calculated as

\[
SNR_{in} = ASL_x - ASL_n^{(RMS)} ,
\]

where \( ASL_x \) is the active speech level of \( x(n) \) in dB. The active speech level measurement is carried out by using ITU-T Rec. P.56 (ITU-T, 1993). The active speech level \( ASL_x \) is measured for the clean speech signal only during active speech, while the root mean square (RMS) level \( ASL_n^{(RMS)} \) is measured for the noise signal for the whole signal employing the RMS option of the active speech level measurement. In analogy, the output SNR (\( SNR_{out} \)) can be computed with the filtered speech signal \( \hat{s}(n) \) and the filtered noise signal \( \hat{n}(n) \) as

\[
SNR_{out} = ASL_{\hat{s}} - ASL_{\hat{n}}^{(RMS)} ,
\]

with \( ASL_{\hat{s}} \) being the active speech level of \( \hat{s}(n) \) and \( ASL_{\hat{n}}^{(RMS)} \) being the RMS level of \( \hat{n}(n) \), respectively. The term \( \Delta SNR \) representing the relative SNR improvement in dB is subsequently determined as

\[
\Delta SNR = SNR_{out} - SNR_{in} .
\]

Thanks to the definition of \( \Delta SNR \), a high \( \Delta SNR \) value indicates a good noise attenuation performance, while a low \( \Delta SNR \) value indicates a poor noise attenuation performance.
3.3.3 White-Box Log Kurtosis Ratio

In the assessment of noise distortion, most of the attention is devoted to a special noise artifact referred to as musical tones. Musical tones, or sometimes called musical noise (Hänsler and Schmidt, 2004, Vary and Martin, 2006), can be described as randomly sounding fluctuations inside the enhanced speech signal. The cause of musical tones can be shortly described as follows: Due to estimation errors of the weighting rule, some isolated peaks can occur in random frequencies and frames in the filtered noise signal of the enhanced speech signal. In the time domain such isolated peaks result in fluctuating sounds, which are short sinusoidal components having random frequencies and appearing and disappearing rapidly within each frame. Though musical tones can be even more disturbing than the undistorted and unattenuated noise signal for listeners (Loizou, 2007), there are only few instrumental musical tones measurements. The amount of musical tones is often still being subjectively evaluated.

Recently, a high correlation of the perceived amount of musical tones with the log kurtosis ratio has been reported by Uemura et al. (2008), evaluating the spectral subtraction approach to noise reduction. In higher-order statistics (Abramowitz and Stegun, 1965), kurtosis measures the shape of the probability density function (PDF) of a random variable. A PDF with a high kurtosis value shows a shape with sharp peak and long tail. An absolute kurtosis value cannot indicate musical tones, since signals with different PDFs can have different kurtosis values, e.g., a speech signal without musical tones shows usually a high kurtosis value. However, the kurtosis ratio between the unprocessed signal and its corresponding processed signal can predict musical tones. In comparison to the PDF of the unprocessed signal, the processed signal with musical tones owns many isolated peaks, which changes the shape of the PDF showing a modified kurtosis value. Therefore, instead of an absolute kurtosis value, a kurtosis ratio has to be applied to show the kurtosis difference originated from the generated isolated peaks (i.e., musical tones). In Uemura et al. (2008), the spectral subtraction approach has been investigated, which is given as

\[
\hat{S}(\ell, k) = \sqrt{|Y(\ell, k)|^2 - v_c \cdot \hat{\Phi}_{NN}(\ell, k) \cdot \exp(j\arg(Y(\ell, k)))},
\]

(3.3.6)

where \(\hat{\Phi}_{NN}(\ell, k)\) is the noise PSD being estimated only in the speech pause using a VAD, \(v_c\) is a subtraction coefficient and \(\arg(Y(\ell, k))\) represents the phase of \(Y(\ell, k)\), respectively. Parameter \(v_c\) controls how much of the estimated noise PSD is subtracted from the noisy signal PSD. The log kurtosis ratio is calculated as the log ratio between the kurtosis of \(|Y(\ell, k)|^2\) and the kurtosis of \(|\hat{S}(\ell, k)|^2\). Unlike the strict definition of kurtosis using the central moment (Abramowitz and Stegun, 1965), Uemura et al. (2008) calculated the kurtosis of a random variable \(x\) using the raw moment as

\[
\Psi_x' = \frac{\text{E}\{x^4\}}{\text{E}\{x^2\}^2},
\]

(3.3.7)

where \(\text{E}\{x^4\}\) and \(\text{E}\{x^2\}\) are the fourth and the second raw moments, respectively. According to Abramowitz and Stegun (1965), the \(n^{th}\)-order raw moment of \(x\) can be computed
as
\[ E\{x^n\} = \int_{-\infty}^{\infty} x^n p(x) dx , \]  
where \( p(x) \) is the PDF of \( x \). In order to calculate the kurtosis, Uemura et al. (2008) assumed the squared speech and noise spectral amplitudes both to be gamma-distributed. Utilizing the properties of the gamma distribution, the kurtosis of \( |Y(\ell, k)|^2 \) can then be estimated as (Uemura et al., 2008)

\[ \Psi_y' = \frac{(\chi + 2)(\chi + 3)}{\chi(\chi + 1)} , \]  
where \( \chi \) can be calculated as

\[ \chi = \frac{3 - \zeta + \sqrt{(\zeta - 3)^2 + 24\zeta}}{12\zeta} , \]  
with \( \zeta = \log(\mathbb{E}\{|Y(\ell, k)|^2\}) - \mathbb{E}\{\log(|Y(\ell, k)|^2)\} \). Employing the gamma distribution and the knowledge of the subtraction coefficient \( v_c \), the lower bound of the kurtosis of the enhanced speech signal \( \hat{S}(\ell, k)^2 \) can be derived as (Uemura et al., 2008)

\[ \Psi_{\hat{s}}' \geq \exp(v_c\chi)\frac{(\chi + 2)(\chi + 3) + v_c\chi(\chi + 2)(\chi - 1) + \frac{(v_c\chi)^2}{2}(\chi - 3)(\chi - 1)}{\chi(\chi + 1)} . \]  
Applying the lower bound of (3.3.11) and (3.3.9), the log kurtosis ratio can then be computed as (Uemura et al., 2008)

\[ \Delta \Psi_{\log}' = \log \left( \frac{\Psi_y'}{\Psi_{\hat{s}}'} \right) = v_c\chi \cdot \log \left( 1 + \frac{v_c\chi(\chi - 1)}{(\chi + 3)} + \frac{(v_c\chi)^2(\chi - 3)(\chi - 1)}{2(\chi + 2)(\chi + 3)} \right) . \]  
Please note that the log kurtosis ratio calculation in (3.3.12) employs the kurtosis of the enhanced speech signal \( \Psi_{\hat{s}}' \), which makes this log kurtosis ratio being dependent on speech distortion and the residual noise component. Later, the log kurtosis ratio has been further extended by Uemura et al. (2009) and Inoue et al. (2011), who applied the unprocessed noise signal and the filtered noise signal to the calculation of the log kurtosis ratio. However, all calculation approaches in (Uemura et al., 2008, 2009) and (Inoue et al., 2011) need internal access to the noise reduction algorithm, i.e., the subtraction coefficient \( v_c \) for spectral subtraction (Uemura et al., 2008, 2009) and for the Wiener filter family (Inoue et al., 2011), which is implemented in a spectral subtraction-like manner. In this thesis, such instrumental measurement approaches, which mandate a specific weighting rule and the knowledge of internal variables, are defined as white-box measurements. In practice, hardware realizations of noise reduction algorithms do not offer any access to their internal variables, preventing the usage of a white-box log kurtosis ratio measure. It is further stated in (Uemura et al., 2009) and (Inoue et al., 2011), that the derivation of an analytical function for calculating the log kurtosis ratio is difficult, therefore, a solution cannot be given for noise reduction algorithms applying the commonly used decision-directed (DD) approach (Ephraim and Malah, 1984) to the a priori SNR estimation. Therefore, these white-box log kurtosis ratio measurements are not applicable to a wide range of noise reduction algorithms utilizing the DD approach.
3.3.4 Black-Box Log Kurtosis Ratio

Due to the drawbacks of the white-box log kurtosis ratio measure for musical tones, a generic instrumental black-box measure is needed, which requires no knowledge of the noise reduction algorithm under test and can be applied to different noise reduction algorithms. Furthermore, no explicit assumption of (squared amplitude) speech and noise PDF is taken, where the assumed gamma distribution in (Uemura et al., 2008, 2009) and (Inoue et al., 2011) is just an approximation. We proposed in (Yu and Fingscheidt, 2011c, 2012c) a new black-box log kurtosis ratio \( \Delta \Psi_{\text{log}} \), which is computed as

\[
\Delta \Psi_{\text{log}} = \ln \left( \frac{\Psi_{\tilde{n}}}{\Psi_n} \right),
\]

with \( \ln(\cdot) \) being the natural logarithm operator, \( \Psi_n \) and \( \Psi_{\tilde{n}} \) being the kurtosis of the noise signal \( n(n) \) and the filtered noise signal \( \tilde{n}(n) \), respectively. \( \Delta \Psi_{\text{log}} \) is applied to quantify the amount of musical tones and is independent of speech distortion and noise attenuation by employing only the noise signal and the filtered noise signal. Unlike in (Uemura et al., 2008, 2009) and (Inoue et al., 2011), where the kurtosis was not strictly defined by applying the raw moment, a strict kurtosis definition according to the theory of higher-order statistics (Abramowitz and Stegun, 1965) is applied as

\[
\Psi_x = \frac{\text{E}\{[x - \text{E}\{x\}]^4\}}{(\text{E}\{[x - \text{E}\{x\}]^2\})^2},
\]

where \( \text{E}\{[x - \text{E}\{x\}]^n\} \) is the \( n \)th-order central moment. Different from (Uemura et al., 2009) and (Inoue et al., 2011), where \( |N(\ell,k)|^2 \) was assumed to be gamma-distributed in the power spectral domain, no such assumption is needed any more. Similar to (3.3.14), an instantaneous kurtosis of squared amplitude noise DFT coefficients for each frame \( \ell \) can be computed as

\[
\Psi_n(\ell) = \frac{1}{K} \sum_{k=0}^{K-1} \left[ |N(\ell,k)|^2 - |\tilde{N}(\ell,k)|^2 \right]^4 \left( \frac{1}{K} \sum_{k=0}^{K-1} \left[ |N(\ell,k)|^2 - |\tilde{N}(\ell,k)|^2 \right]^2 \right)^{-2},
\]

with \( |\tilde{N}(\ell,k)|^2 = \frac{1}{K} \sum_{k=0}^{K-1} |N(\ell,k)|^2 \). \( \Psi_n(\ell) \) can straightforwardly be computed by applying \( |\tilde{N}(\ell,k)|^2 \) in (3.3.15). The respective terms \( \Psi_n \) and \( \Psi_{\tilde{n}} \) can then be calculated as

\[
\Psi_n = \frac{1}{C(\Lambda_{H_0})} \sum_{\ell \in \Lambda_{H_0}} \Psi_n(\ell), \quad \Psi_{\tilde{n}} = \frac{1}{C(\Lambda_{H_0})} \sum_{\ell \in \Lambda_{H_0}} \Psi_{\tilde{n}}(\ell),
\]

where \( C(\Lambda_{H_0}) \) is the number of elements in set \( \Lambda_{H_0} \), which represents the subset of frames with no speech-dominant parts in \( \tilde{N}(\ell,k) \), when a noisy speech signal is applied as input to the noise reduction algorithm. Such a selection can be easily achieved during the evaluation by using the clean speech signal as a mask. If only a noise signal without a speech signal is applied as input signal to the noise reduction algorithm, \( \Lambda_{H_0} \) includes...
then all frames. Inserting $\Psi_n$ and $\tilde{\Psi}_n$ into (3.3.13), the log kurtosis ratio $\Delta \Psi_{\log}$ can finally be computed.

Please note that $\Delta \Psi_{\log}$ defined in (3.3.13) is calculated only from the noise signal $n(n)$ and the filtered noise signal $\tilde{n}(n)$, which allows it to be applicable to all noise reduction algorithms, also those utilizing the DD approach to estimate the a priori SNR. Furthermore, $\Delta \Psi_{\log}$ is computed in a simple direct way without extra knowledge of internal variables of the noise reduction algorithm under test, which means it can be considered as a black-box measurement.

### Performance Analysis

In order to evaluate the black-box log kurtosis ratio, four state-of-the-art noise reduction algorithms with noise signals only as input signals are employed. Our experiments are performed with 18 in-car background noise signals taken from the ETSI noise database (ETSI, 2008a), each sampled with 16 kHz. Within this performance analysis, all noise signals are normalized to –26 dBV according to ITU-T Rec. P.56 (ITU-T, 1993). For computing $\Delta \Psi_{\log}$, a DFT with the length $K = 512$, a root-squared Hann window with the length $L = 512$ as the analysis window, and a frame shift of 50% are applied. The four noise reduction algorithms under test are: the MMSE-SA (SA) estimator (Ephraim and Malah, 1984) and the MMSE-LSA (LSA) estimator (Ephraim and Malah, 1985), the a priori SNR-driven Wiener filter (Scalart and Filho, 1996), and the super-Gaussian joint MAP (SG) estimator (Lotter and Vary, 2005). The corresponding spectral gains are listed in Table 3.2.

<table>
<thead>
<tr>
<th>Name</th>
<th>Spectral gain $G(\ell,k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>$\Gamma_p(1.5) \sqrt{\frac{\vartheta(\ell,k)}{\xi(\ell,k)+1}} \exp(-\vartheta(\ell,k))(1 + \vartheta(\ell,k)I_0(\eta(\ell,k)) + \vartheta(\ell,k)I_1(\eta(\ell,k)))$</td>
</tr>
<tr>
<td>LSA</td>
<td>$\frac{\xi(\ell,k)}{\xi(\ell,k)+1} \exp \left( \frac{1}{2} \int_{\vartheta(\ell,k)}^{\infty} e^{-t} dt \right)$</td>
</tr>
<tr>
<td>WF</td>
<td>$\frac{\xi(\ell,k)}{\xi(\ell,k)+1}$</td>
</tr>
<tr>
<td>SG</td>
<td>$u(\ell,k) + \sqrt{u^2(\ell,k) + \frac{0.126}{2\gamma(\ell,k)}}$</td>
</tr>
</tbody>
</table>

Table 3.2: Spectral gains of the applied noise reduction algorithms, where $\Gamma_p(\cdot)$ denotes the gamma function, $\xi(\ell,k)$ is the a priori SNR, $\gamma(\ell,k)$ is the a posteriori SNR, $\vartheta(\ell,k) = \frac{\xi(\ell,k)\gamma(\ell,k)}{\xi(\ell,k)+1}$, $\eta(\ell,k) = \frac{\vartheta(\ell,k)}{2}$, $u(\ell,k) = 0.5 - \frac{1.74}{4\sqrt{\xi(\ell,k)\gamma(\ell,k)}}$, and $I_n(\cdot)$ denotes the $n$th-order modified Bessel function, respectively.

For all weighting rules, an estimation of the a priori SNR $\xi(\ell,k)$ defined as

$$\xi(\ell,k) = \frac{\mathbb{E}\{S(\ell,k)^2\}}{\mathbb{E}\{N(\ell,k)^2\}}$$

(3.3.17)
3.3.4. Black-Box Log Kurtosis Ratio

is needed, which is successfully addressed by the DD approach (Ephraim and Malah, 1984) as

$$\hat{\xi}'(\ell, k) = \beta \cdot \frac{|\hat{S}(\ell - 1, k)|^2}{\hat{\Phi}_{NN}(\ell - 1, k)} + (1 - \beta) \cdot P[\hat{\gamma}(\ell, k) - 1], \text{ and } \hat{\xi}(\ell, k) = \max\{\hat{\xi}'(\ell, k), \xi_{\min}\},$$

with a smoothing factor $\beta$, the enhanced speech signal of the previous frame $\hat{S}(\ell - 1, k)$, the a posteriori SNR $\hat{\gamma}(\ell, k) = \frac{|Y(\ell, k)|^2}{\hat{\Phi}_{NN}(\ell, k)}$, and $\xi_{\min} = -15$ dB, respectively. The noise power spectrum $\hat{\Phi}_{NN}(\ell, k)$ is estimated by the minimum statistics approach proposed by Martin (2001). Setting $\beta$ close to one yields a strong smoothing of the a priori SNR estimate, which helps to significantly reduce musical tones (Cappé, 1994). To demonstrate the proposed instrumental musical tones measurement, an evaluation with $0 \leq \beta < 1$ is performed. Please note that we change $\beta$ from 0 to 1 only to show the noise log kurtosis ratio measurements for different values of $\beta$, however, no information of $\beta$ is needed for calculating the noise log kurtosis ratio defined in (3.3.13).

The results of the black-box log kurtosis ratio $\Delta\Psi_{\log}$ for SA, LSA, WF and SG are shown in Fig. 3.3(a). Using our proposed $\Delta\Psi_{\log}$ (3.3.13), we observe: With increasing $\beta$, $\Delta\Psi_{\log}$ accordingly increases towards zero, meaning that the kurtosis of $\hat{n}(n)$ becomes more similar to the kurtosis of $n(n)$, which means higher statistical similarity of $n(n)$ and $\hat{n}(n)$, or, less musical tones. We found that by changing $\beta$ in (3.3.18), the higher the noise log kurtosis ratio is, the less musical tones are observed. If $\beta$ is chosen to be greater than 0.9, WF and SG show a more rapid $\Delta\Psi_{\log}$ increase than SA and LSA.

---

6However, as we are mostly interested in the region of $\beta \approx 1$, we define $\beta' = \ln\left(\frac{1+\beta}{1-\beta}\right)$ in analogy to the log-area ratios (Markel and Jr., 1976). A search over $\beta'$ with uniform step-size enables a search over $\beta$ having coarse step-sizes for $\beta$ close to zero and small step-sizes for $\beta \approx 1$. The precondition $0 \leq \beta' \leq 0.999$ leads to $0 \leq \beta' \leq 7.6$. In this thesis, $\beta'$ is varied with a step-size of 0.1.
However, until now $\Delta \Psi_{\text{log}}$ is evaluated only with wideband noise signals. In the next step, $\Delta \Psi_{\text{log}}$ is further investigated with narrowband noise signals, which can be generated by applying the MIRS weighting filter according to ITU-T Rec. P.830 (ITU-T, 1996b) to the same 18 in-car background noise signals and being downsampled to $f_s = 8$ kHz. In the narrowband case, a DFT with the length $K = 256$, a root-squared Hann window with the length $L = 256$ as the analysis window, and a frame shift of 50% are applied. Exactly the same evaluation is repeated with narrowband noise signals as the only inputs. The results of the black-box log kurtosis ratio for narrowband noise signals are depicted in Fig. 3.3(b). Unexpectedly, the $\Delta \Psi_{\text{log}}$ results of narrowband noise signals are inconsistent to the $\Delta \Psi_{\text{log}}$ results of wideband noise signals, although the relation between musical tones and $\beta$ is similar for both narrowband and wideband noise signals. If we only look at $\Delta \Psi_{\text{log}}$ results of narrowband noise signals in Fig. 3.3(b), we thus get the conclusion that the lower the noise log kurtosis ratio is, the less musical tones are observed. Further investigations led us to the finding that the problem of inconsistency lies on the type and the bandwidth of the applied noise signals. The type of the applied noise signals are all in-car background noise signals, which have extremely strong energies in the very low frequency region up to 300 Hz. These energies are just removed by using the MIRS weighting filter for narrowband noise signals (300...3400 Hz). However, they are kept for wideband noise signals (50...7000 Hz) employing the P.341 weighting filter. Since the instantaneous kurtosis calculation in (3.3.15) approximates the kurtosis defined in (3.3.14), which measures the shape of the PDF being very sensitive to extremely large values (Abramowitz and Stegun, 1965). Therefore, the extremely strong energies up to 300 Hz generate a bias effect, which results into two quite different relationships between the $\Delta \Psi_{\text{log}}$ values and the amount of musical tones for narrowband and wideband noise signals.

### 3.3.5 Black-Box Weighted Log Kurtosis Ratio

In order to solve the inconsistency problem of the black-box log kurtosis ratio using the absolute squared amplitude values of noise signals in (3.3.15), a modified approach defined as black-box weighted log kurtosis ratio is proposed now (Yu and Fingscheidt, 2012d,b). The main difference of the new approach is that we modify the calculation of $\Psi_n(\ell)$ in (3.3.15) to become a weighted kurtosis $\Psi_n^w(\ell)$ defined as

$$
\Psi_n^w(\ell) = \frac{1}{K} \sum_{k=0}^{K-1} \left( \frac{\alpha_n(k) \cdot |N(\ell,k)|^2 - \alpha_n(\kappa) \cdot |N(\ell,\kappa)|^2}{\frac{1}{K} \sum_{k=0}^{K-1} \alpha_n(k) \cdot |N(\ell,k)|^2 - \alpha_n(\kappa) \cdot |N(\ell,\kappa)|^2} \right)^4,
$$

(3.3.19)

with $\frac{1}{K} \sum_{\kappa=0}^{K-1} \alpha_n(\kappa) \cdot |N(\ell,\kappa)|^2$. The weighting factor $\alpha_n(k)$ is being calculated for each frequency bin as the inverse mean value of $|N(\ell,k)|^2$ across $C(A_{H_0})$ frames as

$$
\alpha_n(k) = \left( \frac{1}{C(A_{H_0})} \sum_{\ell \in A_{H_0}} |N(\ell,k)|^2 \right)^{-1},
$$

(3.3.20)
3.3.5. Black-Box Weighted Log Kurtosis Ratio

**Figure 3.4:** Black-box weighted log kurtosis ratio $\Delta \Psi_{\log}^w$ of the four weighting rules with $0 \leq \beta < 1$ using (a) wideband noise signals (Yu and Fingscheidt, 2012b), and (b) narrowband noise signals (Yu and Fingscheidt, 2012d) as the only input signals.

The term $\alpha_n(k)$ can thus be interpreted as a normalization factor of $|N(\ell, k)|^2$ for each frequency bin along $C(\Lambda_{H_0})$ frames. $C(\Lambda_{H_0})$ represents a subset of frames with no speech-dominant parts when applying the noisy speech signals as the inputs. By applying noise signals as the only inputs, $C(\Lambda_{H_0})$ is the number of all frames. Employing the normalization factor, the weighted kurtosis $\Psi_n^w(\ell)$ is calculated with the normalized value of $|N(\ell, k)|^2$. Based on the normalized value of $|N(\ell, k)|^2$ for each frequency bin, $\Psi_n^w(\ell)$ becomes independent of the absolute value of $|N(\ell, k)|^2$ and is thus insensitive to the extremely large values of $|N(\ell, k)|^2$ at low frequencies. In the same manner, the weighted kurtosis $\Psi_{\tilde{n}}^w(\ell)$ can be computed by applying $|\tilde{N}(\ell, k)|^2$ in (3.3.20) to calculate $\tilde{\alpha}_n(k)$ and subsequently using $|\tilde{N}(\ell, k)|^2$ and $\alpha_{\tilde{n}}(k)$ in (3.3.19). The respective terms $\Psi_n^w$ and $\Psi_{\tilde{n}}^w$ can then be calculated using $\Psi_n^w(\ell)$ and $\Psi_{\tilde{n}}^w(\ell)$ instead of $\Psi_n(\ell)$ and $\Psi_{\tilde{n}}(\ell)$ in (3.3.16), respectively. Finally, the black-box weighted log kurtosis ratio can be computed as

\[
\Delta \Psi_{\log}^w = \ln \left( \frac{\Psi_{\tilde{n}}^w}{\Psi_n^w} \right).
\]

**Performance Analysis**

The black-box weighted log kurtosis ratio $\Delta \Psi_{\log}^w$ is evaluated for both wideband noise signals (Yu and Fingscheidt, 2012b) and narrowband noise signals (Yu and Fingscheidt, 2012d). Applying the same 18 in-car background noise signals in their wideband and narrowband versions as the only input signals, the results of $\Delta \Psi_{\log}^w$ for the four noise reduction algorithms (see Table 3.2) are presented in Fig. 3.4(a) for wideband noise signals and Fig. 3.4(b) for narrowband noise signals, respectively. Applying $\Delta \Psi_{\log}^w$, a consistent relationship between $\Delta \Psi_{\log}^w$ and $\beta$ in (3.3.18) (i.e., the amount of musical tones) can firstly be observed for both wideband and narrowband noise signals. Secondly, in Figs. 3.4(a) and 3.4(b), two regions can be detected. The first region can be defined for wideband
noise signals with \( \beta > 0.9 \) for LSA and SA, and \( \beta > 0.94 \) for SG and WF, respectively, and for narrowband noise signals with \( \beta > 0.9 \) for all weighting rules. Within the first region, \( \Delta \Psi_{\log}^w \) decreases very fast towards zero with increasing \( \beta \). This means that the weighted kurtosis of \( \tilde{n}(n) \) becomes more similar to the weighted kurtosis of \( n(n) \), meaning higher statistical similarity of \( n(n) \) and \( \tilde{n}(n) \), or, less musical tones. This matches the observations in Cappé (1994) well, where \( \beta \) has been increased from 0.98 to 0.998. However, in Cappé (1994) and Breithaupt and Martin (2011) the smoothing effect of \( \beta \) in (3.3.18) for reducing musical tones was only investigated with \( \beta \) being chosen very close to one. We found that this smoothing comes only strongly in effect in the first region. This corresponds nicely to our above mentioned observation with the weighted log kurtosis ratio. In contrast, in the second region for LSA and SA with \( \beta \) being changed from 0 to 0.9 for both wideband and narrowband noise signals, the perceived amount of musical tones hardly changes for LSA and SA. Also, \( \Delta \Psi_{\log}^w \) values of LSA and SA remain nearly unchanged within this region for narrowband and wideband noise signals. Please note that the second region is an untypical setup of \( \beta \) for the decision-directed approach in (3.3.18). For SG and WF, the second region can be observed as \( 0 < \beta \leq 0.94 \) for wideband noise signals and \( 0 < \beta \leq 0.9 \) for narrowband noise signals. Within this region, in contrast to our first expectation, the weighted log kurtosis ratio increases with increasing \( \beta \) for SG and WF. This is further illustrated in Fig. 3.5(a) for a noise signal

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.5.png}
\caption{Spectrum of a weighted filtered noise signal using SG with (a) \( \beta = 0 \), \( \beta = 0.94 \) and \( \beta = 0.999 \) (from top to bottom) for its wideband version (Yu and Fingscheidt, 2012b), and (b) \( \beta = 0 \), \( \beta = 0.9 \) and \( \beta = 0.999 \) (from top to bottom) for its narrowband version (Yu and Fingscheidt, 2012d)}
\end{figure}

in its wideband version employing SG with three different \( \beta \) values: Using \( \beta = 0 \), much more peaks (red points) are being generated in the spectrum of the weighted filtered noise signal using the weighting factor \( \alpha_n(k) \) as compared to using \( \beta = 0.94 \). However, since for \( \beta = 0 \) many such peaks are generated within each frame \( \ell \), they are not isolated any more, leading to spectral distortions of broader bandwidth. For \( \beta = 0.94 \) we observe less peaks, however, in a more isolated fashion, clearly indicating the effect of musical
tones. This provides the reason why $\Psi_{w}^{\tilde{n}}$ and consequently $\Delta \Psi_{w}^{\log}$ increase for $\beta \to 0.94$. When we compare the case of using $\beta = 0.94$ and $\beta = 0.999$, strong smoothing comes into effect for $\beta \to 0.999$, indicating removal of isolated peaks (musical tones) also by a decrease of $\Delta \Psi_{w}^{\log}$. With this observation we show that the decision-directed approach to a priori SNR estimation (3.3.18) has a different smoothing extent within different $\beta$ ranges. The same phenomenon is shown in Fig.3.5(b) for the same noise signal in its narrowband version employing SG with $\beta = \{0, 0.9, 0.999\}$. Hence, for both wideband and narrowband signals, we found that by changing $\beta$ in (3.3.18), the lower the weighted log kurtosis ratio is, the less musical tones are observed, especially for $\beta \geq 0.94$ (wideband noise signals) and $\beta \geq 0.9$ (narrowband noise signals), which are mostly employed for the decision-directed approach. It is further stated by Breithaupt and Martin (2011), that the decision-directed approach provides much less smoothing effect with SG and WF than with LSA and SA. This is well reflected in Figs.3.4(a) and 3.4(b), where $\Delta \Psi_{w}^{\log}$ of SG and WF are higher than LSA and SA for all $\beta$ values.

Subjective ACR Tests

In order to further validate $\Delta \Psi_{w}^{\log}$, two subjective listening tests in an ACR fashion have been separately conducted for wideband noise signals (Yu and Fingscheidt, 2012b) and for narrowband signals (Yu and Fingscheidt, 2012d), respectively. The setup of the two ACR subjective listening tests is identical: Sixteen test persons (experts and non-experts) had to rate the audibility of musical tones according to an ACR listening scale with seven grades: (1) intolerably audible, (2) loudly audible, (3) rather loudly audible, (4) moderately audible, (5) slightly audible, (6) just audible, (7) inaudible. Three noise signals from the applied 18 in-car background noises have been randomly chosen. Each noise signal has been processed by the SA, LSA, WF and SG weighting rules. Three values of $\beta$ with 0.96, 0.98 (being the optimal value for SA (Ephraim and Malah, 1984)), and 0.993 (being the optimal value for SG (Yu and Fingscheidt, 2011a)) were chosen for each weighting rule. Altogether 36 output (filtered) noise signals had to be rated by each listener. The listening test results are shown in Fig.3.6(a) for wideband noise signals and in Fig.3.6(b) for narrowband noise signals, respectively. Furthermore, the related instrumental $\Delta \Psi_{w}^{\log}$ measurements are shown in Fig.3.7(a) for wideband noise signals and in Fig.3.7(b) for narrowband noise signals, respectively. For both wideband and narrowband noise signals, the ACR results match the $\Delta \Psi_{w}^{\log}$ results nicely for all weighting rules. In order to evaluate $\Delta \Psi_{w}^{\log}$ quantitatively, within each ACR listening test a Pearson’s correlation (Burington and May, 1970) between the subjective $\text{ACR}_i$ and the instrumental measure $\Delta \Psi_{w}^{\log,i}$ is calculated as

$$
\rho = \frac{\sum_i (\text{ACR}_i - \overline{\text{ACR}}_j)(\Delta \Psi_{w}^{\log,i} - \overline{\Delta \Psi_{w}^{\log,i}})}{\sqrt{\sum_i (\text{ACR}_i - \overline{\text{ACR}}_j)^2} \sqrt{\sum_i (\Delta \Psi_{w}^{\log,i} - \overline{\Delta \Psi_{w}^{\log,i}})^2}},
$$

(3.3.22)

where $\overline{\text{ACR}}_j$ and $\overline{\Delta \Psi_{w}^{\log,i}}$ are the mean values of $\text{ACR}_j$ and $\Delta \Psi_{w}^{\log,j}$, respectively. For each weighing rule, Pearson’s correlation is computed by using the three $\overline{\text{ACR}}_i$ values and the
three $\Delta \Psi_{\log,i}^w$ values with $i = 1, 2, 3$ (representing three $\beta$ values) shown in Figs. 3.6 and 3.7, respectively. Finally, all 12 values of $\text{ACR}_i$ and $\Delta \Psi_{\log,i}^w$ are employed to compute an overall $\rho$ for all weighting rules combined. The correlation results for wideband and narrowband noise signals are presented in Table 3.3. It can be seen that for both wideband and narrowband noise signals, the ACR results correlate with the weighted log kurtosis ratio measurements very well for each weighting rule and all weighting rules combined.

**Figure 3.6:** ACR listening test results of the four weighting rules with $\beta = \{0.96, 0.98, 0.993\}$ for (a) wideband noise signals (Yu and Fingscheidt, 2012b), and (b) narrow noise signals (Yu and Fingscheidt, 2012d)

**Figure 3.7:** Black-box weighted log kurtosis ratio $\Delta \Psi_{\log,i}^w$ of the four weighting rules with $\beta = \{0.96, 0.98, 0.993\}$ for (a) wideband noise signals (Yu and Fingscheidt, 2012b), and (b) narrowband noise signals (Yu and Fingscheidt, 2012d)

Both instrumental and subjective results reveal that the larger $\beta$ is, the less musical tones are perceived. The overall correlation between instrumental and subjective results of $|\rho| = 0.9530$ for wideband noise signals and $|\rho| = 0.9880$ for narrowband noise signals
3.3.5. Black-Box Weighted Log Kurtosis Ratio

<table>
<thead>
<tr>
<th>Correlation coef.</th>
<th>Wideband noise signals</th>
<th>Narrowband noise signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{SG}$</td>
<td>-0.9517</td>
<td>-0.9999</td>
</tr>
<tr>
<td>$\rho_{WF}$</td>
<td>-0.9415</td>
<td>-0.9850</td>
</tr>
<tr>
<td>$\rho_{LSA}$</td>
<td>-0.9727</td>
<td>-0.9975</td>
</tr>
<tr>
<td>$\rho_{SA}$</td>
<td>-0.9839</td>
<td>-0.9969</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.9530</td>
<td>-0.9880</td>
</tr>
</tbody>
</table>

Table 3.3: Correlation coefficients of the four weighting rules and corresponding overall correlation coefficients for wideband and narrowband noise signals

have verified that the weighted log kurtosis ratio is an adequate instrumental measure for assessing musical tones within a black-box measurement environment. The black-box weighted log kurtosis ratio has been included as part of an ITU-T Recommendation proposal made by the Study Group 12, Focus Group on Car Communication (FG CarCOM).

Two Further Analyses of Different Calculation Methods and Measurement Setups of the Black-Box Weighted Log Kurtosis Ratio Measure

So far, the performance analysis of the black-box weighted log kurtosis ratio using (3.3.21) for measuring musical tones has been presented. In this section, two further analyses of the weighted log kurtosis ratio are carried out. The first analysis concentrates on assessing different calculation methods of the black-box weighted log kurtosis ratio by evaluating another three alternative calculation methods. The second analysis focuses on assessing different measurement setups of the weighted log kurtosis ratio with different DFT lengths, analysis window types, frame lengths, and frame shifts. Both analyses are performed by utilizing the correlation coefficients as the quality indicator. Actually, a third analysis could also be performed by changing the default setup of different noise reduction algorithms (see Table 3.1) with different DFT lengths, analysis and synthesis window types, frames lengths and frame shifts, while keeping the measurement setup of the weighted log kurtosis ratio unchanged. However, in order to calculate the correlation coefficients, extra ACR subjective listening tests for all possible setups of a noise reduction algorithm would have to be performed, since different filtered noise signals are generated with different setups of the noise reduction algorithm. Unfortunately, this is just infeasible within the scope of this thesis and was therefore not performed.

In the first analysis, it is of interest to investigate some other methods of calculating the weighted log kurtosis ratio between $n(n)$ and $\tilde{n}(n)$. The method using (3.3.19), (3.3.20), and (3.3.21) is defined now as a first variant $\Delta \Psi_{\log}^{w,\text{var1}}$, which at first calculates the instantaneous weighted kurtosis of each frame as $\Psi_{\log}^{w}(\ell)$ and $\Psi_{\log}^{w}(\ell)$ (3.3.19), and the final weighted kurtosis $\Psi_{\log}^{w,\text{var1}}$ and $\Psi_{\log}^{w,\text{var1}}$ are calculated by averaging the instantaneous weighted kurtosis over frames. We now consider a second variant, which firstly calculates
the instantaneous weighted kurtosis of each frequency bin as

\[ \Psi_n^w(k) = \frac{1}{C(\Lambda_{H_0})} \sum_{\ell \in \Lambda_{H_0}} \left[ \alpha_n(\ell) \cdot |N(\ell, k)|^2 - \alpha_n(\lambda) \cdot |N(\lambda, k)|^2 \right]^4 \left( \frac{1}{C(\Lambda_{H_0})} \sum_{\ell \in \Lambda_{H_0}} \left[ \alpha_n(\ell) \cdot |N(\ell, k)|^2 - \alpha_n(\lambda) \cdot |N(\lambda, k)|^2 \right]^2 \right)^2, \]  

(3.3.23)

with

\[ \frac{\alpha_n(\lambda) \cdot |N(\lambda, k)|^2}{C(\Lambda_{H_0})} = \frac{1}{C(\Lambda_{H_0})} \sum_{\lambda \in \Lambda_{H_0}} \alpha_n(\lambda) \cdot |N(\lambda, k)|^2, \]  

(3.3.24)

and

\[ \alpha_n(\ell) = \left( \frac{1}{K} \sum_{k=0}^{K-1} |N(\ell, k)|^2 \right)^{-1}, \quad \ell \in \Lambda_{H_0}. \]  

(3.3.25)

In the same manner, the filtered noise instantaneous weighted kurtosis \( \Psi_n^w(k) \) of the second variant can then be calculated by using \( |\tilde{N}(\ell, k)|^2 \) in (3.3.23), (3.3.24), and (3.3.25), respectively. The weighted kurtosis of \( n(n) \) and \( \tilde{n}(n) \) can then be calculated by averaging \( \Psi_n^w(k) \) and \( \Psi_n^w(k) \) over all frequencies as

\[ \Psi_n^{w, var2} = \frac{1}{K} \sum_{k=0}^{K-1} \Psi_n^w(k), \quad \Psi_n^{w, var2} = \frac{1}{K} \sum_{k=0}^{K-1} \Psi_n^w(k). \]  

(3.3.26)

The second variant can finally be computed as

\[ \Delta \Psi_n^{w, var2} = \ln \left( \frac{\Psi_n^{w, var2}}{\Psi_n^{w, var2}} \right). \]  

(3.3.27)

Furthermore, the third variant \( \Delta \Psi_n^{w, var3} \) can be formulated by combining \( \Psi_n^{w, var1} \) and \( \Psi_n^{w, var2} \) as well as \( \Psi_n^{w, var1} \) and \( \Psi_n^{w, var2} \) by

\[ \Delta \Psi_n^{w, var3} = \ln \left( \frac{\Psi_n^{w, var1} + \Psi_n^{w, var2}}{\Psi_n^{w, var1} + \Psi_n^{w, var2}} \right). \]  

(3.3.28)

The fourth variant \( \Delta \Psi_n^{w, var4} \) can be defined by averaging \( \Delta \Psi_n^{w, var1} \) and \( \Delta \Psi_n^{w, var2} \) as

\[ \Delta \Psi_n^{w, var4} = \frac{1}{2} \Delta \Psi_n^{w, var1} + \frac{1}{2} \Delta \Psi_n^{w, var2}. \]  

(3.3.29)

The same evaluation for the first variant is repeated separately for the second variant (3.3.27), the third variant (3.3.28), and the fourth variant (3.3.29) for wideband and narrowband noise signals, respectively. The correlation coefficients between the three alternative weighted log kurtosis ratio measurements and the ACR listening test results are then separately calculated for wideband and narrowband noise signals. Combined with the results of the first variant (see Table 3.3), all correlation coefficients of the four variants are shown in Tables 3.4 and 3.5 for wideband and narrowband noise signals, respectively.

It can be seen in Tables 3.4 and 3.5 that the other three alternative variants can also provide good correlation coefficients for different weighting rules and all weighting rules
3.3.5. Black-Box Weighted Log Kurtosis Ratio

<table>
<thead>
<tr>
<th>Variant</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.9517</td>
<td>-0.9415</td>
<td>-0.9727</td>
<td>-0.9839</td>
<td>-0.9530</td>
</tr>
<tr>
<td>2</td>
<td>-0.8729</td>
<td>-0.9245</td>
<td>-0.9597</td>
<td>-0.9828</td>
<td>-0.8579</td>
</tr>
<tr>
<td>3</td>
<td>-0.9038</td>
<td>-0.9315</td>
<td>-0.9652</td>
<td>-0.9839</td>
<td>-0.8914</td>
</tr>
<tr>
<td>4</td>
<td>-0.9212</td>
<td>-0.9345</td>
<td>-0.9652</td>
<td>-0.9832</td>
<td>-0.9114</td>
</tr>
</tbody>
</table>

Table 3.4: Correlation coefficients of the four weighted log kurtosis ratio calculation variants for the four weighting rules using wideband noise signals

<table>
<thead>
<tr>
<th>Variant</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.9999</td>
<td>-0.9850</td>
<td>-0.9975</td>
<td>-0.9969</td>
<td>-0.9880</td>
</tr>
<tr>
<td>2</td>
<td>-0.9702</td>
<td>-0.9457</td>
<td>-0.9911</td>
<td>-0.9923</td>
<td>-0.9592</td>
</tr>
<tr>
<td>3</td>
<td>-0.9780</td>
<td>-0.9554</td>
<td>-0.9930</td>
<td>-0.9933</td>
<td>-0.9706</td>
</tr>
<tr>
<td>4</td>
<td>-0.9876</td>
<td>-0.9617</td>
<td>-0.9924</td>
<td>-0.9929</td>
<td>-0.9788</td>
</tr>
</tbody>
</table>

Table 3.5: Correlation coefficients of the four weighted log kurtosis ratio calculation variants for the four weighting rules using narrowband noise signals

combined. It means that the weighted log kurtosis ratio computation is quite insensitive to the sequence of averaging over frequencies and frames for computing the weighted kurtosis utilizing the instantaneous weighted kurtosis. Nevertheless, our original method (first variant) defined in (3.3.21) provides the highest correlation to the ACR listening tests for wideband and narrowband noise signals. Therefore, in the remainder of this thesis, the weighted log kurtosis ratio according to (3.3.21) is employed for instrumentally measuring musical tones.

In the second analysis, we evaluate how sensitive the weighted log kurtosis ratio is according to (3.3.21) to its measurement setup with different DFT and frame lengths, analysis window types, and frame shifts. These parameters are shown in Tables 3.6 and 3.7 for wideband and narrowband noise signals, respectively.

The correlation coefficients results of wideband noise signals employing different measurement setups for computing $\Delta \Psi_{w}^{\log}$ are summarized in Tables 3.8–3.11. Tables 3.12–3.15 then present the correlation coefficients results for narrowband noise signals. Firstly, it can be found that for both wideband and narrowband noise signals, the applied analysis window type has almost no influence on the correlation coefficients for different weighting rules and overall correlation coefficients.

For wideband noise signals within each analysis window type: Under the same DFT length (frame length), the frame shift has nearly no influence on the correlation coefficients for different weighting rules and the overall correlation coefficients. Increasing the DFT length (frame length), decreasing correlation coefficients for all weighting rules except WF can be detected. However, the degree of decrease is very low. Therefore, for wideband noise signals, the weighted log kurtosis ratio can be treated as quite insensitive to different DFT lengths (frame lengths), analysis window types, and frame shifts. Nevertheless,
within each analysis window type, the highest overall correlation coefficient is obtained for $K = L = 512$, $\Delta L = 50\% [\times L]$. Furthermore, among all these highest overall correlation coefficients, the highest overall correlation coefficient is obtained with the root-squared Hann window. Hence, the measurement setup with $K = L = 512$, $\Delta L = 50\% [\times L]$, and the root-squared Hann window, which matches the setup of wideband noise reduction algorithms in Table 3.1, provides the best overall correlation coefficient. Similar results can be also observed for narrowband noise signals. Within each analysis window type: It can be seen that the frame shift has nearly no influence on the correlation coefficients for different weighting rules and the overall correlation coefficients under the same DFT length (frame length). Increasing the DFT length (frame length), increasing correlation coefficients of all weighting rules except WF can be detected, while the correlation coefficients of WF decreases. Still, the degree of modification is very low. Therefore, for narrowband noise signals, the weighted log kurtosis ratio can also be treated as quite insensitive to DFT lengths (frame lengths), analysis window types, and frame shifts. Like the case of wideband noise signals, the highest overall correlation coefficient is obtained with $K = L = 256$, $\Delta L = 50\% [\times L]$ for each analysis window type. In addition, among these highest overall correlation coefficients, the root-squared Hann window provides the highest one. Hence, like for the wideband case, the measurement setup of $K = L = 256$, $\Delta L = 50\% [\times L]$, and the root-squared Hann also yields the best overall correlation coefficient for narrowband noise signals, which matches the setup of narrowband noise reduction algorithms in Table 3.1.

Therefore, the weighted log kurtosis ratio can be verified as nearly insensitive to different measurement setups shown in Tables 3.8–3.11 for wideband noise signals and in Tables 3.12–3.15 for narrowband noise signals. However, among these results, using the same setup of the applied noise reduction algorithms to compute the weighted log kurtosis ratio still provides the best results for wideband and narrowband noise signals. Since we focus on wideband noise reduction algorithms in the remainder of this thesis, the weighted log kurtosis ratio $\Delta \Psi_{\log}^w$ is thus calculated according to (3.3.21) by employing $K = L = 512$, $\Delta L = 50\% [\times L]$, and the root-squared Hann window for wideband noise signals.

### Table 3.6: Measurement setups of $\Delta \Psi_{\log}^w$ with different DFT and frame lengths, analysis window types, and frame shifts for wideband noise signals

<table>
<thead>
<tr>
<th>DFT and frame lengths</th>
<th>$K = L \in {256, 512, 1024}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame shifts</td>
<td>$\Delta L \in {50%, 25%, 12.5%, 6.25%, 3.125%} [\times L]$</td>
</tr>
<tr>
<td>Typical analysis windows</td>
<td>root-squared Hann, Hann, Hamming, Blackman</td>
</tr>
</tbody>
</table>

### Table 3.7: Measurement setups of $\Delta \Psi_{\log}^w$ with different DFT and frame lengths, analysis window types, and frame shifts for narrowband noise signals

<table>
<thead>
<tr>
<th>DFT and frame lengths</th>
<th>$K = L \in {128, 256, 512}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame shifts</td>
<td>$\Delta L \in {50%, 25%, 12.5%, 6.25%, 3.125%} [\times L]$</td>
</tr>
<tr>
<td>Typical analysis windows</td>
<td>root-squared Hann, Hann, Hamming, Blackman</td>
</tr>
</tbody>
</table>
### Table 3.8: Correlation coefficients between the instrumental $\Delta \Psi_{wlog}^w$ measurements and ACR listening tests for wideband noise signals employing a root-squared Hann window and different DFT lengths $K$, frame lengths $L$, and frame shifts $\Delta L$ for computing $\Delta \Psi_{wlog}^w$

<table>
<thead>
<tr>
<th>$K = L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9575</td>
<td>-0.9632</td>
<td>-0.9900</td>
<td>-0.9958</td>
<td>-0.9444</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9576</td>
<td>-0.9631</td>
<td>-0.9905</td>
<td>-0.9955</td>
<td>-0.9447</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9576</td>
<td>-0.9631</td>
<td>-0.9908</td>
<td>-0.9958</td>
<td>-0.9452</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9575</td>
<td>-0.9628</td>
<td>-0.9930</td>
<td>-0.9965</td>
<td>-0.9483</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9575</td>
<td>-0.9628</td>
<td>-0.9934</td>
<td>-0.9966</td>
<td>-0.9486</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9517</td>
<td>-0.9415</td>
<td>-0.9727</td>
<td>-0.9839</td>
<td>-0.9530</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9528</td>
<td>-0.9468</td>
<td>-0.9787</td>
<td>-0.9859</td>
<td>-0.9518</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9529</td>
<td>-0.9471</td>
<td>-0.9795</td>
<td>-0.9860</td>
<td>-0.9515</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9530</td>
<td>-0.9472</td>
<td>-0.9797</td>
<td>-0.9863</td>
<td>-0.9516</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9530</td>
<td>-0.9471</td>
<td>-0.9808</td>
<td>-0.9864</td>
<td>-0.9529</td>
</tr>
<tr>
<td>1024</td>
<td>50%</td>
<td>-0.9447</td>
<td>-0.9504</td>
<td>-0.9701</td>
<td>-0.9776</td>
<td>-0.9374</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9460</td>
<td>-0.9516</td>
<td>-0.9717</td>
<td>-0.9785</td>
<td>-0.9387</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9465</td>
<td>-0.9515</td>
<td>-0.9726</td>
<td>-0.9789</td>
<td>-0.9392</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9465</td>
<td>-0.9516</td>
<td>-0.9727</td>
<td>-0.9791</td>
<td>-0.9393</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9465</td>
<td>-0.9516</td>
<td>-0.9728</td>
<td>-0.9791</td>
<td>-0.9394</td>
</tr>
</tbody>
</table>

### Table 3.9: Correlation coefficients between the instrumental $\Delta \Psi_{wlog}^w$ measurements and ACR listening tests for wideband noise signals employing a Hann window and different DFT lengths $K$, frame lengths $L$, and frame shifts $\Delta L$ for computing $\Delta \Psi_{wlog}^w$

<table>
<thead>
<tr>
<th>$K = L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9586</td>
<td>-0.9671</td>
<td>-0.9921</td>
<td>-0.9955</td>
<td>-0.9430</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9583</td>
<td>-0.9670</td>
<td>-0.9924</td>
<td>-0.9955</td>
<td>-0.9434</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9584</td>
<td>-0.9671</td>
<td>-0.9925</td>
<td>-0.9955</td>
<td>-0.9437</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9583</td>
<td>-0.9669</td>
<td>-0.9943</td>
<td>-0.9957</td>
<td>-0.9472</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9583</td>
<td>-0.9668</td>
<td>-0.9944</td>
<td>-0.9958</td>
<td>-0.9475</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9537</td>
<td>-0.9439</td>
<td>-0.9758</td>
<td>-0.9872</td>
<td>-0.9529</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9539</td>
<td>-0.9499</td>
<td>-0.9801</td>
<td>-0.9878</td>
<td>-0.9498</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9539</td>
<td>-0.9502</td>
<td>-0.9806</td>
<td>-0.9881</td>
<td>-0.9495</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9539</td>
<td>-0.9503</td>
<td>-0.9808</td>
<td>-0.9882</td>
<td>-0.9496</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9539</td>
<td>-0.9502</td>
<td>-0.9816</td>
<td>-0.9884</td>
<td>-0.9509</td>
</tr>
<tr>
<td>1024</td>
<td>50%</td>
<td>-0.9470</td>
<td>-0.9499</td>
<td>-0.9713</td>
<td>-0.9783</td>
<td>-0.9423</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9492</td>
<td>-0.9506</td>
<td>-0.9724</td>
<td>-0.9796</td>
<td>-0.9436</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9497</td>
<td>-0.9508</td>
<td>-0.9727</td>
<td>-0.9803</td>
<td>-0.9430</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9498</td>
<td>-0.9509</td>
<td>-0.9728</td>
<td>-0.9805</td>
<td>-0.9431</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9498</td>
<td>-0.9509</td>
<td>-0.9729</td>
<td>-0.9805</td>
<td>-0.9432</td>
</tr>
</tbody>
</table>
Table 3.10: Correlation coefficients between the instrumental $\Delta \Psi_w$ measurements and ACR listening tests for wideband noise signals employing a Hamming window and different DFT lengths $K$, frame lengths $L$, and frame shifts $\Delta L$ for computing $\Delta \Psi_w$.

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9621</td>
<td>-0.9706</td>
<td>-0.9922</td>
<td>-0.9961</td>
<td>-0.9435</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9613</td>
<td>-0.9704</td>
<td>-0.9917</td>
<td>-0.9952</td>
<td>-0.9437</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9614</td>
<td>-0.9705</td>
<td>-0.9916</td>
<td>-0.9952</td>
<td>-0.9437</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9617</td>
<td>-0.9706</td>
<td>-0.9927</td>
<td>-0.9954</td>
<td>-0.9458</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9618</td>
<td>-0.9706</td>
<td>-0.9929</td>
<td>-0.9955</td>
<td>-0.9460</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9558</td>
<td>-0.9489</td>
<td>-0.9660</td>
<td>-0.9822</td>
<td>-0.9418</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9562</td>
<td>-0.9537</td>
<td>-0.9738</td>
<td>-0.9849</td>
<td>-0.9423</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9560</td>
<td>-0.9541</td>
<td>-0.9746</td>
<td>-0.9851</td>
<td>-0.9424</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9560</td>
<td>-0.9542</td>
<td>-0.9747</td>
<td>-0.9852</td>
<td>-0.9426</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9560</td>
<td>-0.9541</td>
<td>-0.9748</td>
<td>-0.9852</td>
<td>-0.9432</td>
</tr>
<tr>
<td>1024</td>
<td>50%</td>
<td>-0.9478</td>
<td>-0.9553</td>
<td>-0.9701</td>
<td>-0.9791</td>
<td>-0.9360</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9494</td>
<td>-0.9561</td>
<td>-0.9704</td>
<td>-0.9795</td>
<td>-0.9372</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9501</td>
<td>-0.9563</td>
<td>-0.9706</td>
<td>-0.9794</td>
<td>-0.9366</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9497</td>
<td>-0.9561</td>
<td>-0.9705</td>
<td>-0.9792</td>
<td>-0.9365</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9498</td>
<td>-0.9562</td>
<td>-0.9707</td>
<td>-0.9790</td>
<td>-0.9366</td>
</tr>
</tbody>
</table>

Table 3.11: Correlation coefficients between the instrumental $\Delta \Psi_w$ measurements and ACR listening tests for wideband noise signals employing a Blackman window and different DFT lengths $K$, frame lengths $L$, and frame shifts $\Delta L$ for computing $\Delta \Psi_w$.

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9603</td>
<td>-0.9710</td>
<td>-0.9943</td>
<td>-0.9970</td>
<td>-0.9409</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9596</td>
<td>-0.9711</td>
<td>-0.9942</td>
<td>-0.9966</td>
<td>-0.9413</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9597</td>
<td>-0.9713</td>
<td>-0.9945</td>
<td>-0.9967</td>
<td>-0.9415</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9597</td>
<td>-0.9710</td>
<td>-0.9959</td>
<td>-0.9969</td>
<td>-0.9447</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9597</td>
<td>-0.9710</td>
<td>-0.9960</td>
<td>-0.9969</td>
<td>-0.9449</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9550</td>
<td>-0.9470</td>
<td>-0.9785</td>
<td>-0.9900</td>
<td>-0.9506</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9550</td>
<td>-0.9532</td>
<td>-0.9824</td>
<td>-0.9898</td>
<td>-0.9483</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9550</td>
<td>-0.9533</td>
<td>-0.9829</td>
<td>-0.9898</td>
<td>-0.9481</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9551</td>
<td>-0.9534</td>
<td>-0.9830</td>
<td>-0.9899</td>
<td>-0.9482</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9551</td>
<td>-0.9534</td>
<td>-0.9836</td>
<td>-0.9901</td>
<td>-0.9493</td>
</tr>
<tr>
<td>1024</td>
<td>50%</td>
<td>-0.9479</td>
<td>-0.9479</td>
<td>-0.9716</td>
<td>-0.9803</td>
<td>-0.9452</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9506</td>
<td>-0.9485</td>
<td>-0.9733</td>
<td>-0.9809</td>
<td>-0.9471</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9513</td>
<td>-0.9497</td>
<td>-0.9744</td>
<td>-0.9817</td>
<td>-0.9456</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9514</td>
<td>-0.9498</td>
<td>-0.9746</td>
<td>-0.9819</td>
<td>-0.9457</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9514</td>
<td>-0.9499</td>
<td>-0.9747</td>
<td>-0.9819</td>
<td>-0.9458</td>
</tr>
</tbody>
</table>
### Table 3.12:

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>50%</td>
<td>-0.9927</td>
<td>-0.9992</td>
<td>-0.9945</td>
<td>-0.9952</td>
<td>-0.9843</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9929</td>
<td>-0.9992</td>
<td>-0.9941</td>
<td>-0.9945</td>
<td>-0.9838</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9928</td>
<td>-0.9992</td>
<td>-0.9937</td>
<td>-0.9944</td>
<td>-0.9837</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9928</td>
<td>-0.9992</td>
<td>-0.9933</td>
<td>-0.9942</td>
<td>-0.9829</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9928</td>
<td>-0.9992</td>
<td>-0.9931</td>
<td>-0.9942</td>
<td>-0.9829</td>
</tr>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9999</td>
<td>-0.9850</td>
<td>-0.9975</td>
<td>-0.9969</td>
<td>-0.9880</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9997</td>
<td>-0.9846</td>
<td>-0.9965</td>
<td>-0.9956</td>
<td>-0.9865</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9994</td>
<td>-0.9874</td>
<td>-0.9961</td>
<td>-0.9955</td>
<td>-0.9861</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9993</td>
<td>-0.9873</td>
<td>-0.9957</td>
<td>-0.9952</td>
<td>-0.9863</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9994</td>
<td>-0.9871</td>
<td>-0.9952</td>
<td>-0.9951</td>
<td>-0.9855</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9999</td>
<td>-0.9886</td>
<td>-0.9977</td>
<td>-0.9995</td>
<td>-0.9856</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9999</td>
<td>-0.9885</td>
<td>-0.9996</td>
<td>-0.9994</td>
<td>-0.9865</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9999</td>
<td>-0.9884</td>
<td>-0.9996</td>
<td>-0.9995</td>
<td>-0.9871</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9999</td>
<td>-0.9885</td>
<td>-0.9996</td>
<td>-0.9996</td>
<td>-0.9872</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9999</td>
<td>-0.9885</td>
<td>-0.9995</td>
<td>-0.9996</td>
<td>-0.9872</td>
</tr>
</tbody>
</table>

### Table 3.13:

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>50%</td>
<td>-0.9902</td>
<td>-0.9999</td>
<td>-0.9910</td>
<td>-0.9920</td>
<td>-0.9810</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9904</td>
<td>-0.9999</td>
<td>-0.9921</td>
<td>-0.9922</td>
<td>-0.9809</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9904</td>
<td>-0.9999</td>
<td>-0.9917</td>
<td>-0.9923</td>
<td>-0.9809</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9902</td>
<td>-0.9999</td>
<td>-0.9913</td>
<td>-0.9919</td>
<td>-0.9800</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9902</td>
<td>-0.9999</td>
<td>-0.9913</td>
<td>-0.9919</td>
<td>-0.9800</td>
</tr>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9979</td>
<td>-0.9881</td>
<td>-0.9986</td>
<td>-0.9987</td>
<td>-0.9873</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9976</td>
<td>-0.9908</td>
<td>-0.9993</td>
<td>-0.9984</td>
<td>-0.9872</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9975</td>
<td>-0.9909</td>
<td>-0.9992</td>
<td>-0.9983</td>
<td>-0.9871</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9975</td>
<td>-0.9909</td>
<td>-0.9992</td>
<td>-0.9983</td>
<td>-0.9871</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9975</td>
<td>-0.9912</td>
<td>-0.9992</td>
<td>-0.9984</td>
<td>-0.9869</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9999</td>
<td>-0.9885</td>
<td>-0.9999</td>
<td>-0.9992</td>
<td>-0.9862</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9999</td>
<td>-0.9885</td>
<td>-0.9999</td>
<td>-0.9993</td>
<td>-0.9865</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9999</td>
<td>-0.9881</td>
<td>-0.9999</td>
<td>-0.9993</td>
<td>-0.9865</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9999</td>
<td>-0.9881</td>
<td>-0.9999</td>
<td>-0.9993</td>
<td>-0.9865</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9999</td>
<td>-0.9881</td>
<td>-0.9999</td>
<td>-0.9993</td>
<td>-0.9865</td>
</tr>
</tbody>
</table>
3. Instrumental Evaluation Framework

### Table 3.14: Correlation coefficients between the instrumental $\Delta \Psi_w^L$ measurements and ACR listening tests for narrowband noise signals employing a Hamming window and different DFT lengths $K$, frame lengths $L$, and frame shifts $\Delta L$ for computing $\Delta \Psi_w^L$

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>50%</td>
<td>-0.9907</td>
<td>-0.9997</td>
<td>-0.9919</td>
<td>-0.9922</td>
<td>-0.9810</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9911</td>
<td>-0.9998</td>
<td>-0.9919</td>
<td>-0.9925</td>
<td>-0.9805</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9909</td>
<td>-0.9998</td>
<td>-0.9924</td>
<td>-0.9926</td>
<td>-0.9807</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9908</td>
<td>-0.9998</td>
<td>-0.9922</td>
<td>-0.9925</td>
<td>-0.9805</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9908</td>
<td>-0.9998</td>
<td>-0.9921</td>
<td>-0.9925</td>
<td>-0.9805</td>
</tr>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9989</td>
<td>-0.9915</td>
<td>-0.9994</td>
<td>-0.9993</td>
<td>-0.9875</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9987</td>
<td>-0.9901</td>
<td>-0.9999</td>
<td>-0.9999</td>
<td>-0.9874</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9986</td>
<td>-0.9901</td>
<td>-0.9997</td>
<td>-0.9999</td>
<td>-0.9873</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9986</td>
<td>-0.9901</td>
<td>-0.9997</td>
<td>-0.9999</td>
<td>-0.9872</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9986</td>
<td>-0.9901</td>
<td>-0.9997</td>
<td>-0.9999</td>
<td>-0.9872</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9999</td>
<td>-0.9883</td>
<td>-0.9997</td>
<td>-0.9970</td>
<td>-0.9863</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9999</td>
<td>-0.9885</td>
<td>-0.9997</td>
<td>-0.9998</td>
<td>-0.9866</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9999</td>
<td>-0.9884</td>
<td>-0.9997</td>
<td>-0.9998</td>
<td>-0.9869</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9999</td>
<td>-0.9883</td>
<td>-0.9997</td>
<td>-0.9997</td>
<td>-0.9870</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9999</td>
<td>-0.9883</td>
<td>-0.9997</td>
<td>-0.9997</td>
<td>-0.9870</td>
</tr>
</tbody>
</table>

### Table 3.15: Correlation coefficients between the instrumental $\Delta \Psi_w^L$ measurements and ACR listening tests for narrowband noise signals employing a Blackman window and different DFT lengths $K$, frame lengths $L$, and frame shifts $\Delta L$ for computing $\Delta \Psi_w^L$

<table>
<thead>
<tr>
<th>$L$</th>
<th>$\Delta L$</th>
<th>$\rho_{SG}$</th>
<th>$\rho_{WF}$</th>
<th>$\rho_{LSA}$</th>
<th>$\rho_{SA}$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>50%</td>
<td>-0.9869</td>
<td>-0.9999</td>
<td>-0.9883</td>
<td>-0.9900</td>
<td>-0.9770</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9873</td>
<td>-0.9999</td>
<td>-0.9891</td>
<td>-0.9901</td>
<td>-0.9774</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9872</td>
<td>-0.9999</td>
<td>-0.9894</td>
<td>-0.9901</td>
<td>-0.9774</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9870</td>
<td>-0.9999</td>
<td>-0.9890</td>
<td>-0.9894</td>
<td>-0.9764</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9870</td>
<td>-0.9999</td>
<td>-0.9890</td>
<td>-0.9894</td>
<td>-0.9764</td>
</tr>
<tr>
<td>256</td>
<td>50%</td>
<td>-0.9998</td>
<td>-0.9913</td>
<td>-0.9976</td>
<td>-0.9987</td>
<td>-0.9873</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9989</td>
<td>-0.9944</td>
<td>-0.9975</td>
<td>-0.9989</td>
<td>-0.9871</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9988</td>
<td>-0.9944</td>
<td>-0.9972</td>
<td>-0.9988</td>
<td>-0.9869</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9988</td>
<td>-0.9944</td>
<td>-0.9972</td>
<td>-0.9988</td>
<td>-0.9870</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9988</td>
<td>-0.9945</td>
<td>-0.9972</td>
<td>-0.9988</td>
<td>-0.9870</td>
</tr>
<tr>
<td>512</td>
<td>50%</td>
<td>-0.9999</td>
<td>-0.9879</td>
<td>-0.9997</td>
<td>-0.9986</td>
<td>-0.9865</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.9999</td>
<td>-0.9880</td>
<td>-0.9999</td>
<td>-0.9988</td>
<td>-0.9864</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>-0.9999</td>
<td>-0.9882</td>
<td>-0.9999</td>
<td>-0.9992</td>
<td>-0.9867</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>-0.9999</td>
<td>-0.9883</td>
<td>-0.9999</td>
<td>-0.9992</td>
<td>-0.9868</td>
</tr>
<tr>
<td></td>
<td>3.125%</td>
<td>-0.9999</td>
<td>-0.9883</td>
<td>-0.9999</td>
<td>-0.9992</td>
<td>-0.9868</td>
</tr>
</tbody>
</table>
3.4 Instrumental Optimization Using a Figure of Merit (FoM)

The optimization of a noise reduction algorithm can reasonably be conducted with instrumental quality measures for the speech component quality, the level of noise reduction, and the amount of musical tones. In Sections 3.2 and 3.3 we have presented MOS\(\tilde{s}\)\(_{LQO}\) (3.2.3) for instrumentally measuring the speech component quality, \(\Delta\text{SNR}\) (3.3.5) for instrumentally measuring the level of noise attenuation, and \(\Delta\Psi_w\)\(_{\log}\) (3.3.21) for instrumentally measuring the amount of musical tones, respectively. In this section, we show that these three instrumental quality measures are independent of each other and can thus be applied to formulate a figure of merit (FoM) providing a useful means to instrumentally optimize noise reduction algorithms.

3.4.1 Independency of Instrumental Quality Measures

An investigation of the relationship between the three instrumental measures MOS\(\tilde{s}\)\(_{LQO}\), \(\Delta\text{SNR}\), and \(\Delta\Psi_w\)\(_{\log}\) has to be performed, showing their independence, which is particular important when applying them to search for the optimal parameterization for a noise reduction algorithm. First of all, all filtered signals including the enhanced speech signal \(\hat{s}(n)\), the filtered speech signal \(\tilde{s}(n)\), and the filtered noise signal \(\tilde{n}(n)\) must be considered statistically dependent due to the common influence of the spectral gain \(G(\ell,k)\), even if the clean speech signal \(s(n)\) and the noise signal \(n(n)\) are statistically independent. Hence, this property always correlates results of instrumental quality measures that are based on these filtered signals. However, when we consider independency of instrumental quality measures, we mean that the measurement of the speech component quality shall not involve the noise component, the measurement of noise attenuation shall not be dependent on the speech component and on the noise distortion, and the measurement of noise distortion (i.e., the amount of musical tones) shall neither involve the speech component nor being dependent on the level of noise attenuation (at least a fullband factor of noise attenuation).

We can clearly claim such independence between MOS\(\tilde{s}\)\(_{LQO}\) for measuring the speech component quality and \(\Delta\Psi_w\)\(_{\log}\) for measuring musical tones, since MOS\(\tilde{s}\)\(_{LQO}\) (3.2.3) utilizes speech components only and \(\Delta\Psi_w\)\(_{\log}\) (3.3.21) utilizes noise components only. Due to the fact that the MOS\(\tilde{s}\)\(_{LQO}\) measure has a fullband scaling between \(s(n)\) and \(\tilde{s}(n)\) (ITU-T, 2001) and \(\Delta\text{SNR}\) only measures the relative SNR improvements, the terms MOS\(\tilde{s}\)\(_{LQO}\) and \(\Delta\text{SNR}\) can also be treated as independent of each other. However, the independence between \(\Delta\text{SNR}\) and \(\Delta\Psi_w\)\(_{\log}\) is not so obvious, since both terms are computed based on \(n(n)\) and \(\tilde{n}(n)\). Since \(\Delta\Psi_{\log}\) shall only measure noise distortion, but no fullband noise attenuation level, let us assume \(\tilde{N}(\ell,k) = G \cdot N(\ell,k)\) with a fullband noise attenuation factor \(0 \leq G \leq 1\) being constant in time and frequency. The idea is to prove \(\Delta\Psi_w = 0\), meaning no noise distortion according to our musical tones measure. Inserting \(G \cdot N(\ell,k)\) into the equation for \(\alpha_{\tilde{n}}(k)\) analogy to (3.3.20), the normalization factor \(\alpha_{\tilde{n}}(k)\) can be
computed as
\[
\alpha_n(k) = \left( \frac{1}{C(\Lambda_{H_0})} \sum_{\ell \in \Lambda_{H_0}} G^2 \cdot |N(\ell, k)|^2 \right)^{-1} = \left( \frac{1}{C(\Lambda_{H_0})} \cdot G^2 \cdot \log \sum_{\ell \in \Lambda_{H_0}} |N(\ell, k)|^2 \right)^{-1} = G^{-2} \cdot \alpha_n(k) . \tag{3.4.1}
\]

The term \( \Psi_n^w(\ell) \) can then be computed by inserting \( \alpha_n(k) = G^{-2} \cdot \alpha_n(k) \) and \( G \cdot N(\ell, k) \) into the analog formula to (3.3.19) as
\[
\Psi_n^w(\ell) = \frac{1}{K} \sum_{k=0}^{K-1} \left[ G^{-2} \cdot \alpha_n(k) \cdot G^2 \cdot |N(\ell, k)|^2 - G^{-2} \cdot \alpha_n(k) \cdot G^2 \cdot |N(\ell, k)|^2 \right]^4 \left( \frac{1}{K} \sum_{k=0}^{K-1} \left[ G^{-2} \cdot \alpha_n(k) \cdot G^2 \cdot |N(\ell, k)|^2 - G^{-2} \cdot \alpha_n(k) \cdot G^2 \cdot |N(\ell, k)|^2 \right]^2 \right)^2
= \frac{1}{K} \sum_{k=0}^{K-1} \left[ \alpha_n(k) \cdot |N(\ell, k)|^2 - \alpha_n(k) \cdot |N(\ell, k)|^2 \right]^4 \left( \frac{1}{K} \sum_{k=0}^{K-1} \left[ \alpha_n(k) \cdot |N(\ell, k)|^2 - \alpha_n(k) \cdot |N(\ell, k)|^2 \right]^2 \right)^2
= \Psi_n^w(\ell) . \tag{3.4.2}
\]

This results in
\[
\Delta \Psi_n^w = \ln \left( \frac{\Psi_n^w}{\Psi_n^w} \right) = \ln \left( \frac{\Psi_n^w}{\Psi_n^w} \right) = 0 . \tag{3.4.3}
\]

Therefore, the independence between \( \Delta \text{SNR} \) being related to the effective fullband noise attenuation factor \( G \) and \( \Delta \Psi_n^w \) has been shown. Please note that the proof of independence between \( \Delta \text{SNR} \) and \( \Delta \Psi_n^w \) has been derived based on the fullband noise attenuation modeled by \( G \). Note that if \( G = G(k) \) is a frequency-dependent spectral gain we still can show \( \Delta \Psi_n^w = 0 \). However, if a spectral gain \( G = G(\ell, k) \) being dependent on time and frequency is applied, we cannot simply derive \( \Delta \Psi_n^w = 0 \). However, such a relationship should not be given. If \( \Delta \Psi_n^w = 0 \) could be derived for \( G = G(\ell, k) \), we will thus have \( \Psi_n^w(\ell) = \Psi_n^w(\ell) \) by using \( G = G(\ell, k) \) in (3.4.2). This means that by applying any values of \( G(\ell, k) \) no musical tones will be generated. However, musical tones are isolated peaks in the spectral domain which are generated only after noise reduction by using not correctly estimated \( G(\ell, k) \). If we could derive \( \Delta \Psi_n^w = 0 \) by using \( G(\ell, k) \), it would thus imply that \( \Delta \Psi_n^w \) cannot measure noise distortion in terms of musical tones. However, it has been shown in Section 3.3.5 that \( \Delta \Psi_n^w \) can instrumentally measure musical tones. Therefore, \( \Delta \Psi_n^w = 0 \) cannot be derived by using \( G = G(\ell, k) \).

### 3.4.2 FoM-Based Instrumental Optimization Method

During the optimization phase of a noise reduction algorithm, often a high effort is necessary to find the optimal parameterization by manually assessing different instrumental quality measures separately. Based on three independent instrumental measures MOSLQO, \( \Delta \text{SNR} \), and \( \Delta \Psi_n^w \), we propose a figure of merit (FoM), which is a combined entity of these
three independent instrumental measures as (Yu and Fingscheidt, 2011a)

$$\text{FoM}(\theta_p) = A \left( \frac{\text{MOS}^s_{\text{LQO}}(\theta_p)}{\text{MOS}^s_{\text{LQO,ref}}} \right)^a + B \left( \frac{\Delta \text{SNR}(\theta_p)}{\Delta \text{SNR}_{\text{ref}}} \right)^b - C \left( \frac{\Delta \Psi^w_{\text{log,ref}}(\theta_p)}{\Delta \Psi^w_{\text{log,ref}}} \right)^c,$$

(3.4.4)

where $\theta_p$ is an algorithm parameter to be optimized, $A$, $B$, $C$ are weighting factors, and $a$, $b$, $c$ are exponents, respectively. The reference set \{MOS$^s_{\text{LQO,ref}}$, $\Delta \text{SNR}_{\text{ref}}$, $\Delta \Psi^w_{\text{log,ref}}$\} is calculated by applying a reference value of $\theta_p$ in the noise reduction algorithm, which serves as the starting point of the optimization process. This reference set serves to normalize the three components within the FoM. A better value of $\theta_p$ shall yield then a higher FoM value. Hence, the term for $\Delta \Psi^w_{\text{log}}$ is applied as a subtractive term in (3.4.4), since the value of $\Delta \Psi^w_{\text{log}}$ shall be kept as small as possible to achieve musical tones as few as possible. The FoM-based instrumental optimization is then carried out to search for the optimal $\theta_p$ for different $\text{SNR}_{\text{in}}$ conditions. Therefore, we calculate

$$\overline{\text{FoM}}(\theta_p) = \frac{1}{C(\mathcal{I})} \sum_{\text{SNR}_{\text{in}} \in \mathcal{I}} \text{FoM}(\theta_p)\bigg|_{\text{SNR}_{\text{in}}}$$

$$= \frac{1}{C(\mathcal{I})} \left\{ \sum_{\text{SNR}_{\text{in}} \in \mathcal{I}} A \left( \frac{\text{MOS}^s_{\text{LQO}}(\theta_p)}{\text{MOS}^s_{\text{LQO,ref}}} \right)^a + \sum_{\text{SNR}_{\text{in}} \in \mathcal{I}} B \left( \frac{\Delta \text{SNR}(\theta_p)}{\Delta \text{SNR}_{\text{ref}}} \right)^b - \sum_{\text{SNR}_{\text{in}} \in \mathcal{I}} C \left( \frac{\Delta \Psi^w_{\text{log,ref}}(\theta_p)}{\Delta \Psi^w_{\text{log,ref}}} \right)^c \right\}\bigg|_{\text{SNR}_{\text{in}}}$$

$$= A \left( \frac{\text{MOS}_{\text{LQO}}(\theta_p)}{\text{MOS}_{\text{LQO,ref}}} \right)^a + B \left( \frac{\Delta \text{SNR}(\theta_p)}{\Delta \text{SNR}_{\text{ref}}} \right)^b - C \left( \frac{\Delta \Psi^w_{\text{log,ref}}(\theta_p)}{\Delta \Psi^w_{\text{log,ref}}} \right)^c,$$

(3.4.5)

with $\mathcal{I}$ being the set of $\text{SNR}_{\text{in}}$ conditions included in the optimization and $C(\mathcal{I})$ being the cardinality of set $\mathcal{I}$, respectively. Maximizing the value of $\overline{\text{FoM}}(\theta_p)$ according to

$$\theta_{p,\text{opt}} = \arg \max_{\theta_p} \{ \overline{\text{FoM}}(\theta_p) \},$$

(3.4.6)

yields the optimal parameter utilizing the three independent instrumental quality measures for all $\text{SNR}_{\text{in}} \in \mathcal{I}$.

The weighting factors $A$, $B$, $C$ and the exponents $a$, $b$, $c$ have to be appropriately chosen in (3.4.5). Generally, two possibilities exist to choose the weighting factors and exponents:

- Having a number of different noise reduction algorithms at hand, if the optimal parameter $\theta_{p,\text{opt}}$ of one noise reduction algorithm is known, the weighting factors and exponents shall then be chosen in such a manner, that the known optimal parameter results into the maximum $\overline{\text{FoM}}$ in (3.4.6) for this noise reduction algorithm. Subsequently, applying exactly the same weighting factors and exponents, the (hopefully) optimal parameters for the other noise reduction algorithms can then be obtained automatically. This possibility is discussed in detail in Section 3.5, where the optimal smoothing factor $\beta$ and the optimal $a$ priori $\text{SNR}$ floor $\xi_{\text{min}}$ of the DD approach (3.3.18) to estimate the $a$ priori $\text{SNR}$ is known for the SA weighting rule. The
weighting factors and exponents can thus be obtained by achieving the maximum FoM (3.4.6) for SA by using these two well-known optimal parameters. Applying the same weighting factors and exponents, the optimal values of $\beta$ and $\xi_{\text{min}}$ can be obtained automatically for the LSA, WF and SG weighting rules, respectively.

- In case that the optimal parameter $\theta_{p,\text{opt}}$ is unknown to any of noise reduction algorithms at hand, the weighting factors and exponents can also be chosen in such a manner, that the importance of the speech-related instrumental quality measures and the noise-related instrumental quality measures represented by $\text{MOS}_{LQO}^\lambda$, $\Delta\text{SNR}$, and $\Delta\Psi_{w}^\log$ are balanced. This possibility is employed in Chapters 5 and 6 by applying the FoM to optimize the parameters of two new post-filter designs, whose optimal values are just unknown.

### 3.5 FoM-Based Single-Channel Optimization of an A Priori SNR Estimation

For single-channel noise reduction algorithms, the most important component to be estimated is the a priori SNR $\xi(\ell,k)$. It can well be estimated by the decision-directed (DD) approach (Ephraim and Malah, 1984) given in (3.3.18). Within the DD approach, the smoothing factor $\beta$ and the a priori SNR floor $\xi_{\text{min}}$ have to be chosen appropriately. In literature, the heuristically optimized value of $\beta = 0.98$ (Ephraim and Malah, 1984) and $\xi_{\text{min}} = -15$ dB (Cappé, 1994) for the SA weighting rule have been commonly used as the optimal values for different noise reduction algorithms applying the DD approach (sampling frequency $f_s = 8$ kHz and often also for sampling frequency $f_s = 16$ kHz). However, these values are not necessarily the optimal values for other weighting rules employing the DD approach. Therefore, in the following, the values of $\beta$ and $\xi_{\text{min}}$ are jointly optimized for the other three noise reduction algorithms (LSA, WF, and SG) listed in Table 3.2 by applying the FoM-based instrumental optimization. We add each of the 18 ETSI in-car background noise signals to 16 clean speech signals (four male and four female speakers) taken from the NTT speech database (NTT, 1994), each having a length of 8 s. Hence, $16 \times 18 = 288$ noisy signals are obtained for each $\text{SNR}_\text{in}$ condition. The $\text{SNR}_\text{in}$ conditions of the noisy speech signals are taken from the set $\mathcal{S} = \{-5, 0, 5, 10, 15, 20\}$ dB.

#### 3.5.1 Approach

To illustrate the possibility to automatically parameterize the DD a priori SNR estimator w.r.t. $\beta$ and $\xi_{\text{min}}$ using an FoM, we firstly evaluate the four weighting rules employing the three independent instrumental measures $\text{MOS}_{LQO}^\lambda$, $\Delta\text{SNR}$, and $\Delta\Psi_{w}^\log$. Since the value of $\beta$ is typically being chosen around 0.98, a search of $\beta$ is carried out within the interval $\beta \in [0.96, 0.999]$. The value of $\beta$ is increased by applying the same method as described in Section 3.3.4, so that more $\beta$ values are generated when $\beta$ is closer to one. Furthermore, a typical $\xi_{\text{min}}$ is chosen around $-15$ dB. Therefore, a search of $\xi_{\text{min}}$ is performed within
\[ \xi_{\text{min}} \in [-20, -10] \text{dB} \] by applying a step-size of 1 dB. Combing each \( \beta \) value with each \( \xi_{\text{min}} \) value, a joint search can thus be performed for \( \beta \in [0.96, 0.999] \) and \( \xi_{\text{min}} \in [-20, -10] \text{dB} \), respectively.

### 3.5.2 Performance Analysis

Fig. 3.8 shows the MOS\textsuperscript{LQO} results of the SA, LSA, WF, SG weighting rules for different SNR\textsubscript{in} conditions. Within each \( \xi_{\text{min}} \) value, it can be observed that for typical values of the DD approach with \( \beta \geq 0.96 \), MOS\textsuperscript{LQO} scores decrease monotonically when \( \beta \) is being increased. This matches the statement in (Cappé, 1994) that with \( \beta \) being chosen close to one, transient distortions for the speech component will unfortunately occur due to a strong smoothing in (3.3.18). In addition, for each fixed \( \beta \), MOS\textsuperscript{LQO} scores also slightly increase monotonically when \( \xi_{\text{min}} \) is being increased from \(-20 \text{ dB}\) to \(-10 \text{ dB}\). This corresponds to the analysis of Breithaupt and Martin (2011) that \( \xi_{\text{min}} \) indirectly limits the spectral gain \( G(\ell, k) \) of the weighting rule and higher \( \xi_{\text{min}} \), i.e., higher limit for \( G(\ell, k) \), will result into less speech distortion with a lower noise attenuation performance. Therefore, concerning the speech component quality only, a small \( \beta \) close to 0.96 and a high \( \xi_{\text{min}} \) close to \(-10 \text{ dB}\) should be chosen, respectively. Furthermore, it can be seen that SA and LSA can preserve the speech component better than WF and SG by providing higher MOS\textsuperscript{LQO} scores for all SNR\textsubscript{in} conditions with different \( \beta \) and \( \xi_{\text{min}} \) combinations in Fig.3.8.

The results of ASNR for the SA, LSA, WF, SG weighting rules for different SNR\textsubscript{in} conditions are depicted in Fig.3.9. It can be seen that with increasing \( \beta \) the noise attenuation performance increases monotonically. Concerning \( \xi_{\text{min}} \) values, it can be seen that noise attenuation performance decreases for each \( \beta \) value when \( \xi_{\text{min}} \) is being increased from \(-20 \text{ dB}\) to \(-10 \text{ dB}\). This is also due to the indirect influence of \( \xi_{\text{min}} \) on the spectral gain (Breithaupt and Martin, 2011) as just discussed. It turns out that concerning noise attenuation a large \( \beta \) close to one and a low \( \xi_{\text{min}} \) (e.g., \(-20 \text{ dB}\)) should be chosen. In addition, it can be observed that WF and SG provide much better noise attenuation performance than SA and LSA for all SNR\textsubscript{in} conditions in Fig.3.9.

The most cited benefit of employing the DD approach is to reduce musical tones especially in low SNR\textsubscript{in} conditions (i.e., SNR\textsubscript{in} \( \in \{-5, 0\} \text{ dB}\)) by setting \( \beta \) close to one, which carries out a smoothing procedure obtaining a more consistent \( \xi(\ell, k) \) estimate (Cappé, 1994). In addition, applying a \( \xi_{\text{min}} \) value can further eliminate musical tones by limiting the improbable \( \xi(\ell, k) \) estimate (Breithaupt and Martin, 2011). Fig. 3.10 presents the \( \Delta\Psi_{\text{log}}^w \) measurements for different weighting rules and SNR\textsubscript{in} conditions. As expected, due to the strong smoothing effect of \( \beta \) in the region of \( \beta \geq 0.96 \), the value of \( \Delta\Psi_{\text{log}}^w \) decreases rapidly with increasing \( \beta \) for all weighting rules, especially for the lower SNR\textsubscript{in} conditions of SNR\textsubscript{in} \( \in \{-5, 0\} \text{ dB}\). The value of \( \Delta\Psi_{\text{log}}^w \) decreases with \( \xi_{\text{min}} \) being increased from \(-20 \text{ dB}\) to \(-10 \text{ dB}\). Therefore, a large \( \beta \) close to one and a high \( \xi_{\text{min}} \) (e.g., \(-10 \text{ dB}\)) are preferred to suppress musical tones. Furthermore, the lower smoothing effect of SG and WF than of LSA and SA (Breithaupt and Martin, 2011) can also be observed, since \( \Delta\Psi_{\text{log}}^w \) of SG and WF are always higher than LSA and SA for all \( \beta \) values and SNR\textsubscript{in} conditions in Fig.3.10.
Figure 3.8: Speech component quality measure $\text{MOS}_{\text{LQO}}^S(\beta, \xi_{\text{min}})$ for (a) the SA weighting rule, (b) the LSA weighting rule, (c) the WF weighting rule, and (d) the SG weighing rule, respectively, with $\text{SNR}_{\text{in}}$ conditions being taken from the set $\mathcal{S} = \{-5, 0, 5, 10, 15, 20\}$ dB
Figure 3.9: SNR improvement measure $\Delta \text{SNR}(\beta, \xi_{\text{min}})$ for (a) the SA weighting rule, (b) the LSA weighting rule, (c) the WF weighting rule, and (d) the SG weighing rule, respectively, with SNR$_{\text{in}}$ conditions being taken from the set $\mathcal{S} = \{-5, 0, 5, 10, 15, 20\}$ dB
Figure 3.10: Weighted log kurtosis ratio measure $\Delta \Psi_{\log}^w(\beta, \xi_{\text{min}})$ for (a) the SA weighting rule, (b) the LSA weighting rule, (c) the WF weighting rule, and (d) the SG weighing rule, respectively, with SNR$_{\text{in}}$ conditions being taken from the set $\mathcal{S} = \{-5, 0, 5, 10, 15, 20\}$ dB
3.5.2. Performance Analysis

Generally, it can be seen from Figs. 3.8 to 3.10 that $\beta$ can influence the performance of noise reduction algorithms significantly, and $\xi_{\text{min}}$ provides an additional influence beyond the $\beta$ value. However, when we consider the overall noise reduction performance considering the speech component quality, the level of noise attenuation, and the amount of musical tones, there exist a trade-off of choosing the optimal values of $\beta$ and $\xi_{\text{min}}$. Taking the SA weighting rule into account, the widespread choice of $\beta = 0.98$ (Ephraim and Malah, 1984) and $\xi_{\text{min}} = -15\,\text{dB}$ (Cappé, 1994) with the DD approach actually provides a good trade-off between a good speech component quality, a high noise attenuation level and only few musical tones. However, as shown in Figs. 3.8 to 3.10, different weighting rules reveal different degrees of change for the three instrumental quality measurements w.r.t. $\beta$ and $\xi_{\text{min}}$. Therefore, it is worth to optimize both parameters for each weighting rule individually.

FoM-Based Instrumental Optimization

Based on the results of $\text{MOS}^d_{\text{LQO}, \text{ref}}$, $\Delta\text{SNR}$, and $\Delta\Psi^w_{\text{log}, \text{ref}}$ for the SA, LSA, WF, and SG weighting rules, respectively, the FoM-based instrumental optimization of $\beta$ and $\xi_{\text{min}}$ for the DD approach is now carried out. In this case, we will then have different combinations of $\theta_p = (\beta, \xi_{\text{min}})$ with $\beta \in [0.96, 0.999]$ and $\xi_{\text{min}} \in [-20, -10]\,\text{dB}$, respectively. The first step is to determine the weighting factors and exponents for calculating the FoM in (3.4.5). Since $\beta = 0.98$ (Ephraim and Malah, 1984) and $\xi_{\text{min}} = -15\,\text{dB}$ (Cappé, 1994) are the well-known optimal values for the SA weighting rule, the first possibility of choosing the weighting factors and exponents (see Section 3.4.2) by utilizing a reference noise reduction algorithm with known optimal parameters. Therefore, the term $\overline{\text{FoM}}(\beta, \xi_{\text{min}})$ of SA for different SNR in conditions is calculated by employing different combinations of $(A, B, C, a, b, c)$ with $A, B, C$ being chosen from 0.1 to 1 with a step increment of 0.1 and $a, b, c$ being chosen from $\{0.5, 1, 2\}$, respectively. The reference set $\{\text{MOS}^d_{\text{LQO,ref}}, \Delta\text{SNR}_{\text{ref}}, \Delta\Psi^w_{\text{log,ref}}\}$ in (3.4.5) is calculated by applying $\beta = 0.98$ and $\xi_{\text{min}} = -15\,\text{dB}$ to the SA weighting rule. It turns out that for the SA weighting rule in total two combinations of $(A, B, C, a, b, c)$ result in a maximum value of $\overline{\text{FoM}}(\beta, \xi_{\text{min}})$ with $\beta_{\text{opt}} = 0.98$ and $\xi_{\text{min,op}} = -15\,\text{dB}$. These two combinations are listed in Table 3.16.

<table>
<thead>
<tr>
<th>Comb.</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.7</td>
<td>0.3</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.7</td>
<td>0.1</td>
<td>2</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 3.16: Combinations of $(A, B, C, a, b, c)$ yielding the optimal $\beta_{\text{opt}} = 0.98$ and $\xi_{\text{min,op}} = -15\,\text{dB}$ for the SA weighting rule

Applying both combinations of $(A, B, C, a, b, c)$ of Table 3.16, the $\overline{\text{FoM}}(\beta, \xi_{\text{min}})$ (3.4.5) is then computed for the other three weighting rules, namely, LSA, WF, and SG, respectively. For each weighting rule, the reference set $\{\text{MOS}^d_{\text{LQO,ref}}, \Delta\text{SNR}_{\text{ref}}, \Delta\Psi^w_{\text{log,ref}}\}$ is
calculated by applying the commonly used $\beta = 0.98$ and $\xi_{\text{min}} = -15\,\text{dB}$. We found that both combinations of $(A, B, C, a, b, c)$ yield the same optimal values of $\beta_{\text{opt}}$ and $\xi_{\text{min,opt}}$, resulting into the maximum $\text{FoM}(\beta, \xi_{\text{min}})$ for each of the three weighting rules. Choosing the first combination of $(A, B, C, a, b, c) = (0.6, 0.7, 0.3, 1, 0.5, 0.5)$ for the SA and LSA weighting rules as well as for the WF and SG weighting rules, the $\text{FoM}(\beta, \xi_{\text{min}})$ results are shown in Fig. 3.11 and Fig. 3.12, respectively. It can be seen in Fig. 3.11 that the same $\xi_{\text{min,opt}} = -15\,\text{dB}$ can be obtained for SA and LSA. However, no optimal value is proposed for LSA in Ephraim and Malah (1985). Using the FoM-based instrumental optimization, $\beta_{\text{opt}} = 0.975$ instead of 0.98 is obtained for LSA, which is a bit smaller than 0.98. Fig. 3.12 shows that an optimal $\beta_{\text{opt}} \approx 0.99$ is obtained for WF and SG (in detail: $\beta_{\text{opt}} = 0.99$ for WF and $\beta_{\text{opt}} = 0.993$ for SG, respectively). These are very surprising and interesting results, since in many cases, the value $\beta = 0.98$ has been also adopted for the WF and SG weighting rule, e.g., in (Scalart and Filho, 1996), (Erkelens et al., 2007), and (Suhadi, 2012). Furthermore, the optimal $\xi_{\text{min,opt}} = -14\,\text{dB}$ is obtained for WF and SG, which is just slightly smaller than the commonly applied $-15\,\text{dB}$. The newly obtained optimal $\beta_{\text{opt}}$ and $\xi_{\text{min,opt}}$ values are summarized in Table 3.17. As we can see now, the proposed $\beta = 0.98$ and $\xi_{\text{min}} = -15\,\text{dB}$ for SA may not always be the necessary optimal values for other weighting rules, especially for $\beta$.

<table>
<thead>
<tr>
<th></th>
<th>SA</th>
<th>LSA</th>
<th>WF</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{\text{opt}}$</td>
<td>0.98</td>
<td>0.975</td>
<td>0.99</td>
<td>0.993</td>
</tr>
<tr>
<td>$\xi_{\text{min,opt}}$ [dB]</td>
<td>-15</td>
<td>-15</td>
<td>-14</td>
<td>-14</td>
</tr>
</tbody>
</table>

Table 3.17: Optimal $\beta_{\text{opt}}$ and $\xi_{\text{min,opt}}$ resulting into the maximum $\text{FoM}(\beta, \xi_{\text{min}})$ by using $(A, B, C, a, b, c) = (0.6, 0.7, 0.3, 1, 0.5, 0.5)$ in (3.4.5) for the SA, LSA, WF, and SG weighting rules, respectively.

Subjective Listening Test

In order to verify the optimal $\beta_{\text{opt}}$ and $\xi_{\text{min,opt}}$ provided by the FoM-based instrumental optimization, a subjective listening test has been conducted for the weighing rule SG in an ABX fashion with 16 test persons (experts and non-experts) (Yu and Fingscheidt, 2011a). The listeners had to give their preference to either SG with the DD approach using $\beta = 0.98$ (A) or using $\beta = 0.993$ (B). No preference cases were marked by an X. Please note that in (Yu and Fingscheidt, 2011a) only the smoothing factor $\beta$ has been optimized to be also 0.993, while $\xi_{\text{min,opt}}$ was still assumed to be $-15\,\text{dB}$. In this thesis, $\xi_{\text{min,opt}} = -14\,\text{dB}$ is obtained for SG by jointly optimizing $\beta$ and $\xi_{\text{min}}$. Nevertheless, by observing Fig. 3.12, it can be seen that under the same $\beta = 0.993$, the difference between $\text{FoM}(\beta, \xi_{\text{min}})$ by using $\xi_{\text{min}} = -14\,\text{dB}$ and $\xi_{\text{min}} = -15\,\text{dB}$ is quite small. Furthermore, according to informal listening tests, almost no difference can be identified between $\xi_{\text{min}} = -14\,\text{dB}$ and $\xi_{\text{min}} = -15\,\text{dB}$ by using $\beta = 0.993$. Therefore, the ABX listening test results from (Yu and Fingscheidt, 2011a) for the SG weighting rule are adopted here to verify the optimal $\beta_{\text{opt}}$ and $\xi_{\text{min,opt}}$ for the SG weighting rule. In the ABX test, three
Figure 3.11: $\text{FoM}(\beta, \xi_{\text{min}})$ results for (a) the SA weighting rule, and (b) the LSA weighting rule with $\beta \in [0.96, 0.999]$ and $\xi_{\text{min}} \in [-20, -10]$ dB, respectively.
Figure 3.12: FoM(β, ξ_{min}) results for (a) the WF weighting rule, and (b) the SG weighting rule with β ∈ [0.96, 0.999] and ξ_{min} ∈ [-20, -10] dB, respectively.
SNR<sub>in</sub> conditions were applied, namely −5, 5, and 15 dB. For each SNR<sub>in</sub> condition a total of four randomly chosen speech signals (two male and two female speakers) with four different noises were employed. Table 3.18 shows the preference results of all three SNR<sub>in</sub> conditions and the corresponding average results. It can be clearly observed that much more preference had been given to the FoM-based optimal $\beta = 0.993$ compared to $\beta = 0.98$. In average, only in 16.4% of the conditions no preference was reported. This results mostly from the SNR<sub>in</sub> = 15 dB condition, where 37.5% of the listeners were not able to hear differences at all. A significant improvement can be observed for the SG weighting rule by using $\beta = 0.993$ instead of $\beta = 0.98$ for SNR<sub>in</sub> = 5 dB. According to the feedback of most listeners, in SNR<sub>in</sub> = −5 dB condition, a trade-off between speech distortion and the amount of musical tones had to be judged. However, still in this case more listeners prefer the optimal $\beta = 0.993$ showing significant musical tones suppression and acceptable speech distortion. Therefore, this subjective listening test verifies that the FoM-based instrumental optimization is able to identify the surprising new values of the optimal smoothing factor for some noise reduction algorithms.

<table>
<thead>
<tr>
<th>SNR&lt;sub&gt;in&lt;/sub&gt; [dB]</th>
<th>-5</th>
<th>5</th>
<th>15</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ($\beta = 0.98$)</td>
<td>39.1%</td>
<td>6.3%</td>
<td>14.1%</td>
<td>19.8%</td>
</tr>
<tr>
<td>B ($\beta = 0.993$)</td>
<td>57.0%</td>
<td>85.9%</td>
<td>48.4%</td>
<td>63.8%</td>
</tr>
<tr>
<td>X</td>
<td>3.9%</td>
<td>7.8%</td>
<td>37.5%</td>
<td>16.4%</td>
</tr>
</tbody>
</table>

Table 3.18: ABX preference results for the SG weighting rule with different values of $\beta$ and $\xi_{\text{min}} = -15$ dB (Yu and Fingscheidt, 2011a)

3.6 Summary

A unified instrumental evaluation framework for noise reduction algorithms has been presented in its single-channel and multichannel setups. In addition to the noisy speech signal with the clean speech signal and the noise signal, the addressed instrumental evaluation framework can separate the enhanced speech signal into the filtered speech signal and the filtered noise signal. Based on these signals, several state-of-the-art speech- and noise-related instrumental quality measures for both wideband and narrowband noise reduction algorithms are described. In particular, a new weighted log kurtosis ratio has been proposed and subjectively verified to measure the amount of musical tones in a black-box manner. In comparison to the state-of-the-art white-box log kurtosis ratio, the black-box weighted log kurtosis ratio does not require any knowledge of internal variables of the noise reduction algorithms under test. Furthermore, the weighted log kurtosis ratio can be applied to a wide range of noise reduction algorithms including those employing the DD approach to estimate the a priori SNR. Further analyses have shown that the weighted log kurtosis ratio is quite insensitive to different calculation methods and measurement setups.

Combining the three independent instrumental quality measures for the speech com-
ponent quality, the level of noise attenuation, and the amount of musical tones, an entity defined as figure of merit (FoM) has been proposed. Based on the FoM, an instrumental optimization method can be utilized to automatically identify the optimal parameterization of a noise reduction algorithm. Applying the FoM-based instrumental optimization method, undocumented new optimal values of the smoothing factor and the \textit{a priori} SNR floor for the \textit{a priori} SNR estimation using the DD approach have been identified for some noise reduction algorithms, which have also been subjectively verified.
Chapter 4

Multichannel In-Car Speech, Noise, and CAN Bus Database

When dealing with optimization and evaluation of multichannel automotive speech enhancement systems, there is a need for a multichannel in-car speech and noise database. Furthermore, the controller area network bus (CAN bus) of modern cars can provide real-time driving conditions by supplying several key parameters. In this chapter, a multichannel in-car speech, noise and CAN bus database is presented. The application scenario comprising the applied head unit-integrated microphone array and a brief introduction of the CAN bus system is presented in Section 4.1. The data collection of speech signals, noise signals, and CAN bus data is described in Section 4.2. In addition, the post-processing to generate the multichannel database is specified. Finally, the applied data set taken from the multichannel database for the optimization and evaluation of beamforming and post-filtering systems in Chapters 5 and 6 are described in Section 4.3.

4.1 Application Scenario

A multichannel in-car speech, noise and CAN bus database has been acquired in a Volkswagen (VW) Touran. The necessary audio interfaces, amplifiers, and power supply were installed in the luggage compartment (see Fig. 4.1). For acquiring speech and noise signals, a head unit-integrated microphone array consisting of 30 low-cost omnidirectional electret microphones was realized in a prototype representing the Volkswagen infotainment system RNS-510\(^7\). All applied microphones are omnidirectional owning a uniform response to sound pressure coming from all directions up to 16kHz. This head unit-integrated microphone array is shown in Fig. 4.2(a). The distance of adjacent microphones is 1.2 cm. Integration of the microphone array into the head unit turns out to be cost-effective not only due to the use of low-cost microphones, but particularly because extensive wiring to some typical microphone positions as described in Section 1.1

\(^7\)RNS stands for radio and navigation system. In this prototype, the low-cost Monacor MCE-4500 omnidirectional electret microphone capsule with a frequency response up to 16 kHz was applied. This prototype was friendly provided by IAV GmbH.
Figure 4.1: (a) Employed research car VW Touran with (b) audio interfaces, amplifiers, and power supply being installed in the luggage compartment

can be omitted. In this thesis, due to the capacity of the applied audio interfaces and amplifiers, only the left vertical column with 10 microphones and the lower horizontal row with 10 microphones of the microphone array were employed. This is shown in detail in Fig. 4.2(b). Note that the left vertical column with 10 microphones was chosen, since the left vertical column is closer to the driver, who primarily uses the microphone array-based hands-free equipment. Furthermore, a separate hands-free microphone being installed in the overhead light module was also employed. Altogether 21 microphone signals have been recorded simultaneously. The detailed system setup for collecting speech and noise signals with applied audio equipment and hardware are described in Appendix A.

As explained in Chapter 2, the noise coherence function describing the noise field is very important to the multichannel noise reduction system utilizing beamforming and post-filtering. Therefore, it is of great interest to investigate noise coherence functions of different driving conditions.

The CAN bus is a serial automotive bus system developed by Robert Bosch GmbH in 1983 (Lawrenz and Obermüller, 2011). CAN bus provides serial communications between independent electronic control units (ECUs), like the engine control unit, the window control unit, and the door control unit. According to the ISO standards 11898-1 (ISO, 2003) specifying the CAN bus system, a bit rate up to 1 Mbit/s can be achieved for a network length below 40 m. The data sent by each ECU to the CAN bus is event-controlled, meaning that a data will only be sent if the current state of the ECU changes. Since the CAN bus is a serial bus system, the data sent by an ECU with higher priority will be transmitted at first, while an ECU with lower priority has to wait.

The applied VW Touran is equipped with the CAN bus system, which is capable of providing different parameters, some of them showing high acoustical correlates in a real-time manner. By supplying these acoustical-related parameters, the CAN bus is thus able to provide the driving condition in real-time. Therefore, in addition to the multichannel speech and noise signals, the CAN bus data have also been simultaneously acquired.
4.2 Data Collection

4.2.1 Speech Signals

The clean speech signals of the NTT speech database (NTT, 1994) using a sampling frequency of 16 kHz were applied to the multichannel in-car clean speech acquisition. The NTT speech database includes 21 languages with four male and four female native speakers being assigned to each language. For our multichannel in-car clean speech acquisition, two languages, namely English and German were selected. For each language, all eight speakers were employed with four clean utterances per speaker. Therefore, by utilizing two languages, in total 64 clean speech signals have been collected. Each clean speech signal has a length of 8s. During the clean speech acquisition, the VW Touran was parked

Figure 4.2: (a) A head unit-integrated microphone array with 30 omnidirectional microphones; (b) the employed left vertical column with 10 microphones and the bottom horizontal row with 10 microphones for the multichannel in-car speech, noise database collection; in this thesis, the four microphones marked with red color were employed to develop the beamforming and post-filtering systems described in Chapters 5 and 6

Figure 4.3: Setup of the Fostex 6301B loudspeaker for the acquisition of in-car clean speech signals at the driver’s position
Table 4.1: Car driving conditions with different setups of driving speed (Speed), engine state (Engine), air-conditioning (AC) operating level, and window (Window) position, respectively.

<table>
<thead>
<tr>
<th>Cond.</th>
<th>Speed [km/h]</th>
<th>Engine</th>
<th>AC</th>
<th>Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>off</td>
<td>0/7</td>
<td>closed</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>off</td>
<td>3/7</td>
<td>closed</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>off</td>
<td>7/7</td>
<td>closed</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>on</td>
<td>0/7</td>
<td>closed</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>on</td>
<td>3/7</td>
<td>closed</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>on</td>
<td>7/7</td>
<td>closed</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>on</td>
<td>0/7</td>
<td>closed</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>on</td>
<td>0/7</td>
<td>open</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>on</td>
<td>3/7</td>
<td>closed</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>on</td>
<td>3/7</td>
<td>open</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>on</td>
<td>7/7</td>
<td>closed</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>on</td>
<td>7/7</td>
<td>open</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>on</td>
<td>0/7</td>
<td>closed</td>
</tr>
<tr>
<td>14</td>
<td>120</td>
<td>on</td>
<td>0/7</td>
<td>open</td>
</tr>
<tr>
<td>15</td>
<td>120</td>
<td>on</td>
<td>3/7</td>
<td>closed</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>on</td>
<td>3/7</td>
<td>open</td>
</tr>
<tr>
<td>17</td>
<td>120</td>
<td>on</td>
<td>7/7</td>
<td>closed</td>
</tr>
<tr>
<td>18</td>
<td>120</td>
<td>on</td>
<td>7/7</td>
<td>open</td>
</tr>
</tbody>
</table>

in a quiet garage with all doors being closed and the engine being turned off. The same clean speech acquisition process was performed from both driver’s position and co-driver’s position. During each acquisition process, the 64 clean speech signals were played through a Fostex 6301B loudspeaker, which was fixed on the head rest of the corresponding seat. Fig. 4.3 shows the scenario of the loudspeaker being fixed on the head rest of the driver’s seat. Through the installation of the loudspeaker on the head rest, the distance and position of the driver’s head (or co-driver’s head) can be well approximated.

4.2.2 Noise and CAN Bus Signals

The in-car noise signals were acquired for different driving conditions. The driving conditions differentiate from each other with variable parameters including engine state, driving speed, air-conditioning operating level, and window position. Considering different typical combinations of these variable parameters, in total 18 driving conditions have been recorded, which are shown in Table 4.1. The driving conditions with 50 km/h have been recorded with the car being driven on the urban street, while the driving conditions with 120 km/h have been recorded on the highway. During each driving condition, a multi-channel noise acquisition of about 10 minutes was performed.

For each driving condition, the corresponding CAN bus data have also been recorded.
simultaneously. In order to collect the CAN bus data, a module called as CarGate, which was friendly provided by Volkswagen AG, was employed. CarGate works as the control server of the VW Touran and receives the corresponding CAN bus data via a transmission control protocol (TCP) connection. By initializing each CAN bus data collection, a request containing the selected CAN bus parameters was sent from CarGate. Whenever the value of one selected parameter was changed, this value was then collected by the CarGate and saved with the corresponding time stamp. The intervals for checking the value change of each parameter was set to be 100 ms to avoid huge data amounts and data congestion. The 100 ms interval is sufficient, since each driving condition can be treated as stationary with main acoustical-related parameters being nearly unchanged. For the employed VW Touran, 59 different parameters can be collected. However, for our multichannel database only parameters bearing acoustical influence were recorded. A detailed list of the acoustical-related parameters is given in Appendix B.

Note that the CAN bus data collection shall be synchronously performed along with the noise signals collection for each driving condition. However, since the CAN bus internal time is always initialized to 1970-01-01 00:00:00.000 by each power start of the car, a direct alignment between the noise signals and CAN bus data using local system time is not possible. In order to solve the time alignment problem, a simple yet robust method utilizing the horn signal was employed. For each driving condition, the CAN bus data collection was started at first and ended at last. During the simultaneous noise and CAN bus data collection, a horn signal was generated in the beginning and in the end of the noise signals collection. The horn signal was then collected as CAN bus data and was also simultaneously acoustically recorded. By using the CAN bus internal time of the two horn signals, the synchronization between the noise signals and CAN bus data can thus be solved. The minimal wave prorogation delay can be ignored in this case.

4.2.3 Post-Processing

Since the raw in-car speech and noise signals as well as the raw CAN bus data were collected in different lengths and formats, a post-processing was required to generate the multichannel in-car speech, noise, and CAN bus database. The multichannel speech and noise signals were acquired with a sampling frequency of 48 kHz. To achieve wideband signals for the database, the speech and noise signals were firstly downsampled to the sampling frequency of 16 kHz by employing the filter function of ITU-T Rec. G.191 (ITU-T, 2005a). Secondly, the P.341 weighting filter according to ITU-T Rec. P.341 (ITU-T, 1995) was applied to achieve the bandwidth of 50...7000 Hz.

Since the CAN bus data was collected only when the values of the selected parameters were changed, no fixed sampling frequency can be applied to the CAN bus data. In order to match the sampling frequency of the corresponding multichannel noise signals, an interpolation was performed for the CAN bus data. The interpolation was carried out by applying the same value of the CAN bus parameter for the subsequent noise samples until a value change of the CAN bus parameter can be identified. The new value of this parameter was then updated for the next coming noise samples. After matching the
sampling frequency of the CAN bus data with the noise signals, the about 10 minutes of noise signals were cut into 8s segments with their corresponding CAN bus data for each driving condition, which matches the length of the speech signal. Depending on the driving conditions, about 65 to 70 noise signals with a length of 8s were obtained for each driving condition.

4.3 Applied Data Set

In this thesis, we restrict ourselves to multichannel automotive speech enhancement for the driver. For all beamforming and post-filtering systems presented in Chapters 5 and 6, a vertical microphone array employing four microphones (marked with red color in Fig. 4.2(b)) with the distance of adjacent microphones of 3.6 cm is applied. Furthermore, for all performance analyses in Chapters 5 and 6, the following data set chosen from the multichannel in-car speech, noise, and CAN bus database is employed: For each driving condition, eight clean speech signals (four female and four male speakers) from the driver position as well as eight car noise signals are employed, which provides \(8 \times 8 = 64\) noisy signals, each having a length of 8s. The input SNR of the noisy speech signals is set to be \(\text{SNR}_{in} \in \mathcal{S} = \{-5, 0, 5, 10, 15, 20\}\) dB.

4.4 Summary

In this chapter, a multichannel in-car speech, noise, and CAN bus database, which was acquired in a VW Touran, has been described. The acquisition of the in-car speech and noise signals were carried out by applying a head unit-integrated microphone array. In order to achieve a simultaneous in-car speech and noise collection, an audio data collection system has been designed and realized in the VW Touran. The clean speech signals selected from the NTT speech database were recorded separately from the driver’s and co-driver’s positions. The acquisition of noise signals was performed for 18 driving conditions, which cover typical daily driving scenarios. In addition, the CAN bus data providing real-time information about driving conditions were also synchronously collected with noise signals for 18 driving conditions. The collection of the CAN bus data were performed by utilizing the CarGate. After the post-processing, 64 clean speech signals from both driver’s and co-driver’s positions and 65 to 70 noise signals along with CAN bus data were collected for each driving condition. Finally, the applied head unit-integrated vertical microphone array with four microphones and the corresponding data set for optimizing and evaluating beamforming and post-filtering systems in Chapters 5 and 6 have been described.
Chapter 5

Post-Filter Design with Modified PSD Estimation and Hybrid Noise Coherence Functions

For the post-filter estimation, the PSD estimation and the noise coherence function can be treated as the two most important components. Current approaches apply a fixed smoothing factor for the PSD estimation and employ the a priori knowledge of a diffuse noise field even in an automotive environment with varying driving conditions. In this chapter, a modified PSD estimation employing an adaptive smoothing factor and a hybrid noise coherence function are utilized to a new post-filter design. In addition, the CAN bus data of the multichannel in-car database presented in Chapter 4 can provide us the car driving condition in a real-time manner. Therefore, the adaptive smoothing factor for the PSD estimation and the hybrid noise coherence function will be instrumentally optimized for a specific driving condition by using an FoM as described in Section 3.4. The optimized parameters enjoy the potential benefit of being selected by the CAN bus data for the corresponding specific driving condition. In Section 5.1, four state-of-the-art post-filters are evaluated to select the suitable baselines for Chapters 5 and 6. Section 5.2 focuses on the modified PSD estimation along with the FoM-based instrumental optimization of the adaptive smoothing factor. In Section 5.3, the idea of the hybrid noise coherence function, which is a mixture of the diffuse and the measured noise coherence functions for a specific driving condition, is elaborated. Furthermore, the FoM-based instrumental optimization of the hybrid coherence functions for different driving conditions is addressed.

5.1 Baselines

5.1.1 Approaches

In this section, by applying the head unit-integrated vertical microphone array with four microphones as described in Section 4.3, four state-of-the-art post-filtering algorithms as
introduced in Section 2.4 are employed: Zelinski’s (ZE) post-filter (2.4.9) and Simmer’s (SI) post-filter (2.4.11) both utilizing the incoherent noise coherence function (inc), McCowan’s (MC) post-filter (2.4.18) and Lefkimmiatis’s (LE) (2.4.22) post-filter utilizing the diffuse noise coherence function (dif). Since the four post-filters apply a fixed smoothing factor $\alpha$ to the auto- and cross-PSD estimation of input microphone signals in (2.4.5) and (2.4.6), respectively, a fixed (F) $\alpha = 0.8$ is applied. Additionally, the floor value employed to different post-filters is set to be $H_{\text{min}} = 0.15$ for masking musical tones by keeping a small level of residual noise. Please note that in this thesis, the same floor value of $H_{\text{min}} = 0.15$ is applied to all post-filter designs. Therefore, the coefficients of all post-filters are limited to the interval of $[0,1]$. The aforementioned post-filters are combined with the delay-and-sum beamformer (DS) (2.3.23) or the constrained superdirective beamformer (CS) (2.3.27) as proposed in their original publications. In the case of CS, a value of $\varrho = 0.01$ is used. The four baseline systems are listed in Table 5.1.

<table>
<thead>
<tr>
<th>System</th>
<th>Beamforming</th>
<th>Post-filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-ZE/F, inc</td>
<td>DS (2.3.23)</td>
<td>ZE (2.4.9)</td>
</tr>
<tr>
<td>(Zelinski, 1988)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS-SI/F, inc</td>
<td>DS (2.3.23)</td>
<td>SI (2.4.11)</td>
</tr>
<tr>
<td>(Simmer and Wasiljeff, 1992)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-MC/F, dif</td>
<td>CS (2.3.27)</td>
<td>MC (2.4.18)</td>
</tr>
<tr>
<td>(McCowan and Bourlard, 2003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-LE/F, dif</td>
<td>CS (2.3.27)</td>
<td>LE (2.4.22)</td>
</tr>
<tr>
<td>(Lefkimmiatis and Maragos, 2006)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Four baseline beamforming and post-filtering systems

5.1.2 Performance Analysis

Along with the three instrumental measures $\Delta\text{SNR}$, $\text{MOS}_{\text{LQO}}^\delta$, and $\Delta\Psi_{\text{log}}^\delta$ as introduced in Chapter 3, an $\text{FoM}$ (3.4.5) is also applied. Please note that the $\text{FoM}$ is applied in this section only for the evaluation of the four baseline systems, not for the purpose of optimization. In order to calculate the $\text{FoM}$ in (3.4.5), the weighting factors and exponents shown in Table 5.2 are chosen, so that the weight $B$ for the SNR improvement $\Delta\text{SNR}$ equals the sum of the weights $A$ and $C$ for the speech component quality $\text{MOS}_{\text{LQO}}^\delta$ and the noise component quality (i.e., musical tones) $\Delta\Psi_{\text{log}}^\delta$, respectively. Therefore, the values of $(A, B, C, a, b, c)$ are chosen here by applying the second method as described in Section 3.4.2 for balancing the speech-related and noise-related instrumental measures. This method is used since the same $\text{FoM}$ is also applied to optimize the new post-filter designs in Chapters 5 and 6, whose optimal parameters are just unknown. In this section, the reference set $\{\text{MOS}_{\text{LQO},\text{ref}}^\delta, \Delta\text{SNR}_{\text{ref}}, \Delta\Psi_{\text{log},\text{ref}}^\delta\}$ in (3.4.5) is calculated by applying the (CS-MC/F, dif) system, leading to $\overline{\text{FoM}} = 1$. The averaged instrumental measurement results of $\text{MOS}_{\text{LQO}}^\delta$, $\Delta\text{SNR}$, $\Delta\Psi_{\text{log}}^\delta$, and $\text{FoM}$ of all driving conditions are shown in Fig. 5.1. Observing the $\overline{\text{FoM}}$ values, it can be found that the (CS-LE/F, dif) system
5.1.2. Performance Analysis

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.2: Chosen values of \((A, B, C, a, b, c)\) for the FoM computation (3.4.5) applied in Chapters 5 and 6

and the \((\text{CS-MC/F, dif})\) system achieve much better overall noise reduction performance than the \((\text{DS-SI/F, inc})\) system and the \((\text{DS-ZE/F, inc})\) system. Although the CS beamformer with \(\varrho = 0.01\) can provide slightly better directivity index and beam pattern than the DS beamformer for low frequencies as shown in Fig. 2.9(a) and Fig. 2.10, the noise reduction performance of the MVDR beamformer (i.e., the DS and the CS beamformer) can be ignored for car noise (Simmer and Wasiljeff, 1992). The main reason for

![Graph showing performance analysis results](image-url)

Figure 5.1: Average instrumental measurement results of all driving conditions for the \((\text{CS-LE/F, dif})\) system, the \((\text{CS-MC/F, dif})\) system, the \((\text{DS-SI/F, inc})\) system, and the \((\text{DS-ZE/F, inc})\) system, respectively

the difference of the overall noise reduction performance lies in the different post-filters (McCowan and Bourlard, 2003). Therefore, in the following, the performance analysis will mainly focus on the post-filters. As expected, employing the incoherent noise coherence function which does not satisfy the spatial characteristics of car noise, Simmer’s post-filter \((\text{DS-SI/F, inc})\) and Zelinski’s post-filter \((\text{DS-ZE/F, inc})\) hardly provide any SNR improvement indicated by a \(\Delta\text{SNR}\) value of about 0 dB for all \(\text{SNR}_{\text{in}}\) conditions. Therefore, even though Simmer’s and Zelinski’s post-filters can preserve the speech component
5. Post-Filter Design with Modified PSD Estimation and Hybrid Noise Coherence Functions

quite well by having a MOS\textsubscript{LQO} score of more than 3.9 for the (DS-SI/F, inc) system and more than 3.7 for the (DS-ZE/F, inc) system for all SNR\textsubscript{in} conditions, the overall noise reduction performances of both systems indicated by the FoM are kept still very low. Utilizing the \textit{a priori} knowledge of the diffuse noise field approximating the spatial characteristics of car noise, Lefkimmiatis’s post-filter (CS-LE/F, dif) and McCowan’s post-filter (CS-MC/F, dif) show significantly improved ΔSNR values against Simmer’s and Zelinski’s post-filters. Between Lefkimmiatis’s and McCowan’s post-filters, there is only a small difference of the ΔSNR values for different SNR\textsubscript{in} conditions. Concerning the speech component quality, a performance degradation of Lefkimmiatis’s post-filter can be detected by showing about 0.3 lower MOS\textsubscript{LQO} scores than McCowan’s post-filter for different SNR\textsubscript{in} conditions. However, in comparison to McCowan’s post-filter, Lefkimmiatis’s post-filter generates much less musical tones by showing much lower ΔΨ\textsubscript{log} values for different SNR\textsubscript{in} conditions. Therefore, a trade-off exists between the speech component quality and the amount of musical tones for Lefkimmiatis’s and McCowan’s post-filters. Nevertheless, both Lefkimmiatis’s and McCowan’s post-filters can deliver improved overall noise reduction performance when compared to Simmer’s and Zelinski’s post-filters by exploiting the diffuse noise coherence function.

Based on the performance analysis, Lefkimmiatis’s post-filter with constrained superdirective beamforming (CS-LE/F, dif) and McCowan’s post-filter with constrained superdirective beamforming (CS-MC/F, dif), both applying a fixed smoothing factor α = 0.8 and the diffuse noise coherence function are employed as two baseline systems for evaluating the new post-filter designs in Chapters 5 and 6.

5.2 PSD Estimation with an Adaptive Smoothing Factor

As described in the previous section, state-of-the-art approaches employ a fixed smoothing factor α to the auto- and cross-PSD estimation of input microphone signals in (2.4.5) and (2.4.6), respectively. In this section, a new modified PSD estimation by applying an adaptive smoothing factor is addressed (Yu and Fingscheidt, 2010c,b). Furthermore, the FoM-based instrumental optimization of the adaptive smoothing factor for different driving conditions is also presented.

5.2.1 Approach

Commonly, a fixed smoothing factor close to one is employed to achieve a smooth auto- and cross-PSD estimation of input microphone signals in (2.4.5) and (2.4.6) with reduced variance for the post-filter estimation. This results into less musical tones, however, the more instationary speech signal is then also strongly affected during active speech frames, which results into perceivable speech distortion and a reverberation-like effect. On the contrary, applying a smaller fixed smoothing factor will lead to less speech distortion dur-
5.2.2 Performance Analysis

ing active speech frames, however, it will then generate much more musical tones especially during speech pause frames. Therefore, an adaptive smoothing factor computation based on different SNR conditions, which can differentiate between speech pause and active speech frames, is much more preferred. According to the analysis of Guerin et al. (2003), the smoothing factor can be adaptively estimated as

\[ \alpha(\ell,k) = \alpha_1 - \alpha_2 \cdot \frac{\text{SNR}(\ell,k)}{1 + \text{SNR}(\ell,k)}, \]  

(5.2.1)

with \( \text{SNR}(\ell,k) \) being the signal-to-noise ratio at the beamformer output. For a low \( \text{SNR}(\ell,k) \) indicating a speech pause frame, \( \alpha(\ell,k) \) will reach its upper limit \( \alpha_1 \), leading to a smooth estimation of the auto- and cross-PSD of input microphone signals. This limits then the occurrence of musical tones (Guerin et al., 2003). When \( \text{SNR}(\ell,k) \) is high indicating an active speech frame, \( \alpha(\ell,k) \) will reach its minimum \( \alpha_1 - \alpha_2 \), leading to a good estimation following the fast speech variation.

However, \( \text{SNR}(\ell,k) \) still has to be estimated. Comparing (5.2.1) with the general post-filter estimation \( H_{PF}(\ell,k) \) in (2.4.1), it can be found that both terms can be treated as a Wiener filter. Since \( \text{SNR}(\ell,k) \) needs to be firstly estimated for calculating \( H_{PF}(\ell,k) \), and since \( \text{SNR}(\ell,k) \) can be assumed to be almost unchanged from frame to frame, the term \( \text{SNR}(\ell,k) \) in (5.2.1) can thus be approximated as (Fertl, 2005)

\[ \frac{\text{SNR}(\ell,k)}{1 + \text{SNR}(\ell,k)} \approx H_{PF}(\ell - 1, k). \]  

(5.2.2)

Inserting (5.2.2) into (5.2.1), the adaptive smoothing factor for the PSD estimation within the post-filter estimation can be realized as

\[ \alpha(\ell,k) = \alpha_1 - \alpha_2 \cdot H_{PF}(\ell - 1, k), \]  

(5.2.3)

In this thesis, all post-filters are limited to the interval \([0.15, 1]\), which lead \( \alpha(\ell,k) \) in (5.2.3) to be limited to \([\alpha_1 - \alpha_2, \alpha_1 - \alpha_2 \cdot 0.15]\). Care has to be taken in choosing the parameters \((\alpha_1, \alpha_2)\) to achieve the best result concerning the speech component quality and the amount of musical tones. In this work, the optimal \((\alpha_1, \alpha_2)\) has been obtained only heuristically as \((0.8, 0.5)\), see (Yu and Fingscheidt, 2010c,b).

5.2.2 Performance Analysis

As analyzed in Section 3.5, the FoM-based instrumental optimization is capable of automatically finding the optimal parameterization of a noise reduction algorithm. Therefore, in this section, the FoM using the weighting factors and exponents shown in Table 5.2 is applied to identify the unknown optimal pair \((\alpha_{1,\text{opt}}, \alpha_{2,\text{opt}})\) for the post-filter estimation within each driving condition. In the \((\alpha_1, \alpha_2)\) optimization procedure, the term \( \theta_p \) within the FoM calculation in (3.4.5) indicates different \((\alpha_1, \alpha_2)\) combinations for searching for the optimal \((\alpha_{1,\text{opt}}, \alpha_{2,\text{opt}})\). The optimization is performed with the (DS-MC/A, dif) system (Yu and Fingscheidt, 2010c,b). The (DS-MC/A, dif) system utilizes the delay-and-sum beamformer (DS), McCowan’s post-filter (MC) by utilizing the adaptive (A)
5. Post-Filter Design with Modified PSD Estimation and Hybrid Noise Coherence Functions

Figure 5.2: Averaged instrumental measurement results of all driving conditions by using the (DS-MC/A, dif) system (Yu and Fingscheidt, 2010c,b) with different combinations of $(\alpha_1, \alpha_2)$ smoothing factor $\alpha(\ell,k)$ (5.2.3), and the diffuse noise coherence function (dif). Please note that in addition to the adaptive smoothing factor, the second difference to the baseline (CS-MC/A, dif) system is that, instead of the constrained superdirective beamformer, the delay-and-sum beamformer is applied, since the delay-and-sum beamformer behaves much more robust in practice with low-cost microphones (Bitzer and Simmer, 2001) and is easier to be implemented. Setting $\alpha_1 \in \{0.7, 0.8, 0.9, 1.0\}$ and $\alpha_2 \in \{0.2, 0.3, 0.4, 0.5\}$, altogether $4 \times 4 = 16$ combinations of $(\alpha_1, \alpha_2)$ are simulated using the (DS-MC/A, dif) system to search for $(\alpha_{\text{opt},1}, \alpha_{\text{opt},2})$. The reference set $\{\text{MOS}_{LQO,\text{ref}}, \Delta\text{SNR}_{\text{ref}}, \Delta\Psi_{\text{log,ref}}\}$ is calculated by the (DS-MC/A_{\text{ref}}, dif) system using the heuristically found optimal value $(\alpha_{\text{opt},1}, \alpha_{\text{opt},2}) = (0.8, 0.5)$ ($A_{\text{ref}}$) as proposed in (Yu and Fingscheidt, 2010c,b), leading to $\text{FoM} = 1$. It turns out that $(\alpha_{\text{opt},1}, \alpha_{\text{opt},2}) = (1.0, 0.3)$ provides the maximum $\text{FoM}$ for all driving conditions. The averaged instrumental measurement results along with $\text{FoM}$ of all driving conditions for different combinations of $(\alpha_1, \alpha_2)$ are shown in Fig. 5.2. It can be observed that the maximum $\text{FoM} = 1.32$ is achieved by applying $(\alpha_{\text{opt},1}, \alpha_{\text{opt},2}) = (1.0, 0.3)$. As illustrated in Fig. 5.2, the differences of $\Delta\text{SNR}$ values for different $(\alpha_1, \alpha_2)$ combinations are quite low, since the adaptive smoothing factor is actually utilized to improve the speech component quality and reduce musical tones at the same time. Independently of $\alpha_2$, a larger $\alpha_1$ is much more preferred by showing lower $\Delta\Psi_{\text{log}}$ values as well as higher MOS$_{LQO}$ scores than a smaller $\alpha_1$. In the case of $\alpha_1 = 1.0$, there is a trade-off between
ΔSNR, MOSLQO, and ΔΨw by applying different values of α2. Actually, the optimal found combination (α1,opt, α2,opt) = (1.0, 0.3) showing the maximum FoM is thus also reasonable by providing the best trade-off between ΔSNR, MOSLQO, and ΔΨw against all other (α1, α2) combinations. These results can also be confirmed by informal listening tests.

5.3 Hybrid Noise Coherence Functions

In this section, a novel hybrid noise coherence function combining the diffuse and the measured noise coherence functions for a specific driving condition, is derived and analyzed (Yu and Fingscheidt, 2012a). Furthermore, the optimization of the hybrid noise coherence function for the post-filter estimation, which is performed for different driving conditions using the same FoM calculation as in Section 5.2.2, is also presented.

5.3.1 Noise Coherence Function Measurement

As mentioned in Section 2.4, the classification of different post-filter estimations can be achieved by using the noise coherence function. Therefore, the noise coherence function is of enormous interest to the post-filter estimation. While the diffuse noise coherence function is treated as a good approximation model for car noise (Meyer and Simmer, 1997), it is still worth to validate the assumption of the diffuse noise coherence function for different driving conditions, e.g., different driving speeds, different air-conditioning operation levels, or windows being open or closed. Motivated by the potential benefit of accessing the real-time driving condition through CAN bus data, it is worth to measure the noise coherence function for different driving conditions.

Welch-Periodogram Method

In order to measure the noise coherence function as defined in (2.1.10), the auto- and cross-power spectral densities of noise signals from the microphone pair i and j have to be measured first. In this thesis, all derived post-filters are frame-wise operated in the frequency domain. Therefore, after applying the analysis window wa(n) and DFT, the auto- and cross-periodograms can be estimated by the modified periodograms proposed by Welch (1967) as

\[ I_{N_iN_i}(\ell,k) = \frac{1}{N_a \cdot U} N_i(\ell,k)N_i^*(\ell,k) , \]

\[ I_{N_iN_j}(\ell,k) = \frac{1}{N_a \cdot U} N_i(\ell,k)N_j^*(\ell,k) , \]

where \( N_a \) indicates the length of \( w_a(n) \) and the factor \( N_a \cdot U \) normalizes the modified periodograms to the energy of the analysis windows \( w_a(n) \) with

\[ U = \frac{1}{N_a \sum_{n=0}^{N_a-1} w_a^2(n)} . \]
Since each noise signal of the multichannel in-car noise database has the length of 8 s resulting in a set of \( \mathcal{L} \) frames, the auto- and cross-power spectral densities can thus be measured by averaging the corresponding auto- and cross-periodograms through all frames in \( \mathcal{L} \) as (Martin, 1995)

\[
\tilde{\Phi}_{N_i,N_i}(k) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} I_{N_i,N_i}(\ell,k),
\]

(5.3.4)

\[
\tilde{\Phi}_{N_j,N_j}(k) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} I_{N_j,N_j}(\ell,k),
\]

(5.3.5)

\[
\tilde{\Phi}_{N_i,N_j}(k) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} I_{N_i,N_j}(\ell,k),
\]

(5.3.6)

with \(|\mathcal{L}|\) being the total number of frames in set \( \mathcal{L} \). Inserting \( \tilde{\Phi}_{N_i,N_i}(k), \tilde{\Phi}_{N_j,N_j}(k), \) and \( \tilde{\Phi}_{N_i,N_j}(k) \) into (2.1.10), the noise coherence function of the noise signal pair with a length of 8 s can then be computed as

\[
\tilde{\Gamma}^{8s}_{i,j}(k) = \frac{\tilde{\Phi}_{N_j,N_j}(k)}{\sqrt{\tilde{\Phi}_{N_i,N_i}(k)\tilde{\Phi}_{N_j,N_j}(k)}} \in \mathbb{C}.
\]

(5.3.7)

The applied parameters for the modified periodogram estimation are listed in Table 5.3. Please note that the Hann window and the frame shift \( \Delta L = 50\% \times L \) are chosen as suggested by Martin (1995) to achieve a good compromise between the frequency resolution, the variance of the noise coherence function measurement, and the computation efficiency for a fixed signal length. For each driving condition of the multichannel in-car noise database, more than 64 noise signals with a length of 8 s have been recorded. Therefore, in order to further improve the noise coherence function measurement with reduced variance, more than one noise signal with the length of 8 s can be employed to measure the noise coherence function as

\[
\tilde{\Gamma}^{Q}_{i,j}(k) = \frac{1}{Q} \sum_{q=1}^{Q} \tilde{\Gamma}^{8s,q}_{i,j}(k),
\]

(5.3.8)

where \( Q \) is the applied number of noise signals.

### Statistical Stability Test

In the previous section, the measurement of the noise coherence function \( \tilde{\Gamma}^{Q}_{i,j}(k) \) is expressed in (5.3.8) by applying \( Q \) noise signals. However, it is unclear how many noise...
5.3.1. Noise Coherence Function Measurement

**Figure 5.3:** Square error \( E_{1,2}^{(Q,0.5Q)}(k) \) by applying different \( Q \) values to the \( \text{Re}\{ \tilde{\Gamma}_{1,j}^Q(k) \} \) measurement in (5.3.9) for driving condition 8; frequency \( f \in [0, 8000] \text{ Hz} \) is equivalent to \( k \in [0, 256] \) with a DFT length of 512 and \( f_s = 16 \text{ kHz} \).

signals shall be applied to achieve a correct (i.e., stable) noise coherence function measurement. In order to obtain the suitable number of noise signals, a similar statistical stability test as in (He et al., 2009) by doubling the sample size is employed. The sample number can be treated as suitable to measure the coherence function, as soon as no significant (change in the) square error can be identified (He et al., 2009). Therefore, different \( Q \in \{1, 2, 4, 8, 16, 32, 64\} \) values are used in (5.3.8) to compute \( \tilde{\Gamma}_{1,j}^Q(k) \). Based on the noise coherence function measurement applying different \( Q \) values, the square error between each adjacent \( Q \) values can be calculated as

\[
E_{i,j}^{(Q,0.5Q)}(k) = \left( \text{Re}\{ \tilde{\Gamma}_{1,j}^Q(k) \} - \text{Re}\{ \tilde{\Gamma}_{1,j}^{0.5Q}(k) \} \right)^2 .
\]

(5.3.9)

Fig. 5.3 depicts \( E_{1,2}^{(Q,0.5Q)}(k) \) for driving condition 8 (see Table 4.1). It can be seen that using small \( Q \) values up to \( Q = 16 \), a considerable difference can be observed between \( Q \) and \( 0.5Q \), especially for \( k \geq 128 \) (i.e., \( f \geq 4000 \text{ Hz} \)). However, applying \( Q \geq 32 \), no more significant difference can be observed between \( Q \) and \( 0.5Q \), (see \( E_{1,2}^{(32,16)}(k) \) and \( E_{1,2}^{(64,32)}(k) \) in Fig. 5.3). In addition, the sum of the square error \( E_{i,j}^{(Q,0.5Q)}(k) \) (5.3.9) through
5. Post-Filter Design with Modified PSD Estimation and Hybrid Noise Coherence Functions

Figure 5.4: $E_{i,j}^{(Q,0.5Q)}$ values for all noise coherence function measurements in (5.3.9) by applying different $(Q,0.5Q)$ values for driving condition 8

all frequency bins can be computed as

$$E_{i,j}^{(Q,0.5Q)} = \sum_{k=0}^{K-1} E_{i,j}^{(Q,0.5Q)}(k)$$

(5.3.10)

The results of $E_{i,j}^{(Q,0.5Q)}$ for all noise coherence function measurements for driving condition 8 are illustrated in Fig. 5.4. It can be seen that for different noise coherence function measurements, when $Q \geq 32$ is employed, the value of $E_{i,j}^{(Q,0.5Q)}$ is kept to a stable small value within the pair $(Q,0.5Q)$ meaning no significant difference between the noise coherence function measurement employing $Q$ and $0.5Q$. The same results can also be observed for other driving conditions. Therefore, it is sufficient for the noise coherence function measurement by using a $Q$ value satisfying $Q \geq 32$ in (5.3.8). In this thesis, the value $Q = 32$ is applied to provide correct and stable noise coherence function measurements for different driving conditions.

5.3.2 Approach

Applying $Q = 32$ in (5.3.8), we are now able to measure the noise coherence function of the applied head unit-integrated vertical microphone array for different driving conditions. Therefore, a direct comparison between the measured noise coherence function $\tilde{\Gamma}_{i,j}(k)$ and the diffuse noise coherence function $\Gamma_{i,j}^{\text{diff}}(k)$ (2.1.18) can be carried out. The
5.3.2. Approach

Figure 5.5: The measured noise coherence functions $\text{Re}\left\{\tilde{\Gamma}_{1,2}(k)\right\}$ for driving conditions 4 and 9, and the diffuse noise coherence function $\text{Re}\left\{\Gamma_{1,2}^{\text{dif}}(k)\right\}$, respectively.

results of the measured noise coherence function $\text{Re}\left\{\tilde{\Gamma}_{1,2}(k)\right\}$ and the diffuse noise coherence function $\text{Re}\left\{\Gamma_{1,2}^{\text{dif}}(k)\right\}$ between microphones 1 and 2 for driving conditions 4 and 9 are depicted in Fig. 5.5. It can be observed that for driving condition 9, the measured noise coherence function matches the diffuse noise coherence function quite well up to about 1000 Hz. However, for driving condition 4, the measured noise coherence function matches the diffuse noise coherence function only in the low frequency region until about 250 Hz, while there is a considerable mismatch between these two noise coherence functions for frequencies beyond 250 Hz. Investigating other driving conditions, a considerable mismatch between the measured and the diffuse noise coherence functions within different frequency regions can be observed. Accordingly, we propose a hybrid noise coherence function by mixing the diffuse and the measured noise coherence functions. The hybrid noise coherence function is then defined as (Yu and Fingscheidt, 2012a)

$$\Gamma_{i,j}^{\text{hyb}}(k) = (1 - \delta(k)) \cdot \Gamma_{i,j}^{\text{dif}}(k) + \delta(k) \cdot \tilde{\Gamma}_{i,j}(k), \quad (5.3.11)$$

where $\delta(k) \in [0, 1]$ is a mixing factor defined as

$$\delta(k) = \begin{cases} \frac{k}{k_{\text{up}}} & , \text{if } 0 \leq k \leq k_{\text{up}} \ , \\ 1 & , \text{if } k > k_{\text{up}} \ . \end{cases} \quad (5.3.12)$$

Since for different driving conditions the mismatch between the measured and the diffuse noise coherence functions appears only beyond a certain frequency $k_{\text{up}}$, the mixing factor $\delta(k)$ is thus set to be linearly increased from $\delta(0) = 0$ to $\delta(k_{\text{up}}) = 1$ within $0 \leq k \leq k_{\text{up}}$. In this range, a linear combination of $\Gamma_{i,j}^{\text{dif}}(k)$ and $\tilde{\Gamma}_{i,j}(k)$ is applied, since both noise coherence
functions can be treated as nearly identical. For \( k > k_{up} \), only the measured \( \tilde{\Gamma}_{i,j}(k) \) is applied, showing a better match to the real noise coherence function of the current driving condition. However, as we can see in Fig. 5.5, the difference between the measured and the diffuse noise coherence functions is dependent on driving conditions. Therefore, the value of \( k_{up} \) has to be optimized individually for each driving condition.

### 5.3.3 Performance Analysis

In order to identify the optimal \( k_{up, opt} \), the same FoM-based instrumental optimization using the identical weighting factors and exponents as given in Table 5.2 for the FoM computation is applied. In this case, the term \( \theta_p \) within the FoM calculation in (3.4.5) indicates different \( k_{up} \) values within

\[
k_{up} \in \{4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 88, 96, 104, 112, 120, 128\},
\]

which are equivalent to

\[
f \in \{125, 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500, 3750, 4000\} \text{ Hz},
\]

with \( f_s = 16 \text{ kHz} \) and a DFT length of 512. The search of \( k_{up} \) is performed only up to 4000 Hz, since a considerable difference between the measured and the diffuse coherence functions for different driving conditions can only be detected up to 4000 Hz. Using different values of \( k_{up} \) as given in (5.3.13), the corresponding hybrid noise coherence function \( \Gamma_{i,j}^{hyb}(k) \) expressed in (5.3.11) can be generated. The new beamformer and post-filter system is denoted as the (DS-MC/\( A_{opt, hyb} \)) system. It uses the delay-and-sum beamformer (DS), McCowan’s post-filter (MC) applying the hybrid noise coherence function (hyb), and \((\alpha_{1,opt}, \alpha_{2,opt}) = (1.0, 0.3) \) (\( A_{opt} \)) as optimized in Section 5.2.2. Furthermore, instead of the hybrid noise coherence function, the new system employing the diffuse noise coherence function \( \Gamma_{i,j}^{dif}(k) \) only and the measured noise coherence function \( \tilde{\Gamma}_{i,j}(k) \) only are also simulated. The same reference set \{MOS^s_{LQO,ref}, \Delta \text{SNR}_{ref}, \Delta \Psi^{w, log}_{ref}\} calculated by the (DS-MC/\( A_{opt, dif} \)) system by using the diffuse noise coherence function and \((\alpha_1, \alpha_2) = (0.8, 0.5) \) as employed in Section 5.2.2 for optimizing the adaptive smoothing factor is also used here. Therefore, the (DS-MC/\( A_{ref, dif} \)) system results into \( \text{FoM} = 1 \). The FoM-based optimization is individually performed for each driving condition. The \( k_{up, opt} \) value resulting in the maximum FoM for each driving condition are shown in Table 5.4. It is obvious that the diffuse noise coherence function is not always the ideal model for different driving conditions, only driving condition 3 can be identified with the diffuse noise coherence function being the optimal one. It can be observed that at higher speeds a higher value of \( k_{up, opt} \) is generally obtained. Furthermore, we found that in the driving conditions with 120 km/h, \( k_{up, opt} \) will increase with the window being open compared to the window being closed. In both cases, it means that once \( k_{up, opt} \) increases, the noise field is getting more diffuse. The obtained \( k_{up, opt} \) values also match the results in Fig. 5.5 nicely, where the measured noise coherence function is compared to the diffuse noise coherence function for driving conditions 4 and 9, respectively. As can be seen in Fig. 5.5, there is
### Table 5.4: Results of $k_{\text{up, opt}}$ for different driving conditions, if the hybrid noise coherence function is optimal (otherwise the diffuse noise coherence function is optimal)

<table>
<thead>
<tr>
<th>Cond.</th>
<th>Speed [km/h]</th>
<th>Engine</th>
<th>AC</th>
<th>Window</th>
<th>$k_{\text{up, opt}}$ [bin] ([Hz])</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>off</td>
<td>0/7</td>
<td>closed</td>
<td>8 (250)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>off</td>
<td>3/7</td>
<td>closed</td>
<td>24 (750)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>off</td>
<td>7/7</td>
<td>closed</td>
<td>——</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>on</td>
<td>0/7</td>
<td>closed</td>
<td>8 (250)</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>on</td>
<td>3/7</td>
<td>closed</td>
<td>16 (500)</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>on</td>
<td>7/7</td>
<td>closed</td>
<td>32 (1000)</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>on</td>
<td>0/7</td>
<td>closed</td>
<td>24 (750)</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>on</td>
<td>0/7</td>
<td>open</td>
<td>32 (1000)</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>on</td>
<td>3/7</td>
<td>closed</td>
<td>32 (1000)</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>on</td>
<td>3/7</td>
<td>open</td>
<td>48 (1500)</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>on</td>
<td>7/7</td>
<td>closed</td>
<td>32 (1000)</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>on</td>
<td>7/7</td>
<td>open</td>
<td>16 (500)</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>on</td>
<td>0/7</td>
<td>closed</td>
<td>24 (750)</td>
</tr>
<tr>
<td>14</td>
<td>120</td>
<td>on</td>
<td>0/7</td>
<td>open</td>
<td>64 (2000)</td>
</tr>
<tr>
<td>15</td>
<td>120</td>
<td>on</td>
<td>3/7</td>
<td>closed</td>
<td>24 (750)</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>on</td>
<td>3/7</td>
<td>open</td>
<td>64 (2000)</td>
</tr>
<tr>
<td>17</td>
<td>120</td>
<td>on</td>
<td>7/7</td>
<td>closed</td>
<td>48 (1500)</td>
</tr>
<tr>
<td>18</td>
<td>120</td>
<td>on</td>
<td>7/7</td>
<td>open</td>
<td>48 (1500)</td>
</tr>
</tbody>
</table>
5. Post-Filter Design with Modified PSD Estimation and Hybrid Noise Coherence Functions

![Diagram](image)

**Figure 5.6:** Averaged instrumental measurement results of the (DS-MC/A\textsubscript{opt}, hyb\textsubscript{opt}) system using the individually optimized hybrid noise coherence function (\(k_{\text{up, opt}}\) in Table 5.4) and \((\alpha_{1, \text{opt}}, \alpha_{2, \text{opt}}) = (1.0, 0.3)\) for all driving conditions; in comparison to the (DS-MC/A\textsubscript{opt}, dif) system applying \((\alpha_{1, \text{opt}}, \alpha_{2, \text{opt}}) = (1.0, 0.3)\), the (DS-MC/A\textsubscript{ref}, dif) system using \((\alpha_{1}, \alpha_{2}) = (0.8, 0.5)\) (Yu and Fingscheidt, 2010c,b), and the baseline (CS-MC/F, dif) system (McCowan and Bourlard, 2003). \(\Gamma_{ij}(k)\) is applied to the last three approaches.

Fig. 5.6 depicts the average instrumental measurement results for all driving conditions by the new (DS-MC/A\textsubscript{opt}, hyb\textsubscript{opt}) system. The new system employs the individually optimized hybrid noise coherence function (hyb\textsubscript{opt}) for each driving condition with \(k_{\text{up, opt}}\) as given in Table 5.4 and \((\alpha_{1, \text{opt}}, \alpha_{2, \text{opt}}) = (1.0, 0.3)\) as optimized in Section 5.2.2. In addition, the (DS-MC/A\textsubscript{opt}, dif) system using \((\alpha_{1, \text{opt}}, \alpha_{2, \text{opt}}) = (1.0, 0.3)\), the (DS-MC/A\textsubscript{ref}, dif) system using the heuristically optimized \((\alpha_{1}, \alpha_{2}) = (0.8, 0.5)\) (Yu and Fingscheidt, 2010c,b), and the baseline McCowan’s approach (CS-MC/F, dif) (McCowan and Bourlard, 2003) are also shown in Fig. 5.6. Please note that the diffuse noise coherence function \(\Gamma_{ij}^{\text{dif}}(k)\) is applied in the last three approaches. By using the individually optimized hybrid noise coherence functions and the optimal adaptive smoothing factor, the (DS-MC/A\textsubscript{opt}, hyb\textsubscript{opt}) system achieves the maximum FoM result of 1.43. Under the assumption of the dif-
5.3.3. Performance Analysis

![Graphs showing performance analysis results](image)

Figure 5.7: Averaged instrumental measurement results of the (DS-LE/A\_opt, hyb\_opt) system using the individually optimized hybrid noise coherence function \((k_{up,\text{opt}})\) in Table 5.4 and \((\alpha_{1,\text{opt}}, \alpha_{2,\text{opt}}) = (1.0, 0.3)\) for all driving conditions; in comparison to the (DS-LE/A\_opt, dif) system applying \((\alpha_{1,\text{opt}}, \alpha_{2,\text{opt}}) = (1.0, 0.3)\), the (DS-LE/A\_ref, dif) system employing \((\alpha_1, \alpha_2) = (0.8, 0.5)\) (Yu and Fingscheidt, 2010c,b), and the baseline (CS-LE/F, dif) system (Lefkimmiatis and Maragos, 2006). \(\Gamma_{i,j}^{\text{dif}}(k)\) is applied to the last three approaches.

Fuse noise coherence function \(\Gamma_{i,j}^{\text{dif}}(k)\), the (DS-MC/A\_opt, dif) system employing the optimized adaptive smoothing factor improves the FoM from 1.0 to 1.32 compared to the (DS-MC/A\_ref, dif) system employing only the heuristically optimized adaptive smoothing factor. However, using the heuristically optimized adaptive smoothing factor, the (DS-MC/A\_ref, dif) system still yields a slightly better FoM than the baseline (CS-MC/F, dif) system (1.0 vs. 0.99). Comparing the three instrumental measurement results separately, it can also be observed that the (DS-MC/A\_opt, hyb\_opt) system achieves the best results for \(\Delta\text{SNR}\) and MOS\(^L_{\text{LQO}}\), while still maintaining nearly the same lowest \(\Delta\Psi_{\log}^a\) values with the (DS-MC/A\_opt, dif) system. Therefore, in comparison to the baseline (CS-MC/F, dif) system, the (DS-MC/A\_opt, hyb\_opt) system obtained by the FoM-based instrumental optimization improves the overall noise reduction performance significantly. These results can also be confirmed by informal listening tests.

In order to validate the generalization of the optimized hybrid noise coherence functions and the optimized adaptive smoothing factor to other beamforming and post-filtering
systems, the baseline (CS-LE/F, dif) system using Lefkimmiatis’s post-filter is also evaluated. Therefore, a new (DS-LE/A_{opt}, hyb_{opt}) system is simulated by employing the delay-and-sum beamformer and Lefkimmiatis’s post-filter utilizing the optimized hybrid noise coherence function with $k_{up,opt}$ being given in Table 5.4 and the optimized adaptive smoothing factor using $(\alpha_{1,opt}, \alpha_{2,opt}) = (1.0, 0.3)$. Furthermore, the (DS-LE/A_{opt}, dif) system using only $(\alpha_{1,opt}, \alpha_{2,opt}) = (1.0, 0.3)$, the (DS-LE/A_{ref}, dif) system using the heuristically found $(\alpha_{1,opt}, \alpha_{2,opt}) = (0.8, 0.5)$ (Yu and Fingscheidt, 2010c,b), and the baseline Lefkimmiatis’s approach (CS-LE/F, dif) system (Lefkimmiatis and Maragos, 2006) are also simulated. Please note that the diffuse noise coherence function $\Gamma_{dif,i,j}^{\delta}(k)$ is applied to the last three approaches. The reference set \{MOS_{LQO,ref}, \Delta SNR_{ref}, \Delta \Psi_{log,ref}\} is being calculated with the (DS-LE/A_{ref}, dif) system, leading to $\text{FoM} = 1$. Fig. 5.7 shows the average instrumental measurement results of the four approaches for all driving conditions. As in Fig. 5.6, applying the optimized hybrid noise coherence function and the optimized adaptive smoothing factor, the (DS-LE/A_{opt}, hyb_{opt}) system also achieves the maximum $\text{FoM}$ result of about 1.3. This result is also reasonable, since the (DS-LE/A_{opt}, hyb_{opt}) system also yields the best $\Delta SNR$ and MOS_{LQO} results, while maintaining the second lowest $\Delta \Psi_{log}$ result. Furthermore, under the same diffuse noise coherence function, the (DS-LE/A_{opt}, dif) system utilizing $(\alpha_{1,opt}, \alpha_{2,opt}) = (1.0, 0.3)$ achieves also the second largest $\text{FoM}$ result of 1.19 compared to the (DS-LE/A_{ref}, dif) system using the heuristically optimized $(\alpha_1, \alpha_2) = (0.8, 0.5)$. Still, the (DS-LE/A_{ref}, dif)-system yields a better $\text{FoM}$ result of 1 than the baseline (CS-LE/F, dif) system. All results can also be confirmed by informal listening tests.

Therefore, the hybrid noise coherence function and the adaptive smoothing factor, which are optimized with McCowan’s post-filter and validated by Lefkimmiatis’s post-filter, can be treated as two generalized optimal parameters for the general post-filter designs based on the noise coherence function.

### 5.4 Summary

In this chapter, a novel post-filter design utilizing an adaptive smoothing factor for the PSD estimation as well as a hybrid noise coherence function has been presented. In order to solve the compromise between speech distortion and noise distortion in term of musical tones, an adaptive smoothing factor has been proposed to the PSD estimation. Having the multichannel speech, noise and CAN bus database for different driving conditions at our disposal, the adaptive smoothing factor has been individually optimized for each driving condition by applying the $\text{FoM}$-based instrumental optimization method.

Based on the Welch-periodogram method, all possible noise coherence functions have been measured for each driving condition. Statistical stability investigations show that 32 noise signals with a length of 8 s are sufficient to provide a robust and reliable noise coherence function measurement for each driving condition. Since a considerable mismatch between the diffuse and the measured noise coherence functions can be observed for different driving conditions, a hybrid noise coherence function has been proposed. The hybrid
noise coherence function is a mixture of the diffuse and the measured noise coherence function for a specific driving condition. In a similar manner, the hybrid noise coherence function has been individually optimized for different driving conditions by employing the FoM-based instrumental optimization method. In a car application, the optimal hybrid noise coherence function can then be selected by the CAN bus data for the corresponding driving condition in a real-time manner. Applying the optimal adaptive smoothing factor and the optimal hybrid noise coherence function to the new post-filter design, significant improvement of the overall noise reduction performances has been achieved.
5. Post-Filter Design with Modified PSD Estimation and Hybrid Noise Coherence Functions
Chapter 6

Post-Filter Design with Multichannel Decision-Directed Approach to \textit{A Priori} SNR Estimation

In noise reduction algorithms, the \textit{a priori} SNR estimation plays a very important role. In the field of single-channel noise reduction, the decision-directed (DD) approach proposed by Ephraim and Malah (1984) is commonly applied to estimate the \textit{a priori} SNR. The \textit{temporal} smoothing of the single-channel DD approach results into a good suppression of musical tones, however, speech distortion unfortunately occurs particularly in speech onsets. On the other hand, applying an MVDR beamformer with a post-filter being estimated by using \textit{spatial} smoothing, good preservation of the speech component can be provided, however, along with disturbing musical tones. In this chapter, a new multichannel DD \textit{a priori} SNR estimation based on both temporal and spatial smoothing is presented, which can be applied to a new post-filter design. In Section 6.1, the single-channel baseline \textit{a priori} SNR estimation using temporal smoothing is introduced. The multichannel baseline \textit{a priori} SNR estimation based on spatial smoothing is then described in Section 6.2.1. Based on the temporal and spatial smoothing, the new multichannel DD \textit{a priori} SNR estimation is then addressed in Section 6.2.2. Applying the multichannel DD \textit{a priori} SNR estimate to formulate a new post-filter design, the weighting factors of the multichannel DD \textit{a priori} SNR estimation are optimized by using the FoM-based instrumental optimization as applied in Chapter 5.

6.1 Single-Channel Baseline Approach by Temporal Smoothing

For the beamforming and post-filtering system discussed in this thesis, the single-channel Wiener post-filter taking the MVDR beamformer output $\hat{S}_{BF}(\ell, k)$ as the input signal (see
6. Post-Filter Design with Multichannel Decision-Directed Approach to A Priori SNR Estimation

Fig. 2.12) can also be realized by utilizing the a priori SNR as (Scalart and Filho, 1996)

\[ H_{WF} = \frac{\xi_{PF}(\ell, k)}{\xi_{PF}(\ell, k) + 1}, \]

(6.1.1)

where the term \( \xi_{PF}(\ell, k) \) denotes the a priori SNR for the post-filter estimation, which can be expressed as

\[ \xi_{PF}(\ell, k) = \frac{E\{|S(\ell, k)|^2\}}{E\{|N_{BF}(\ell, k)|^2\}}, \]

(6.1.2)

where \( S(\ell, k) \) is the clean speech signal and \( N_{BF}(\ell, k) \) is the noise signal after the MVDR beamformer, i.e., the input noise signal for the Wiener post-filter, respectively. Since the MVDR beamformer is applied, the clean speech signal PSD remains unchanged after the MVDR beamformer (cf. (2.2.16)). The estimation of \( \xi_{PF}(\ell, k) \) can be performed by the single-channel decision-directed approach proposed by Ephraim and Malah (1984) as

\[ \xi_{PF}^{DD'}(\ell, k) = \beta_1 \cdot \frac{|\hat{S}(\ell, k)|^2}{\Phi_{N_{BF}N_{BF}}(\ell, k)} + \beta_2 \cdot P[\hat{\gamma}_{BF}(\ell, k) - 1], \]

\[ \xi_{PF}^{DD}(\ell, k) = \max\{\xi_{PF}^{DD'}(\ell, k), \xi_{min}\}, \]

(6.1.3)

with \( \beta_1 = 1 - \beta_2 \), \( \hat{S}(\ell, k) \) denoting the enhanced speech signal of the post-filter in the previous frame, \( \Phi_{N_{BF}N_{BF}}(\ell, k) \) being the estimated noise PSD computed via minimum statistics (Martin, 2001) in the previous frame, and \( \hat{\gamma}_{BF}(\ell, k) \) being the estimated a posteriori SNR, respectively. According to Ephraim and Malah (1984), the a priori SNR can be formulated as

\[ \xi_{PF}(\ell, k) = \frac{E\{|S(\ell, k)|^2\}}{E\{|N_{BF}(\ell, k)|^2\}} = E\left\{ \frac{|S(\ell, k)|^2}{E\{|N_{BF}(\ell, k)|^2\}} \right\} \approx \frac{|\hat{S}(\ell, k)|^2}{\Phi_{N_{BF}N_{BF}}(\ell, k)}, \]

(6.1.4)

Furthermore, based on the assumption of statistical independence between the clean speech signal and the noise signal, we obtain the relation between the a priori SNR and the a posteriori SNR as (Ephraim and Malah, 1984)

\[ \xi_{PF}(\ell, k) = E\{\gamma_{BF}(\ell, k)\} - 1 = \frac{E\{\hat{S}_{BF}(\ell, k)^2\}}{E\{|N_{BF}(\ell, k)|^2\}} - 1 \approx \frac{\hat{S}_{BF}(\ell, k)^2}{\Phi_{N_{BF}N_{BF}}(\ell, k)} - 1, \]

(6.1.5)

Comparing (6.1.3) to (6.1.4) and (6.1.5), it can be seen that the single-channel decision-directed approach combines two a priori SNR estimates in a linear fashion: The first one is the a priori SNR estimate based on the speech and noise PSD estimations of the previous frame (6.1.4). The second one is based on the a posteriori SNR estimate in the current frame (6.1.5). Applying the weighting factors \( \beta_1 \) and \( \beta_2 \), a recursive smoothing over time for the a priori SNR estimation is achieved by the decision-directed approach in (6.1.3). In Section 3.5, the instrumental measurements have been shown for different weighting rules with \( \beta_1 \) being chosen close to one. It can be seen that applying a large \( \beta_1 \) value, the musical tones can be significantly reduced compared to a small \( \beta_1 \) value (see Fig. 3.10), however, the speech component is unfortunately much more distorted (see Fig. 3.8). Actually, applying a large \( \beta_1 \) close to one, the single-channel decision-directed approach indeed
yields a strong temporally smoothed \textit{a priori} SNR estimation with reduced variance, which helps to significantly reduce musical tones (Cappé, 1994). Unfortunately, the strong temporal smoothing also reduces the capability of the single-channel decision-directed approach to follow especially speech onsets, which leads to speech distortion in these cases.

6.2 Multichannel Approaches

In order to deal with the drawback of the single-channel decision-directed \textit{a priori} SNR estimation in (6.1.3), a new multichannel decision-directed \textit{a priori} SNR estimation is proposed (Yu and Fingscheidt, 2010a, 2011b), which is based on both temporal and spatial smoothing. Before formulating the multichannel decision-directed \textit{a priori} SNR estimator in Section 6.2.2, the \textit{a priori} SNR estimation utilizing only spatial smoothing is introduced in Section 6.2.1

6.2.1 Baseline Approach by Spatial Smoothing

In Section 2.4.2, McCowan’s post-filter and Lefkimmiatis’s post-filter employing the diffuse noise coherence function have been derived. Both of these two post-filters are also single-channel Wiener post-filters. The only difference between the two post-filters is that McCowan’s post-filter is based on the theoretically sub-optimal post-filter estimation (cf. (2.4.7)) using the microphone noise signal PSD, while Lefkimmiatis’s post-filter is based on the theoretically optimal post-filter estimation (cf. (2.4.1)) using the beamformer output noise PSD.

In the estimation of McCowan’s post-filter, comparing the theoretically sub-optimal Wiener post-filter estimation (2.4.7) to the \textit{a priori} SNR-driven Wiener post-filter estimation (6.1.1), we can obtain

\[ \xi_{\text{PF}}^{MC}(\ell,k) = \frac{\Phi_{SS}(\ell,k)}{\Phi_{NN}(\ell,k)}, \]  

where \( \Phi_{SS}(\ell,k) \) and \( \Phi_{NN}(\ell,k) \) can be estimated by spatially smoothing the estimated clean speech PSD \( \hat{\Phi}_{SS}^{(i,j)}(\ell,k) \) (2.4.16) and the noise PSD \( \hat{\Phi}_{NN}^{(i,j)}(\ell,k) \) (2.4.19) over all possible \((i,j)\) microphone signal combinations. Therefore, we are able to derive the \textit{a priori} SNR estimate \( \hat{\xi}_{\text{PF}}^{MC}(\ell,k) \) as (Yu and Fingscheidt, 2011b)

\[ \hat{\xi}_{\text{PF}}^{MC}(\ell,k) = \frac{\hat{\Phi}_{SS}(\ell,k)}{\Phi_{NN}(\ell,k)} = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{\hat{\Phi}_{SS}^{(i,j)}(\ell,k)}{\Phi_{NN}^{(i,j)}(\ell,k)}, \]  

which is a variance-reduced \textit{a priori} SNR estimation by using spatial smoothing. Please note that even though the noise PSD \( \Phi_{NN}(\ell,k) \) is employed instead of the beamformer
output noise PSD $\hat{\Phi}_{NBF-NBF}(\ell, k)$, $\hat{\xi}_{PF}^{MC}(\ell, k)$ can still be treated as a good estimate, since the noise attenuation effect of the MVDR beamformer can be ignored for car noise, which has most of its energy at low frequencies (Simmer and Wasiljeff, 1992).

In the same manner, comparing Lefkimmiatis’s post-filter estimation based on the theoretically optimal Wiener post-filter estimation (cf. (2.4.1)) to the a priori SNR-driven Wiener post-filter estimation (6.1.1), we can thus get

$$\hat{\xi}_{PF}^{LE}(\ell, k) = \frac{\Phi_{SS}(\ell, k)}{\hat{\Phi}_{NBF-NBF}(\ell, k)}, \quad (6.2.3)$$

where the beamformer output noise PSD $\hat{\Phi}_{NBF-NBF}(\ell, k)$ is applied by taking the very small noise attenuation effect of the MVDR beamformer into account, which makes the only difference to the a priori SNR estimate $\hat{\xi}_{PF}^{MC}(\ell, k)$ in the denominator. Therefore, inserting the beamformer output noise PSD estimate $\hat{\Phi}_{NBF-NBF}(\ell, k)$ given by (2.4.21) and the numerator of (6.2.2) into (6.2.3), we can derive

$$\hat{\xi}_{PF}^{LE}(\ell, k) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\Phi}_{SS}^{(i,j)}(\ell, k) \left[ \hat{\Phi}_{\alpha}^{\beta,\gamma}(\ell, k) \right] \left[ \hat{\Phi}_{\alpha}^{\beta,\gamma}(\ell, k) \right], \quad (6.2.4)$$

which is also a variance-reduced a priori SNR estimate by using spatial smoothing.

Unlike the strong temporal smoothing applied in the single-channel decision-directed approach (6.1.3), the terms $\hat{\xi}_{PF}^{MC}(\ell, k)$ and $\hat{\xi}_{PF}^{LE}(\ell, k)$ are only indirectly based on a temporal smoothing for the auto- and cross-PSD estimation of microphone signals as given in (2.4.5) and (2.4.6), respectively, which is, however, quite relaxed with its adaptive smoothing factor $\alpha(\ell, k)$ being optimized in Section 5.2 with $(\alpha_{1,\text{opt}}, \alpha_{2,\text{opt}}) = (1.0, 0.3)$. In a direct manner, they represent multichannel spatial smoothing rather than single-channel temporal smoothing. Accordingly, we can also interpret $\hat{\xi}_{PF}^{MC}(\ell, k)$ and $\hat{\xi}_{PF}^{LE}(\ell, k)$ as two instantaneous a priori SNR estimates, which can follow a transient a priori SNR change much better and provide better speech quality.

### 6.2.2 Approach by Temporal and Spatial Smoothing

Combining the single-channel decision-directed a priori SNR estimate utilizing temporal smoothing given in (6.1.3) with the multichannel a priori SNR estimates utilizing spatial smoothing given in (6.2.2) and (6.2.4), respectively, a new multichannel decision-directed a priori SNR estimation can be derived as (Yu and Fingscheidt, 2011b)

$$\hat{\xi}_{PF}^{NEW'}(\ell, k) = \beta_1 \cdot \frac{|\hat{S}(\ell-1, k)|^2}{\hat{\Phi}_{NBF-NBF}(\ell-1, k)} + \beta_2 \cdot P[\hat{\gamma}_{BF}(\ell, k) - 1] + \beta_3 \cdot \hat{\xi}_{PF}^{(XX)}(\ell, k),$$

$$\hat{\xi}_{PF}^{NEW}(\ell, k) = \max\{\hat{\xi}_{PF}^{NEW'}(\ell, k), \xi_{\text{min}}\}, \quad (6.2.5)$$

where the term $\hat{\xi}_{PF}^{(XX)}(\ell, k)$ can be replaced by $\hat{\xi}_{PF}^{MC}(\ell, k)$ (6.2.2) or $\hat{\xi}_{PF}^{LE}(\ell, k)$ (6.2.4), respectively. Three weighting factors are now used in (6.2.5): $\beta_1$ and $\beta_2$ correspond to the
6.2.3. Performance Analysis

In order to obtain the optimal vector of weighting factors $\boldsymbol{\beta} = (\beta_1 \beta_2 \beta_3)$ for different driving conditions, the same FoM-based instrumental optimization method as used in Chapter 5 is also applied. The $\theta_p$ within the $\text{FoM}$ calculation in (3.4.5) represents now different possible values of $\boldsymbol{\beta} = (\beta_1 \beta_2 \beta_3)$. The values $(A, B, C, a, b, c) = (0.5, 1, 0.5, 1, 1, 1)$. The optimization is performed with the new $(\text{DS-MC/WF}, \hat{\zeta}_{\text{PF}}^{\text{NEW}})$ system, where the DS

Figure 6.1: Triangle of possible vectors of weighing factors $\boldsymbol{\beta} = (\beta_1 \beta_2 \beta_3)$, where point A represents $\boldsymbol{\beta} = (0 0 1)$ resulting in McCowan’s post-filter (2.4.18) by using $\hat{\zeta}_{\text{PF}}^{\text{MC}}(\ell, k)$ in (6.2.5), and in Lefkimmiatis’s post-filter (2.4.22) by using $\hat{\zeta}_{\text{PF}}^{\text{LE}}(\ell, k)$ in (6.2.5), respectively; point B represents $\boldsymbol{\beta} = (0.98 0.02 0)$ resulting in the classical Wiener filter (6.1.1) according to Scalart and Filho (1996), and point C represents a newly optimized $\boldsymbol{\beta} = (0.8 0.02 0.18)$ for estimating $\hat{\zeta}_{\text{PF}}^{\text{NEW}}(\ell, k)$ in (6.2.5) by using $\hat{\zeta}_{\text{PF}}^{\text{MC}}(\ell, k)$ or $\hat{\zeta}_{\text{PF}}^{\text{LE}}(\ell, k)$ components based on temporal smoothing, while $\beta_3$ is assigned to the component based on spatial smoothing, respectively. Fig. 6.1 shows the possible vectors of weighting factors $\boldsymbol{\beta} = (\beta_1 \beta_2 \beta_3)$ as a shadowed triangle in a unity cubic space, fulfilling $\sum_{i=1}^{3} \beta_i = 1$ and $\beta_i \geq 0$. Applying $\hat{\zeta}_{\text{PF}}^{\text{MC}}(\ell, k)$ in (6.2.5) and setting the vector of weighting factors to $\boldsymbol{\beta} = (0 0 1)$, which is marked as point A in Fig. 6.1, we exactly obtain $\hat{\zeta}_{\text{PF}}^{\text{MC}}(\ell, k)$ (6.2.2). Inserting $\hat{\zeta}_{\text{PF}}^{\text{MC}}(\ell, k)$ into (6.1.1), we thus obtain McCowan’s post-filter (2.4.18). In the same way, applying $\hat{\zeta}_{\text{PF}}^{\text{LE}}(\ell, k)$ in (6.2.5) and setting the vector of weighting factors to $\boldsymbol{\beta} = (0 0 1)$, which is marked as point A in Fig. 6.1, yields exactly $\hat{\zeta}_{\text{PF}}^{\text{LE}}(\ell, k)$ (6.2.4). Inserting $\hat{\zeta}_{\text{PF}}^{\text{LE}}(\ell, k)$ into (6.1.1), Lefkimmiatis’s post-filter (2.4.22) can then be obtained. Setting $\boldsymbol{\beta} = (0.98 0.02 0)$ in (6.2.5), which is marked as point B in Fig. 6.1, yields exactly the classical single-channel $\hat{\zeta}_{\text{PF}}^{\text{DD}}(\ell, k)$ as given in (6.1.3). Inserting $\hat{\zeta}_{\text{PF}}^{\text{DD}}(\ell, k)$ into (6.1.1), the a priori SNR-driven Wiener filter according to Scalart and Filho (1996) can be realized. Hence, our newly proposed multichannel decision-directed a priori SNR estimate $\hat{\zeta}_{\text{PF}}^{\text{NEW}}(\ell, k)$ provides a framework, which is capable of continuously combining the benefits of both temporal and spatial smoothing.

6.2.3 Performance Analysis

In order to obtain the optimal vector of weighting factors $\boldsymbol{\beta} = (\beta_1 \beta_2 \beta_3)$ for different driving conditions, the same FoM-based instrumental optimization method as used in Chapter 5 is also applied. The $\theta_p$ within the $\text{FoM}$ calculation in (3.4.5) represents now different possible values of $\boldsymbol{\beta} = (\beta_1 \beta_2 \beta_3)$. The values $(A, B, C, a, b, c) = (0.5, 1, 0.5, 1, 1, 1)$. The optimization is performed with the new $(\text{DS-MC/WF}, \hat{\zeta}_{\text{PF}}^{\text{NEW}})$ system, where the DS
6. Post-Filter Design with Multichannel Decision-Directed Approach to A Priori SNR Estimation

Figure 6.2: Averaged instrumental measurement results for different driving conditions of the (DS-MC/WF, \(\hat{\xi}_{PF}^{NEW}\)) system using the Wiener post-filter (6.1.1) utilizing \(\hat{\xi}_{PF}^{NEW}\) with the optimized \(\beta = (0.8 \ 0.02 \ 0.18)\) and \(\hat{\xi}_{PF}^{MC}(\ell,k)\); in comparison to the (DS/WF, \(\hat{\xi}_{PF}^{DD}\)) system applying \(\hat{\xi}_{PF}^{DD}(\ell,k)\) to the Wiener post-filter; the (DS-MC/A\(_{opt}\), dif) system and the (DS-MC/A\(_{ref}\), dif) system of Section 5.3.3 are also employed for the comparison beamformer is applied and the new post-filter is computed with \(\hat{\xi}_{PF}^{NEW}\) being inserted in the a priori SNR-driven Wiener post-filter (6.1.1). For the computation of \(\hat{\xi}_{PF}^{NEW}\) in the (DS-MC/WF, \(\hat{\xi}_{PF}^{NEW}\)) system, the multichannel a priori SNR estimate \(\hat{\xi}_{PF}^{MC}(\ell,k)\) is utilized and \(\xi_{min} = -15\) dB is applied in (6.2.5). Furthermore, the optimized smoothing factor \((\alpha_1, \alpha_2) = (1.0, 0.3)\) and the diffuse noise coherence function \(\Gamma_{\ell,k}^{dif}(k)\) is employed to estimate \(\hat{\xi}_{PF}^{MC}(\ell,k)\) in (6.2.2). A full search of \(0 \leq \beta_i \leq 1, i = 1, 2, 3\), with increments of 0.02, \(\sum_{i=1}^{3} \beta_i = 1\) is performed for each driving condition. The reference set \(\{\text{MOS}^3_{\text{LQO,ref}}, \Delta\text{SNR}_{\text{ref}}, \Delta\text{Ψ}^w_{\text{log,ref}}\}\) is calculated by using \(\beta = (0.98 \ 0.02 \ 0.0)\).

After the simulation, we found that the vector of weighting factors \(\beta = (0.8 \ 0.02 \ 0.18)\) yields the maximum FoM for the (DS-MC/WF, \(\hat{\xi}_{PF}^{NEW}\)) system within each driving condition. Please note that with \(\beta_1 = 0.8\) in (6.2.5), which is smaller than the common value of \(\beta_1 = 0.98\), the temporal smoothing is quite relaxed in the multichannel decision-directed approach. Furthermore, the term \(\hat{\xi}_{PF}^{MC}(\ell,k)\) is now weighted with \(\beta_3 = 0.18\), which has a larger weight than the term \(P[\gamma_{BF}(\ell,k) - 1]\) with \(\beta_2 = 0.02\). Hence, \(\hat{\xi}_{PF}^{MC}(\ell,k)\) thus provides a much better instantaneous a priori SNR estimate than \(P[\gamma_{BF}(\ell,k) - 1]\).

The averaged instrumental measurement results for different driving conditions of
the optimized (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system is plotted in Fig. 6.2. For comparison, the (DS/WF, $\xi_{\mathcal{PF}}^{\text{DD}}$) system using $\hat{\xi}_{\mathcal{PF}}^{\text{DD}}(\ell, k)$ in (6.1.1) to calculate the Wiener post-filter is also simulated. This is equivalent to set $\beta = (0.98 \ 0.02 \ 0)$ in (6.2.5) by using only temporal smoothing. Furthermore, the (DS-MC/A_{opt}, dif) system and the (DS-MC/A_{ref}, dif) system of Section 5.3.3 are also employed. These two post-filters are equivalent to set $\beta = (0 \ 0 \ 1)$ in (6.2.5) by using spatial smoothing only. Please note that for all post-filter estimations, the lower floor $\Delta \Psi_{\mathcal{LQO}, \mathcal{PF}}^{\text{new}}$ is utilized instead of the multichannel decision-directed approach with $\beta = (0.8 \ 0.02 \ 0.18)$, the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system has obtained the maximum FoM value of 1.8. In comparison to the (DS/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{DD}}$) system using only temporal smoothing, the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system has improved the noise attenuation performance by showing a higher $\Delta\text{SNR}$ value of about 0.5 dB for all SNR_{in} conditions. Compared to the (DS/MC, $\xi_{\mathcal{PF}}^{\text{DD}}$) system, the speech component quality has been improved by the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system, especially in low SNR_{in} conditions from -5 dB to 5 dB. In addition, the amount of musical tones has been kept in the same level by showing almost the same $\Delta \Psi_{\mathcal{log}, \mathcal{PF}}^{\text{new}}$ values for all SNR_{in} conditions. In comparison to the (DS-MC/A_{opt}, dif) system and the (DS-MC/A_{ref}, dif) system using only spatial smoothing, the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system has significantly improved in noise attenuation performance for all SNR_{in} conditions. Furthermore, the amount of musical tones has also been reduced, especially compared to the (DS-MC/A_{ref}, dif) system, where the heuristically optimized $(\alpha_1, \alpha_2) = (0.8, 0.5)$ (Yu and Fingscheidt, 2010c,b) is applied. For low SNR_{in} conditions, the speech component quality of the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system is more degraded than that of the (DS-MC/A_{opt}, dif) system and the (DS-MC/A_{ref}, dif) system using only spatial smoothing. Nevertheless, the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system still provides an acceptable MOS_{\mathcal{LQO}}^d score of more than 2.5. In summary, by applying $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}(\ell, k)$ utilizing the optimized $\beta = (0.8 \ 0.02 \ 0.18)$ in (6.2.5) with temporal and spatial smoothing, the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system has shown a much better trade-off between the speech component quality, the level of noise attenuation, and the amount of musical tones than using temporal smoothing only (DS-WF, $\hat{\xi}_{\mathcal{PF}}^{\text{DD}}$) and using spatial smoothing only (DS-MC/A_{opt}, dif) and (DS-MC/A_{ref}, dif). These results can also be confirmed by informal listening tests.

Furthermore, a different $a$ priori SNR estimate $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}(\ell, k)$ based on temporal and spatial smoothing can be calculated by inserting $\xi_{\mathcal{PF}}^{\text{LE}}(\ell, k)$ according to (6.2.4) into (6.2.5). Based on this $a$ priori SNR estimate, the (DS-LE/WF, $\xi_{\mathcal{PF}}^{\text{NEW}}$) system can be realized. The only difference to the optimized (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system is that the $a$ priori SNR estimate $\hat{\xi}_{\mathcal{PF}}^{\text{LE}}(\ell, k)$ is utilized instead of the $a$ priori SNR estimate $\hat{\xi}_{\mathcal{PF}}^{\text{MC}}(\ell, k)$ in (6.2.5). The same optimization of $\beta = (\beta_1 \ \beta_2 \ \beta_3)$ as for the (DS-MC/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system is also carried out for the (DS-LE/WF, $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}$) system. It turns out that for each driving condition, again $\beta = (0.8 \ 0.02 \ 0.18)$ is the optimal vector of weighting factors to apply $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}(\ell, k)$ in (6.2.5) to compute $\hat{\xi}_{\mathcal{PF}}^{\text{NEW}}(\ell, k)$. This result is actually also reasonable, since in comparison to $\xi_{\mathcal{PF}}^{\text{MC}}(\ell, k)$, $\hat{\xi}_{\mathcal{PF}}^{\text{LE}}(\ell, k)$ makes the only difference in considering the beamformer noise attenuation performance, which can be neglected for car noise.

The averaged instrumental measurement results for different driving conditions of
Figure 6.3: Averaged instrumental measurement results for different driving conditions of the (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system using the Wiener post-filter (6.1.1) utilizing \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \) with the optimized \( \beta = (0.8 \ 0.02 \ 0.18) \) and \( \xi_{\text{PF}}^{\text{LE}}(\ell,k) \); in comparison to the (DS/WF, \( \hat{\xi}_{\text{DD}}^{\text{PF}} \)) system applying \( \xi_{\text{PF}}^{\text{DD}}(\ell,k) \) to the Wiener post-filter; the (DS-LE/A\(_{\text{opt}}\), dif) system and the (DS-LE/A\(_{\text{ref}}\), dif) system of Section 5.3.3 are also employed for the comparison.

The optimized (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system as shown in Fig. 6.3. The same (DS/WF, \( \hat{\xi}_{\text{PF}}^{\text{DD}} \)) system as shown in Fig. 6.2 by using temporal smoothing only is also applied for comparison. In addition, the (DS-LE/A\(_{\text{opt}}\), dif) system and the (DS-LE/A\(_{\text{ref}}\), dif) system of Section 5.3.3 are applied as the baseline systems using spatial smoothing only, i.e., equivalent to \( \beta = (0 \ 0 \ 1) \) by using \( \hat{\xi}_{\text{PF}}^{\text{LE}}(\ell,k) \) in (6.2.5). It can be seen that the optimized (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system by using \( \beta = (0.8 \ 0.02 \ 0.18) \) has achieved the best overall noise reduction performance by yielding the maximum \( \text{FoM} \) value. Compared to the (DS/WF, \( \hat{\xi}_{\text{PF}}^{\text{DD}} \)) system, all three quality aspects have been improved. Particularly, the (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system can preserve the speech component much better than the (DS/WF, \( \hat{\xi}_{\text{PF}}^{\text{DD}} \)) system by showing higher MOS\(_L^2\)QO scores in low SNR\(_{in}\) conditions from -5 dB to 5 dB. In comparison to the (DS-LE/A\(_{\text{opt}}\), dif) system and the (DS-LE/A\(_{\text{ref}}\), dif) system, the (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system has significantly improved the noise attenuation performance by showing a much higher \( \Delta\text{SNR} \) value for each SNR\(_{in}\) condition. Furthermore, compared to the (DS-LE/A\(_{\text{ref}}\), dif) system, much more musical tones can be suppressed by the (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system. In low SNR\(_{in}\) conditions, the (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \)) system has degraded the speech component compared to the (DS-LE/A\(_{\text{opt}}\), dif) system and the (DS-LE/A\(_{\text{ref}}\), dif) system. However, the (DS-LE/WF, \( \hat{\xi}_{\text{NEW}}^{\text{PF}} \))
system can still provide an acceptable MOS$_{LQO}^S$ score of more than 2.4 and even improves the MOS$_{LQO}^S$ scores for high SNR in conditions. Therefore, the results presented in Fig. 6.3 again prove that by applying the multichannel decision-directed a priori SNR estimate $\hat{\xi}_{PF}^{\text{NEW}}(\ell,k)$ based on temporal and spatial smoothing (DS-LE/WF, $\hat{\xi}_{PF}^{\text{NEW}}$), a much better trade-off between the speech component quality, the level of noise attenuation, and the amount of musical tone can be achieved than using temporal smoothing only (DS/WF, $\hat{\xi}_{PF}^{\text{DD}}$) and using spatial smoothing only (DS-LE/$A_{\text{opt}}$, dif) and (DS-LE/$A_{\text{ref}}$, dif). These results can also be confirmed by informal listening tests.

It is important to note that only slight differences between the results of the (DS-LE/WF, $\hat{\xi}_{PF}^{\text{NEW}}$) system in Fig. 6.3 and the (DS-MC/WF, $\hat{\xi}_{PF}^{\text{NEW}}$) system in Fig. 6.2 can be observed. Hence, both $\hat{\xi}_{PF}^{\text{LE}}(\ell,k)$ and $\hat{\xi}_{PF}^{\text{MC}}(\ell,k)$ can be applied in (6.2.5) to calculate the new a priori SNR $\xi_{PF}^{\text{NEW}}$ without performance degradation.

6.3 Summary

In this chapter, a new post-filter design based on a multichannel decision-directed a priori SNR estimator exploiting both temporal and spatial smoothing has been proposed. The classical single-channel decision-directed a priori SNR estimator based on temporal smoothing can perform well in suppression of musical tones, however, the speech component is unfortunately also degraded. On the other hand, the multichannel a priori SNR estimator based on spatial smoothing can provide good speech component quality, while the noise attenuation performance is moderate and the suppression of musical tones is not enough. Motivated by these factors, a new multichannel decision-directed a priori SNR estimator is formulated, which combines the benefits of the single-channel decision-directed a priori SNR estimator with the multichannel a priori SNR estimator.

Employing the multichannel decision-directed a priori SNR estimate based on temporal and spatial smoothing to formulate a new post-filter design, the weighting factors of the multichannel decision-directed a priori SNR estimator have been optimized for different driving conditions by utilizing the FoM-based instrumental optimization method. Applying the optimized new post-filter design, a much better overall noise reduction performance can be achieved. Particularly, in comparison to the Wiener post-filter using the a priori SNR estimate based on temporal smoothing only, a much better speech component quality can be achieved. Furthermore, in comparison to the baseline McCowan’s and Lefkimmiatis’s post-filters using the a priori SNR estimate based on spatial smoothing only, the noise attenuation performance can be significantly improved and more musical tones can be suppressed.
6. Post-Filter Design with Multichannel Decision-Directed Approach to *A Priori* SNR Estimation
Chapter 7

Conclusions

In this thesis, multichannel automotive speech enhancement utilizing beamforming and post-filtering has been addressed. The main objectives of this thesis are new post-filter designs and their optimization in an automotive environment. Two novel post-filter designs have been presented: The first post-filter design utilizes a modified PSD estimation using an adaptive smoothing factor as well as a hybrid noise coherence function. The second post-filter design employs a multichannel decision-directed a priori SNR estimator based on both temporal and spatial smoothing. The optimization of the two new post-filters has been achieved for different driving conditions by employing a figure of merit (FoM), which is an entity of three independent instrumental quality measures covering the speech component quality, the level of noise reduction, and the amount of musical tones. Particularly, in this thesis, a new weighted log kurtosis ratio has been proposed to instrumentally assess musical tones in a black-box manner. In this thesis, the following results have been achieved:

Based on a unified instrumental evaluation framework, which can be applied to both single-channel and multichannel noise reduction algorithms, several speech- and noise-related instrumental quality measures have been defined. In particular, a new black-box weighted log kurtosis ratio for instrumentally measuring musical tones has been presented. Compared to the white-box log kurtosis ratio, the black-box weighted log kurtosis ratio does not require any knowledge of internal variables of the noise reduction algorithm under test. In addition, the weighted log kurtosis ratio can be applied to a wide range of noise reduction algorithms including those employing the decision-directed a priori SNR estimation. The weighted log kurtosis ratio has been evaluated on both wideband and narrowband signals by employing different noise reduction algorithms. Furthermore, a subjective ACR listening test to assess musical tones has been carried out separately for wideband and narrowband signals. The weighted log kurtosis ratio has been shown to be quite insensitive to different calculation methods and measurement setups. Overall correlations of $|\rho| = 0.953$ and $|\rho| = 0.988$ between the instrumental and subjective measurements have been achieved for assessing musical tones for wideband and narrowband signals, respectively. Therefore, the weighted log kurtosis ratio is capable of instrumentally assessing musical tones for wideband and narrowband signals.
In order to optimize noise reduction algorithms, three instrumental quality measurements are taken into consideration: The speech component quality, the level of noise attenuation, and the amount of musical tones. In this thesis, a figure of merit (FoM) has been proposed, which is an entity of combining three independent instrumental quality measures, namely PESQ MOS ($\text{MOS}_{\text{LQO}}$) measuring the speech component quality, SNR improvement ($\Delta \text{SNR}$) measuring the effective level of noise attenuation, and weighted log kurtosis ratio ($\Delta \Psi_{\text{log}}^w$) measuring the amount of musical tones. Thanks to some independence between the three instrumental quality measures, the FoM provides a useful means to instrumentally optimize noise reduction algorithms. As an application example, we have shown how to automatically search for the optimal smoothing factor and the optimal a priori SNR floor of the single-channel decision-directed a priori SNR estimation using an FoM. After the optimization, yet unknown optimal values of the smoothing factor and the a priori SNR floor have been identified and subjectively verified for some noise reduction algorithms.

For the optimization and evaluation of different post-filter designs, a new multichannel in-car speech, noise and CAN bus database has been acquired. The new multichannel database includes 18 driving conditions and has been acquired by using a head unit-integrated microphone array in a VW Touran. Different to conventional multichannel in-car speech and noise databases, the CAN bus data describing driving conditions in a real-time manner have also been synchronously collected. The CAN bus data gives the potential benefit of providing car driving conditions in real-time, which inspires the optimization of the new post-filter designs for different driving conditions.

In the context of multichannel noise reduction algorithms, delay-and-sum as well as constrained superdirective beamformers, which belong to the class of MVDR beamformers, are usually applied to reduce car noise. However, instrumental evaluation by using different beamforming measures has shown that both beamformers are not able to provide sufficient noise attenuation at low frequencies, where car noise energies are concentrated most. Therefore, a multichannel Wiener filter comprising an MVDR beamformer and a post-filter has to be applied. Current state-of-the-art post-filter estimations employ a fixed smoothing factor to the speech and noise PSD estimations, which faces a compromise between speech distortion and noise distortion in terms of musical tones. To address this problem, an adaptive smoothing factor is proposed, which adapts between active speech and speech pause frames. Furthermore, the a priori knowledge of a diffuse noise coherence function is usually applied to car noise. However, in this thesis, a considerable mismatch has been detected between the diffuse and the measured noise coherence functions for different driving conditions. Based on this observation, a hybrid noise coherence function, which is a mixture of the diffuse noise coherence function and the measured noise coherence function for a specific driving condition, has been proposed. Applying the adaptive smoothing factor and the hybrid noise coherence function, a new post-filter design is realized. Applying an FoM, the adaptive smoothing factor and the hybrid noise coherence function have been individually optimized for different driving conditions. Compared to state-of-the-art post-filters using the fixed smoothing factor and the diffuse noise coherence function throughout, a significant improvement of overall noise

Finally, another new post-filter design has been proposed, which focuses on the \textit{a priori} SNR estimation. Applying the multichannel inputs, an \textit{a priori} SNR estimate can be calculated by utilizing spatial smoothing. Combining this multichannel \textit{a priori} SNR estimate with the single-channel decision-directed \textit{a priori} SNR estimation using temporal smoothing, a multichannel decision-directed \textit{a priori} SNR estimation exploiting both temporal and spatial smoothing has been formulated. Thanks to the multichannel decision-directed approach, it is thus possible to suppress more musical tones than using spatial smoothing only and to improve the speech component quality more than using temporal smoothing only — both at the same time. However, the weighting factors for the temporal and spatial smoothing parts within the multichannel decision-directed approach have to be carefully chosen. In analogy to the optimization of the adaptive smoothing factor and the hybrid noise coherence function, the same FoM has been applied to automatically identify the optimal weighting factors. Employing the optimized weighting factors, the post-filter utilizing the multichannel decision-directed \textit{a priori} SNR estimate provides better overall noise reduction performance than those post-filters using single-channel decision-directed \textit{a priori} SNR estimate only (temporal smoothing) and using multichannel \textit{a priori} SNR estimate only (spatial smoothing).
Appendix A

In-Car Audio Data Collection

A.1 Audio Data Collection System Design

The following audio equipments and hardwares were applied to the speech and noise signals acquisition:

- Two audio interfaces of type RME Fireface 400
- Two microphone amplifiers of type RME OctaMic II
- One laptop of type IBM ThinkPad T61 installed with the audio acquisition software Audition 3.0
- 20 omnidirectional electret microphones (type Monacor MCE-4500) of the head unit-integrated microphone array (see Fig. 4.2(b)) and one hands-free microphone being positioned in the overhead light module

The block diagram of the audio data collection system is shown in Fig. A.1. Two pairs of RME Fireface 400 and RME OctaMic II were connected for the audio data collection. For each electret microphone, an amplifier is required. Each RME Fireface 400 can provide 2 channels with one amplifier per channel, and each OctaMic II can provide 8 channels with one amplifier per channel. Within each pair, the 8 amplified microphone signals acquired by OctaMic II were transmitted to the corresponding RME Fireface 400 via an optical ADAT connection. Along with the two amplified channels of RME Fireface 400, each pair can thus provide 10 amplified channels. Therefore, 20 electret microphones of the microphone array can be simultaneously acquired by two RME Fireface 400 and two RME OctaMic II. The synchronization between these four audio equipments has been solved by utilizing one RME Fireface 400 as the master interface, which provided the master clock via a BNC connection to the slave RME Fireface 400 and the two RME OctaMic II. The signals of the slave RME Fireface 400 and its corresponding OctaMic II were transmitted to the master Fireface 400 via a FireWire connection. In addition, the hands-free microphone signal was acquired by the master
A. In-Car Audio Data Collection

Figure A.1: Setup of the audio data collection system using two audio interfaces (RME Fireface 400), two microphone amplifiers (RME OctaMic II), and IBM ThinkPad T61 installed with the audio acquisition software Audition 3.0.

RME Fireface 400. Finally, the master RME Fireface 400 acted as the external interface to the laptop by transmitting all 21 synchronized microphone signals via a FireWire connection. The multichannel audio data collection was then performed by applying the audio acquisition software Audition 3.0.

A.2 In-Car Realization of the Audio Data Collection System

In order to realize the audio data collection system as depicted in Fig. A.1 in the VW Touran, a power supply system was implemented. The power supply system was constructed by employing an automotive battery (12 V, 88 Ah) and an AC/DC power inverter (350 W, 12 V/230 V). The 12 VDC electricity provided by the automotive battery can thus be converted to the 230 VAC electricity, which provided the AC power to the audio data collection system. All audio equipments along with the power supply system were installed in the luggage compartment (see Fig. 4.1(b)).
A.2. In-Car Realization of the Audio Data Collection System

Since the output voltage of the applied electret microphone (type *Monacor MCE-4500*) is too small, an additional preamplifier providing further amplification was applied to each electret microphone of the microphone array. The preamplifiers were installed inside the box of the prototype representing the head unit (see Fig. A.2(a)). The transmission of the multichannel audio signals were realized in two stages. In the first stage, the acquired multichannel audio signals were transmitted via flat ribbon cables from the head unit-integrated microphone array to the circular connectors in the glove compartment. For the multichannel audio signals with 20 channels, three flat ribbon cables with three circular connectors were applied. Fig. A.2(a) shows the connection of the flat ribbon cables to the microphone array. Please note that the microphone array was taken out of the head unit position to show the connection at the backside. During the audio data acquisition, the flat ribbon cables were laid from the backside of the microphone array to the circular connectors in the glove compartment (see Fig. A.2(b)). In the second stage, the multichannel audio signals were transmitted via audio cables from the glove compartment to the luggage compartment. In the luggage compartment, the multichannel signals were then transmitted via the circular connectors to separate XLR connectors.

**Figure A.2:** (a) Back view of the applied head unit-integrated microphone array prototype with preamplifiers and being connected to flat ribbon cables, (b) circular connectors in the glove compartment, and (c) circular connectors in the luggage compartment
The circular connectors in the luggage compartment is shown in Fig. A.2(c). Finally, the multichannel audio signals were transmitted to the two pairs of RME Fireface 400 and RME OctaMic II as shown in Fig. A.1 via the XLR connectors. The applied audio cables using the circular connector with the XLR connectors as well as the circular connector with the flat ribbon cable and its connector are shown in Fig. A.3.
Appendix B

Acoustical-Related CAN Bus Parameters

In the following, the selected acoustical-related CAN bus parameters are listed:

- **BrakeActuation**: Activation of the brake pedal (True, False)
- **ClutchSwitchActuation**: Activation of the clutch switch (True, False)
- **DashboardIndicators**: Activation of each dash board indicator displayed on the dash board (True, False)
- **DisplayedVehicleSpeed**: Displayed vehicle speed on the tachometer including the deviation caused by indicator needle (0…450 km/h)
- **DoorPosition**: Status of all doors including the four side doors and the door of the luggage compartment (True, False, Ajar\(^8\))
- **EngineSpeed**: Engine rotation speed (0…10000 r/min)
- **Horn**: Activation of the horn (True, False)
- **IgnitionKey**: Position of the ignition key (NoKeyPresent, KeyRemovable, ElectricityOnMotorOff, MotorOn, StarterActive)
- **SteeringWheelAngle**: Rotation angles of the steering wheel (\(-1500^\circ...1500^\circ\)), where positive values stand for a clockwise rotation and negative values stand for a counterclockwise rotation.
- **SteeringWheelAngularVelocity**: Rotation angular rate of the steering wheel (-1500...1500°/s)
- **VehicleSpeed**: Vehicle speed (0…450 km/h)
- **WindowAperture**: The aperture of four windows (0…100%)

\(^8\)Ajar means that a door is not completely closed.
• **WiperControl**: Status of the front and rear wipers (Idle, WipeOnce, Intermittent, Normal, Fast, WasherActive)

Unfortunately, no CAN bus parameter describing the air-conditioning operation level is provided by the VW Touran. Therefore, air-conditioning operation levels of each driving condition have been manually collected.
## List of Abbreviations and Symbols

### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACR</td>
<td>Absolute category rating</td>
</tr>
<tr>
<td>AG</td>
<td>Array gain</td>
</tr>
<tr>
<td>AMNOR</td>
<td>Adaptive microphone array system for noise reduction</td>
</tr>
<tr>
<td>BP</td>
<td>Beam pattern</td>
</tr>
<tr>
<td>CAN</td>
<td>Controller area network</td>
</tr>
<tr>
<td>CS</td>
<td>Constrained superdirective</td>
</tr>
<tr>
<td>dBov</td>
<td>Decibel overload</td>
</tr>
<tr>
<td>DD</td>
<td>Decision-directed</td>
</tr>
<tr>
<td>DI</td>
<td>Directivity index</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier transform</td>
</tr>
<tr>
<td>DS</td>
<td>Delay-and-sum</td>
</tr>
<tr>
<td>ECU</td>
<td>Electronic control unit</td>
</tr>
<tr>
<td>FoM</td>
<td>Figure of merit</td>
</tr>
<tr>
<td>GSVD</td>
<td>Generalized singular value decomposition</td>
</tr>
<tr>
<td>IDFT</td>
<td>Inverse discrete Fourier transform</td>
</tr>
<tr>
<td>LC</td>
<td>Linearly constrained</td>
</tr>
<tr>
<td>LCMV</td>
<td>Linearly constrained minimum variance</td>
</tr>
<tr>
<td>LE</td>
<td>Lefkimmiatis’s post-filter</td>
</tr>
<tr>
<td>LMS</td>
<td>Least mean square</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear predictive coding</td>
</tr>
<tr>
<td>LSA</td>
<td>Log-spectral amplitude</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a posteriori</td>
</tr>
<tr>
<td>MC</td>
<td>McCowan’s post-filter</td>
</tr>
<tr>
<td>MIRS</td>
<td>Modified intermediate reference system</td>
</tr>
<tr>
<td>MMSE</td>
<td>Minimum mean square error</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean opinion score</td>
</tr>
<tr>
<td>MOS_LQS</td>
<td>MOS listening quality subjective</td>
</tr>
<tr>
<td>MOS_LQO</td>
<td>MOS listening quality objective</td>
</tr>
<tr>
<td>MOS_raw</td>
<td>Raw MOS</td>
</tr>
<tr>
<td>MV</td>
<td>Minimum variance</td>
</tr>
<tr>
<td>MVDR</td>
<td>Minimum variance distortionless response</td>
</tr>
<tr>
<td>NA</td>
<td>Noise attenuation</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NL</td>
<td>Noise level</td>
</tr>
<tr>
<td>NLMS</td>
<td>Normalized least mean square</td>
</tr>
<tr>
<td>N-MOS</td>
<td>Noise mean opinion score</td>
</tr>
<tr>
<td>OLA</td>
<td>Overlap-and-add</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability density function</td>
</tr>
<tr>
<td>PESQ</td>
<td>Perceptual evaluation of speech quality</td>
</tr>
<tr>
<td>PF</td>
<td>Post-filter</td>
</tr>
<tr>
<td>PSD</td>
<td>Power spectrum density</td>
</tr>
<tr>
<td>PSQM</td>
<td>Perceptual speech quality measure</td>
</tr>
<tr>
<td>RA</td>
<td>Relative approach</td>
</tr>
<tr>
<td>RMS</td>
<td>Root mean square</td>
</tr>
<tr>
<td>RNS</td>
<td>Radio navigation system</td>
</tr>
<tr>
<td>SA</td>
<td>Spectral amplitude</td>
</tr>
<tr>
<td>SD</td>
<td>Superdirective</td>
</tr>
<tr>
<td>SG</td>
<td>Super-Gaussian</td>
</tr>
<tr>
<td>SI</td>
<td>Simmer’s post-filter</td>
</tr>
<tr>
<td>S-MOS</td>
<td>Speech mean opinion score</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>SSDR</td>
<td>Speech-to-speech distortion ratio</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission control protocol</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice activity detector</td>
</tr>
<tr>
<td>WF</td>
<td>Wiener filter</td>
</tr>
<tr>
<td>WNG</td>
<td>White noise gain</td>
</tr>
<tr>
<td>ZE</td>
<td>Zelinski’s post-filter</td>
</tr>
</tbody>
</table>
## List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\cdot)^*$</td>
<td>Complex conjugate operator</td>
</tr>
<tr>
<td>$(\cdot)^H$</td>
<td>Hermitian operator</td>
</tr>
<tr>
<td>$(\cdot)^T$</td>
<td>Transpose operator</td>
</tr>
<tr>
<td>$</td>
<td>\cdot</td>
</tr>
<tr>
<td>$\partial$</td>
<td>Partial differentiation operator</td>
</tr>
<tr>
<td>$\text{arg}(\cdot)$</td>
<td>Phase operator</td>
</tr>
<tr>
<td>$a, b, c$</td>
<td>Exponents of an FoM</td>
</tr>
<tr>
<td>$A, B, C$</td>
<td>Weighting factors of an FoM</td>
</tr>
<tr>
<td>$\text{ASL}_x$</td>
<td>Active speech level of $x(n)$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Fixed smoothing factor for the PSD estimation</td>
</tr>
<tr>
<td>$\alpha(\ell, k)$</td>
<td>Adaptive smoothing factor for the PSD estimation</td>
</tr>
<tr>
<td>$(\alpha_1, \alpha_2)$</td>
<td>Coefficients used for computing $\alpha(\ell, k)$</td>
</tr>
<tr>
<td>$\alpha_n(k)$</td>
<td>Normalization factor</td>
</tr>
<tr>
<td>$\textbf{B}(\ell, k)$</td>
<td>Blocking matrix</td>
</tr>
<tr>
<td>$c$</td>
<td>Speed of sound</td>
</tr>
<tr>
<td>$\mathbb{C}$</td>
<td>Set of complex numbers</td>
</tr>
<tr>
<td>$C$</td>
<td>Constraint matrix</td>
</tr>
<tr>
<td>$d$</td>
<td>Distance of adjacent microphones within a linear equidistant microphone array</td>
</tr>
<tr>
<td>$d_{i,j}$</td>
<td>Distance between two microphones $i$ and $j$</td>
</tr>
<tr>
<td>$d$</td>
<td>Relative position difference vector between two points</td>
</tr>
<tr>
<td>$\textbf{D}(k)$</td>
<td>Propagation vector</td>
</tr>
<tr>
<td>$\delta(k)$</td>
<td>Mixing factor of the hybrid noise coherence function</td>
</tr>
<tr>
<td>$\Delta L$</td>
<td>Frame shift</td>
</tr>
<tr>
<td>$\Delta RA_{x,y}(\ell, k)$</td>
<td>Difference of the relative approach values of $x(n)$ and $y(n)$</td>
</tr>
<tr>
<td>$\Delta \text{SNR}$</td>
<td>Relative SNR improvement</td>
</tr>
<tr>
<td>$\Delta \Omega$</td>
<td>Phase shift</td>
</tr>
<tr>
<td>$\Delta \Psi_{\log}$</td>
<td>Black-box log kurtosis ratio</td>
</tr>
<tr>
<td>$\Delta \Psi'_{\log}$</td>
<td>White-box log kurtosis ratio</td>
</tr>
<tr>
<td>$\Delta \Psi''_{\log}$</td>
<td>Black-box weighted log kurtosis ratio</td>
</tr>
<tr>
<td>$E{\cdot}$</td>
<td>Expectation operator</td>
</tr>
<tr>
<td>$E_{i,j}^{(Q,0.5Q)}(k)$</td>
<td>Square error between the measured noise coherence functions using $Q$ noise signals and $0.5Q$ noise signals</td>
</tr>
<tr>
<td>$\exp(\cdot)$</td>
<td>Exponential function</td>
</tr>
<tr>
<td>$f_N$</td>
<td>Nyquists frequency</td>
</tr>
<tr>
<td>$f_s$</td>
<td>Sampling frequency</td>
</tr>
<tr>
<td>$f_w$</td>
<td>Wave frequency</td>
</tr>
<tr>
<td>$\overline{\text{FoM}}$</td>
<td>Mean FoM value</td>
</tr>
<tr>
<td>$G$</td>
<td>Fullband attenuation factor</td>
</tr>
<tr>
<td>$G(\ell, k)$</td>
<td>Spectral gain of a weighting rule</td>
</tr>
<tr>
<td>$\gamma(\ell, k)$</td>
<td>A posteriori SNR</td>
</tr>
<tr>
<td>$\Gamma_{i,j}(k)$</td>
<td>Noise coherence function of points $i$ and $j$</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$\Gamma_{\text{inc}}_{ij}(k)$</td>
<td>Incoherent noise coherence function of points $i$ and $j$</td>
</tr>
<tr>
<td>$\Gamma_{\text{dif}}_{ij}(k)$</td>
<td>Spherically isotropic (diffuse) noise coherence function of points $i$ and $j$</td>
</tr>
<tr>
<td>$\Gamma_{\text{cyl}}_{ij}(k)$</td>
<td>Cylindrically isotropic noise coherence function of points $i$ and $j$</td>
</tr>
<tr>
<td>$\Gamma_{\text{hyb}}_{ij}(k)$</td>
<td>Hybrid noise coherence function of points $i$ and $j$</td>
</tr>
<tr>
<td>$\Gamma_{\text{NN}}(k)$</td>
<td>Noise coherence matrix</td>
</tr>
<tr>
<td>$\Gamma_{\text{inc}}_{\text{NN}}(k)$</td>
<td>Incoherent noise coherence matrix</td>
</tr>
<tr>
<td>$\Gamma_{\text{dif}}_{\text{NN}}(k)$</td>
<td>Diffuse noise coherence matrix</td>
</tr>
<tr>
<td>$\tilde{\Gamma}<em>{\text{Q}}</em>{ij}(k)$</td>
<td>Measured noise coherence function of microphone $i$ and $j$ using $Q$ noise signals with a length of 8 s</td>
</tr>
<tr>
<td>$H_{\text{PF}}(\ell,k)$</td>
<td>Post-filter coefficient</td>
</tr>
<tr>
<td>$H_{\text{min}}$</td>
<td>Lower floor of $H_{\text{PF}}(\ell,k)$</td>
</tr>
<tr>
<td>$I$</td>
<td>Identity matrix</td>
</tr>
<tr>
<td>$I_{N_{\text{c}}}(\ell,k)$</td>
<td>Modified auto-periodogram of $N_{i}(\ell,k)$</td>
</tr>
<tr>
<td>$I_{N_{\text{c}}N_{\text{c}}}(\ell,k)$</td>
<td>Modified cross-periodogram of $N_{i}(\ell,k)$ and $N_{j}(\ell,k)$</td>
</tr>
<tr>
<td>$J_{0}(\cdot)$</td>
<td>Zero-order Bessel function of the first kind</td>
</tr>
<tr>
<td>$k$</td>
<td>Frequency bin</td>
</tr>
<tr>
<td>$k_{\text{up}}$</td>
<td>Upper frequency bin of the hybrid noise coherence function</td>
</tr>
<tr>
<td>$k$</td>
<td>Wavenumber vector</td>
</tr>
<tr>
<td>$K$</td>
<td>DFT length</td>
</tr>
<tr>
<td>$\ell$</td>
<td>Frame index</td>
</tr>
<tr>
<td>$L$</td>
<td>Frame length</td>
</tr>
<tr>
<td>$\mathcal{L}$</td>
<td>Set of frames in a database</td>
</tr>
<tr>
<td>$\ln(\cdot)$</td>
<td>Natural logarithm operator</td>
</tr>
<tr>
<td>$\Lambda_{H_{0}}$</td>
<td>Speech pause frames (containing no speech dominant parts)</td>
</tr>
<tr>
<td>$\Lambda_{H_{1}}$</td>
<td>Speech active frames</td>
</tr>
<tr>
<td>$\lambda_{s}$</td>
<td>Segment index</td>
</tr>
<tr>
<td>$\lambda_{w}$</td>
<td>Wavelength</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of microphones</td>
</tr>
<tr>
<td>$\text{MOS}^{\text{L}}_{\text{QO}}$</td>
<td>PESQ MOS for measuring the speech component quality</td>
</tr>
<tr>
<td>$\mu_{x}$</td>
<td>Mean value of $x(n)$</td>
</tr>
<tr>
<td>$n$</td>
<td>Discrete time index</td>
</tr>
<tr>
<td>$n(n)$</td>
<td>Single-channel noise signal</td>
</tr>
<tr>
<td>$N(\ell,k)$</td>
<td>Single-channel noise signal in the frequency domain</td>
</tr>
<tr>
<td>$n_{i}^{\ell}(n)$</td>
<td>Noise signal of microphone $i$</td>
</tr>
<tr>
<td>$N_{i}^{\ell}(\ell,k)$</td>
<td>Noise signal of microphone $i$ in the frequency domain</td>
</tr>
<tr>
<td>$N(\ell,k)$</td>
<td>Vector of noise signals of all microphones in the frequency domain</td>
</tr>
<tr>
<td>$N_{i}(\ell,k)$</td>
<td>Delay-compensated noise signal of microphone $i$ in the frequency domain</td>
</tr>
<tr>
<td>$N(\ell,k)$</td>
<td>Vector of delay-compensated noise signals of all microphones in the frequency domain</td>
</tr>
<tr>
<td>$\tilde{\eta}(n)$</td>
<td>Filtered noise signal</td>
</tr>
<tr>
<td>$\nabla^{2}$</td>
<td>Laplace operator</td>
</tr>
<tr>
<td>$N_{c}$</td>
<td>Number of linearly independent constraints</td>
</tr>
<tr>
<td>$\omega_{w}$</td>
<td>Angular frequency of a wave</td>
</tr>
<tr>
<td>$\mathbf{p}$</td>
<td>Position vector in an Euclidean space</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>$p(p, t)$</td>
<td>Sound pressure of a wave</td>
</tr>
<tr>
<td>$p(p, n)$</td>
<td>A point being modeled as a random process in a noise field</td>
</tr>
<tr>
<td>$p(x)$</td>
<td>Probability density function of a random variable $x$</td>
</tr>
<tr>
<td>$\varphi_{i}(n)$</td>
<td>Auto-correlation function of the random point $i$</td>
</tr>
<tr>
<td>$\varphi_{i,j}(d, n)$</td>
<td>Cross-correlation function of the random points $i$ and $j$</td>
</tr>
<tr>
<td>$\Phi_{i}(k)$</td>
<td>Auto-PSD of the random point $i$</td>
</tr>
<tr>
<td>$\Phi_{i,j}(k)$</td>
<td>Cross-PSD of the random points $i$ and $j$</td>
</tr>
<tr>
<td>$\Phi_{o}(k)$</td>
<td>Auto-PSD of the source signal</td>
</tr>
<tr>
<td>$\Phi_{SS}(\ell, k)$</td>
<td>Clean speech signal PSD</td>
</tr>
<tr>
<td>$\Phi_{NN}(\ell, k)$</td>
<td>Noise signal PSD</td>
</tr>
<tr>
<td>$\Phi_{Y'S}(\ell, k)$</td>
<td>Cross-PSD vector of $Y'(\ell, k)$ and $S(\ell, k)$</td>
</tr>
<tr>
<td>$\Phi_{Y'Y'}(\ell, k)$</td>
<td>Auto-PSD matrix of $Y'(\ell, k)$</td>
</tr>
<tr>
<td>$\Phi_{EE}(\ell, k)$</td>
<td>Error signal PSD</td>
</tr>
<tr>
<td>$\Phi_{N'S}(\ell, k)$</td>
<td>Cross-PSD vector of $N'(\ell, k)$ and $S(\ell, k)$</td>
</tr>
<tr>
<td>$\Phi_{SS_S'B'F}(\ell, k)$</td>
<td>Clean speech signal PSD after the beamformer</td>
</tr>
<tr>
<td>$\Phi_{NN_S'B'F}(\ell, k)$</td>
<td>Noise signal PSD after the beamformer</td>
</tr>
<tr>
<td>$\Phi_{Y_1Y_2}(\ell, k)$</td>
<td>Cross-PSD of $Y_1(\ell, k)$ and $Y_2(\ell, k)$</td>
</tr>
<tr>
<td>$\Phi_{SS_S'B'F}(\ell, k)$</td>
<td>Auto-PSD of the beamformer output $\hat{S}_{BF}(\ell, k)$</td>
</tr>
<tr>
<td>$\Psi_x$</td>
<td>Kurtosis of a random variable $x$ using raw moments</td>
</tr>
<tr>
<td>$\Psi_x$</td>
<td>Kurtosis of a random variable $x$ using central moments</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>Set of real numbers</td>
</tr>
<tr>
<td>$RA_x(\ell, k)$</td>
<td>Relative approach value of $x(n)$</td>
</tr>
<tr>
<td>$\text{Re}{\cdot}$</td>
<td>Real operator</td>
</tr>
<tr>
<td>$(r, \theta, \phi)$</td>
<td>Spherical coordinate system</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Pearson’s correlation coefficient</td>
</tr>
<tr>
<td>$s(n)$</td>
<td>Single-channel clean speech signal</td>
</tr>
<tr>
<td>$S(\ell, k)$</td>
<td>Single-channel clean speech signal in the frequency domain</td>
</tr>
<tr>
<td>$s_i'(n)$</td>
<td>Clean speech signal of microphone $i$</td>
</tr>
<tr>
<td>$S_i'(\ell, k)$</td>
<td>Clean speech signal of microphone $i$ in the frequency domain</td>
</tr>
<tr>
<td>$S'(\ell, k)$</td>
<td>Vector of clean speech signals of all microphones in the frequency domain</td>
</tr>
<tr>
<td>$S_i(\ell, k)$</td>
<td>Delay-compensated clean speech signal of microphone $i$ in the frequency domain</td>
</tr>
<tr>
<td>$S(\ell, k)$</td>
<td>Vector of delay-compensated clean speech signals of all microphones in the frequency domain</td>
</tr>
<tr>
<td>$\hat{s}(n)$</td>
<td>Enhanced speech signal</td>
</tr>
<tr>
<td>$\hat{S}(\ell, k)$</td>
<td>Enhanced speech signal in the frequency domain</td>
</tr>
<tr>
<td>$\hat{S}_{BF}(\ell, k)$</td>
<td>Enhanced speech signal of the beamformer in the frequency domain</td>
</tr>
<tr>
<td>$\hat{s}(n)$</td>
<td>Filtered speech signal</td>
</tr>
<tr>
<td>$\text{SNR}_{BF}(\ell, k)$</td>
<td>SNR of the beamformer output</td>
</tr>
<tr>
<td>$\text{SSDR}_{\text{max}}$</td>
<td>Maximum SSDR value</td>
</tr>
<tr>
<td>$\text{SSDR}(\lambda_s)$</td>
<td>SSDR at the segment $\lambda_s$</td>
</tr>
<tr>
<td>$\text{SSDR}_{\text{seg}}$</td>
<td>Segmental SSDR</td>
</tr>
<tr>
<td>$\sigma^2_x$</td>
<td>Variance of $x(n)$</td>
</tr>
</tbody>
</table>
sinc(·) Sinc function
t Continuous time index
θo Angle of arrival
τi,j Time delay between two microphones i and j
u Unit vector
V Amplitude of a plane wave
\( W(\ell,k) \) Beamformer coefficients vector
\( W_{\text{INT}}(\ell,k) \) Interference cancellation module coefficients vector
\( W_{\text{opt}}(\ell,k) \) Optimal multichannel Wiener filter coefficients vector
(\( x, y, z \)) Cartesian coordinate system
\( \xi(\ell,k) \) A priori SNR
\( \xi_{\text{min}} \) A priori SNR floor
\( \Xi(\ell,k) \) Vector applied to the constraint matrix to form linearly independent constraints
\( y(n) \) Single-channel noisy speech signal
\( Y(\ell,k) \) Single-channel noisy speech signal in the frequency domain
\( y'_i(n) \) Noisy speech signal of microphone i
\( Y'_i(\ell,k) \) Noisy speech signal of microphone i in the frequency domain
\( Y'_i(\ell,k) \) Vector of noisy speech signals of all microphones in the frequency domain
\( Y_i(\ell,k) \) Delay-compensated noisy speech signal of microphone i in the frequency domain
\( Y(\ell,k) \) Vector of delay-compensated noisy speech signals of all microphones in the frequency domain
Bibliography


ETSI (2008a). ETSI EG 202 396-1 Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise Part 1: Background Noise Simulation Technique and Background Noise Database.


Bibliography


Own Publications


