Department of Electrical and Computer Engineering

Speech Enhancement in Binaural Hearing Protection Devices

Pei Chee Yong

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

October 2013

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: Yong Pei Chee .....

Date: 24/02/2014 .....

Acknowledgements

First and foremost, I thank God who gave me the grace and privilege to pursue my PhD, which turned out to be quite a remarkable journey in spite of many challenges faced. No man is an island, and it is a journey I have travelled along with the support and encouragement of numerous people who contributed in many ways and made this thesis possible. At the end of this thesis, I would like to express my heartfelt gratitude to all of them, especially the selected few I am about to mention, for without their help, I would not have gotten this far.

I am deeply indebted to my supervisor Prof. Sven Nordholm for his enthusiasm, patience, insight, and every bit of guidance, assistance, involvement and encouragement that he has given me throughout the years to produce this thesis. I would also like to thank my co-supervisor Dr. Hai Huyen Dam for her knowledge, encouragement, patience, and valuable suggestions and feedbacks in various matters. The countless inspirational discussions with two of these great people (and friends) have been beautiful memories that I cherished and enjoyed a lot.

Further, I would like to thank Dr. Yee Hong Leung for his enthusiastic introduction to the world of digital signal processing during my undergraduate course of study. I would also like to thank the following researchers for their fine collaborations and interesting discussions over the years, namely Dr. Kit Yan Chan, Mr. Chiong Ching Lai, and Mr. Renato Nakagawa. I would like to extend my deepest appreciation to my proof-readers in their painstaking efforts to polish this thesis.

In addition, I would like to acknowledge the Australian government for the generous research scholarships that they have granted to me, namely the CIPRS and subsequently the APA. My deep gratitude also goes to Curtin University for

providing the research support and the student travel grant, and to Sensear Pty Ltd for providing the testing equipment and environment.

Finally, I would like to thank my family and friends. To my wonderful parents, Yong Kwong and Tie Hung Kiong, my younger sisters Yong Ai Shing and Yong Ai Ping, and my little brother Yong Pei Jie, thank you for loving me and believing in me. To my best friend, Wee Lih Lee, who pursued his dream together with me since we were young, till now the end of our PhD journey, I thank you for your companion and your encouragement, and I wish you all the best in the completion of your thesis. To the brothers and sisters in Immanuel Methodist Church, and to people whose life intersects with mine during this journey, I thank you for the laughter and sorrow that we spared together. They have shaped me to whom I am today.

Pei Chee Yong

October 2013

Perth, WA, Australia

Abstract

Hearing protection is essential to industries operating under extreme noise conditions. This thesis aims at developing efficient binaural speech enhancement for communication hearing protectors that is capable of giving wearer a perception of the surrounding sound field while providing an adequate amount of noise suppression. For this purpose, two binaural noise reduction frameworks are adopted. The first approach uses a differential microphone array (DMA) at each side of the ears to suppress the noise from behind, followed by a binaural gain function to attenuate excessive residual noise from surrounding. Another one is a binaural multi-channel Wiener filter (MWF). This work seeks improvement in both frameworks by exploiting alternative solutions for different blocks in the algorithms.

A sigmoid (SIG) function is investigated as an alternative gain function for real-time single-channel speech enhancement. Besides having tunable parameters for altering the slope and the mean of its curve, a key benefit of using this function is the ability to preserve more speech signal at high signal-to-noise ratio (SNR) level. The parameters of the SIG function is optimised by studying the relationship between the gain function and the *a posteriori* SNR estimate, and through minimising a cost function comprising two objective measures.

The mapping between the SIG gain function and the *a priori* SNR is then investigated. As the widely-used decision-directed (DD) *a priori* SNR estimate has a one-frame delay that leads to the degradation of speech quality, an *a priori* SNR estimator is proposed to overcome this delay. A modified sigmoid (MSIG) gain function is also proposed, with three parameters optimised to match conventional gain functions. Performance evaluation utilises an objective evaluation metric that measures the trade-off among the noise reduction, speech distortion

and musical noise in the enhanced signal. Results are compared with more objective measures and subjective listening tests.

Noise estimation is one of the most crucial part in any speech enhancement algorithm. Two novel single-channel noise estimation algorithms are proposed in this thesis. First method tracks noise variations with low computational complexity, and is robust to speech onsets. It computes the noise estimate by comparing the estimate with short-term noise and speech at every time frame, and updating it using an optimised step-size. The second approach is a speech presence probability (SPP)-based method. The SIG function, with slope and mean that can be adjusted independently, is used to provide better trade-off between noise overestimation and underestimation. Harder decisions based on conditional smoothing is then employed on top of the SIG function for better noise tracking capability.

This thesis also studies the MWF algorithm for speech enhancement, which suffers from performance degradation due to the lack of robustness against estimation errors of the second-order statistics. The reasons are twofold: (i) they rely on real voice activity detection (VAD), and (ii) they involve estimation of the second order clean speech statistics. A MWF formulation that requires neither VAD nor clean speech statistics is presented. The aforementioned single-channel framework is employed to obtain the clean speech estimate at a reference channel. A rank-one formulation is also included. Results show that the proposed method outperforms the conventional approach in speech quality.

Finally, the new MWF is extended to the binaural configuration for speech enhancement. The blocks in DMA with a binaural gain function (DMA-BPF) algorithm are also altered by tracking the noise estimates using two identical single-channel noise estimation algorithm, one at each side of the ear. Novel single-channel algorithm presented in this thesis is employed in the DMA-BPF for better speech quality performance. Results show that the proposed binaural MWF algorithm outperforms the conventional binaural MWF both in speech quality improvement and binaural cues preservation. The improved DMA-BPF can also preserve the binaural cues of the DMA outputs, and provide more noise reduction capability when compared to the reference methods.

Contents

List of Figures

List of Tables

List of Terms and Acronyms

Mathematical Symbols

$\lvert \cdot \rvert$	absolute value
$\approx$	approximately
$\lvert\lvert \cdot \rvert\rvert$	Euclidean norm
$\mathrm{e}^{[\cdot]}$	exponential function $\exp(\cdot)$
$\log[\cdot]$	natural logarithm
$\min[\cdot]$	minimum value
$E(\cdot)$	expected value operator
$P[\cdot]$	probability density function
$\Re\{\cdot\}$	real part of a complex variable
$\sum$	summation of
$(\cdot)^H$	conjugate transposition
$(\cdot)^*$	complex conjugation
$(\cdot)^T$	transposition
$\mathrm{Tr}[\cdot]$	trace of a matrix

$A(k,m)$	ATF in STFT domain
$\mathcal{A}(k,m)$	STFT of the clean speech magnitude
$\alpha(\mathrm{e}^{j\Omega},\phi_{sq})$	parameter modelling the ILDs for DMA
$\alpha(k,m)$	smoothing constant
$\alpha_v(k,m)$	smoothing constant for noise estimate
$\alpha_{vv}(k,m)$	smoothing constant for noise correlation matrix
$\alpha_y(k,m)$	smoothing constant for noisy speech estimate
$\alpha_{yy}(k,m)$	smoothing constant for speech plus noise correlation matrix
$A_{ql}(\mathrm{e}^{j\Omega},\phi_{sq})$	ATF for the q-th source and the l-th microphone
$B_k(k,m)$	speech intelligibility weight from ANSI SII standard
$D(k,m)$	binary decision from MDSVAD
d	distance between microphones
δ	step-size for SSC method
ϵ	noise floor employed in spectral gain function
f_s	sampling frequency
$G(k,m)$	single-channel spectral gain function
$\mathcal{G}(k,m)$	binaural gain function
$\gamma(k,m)$	*a posteriori* SNR
$\Gamma_{v_\mathrm{left}v_\mathrm{right}}(k,m)$	coherence between the noise contained in the left and right DMA output

$G_{\mathrm{MSIG}}(k,m)$	gain function for LSA estimator
$G_{\mathrm{MS}}(k,m)$	gain function for magnitude subtraction
$G_{\mathrm{MSIG}}(k,m)$	modified sigmoid gain function
$G_{\mathrm{PS}}(k,m)$	gain function for power subtraction
$G_{\mathrm{SIG}}(k,m)$	sigmoid gain function
$G_{\mathrm{SS}}(k,m)$	gain function for spectral subtraction
$G_{\mathrm{WF}}(k,m)$	gain function for Wiener filter
$\mathcal{H}_0(k,m)$	hypothesis for speech absence
$\mathcal{H}_1(k,m)$	hypothesis for speech presence
$H(\mathrm{e}^{j\Omega},\phi_{sq})$	HRTF
$\hat{\xi}_{\mathrm{DD}}(k,m)$	*a priori* SNR estimate using DD approach
$\hat{\xi}_{\mathrm{MDD}}(k,m)$	*a priori* SNR estimate using MDD approach
K	number of frequency bins
k	frequency bin index
L	number of microphones
λ	trade-off parameter for proposed MWF
$\lambda_v(k,m)$	noise PSD
$\lambda_x(k,m)$	clean speech PSD
$\lambda_y(k,m)$	noisy speech PSD
M	number of frames
m	frame index

μ	trade-off parameter for SDW-MWF
N	frame duration in samples for STFT
n	discrete time index
Ω	normalized frequency
ϕ	direction of arrival of an acoustic wave signal
ϕ_s	angular position of the sound source
Φ_{ss}	power of the speech signal (single source case)
$\Phi_{v_i v_i}(k,m)$	PSD of the noise components for the left/right DMA output in the STFT domain
$\Phi_{v_{\text{left}} v_{\text{right}}}(k,m)$	cross-PSD of the noise components between the left and right DMA output in the STFT domain
$\Phi_{\hat{v}\hat{v}}(k,m)$	PSD of the joint noise estimate in STFT domain
$\Phi_{xx}(k,m)$	PSD of a speech signal in STFT domain
$\Phi_{y_{\text{left}} y_{\text{right}}}(k,m)$	cross-PSD between the left and right DMA output in the STFT domain
$p(k,m)$	*a posteriori* SPP
Q	number of point sources
q	point source index
R	frame shift in samples for STFT
$\mathbf{R}_v(k,m)$	noise correlation matrix
$\mathbf{R}_y(k,m)$	speech plus noise correlation matrix
$\mathbf{r}_{yx}$	cross-correlation vector

$\sigma_v^2(k,m)$	smoothed true noise PSD
$S_q(\mathrm{e}^{j\Omega})$	frequency response of the q-th source signal
τ_{DMA}	delay applied in a DMA
$\tau(\phi_{sq})$	parameter modelling the ITDs
$t_i(k,m)$	different time averaging constant for i
$t_v(k,m)$	time averaging constant for noise estimate
$t_y(k,m)$	time averaging constant for noisy speech estimate
$\Upsilon(k)$	ratio between the left and right HRTF
$V(k,m)$	STFT of the noise
$v(n)$	noise in discrete-time domain
ϖ	golden ratio for GSS
ϱ	narrowband output SNR for MWF
$V_{\mathrm{DMA}}(k,m)$	STFT of the DMA output noise
$Y_l(k,m)$	STFT of the speech component in l-th microphone signal
$V_l(\mathrm{e}^{j\Omega})$	frequency response of the noise component at the l-th microphone signal
$\mathbf{w}(k,m)$	complex weight vectors
$w_a(n)$	analysis window function
$W_l(\mathrm{e}^{j\Omega})$	frequency response of the filter applied to l-th channel
$w(n)$	window function

$w_s(n)$	synthesis window function
$X(k,m)$	STFT of the clean speech signal
$x(n)$	clean speech signal in discrete-time domain
$X(\mathrm{e}^{j\Omega})$	frequency response of the microphone signal
$\hat{x}(n)$	clean speech signal estimate
$\xi(k,m)$	*a priori* SNR
$Y_l(k,m)$	STFT of the speech component in l-th microphone signal
$X_l(\mathrm{e}^{j\Omega})$	frequency response of the speech component at the l-th microphone signal
$\mathbf{x}(m)$	vector of clean speech frame
$Y(k,m)$	STFT of the noisy speech signal
$y(n)$	noisy speech signal in discrete-time domain
$Y_{\mathrm{DMA}}(k,m)$	STFT of the DMA output signal
$Y_l(k,m)$	STFT of the l-th microphone signal
$X_l(\mathrm{e}^{j\Omega})$	frequency response of the l-th microphone signal
$Z(\mathrm{e}^{j\Omega})$	frequency response of the output signal

Acronyms

AMS	analysis-modification-synthesis
ANC	adaptive noise canceller

ATF	acoustic transfer function
BSS	blind signal separation
CDF	cumulative distribution function
dB	decibel
dB(A)	decibel A-weighting
DD	decision-directed
DFT	discrete Fourier transform
DMA	differential microphone array
DMA-BPF	DMA with a binaural gain function
DoA	direction of arrival
DSP	digital signal processing
GerkSPP	Gerkmann's SPP
GSC	generalized sidelobe canceller
GSS	golden section search
HenMMSE	Hendriks' MMSE
HPD	hearing protection device
HRTF	head-related transfer function
IDFT	inverse DFT
IFWSNRseg	speech intelligibility weighted segmental SNR in frequency domain
ILD	interaural level difference

IMCRA improved minima controlled recursive averaging

ITD interaural time difference

KurtR kurtosis ratio

LCMV linearly constrained minimum variance

LLR log-likelihood ratio

LMS least mean square

LogErr symmetric logarithmic-error distortion measure

LSA log spectral amplitude

MAP maximum a posteriori

MDD modified decision-directed

ML maximum likelihood

MMSE minimum mean square error

MS minimum statistics

MSE mean square error

MSIG modified sigmoid

MVDR minimum variance distortionless response

MWF multi-channel Wiener filter

NATTseg segmental noise attenuation

NRR noise reduction ratio

ONIHL occupational noise-induced hearing loss

PDF probability density function

PESQ	perceptual evaluation of speech quality
PSD	power spectral density
SDW-MWF	speech distortion weighted MWF
SIG	sigmoid
SNR	signal-to-noise ratio
SNRseg	segmental SNR
SPL	sound pressure level
SPP	speech presence probability
SPREseg	segmental speech preservation
SS	spectral subtraction
SSC	step-size controlled
STFT	short-time Fourier transform
SVAD	soft VAD
VAD	voice activity detection
WF	Wiener filter
WGN	white Gaussian noise
WHO	World Health Organization

Chapter 1

Introduction

Motivation is what gets you started.
Habit is what keeps you going.
– Jim Rohn

1.1 Introduction

Voice communication is the most fundamental form of human communication in delivering messages in everyday lives, whether it is face-to-face speech communication or speech communication through electronic devices. However, noise is omnipresent and this impacts speech communication in terms of speech quality and intelligibility. In many industrial environments, such as mine sites and oil-and-gas industry, workers have to stay in extreme noise conditions with sound pressure levels exceeding 100 decibel A-weighting (dB(A)) for approximately 8 hours daily [1]. In Australia alone, an estimated 1 million employees may be potentially exposed to hazardous levels of noise at work [2]. Under such environments, speech communication could cease, or only be made with raised volumes. In addition to that, long exposure to high noise levels on a daily basis not only causes permanent damage to the hearing [3], but also serious injury or death as safety could be compromised. As a result, there is legislation in many countries, which requires hearing protection devices to be worn in noisy industrial environments where the noise level exceeds 85 dB(A).

Conventional hearing protectors, however, not only suppress noise, but also

attenuate the desired speech. This leads to a situation where many workers refuse to wear hearing protection and shout into each other's ears because they would rather bear the health risks than sacrifice the ability to communicate with co-workers. As a consequence, providing intelligible speech communication capability in hearing protection under extreme noise environments is of great importance. The ongoing progress in the development of smaller, more efficient, more powerful and less cosy electronics makes it possible to include digital speech enhancement systems in hearing protectors. Commercial works had started with the implementation of active noise reduction techniques into communication headsets to reduce more background noise so that conversation can be made. The drawback of such techniques is that they also reduce the users' ability to hear speech and remain aware of the surroundings. In fact, there is an inherent flaw in the application of those techniques in extremely noisy environments where hearing protection is required in the first place.

A more advanced type of digital hearing protectors, named assistive listening devices, aims to suppress all unwanted noise components while preserving the target speech. These devices first capture the surrounding acoustic environment with the use of two or more microphones embedded into the hearing protectors to process the speech and noise through speech enhancement techniques. The enhanced signals are then delivered to the ears via two loudspeakers embedded in the earmuffs or earplugs. An important aspect of this type of noise suppression is the ability to maintain the spatial cues of the acoustic scene. That means the enhanced signals should have maintained the same spatial characteristics as received by the ears without hearing protection. For this reason, the microphones should be placed close to the ears such that similar shadowing effects can be obtained.

Besides the placement of the microphones, different signal processing techniques can also alter the spatial impression of the acoustic environment. In the early stages of assistive listening, particularly in hearing aids, the microphone signals on each side were processed independently of each other. This type of processing is referred to as bilateral processing and may distort the so-called spatial cues, i.e., interaural time and level differences, which are necessary for

a correct localisation of sound sources in the horizontal plane. Distorted cues may, however, cause a mismatch between the visual and auditive perception of the environment and possibly result in an abnormal impression. In order to account for the binaural aspect of the auditory system, modern systems exploit the microphone signals of both sides for binaural processing.

Much research effort has been put into the improvement of binaural noise suppression systems for hearing aids, which help impaired people to hear better and understand what other people are saying in everyday life [4]. In this context, scenarios with multi-talker, or known as the cocktail party effects, are frequently considered, where many people are simultaneously talking and the target signal is to be extracted out of mixed speech signals. In the scope of this work, the conditions are different and the desired speech signal is corrupted by industrial noise. The goal of this thesis is to design better speech enhancement techniques to extract the desired target speech and suppress high level background industrial noise. Although there has been many existing binaural noise-reduction algorithms reported in the literature, most of them are rather sophisticated and require complex architectures, which turn out to be prohibitive for binaural hearing protection devices. Therefore, the design of the algorithms also aims to take care of two crucial aspects in the processing, namely the computational complexity and the latency. A lower computational complexity implies less power consumption and thus longer battery life, which enables the usage of the hearing protectors for long working hours without recharging the battery. Whereas, a minimum latency is required to avoid the delivery of unpleasant sounds to the ear.

Hence, this dissertation aims at utilising promising methods which have the potential to be implemented in a binaural hearing protection devices, and proposes solutions that aim at reducing the computational complexity of a speech enhancement framework. Likewise, investigation is carried out to evaluate two efficient binaural noise suppression frameworks for industrial noise scenarios. Novel contributions have been proposed to improve both of these binaural processing techniques in terms of both speech quality and intelligibility by comprising blocks of multi-channel and single-channel speech enhancement algorithms. The motivation is that although multi-channel methods often lead to better performance

than single-microphone methods, the usability is limited by additional costs such as power usage, computational complexity, and size demands. On the other hand, single-channel algorithms can improve quality aspects of the signals (in terms of SNR), which helps to increase the comfort and reduce listeners fatigue.

1.2 Objective

The objective of this thesis is to investigate and provide novel single-channel and multi-channel solutions for binaural hearing protectors to enhance speech signals in noisy environments and to protect the cues of spatial hearing. Desirably, the end-product should be capable of introducing as less speech distortion as possible whilst reducing as much ambient noise as desirable. This requires the hearing protection device to be a cross-fertilisation between the two distinct types of techniques (single-channel and multi-channel speech enhancement techniques). Also, the new structures should be able to track and estimate the surrounding noise continuously. Ideally, in order to build an implementable binaural speech enhancement framework for binaural hearing protection devices, it should have the following traits

- use only a few microphones (e.g., 2 to 4),
- no knowledge of the geometry of the array,
- no knowledge of the physical location of the target signal,
- no assumptions about the distribution of the target signal,
- no need for a VAD, i.e., no additional algorithm to detect speech active or speech inactive periods, and
- low computational complexity and short processing delays.

1.3 Contribution of the Thesis

The work has mainly focused on designing speech enhancement frameworks that are feasible for real-time implementation. The original contributions of this dissertation include

- Identified the problems in conventional approaches and investigated a flexible spectral weighting gain function, the sigmoid function for single-channel speech enhancement framework [5]. The parameters of the gain function are determined based on the mapping with the SNR estimate, and a cost function that comprises two speech quality objective measures.

- Proposed a modified sigmoid function and an SNR estimate for single-channel speech enhancement [6, 7]. The former provides a better function that can be optimised to match the conventional gain curves to deal with the trade-off among speech distortion, noise reduction and musical noise; while the latter reduces the speech transient distortion that occurs in the conventional decision-directed approach. A kurtosis measure is introduced to evaluate the amount of musical noise generated, and is used together with other objective measures to quantify the overall performance of the proposed algorithms.

- Two noise PSD estimation methods are proposed, namely (i) the SSC noise PSD estimator [8], and (ii) the soft VAD (SVAD) noise PSD estimator [9]. The former offers an estimate that is robust to outliers (speech) with very low computational complexity relative to the conventional methods, while the latter offers faster noise tracking capability and better speech quality when compared to reference algorithms.

- A multi-channel Wiener filter solution is proposed to deal with the non-stationary of speech and noise, which then improves the speech quality performance [10]. The solution avoids the poor estimation of clean speech correlation matrix by incorporating a single-channel framework in the estimation of the clean speech signal in the reference channel, which involves a noise power spectral estimation method utilising a conditional speech presence probability. The conditional speech presence probability is then used to estimate the second order statistics of noisy speech and noise.

- Extended the proposed algorithms into binaural configurations for speech enhancement. Two frameworks are considered, both comprising blocks of single-channel and multi-channel algorithms. The most important finding

is that the proposed multi-channel Wiener filter is capable of preserving the speech and noise cues based on objective measurement.

1.4 Thesis Outline

The focus of this thesis is speech enhancement in hearing protection device (HPD) and the flow of this thesis is intended to resemble the author's research and development progress. Each chapter provides a stepping stone for the subsequent chapter or complements one another. The thesis is organised as follows.

Chapter 2 gives a background on the impact of long exposure to extremely loud noise and the importance of having speech enhancement system in a hearing protection device. Several single-channel and multi-channel speech enhancement techniques are reviewed and discussed.

Chapter 3 gives a brief background on the derivation of optimal gain function in the literature, and shows that a sigmoid function can be utilised as a gain function with its built in flexibility. Instead of trying to optimise the gain in mathematical sense, several objective evaluation tools for speech quality are utilised to find optimal parameters for sigmoid function.

Chapter 4 proposes a modified sigmoid function with a new *a priori* SNR estimate for speech enhancement. The sigmoid function can be flexibly fitted to the conventional gain functions, with two smoothing parameters from the *a priori* SNR estimate. Optimal parameters for modified sigmoid function are obtained to improve speech quality in terms of trade-off among noise reduction, speech distortion and musical noise.

Chapter 5 introduces two novel estimation algorithms for noise power spectrum. The first one utilises a step-size control mechanism to update the noise PSD estimate based on the difference between the estimate and the noisy observation at the previous frame. The second method is a soft voice activity detection based algorithm that updates the noise PSD estimate by employing conditional smoothing towards a sigmoid function that offers flexibility in controlling the amount of noise power overestimation and underestimation.

Chapter 6 proposes a new MWF that avoids the estimation of speech correlation matrix and employs a single-channel speech enhancement algorithm to estimate the desired signals. The conditional SPP is employed to estimate the noise PSD in the single-channel gain function and to update the second-order statistics. The proposed method outperforms the reference method in terms of noise reduction and speech quality.

Chapter 7 revises and investigates two possible binaural noise reduction techniques to be implemented in hearing protectors. The problems of each approach are addressed, and partly tackled with proposed solutions presented in previous chapters, which include the noise PSD estimation algorithm in Chapter 3, the proposed single-channel speech enhancement algorithm in Chapter 5, and the multi-channel speech enhancement framework in Chapter 6. It has been concluded that the proposed methods can improve the performance in terms of noise reduction while maintaining the spatial awareness of both speech and noise.

Chapter 8 gives conclusions and recommendations for future work.

1.5 List of Peer-Reviewed Publications

The following papers are published in conjunction with this thesis.

I P. C. Yong, S. Nordholm, H. H. Dam, and S. Y. Low, "On the optimization of sigmoid function for speech enhancement," in *Proc. 19th European Signal Process. Conference (EUSIPCO'11)*, Barcelona, Spain, Aug. 2011, pp. 211-215.

II P. C. Yong, S. Nordholm, and H. H. Dam, "Noise estimation with low complexity for speech enhancement," in *Proc. IEEE Workshop Applications of Signal Process. to Audio and Acoust. (WASPAA'11)*, New Paltz, USA, Oct. pp. 109-112.

III P. C. Yong, S. Nordholm, and H. H. Dam, "Trade-off evaluation for speech enhancement algorithms with respect to the a priori SNR estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'12).* Kyoto, Japan, Mar. 2012, pp. 4657-4660.

IV P. C. Yong, S. Nordholm, and H. H. Dam, "Noise estimation based on soft decisions and conditional smoothing for speech enhancement," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC'12)*, Aachen, Germany, Sep. 2012, pp. 4640-4643.

V P. C. Yong, S. Nordholm, and H. H. Dam, "Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement," *Speech Communication*, vol. 55, no. 2, pp. 358-376, Feb. 2013.

VI P. C. Yong, S. Nordholm, H. H. Dam, Y. H. Leung, and C. C. Lai, "Incorporating multi-channel Wiener filter with single-channel speech enhancement algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'13)*. Vancouver, Canada, May. 2013.

VII K. Y. Chan, P. C. Yong, S. Nordholm, C. K. F. Yiu, and H. K. Lam, "A hybrid noise suppression filter for accuracy enhancement of commercial speech recognizers in varying noisy conditions," *Applied Soft Computing*, 2013.

VIII P. C. Yong, S. Nordholm, and H. H. Dam, "Effective binaural multi-channel processing algorithm for improved environmental presence," *Submitted to IEEE Trans. on Audio, Speech, and Language Process.*

Chapter 2

Background Literature

If the only tool you have is a hammer,
you tend to see every problem
as a nail.
– Abraham Maslow

2.1 Introduction

Speech enhancement in adverse acoustic environments with regards hearing aids remains an area of intensive research over decades [4, 11–17]. However, only little attention has been paid in designing speech enhancement algorithms for hearing protection devices, which is a crucial application to avoid hearing impairment and to enable speech communication among workers. Recall from Chapter 1 that such devices have to be able to suppress high level industrial noise while maintaining the target speech signal and preserving the spatial cues of both the target speech and background noise. In addition, there are two additional important elements to be associated with the design criteria, namely the low latency and low computational complexity.

There are largely two approaches to address speech enhancement, i.e., single-channel and multichannel approaches. On one hand, the single-channel method is essentially easier to be implemented in hardware with less computational effort required. Multi-channel algorithms on the other hand, offers invaluable spatial diversity which provides further improvement in speech quality. Both of

the two solutions, however, have their own fundamental limitations. In general, single-channel approaches tend to generate noise artefacts while multi-channel algorithms cannot suppress as much noise compared to the single-channel methods.

This chapter investigates the need of assistive listening devices in industrial environment and reviews state-of-the-art single and multiple channel solutions for speech enhancement. It begins by introducing noise and noise induced hearing loss, and followed by the review of the hearing protectors. Next, the human spatial hearing system is briefly described before moving into the discussion of the noise reduction techniques. The style of writing is intended to give the readers a quick idea of the problem and compendious review of conventional speech enhancement techniques and how they relate to the eventual proposed structures.

2.2 Noise and Hearing Loss

Noise is often defined simply as unwanted sound, which is however misleading as any loud sound, whether wanted or not, can damage hearing. The relationship between noise and hearing loss has been the focus of numerous studies since 1950s. Conclusive evidence for greater hearing loss was found between workers exposed to elevated noise levels compared to workers in quieter jobs [18]. An Estimation from the 2000 Global Burden of Disease study has stated that worldwide 70% of mild or greater hearing loss, 75% of moderate or greater hearing loss and 87% of severe or greater hearing loss is adult-onset [19], with the two major causes of adult-onset hearing loss are ageing and exposure to loud noise [20]. Age-related hearing loss is mostly the damage to the hearing-related structures and nerves that occurred from various sources over time rather than from biological deterioration alone [20]. On the other hand, exposure to excessive levels of noise (i.e., levels considered hazardous to the hearing of most people, see Table 2.1 for examples of hazardous noise levels) affects hearing by changing the physiology of the inner ear, which can lead to severe hearing impairment, and even complete deafness.

The term deafness can sometimes be confusing, as in some places it is only used to describe those who are totally deaf while in others it also includes those

Painful
150 dB = jet take-off (at 25 meters)
140 dB = fire arms, air raid siren, jet engine
130 dB = jack-hammer, motorcycles
120 dB = chain saw, oxygen torch
Extremely Loud
110 dB = rock music, model aeroplane, steel mill
100 dB = pneumatic drill, farm tractor
90 dB = lawnmower, shop tools, truck traffic, subway
Very Loud
80 dB = heavy city traffic, busy street, average factory
70 dB = busy traffic, vacuum cleaner
60 dB = conversation in restaurant, office, dishwasher
Moderate
50 dB = moderate rainfall
45 dB = humming of a refrigerator
40 dB = quiet room
30 dB = whisper, quiet library

Table 2.1: Hazardous noise levels and examples.

who experience difficulty in hearing [21]. As a consequence, the grades of hearing impairment often cannot be compared directly across studies. World Health Organization (WHO) defines permanent hearing impairment in adults as having difficulty in listening at hearing threshold level of 41 decibel (dB) or higher (refer to Table 2.2), based on the un-aided hearing threshold in the better ear and is averaged over the 0.5, 1, 2, and 4 kHz frequencies. A hearing threshold level of 41 to 60 dB is considered as the beginning of hearing impairment because at this level of impairment, an individual can only distinguish words spoken at one metre and only if they are spoken in a raised voice [22]. This is also the level of impairment where hearing aids are usually required by the individual [23].

Depending on the intensity of the noise and the duration of exposure, either exposure to very loud noise for a short period of time or repeated exposure to moderately loud noise, noise-induced hearing loss can begin immediately or gradually and may be temporary or permanent. Such hearing loss may affect one or both ears, although not always to the same extent [24]. The first sign of noise-induced hearing loss is typically a shift in the pure-tone threshold in 3-6 kHz

Degree of hearing loss	Hearing loss range (dB)	Effect
None	≤ 25	can hear whispers
Slight	26 to 40	can hear words at 1m in normal voice
Moderate	41 to 60	can hear words at 1m in raised voice
Severe	61 to 80	can hear words if shouted into ear
Profound	≥ 80	cannot hear shouted words

Table 2.2: WHO grades of hearing impairment.

frequencies. That means a notably louder tone than previous is required for an individual to detect a tone at the range of the frequencies [19]. Threshold shifts in these frequencies usually indicate a hearing impairment in the upper part of frequency range for human voices [25]. Once the individual workers slowly lose their ability to hear voices, they would suffer from the inability to communicate with co-workers under those extremely noisy circumstances. Research in Sweden has shown that in a noisy environment, only 40% of a conversation can be heard by someone with moderate noise-induced hearing loss when compared to 75% by someone with unimpaired hearing [26]. Whereas 90% of a conversation in a quiet environment can be heard by a moderate hearing-impaired person compared with 98% by another with unimpaired hearing [26]. However, even if a hearing impaired person can find another job in less noisy workplaces, noise is ubiquitous thus exposure to noise is always unavoidable.

Exposure to noise in the workplace (occupational noise) has been estimated to account for about 10% of the burden of adult hearing loss in Western countries, and is well-known as occupational noise-induced hearing loss (ONIHL) [20, 25]. Loss of hearing caused by ONIHL is a significant problem in terms of not only health but economically. As for Australia, there were about 16,500 successful compensation claims from industrial workers for permanent impairment due to noise from the year 2002 to 2007 [19]. This means not only individual workers but all their families, business owners, the whole industry and the wider society have to bear the burden of ONIHL. Besides the harmful effects on hearing from

the exposure to loud noise, it also leads to annoyance and fatigue in the individuals, which could result in serious health conditions such as hypertension and heart disease [19]. The ONIHL and many of its effects can be prevented often by proper workplace design, with sufficient equipment and training provided to control occupational noise levels and workers exposures. A good solution to reduce the ONIHL is to protect the ears from exposure to intense environmental noise by using hearing protection devices (HPDs).

2.3 Hearing Protection

The engineering of conventional hearing protectors is considered a relatively mature technology that has been continuously evolving over the past one and a half centuries. They include designs with frequency-independent attenuation, with attenuation that increases with sound pressure level (SPL) to provide protection against impulsive sounds, with attenuation that is specifically designed to the application, and with earplugs that can be fitted to individual ear canals [27]. A common drawback of those passive HPDs is that they suppress both noise and speech, which prohibit the face-to-face communication under extremely noisy environments such as in factories. This leads to the lack of situational awareness resulting in serious injury and death. In order to communicate with coworkers under such conditions, people would choose not to use hearing protection and expose themselves back to the danger of ONIHL.

Of more recent development is the incorporation of low cost with high performance digital signal processors to allow speech communication and/or situational awareness in extreme noise [28]. One way to do this is the use of electronic components that amplify the sounds reaching the ear inside the hearing protectors, particularly earmuffs (which can provide overprotection from environmental noise with passive attenuation), up to specified limits. Although such level-dependent HPDs can improve the user's audibility of speech and performance at work [29–31], they are still unable to restore all dimensions of situational awareness compared to human hearing. Another signal processing strategy for HPDs is to employ an active noise reduction system to detect the surrounding

sounds with microphones attached to the earmuff or earplug, then invert the sounds in phase, and broadcast them into the ears using loudspeakers [32]. Such technique can be categorised into two methods, namely feedforward (a microphone is placed outside each earcup) and feedback (microphones are placed inside earcups) noise cancellations [33]. Although the HPDs incorporating the active noise reduction techniques can provide additional noise reduction at frequencies from approximately 32 Hz to 500 Hz compared to passive hearing protectors, acoustical constraints make it difficult to achieve better noise reduction at higher frequencies. Furthermore, they also suppress speech together with noise, and impede source localisation due to the directional characteristics and positions of the microphones [34].

This thesis exploits the possibility of employing speech enhancement in hearing protectors to reduce as much noise as possible while maintaining the speech signals. By isolating and enhancing speech while reducing harmful background noise, users are able to hear speech and stay protected whilst remaining aware of their surrounding in high noise environments. Such technology aims at adapting quickly to any changes in background noise, so that users can communicate effectively with others under every kind of scenarios, without the need to take on and off their HPDs. This type of electronic HPDs can be termed as assistive listening devices, as shown in Figure 2.1 [1]. The focus point of such devices is to develop speech enhancement algorithms which can attenuate as much noise power as possible, while leaving speech as undistorted as possible. By starting from the spatial cues, new techniques can be investigated and developed to provide very little alteration of spatial information. This shall allow a high noise suppression while still providing the desirable spatial and low distortion properties of the desirable information.

[1]The devices shown in the figure are developed and manufactured by Sensear Pty. Ltd. The company's website can be found at http://www.sensear.com/.

Figure 2.1: Assistive listening devices.

2.4 Human Spatial Hearing

This section discusses the different localisation cues utilised by human auditory system to localise sound sources in three dimensions. Such understanding is important when building binaural HPD that is capable of maintaining situation awareness in extremely noisy environment.

There are three distinct perceptual effects that support the fact that an improvement of speech intelligibility in noise can be achieved by listening with two ears rather than with one. The first being the *Binaural summation* effect, which states that one can identify more easily the target speech component when two identical mixtures of speech and noise (with same input signal-to-noise ratios (SNRs)) are presented to the two ears. The equivalent SNR benefit by the binaural summation effect is in the range of 0.5 to 2 dB [35]. Secondly, the *head shadow* effect, which is purely a physical effect caused by the direction of the speech and noise sources with respective to the head. In many real scenarios the speech source would be closer to one ear than to the other, creating a "best-ear" side where the SNR is higher (up to 10 dB have been recorded [35]). Due to this head shadow effect, the listener can use the best-ear to listen the presented speech. By the definition of the third effect, namely the *binaural unmasking*

effect, the other ear (with lower SNR) can lead to further speech intelligibility improvements, which is purely caused by the spatial processing of the binaural cues inside human ears [35].

The primary cues that influence the spatial sound perception by people include the following auditory cues [36]

1. the interaural time difference (ITD),
2. the interaural level difference (ILD),
3. monaural spectral cues depending on the shape of the outer ear (pinna),
4. cues from torso reflection and diffraction,
5. the ratio of direct to reverberant energy,
6. cue changes by voluntary head movements, and
7. familiarity with the sound source,

as well as visual cues and other non-auditory cues [37]. It should be noted that except for source familiarity, all the auditory cues come from the physics of sound propagation and vary with azimuth, elevation, range, and frequency [36]. Although all of the cues should be present and consistent for optimum sound reproduction, some of these cues are stronger than the others, particularly the ITD and ILD cues, typically also known as the binaural cues. In this thesis, we will only consider these two binaural cues as they are the main cues responsible for localisation in the azimuthal (horizontal) plane, which is crucial in the workplaces compared to other cues.

The binaural cues are obtained from a sound wave impinging on the two ears from a certain direction, with a particular time and intensity difference between the two ears, which are independent of the source spectrum (see Figure 2.2 [38]) [39]. The well-known duplex theory by Lord Rayleigh [40] states that ITD cues are the dominant localisation cues in the lower frequencies ($< 1 - 1.5$ kHz), while ILD cues are dominant in the higher frequencies. The fact that ITD cues are only usable in the lower frequencies is based on the observation that the wavelength of the sound signal becomes smaller than the diameter of the head in the range

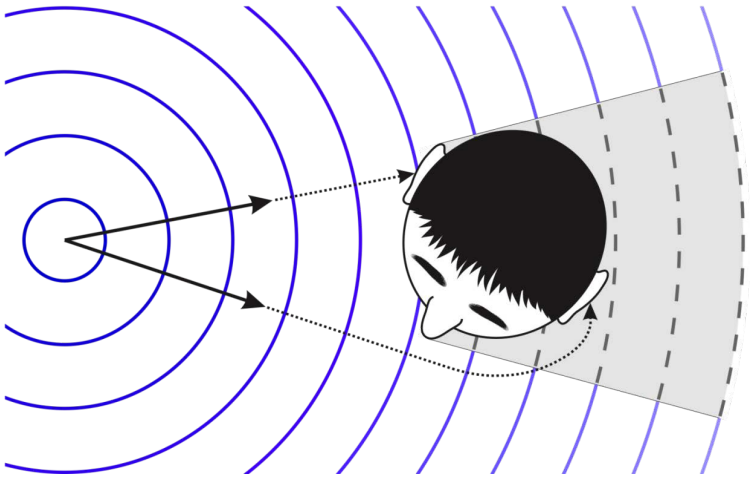

Figure 2.2: Sound wave arriving at human ears, generating ITD and ILD cues.

of 1-1.5 kHz. The ILDs do occur over the entire frequency range, but usually have a large magnitude only in the higher frequency range. However, if ILDs are artificially introduced at lower frequencies, they can introduce a spatial percept, while artificially introducing ITDs at higher frequencies do not lead to a spatial percept [41]. The intention of this dissertation is not artificially creating binaural cues, but to investigate methods to preserve both ITD and ILD when binaural speech enhancement algorithms are employed in assistive listening devices.

2.5 Noise Reduction Techniques

Binaural signal processing for speech enhancement, to be discussed in Chapter 7, comprises two types of noise reduction techniques, namely the single-microphone noise reduction and the multi-microphone noise reduction. Some of the state-of-the-art techniques will be discussed in the following sub-sections. As many noise reduction techniques rely on a voice activity detection (VAD) algorithm, which classifies audio frames as noise-only or speech plus noise frames. An overview of VAD algorithms is also presented.

2.5.1 Multi-channel Noise Reduction

Multi-channel noise reduction approaches combine and process input signals from different microphones in order to achieve an SNR improvement. The fact that different sources originate from different positions in space is hereby exploited: constructive interference is created in the direction of the target speech source, while destructive interference is created in other directions, which represents the direction of arrival (DoA) of the interfering noise sources. As such, multi-channel noise reduction techniques have the capability to spatially accept or reject sources at a specific point in space. This concept is referred to as beamforming, which was originally applied in antenna and radar [42], and then utilised for microphone array in applications such as mobile communications and hearing aids [12, 43–46].

In general, there are two types of multi-channel techniques, namely geometry dependent and geometry independent approaches. The former refers to conventional beamforming methods where some *a priori* information is required to form a beam towards the target signal. Beamforming itself can be broadly classified into fixed beamforming and adaptive beamforming. Fixed beamforming is data independent as a time-invariant filter is already pre-designed to capture the target signal. Hence, optimal fixed beamformers can be constructed provided that no model mismatch is present. As for adaptive beamforming, it is data-dependent as it combines spatial filtering with adaptive noise suppression capability to track variations as well as to compensate for model mismatch. Unlike the geometry dependent methods, geometry independent techniques do not require information about the array geometry or the source localisation. In the following subsections, some examples of both geometry dependent and geometry independent approaches are discussed.

Fixed Beamforming

Fixed beamforming is the most traditional geometry dependent multi-microphone approach for speech enhancement. One of the most popular fixed beamformers is the delay-and-sum beamformer [42], followed by other techniques such as the directional microphones and the super-directive beamformers [47]. As fixed beamformers are data-independent techniques, an assumption has to be made about

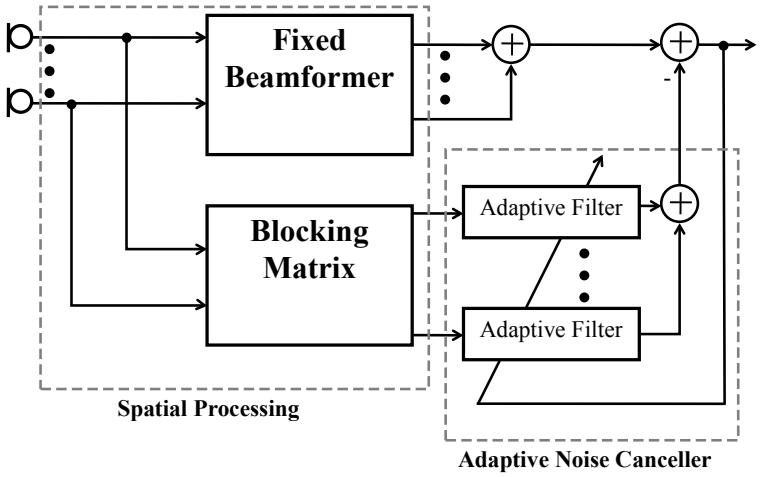

Figure 2.3: A general sidelobe canceller.

the target speech source location. Most often, it is assumed that the target speech source is in the frontal direction. Unfortunately, this assumption is not always fulfilled in some commonly encountered practical scenarios. In these cases, the direction of the target speech source may need to be estimated using source localisation algorithms. A disadvantage of fixed beamformers for higher resolution applications is the sensitivity to model imperfections such as microphone mismatch, caused by environmental influences or ageing of equipments. Also, a fixed beamformer usually requires a large number of sensors and thus it requires a high computational cost to achieve good performance.

Adaptive Beamforming

In contrast to fixed beamformers, adaptive beamformers are data-dependent techniques that can adapt to changing noise scenarios. Therefore, they are capable of giving a better noise reduction performance in certain scenarios. The most commonly used and computationally cheapest adaptive beamforming technique is the so-called adaptive directional microphone (ADM) [48]. The ADM consists of two fixed directional microphones, with one having a forward-facing spatial pattern and the other having a backward-facing spatial pattern, followed by an adaptive block. The adaptive part combines the two fixed spatial patterns so that a null is formed in the direction of the strongest noise interferer.

A general class of adaptive beamforming techniques has been developed to

minimise the total output power of the beamformer under the constraint that the response towards the target speech directions is preserved. The so-called minimum variance distortionless response (MVDR) beamformer [49] is the pioneering technique that constrains the response towards a single target direction, which is then extended by the linearly constrained minimum variance (LCMV) beamformer [50] to a set of linear constraints. A well-known implementation of these beamformers is the generalized sidelobe canceller (GSC) algorithm [51], which transforms the constrained optimisation problem into an equivalent but easier unconstrained optimisation problem. Here, the problem is transformed into two main parts, i.e., non-adaptive constrained and adaptive unconstrained components. As depicted in Figure 2.3, the non-adaptive constrained component is known as the spatial preprocessing stage that has two branches operating on orthogonal subspaces: a fixed beamfomer which passes the target signal (maintains the desired constraint) and generates a so-called speech reference, and a blocking matrix which blocks the target signal (nulls the desired constraint) and generates so-called noise references. The adaptive unconstrained component is an adaptive noise canceller (ANC) that cancels the residual noise from the fixed beamformer's output. It can be viewed as that the dimension of the adaptation subspace is reduced by the blocking matrix, which enables the use of simple adaptive algorithm such as the least mean square (LMS) algorithm.

Regrettably, GSC might also cancel out the target signal due to the presence of steering vector errors, which are caused by wrong modelling on the target speech location, room reverberation and microphone characteristics. The blocking matrix fails to block the target signal entirely and causes speech signal leakage into the ANC, which leads to the cancellation of the speech signal. As a result, the performance may degrade significantly [52]. Several extensions were proposed to mitigate the leakage problem [53–57]. One direction of the works was in improving the spatial preprocessing stage (fixed beamformer with the blocking matrix) by applying space constraints to protect the target signal's area [58–60]. For the noise cancelling stage, a way to tackle the problem is to make use of a VAD algorithm: the ANC is only activated in noise-only periods in order to avoid signal cancellation of the target speech signal [61]. Similar noise reduction

approach entails the use of a non-coherent approach, i.e., a single-channel speech enhancement technique as a post-filter for the beamformer [60].

Multi-channel Wiener Filter

One of the geometry independent classes of multi-microphone noise reduction techniques is based on multi-channel Wiener filtering (MWF) [59, 62–66], which is basically a generalisation of single-microphone techniques [67]. The MWF produces a minimum mean square error (MMSE) estimate of the speech component in a reference microphone signal, by exploiting speech and noise correlation matrices (second-order statistics). The formulation was extended to the speech distortion weighted MWF (SDW-MWF) to provide an explicit trade-off between speech distortion and noise reduction [64–66]. This extension is equivalent to applying additional single-microphone noise reduction to the output of a multi-microphone noise reduction algorithm. In contrast to the previously discussed adaptive beamformers, the SDW-MWF can adapt to both changing speech and changing noise scenarios. In principle, the SDW-MWF does not require *a priori* knowledge or assumptions about the target speech location and microphone characteristics, unlike the GSC. It is therefore expected to be more robust to the beamforming approaches, which was indeed presented in [52].

Blind Signal Separation

An alternative for the geometry independent multi-channel approach is the blind signal separation (BSS), which was originally suggested by Herault and Jutten in the mid-80s [68]. As the name suggests, BSS attempts to recover a set of unobserved signals from several observed mixtures with no *a priori* information about the array geometry and the localisation [69]. In speech processing context, BSS techniques consider the microphone signals as different observations of mixtures of audio signals, whereby each audio signal is a filtered version of a certain source signal. Under the assumptions that the sources are statistically independent, the number of microphones is at least equal to the number of sources, and at most one source is Gaussian distributed, the original source signals can be recovered (up to a scaling and permutation) using BSS algorithms such as Independent

Component Analysis (ICA) [70, 71].

The BSS techniques for acoustic signal separation can be divided into three classes, namely, higher-order-based BSS [71, 72], second-order-based BSS [73–75] and the latest time-frequency masking technique [76, 77]. These three approaches require different assumptions regarding the signal statistics. For example, higher-order-based BSS generally requires assumption about the density functions of the sources [78], while second-order based BSS, on the other hand, requires assumption about the second-order statistics such as nonstationarity or nonwhiteness [74]. In contract, the time-frequency masking technique exploits the sparseness in the time-frequency spectrum of the sound source to achieve source separation. It is known to be able to separate an arbitrary number of sources with just two anechoic mixtures provided there is not much overlap in the time-frequency representations of the sources, which is often true for speech signals [77]. The BSS problems can be solved in either frequency-domain or/and time-domain. For BSS in frequency-domain, the observed convolutive mixture per frequency bin can be approximated as an easier instantaneous mixture, so that the BSS problem is solvable by the standard BSS algorithm [71]. However, as each frequency bin is processed separately, the inherent BSS permutation and scaling ambiguities impose a problem, such that different frequency bins have to be re-aligned in a postprocessing stage [79]. For time-domain techniques, the BSS techniques considers the original convolutive BSS problem in the time domain [75], so that permutation and scaling ambiguities are not an issue. Even so, a disadvantage is that the algorithms in time-domain involve computationally expensive operations on large matrices.

Further works have been proposed to combine the knowledge in time domain and frequency domain for BSS. For frequency-domain BSS techniques, the problem was formulated as constrained optimisation problems where the time domain constraints on the unmixing matrices were added to ease the permutation effects associated with convolutive mixtures [80, 81]. For time-domain techniques, a computationally cheaper variant using block-online adaptation combined with frequency-domain fast convolution techniques was therefore proposed [82]. Nevertheless, as far as speech enhancement is concerned, BSS is not the most relevant

technique because all separated sources from all observations are of interest to BSS. However, in speech enhancement, there is usually only one source of interest in a noisy environment. As such, work have been done to transform BSS into a speech enhancement tool, i.e., BSS becomes blind signal extraction (BSE) which acts like a self-designed GSC beamformer [83]. However, when compared to the other afore-mentioned multi-channel approaches, the computation complexity of BSS makes this popular class of technique less appealing for a hearing protection device.

2.5.2 Single-channel Noise Reduction

Although multi-microphone methods often lead to better performance when compared to the single-microphone methods, their usability is often limited by the additional costs, usually for matched microphones, power usage, computational complexity, and size demands. Most of the time, only a single-microphone solution is preferred for wearable hearing devices. Although single-channel noise reduction algorithms usually provide very modest or hardly any improvement in intelligibility [84], they do improve quality aspects of the signals [85], which helps to increase the comfort and reduce listener's fatigue in extreme noise conditions.

As previously mentioned, single-channel noise reduction techniques are also important in the context of multi-microphone systems, because these can often be decomposed into a concatenation or a postfilter of a beamformer or a BSS algorithm. The typical aim of single-microphone method in these techniques is to reduce the remaining noise coming from the same direction as the target, and to adapt to the temporal changes of the statistics of the sound sources. As such, single-channel noise reduction systems play an important role in stand-alone systems or as a post-processing scheme for multi-microphone methods.

The focus of the following text is on the concise description of the various building blocks in the state-of-the-art speech enhancement system, as illustrated in Figure 2.4.

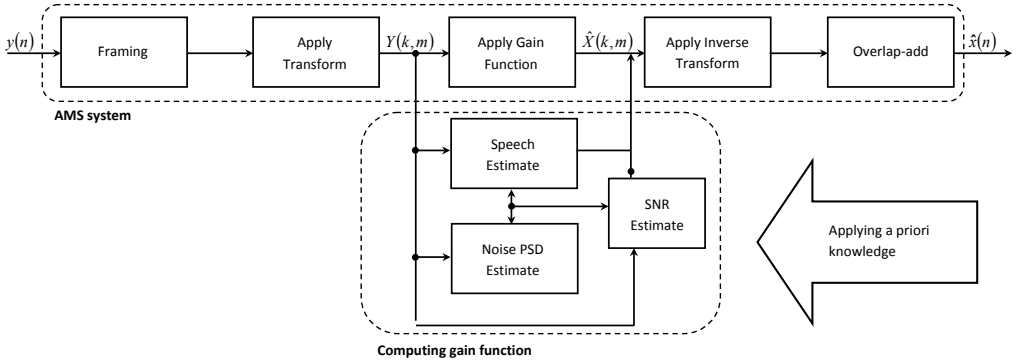

Figure 2.4: General single-channel speech enhancement system.

Speech Enhancement Framework

Figure 2.4 shows an analysis-modification-synthesis (AMS) system for speech enhancement. The system produces the estimate $\hat{x}(n)$ of the underlying clean speech signal $x(n)$ based on a noisy input speech signal $y(n)$. This begins by dividing the noisy input signal into generally overlapping signal frames within which underlying statistical properties can be assumed invariant. Then, the frames are windowed by an analysis window before a transformation is applied to the windowed frame to produce transform coefficients $Y(k, m)$ where k and m are the transform coefficient index and frame index, respectively. The noisy transform coefficients are then modified by applying a scalar gain to each coefficient independently before the enhanced transform coefficients are inversely transformed. The resulting time frames are overlap-added using a synthesis window to produce the enhanced noise reduced estimate $\hat{x}(n)$.

The transformation serves several purposes. First, it may act as a decorrelator to deliver noisy transform coefficients which are uncorrelated or even statistically independent. Most of the existing systems compute short-time Fourier transform (STFT) using a discrete Fourier transform (DFT), because the DFT can be implemented computationally efficient, which delivers approximately uncorrelated transform coefficients. This leads to an enhancement in the performance and a straightforward interpretation in terms of spectral signal content. The latter point is important because some systems are based on models of speech production [86] or based on auditory models [87]. In the general AMS system, the task

is to find the target speech signal estimate by building a spectral weighting gain function that contains a noise power spectral density (PSD) estimate and an SNR estimate.

Finding the Target Estimate

The critical point in the speech enhancement system is finding the gain values that lead to a suitably enhanced output signal with only the noisy input signal available at the microphone. Often, estimates of the target signal are derived under the assumption that target and noise are both present in the noisy observation. However, this is not always the case since speech signals have pauses between syllables and words. Furthermore, many speech transform coefficients are essentially zero, even during speech activity. Thus, the gain function has to be carefully designed taking into account these problems. Continuous noise PSD estimation and SNR estimation have to be performed in order to ensure that the distortion of target speech can be maintained as low as possible. All these blocks will be explicitly discussed in Chapters 3, 4, and 5.

A Priori Knowledge and Assumptions

In order to estimate a target signal from a noisy observation, one should have some knowledge about the signals, for example that noise is additive and independent of the target signal. The earlier proposed speech enhancement methods are based on these assumptions only. For example, the spectral subtraction approach [88, 89] is a method for finding an estimate of the target STFT. The spectral subtraction approach is computationally extremely simple, but is often not optimal in any sense, and is hardly ever used as a stand-alone algorithm, as it tends to produce an annoying musical-like residual noise.

Generally speaking, any available *a priori* knowledge could be used in the estimation process. For single-channel speech enhancement, this *a priori* knowledge comes in three different classes, namely, i) knowledge about the target signal, ii) knowledge about the noise signal, and iii) the knowledge (or assumption) that the enhanced signal is to be listened by a human. These classes of knowledge are not mutually exclusive; in fact, enhancement systems exist which utilise knowledge

from all three categories. However, while systems which rely heavily on a *a priori* knowledge or assumptions may perform very well when the assumptions are valid, they may perform poorly when the acoustical situation does not match the underlying assumptions. Thus, exploiting *a priori* knowledge involves a trade-off between optimum system performance and robustness to changing acoustical scenarios. In the context of this thesis, the focus is more on the stochastic processes and the actual characteristics of the observed signals instead of the *a priori* assumptions. As such, solutions with as less *a priori* information as possible are investigated in Chapters 3, 4, and 5. This is to build a robust single-channel speech enhancement system that is capable to perform similarly well in different adverse noisy environments.

2.5.3 Voice Activity Detection (VAD)

A VAD algorithm classifies audio frames as either noise-only or speech plus noise frames. The VAD makes a binary decision on a frame-by-frame basis, whereby frames are typically 20-40 ms. VAD algorithms are commonly used in both single and multi-channel noise reduction algorithms. In single-microphone noise reduction, where an estimate of the noise spectrum is often required, a VAD can be used so that the noise spectrum is updated during noise-only periods. For adaptive beamforming, it was noted that the ANC is usually only updated in noise-only periods to avoid speech cancellation, again by using a VAD. Finally, the multi-channel Wiener filter (MWF) technique requires a VAD to estimate the speech plus noise and noise-only correlation matrices (second-order statistics) in periods of speech plus noise and noise-only.

As VADs are so widely used, many algorithms have been proposed in literature [67]. In essence, VADs exploit the different properties and features of speech and noise, e.g., short-term energy, zero-crossing rate, speech and noise probability distributions, or combinations of properties. Many VAD algorithms are known to perform poorly if the noise is non-stationary, and at low input SNR levels. For single-microphone noise reduction, some techniques have therefore been developed to obtain an estimate of the noise spectrum without using a VAD. The

best-known techniques are probably the Minimum Statistics [90] and Minima-Controlled Recursive Averaging [91] techniques. These techniques could indeed provide good alternatives to VAD-based estimation, as the noise spectrum can be continuously updated and thus better tracked, compared to only updating during noise-only periods.

In this thesis, the use of VAD is avoided to ensure that more robust online speech enhancement algorithms can be developed. The ideas of developing computationally cheap noise spectrum estimation algorithms without using a VAD have been proposed in Chapter 5.

2.6 Summary

In this chapter, the background of the problems to be tackled in this thesis has been given. In a nutshell, HPDs are very important for individuals that work in extremely noisy environment. The traditional passive HPDs block out all sounds coming from the surrounding, including the desired speech. In this case, digital signal processing (DSP) techniques can be implemented into HPDs to allow speech communication to take place among workers. A speech enhancement framework can be considered as it can reduce the background noise while preserving the speech signal. It is also possible to have algorithms with low computational complexity in speech enhancement algorithms to ensure low battery consumption and alleviate the need to charge battery during working hours. An important aspect of the speech enhancement algorithm is to ensure that the spatial awareness can be preserved. The following chapters focus on proposing new methods for speech enhancement algorithms, which aim at developing efficient binaural frameworks in Chapter 7 that are capable of reducing as much noise as possible without introducing speech distortion and without distorting the spatial cues. In the next chapter, a new single-channel speech enhancement algorithm without requiring *a priori* information will be proposed.

Chapter 3

Speech Enhancement using A Posteriori SNR and a Sigmoid Function

Science progresses best when observations force us
to alter our preconceptions.
– Vera Rubin

3.1 Introduction

It is well known that the single-channel speech enhancement solutions have a *classic trade-off* between signal-to-noise ratio (SNR) and speech distortion [67]. Moreover, SS based algorithms are prone to generate speech artifacts commonly known as musical tones, a phenomenon due to errors in noise statistics estimation [67]. The challenge in noise estimation is to control the update so it is not affected by the speech. Consequently, when speech is coming into the noise estimate, it will be biased. One of the solutions for noise estimation is to employ voice activity detection (VAD) based algorithms [92]. However, VAD algorithms often miss-detect speech onsets at low SNR and cause the noise estimate to be affected by the speech energy [91]. There are a multitude of methods suggested to control of noise update [90, 91, 93]. All of them can be employed in this work but we assume an ideal estimation in order to highlight the work in this study.

A main task for speech enhancement techniques to be deemed practical is the mapping mechanism of the SNR measure to the gain function that is applied on the input data. The *a posteriori* SNR and the *a priori* SNR, which contain the estimation of speech and noise probability density functions (PDFs), are the common SNR measures used for this purpose. Since temporal averaging is often needed for both speech and noise estimates, this would change the distribution of the SNR estimates. Therefore, a more flexible gain function is required so that it can be mapped effectively to the SNR estimate to provide a good trade-off among noise reduction, speech distortion and musical noise. Instead of using the *a priori* SNR as in [94, 95], we propose to use the *a posteriori* SNR estimate in this chapter, as it provides an efficient way to optimise the gain function as well as the noise floor.

The use of the sigmoid (SIG) function for speech enhancement has been proposed in [95]. The study showed that SIG function has benefits for hearing impaired people. A more comprehensive description of the use of the SIG function for speech enhancement is found in [94]. Even though both [94, 95] use the *a priori* SNR estimation in the gain function, they did not provide a clear picture on how the mean and the slope should be estimated. Sigmoid functions naturally maps the SNR estimate into a gain function between zero and one. Thus, we propose to investigate the optimisation of the parameters in SIG function in order to have a full use of the gain function and its applicability over a wide range of scenarios. More specifically, SIG function is optimised based on the perceptual evaluation of speech quality (PESQ) measure and the log-likelihood ratio (LLR) measure. Both of these measures correlates well with subjective listening evaluations when compared to other objective measures [96]. This has also been verified in subjective evaluations.

The contributions in this chapter include the direct use of the *a posteriori* SNR in the SIG function and the establishment of the relationship between the SNR estimate and the gain functions. This study has direct impact for other speech enhancement techniques and gives a framework for finding new and improved enhancement functions.

3.2 Conventional Spectral Gain Function

3.2.1 Signal Model

For single-channel applications, the corrupted speech sequence can be represented by an additive observation model $y(n) = x(n) + v(n)$, where $y(n)$ represents the observed signal at discrete-time index n, $x(n)$ is the clean speech signal and $v(n)$ is the additive random noise, uncorrelated with the clean signal. The goal of speech enhancement is to form an estimate $\hat{x}(n)$ of the original clean speech signal $x(n)$ based on the observed signal $y(n)$.

As speech signals can assume to be short-time stationary, i.e., stationary within short periods of time (typically $5 - 30$ms), the signals have to be processed in frames of proper length. The principle of the frame-based processing is explained using the prominent example of the discrete Fourier transform (DFT)-based short-time Fourier transform (STFT), which is defined as

$$Y(k,m) = \sum_{n=1}^{N} y(mR+n)\, w_a(n) \exp\left(\frac{-j2\pi kn}{N}\right) \tag{3.1}$$

where m and k are the frame number and frequency bin index, respectively. The analysis window function is denoted by $w_a(n)$, while the frame shift between two consecutive frames in samples specified by R, with N denotes the frame duration in samples. The frame is processed in the frequency domain and is then applied with an inverse DFT (IDFT) and a synthesis window $w_s(n)$. Since the frames are overlapping, the respective samples of the current frame have to be added to the previous frame, which is known as the overlap-add method. The benefit of overlapping frames is that a smooth transition between consecutive frames can be achieved. The window functions $w_a(n) = w_s(n) = w(n)$ and frame shift R have to be chosen such that perfect reconstruction is achieved.

Noise reduction can be viewed as the application of a non-negative real-valued spectral weighting gain $G(k,m)$, to each frame m and each frequency bin k of the observed signal spectrum

$$Y(k,m) = X(k,m) + V(k,m) \tag{3.2}$$

where $X(k,m)$ and $V(k,m)$ are zero-mean (complex-valued) random variables

representing DFT coefficients of the the speech target and additive noise, respectively. The DFT coefficients of both $X(k,m)$ and $V(k,m)$ can be assumed to be statistically independent across time and frequency, only when the signal frames are sufficiently long, with their overlap being sufficiently small. The spectral variances are denoted as $\lambda_x(k,m) = E(|X(k,m)|^2)$ and $\lambda_v(k,m) = E(|V(k,m)|^2)$. Let $\xi(k,m) = \lambda_x(k,m)/\lambda_v(k,m)$ and $\gamma(k,m) = |Y(k,m)|^2/\lambda_v(k,m)$ denote the *a priori* and *a posteriori* SNR, respectively. Then, an estimate $\hat{X}(k,m)$ of the original signal spectrum can be formed as

$$\hat{X}(k,m) = G(k,m)\,Y(k,m). \tag{3.3}$$

Since, generally, the statistical properties of speech signals can change for every single frame, and the gain function $G(k,m)$ is time-variant, this implies that its coefficients are to be adapted continuously over time.

3.2.2 Wiener Filter

In literature, the gain function is often built by optimally deriving an estimate of the underlying clean DFT coefficient in some statistical sense, by using e.g., the maximum likelihood (ML) estimators, the maximum a posteriori (MAP) estimators and the minimum mean square error (MMSE) estimators. Among the different approaches, the simplest approach is to use a linear MMSE estimator to predict the clean speech DFT coefficients, which is often referred to as the Wiener filter (WF) [97]. It can be derived by minimising the mean square error (MSE) while putting a constraint to the estimator to be linear in $Y(k,m)$. This leads to the simple WF gain function

$$G_{\mathrm{WF}}(k,m) = \frac{\lambda_x(k,m)}{\lambda_x(k,m) + \lambda_v(k,m)} = \frac{\xi(k,m)}{1+\xi(k,m)}. \tag{3.4}$$

However, the estimation of the clean signal variance is not trivial. By using the ML estimator $\lambda_x(k,m) \approx |Y(k,m)|^2 - \lambda_v(k,m)$, it leads to

$$G_{\mathrm{WF}}(k,m) = \frac{|Y(k,m)|^2 - \lambda_v(k,m)}{|Y(k,m)|^2} = 1 - \frac{1}{\gamma(k,m)}. \tag{3.5}$$

The WF gain function thus only depends on the PDF of the actual signals, and is independent of the actual distribution of speech and noise DFT coefficients. If the speech and noise DFT coefficients are complex Gaussian, then WF is optimal amongst all estimators.

3.2.3 Spectral Subtraction

Next, the concept of spectral subtraction, which is very similar to the WF, is illustrated. The basic idea is to subtract the noise power in each frequency bin from the power of the noisy signal

$$\begin{aligned}\lambda_x(k,m) &= |Y(k,m)|^2 - \lambda_v(k,m) \\ &= |Y(k,m)|^2 \left(1 - \frac{\lambda_v(k,m)}{|Y(k,m)|^2}\right) \\ &= |Y(k,m)|^2 |G_{\mathrm{PS}(k,m)}|^2\end{aligned} \tag{3.6}$$

which is referred to as power subtraction. Comparing Eqs. (3.5) and (3.6), the relationship between the WF and spectral subtraction can be seen as

$$G_{\mathrm{PS}}(k,m) = \sqrt{1 - \frac{\lambda_v(k,m)}{|Y(k,m)|^2}} = \sqrt{G_{\mathrm{WF}}(k,m)}. \tag{3.7}$$

Hence, the concept of power subtraction is identical to applying the square root of the WF.

Another commonly used spectral subtraction rule is known as magnitude subtraction, which differs from power subtraction in that the subtraction is based on magnitude spectra instead of power spectra

$$\begin{aligned}E(|X(k,m)|) &= |Y(k,m)| - \sqrt{\lambda_v(k,m)} \\ &= |Y(k,m)| \left(1 - \frac{\sqrt{\lambda_v(k,m)}}{|Y(k,m)|}\right) \\ &= |Y(k,m)| G_{\mathrm{MS}(k,m)}.\end{aligned} \tag{3.8}$$

Both subtraction rules and the Wiener filter can be subsumed in a generalised spectral gain function

$$G_{\mathrm{SS}}(k,m) = \left(1 - \left(\frac{\lambda_v(k,m)}{|Y(k,m)|^2}\right)^{p_1}\right)^{p_2} \tag{3.9}$$

where the parameters p_1 and p_2 are to be chosen according to Table 3.1.

	p_1	p_2
Wiener filter	1	1
power subtraction	1	0.5
magnitude subtraction	0.5	1

Table 3.1: Parameters for different subtraction rules.

3.2.4 Statistically Motivated Estimators

Since the year 1984, Ephraim and Malah [98] presented the very first work of statistical model-based approach for speech enhancement with an MMSE magnitude DFT estimator under the assumptions that both speech and noise complex DFT coefficients have a complex-Gaussian distribution. However, literature studies show that the speech DFT coefficients are better modelled by heavy-tailed super-Gaussian distributions such as the Gamma and the Laplace distributions, suggesting that better estimators could be found if these statistical characteristics were taken into account. Hence, complex DFT MMSE, MDFT MMSE, and magnitude DFT MAP estimators have been presented under a super-Gaussian PDF for the speech DFT coefficients [99–101].

Non-linear estimators can be derived in a Bayesian framework by minimising a non-negative cost-function, such as the square-error cost function, which leads to the MMSE estimator [102]. Besides the square-error function, many other cost-functions, which were stated to be perceptually more relevant to speech processing, have also been proposed. Table 3.2 shows different cost functions derived for estimating the complex or the magnitude DFT coefficients, where ζ and ν indicate the shape parameters of the assumed clean speech distribution [103]. Here, $\mathcal{B}$ denotes the power law applied to the magnitude DFT coefficients of real and estimated clean speech in the squared difference. The exponent $\mathscr{P}$ is applied to the clean speech magnitude DFT coefficients to perceptually weight the squared difference. The cost-functions listed Table 3.2 usually lead to estimators that can be formulated as variants of conditional mean estimators. It has to be mentioned that the development of super-Gaussian complex and magnitude DFT estimators was made under conflicting assumptions. The complex DFT estimators were derived by assuming that the real and imaginary parts of DFT coefficients are independent, while the magnitude MMSE estimators was derived by making assumptions that the speech phase is uniformly distributed and the magnitude DFT is generalised-Gamma distributed [103]. Therefore, the work in [104] proposed a unified framework that allows to derive both complex and magnitude DFT estimators under the same consistent statistical assumptions.

Magnitude DFT Estimators		
Cost function	Assumed distribution	ref.
$E\left(\left(\mathcal{A}(k,m)-\hat{\mathcal{A}}(k,m)\right)^2\right)$	$\zeta=2$, $\nu=1$	[98]
$E\left(\left(\mathcal{A}(k,m)-\hat{\mathcal{A}}(k,m)\right)^2\right)$	$\zeta=1$ and $\zeta=2$	[106–108]
$E\left(\left(\mathcal{A}(k,m)^2-\hat{\mathcal{A}}(k,m)^2\right)^2\right)$	$\zeta=2$, $\nu=1$	[109]
$E\left(\left(\mathcal{A}(k,m)^2-\hat{\mathcal{A}}(k,m)^2\right)^2\right)$	$\zeta=2$	[110]
$E\left(\left(\log\{\mathcal{A}(k,m)\}-\log\{\hat{\mathcal{A}}(k,m)\}\right)^2\right)$	$\zeta=2$, $\nu=1$	[111]
$E\left(\left(\log\{\mathcal{A}(k,m)\}-\log\{\hat{\mathcal{A}}(k,m)\}\right)^2\right)$	$\zeta=2$	[112]
$E\left(\left(\mathcal{A}(k,m)^{\mathcal{B}}-\hat{\mathcal{A}}(k,m)^{\mathcal{B}}\right)^2\right)$	$\zeta=2$, $\nu=1$	[113]
$E\left(\left(\mathcal{A}(k,m)^{\mathcal{B}}-\hat{\mathcal{A}}(k,m)^{\mathcal{B}}\right)^2\right)$	$\zeta=2$	[114, 115]
$E\left(\mathcal{A}(k,m)^{\mathscr{P}}\left(\mathcal{A}(k,m)-\hat{\mathcal{A}}(k,m)\right)^2\right)$	$\zeta=2$, $\nu=1$	[116]
$E\left(\mathcal{A}(k,m)^{-2\mathscr{P}}(\mathcal{A}(k,m)^{\mathcal{B}}-\hat{\mathcal{A}}(k,m)^{\mathcal{B}})^2\right)$	$\zeta=2$, $\nu=1$	[117]
Complex DFT Estimators		
$E\left(\left(S(k,m)-\hat{S}(k,m)\right)^2\right)$	$\zeta=1$ and $\zeta=2$	[104, 118]

Table 3.2: Overview of non-linear DFT estimators presented in literature.

Although much work had been done to propose non-linear MMSE estimators based on *a priori* assumptions about the speech and noise distributions, the cost functions often lead to solutions without closed-form expression. Also, when short and/or overlapping frames are used for spectral analysis, the assumption about the inter-frame independence will not be valid anymore [105]. This would cause a conflict since most statistically motivated estimators are derived by assuming the frames are statistically independent. By taking the assumption about the dependency between successive frames into account would again result in estimators without closed-form solutions. As such, they are often prohibitive to be implemented in mobile devices. Also, since speech is highly non-stationary and a wide range of different noises can be encountered in a real environment, their DFT distributions would change from time to time and from place to place. In those cases, the solutions derived from explicit mathematical expressions may have poorer performance than expectation.

3.3 Behaviour of the A Posteriori SNR

As can be seen in the previous sub-section, the *a posteriori* SNR involves calculation of ensemble averages, which cannot be done and has to be estimated since only a limited number of sample functions of the random process is available in practice. This is usually done by using first recursively averaging periodograms, such that the estimate of the *a posteriori* SNR is given by

$$\hat{\gamma}(k,m) = \frac{\hat{\lambda}_y(k,m)}{\hat{\lambda}_v(k,m)} \tag{3.10}$$

where both the noisy speech estimate $\hat{\lambda}_y(k,m)$ and the noise estimate $\hat{\lambda}_v(k,m)$ can be obtained as

$$\hat{\lambda}_y(k,m) = \alpha\hat{\lambda}_y(k,m-1) + (1-\alpha)|Y(k,m)|^2 \tag{3.11}$$

$$\hat{\lambda}_v(k,m) = \alpha\hat{\lambda}_v(k,m-1) + (1-\alpha)|V(k,m)|^2 \tag{3.12}$$

and α is the averaging constants. Ideally, the averaging constants should be chosen as closed to one as possible, e.g., $\alpha \approx 1$ to obtain reliable estimate with low variance. However, this estimate is not capable of capturing the non-stationary nature of speech signals, and will introduce undesired reverberant effects in the output signals. Choosing an α value close to 0 means that the SNR estimate depends only on the current frame, which contains large fluctuations and may result in an increase of noise artifacts. Therefore, a compromise is required in practice in the choice of the smoothing constants.

Practically, two different averaging parameters, α_y and α_v, have to be used for noisy speech signal and noise, respectively. The value of α_y should be notably smaller than α_v since noise can be assumed to be much more short-time stationary than speech. As such, when considering that the noise reference is not perfect and the speech is leaking into the noise estimate, the speech components in the noise estimate can be smoothed by the longer averaging time. Since the averaging time in the speech estimate is shorter, the speech components in the numerator stands out more, resulting in a less biased *a posteriori* SNR estimate.

Now consider the case during speech pauses, i.e., $Y(k,m) \equiv V(k,m)$, the *a posteriori* SNR depends on both the smoothing parameters. More precisely,

the SNR estimate ideally should be $\hat{\gamma}(k,m) = 1$ (0 decibel (dB)) when speech is absent, by applying different values of smoothing factors, $\alpha_y > \alpha_v$ however leads to an SNR estimate spreading around 0 dB. Such distribution has been investigated on many occasions. If the noise power spectral density (PSD) is assumed to be complex Gaussian distributed, it can be shown that the numerator and the denominator of the *a posteriori* SNR estimate are random variables of Chi-Square distributions with two degrees of freedom [119]. Since the PSD estimate is made up of a sum of M Chi-Square random distributions with two degrees of freedom each, this results in a Chi-Square distribution with $2M$ degrees of freedom [92]. The relationship between M and α can be derived as [120]

$$M = \frac{1+\alpha}{1-\alpha}\frac{1}{1+2\sum_{m=1}^{\infty}\alpha^m\psi(m)} \tag{3.13}$$

where $\psi(m)$ can be computed by

$$\psi(l) = \frac{(\sum_{\mu=1}^{K-1} w(\mu)w(\mu+lR))^2}{(\sum_{\mu=1}^{K-1} w^2(\mu))^2}. \tag{3.14}$$

Then, it is shown that quotient of two Chi-Square variables can be represented by Fisher's F-distribution with $Q_y = 2M_y$ and $Q_v = 2M_v$ degrees of freedom [121].

In order to validate this, the SNR measure was experimentally calculated in a range of noise environments and was compared with the F-distribution. As depicted in Figures 3.1 and 3.2, the measure was found to follow closely with the F distribution in pink noise and factory noise as taken from the NOISEX-92 database [122]. It can be observed, however, that there are small but significant deviation for the upper tail at higher frequency range. This is the region that determines the amount of trade-off among speech distortion, noise reduction and musical noise to be generated in the speech enhancement system. If a spectral gain function is built based on statistical assumptions of speech and noise DFT distributions, the performance can be very limited by such false alarms in the distribution mapping. The problem can be exaggerated in highly variable environments such as babble noise, as can be seen in Figure 3.3 where the assumption is totally violated.

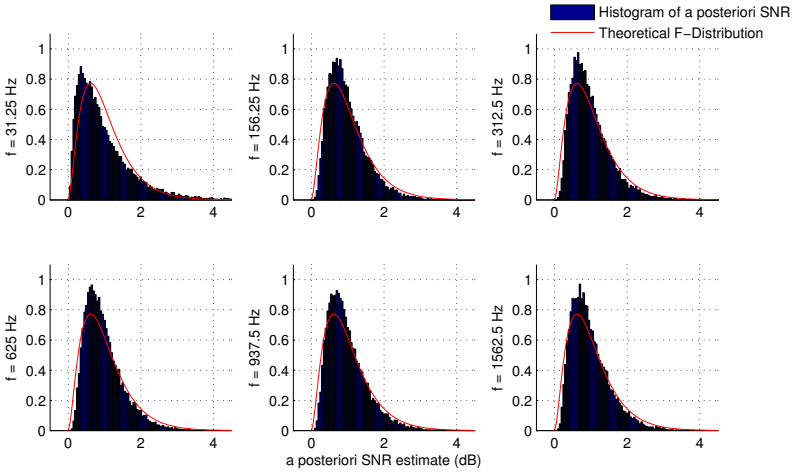

Figure 3.1: Comparison between theoretical F-distribution and histograms at different frequency bins in pink noise environment.

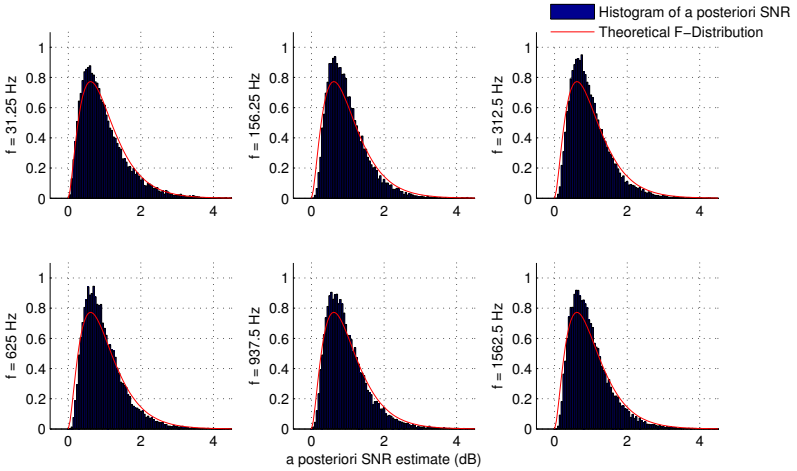

Figure 3.2: Comparison between theoretical F-distribution and histograms at different frequency bins in factory noise environment.

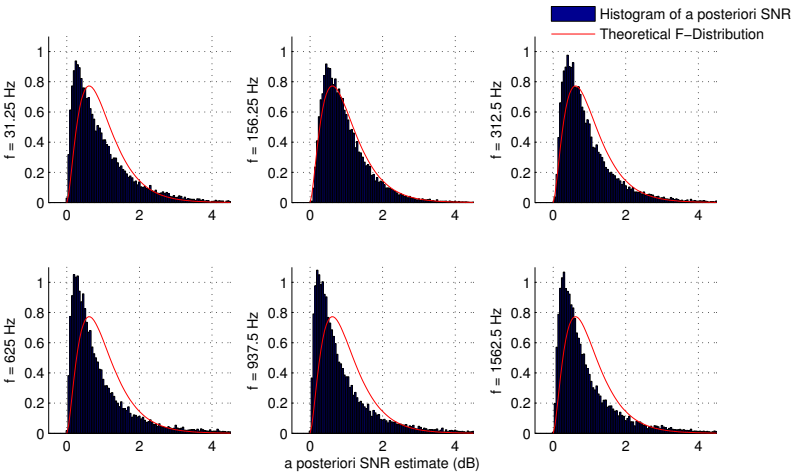

Figure 3.3: Comparison between theoretical F-distribution and histograms at different frequency bins in babble noise environment.

3.4 SNR Estimation and Gain Function

For spectral gain functions that only utilises the *a posteriori* SNR estimate, the main problem is that both the SNR estimate and the gain function comprise too many variations, which leads to musical noise. There have been many works proposed to suppress musical noise in literature. One notable work is to use Bartlett averaging periodogram for the SNR estimate and to use an adaptive exponential averaging to smooth the spectral subtraction (SS) gain function [123]. In this work, the former problem is dealt with by averaging the magnitude spectrum instead of the power spectrum estimate over time. In this case, the *a posteriori* SNR can be estimated as

$$\hat{\gamma}(k,m) = \frac{\hat{\lambda}_y(k,m)}{\hat{\lambda}_v(k,m)} \tag{3.15}$$

where both the speech estimate $\hat{\lambda}_y(k,m)$ and the noise estimate $\hat{\lambda}_v(k,m)$ can be obtained as

$$\hat{\lambda}_y(k,m) = \alpha_y \hat{\lambda}_y(k,m-1) + (1-\alpha_y)|Y(k,m)| \tag{3.16}$$

$$\hat{\lambda}_v(k,m) = \alpha_v \hat{\lambda}_v(k,m-1) + (1-\alpha_v)|V(k,m)| \tag{3.17}$$

and α_v and α_y can be obtained by

$$\alpha_y = \exp\left(\frac{-2.2R}{t_x f_s}\right) \tag{3.18}$$

and

$$\alpha_v = \exp\left(\frac{-2.2R}{t_v f_s}\right) \tag{3.19}$$

where f_s is the sampling rate, t_y and t_v denote the time averaging constants for both speech and noise, respectively.

The variability of the gain function is then studied by investigating how the SNR estimate can be mapped by the speech estimation gain function such that noise components will be attenuated while the speech components will be maintained, without generating large musical noise. In this case, the parameters of the gain function can be tuned. It is also natural for a gain function to operate between zero and one. In this chapter, two gain functions have been studied. We

consider the SS function, $G_{\mathrm{SS}}(k,m)$, given here as

$$G_{\mathrm{SS}}(k,m) = \max\left(\epsilon, 1-\beta\frac{1}{\hat{\gamma}(k,m)^p}\right) \tag{3.20}$$

where β and p are the oversubtraction factor and the power factor, respectively. The factor β is used to control the amount of speech spectral distortion, while the power factor $p < 1$ can be used to achieve high noise suppression under low SNR [124]. In addition to that, a lower p value can also allow more variations in the noise estimate. However, for $p < 1$, the gain function at high SNR region will also be attenuated and will not approach unity gain. Thus, the noise floor ϵ is introduced to control the amount of perceived residual noise and to avoid annoying musical noise [89]. The amount of musical noise is depending on the slope of the gain function and how often the SNR values come above the noise floor during noise only periods. A lower ϵ threshold can be chosen for a larger β value, which gives higher noise suppression with little musical noise, but at the same time suppresses low energy speech parts.

In order to provide a higher flexibility for speech enhancement and to control the shape of the gain function, the SIG gain function is investigated. It is given by

$$G_{\mathrm{SIG}}(k,m) = \max\left(\epsilon, \frac{1}{1+\exp\left[-a\left(\hat{\gamma}(k,m)-c\right)\right]}\right) \tag{3.21}$$

or

$$G_{\mathrm{SIG}}(k,m) = \max\left(\epsilon, 1-\left(\frac{1-\tanh\left(\frac{a(\hat{\gamma}(k,m)-c)}{2}\right)}{2}\right)\right) \tag{3.22}$$

where a and c are the slope and the mean, respectively. This function allows the control of the SIG function's mean and slope. As such it provides a mean to suppress the noise as well as to maintain the unity gain at high SNR region.

The parameters of a gain function are optimised in terms of the level of noise suppression and the amount of musical noise generated. As such, the gain function is highly sensitive to changes in the SNR estimates when speech is active but has a constant value for noise only periods. According to Figure 3.4, which plots the PDF of SNR estimate for white noise at 938 Hz mapped with several gain functions, the SNR estimate at noise only periods is distributed approximately between 0.5 and 1.5. This means that attenuation shall only be performed when the SNR estimate falls within that region. From the figure, the blue dotted line

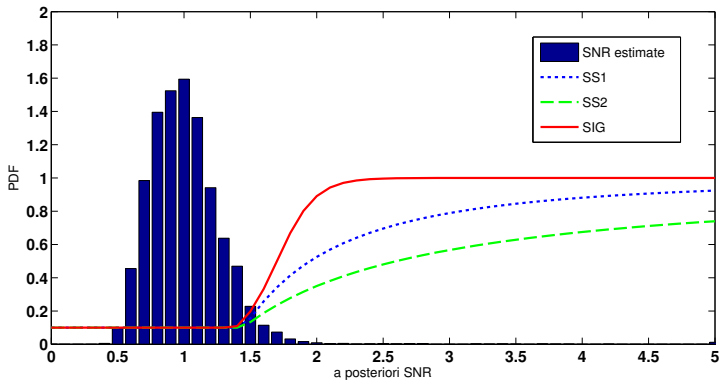

Figure 3.4: PDF of SNR estimate for white noise at 938 Hz mapped with (i) a spectral subtraction function with power spectrum estimates (SS1, $p = 2$, $\beta = 1.9$), (ii) a spectral subtraction function with amplitude spectrum estimates (SS2, $p = 1$, $\beta = 1.3$), and (iii) a sigmoid function (SIG).

and the green dashed line indicate a SS function with power spectrum estimates with $p = 2$, $\beta = 1.9$ (SS1) and a SS function with amplitude spectrum estimates with $p = 1$, $\beta = 1.3$ (SS2), respectively. The SIG function is represent by the red solid line. It can be observed that the gain values for both SS function are significantly lower than the SIG function when the *a posteriori* SNR is larger than 1.5. For SIG function in Eq. (3.21) or (3.22), by mapping the gain function to the SNR estimate, the parameters a and c can be optimised for varying noise types and SNRs. The mean value c can be optimised based on objective evaluation and the SNR estimate, while the slope a is more of a challenge since a larger a indicates more speech distortion while a smaller a indicates lower noise reduction. Furthermore, the optimisation problem is exacerbated by the type of noise that corrupts the noisy speech. Thus, it is important to understand the complication in parameters selection considering the wide options.

3.5 Optimisation

The performance of a single channel speech enhancement is often determined by how much and how well noise is suppressed without introducing any speech distortion. The evaluation is always done by subjective listening tests, or objective measurement that can more or less represent the listening experience. This section aims to optimise the parameters in the gain function based on a proposed

multi-objective optimisation algorithm, which can be formulated as

$$\max_{a} \quad W \text{ PESQ} - (1 - W) \text{ LLR} \tag{3.23}$$

where W denotes a trade off between two objective measures, $0 \leq W \leq 1$. This measure is used as a composite measure since conventional objective measures do not correlate highly with overall speech quality [96]. Here, PESQ and LLR measures are used as the criteria of the optimisation problem. PESQ measure has been proposed in ITU-T Recommendation P.862 and has recently been suggested to be more reliable than other traditional objective measures for speech quality [96]. It was implemented based on the steps in [67], which consists of pre-processing and filtering, time alignment, auditory transformation, computation of the difference between loudness spectra and time averaging of both reference and test signals. A higher PESQ score yields a better perceived speech quality [67]. The LLR measure was reported in [96] as a reliable objective measure for speech distortion. It is a speech quality objective measure that evaluate the dissimilarity of the all-pole models between the clean and the processed speech signals [67],

$$d_{\text{LLR}}\left(\vec{l}_{\widehat{x}}, \vec{l}_{x}\right) = \frac{\vec{l}_{\widehat{x}} \mathrm{R}_{x} \vec{l}_{\widehat{x}}^{T}}{\vec{l}_{x} \mathrm{R}_{x} \vec{l}_{x}^{T}} \tag{3.24}$$

where $\vec{l}_{x}$ and $\vec{l}_{\widehat{x}}$ are the linear predictive coding (LPC) coefficients of the clean speech signal and the processed speech signal respectively, and R_{x} is the autocorrelation matrix of the clean speech signal. A lower LLR score indicates a better speech quality.

We begin with the evaluation of (i) SS function with power spectrum estimates (SS1); (ii) SS function with amplitude spectrum estimates (SS2) and (iii) SIG function. By mapping the gain functions to the PDF of SNR estimate as shown in Figure 3.4, parameters β, a and c were chosen such that the gain functions would stay constant during noise only periods to avoid musical noise. Since the SNR estimate is distributed mainly between 0.5 and 1.5 during noise only periods, the gain functions should be constant up to SNR = 1.5 in order to minimise the amount of musical noise. With this in mind, the parameters for SS1 and SS2 are optimised based on the mapping of the gain functions to the fitted distribution in Figure 3.4, the objective measures (PESQ and LLR) and the informal listening

tests. Hence, the optimal parameters are $\beta = 1.9$ for SS1 with $p = 2$, and $\beta = 1.3$ for SS2 with $p = 1$. This is consistent with the findings in [67] that the oversubtraction factor should range from 1.3 to 2.0 for low SNR conditions. The noise floor for the gain functions was set as a constant value $\epsilon = -20$ dB.

For SIG function, from an exhaustive study based on the similar procedure used for SS function, by using the NOISEX-92 database, and with the help from the objective measure in Eq. (3.23), we have obtained the optimised mean value as $c = 1.7$ from the distribution of SNR estimate. The slope can be set to $a = 7$ to achieve the same amount of noise suppression when compared to SS1 and SS2 as shown in Figure 3.4. However, in order to find the optimal performance of SIG function in different noise conditions and SNRs, we optimise a based on the objective function defined in Eq. (3.23).

3.6 Experimental Results

The evaluation of the speech enhancement gain functions was done by using a database of noisy speech corpus named NOIZEUS [67]. The database contains 30 IEEE sentences produced by 3 male and 3 female speakers and corrupted by 8 different types of noise at global SNR levels of 0 dB, 5 dB, 10 dB and 15 dB. In this work, white noise, pink noise and factory noise were used for evaluation.

The recursive averaging constants were chosen as $\alpha_V = 0.9912$ with 1 second averaging time and $\alpha_X = 0.8636$ with 60 millisecond averaging time. The frame size was chosen as $K = 256$ with frame rate $R = 64$. A sampling frequency of $fs = 8000$ Hz and a 256 points Hamming Window were applied.

By using the cost function as defined in Eq. (3.23), with $W = 0.5$, the optimal points for SIG function (SIGopt) at different noise conditions and SNRs were determined. Tables 3.3, 3.4, and 3.5 summarise the mean value for 30 NOIZEUS sentences that had been corrupted by white, pink and factory noise, respectively. In these three tables, SIGopt at different SNRs and noise conditions are compared to the corresponding results obtained from the the noisy signal, SS1 and SS2. The optimum points for SIG function with the corresponding slope value a, which were obtained from Eq. (3.23), can be identified from Figure 3.5. As observed, with

SNR	**Noisy**	**SS1**	**SS2**	**SIGopt**
0	-0.0378	0.1575	0.1303	0.2726
5	0.1235	0.2569	0.2352	0.4082
10	0.3346	0.3930	0.3925	0.5670
15	0.5739	0.5307	0.5577	0.7264

Table 3.3: Optimal value of objective function, white noise.

SNR	**Noisy**	**SS1**	**SS2**	**SIGopt**
0	0.0855	0.4423	0.4005	0.2910
5	0.3123	0.5502	0.7288	0.7736
10	0.5579	0.6359	0.7903	0.9265
15	0.8035	0.7399	0.8814	1.0580

Table 3.4: Optimal value of objective function, pink noise.

SNR	**Noisy**	**SS1**	**SS2**	**SIGopt**
0	0.4446	0.5559	0.6405	0.7713
5	0.6846	0.6764	0.7444	0.9111
10	0.9192	0.7960	0.8592	1.056
15	1.1379	0.9396	1.0018	1.217

Table 3.5: Optimal value of objective function, factory noise.

the flexibility of the parameters a and c, the optimal values of SIGopt are much higher than the results of both the SS functions. However, for pink noise at 0 dB SNR, the performance of SIG is slightly lower than SS1 and SS2. Despite that, the tables show that there are significant improvements between SIGopt at 0 dB SNR pink noise and the noisy signal. The possible solution to increase the objective scores for SIG function at 0 dB SNR pink noise is to increase its mean value, c. Besides that, we can also observe that the optimal a becomes smaller with an increase of SNR. This is because the cost function finds the optimal points of the gain function that minimise speech distortion. Although a larger a leads to a higher noise reduction, it will increase the amount of speech distortion in the enhanced speech signal. This indicates that the gain function can be less aggressive at higher SNR for lower speech distortion.

Table 3.6 shows the results from both LLR and PESQ measures. Similarly, for both individual objective measures, the performance of SIGopt is slightly better than the performance of SS1 and SS2 except for 0 dB SNR pink noise that acts as an outlier in these results. From the tables, it can be observed that

LLR					PESQ			
white noise								
SNR	Noisy	SS1	SS2	SIGopt	Noisy	SS1	SS2	SIGopt
0	1.5978	1.4843	1.4949	1.4340	1.5221	1.7994	1.7556	1.9800
5	1.4978	1.4383	1.4352	1.3480	1.7448	1.9521	1.9056	2.1680
10	1.3708	1.3609	1.3471	1.2570	2.0401	2.1470	2.1320	2.3960
15	1.2245	1.2601	1.2386	1.1530	2.3724	2.3214	2.3541	2.6170
pink noise								
0	1.4225	1.3221	1.2659	1.3100	1.5935	2.2067	2.0669	1.8920
5	1.2867	1.2429	1.1057	1.0750	1.9114	2.3432	2.5633	2.6260
10	1.1383	1.1794	1.0915	0.9941	2.2541	2.4512	2.6721	2.8530
15	0.9920	1.1021	1.0441	0.9189	2.5991	2.5820	2.8068	3.0580
factory noise								
0	1.1480	1.2071	1.1066	1.0220	2.0371	2.3188	2.3876	2.5680
5	1.0047	1.0873	1.0339	0.9331	2.3740	2.4401	2.5226	2.7630
10	0.8680	0.9897	0.9498	0.8395	2.7065	2.5818	2.6683	2.9700
15	0.7519	0.9002	0.8682	0.7448	3.0277	2.7794	2.8718	3.2240

Table 3.6: Average results for LLR and PESQ measures for 3 types of noise.

both measures show similar behaviour in defining the quality of a speech signal in terms of speech distortion.

In order to validate the performance of objective measures, informal subjective listening tests had been performed in factory noise at both 0 dB and 10 dB SNRs. The listening tests were conducted with closed-headphones on ten listeners with age ranged from 20-25 years old. According to the amount of perceived noise and speech distortion, each listener was required to rate each signal from a scale between one and five: 5 = Excellent, 4 = Good, 3 = Fair and 1 = Bad. All the listening results were averaged to obtain a mean opinion score (MOS) as described in [67]. For each listener, the applied procedures are: (1) clean speech and noisy speech were played and repeated upon request; (2) test signals were randomly played. The used parameters are: $b = 1.9$ for SS1, $b = 1.3$ for SS2, $a = 1.1, c = 1.7$ for SIG at 0 dB factory noise, and $a = 1.7, c = 1.7$ for SIG at 10 dB factory noise.

Table 3.7 lists the subjective MOS results of the noisy speech signal and the enhanced signals. For factory noise at both 0 dB and 10 dB, a human listener prefers the SIGopt approach when compared to SS1 and SS2 methods. These results match the performance of LLR and PESQ measures in Table 3.6.

SNR	Noisy	SS1	SS2	SIGopt
0 dB	1.83	2.08	2.33	2.75
10 dB	3.33	3.33	2.83	3.50

Table 3.7: Subjective MOS of the speech signals in factory noise.

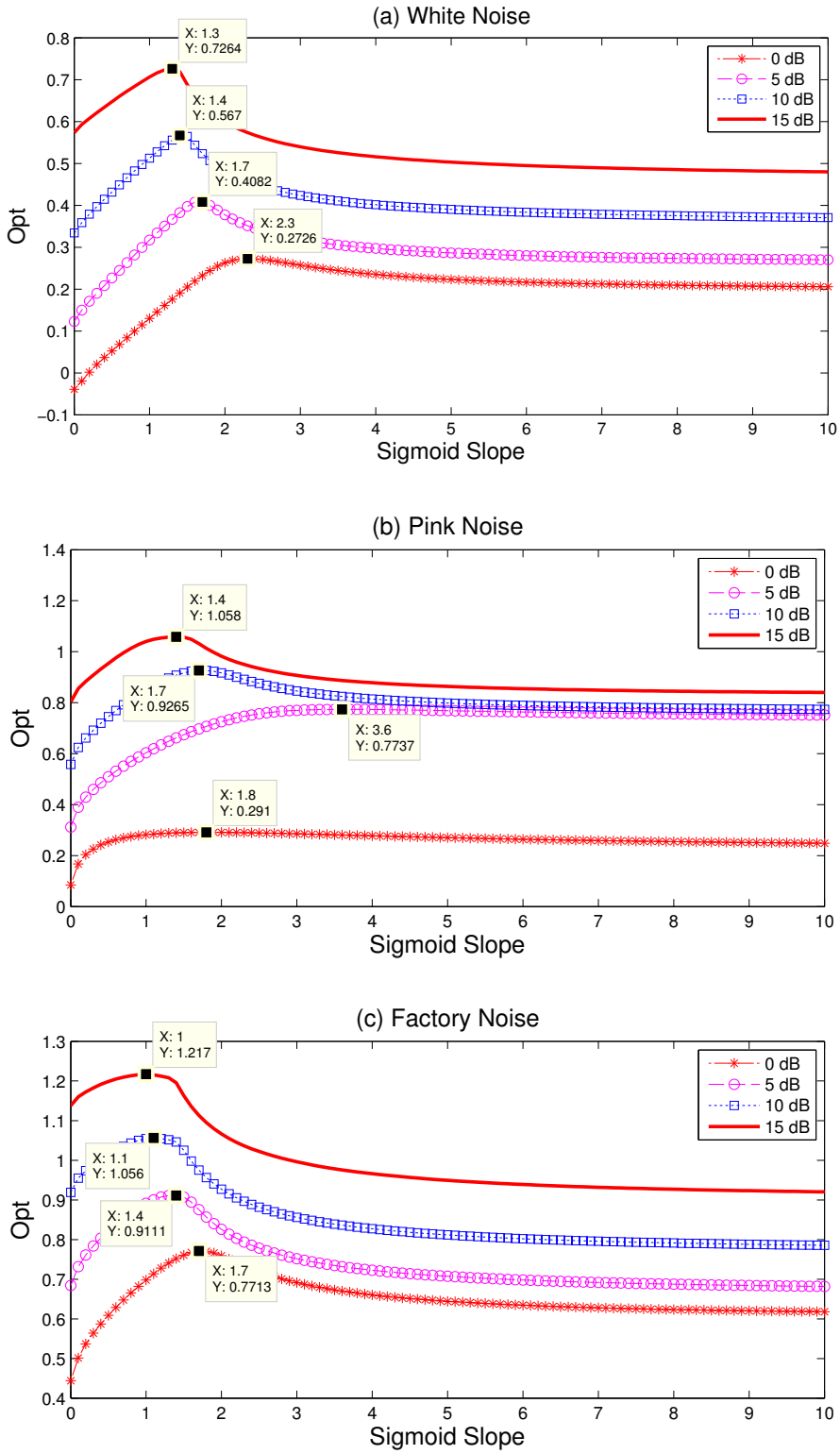

Figure 3.5: Cost function for optimisation for different slope a with mean $c = 1.7$ at different SNRs: (a) white noise; (b) pink noise; and (c) factory noise.

3.7 Summary

This chapter reviews the single-channel speech enhancement gain functions in the literature and presents a methodology to optimise the mean and the slope of the sigmoid function, which does not require any *a priori* information about the speech signal. It was shown that the SNR estimate and the gain function impact the objective measures and provide varying subjective quality. The gain function parameters were designed such that during the noise only periods it provides a constant suppression thus avoiding annoying non-linear artefacts (musical noise). This was done by mapping the function to the distribution of the SNR estimate. Optimisation of the sigmoid function was done based on two widely used objective measures: PESQ and LLR. Experimental results prove that with the proper choice of parameters, the sigmoid function can be optimised to enhance the quality of a noisy speech while maintaining more energy of the speech components when compared to the state-of-the-art spectral subtraction function. In the next chapter, a solution for SIG function with *a priori* SNR estimate, together with a metric to measure the trade-off between noise reduction, speech distortion and musical noise, will be proposed.

Chapter 4

Speech Enhancement using a Modified Sigmoid (MSIG) Function with A Priori SNR Estimate

The important thing in science is not so much to obtain
new facts as to discover new ways of
thinking about them.
– Sir William Bragg

4.1 Introduction

Over the past five decades, a vast amount of short-time spectral domain speech enhancement algorithms have been published and developed for applications such as mobile phones and hearing aids. In terms of single channel approach, the best known methods are the spectral subtraction (SS) [88, 89, 93], the minimum mean square error (MMSE) based estimator [98, 111], and the Wiener filter (WF) [125]. Among these algorithms, SS is more often utilised in real world implementation due to its relative simplicity, which only requires an estimate of the noise power spectrum for computing the *a posteriori* signal-to-noise ratio (SNR). For MMSE based algorithms and WF method, *a priori* SNR, which involves an estimate of

the clean speech signal, is required. Although this increases the complexity of the problem, it was stated in the literature that the performance of the gain functions is mainly determined by the *a priori* SNR, while the *a posteriori* SNR acts only as a correction parameter for low *a priori* SNR [126]. Since SS employs only the *a posteriori* SNR without utilising the statistics and the distributions of the stochastic signal process, its performance is limited, which results in audible sound artifacts in the enhanced speech signal known as the musical noise. In order to solve this, a speech enhancement scheme in the modulation domain rather than in the conventional acoustic domain has been proposed in the literature [127, 128]. However, a low delay solution to reduce the musical noise can be achieved by improving the *a priori* SNR estimate in the acoustic domain.

The most widely used approach for estimating the *a priori* SNR is the decision-directed (DD) approach [98]. The DD approach performs a linear combination of two components: one being an estimate of previous *a priori* SNR and another being the maximum likelihood (ML) SNR estimate. By applying a weighting factor close to unity of the past *a priori* SNR estimate, the DD approach corresponds to a highly smoothed version of the *a posteriori* SNR, which reduces the musical noise [126]. The drawback of reducing the variance in the *a priori* SNR estimate is that it cannot react quickly to abrupt changes in the instantaneous SNR. This gives rise to a performance degradation in speech enhancement scheme due to the speech transient distortion. In order to reduce the transient distortion, many algorithms have been proposed in the literature [129–134]. Most of them have outperformed the traditional DD approach in terms of objective evaluations [135].

In general, the performance of a speech enhancement scheme depends on the joint temporal dynamics between the SNR estimate and the gain function [115]. For instance, the MMSE-log spectral amplitude (LSA) estimator with the DD *a priori* SNR estimate can generate speech signals without audible musical noise, provided that the weighting factor is close to unity [67, 111]. Unlike the LSA approach, WF with the DD approach generates more speech distortion and musical noise. The main reason behind this is that the WF is a more aggressive gain function for low SNR. The result is a tendency to suppress more weak-speech components together with the residual noise. Thus, when compared to WF method,

the LSA approach is preferred for less musical noise and speech distortion. In addition to that, much progress has been made in the development of MMSE estimators based on different cost functions and/or different statistical prior models to improve speech quality [100, 114, 115, 136]. However, these algorithms involve the calculation of the confluent hypergeometric functions, which require a lot more computational complexity to implement when compared to WF and LSA methods.

Here, we are interested in developing a low complexity gain function, which employs the *a priori* SNR estimate with good noise suppression performance for real-time implementation. As such, the WF and LSA approaches will be used as the benchmark. Another method to obtain the gain function is to use a sigmoid (SIG) function. The rationale for using SIG function as a gain function is that it is a logistic function and can be viewed as a general cumulative distribution function (CDF) [5]. This gain function provides several parameters that can be adjusted to flexibly model exponential distributions. By optimising the parameters of a SIG function, a well-balanced trade-off between noise reduction, speech distortion and musical noise can be achieved [5]. Although this can also be achieved by employing an over-subtraction parameter on WF, only the mean of WF will be shifted while the shape will remain unchanged. This will give different sensitivity in the feedback when the gain is applied to the *a priori* SNR estimate. Therefore, a modified WF is not preferable as it does not provide as much flexibility offered by SIG function.

In this work, a modified sigmoid (MSIG) gain function has been proposed to increase the flexibility of the speech enhancement gain function and to provide a vehicle to enhance the speech quality in high noise conditions. The MSIG function combines a logistic function with a hyperbolic tangent function, providing a more adjustable gain function with three controllable parameters. Optimization has been performed on these parameters to fit the MSIG function to either the LSA approach, WF method or SIG function to demonstrate the flexibility of the proposed gain function. In addition, a modified decision-directed (MDD) SNR estimator, which basically nullifies the one frame delay in the conventional DD

SNR estimator without increasing the computational complexity, has been developed. This is achieved by matching the estimate of the clean speech spectrum and the *a priori* SNR estimate with the noisy speech spectrum in the current frame rather than the previous one.

The remainder of this chapter is organised as follows. Section 4.2 gives a system overview. Section 4.3 shows the proposed SNR estimate. Section 4.4 demonstrates the modified gain function. Section 4.5 outlines the objective measures used for performance evaluation and Section 4.6 presents the results. Section 4.7 concludes the chapter.

4.2 System Overview

The goal of speech enhancement is to compute the enhanced speech signal $\hat{x}(n)$, given a noisy signal $y(n) = x(n) + v(n)$, where $x(n)$ is the clean speech signal and $v(n)$ is the uncorrelated additive noise. By using the short-time Fourier transform (STFT), the spectral components of the noisy speech $Y(k, m)$ can be obtained by[1]

$$Y(k,m) = \sum_{n=1}^{N} y(mR+n)\, w(n) \exp\left(\frac{-j2\pi kn}{N}\right) \tag{4.1}$$

where k is the frequency bin index, m is the frame index, N is the frame duration in samples, R is the frame shift in samples, and $w(n)$ is a window function. The clean speech spectrum estimate $\hat{X}(k, m)$ is then obtained by

$$\hat{X}(k,m) = G(k,m)Y(k,m) \tag{4.2}$$

where $G(k, m)$ is a non-linear spectral gain function. The gain function can be expressed as a function of the *a priori* SNR

$$\xi(k,m) = \frac{E\{|X(k,m)|^2\}}{E\{|V(k,m)|^2\}} = \frac{\lambda_x(k,m)}{\lambda_v(k,m)} \tag{4.3}$$

and/or the *a posteriori* SNR

$$\gamma(k,m) = \frac{|Y(k,m)|^2}{E\{|V(k,m)|^2\}} = \frac{|Y(k,m)|^2}{\lambda_v(k,m)} \tag{4.4}$$

where $\lambda_x(k, m)$ and $\lambda_v(k, m)$ denote clean speech power spectral density (PSD) and noise PSD, respectively.

[1]For convenience, we reproduce the equations from Chapter 3.

The gain function is often derived from MMSE optimisation criteria. One of those is the WF method, which minimises the expected value $\mathrm{E}\{|X(k,m) - \hat{X}(k,m)|^2\}$. It can be computed using the *a priori* SNR as

$$G_{\mathrm{WF}}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)}. \tag{4.5}$$

Other widely used algorithms are based on a direct estimate of the clean speech spectral magnitude. One of them is the LSA estimator, which can be obtained by minimising $\mathrm{E}\{[\log(|X(k,m)|) - \log(|\hat{X}(k,m)|)]^2\}$ [111]. The resulting gain function for the LSA approach can be obtained as

$$G_{\mathrm{LSA}}(k,m) = \min\left\{\varsigma, \frac{\xi(k,m)}{1+\xi(k,m)}\left[\frac{1}{2}\int_{\nu(k,m)}^{\infty}\frac{e^{-t}}{t}dt\right]\right\} \tag{4.6}$$

with

$$\nu(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)}\gamma(k,m) \tag{4.7}$$

where ς denotes a practical upper bound used to prevent a large gain value at low *a posteriori* SNR. Here, we choose $\varsigma = 10$.

4.3 A Priori SNR Estimation

Prior to the computation of the spectral gain function, two parameters have to be estimated: the noise PSD and the *a priori* SNR. In this work, the noise PSD is estimated by using the MMSE noise estimator in [137]. Meanwhile, for the estimation of the *a priori* SNR, the most widespread method is the DD approach, given by [98]

$$\hat{\xi}_{\mathrm{DD}}(k,m) = \max\left\{\beta\frac{|\hat{X}(k,m-1)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1], \xi_o\right\} \tag{4.8}$$

where $\hat{\lambda}_v(k,m)$ and $\hat{X}(k,m-1)$ denote the estimated noise PSD and the estimated clean speech spectrum from the preceding frame, respectively. The parameter β denotes the smoothing factor, $P[.]$ denotes the half-wave rectification and ξ_o denotes a SNR floor. As observed from the equation, the first term is the *a priori* SNR estimate in the previous frame while the second term is an ML estimate computed from the *a posteriori* SNR. The advantage of the DD approach is its capability to eliminate musical noise based on the choice of β in the conditional

smoothing procedure [126]. It was suggested to set β close to unity so that the musical noise can be eliminated, particularly in conjunction with the MMSE estimator approach. However, this leads to a slow update of the *a priori* SNR estimate, resulting in speech transient distortion, especially in speech onsets. This is due to little influence of the second term $(1-\beta)P[\gamma(k,m)-1]$ in the update.

In addition, the DD approach based on Eq. (4.8) has an extra frame delay during speech transients since it employs the previous frame clean speech spectrum estimate. As a consequence, the gain function matches the previous frame instead of the current one. Although the delay can be reduced by choosing a smaller β in Eq. (4.8), more musical noise will be perceived since $(1-\beta)P[\gamma(k,m)-1]$ is usually unsmoothed. Thus, we propose a MDD approach to reduce the delay in speech transients by matching both estimates of the clean speech spectrum and the *a priori* SNR estimate with the current noisy speech spectrum. The first term of the DD approach is modified such that the gain function at previous frame is mapped with the current noisy speech spectrum rather than the previous one. As such, the MDD approach is given by

$$\hat{\xi}_{\mathrm{MDD}}(k,m) = \max\left\{\beta\frac{|G_{(.)}(k,m-1)Y(k,m)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1], \xi_o\right\} \tag{4.9}$$

where $G_{(.)}$ indicates that the same gain function is used to obtain both the *a priori* SNR estimate and the speech estimate. The advantage of this new approach is that it has the same computational complexity as the DD approach while having a better enhanced speech quality, which makes it suitable for real-time implementation.

According to Eq. (4.9), the first term of the MDD approach does not contain an estimate of the *a priori* SNR at previous frame when compared to the original method. Therefore, the MDD approach is no longer representing a conditional first order recursive averaging algorithm as in Eq. (4.8). This means that it increases the sensitivity of the *a priori* SNR estimate towards abrupt changes in speech signal, which directly reduces the speech transient distortion. However, such variance in the *a priori* SNR estimate can lead to audible musical noise due to the higher sensitivity to changes. In order to reduce, or eliminate such musical noise, a first order recursive smoothing procedure can be applied in the

a posteriori SNR estimation in Eq. (4.4) as [5]

$$\hat{\gamma}(k,m) = \frac{\lambda_y(k,m)}{\hat{\lambda}_v(k,m)} \tag{4.10}$$

where

$$\lambda_y(k,m) = \alpha_y \lambda_y(k,m-1) + (1-\alpha_y)|Y(k,m)|^2. \tag{4.11}$$

The parameter λ_y is the noisy speech PSD, which is obtained by smoothing the magnitude square of the noisy signal. The smoothing constant is defined as $\alpha_y = \exp\left(\frac{-2.2R}{t_y f_s}\right)$, where R is the frame rate from Eq. (4.1), while t_y and f_s denote the time averaging constant and the sampling rate, respectively.

4.4 Modified Sigmoid Gain Function

Most of the gain functions developed for speech enhancement schemes are based on optimisation criteria, such as the LSA approach, the WF method and all other MMSE-based algorithms [100, 114, 136]. The problem is that the optimisation of the criteria is made under certain model conditions such as stationarity and certain distributions. Ideally it is desirable to have a gain function that offers optimal performance in all scenarios. This will lead to different cost functions and gain functions giving different performance in different background noise scenarios. Apart from that, some of them involve complex mathematical equations that require large computational load to solve, making them sometimes less attractive in real world scenarios. Thus, we propose to design a gain function with low complexity and high flexibility, while having similar or better performance when compared to most of the MMSE estimators. The important consideration is to have control over the shape of the gain function. For this purpose, a flexible sigmoid-shape function is utilised. By designing the SIG function with different shapes, a similar performance as the MMSE-based estimators will be obtained. The rationale behind using the SIG function is that it has a general CDF function with a shape that can be adjusted by several tunable parameters. In that case, the quality of the enhanced speech can be improved.

In previous work [5], a SIG function has been presented to be mapped with the *a posteriori* SNR. However, instead of mapping with the *a posteriori* SNR, which

limits the performance of the gain function, here the SIG function is mapped with the *a priori* SNR estimate. The gain function is given as

$$G_{\mathrm{SIG}}(k,m) = \frac{1}{1+\exp\left[-a\left(\hat{\xi}\left(k,m\right)-c\right)\right]} \tag{4.12}$$

where a and c are parameters that control the slope and the mean of the gain curve, respectively. Both parameters control the amount of musical noise, speech distortion and noise reduction in the enhanced speech. In order to obtain a balanced trade-off among them, the sigmoid slope has to be sensitive towards speech and less sensitive towards the variation of noise. In this case, the mean of the SIG function has to be less than 1. This is not plausible as when the mean value is approaching zero, the gain value will not reach zero until a very small SNR value, which leads to insufficient noise reduction. To provide more noise reduction at low SNR conditions, a MSIG function, which has three parameters that can be adjusted or optimised for better enhanced speech quality, is developed. The proposed function is obtained by multiplying the original logistic function in Eq. (4.12) with a hyperbolic tangent function as

$$G_{\mathrm{MSIG}}(k,m) = \frac{1-\exp\left(-a_1\hat{\xi}\left(k,m\right)\right)}{1+\exp\left(-a_1\hat{\xi}\left(k,m\right)\right)}\left(\frac{1}{1+\exp\left(-a_2\left(\hat{\xi}\left(k,m\right)-c\right)\right)}\right). \tag{4.13}$$

By changing the parameter values, the behaviour of the MSIG function can be made similar to the different conventional gain functions, such as the LSA, the WF and the SIG approaches. To achieve this, an optimisation problem has been set up. The problem can be formulated as

$$\min_{\mathbf{z}} \left\| G_{\mathrm{MSIG}}\left(\mathbf{z},x\right) - D(x) \right\|_2^2 \tag{4.14}$$

where $\mathbf{z} = [a_1\ a_2\ c]$ and $D(x)$ is a gain function chosen from WF, LSA and SIG. The optimisation problem in Eq. (4.14) is a non-linear optimisation problem in terms of the parameter $\mathbf{z}$. A solution for the problem can be obtained by using the minimisation function *lsqnonlin* in MATLAB. This solves the non-linear least-square curve fitting problem by using a trust region reflective Newton method. As such, MSIG parameters that best fit the gain function in $D(x)$ in the least-square sense can be found in Section 4.6.

4.5 Representative Objective Measures

Many objective measurement algorithms have been derived in the literature for evaluating the performance of speech enhancement algorithms [67, 96]. The most widely used methods include the perceptual evaluation of speech quality (PESQ) measure [138] and the segmental SNR (SNRseg) measure [139]. The PESQ measure, which was not originally designed to evaluate the performance of speech enhancement algorithms, has been found to have a good correlation overall with mean opinion score (MOS) [96]. It predicts the MOS scores which yields a result from 1 to 5, where a higher score indicates a better speech quality. Meanwhile, the SNRseg measure is also preferred among the vast amount of objective measures as it has been found to correlate best with background noise reduction [96].

In this chapter, both the PESQ measure and the SNRseg measure were used to evaluate the performance of the proposed algorithms. The PESQ measure was implemented based on the procedures presented in [67]. The SNRseg measure is defined as [67]

$$\text{SNRseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{||\mathbf{x}(m)||^2}{||\mathbf{x}(m)-\hat{\mathbf{x}}(m)||^2} \tag{4.15}$$

where the vector $\mathbf{x}(m)$ represents a clean speech (time-domain) frame, and $\hat{\mathbf{x}}(m)$ is the enhanced speech frame. In order to discard non-speech frames, each frame was threshold by a $-10\,\text{dB}$ lower bound and a $35\,\text{dB}$ upper bound.

The performance of the speech enhancement scheme has a trade-off among musical noise, speech distortion and noise reduction. The PESQ measure and the SNRseg measure can not represent the whole picture of these trade-offs. Therefore, an objective evaluation metric is also utilised to evaluate and compare the results among the amount of musical noise, speech distortion and noise reduction generated from the speech enhancement scheme.

First of all, the musical noise and the noise reduction should only be calculated during noise-only periods in short-time spectral domain. Since in practical situations the true noise PSD is often unknown, a reference VAD for the clean speech is required for performance evaluation without the knowledge of noise characteristics. In order to obtain the VAD decisions at different frames and frequency bins, a multi decisions sub-band VAD (MDSVAD) is utilised [140]. Given two

hypotheses, $\mathcal{H}_0(k,m)$ and $\mathcal{H}_1(k,m)$, which indicate speech absence and presence, respectively in the k^{th} frequency bin of the m^{th} frame, the MDSVAD is given by

$$D(k,m) = \begin{cases} 1 & \text{if } \mathcal{H}_0(k,m) \\ 0 & \text{if } \mathcal{H}_1(k,m). \end{cases} \tag{4.16}$$

The amount of musical noise is believed to be highly correlated with the number of isolated spectral components and their level of isolation [141]. Since such components have relatively high power, they can be perceived as tonal sounds that are strongly related to the weight of skirt of the probability density function (PDF). A signal with skirt can be identified using higher-order statistics, such as kurtosis. However, in order to identify only the musical-noise components, a kurtosis ratio (KurtR) is used to measure the change in kurtosis between the noisy signal and the enhanced signal. In [141], this ratio is derived as a function controlled by the over-subtraction factor in the SS function as well as the shape parameters from the distribution model of the speech or noise signal. The KurtR in this chapter is determined by the actual noisy speech signal and the enhanced speech signal during noise-only periods. Such measure is defined as

$$\text{KurtR} = E\left\{\frac{\mathcal{K}_{\hat{x}}(k)}{\mathcal{K}_y(k)}\right\} \tag{4.17}$$

where $\mathcal{K}_{\hat{x}}(k)$ and $\mathcal{K}_y(k)$ denote the kurtosis of the enhanced signal and the noisy signal, respectively at k-th frequency bin. Both of them are computed only during speech absence periods, as given by

$$\mathcal{K}_{\hat{x}}(k) = \frac{\sum_{m=0}^{M-1} |\hat{X}(k,m)\, D(k,m)|^4}{\left[\sum_{m=0}^{M-1} |\hat{X}(k,m)\, D(k,m)|^2\right]^2} - 2 \tag{4.18}$$

and

$$\mathcal{K}_y(k) = \frac{\sum_{m=0}^{M-1} |Y(k,m)\, D(k,m)|^4}{\left[\sum_{m=0}^{M-1} |Y(k,m)\, D(k,m)|^2\right]^2} - 2. \tag{4.19}$$

A smaller value of KurtR in Eq. (4.17) indicates less musical noise.

Meanwhile, the amount of noise reduction can be defined as the input noise power in dB minus the output noise power in dB. This noise reduction ratio (NRR) is defined during noise-only periods as

$$\text{NRR [dB]} = 10\log_{10}\frac{\sum_{m=0}^{M-1}\sum_{k=1}^{K} |Y(k,m)\, D(k,m)|^2}{\sum_{m=0}^{M-1}\sum_{k=1}^{K} |\hat{X}(k,m)\, D(k,m)|^2}. \tag{4.20}$$

For speech distortion measure, the log-likelihood ratio (LLR) measure is used. It is a spectral distance measure that models the mismatch between the formants of the clean and enhanced speech signals [142]. Similar to Eq. (3.24), the LLR measure is defined as [142]

$$d_{\mathrm{LLR}}\left(\vec{l}_{\widehat{x}}, \vec{l}_{x}\right) = \frac{\vec{l}_{\widehat{x}} \mathrm{R}_{x} \vec{l}_{\widehat{x}}^{T}}{\vec{l}_{x} \mathrm{R}_{x} \vec{l}_{x}^{T}} \tag{4.21}$$

where $\vec{l}_{x}$ and $\vec{l}_{\widehat{x}}$ are the linear predictive coding (LPC) coefficient vectors of the clean speech signal and the enhanced speech signal respectively, and R_{x} is the autocorrelation matrix of the clean speech signal. A lower LLR score indicates a better speech quality.

The objective evaluation metric provides a multi-criteria evaluation for the various parameters that define the speech enhancement methods. This helps to identify reasonable parameters and to provide an indication of parameter sensitivity. As such the objective evaluation has been used to obtain reasonable candidates for listening tests.

4.6 Experimental Evaluation

4.6.1 Parameter Optimisation of the MSIG Function

The parameters of the MSIG function were fitted to either WF, LSA or SIG function by using Eq. (4.14). For the SIG function, the parameters chosen were $a = 1$ and $c = 0.7$. In the optimisation procedure, the initial estimates of MSIG parameters were set as $\mathbf{z}_0 = [0\ 0\ 0]$. An upper bound constraint was employed in the optimisation procedure, such that $\mathbf{z}_{ub} = [15\ 1\ 1]$. The curve fitting was done under the condition that the instantaneous SNR is equal to the *a priori* SNR. The results of the curve fitting can be observed in Table 4.1, where the optimised parameters for three different MSIG curves from Eq. (4.13) are displayed.

Figure 4.1 plots the MSIG curves as functions of the *a priori* SNR, together with the WF gain in Eq. (4.5), the LSA approach in Eq. (4.6) and the SIG function from Eq. (4.12). From the figure, MSIG-fix1 is fitted to WF, but with slightly higher attenuation at low SNR conditions (below -4 dB). Also, MSIG-fix2 is fitted to the LSA estimator but is a more aggressive gain function below

Functions	Parameters			Fitted Curve
	a_1	a_2	c	
MSIG-fix1	2.3918	0.2120	-1.7071	LSA
MSIG-fix2	11.6869	0.4337	0.7556	WF
MSIG-fix3	15	0.6351	0.2243	SIG

Table 4.1: MSIG parameters.

$-10\,\text{dB}$ SNR. The least aggressive gain function can be found in MSIG-fix3, which does not really match the SIG function. This is because of the upper bound constraint of the parameter a_1, which offers sufficiently small gain values at lower SNR region. If the upper bounds were not imposed, MSIG will fully fit the SIG function at $a_1 = 92043$, $a_2 = 1$ and $c = 0.7$. An advantage of all the MSIG functions over the conventional methods is a lower gain value at low SNR, while having a larger gain value at high SNR region. This allows more noise to be suppressed and more speech components to be preserved.

4.6.2 Experimental Setup

For objective evaluation, 30 IEEE speech sequences were taken from NOIZEUS speech database [67] and were added with pink noise. The tests were done with 0.01 step for both $0 \leq t_x \leq 0.1$ and $0.9 \leq \beta \leq 0.99$. The smoothing constant α_y was plotted instead of t_x, in conjunction to β for consistency in terms of the frame rate. The reference MDSVAD in Eq. (4.16) were generated from the same speech sequences but with 50 dB global SNR to reduce miss-detections of speech. All results were generated with $K = 512$ frequency bins. A square-root Hann window was used for $w(n)$ with 50% overlap ($R = 256$). The value of the *a priori* SNR floor was chosen as $\xi_o = -25\,\text{dB}$. In addition, a constant residual noise floor, $\epsilon = -15\,\text{dB}$ was employed for all the gain functions, such that

$$G_{(.)}(k,m) = \max\left\{\epsilon, G_{(.)}(k,m)\right\}. \tag{4.22}$$

The results of the performance evaluation will differ with the noise estimate $\hat{\lambda}_v(k,m)$. In this experiment, the results were generated with the MMSE noise estimator in [137]. As such, a consistent simulation could be run, where the only changes in the system were the gain functions and the SNR estimators.

The performance of the proposed approach was further verified by subjective listening tests, where the listeners provided ratings for each individual component of a noisy speech signal - the speech signal, the background noise, and the musical noise [96]. The listener was prompted to rate the noisy and the enhanced speech signal on the following three criteria:

1. SPCH: the speech signal using a 5 point scale of signal distortion;
2. NSE: the noise using a 5 point scale of background intrusiveness;
3. MUSIC: the musical noise using a 5 point scale of musicalness.

The SPCH, NSE and MUSIC scales are described in Table 4.2. Note that those numbers are not absolute scales but serve as an indication of hearing experience to evaluate speech quality. A total of eight listeners (six males, two females aged between 20-30) were recruited for the listening tests. Five separate sentences, consisting of three male spoken sentences and two female spoken sentences, were included for the tests.They were taken from the same NOIZEUS database used for objective evaluations. Each of them were corrupted with either pink noise or factory noise, at 0 dB and 15 dB SNRs. Tests were performed using audio-technica ATH-ESW9 headphones. The subjects were categorised into two groups: one group was required to start the tests with 0 dB SNR cases, while the other group began the tests from 15 dB SNR cases. During the tests, the listeners were not given information about the type of the gain function and the method for the SNR estimate used in each speech. The listeners were allowed to listen to each sentence several times with access to the clean speech signal and noisy signal references. The average duration of a test was approximately 2 hours per subject.

4.6.3 Evaluation of the Proposed MDD in Estimating A Priori SNR

The performance of the MDD approach in estimating the *a priori* SNR is compared with the DD approach and a reference method from [132]. Instead of using a fixed smoothing factor β, the reference method, ξ_{ref} modifies the DD approach by

SPCH	
Rating	Description
5	No degradation, very natural
4	Little degradation, fairly natural
3	Somewhat degraded, somewhat natural
2	Fairly degraded, fairly unnatural
1	Very degraded, very unnatural

NSE	
Rating	Description
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive

MUSIC	
Rating	Description
5	Not perceptible
4	Somewhat perceptible
3	Perceptible but not annoying
2	Fairly conspicuous, somewhat annoying
1	Very conspicuous, very annoying

Table 4.2: Description of the SPCH, NSE and MUSIC scales used in the listening tests.

using a sigmoid function to control the weighting value for β. Figure 4.2 shows an example of the *a priori* SNR estimation for ξ_{DD}, ξ_{ref} and ξ_{MDD} approaches with different gain functions when speech is detected. It can be observed that the *a priori* SNR for all the gain functions can be represented as the smoothed version of the *a posteriori* SNR. For the conventional DD approach, $\hat{\xi}_{\text{DD}}(k,m)$ follows the $\gamma(k,m)-1$ estimate with one frame delay in speech frames when β is close to 1, ($\beta = 0.98$). Both the reference method and the MDD approach improve the DD approach in terms of reduction in the delays in speech onsets. However, when the speech stops, ξ_{MDD} follows the *a posteriori* SNR estimate but ξ_{ref} gets a one frame delay. Thus, the proposed method is superior in estimating the *a priori* SNR. It is a direct yet effective solution to reduce and eliminate the distortion at speech transients. The same patterns of improvement can be seen

for all the evaluated gain functions. In addition, the proposed method has potentially lower computational complexity when compared to the reference method in [132], which is beneficial for many real-time applications.

4.6.4 Objective Performance Evaluation

Evaluation is performed for both DD and MDD SNR estimators, for different gain functions, which include the WF, LSA, SIG, MSIG-fix1, MSIG-fix2, and MSIG-fix3 methods. The measurements employed are the PESQ measures, the SNRseg measures, and the evaluation metric which include the KurtR, NRR and LLR measures. For PESQ, SNRseg and NRR, higher scores indicate better results and better speech quality. Meanwhile, lower KurtR and LLR scores mean less musical noise and less speech distortion, respectively.

4.6.4.1 PESQ Evaluation

Figures 4.3-4.4 and 4.5-4.6 show the PESQ results for the DD and MDD approaches, respectively, for 0 dB and 15 dB SNRs. The PESQ scores obtained from the WF, LSA and MSIG-fix1 gain functions have a similar trend. They have better results at small values of β and α_y, while having performance drop when both β and α_y are increasing. This is because speech starts to sound unnatural and degraded when more smoothing is applied to the WF, LSA and MSIG-fix1 approaches. In addition, the decreasing rate of the PESQ scores for the MDD approach for an increasing β and α_y is slower than the DD approach, resulting in a better PESQ results for the MDD approach when both β and α_y are large values.

For the SIG function, while the DD approach follows the previously described trend, the MDD approach has a different trend. At 0 dB SNR, the SIG function with the MDD approach has better PESQ results for all β values at a large α_y. While at 15 dB SNR, it has the optimal PESQ scores at a smaller α_y, of which the values are the same for all β. This indicates that even with the same values of β and α_y, the amount of smoothing varies for different gain functions. Since the MDD approach provides better PESQ scores at most β values, particularly for $\beta > 0.94$, it is the preferred choice for SNR estimate when compared to the

DD approach.

The contour shape obtained from MSIG-fix2 is similar to MSIG-fix3, but totally different from WF, LSA, MSIG-fix1 and SIG. Both of them have better PESQ results recorded when β and α_y are large. This is because both MSIG-fix2 and MSIG-fix3, together with the SIG function, are non-aggressive gain functions. As shown in Figure 4.1, they do not provide much noise suppression at low SNR. Therefore, by providing more smoothing to the SNR estimates for these two gain functions help to reduce noise variations, which leads to better PESQ scores. In terms of the comparison between the SNR estimates, MSIG-fix2 has better PESQ scores in conjunction with the MDD approach when compared to the DD approach, while MSIG-fix3 has better PESQ scores when the DD approach is employed.

As observed from the figures, both the WF and MSIG-fix1 have the best PESQ results among all gain functions, with the MDD approach having the least parameter sensitivity over a wide range of parameters and the best scores for 15 dB global SNR conditions. By taking this into account with the observation that better PESQ results have been obtained for the MDD approach for increasing β and α_y, MDD performs better than DD, particularly for WF, LSA and MSIG-fix1 in terms of the PESQ measure.

4.6.4.2 SNRseg Evaluation

Figures 4.7-4.10 present the SNRseg results for the DD and MDD approaches with 0 dB and 15 dB SNRs. All evaluated gain functions give better SNRseg results for the MDD approach when compared with the DD approach for large value of β. In particular, the MDD approach has a significant improvement over the DD approach for the WF, LSA, MSIG-fix1 and SIG gain functions for all β and α_y. This indicates that the segmental SNR increases when the delay in speech transients is reduced and removed. Also, the WF and MSIG-fix1 perform best when β and α_y are small, while LSA has better SNRseg results at high smoothing setting.

For MSIG-fix2 at 0 dB SNR, apart from having the optimal point at different smoothing parameters, the SNRseg results for both DD and MDD approaches are

very similar. For MSIG-fix3, it has poorer SNRseg scores for the MDD approach at 0 dB when compared to the DD approach. Despite that, at 15 dB SNR both MSIG-fix2 and MSIG-fix3 give better SNRseg results for the MDD approach. For SIG function, the MDD approach has similar performance with the DD approach when less smoothing is applied, but it becomes better than the DD approach for increasing β and α_y.

4.6.4.3 Objective Speech Distortion, Musical Noise and Noise Suppression Evaluation

Figures 4.11-4.12 and 4.13-4.14 present the results from the objective evaluation metric for the DD and MDD approaches at 0 dB and 15 dB SNRs, respectively. In terms of the amount of musical noise generated, KurtR decreases with increasing values of β and α_y for all evaluated gain functions with one exception (MSIG-fix3). The MSIG-fix3 function is a rather flat function over a range of input SNRs that lacks distinctiveness resulting in poor performance in terms of musical noise. For the rest of the evaluated gain functions, the DD approach performs better than MDD for KurtR with a few exceptions, which is due to the higher sensitivity to changes in MDD that provides capability to track speech onsets. For WF and MSIG-fix1, both the DD and MDD approaches have almost identical results for the KurtR measure at $\beta > 0.98$. Meanwhile, at 15 dB SNR, there is a significant improvement in performance for the MDD approach over the DD approach in conjunction with both WF and MSIG-fix1 functions when β is large. These results are the best among all evaluated gain functions. Since less musical noise is generated with large smoothing parameters, the MDD approach is a better choice for both WF and MSIG-fix1 approaches.

While WF and MSIG-fix1 are the best gain functions to be used for the MDD approach in terms of the KurtR measure, LSA performs the best for the DD approach with large smoothing parameters. On the other hand, for less smoothing applied, SIG is the best gain functions among all. Apart from that, MSIG-fix2 and MSIG-fix3 perform worst in KurtR results among all evaluated gain functions. This is because both gain functions are less aggressive and are not directly propositional towards the *a priori* SNR as shown in Figure 4.1. There

is a drop in gain values between $\xi = -5\,\text{dB}$ and $\xi = 0\,\text{dB}$, which should be the main factor that isolated spectral components were formed after the processing.

The results from the NRR measure are almost inversely proportional to the KurtR measure. All evaluated gain functions show poorer performance for the MDD approach when compared to the DD approach. This can be explained as MDD has more variations in noise when compared to the DD approach. However, the remaining residual noise in the enhanced signal helps to mask the musical noise. This acts as a good compromise between the amount of musical noise and noise reduction for the MDD approach.

In terms of the amount of speech distortion generated, LLR is almost directly proportional to the NRR measure. From the figures, a small LLR can be obtained for a decreasing β and α_y. At small β and α_y values, the DD approach performs better than the MDD approach. Meanwhile, at large β and α_y values, the MDD approach generally gives less speech distortion when compared to the DD approach. This indicates that with a large smoothing parameters, the MDD approach performs better than the DD approach, particularly for the WF, LSA and MSIG-fix1 gain functions.

4.6.5 Subjective Listening Tests

Subjective listening tests were performed to validate the results from the objective measures. The description of the tests setup can be found in Section 4.6.2. Prior to the subjective tests, appropriate choice of parameters β and α_y has to be determined.

4.6.5.1 Selection of Smoothing Parameters

Tables 4.3 and 4.4 summarise the best smoothing parameters for different gain functions and different SNR levels. As can be seen in the tables, a small value for β is preferred for lower speech distortion in the case of WF, LSA, SIG and MSIG-fix1 gain functions. This is validated from the LLR, PESQ and SNRseg results. However, there is a trade-off because of the resulting high level of annoying musical noise in the output when both β and α_y are small. Also, small values in β

and α_y give low NRR values, which is not desirable in most situations. In addition, an increment of both β and α_y does not give a direct drop in speech quality. This means that by choosing appropriate values for both smoothing parameters, a balanced trade-off can be obtained. Meanwhile, for MSIG-fix2 and MSIG-fix3, better PESQ, SNRseg and LLR results are recorded at larger values for β and α_y. This motivates the use of MSIG functions as they are able to show similar preferred smoothing parameters from all the objective measurement results.

The last columns in Tables 4.3 and 4.4 also show the smoothing parameters chosen for the listening test. The value β for each gain function and each SNR condition was chosen such that the smallest value was selected so that the KurtR was at its minimum value. This is to keep the speech distortion as low as possible while having the lowest possible level of audible musical noise. The value α_y was chosen as the mean of α_y values from all the objective measures. This have been found to be a reasonable compromise between the aforementioned trade-offs.

4.6.5.2 Objective Measurement with Selected Smoothing Parameters

Objective measurement was also performed for the selected parameters listed in Tables 4.3 and 4.4, using the NOIZEUS database. Tables 4.5 lists the average results for pink noise, where Δ indicates the improvement between the enhanced signals and the noisy signals. According to the table, WF, LSA, and MSIG-fix1 have better performance particularly in terms of noise reduction as reflected by the results from SNRseg and NRR measures. When comparing both DD and MDD *a priori* SNR estimates, WF and MSIG-fix1 perform best with MDD approach, as indicated by larger NRR results while having smaller KurtR values. While for LSA, DD has slightly less musical noise when compared to MDD, with approximately similar NRR results. For MSIG-fix2 and MSIG-fix3, their results indicate less noise reduction when compared to WF, LSA and MSIG-fix1. In terms of KurtR measure, MSIG-fix2 with DD approach has similar amount of musical noise when compared to WF and MSIG-fix1. For MSIG-fix2 with MDD approach and MSIG-fix3 with both DD and MDD approaches, KurtR results are large. In terms of the least amount of musical noise generated, the SIG function has the smallest KurtR values, but with the least amount of noise reduction

from SNRseg and NRR measures. As for the amount of speech distortion, all evaluated gain functions have almost similar performance, as shown by PESQ and LLR scores, except for SIG with slightly poorer performance when compared to others.

Table 4.6 shows the average results for the factory noise. The results are similar to the results from pink noise, except for some cases from the KurtR values. As observed in 15 dB SNR, the DD approach has smaller KurtR results for all evaluated gain functions when compared to the MDD approach for factory noise. Here, MSIG-fix3 has the smallest KurtR results, which is similar to the results from the SIG function. However, since MSIG-fix3 has larger SNRseg and similarly small NRR results when compared with the SIG function, more distorted noise would be perceived as musical noise.

4.6.5.3 Evaluation for Listening Tests

Tables 4.7 and 4.8 tabulate the average results of the subjective listening tests, and also the overall performance of each gain function by taking the average of SPCH, NSE and MUSIC. From the overall scales, it can be observed that although the difference is not that significant, the listeners preferred the signals with WF and MSIG-fix1 gain functions, both combined with the MDD approach. The results are consistent for both noise types. For the LSA method, the overall performance between the DD and the MDD approach are almost identical, while LSA with DD approach has recorded the least amount of musical noise particularly in the pink noise. For the MSIG-fix2 and MSIG-fix3 functions, the audible musical noise is more prominent which is reflected in the MUSIC column. This trend is particularly obvious with the MDD approach, which aligns with the objective results. As for the SIG function, there is less musical noise in the output, but also a very small amount of noise suppression. The MDD approach helps to reduce audible noise and increase noise reduction for aggressive gain functions such as WF and MSIG-fix1. While for other less aggressive gain functions, the DD approach can be sufficient. As for factory noise, particularly at 15 dB SNR, the subjective results are almost the same for each *a priori* SNR estimate and every gain function. The suggested reason behind this is that the factory noise is

less intrusive than the pink noise. Thus, less noise with less audible musical noise was perceived when the noise floor was fixed at $-15\,\mathrm{dB}$ for the gain functions.

4.7 Summary

In this chapter, a new MSIG has been developed to provide flexibility to the gain function that can be optimised to match various criteria to achieve a compromised trade-off among speech distortion, noise reduction and musical noise. In addition, a new approach to estimate the *a priori* SNR has been proposed for the MDD, which reduces and eliminates the speech transient distortion. The musical noise has been further reduced by applying more smoothing to the *a posteriori* SNR by using a recursive averaging algorithm. As such, the level of smoothing is controlled by the parameters β and α_y. Performance evaluation shows that the proposed MDD performs better than the traditional DD. The MSIG-fix1 function has similar performance compared to the conventional gain functions. At large smoothing parameters, MSIG-fix2 and MSIG-fix3 generate the lowest speech distortion among all evaluated gain functions. Finally, subjective listening tests verify the findings from the objective measurements and give confidence that the chosen objective measures reflect the true subjective experience. The proposed algorithms and the findings in the performance measures help to develop a promising binaural speech enhancement algorithm, which will be discussed in Chapter 7.

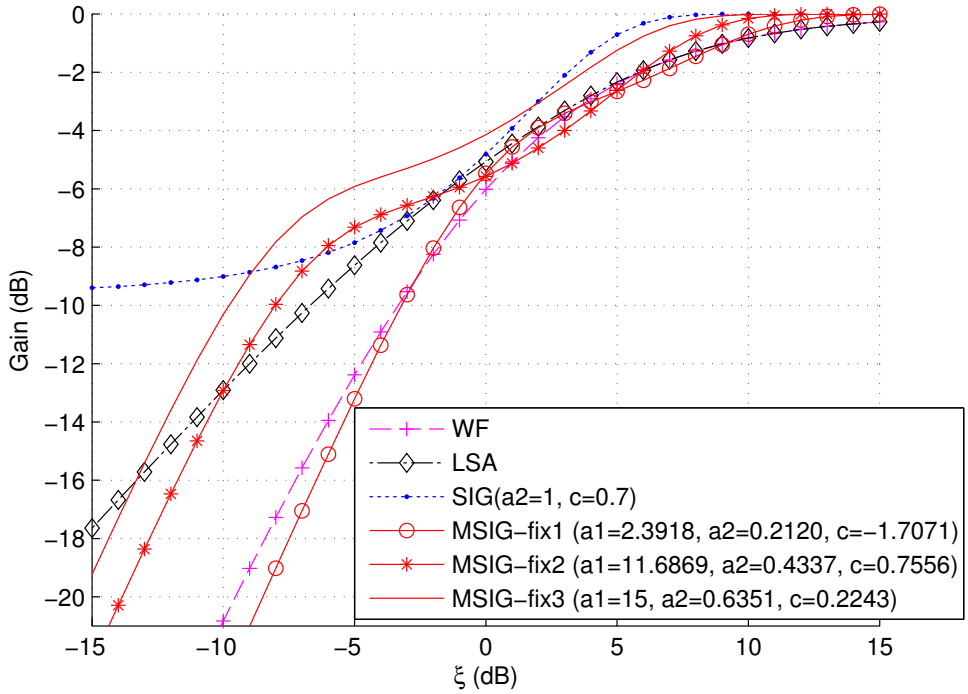

Figure 4.1: Gain curves of different algorithms, as functions of the *a priori* SNR $\xi(k,m)$, where $\gamma(k,m) = \xi(k,m) + 1$.

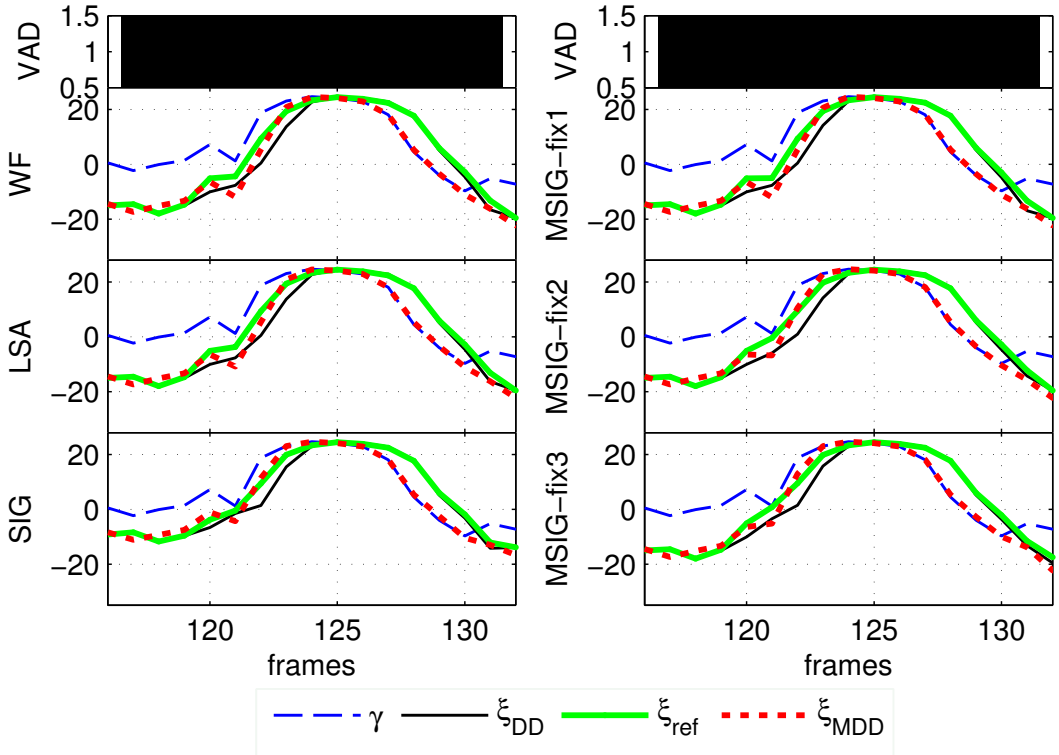

Figure 4.2: Comparison between MDSVAD decisions, $\gamma(k,m) - 1$ (blue dashed line), $\hat{\xi}_{\mathrm{DD}}(k,m)$ (black solid line), $\hat{\xi}_{\mathrm{ref}}(k,m)$ (green solid line) and $\hat{\xi}_{\mathrm{MDD}}(k,m)$ (red dotted line) at 937.5 Hz and 15 dB SNR. Here, $\beta = 0.98$ were applied for $\hat{\xi}_{\mathrm{DD}}(k,m)$ and $\hat{\xi}_{\mathrm{MDD}}(k,m)$, while $\alpha_y = 0.3$ were employed for all evaluated *a priori* SNR estimators.

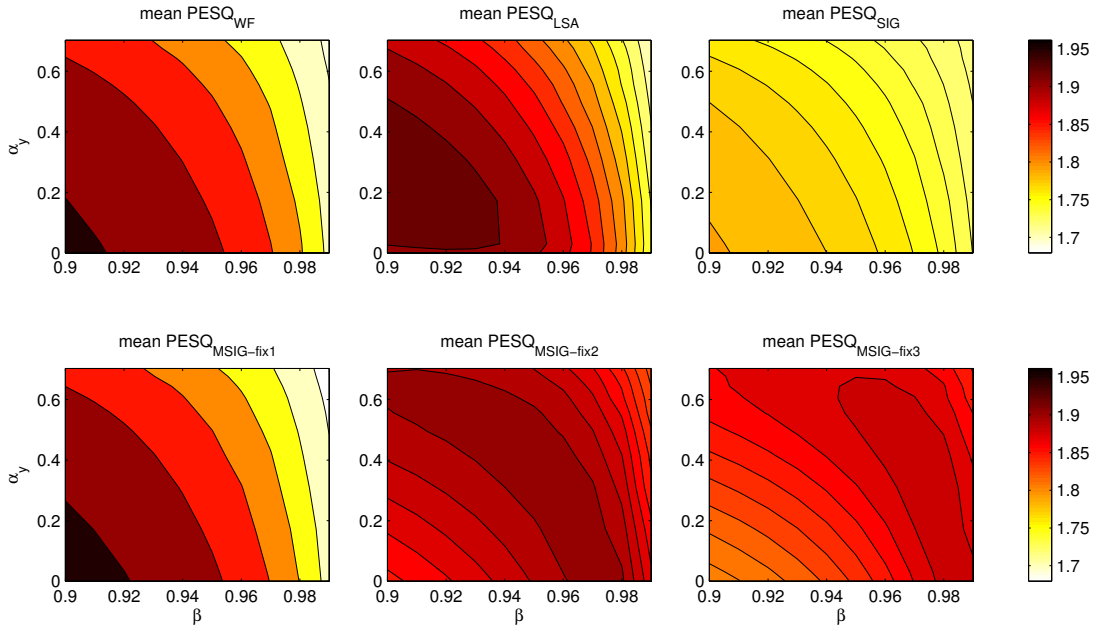

Figure 4.3: Average PESQ scores with $\hat{\xi}_{\mathrm{DD}}$ at SNR $= 0\,\mathrm{dB}$.

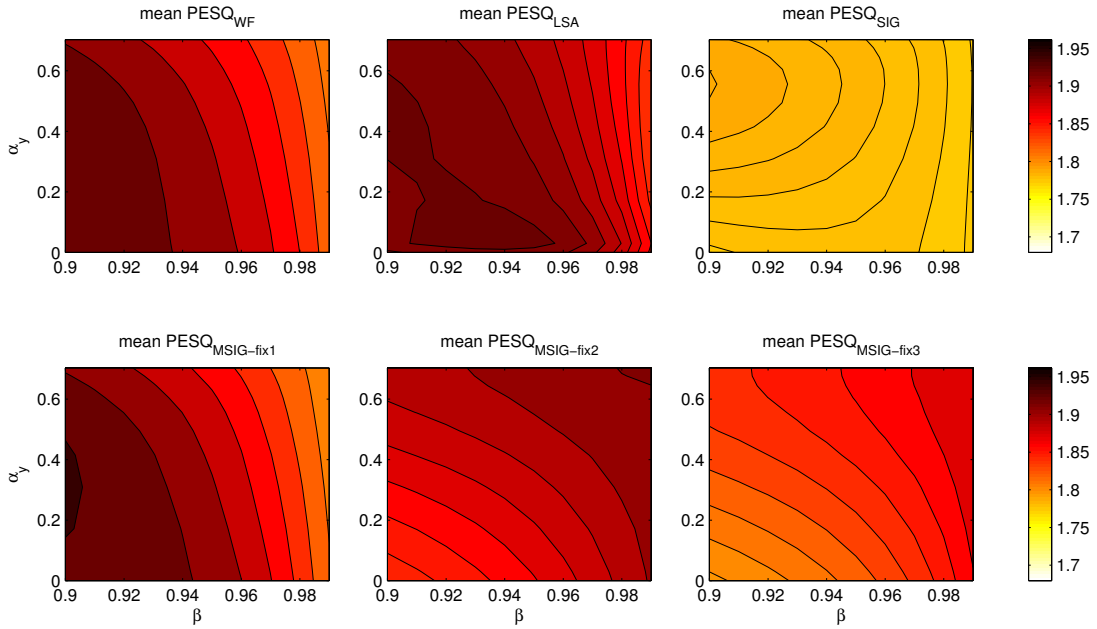

Figure 4.4: Average PESQ scores with $\hat{\xi}_{\mathrm{MDD}}$ at SNR $= 0\,\mathrm{dB}$.

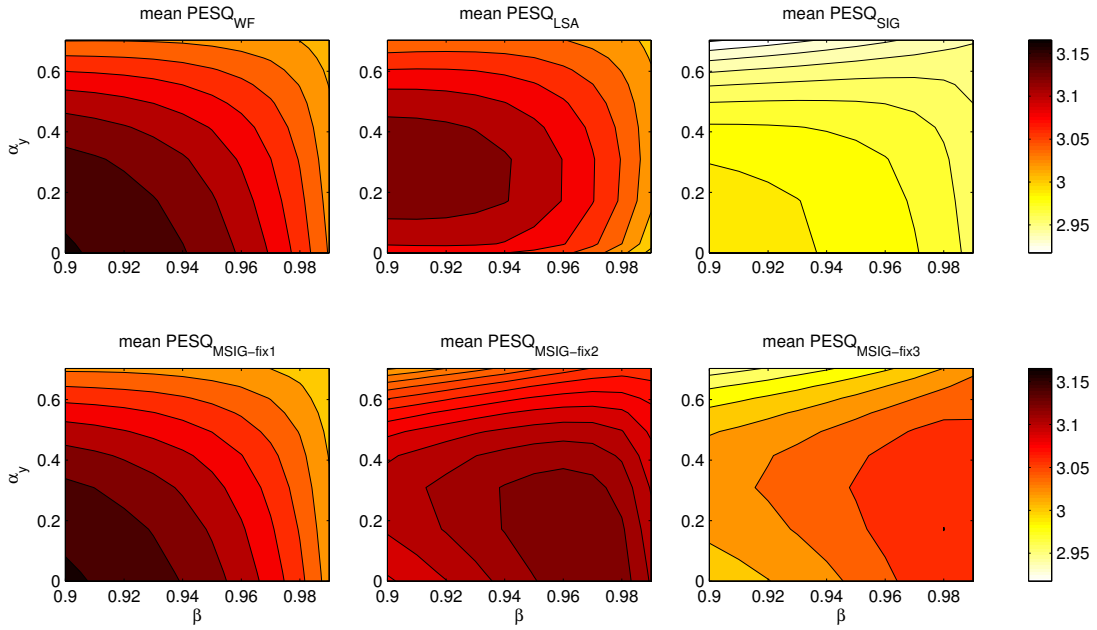

Figure 4.5: Average PESQ scores with $\hat{\xi}_{\mathrm{DD}}$ at SNR $= 15\,\mathrm{dB}$.

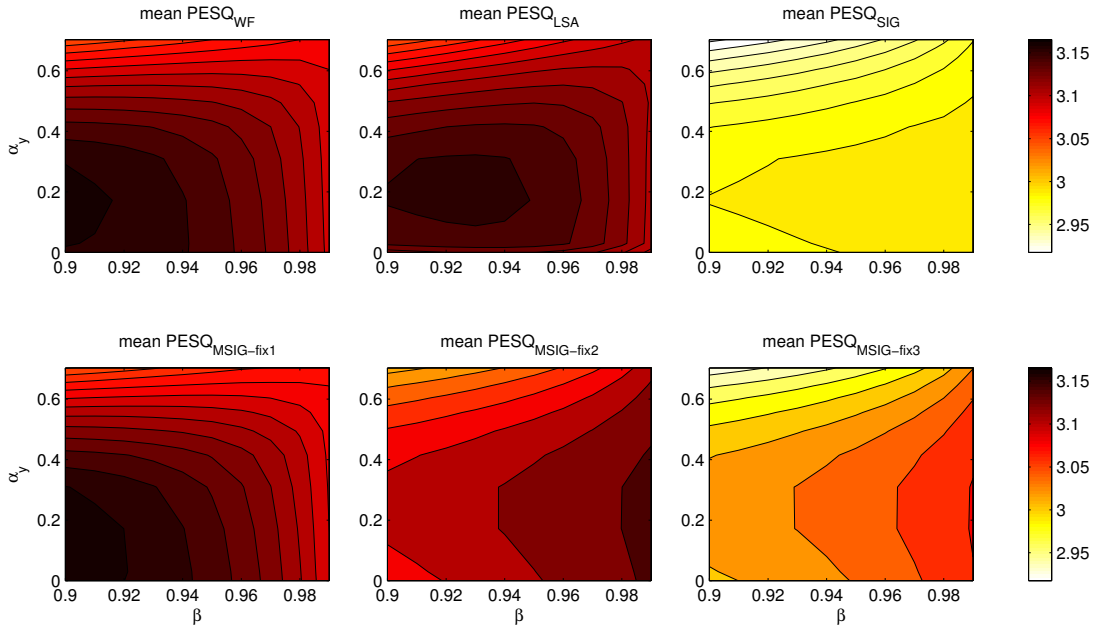

Figure 4.6: Average PESQ scores with $\hat{\xi}_{\mathrm{MDD}}$ at SNR $= 15\,\mathrm{dB}$.

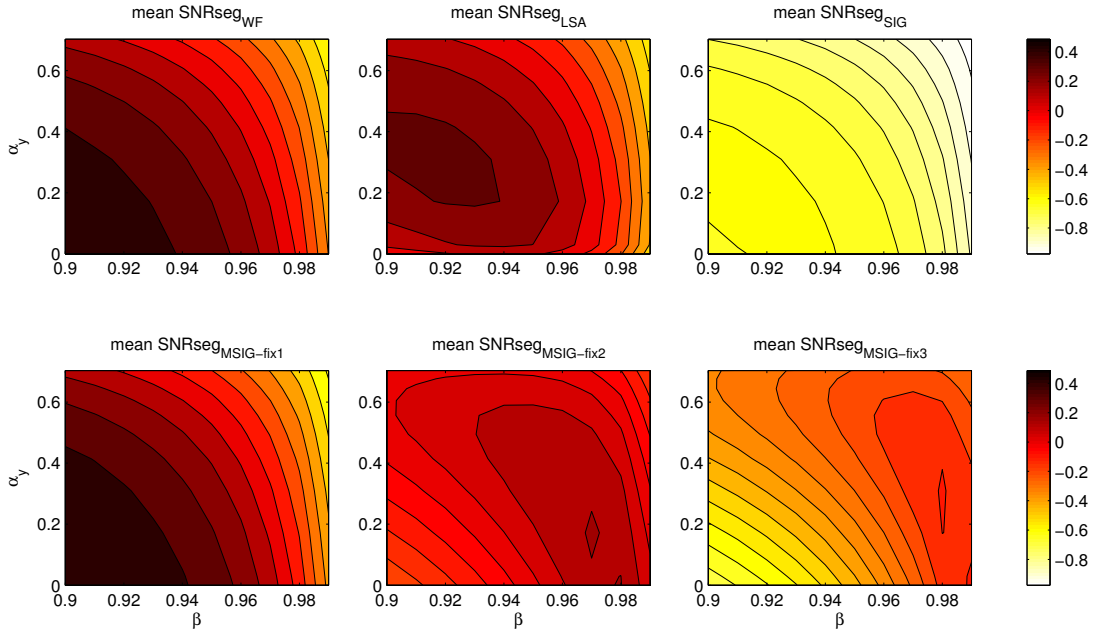

Figure 4.7: Average SNRseg values with $\hat{\xi}_{\mathrm{DD}}$ at SNR $= 0\,\mathrm{dB}$.

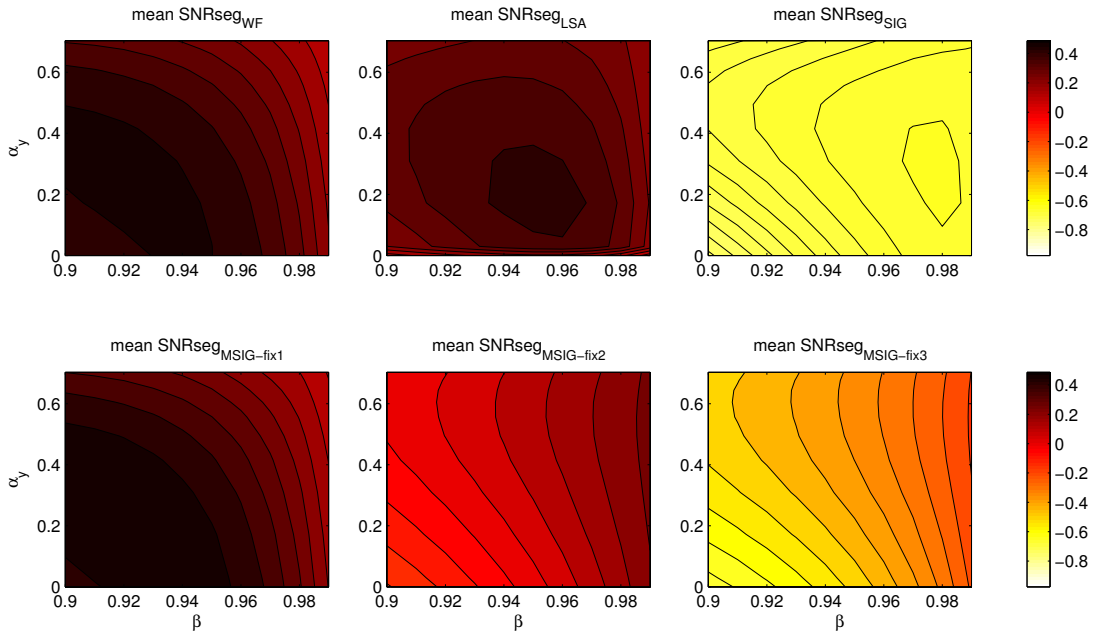

Figure 4.8: Average SNRseg values with $\hat{\xi}_{\mathrm{MDD}}$ at SNR $= 0\,\mathrm{dB}$.

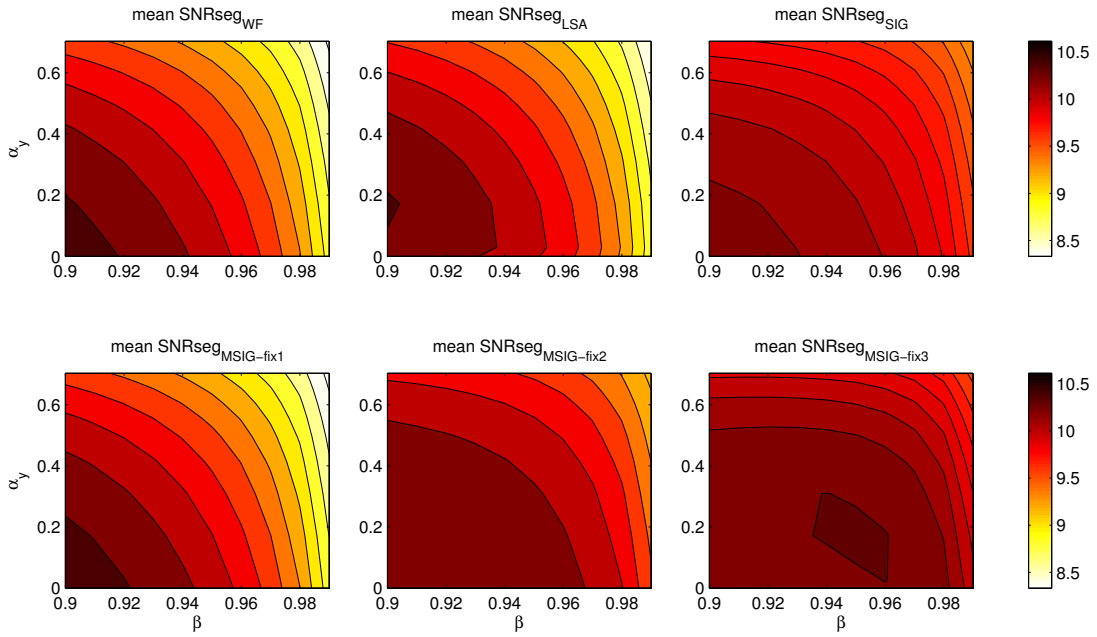

Figure 4.9: Average SNRseg values with $\hat{\xi}_{\mathrm{DD}}$ at SNR $= 15\,\mathrm{dB}$.

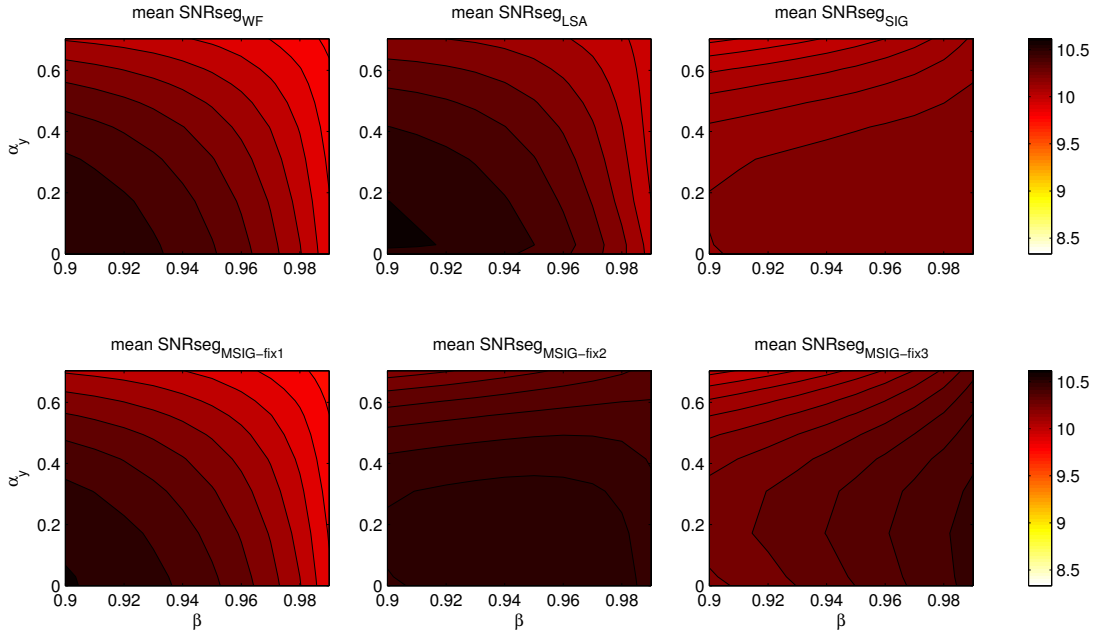

Figure 4.10: Average SNRseg values with $\hat{\xi}_{\mathrm{MDD}}$ at SNR $= 15\,\mathrm{dB}$.

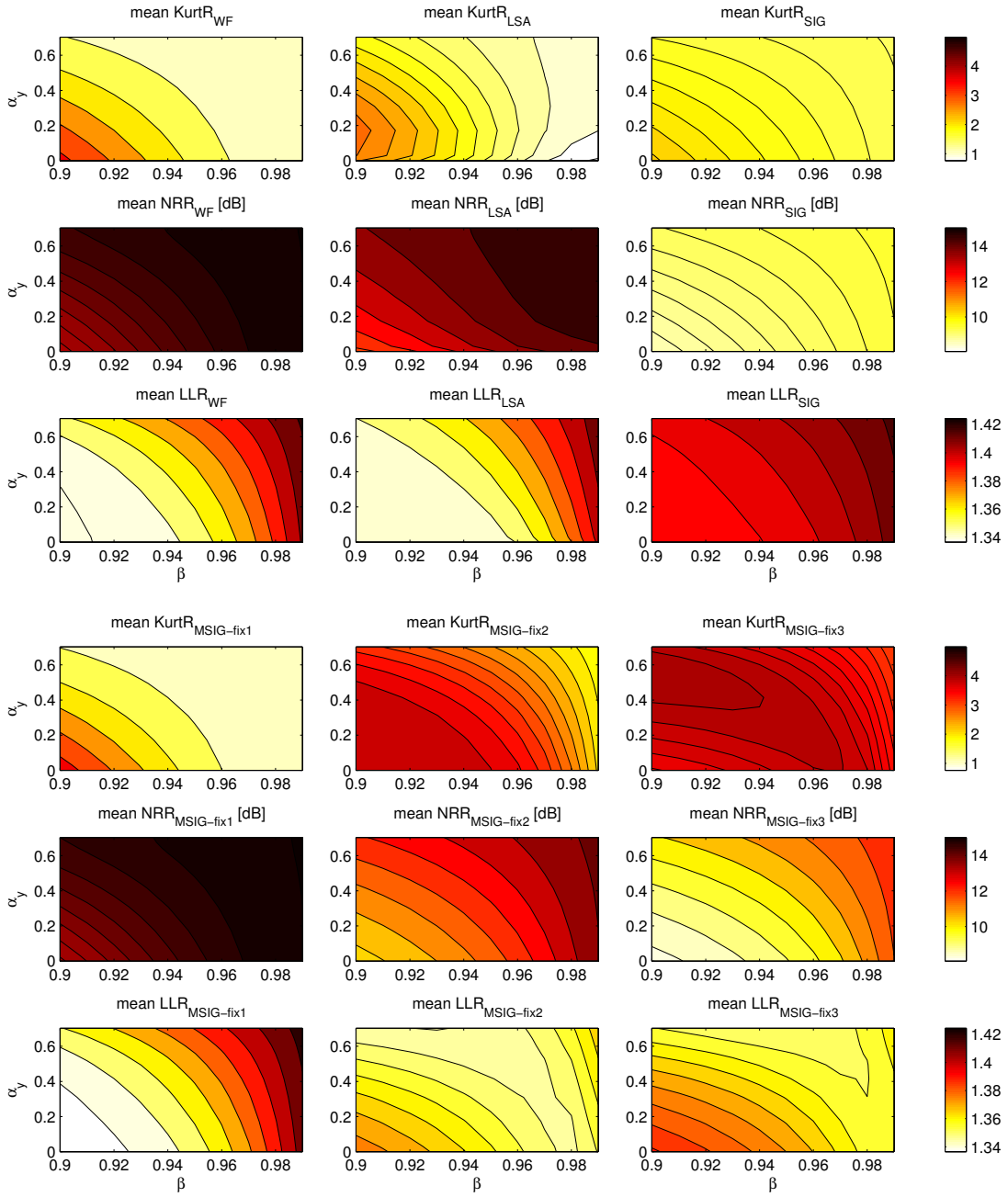

Figure 4.11: Average results for KurtR, NRR and LLR measures with $\hat{\xi}_{\mathrm{DD}}$ at SNR $= 0\,\mathrm{dB}$.

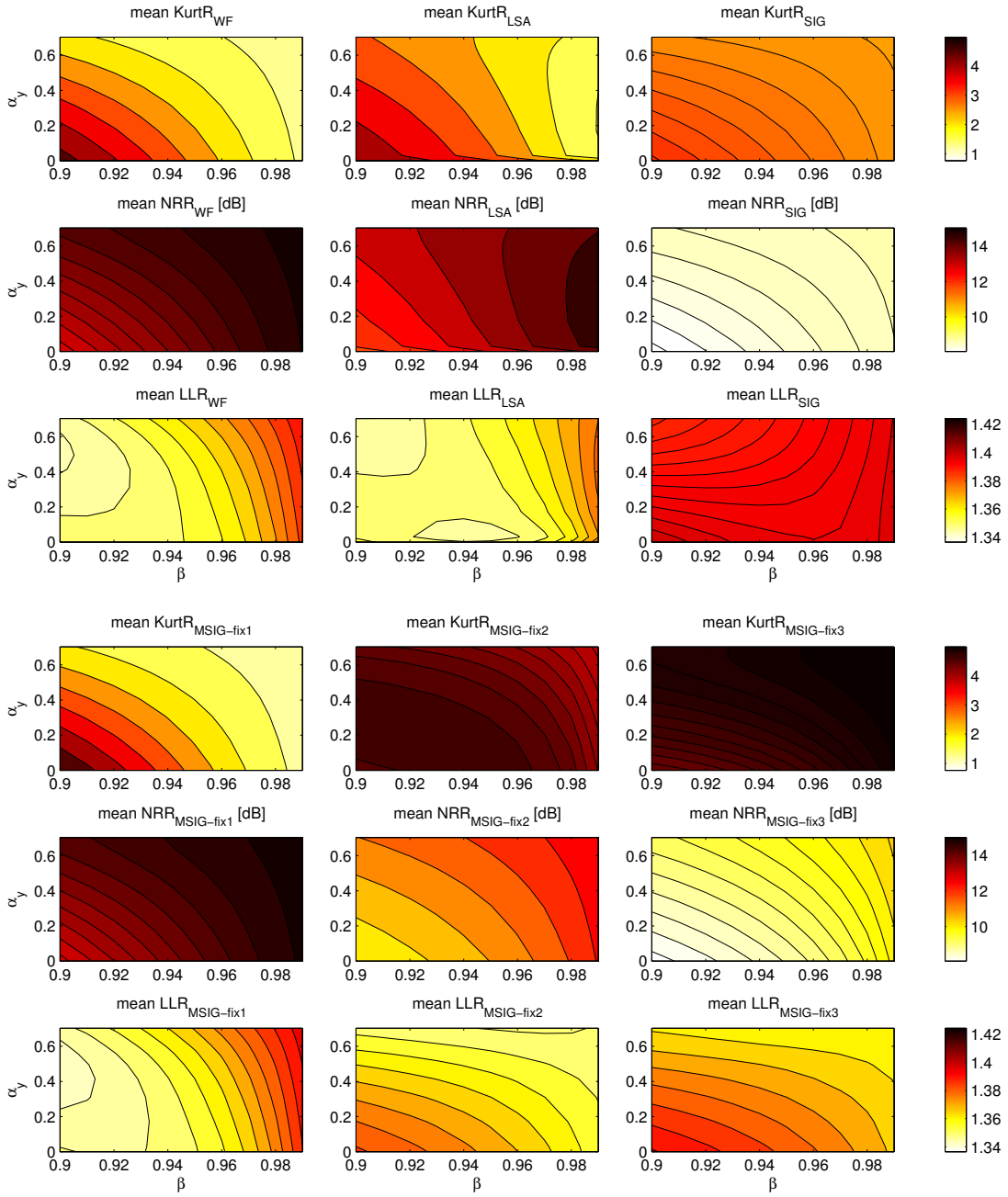

Figure 4.12: Average results KurtR, NRR and LLR measures with $\hat{\xi}_{\mathrm{MDD}}$ at SNR $= 0\,\mathrm{dB}$.

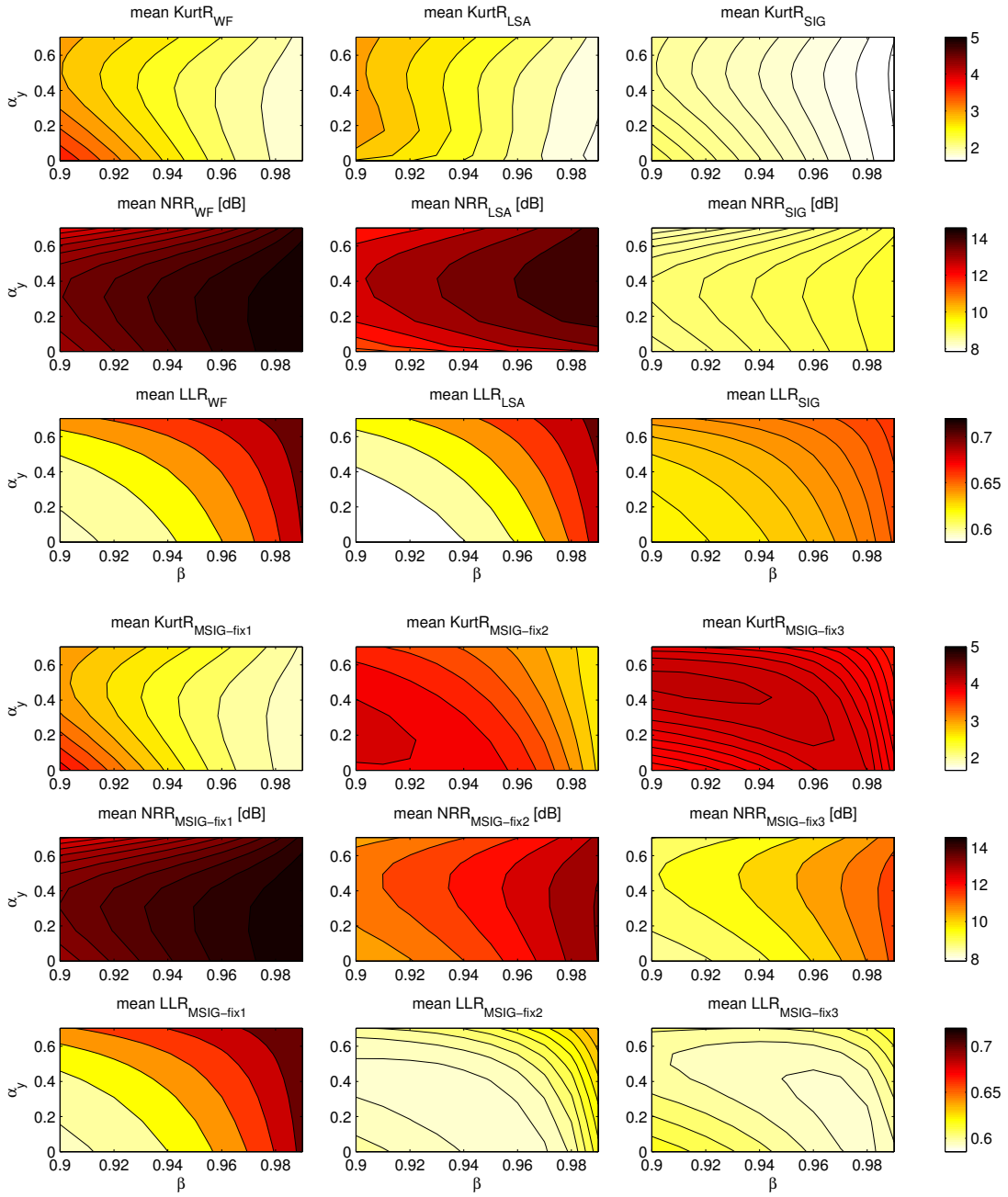

Figure 4.13: Average results KurtR, NRR and LLR measures with $\hat{\xi}_{\mathrm{DD}}$ at SNR = 15 dB.

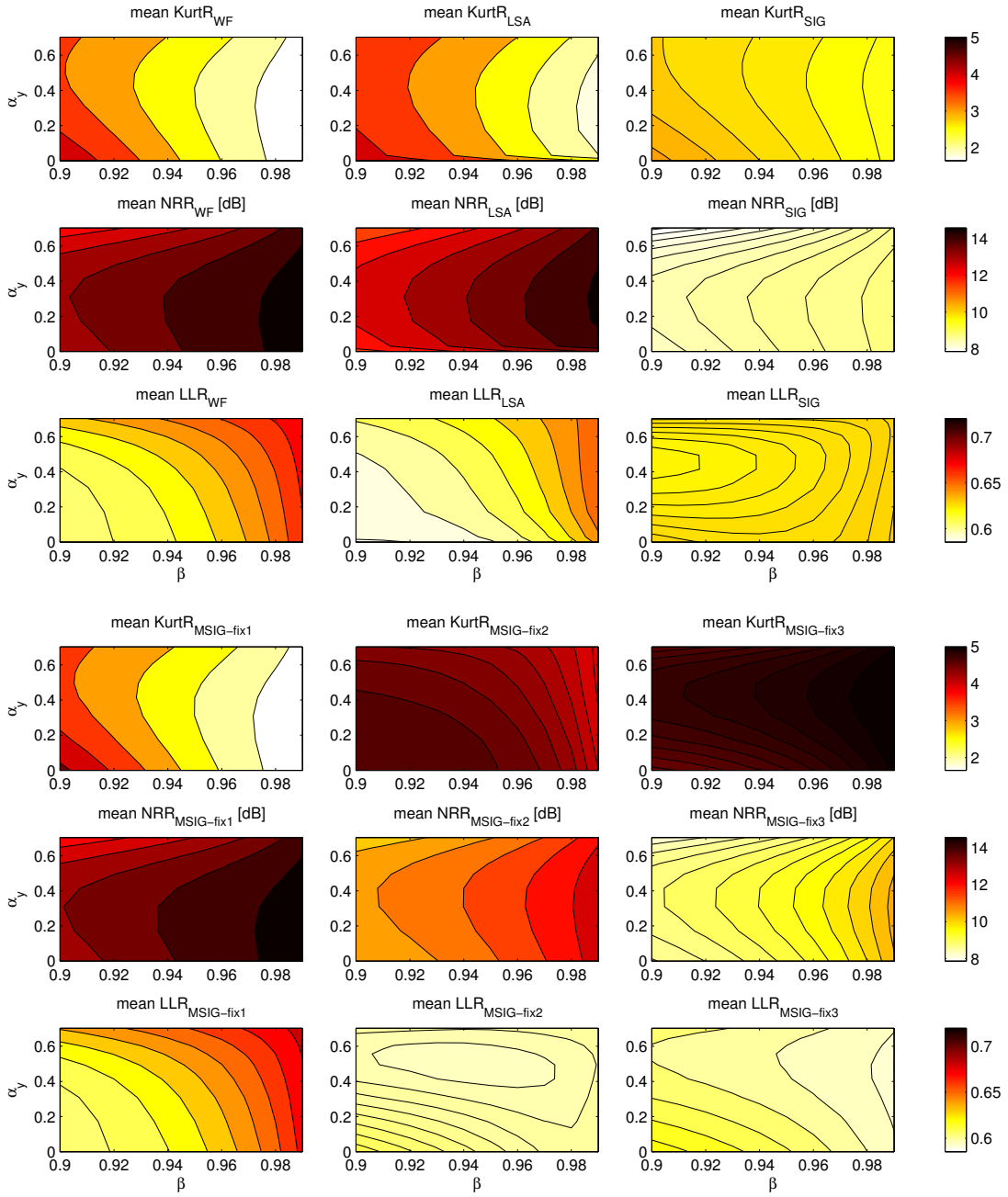

Figure 4.14: Average results KurtR, NRR and LLR measures with $\hat{\xi}_{\mathrm{MDD}}$ at SNR = 15 dB.

		SNR	PESQ	SNRseg	KurtR	NRR	LLR	Test
WF	β	0 dB	0.90	0.90	0.91	0.94	0.90	0.91
		15 dB	0.90	0.90	0.98	0.95	0.90	
	α_y	0 dB	0-0.2	0-0.4	0.7	0.7	0-0.3	0.35
		15 dB	0	0-0.2	0-0.5	0.2-0.3	0-0.2	
LSA	β	0 dB	0.90	0.90	0.97	0.95	0.90	0.95
		15 dB	0.90	0.90	0.99	0.96	0.90	
	α_y	0 dB	0.1-0.5	0.3-0.4	0	0.6-0.7	0-0.7	0.39
		15 dB	0.1-0.4	0.2	0-0.2	0.3-0.4	0-0.4	
SIG	β	0 dB	0.90	0.90	0.99	0.98	0.90	0.98
		15 dB	0.90	0.90	0.99	0.98	0.90	
	α_y	0 dB	0-0.1	0-0.1	0.6	0.6-0.7	0-0.6	0.40
		15 dB	0-0.3	0-0.2	0.2-0.6	0-0.5	0-0.3	
MSIG-fix1	β	0 dB	0.90	0.90	0.90	0.94	0.90	0.90
		15 dB	0.90	0.90	0.98	0.97	0.90	
	α_y	0 dB	0-0.2	0-0.4	0.7	0.6-0.7	0-0.4	0.35
		15 dB	0	0-0.2	0-0.5	0.1-0.3	0-0.1	
MSIG-fix2	β	0 dB	0.90	0.93	0.98	0.99	0.90	0.98
		15 dB	0.94	0.90	0.98	0.99	0.90	
	α_y	0 dB	0.6-0.7	0.5	0.7	0.2-0.7	0.6-0.7	0.52
		15 dB	0.2-0.3	0-0.5	0.7	0-0.4	0	
MSIG-fix3	β	0 dB	0.95	0.96	0.99	0.97	0.90	0.97
		15 dB	0.95	0.94	0.99	0.99	0.90	
	α_y	0 dB	0.5-0.6	0.5-0.6	0.6-0.7	0.6-0.7	0.7	0.52
		15 dB	0.3	0.2-0.3	0.6-0.7	0-0.6	0	

Table 4.3: Approximated best smoothing values for different gain functions with DD approach.

		SNR	PESQ	SNRseg	KurtR	NRR	LLR	Test
WF	β	0 dB	0.90	0.90	0.96	0.98	0.90	0.96
		15 dB	0.90	0.90	0.98	0.98	0.90	
	α_y	0 dB	0-0.6	0.2-0.4	0.7	0.6	0.4-0.6	0.49
		15 dB	0-0.3	0-0.3	0-0.6	0-0.4	0-0.4	
LSA	β	0 dB	0.90	0.94	0.97	0.98	0.90	0.97
		15 dB	0.90	0.90	0.98	0.99	0.90	
	α_y	0 dB	0.3-0.6	0.2-0.3	0.5	0.2-0.4	0.4-0.6	0.36
		15 dB	0.2	0-0.2	0.2-0.4	0.2-0.4	0	
SIG	β	0 dB	0.90	0.97	0.98	0.97	0.90	0.97
		15 dB	0.90	0.90	0.98	0.98	0.90	
	α_y	0 dB	0.6	0.2-0.4	0.7	0.6	0.7	0.49
		15 dB	0.2	0.1-0.2	0.2-0.6	0.1-0.4	0.4-0.5	
MSIG-fix1	β	0 dB	0.90	0.90	0.96	0.97	0.90	0.96
		15 dB	0.90	0.90	0.98	0.98	0.90	
	α_y	0 dB	0.2-0.4	0-0.5	0.6	0.7	0.3-0.6	0.48
		15 dB	0-0.3	0-0.3	0-0.6	0-0.4	0-0.4	
MSIG-fix2	β	0 dB	0.93	0.99	0.99	0.97	0.94	0.97
		15 dB	0.98	0.90	0.99	0.98	0.91	
	α_y	0 dB	0.7	0.2-0.7	0.7	0.7	0.7	0.56
		15 dB	0.2-0.3	0-0.2	0.4-0.7	0.3	0.5-0.6	
MSIG-fix3	β	0 dB	0.97	0.99	0.90	0.99	0.90	0.90
		15 dB	0.99	0.98	0.90	0.99	0.99	
	α_y	0 dB	0.6-0.7	0.4-0.7	0	0.7	0.7	0.50
		15 dB	0.1-0.3	0.2	0.7	0.1-0.4	0.1-0.6	

Table 4.4: Approximated best smoothing values for different gain functions with MDD approach.

Signal	ΔPESQ		ΔSNRseg		KurtR		NRR		ΔLLR	
	DD	MDD	DD	MDD	DD	MDD	DD	MDD	DD	MDD
0 dB SNR										
WF	0.5	0.5	3.9	3.9	2.3	1.8	14.2	14.6	-0.2	-0.2
LSA	0.5	0.5	3.7	3.9	1.5	2.1	14.4	14.3	-0.2	-0.2
SIG	0.3	0.4	2.7	2.9	1.6	2.6	9.5	8.7	-0.1	-0.1
MSIG-fix1	0.5	0.5	4.0	3.9	2.5	1.7	14.1	14.6	-0.2	-0.2
MSIG-fix2	0.5	0.5	3.6	3.7	2.2	4.3	13.8	12.3	-0.2	-0.2
MSIG-fix3	0.5	0.4	3.4	3.1	3.7	4.8	11.5	9.0	-0.2	-0.2
15 dB SNR										
WF	0.6	0.6	3.5	3.4	2.9	2.3	13.7	14.1	-0.3	-0.3
LSA	0.6	0.6	3.1	3.5	2.3	2.3	13.9	14.1	-0.3	-0.3
SIG	0.5	0.5	3.0	3.4	1.7	2.5	9.3	8.7	-0.2	-0.3
MSIG-fix1	0.6	0.6	3.5	3.4	3.2	2.3	13.6	14.1	-0.3	-0.2
MSIG-fix2	0.6	0.6	2.9	3.6	2.9	4.3	13.1	12.0	-0.3	-0.3
MSIG-fix3	0.5	0.5	3.3	3.4	4.0	4.8	10.9	8.9	-0.3	-0.3

Table 4.5: Objective results for pink noise with selected parameters.

Signal	ΔPESQ		ΔSNRseg		KurtR		NRR		ΔLLR	
	DD	MDD	DD	MDD	DD	MDD	DD	MDD	DD	MDD
0 dB SNR										
WF	0.6	0.6	4.6	4.5	4.1	4.1	13.9	14.3	-0.1	-0.1
LSA	0.6	0.6	4.5	4.6	3.4	4.0	14.2	14.1	-0.1	-0.1
SIG	0.4	0.4	3.4	3.4	2.2	3.1	9.8	9.0	-0.1	-0.2
MSIG-fix1	0.6	0.6	4.6	4.5	4.2	4.1	13.8	14.4	-0.1	-0.1
MSIG-fix2	0.6	0.6	4.4	4.5	3.4	4.1	13.4	12.2	-0.1	-0.2
MSIG-fix3	0.5	0.5	4.3	4.0	3.4	3.9	11.4	9.3	-0.2	-0.2
15 dB SNR										
WF	0.5	0.5	5.0	4.9	3.8	4.2	12.3	12.5	0.1	0.1
LSA	0.5	0.6	4.7	4.9	3.6	4.3	12.4	13.0	0.1	0.1
SIG	0.4	0.4	4.3	4.4	2.4	2.9	9.1	8.5	0	0
MSIG-fix1	0.5	0.5	5.0	4.9	3.8	4.3	12.1	12.5	0.1	0.1
MSIG-fix2	0.5	0.5	4.6	5.1	3.3	3.4	11.6	10.7	0.1	0
MSIG-fix3	0.4	0.4	5.0	5.0	2.8	2.9	9.8	8.3	0	0

Table 4.6: Objective results for factory noise with selected parameters.

SNR	Signal	SPCH		NSE		MUSIC		Overall	
		DD	MDD	DD	MDD	DD	MDD	DD	MDD
	Noisy	3.0		1.0		5.0		2.98	
	WF	3.5	3.6	2.6	2.8	3.4	3.5	3.16	3.28
	LSA	3.3	3.5	2.2	2.4	4.1	3.8	3.21	3.22
0 dB	SIG	3.0	3.2	1.4	1.5	4.6	4.1	2.97	2.92
	MSIG-fix1	3.7	3.7	2.9	2.9	3.1	3.5	3.20	3.35
	MSIG-fix2	3.8	4.0	3.0	3.7	2.5	1.7	3.08	3.12
	MSIG-fix3	4.1	4.2	4.0	4.2	1.3	1.0	3.12	3.12
	Noisy	4.0		1.6		5.0		3.51	
	WF	4.6	4.6	3.5	3.6	3.7	4.0	3.91	4.09
	LSA	4.4	4.4	3.1	3.3	4.2	4.1	3.91	3.94
15 dB	SIG	4.1	4.2	2.1	2.4	4.8	4.1	3.66	3.54
	MSIG-fix1	4.7	4.9	3.8	3.7	3.6	4.0	3.99	4.18
	MSIG-fix2	4.9	4.9	3.8	4.3	3.4	2.5	4.00	3.86
	MSIG-fix3	4.9	4.9	4.4	4.5	2.4	1.5	3.86	3.58

Table 4.7: Subjective results for pink noise.

SNR	Signal	SPCH		NSE		MUSIC		Overall	
		DD	MDD	DD	MDD	DD	MDD	DD	MDD
	Noisy	3.9		1.7		5.0		3.52	
	WF	4.6	4.6	3.5	3.6	4.5	4.6	4.18	4.26
	LSA	4.4	4.5	3.3	3.3	4.7	4.6	4.14	4.14
0 dB	SIG	4.1	4.2	2.4	2.4	4.9	4.6	3.79	3.73
	MSIG-fix1	4.7	4.7	3.6	3.6	4.3	4.5	4.18	4.28
	MSIG-fix2	4.7	4.7	3.7	3.9	4.0	3.1	4.13	3.90
	MSIG-fix3	4.7	4.7	3.9	4.0	2.6	2.1	3.75	3.62
	Noisy	4.5		2.8		5.0		4.09	
	WF	4.9	4.9	4.1	4.1	4.9	4.9	4.61	4.62
	LSA	4.9	4.9	4.0	4.1	4.9	4.9	4.58	4.59
15 dB	SIG	4.7	4.7	3.4	3.4	4.9	4.7	4.33	4.29
	MSIG-fix1	4.9	4.9	4.1	4.1	4.8	4.9	4.60	4.63
	MSIG-fix2	4.9	4.9	4.2	4.2	4.8	4.3	4.60	4.45
	MSIG-fix3	4.9	4.9	4.2	4.3	4.1	3.7	4.40	4.27

Table 4.8: Subjective results for factory noise.

Chapter 5

Noise Estimation for Speech Enhancement

Research is to see what everybody else has seen,
and to think what nobody else has thought.
– Albert Szent-Gyorgyi

5.1 Introduction

This chapter will focus on noise estimation for speech enhancement. As the noise power may change rapidly over time, its estimate has to be updated as often as possible. Using an overestimate or an underestimate of the true, but unknown, spectral noise power will lead to an over-attenuation or under-attenuation of the noisy signal. This might lead to a large amount of speech distortion and remaining noise when employed in a speech enhancement framework. One way to estimate the spectral noise power is during periods where speech is present or absent. This requires detection of speech presence by using a voice activity detection (VAD), see e.g., [92, 143, 144]. However, in non-stationary noise scenarios, this detection is difficult, as a sudden rise in the noise power can be treated as a speech onset. In addition, if the noise spectral power changes during speech presence, this change can only be detected with a delay.

To improve estimation of the spectral noise power, several approaches have been proposed in literature. Among the most established estimators are those

based on minimum statistics (MS) [90, 91, 145]. In [90] the power spectrum of the noisy signal is estimated on a frame basis and observed over a finite window of about 1-3 seconds time-span. In general, MS based spectral noise power estimators utilises the assumption that speech is absent during at least a small part within the observed window. The spectral noise power is then obtained from the minimum of the estimated power spectrum of the noisy signal. This makes MS robust, however, if the noise power rises within the observed window, noise spectrum will be underestimated or tracked with a certain delay. The amount of delay generally depends on the buffer length chosen for the finite window. A shorter buffer length results in a shorter maximum delay. However, decreasing the buffer length is not plausible as this increases the chance that speech is not absent within this observed window. The consequence of this is that the spectral noise power may be overestimated, as the estimator might track instances of the noisy spectral power instead of the noise spectral power. Thus in [90] mechanisms are proposed to enable tracking of rising noise powers within the observed window, but still with a rather large delay. The local underestimation of the noise power is likely to result in annoying artifacts, like residual noise and musical noise, when the noise power estimate is applied in a speech enhancement framework.

The methods in [91, 145, 146] are based on a recursive averaging of the noisy spectral power using the speech presence probability (SPP), which is obtained from the ratio of the likelihood functions of speech presence and speech absence. As oppose to the likelihood of speech absence, the likelihood of speech presence contains the *a priori* signal-to-noise ratio (SNR). In case the *a priori* SNR is zero, both likelihood functions overlap such that there will be no distinction between speech presence and absence. In [91, 145, 146], the *a priori* SNR is estimated adaptively on a short-time scale. In speech absence the adapted *a priori* SNR estimate is close to zero, and the two likelihood functions eventually overlap. The resulting *a posteriori* SPP yields only the *a priori* SPP, which is independent of the observed signal. This problem is tackled in [91], where low values for the *a posteriori* SPP are enabled by an additional adaptation of the *a priori* SPP with respect to the observation. However, as the methods in [145] and [146] are based on MS principles, they show also a delay in tracking the rising spectral

noise powers similarly to [90].

Recent works in spectral noise power estimation generally focus on tracking of the spectral noise power with a shorter delay and lower complexity, i.e., the minimum mean square error (MMSE)-based approaches [137, 147]. They are computationally less demanding and at the same time robust to increasing noise levels as shown in a comparison presented in [148]. In the MMSE-based estimator, first a limited maximum likelihood (ML) estimate of the *a priori* SNR is used to obtain an MMSE estimate of the noise periodogram. Nevertheless, under the given *a priori* SNR estimate, the resulting MMSE estimate exhibits a bias which can be computed analytically. In order to compensate for the bias, a second estimate of the *a priori* SNR is required. It has been shown in [149] that under the given ML *a priori* SNR estimator, the MMSE-based spectral noise power estimator can be interpreted as a VAD-based estimator. To improve the MMSE-based spectral noise power estimator, the algorithm in [137] was modified in [149] such that it evolves into a soft SPP instead of a hard SPP (i.e., VAD)-based estimator, which automatically makes the estimator unbiased. The proposed estimator exhibits a computational complexity that is even lower than that of the MMSE-based approach [137] while maintaining its fast noise tracking performance without requiring a bias compensation.

In this chapter, two algorithms for noise estimation, which focus on low tracking delay and low computational complexity for hearing protection devices, are proposed. The first method, namely step-size controlled (SSC) noise estimator, does not require the estimation of statistical properties of noise and speech. An estimation procedure is developed by comparing the noise estimate with the smoothed noisy speech spectrum at every time frame. Based on this comparison, the noise estimate is updated from its own feedback using a defined step-size. The step-size is optimised such that it can track the true noise estimate while still providing robustness to speech onsets for varying noise conditions and varying SNRs. The step-size is optimised by using the golden section search (GSS) approach, with cost function based on the symmetric logarithmic-error distortion measure (LogErr) and the perceptual evaluation of speech quality (PESQ) objective measure. The advantage of the proposed method is that the noise tracking

does not depend on the search window length and does not require any bias compensation. The second method, called the soft VAD (SVAD) approach, does not really mean a VAD but a recursive averaging that relies on a modified SPP. For this method, a flexible sigmoid function can be used to replace the SPP algorithm in [149]. This function offers the possibility to alter the slope and the mean of the SPP independently to achieve a desired trade-off between noise overestimation and underestimation. We also argue that a soft SPP is insufficient for the noise tracking, and so improve it by employing harder decisions based on conditional smoothing.

5.2 Step-size Controlled Noise PSD Estimator

A SSC noise power spectral density (PSD) estimator is proposed in this section. It is motivated by the well-known sigma-delta modulation for encoding analog signals into digital signals or higher-resolution digital signals into lower-resolution digital signals. As the low-resolution signal typically changes more quickly than the high-resolution signal, the conversion is done using error feedback, where the difference between the two signals is measured, and the low-resolution signal can be filtered to recover the high-resolution signal with little or no loss of fidelity. With this in mind, to obtain a noise estimate without using any prior knowledge of the noise statistics, lets consider the following two hypotheses

$$\begin{aligned} \mathcal{H}_0(k,m): &\quad Y(k,m) = V(k,m) \\ \mathcal{H}_1(k,m): &\quad Y(k,m) = X(k,m) + V(k,m). \end{aligned} \tag{5.1}$$

Let $\Lambda(k,m)$ denote the noise estimate. The hypotheses in Eq. (5.1) have the following properties

$$\begin{aligned} &\Lambda(k,m)|\mathcal{H}_0(k,m): \text{ noise level only changes over long period of time} \\ &\Lambda(k,m)|\mathcal{H}_1(k,m): \text{ speech has large changes in the envelope.} \end{aligned} \tag{5.2}$$

Now consider an estimator that can track slow variations in $\Lambda(k,m)|\mathcal{H}_0(k,m)$ but is not fast in tracking large deviations within short durations as in the case of $\Lambda(k,m)|\mathcal{H}_0(k,m)$. Such an estimator can be found by controlling the feedback of the estimate $\Lambda(k,m)$ from previous frame at current frame with a constant

step-size. This is done by firstly initialising $\Lambda(k,m)$ to be equal to the noisy speech estimate $\lambda_y(k,m)$, given by

$$\lambda_y(k,m) = \alpha_y \lambda_y(k,m-1) + (1-\alpha_y)|Y(k,m)|^2 \tag{5.3}$$

where $\alpha_y = \exp(-2.2R)/(t_y f_s)$ is the smoothing constant, with R denotes the frame rate and f_s denotes the sampling frequency. This smoothing constant determines the tracking capability and the estimation error of the speech estimate. Since $Y(k,m)$ contains both speech and noise, α_y is chosen such that it represents a short-term speech estimate.

After that, the following steps are computed for every time frame

$$\begin{aligned} &\text{If} && \Lambda(k,m-1) \leq \lambda_y(k,m) \\ &\text{Then} \\ &&& \Lambda(k,m) = (1+\delta_1)\Lambda(k,m-1) \\ &\text{Otherwise} \\ &&& \Lambda(k,m) = (1-\delta_2)\Lambda(k,m-1) \end{aligned} \tag{5.4}$$

where δ_1 and δ_2 denote the step-size constants. The noise estimate $\Lambda(k,m)$ can be smoothed again using Eq. (5.3) with a larger time constant to reduce variations

$$\lambda_v(k,m) = \alpha_v \lambda_v(k,m-1) + (1-\alpha_v)|\Lambda(k,m)| \tag{5.5}$$

where α_v is calculated same way as α_y, but with longer averaging, such that $\alpha_v > \alpha_y$. A motivation for the update in Eq. (5.4) can be illustrated in Figure 5.1, which shows the noisy speech estimate $\lambda_y(k,m)$ and the noise estimates, $\Lambda(k,m)$ and $\lambda_v(k,m)$, respectively, for noisy speech corrupted by factory noise. By using the step-size constants δ_1 and δ_2, the noise estimate $\Lambda(k,m)$ can track variations in the background noise but is not sensitive to large deviations, which represent the speech onsets. The long-term estimate $\lambda_v(k,m)$ is shown to have much less variations than $\Lambda(k,m)$.

The noise estimator becomes unbiased when both step-size constants are equal, i.e., $\delta_1 = \delta_2 = \delta$. As such, the performance of this estimator depends on neither a bias compensation nor the window length. The resulting $\Lambda(k,m)$ will be very robust against speech onsets since the comparison only measures if the current short-term noisy speech estimate $\lambda_y(k,m)$ is larger or smaller than the

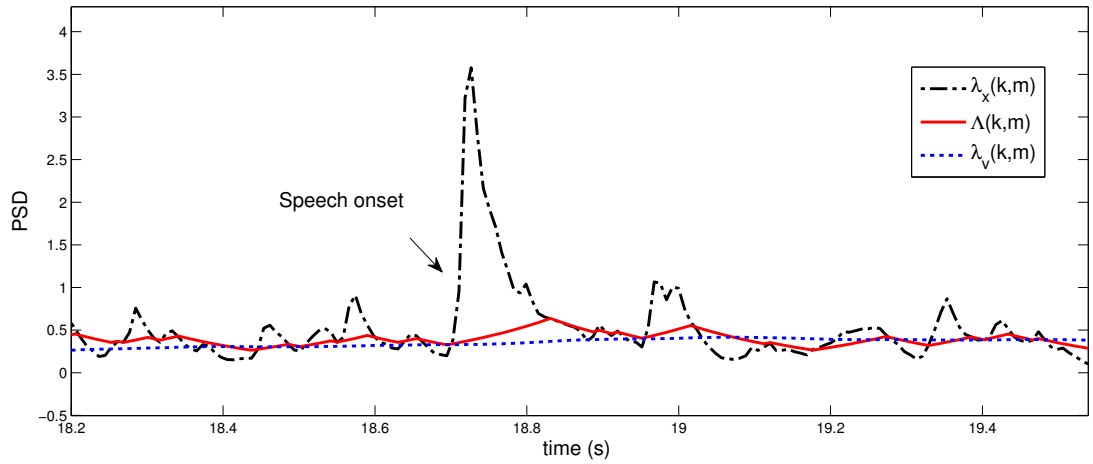

Figure 5.1: Speech corrupted by factory noise: comparison between short-term noisy speech estimate $\lambda_y(k,m)$ (dotted line) and noise estimate before smoothing $\Lambda(k,m)$ (solid line) and after smoothing $\lambda_v(k,m)$ (dash-dot line) at 1562.5 Hz.

previous noise estimate $\Lambda(k,m-1)$. Also, the simplicity of this noise estimator makes it applicable for real time application.

5.2.1 Step-size Optimisation

The objective is to obtain the optimal step-size $\delta_{\text{opt.}}$, which can be used for different types and levels of noise. This is done based on the performance of noise tracking and speech quality. The noise tracking capability is evaluated by using the LogErr measure [137]

$$\text{LogErr} = \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M}\left|10\log_{10}\frac{\sigma_v^2(k,m)}{\Lambda(k,m)}\right| \quad \text{[dB]} \tag{5.6}$$

where $\sigma_v^2(k,m)$ denotes the true noise power spectrum obtained by recursive smoothing the noise periodogram $|V(k,m)|^2$

$$\sigma_v^2(k,m) = 0.9\sigma_v^2(k,m-1) + 0.1|V(k,m)|^2. \tag{5.7}$$

The speech quality performance is evaluated by using the PESQ measure. PESQ has been proposed in ITU-T Recommendation P.862 and has been suggested to be a reliable objective measure for speech quality [96].

Thus, the optimisation problem is formulated as

$$F(\delta) = \min_{\delta} W\text{LogErr}(\delta) - (1-W)\text{PESQ}(\delta) \tag{5.8}$$

where $0 \leq W \leq 1$ represents the trade-off between two objective measures. Here W is chosen as $W = 0.5$. The search of $\delta_{\text{opt.}}$ is done by using the GSS method. Let the search interval be $[\delta_i, \delta_j]$, the measurement points within the first interval $[\delta_i, \delta_j]$ are

$$\delta_a = \quad \varpi\delta_i + (1 - \varpi)\,\delta_j \qquad \delta_b = (1 - \varpi)\,\delta_i + \varpi\delta_j \tag{5.9}$$

where $\varpi = \frac{\sqrt{5}-1}{2}$ denotes the golden search ratio. The search intervals $[\delta_i, \delta_j]$ and the measurement points $[\delta_a, \delta_b]$ are repeatedly updated by using the following steps

$$\begin{aligned} &\text{If} \quad F(\delta_a) \leq F(\delta_b) \quad \text{then} \\ &\text{Update} \quad \delta_j = \delta_b \quad \delta'_b = \delta_a \quad \delta'_a = \varpi\delta_i + (1 - \varpi)\,\delta_j \\ &\text{Otherwise} \\ &\text{Update} \quad \delta_i = \delta_a \quad \delta'_a = \delta_b \quad \delta'_b = (1 - \varpi)\,\delta_i + \varpi\delta_j. \end{aligned} \tag{5.10}$$

The GSS is terminated if $|\delta_b - \delta_a| < \varsigma$, where ς denotes the error tolerance of $\delta_{\text{opt.}}$. The optimal value $\delta_{\text{opt.}}$ is set as the smallest value among δ_a and δ_b.

In order to obtain an optimal $\delta_{\text{opt.}}$ for SSC noise PSD estimator, 24 English spoken utterances were used. The speech enhancement system in [5] was employed. The parameters used for the step-size optimisation are chosen from empirical studies, such that $\delta_i = 0.01$, $\delta_j = 0.05$, and $\varsigma = 0.000001$. The optimal step-size $\delta_{\text{opt.}}$ is obtained by computing $F(\delta)$ for each speech sentence in the database at each global SNR. Table 5.1 shows the mean values of $\delta_{\text{opt.}}$ with their standard deviations. As observed, $\delta_{\text{opt.}}$ decreases when the input SNR increases. This follows naturally from the fact that higher noise levels would have larger variations in the PSD estimate, thus a larger δ is required under lower input SNR. However, for factory noise, the mean value at 10 dB SNR is larger when compared to the mean value at 5 dB SNR. This is due to the fact that the noise variations is small at high SNR. As such, the noise estimate becomes more sensitive to large variations in speech onsets. This is particularly noticeable when the factory noise is non-stationary and is usually concentrated in the low frequency range. The results also show that the standard deviation for factory noise is much larger compared to the standard deviation for pink noise. By taking the average of the values in Table 5.1, step-size is set as $\delta = 0.04$ for the performance evaluation later in this chapter.

SNR	Pink Noise	Factory Noise
-5	0.0483±0.0011	0.0415±0.0115
0	0.0478±0.0014	0.0344±0.0156
5	0.0476±0.0023	0.0316±0.0187
10	0.0418±0.0131	0.0341±0.0192

Table 5.1: Optimal step-size values $\delta_{\text{opt.}}$ (mean ± standard deviation).

5.2.2 Varying Smoothing Factors

As mentioned in the previous section, the SSC noise estimate tends to become more sensitive to large variations in speech onsets at high input SNR, and tracks the speech estimate instead of the noise estimate. One way to avoid this is to let δ vary depending on the noise types and input SNRs, which is not what we desired. Here, we utilise the fact that since the constant step-size δ follows the noisy speech estimate $\lambda_y(k,m)$ in order to update the estimate $\Lambda(k,m)$, the amount of fluctuations in $\lambda_y(k,m)$ decides the performance of the noise tracker. By employing a shorter averaging for α_y, the estimate $\Lambda(k,m)$ would track the small variations in noise estimate and would be less sensitive to the large fluctuations in speech estimate under high input SNR. However, in this case if the variations of $|Y(k,m)|^2$ is large, i.e., when the input SNR is very low, $\Lambda(k,m)$ will also have a large variation and will probably underestimate the high level noise since the step-size $\delta = 0.04$ is small. Thus, a mechanism is proposed to use a longer averaging at low SNR and a shorter one at high SNR, given by

$$
\begin{aligned}
&\text{If} \quad 5\text{dB} < 10\log_{10} E\left(\tfrac{\lambda_y(k,m)}{\lambda_v(k,m)}\right) \leq 10\text{dB} \quad \text{then} \\
&\quad t_y = -\left(\tfrac{t_{y,2}-t_{y,1}}{\text{SNR}_2-\text{SNR}_1}\right)\text{SNR}_1 + t_{y,1} + \left(\tfrac{t_{y,2}-t_{y,1}}{\text{SNR}_2-\text{SNR}_1}\right)\left(10\log_{10} E\left(\tfrac{\lambda_y(k,m)}{\lambda_v(k,m)}\right)\right) \\
&\text{Otherwise} \\
&\qquad \text{If} \quad 10\log_{10} E\left(\tfrac{\lambda_y(k,m)}{\lambda_v(k,m)}\right) \leq 5\text{dB} \quad \text{then} \qquad t_y = t_{y,1} \\
&\qquad \text{Otherwise} \qquad\qquad\qquad\qquad\qquad\quad t_y = t_{y,2}
\end{aligned}
\tag{5.11}
$$

where $t_{y,1} > t_{y,2}$ and $\text{SNR}_1 < \text{SNR}_2$ are different fixed time constants and input SNR in dB, respectively. Similar to Eq. (5.11), shorter averaging is also applied to $\lambda_v(k,m)$ at higher input SNR, with $t_{v,1} > t_{v,2}$ and $\text{SNR}_1 < \text{SNR}_2$.

5.3 Soft VAD Noise PSD Estimator

Now, we will focus on the second main study in this chapter, namely the SVAD Noise PSD Estimator. Again, given two hypothesis, $\mathcal{H}_0(k,m)$ and $\mathcal{H}_1(k,m)$, which denote, respectively, speech absence and speech presence in the k^{th} frequency bin of the m^{th} frame. Under the assumption that the STFT coefficients of both speech and noise are complex Gaussian distributed, the conditional probability density functions (PDFs) of the observation are given by [91]

$$P\left(Y(k,m)|\mathcal{H}_0(k,m)\right) = \frac{1}{\lambda_v(k,m)\pi}\exp\left(-\frac{|Y(k,m)|^2}{\lambda_v(k,m)}\right) \tag{5.12}$$

$$P\left(Y(k,m)|\mathcal{H}_1(k,m)\right) = \frac{1}{\left(\lambda_x(k,m)+\lambda_v(k,m)\right)\pi}\exp\left(-\frac{|Y(k,m)|^2}{\lambda_x(k,m)+\lambda_v(k,m)}\right) \tag{5.13}$$

where $\lambda_x(k,m) = E\left\{|X(k,m)|^2\,|\mathcal{H}_1(k,m)\right\}$ is the speech power spectrum and $\lambda_v(k,m) = E\left\{|V(k,m)|^2\right\}$ indicates the noise power spectrum. As such, the *a posteriori* SNR and the *a priori* SNR can be defined respectively as $\gamma = \frac{|Y(k,m)|^2}{\lambda_v(k,m)}$ and $\xi = \frac{\lambda_x(k,m)}{\lambda_v(k,m)}$. By applying Bayes' theorem, the *a posteriori* SPP $p(k,m) = P\left(\mathcal{H}_1(k,m)|Y(k,m)\right)$ can be obtained from Eqs. (5.12) and (5.13) at every time-frequency point as [91]

$$p(k,m) = \left\{1 + \frac{P\left(\mathcal{H}_0(k,m)\right)}{P\left(\mathcal{H}_1(k,m)\right)}\left(1+\xi(k,m)\right)\exp\left(-\hat{\gamma}(k,m)\frac{\xi(k,m)}{1+\xi(k,m)}\right)\right\}^{-1} \tag{5.14}$$

where $P\left(\mathcal{H}_0(k,m)\right)$ and $P\left(\mathcal{H}_1(k,m)\right)$ denote, respectively, the *a priori* probabilities for speech absence and speech presence.

As the noise power spectrum $\lambda_v(k,m)$ is practically unknown, it has to be estimated. A direct way of obtaining the estimate $\hat{\lambda}_v(k,m)$ is by applying a temporal recursive smoothing to the noisy observation during speech absence periods only, such that

$$\begin{aligned}
&\mathcal{H}_0(k,m):\ \hat{\lambda}_v(k,m) = \alpha_v\hat{\lambda}_v(k,m-1) + (1-\alpha_v)\,|Y(k,m)|^2 \\
&\mathcal{H}_1(k,m):\ \hat{\lambda}_v(k,m) = \hat{\lambda}_v(k,m-1)
\end{aligned} \tag{5.15}$$

where $\alpha_v(k,m)$ denotes the smoothing factor for the noise PSD estimate. By applying the *a posteriori* SPP from Eq. (5.14) to Eq. (5.15), the recursive averaging

becomes [91]

$$\begin{aligned}\hat{\lambda}_v(k,m) = \; & p(k,m)\hat{\lambda}_v(k,m-1) \\ & + (1-p(k,m))\left[\alpha_v\hat{\lambda}_v(k,m-1) + (1-\alpha_v)\left|Y(k,m)\right|^2\right].\end{aligned} \tag{5.16}$$

Whereas such SPP based noise power estimator can also be derived in the MMSE sense as in [149]. That results in

$$\begin{aligned}\hat{\lambda}_v(k,m) = \; & \alpha_v\hat{\lambda}_v(k,m-1) \\ & + (1-\alpha_v)\left[p(k,m)\hat{\lambda}_v(k,m-1) + (1-p(k,m))\left|Y(k,m)\right|^2\right]\end{aligned} \tag{5.17}$$

which is actually identical to Eq. (5.16) with different arrangement of parameters. In this case, the main factor that updates the noise power estimate lies in the *a priori* estimation of ξ, $P\left(\mathcal{H}_0(k,m)\right)$ and $P\left(\mathcal{H}_1(k,m)\right)$ in Eq. (5.14), where $P\left(\mathcal{H}_1(k,m)\right) = 1 - P\left(\mathcal{H}_0(k,m)\right)$.

5.3.1 Speech Presence Probability

The *a posteriori* SPP in Eq. (5.14) requires an estimate of the *a priori* SNR, $\xi(k,m)$, which tends to zero in speech absence, and is gradually increasing with speech power spectrum. However, when $\xi(k,m) = 0$, i.e., $\lambda_x(k,m) = 0$, the likelihoods of speech absence in Eq. (5.12) and speech presence in Eq. (5.13) become identical. As such, the *a posteriori* SPP becomes

$$p(k,m) = P\left(\mathcal{H}_1(k,m)|Y(k,m),\xi(k,m)=0\right) = P\left(\mathcal{H}_1(k,m)\right) \tag{5.18}$$

which is independent of the noisy observation, $Y(k,m)$. In this case, speech absence cannot be detected unless $P\left(\mathcal{H}_1(k,m)\right)$ is modified based on the observation signal, as shown in [91].

To solve the problem, it is proposed in [149] to replace the *a priori* SNR, $\xi(k,m)$ by a fixed parameter, $\xi_{\mathcal{H}_1}$, which represents the typical *a priori* SNR value when speech is active. Such fixed optimal is obtained at $10\log_{10}\left(\xi_{\mathcal{H}_1}\right) = 15\text{dB}$ by minimising the probability of error, P_ϵ, which is a function of $P\left(\mathcal{H}_0(k,m)\right)$ and $P\left(\mathcal{H}_1(k,m)\right)$ [149]. However, according to the result in [149], it can be seen that $12\text{dB} \leq 10\log_{10}\left(\xi_{\mathcal{H}_1}\right) \leq 18\text{dB}$ yields about similar probability of error. While P_ϵ was obtained by assuming uniform priors for speech probability, i.e.,

$P(\mathcal{H}_0(k,m)) = P(\mathcal{H}_1(k,m)) = 0.5$, that indicates a worst case scenario and may not hold in practice. This reflects that $\xi_{\mathcal{H}_1} = 15\text{dB}$ is not the only possible value to be used to estimate $p(k,m)$. By choosing different values of $\xi_{\mathcal{H}_1}$, the speed of $p(k,m)$ to switch between speech absence and presence can be altered.

In addition, by employing fixed priors, the *a posteriori* SPP will be mapped directly to the *a posteriori* SNR estimate $\hat{\gamma}(k,m)$. As such, $p(k,m)$ yields a value that is close to zero when $\hat{\gamma}(k,m)$ is small, and close to one when $\hat{\gamma}(k,m)$ is sufficiently large. In between zero and one is a soft transition determined by $\xi_{\mathcal{H}_1}(k,m)$, $P(\mathcal{H}_0(k,m))$ and $P(\mathcal{H}_1(k,m))$. However, since $\hat{\gamma}(k,m)$ is an estimate based on the magnitude power of the raw observation $|Y(k,m)|^2$, variations can grow very large. By tracking these variations with the soft decisions between zero and one, $p(k,m)$ becomes an estimate with large fluctuations as well. This increases the probability of noise being overestimated or underestimated locally, which results in speech distortion or annoying artifacts perceived as musical noise.

5.3.2 Interpretation of SPP as a Flexible Sigmoid Function

To avoid the procedures of finding another so-called optimal $\xi_{\mathcal{H}_1}$ with different assumptions of $P(\mathcal{H}_0(k,m))$ and $P(\mathcal{H}_1(k,m))$, a sigmoid function is employed to model the *a posteriori* SPP in Eq. (5.14). This sigmoid function is defined as

$$p_{\text{sig}}(k,m) = \left\{1 + \exp\left(-a_{\text{sig}}(k,m)\left(\hat{\gamma}(k,m) - c_{\text{sig}}(k,m)\right)\right)\right\}^{-1} \tag{5.19}$$

where $a_{\text{sig}}(k,m)$ and $c_{\text{sig}}(k,m)$ indicate, respectively, the slope and the mean of the sigmoid function. Both are given by

$$a_{\text{sig}}(k,m) = \frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}, \tag{5.20}$$

$$c_{\text{sig}}(k,m) = \log\left(\frac{P(\mathcal{H}_0(k,m))}{P(\mathcal{H}_1(k,m))}\left(1+\xi_{\mathcal{H}_1}\right)\right)\frac{1+\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}}. \tag{5.21}$$

As mentioned in the previous section, when $P(\mathcal{H}_0(k,m)) = P(\mathcal{H}_1(k,m))$, $\xi_{\mathcal{H}_1}$ can be any value of $12\text{dB} \leq 10\log_{10}(\xi_{\mathcal{H}_1}) \leq 18\text{dB}$. By substituting these values into the slope and the mean of the sigmoid function, one obtains $0.94 \leq a_{\text{sig}}(k,m) \leq 0.98$ and $3 \leq c_{\text{sig}}(k,m) \leq 4.23$. This reflects that as long as the

range of $a_{\text{sig}}(k,m)$ and $c_{\text{sig}}(k,m)$ is satisfied, any value of $\xi_{\mathcal{H}_1}$, $P\left(\mathcal{H}_0(k,m)\right)$ and $P\left(\mathcal{H}_1(k,m)\right)$ can be used, provided that $10\log_{10}\left(\xi_{\mathcal{H}_1}\right) \in [12\text{dB}, 18\text{dB}]$ and $P\left(\mathcal{H}_0(k,m)\right) + P\left(\mathcal{H}_1(k,m)\right) = 1$. Since different values of $\xi_{\mathcal{H}_1}$ can be used, such that ξ_a for $a_{\text{sig}}(k,m)$ and ξ_c for $c_{\text{sig}}(k,m)$, the slope and the mean can be controlled independently. As such, the *a posteriori* SPP can be fully controlled to achieve a desired trade-off between noise overestimation and underestimation.

5.3.3 Conditional Smoothing for the Sigmoid Function

Literally speaking, rather than having a soft transition between the absolute speech absence and speech presence, the *a posteriori* SPP can be categorised into three regions, as

$$p'(k,m) = \begin{cases} \text{less likely speech presence,} & p(k,m)_{\text{sig}}(k,m) \leq p_1(k,m) \\ \text{more likely speech presence,} & p_1(k,m) < p_{\text{sig}}(k,m) \leq p_2(k,m) \\ \text{most likely speech presence,} & p_{\text{sig}}(k,m) > p_3(k,m) \end{cases} \tag{5.22}$$

where $0 < p_1(k,m) < p_2(k,m) < p_3(k,m) \leq 1$ are different values of the sigmoid function. For the region where speech is less likely to present, i.e., when $\hat{\gamma}(k,m) \approx 0$, in case speech is active, $p_{\text{sig}}(k,m)$ should be prevented from getting too close to zero. By doing so, the noise PSD estimate yields a comprised weight of sum between the estimated noise PSD at previous frames and the instance noisy observations. The result is an even smoothed estimate compared to the original noise PSD estimate when $\hat{\gamma}(k,m)$ is small, which reduces the likelihood of noise being overestimated and underestimated locally. While for the regions where speech is either more likely or most likely to present, the soft transitions of $p_{\text{sig}}(k,m)$ might not be sufficient for the noise PSD estimate to change from using the previous noise PSD estimates to tracking the current noisy observations and vice versa. Accordingly, to avoid those pitfalls, multi-level decisions are imposed on $p_{\text{sig}}(k,m)$ to realise an improved *a posteriori* SPP estimate. A solution for this

is proposed as

$$p'(k,m) = \begin{cases} \mathcal{P}_1, & p_{\text{sig}}(k,m) \leq 0.3 \\ \mathcal{P}_2, & 0.3 < p_{\text{sig}}(k,m) \leq 0.6 \\ \min\left\{\mathcal{P}_3,\, p_{\text{sig}}(k,m)\right\}, & p_{\text{sig}}(k,m) > 0.6 \end{cases} \tag{5.23}$$

where $\mathcal{P}_i = \exp\left(-2.2R\right) / \left(t_i f_{\text{s}}\right)$ indicates the exponential smoothing constant, with t_i, $i = [1,2,3]$ denotes the averaging time constant, with $t_1 < t_2 \ll t_3$. Note that instead of having a fixed smoothing constant, $p'(k,m)$ is assigned with $p_{\text{sig}}(k,m)$ for $p_{\text{sig}}(k,m) > 0.6$ to keep the noise PSD estimate as robust to speech onsets as possible. While the noise power estimate may cease to update when the noise level would make an abrupt step from one sample to another, such that $p'(k,m) = 1$, the upper bound $\mathcal{P}_3$ is employed to prevent such stagnation in the estimate.

5.4 Experimental Results

The proposed methods (SSC and SVAD) are compared to four reference methods, namely the MS algorithm [90], the improved minima controlled recursive averaging (IMCRA) method [91], the Hendriks' MMSE (HenMMSE) approach [137] and the Gerkmann's SPP (GerkSPP) with fixed priors approach [149]. The parameters used for the proposed method are: for SSC, $t_{y,1} = 0.06$, $t_{y,1} = 0$, $t_{v,1} = 1$, $t_{v,2} = 0.01$, $\text{SNR}_1 = 0\,\text{dB}$ $\text{SNR}_2 = 5\,\text{dB}$; while for SVAD, $\xi_a = \xi_b = 17\text{dB}$, $P\left(\mathcal{H}_0(k,m)\right) = 0.3$, $P\left(\mathcal{H}_1\right) = 0.7$, $t_1 = 0.05$, $t_2 = 0.08$ and $t_3 = 240$. Performance evaluations were conducted using NOIZEUS database which contains 30 IEEE sentences spoken by 3 male and 3 female speakers [67]. The signals were corrupted by pink noise or a white Gaussian noise (WGN) at input SNRs of -5, 0, 5, 10, and 15dBs. The modulated noise were created by modulating WGN using the function $f\left(n\right) = 1 + 0.5\sin\left(2\pi n f_{\text{mod}}/f_{\text{s}}\right)$, where n is the time-sample index and $f_{\text{mod}} = 0.5\text{Hz}$ is the modulation frequency. The LogErr was used to measure the noise tracking performance [137]. Unlike Eq. (5.6), in order to evaluate both overestimation and underestimation, it is defined here as $\text{LogErr} = \text{LE}_{\text{Ov}} + \text{LE}_{\text{Un}}$, where LE_{Ov} and LE_{Un} denote, respectively, the LogErr

for noise power overestimation and noise power underestimation. Both are given as [149]

$$\mathrm{LE_{Ov}} = \frac{1}{KM}\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\left|\min\left[0, 10\log_{10}\left(\frac{\lambda_v(k,m)}{\hat{\lambda}_v(k,m)}\right)\right]\right| \tag{5.24}$$

and

$$\mathrm{LE_{Un}} = \frac{1}{KM}\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\max\left[0, 10\log_{10}\left(\frac{\lambda_v(k,m)}{\hat{\lambda}_v(k,m)}\right)\right] \tag{5.25}$$

where K and M values are, respectively, the total number of frequency bins and frames. A larger value of LogErr indicates a better performance. All signals are sampled at $f_\mathrm{s} = 16\mathrm{kHz}$. For STFT a square-root-Hann window was applied with $K = 512$ and $R = 256$ samples. Instead of taking the direct estimate of the noise periodograms, the true noise PSD, $\lambda_v(k,m)$ was obtained by a recursive smoothing given as

$$\lambda_v(m) = 0.9\lambda_v(m-1) + 0.1|V(m)|^2. \tag{5.26}$$

Apart from that, the estimated noise power was evaluated in a speech enhancement system in terms of the kurtosis ratio (KurtR), the noise reduction ratio (NRR) and the log-likelihood ratio (LLR) [6]. They measure the trade-off between the amount of musical noise, noise reduction and speech distortion generated after the speech signal processing. The PESQ measure was also utilised as an indication for overall perceptual performance of the processed speech signals. Lower values of KurtR and LLR with larger NRR and larger PESQ are required for an improved performance. The modified sigmoid (MSIG) function [6] with parameters $a_1 = 3$, $a_2 = 1$ and $c = 0.7$ and the modified decision-directed (MDD) algorithm [6] with smoothing constants $\beta = 0.98$ and $\alpha_y = 0.172$ were employed correspondingly for the gain function and the *a priori* SNR estimate in the speech enhancement framework.

Before evaluating the proposed and reference methods, we show the effects of employing the varying smoothing factors for SSC approach, as illustrated in subsection 5.2.2. Figure 5.2 shows the noise PSD tracking performance of the SSC method with different parameters as described in the figure's caption. A speech sequence of 26 seconds was utilised and was corrupted by pink noise from 0 dB SNR to 10 dB SNR and then back to 0 dB SNR. The depicted results show that SSC_3 improves the performance of SSC with less overestimation of noise PSD at

high SNR when compared to SSC_1, and with less underestimation of noise PSD at low SNR when compared to SSC_2. Consistent results have also been shown from LogErr measure in sub-plots (a) and (b) in Figure 5.3, where pink noise is used for evaluation in NOIZEUS database under global input SNRs from $-5\,\text{dB}$ to $15\,\text{dB}$. It also shows that SSC_3 improves the performance of SSC_1 with slightly larger PESQ scores under high input SNR, where Δ indicates the improvement in PESQ scores between the processed speech signals and the observed noisy signals. Although SSC_2 has the largest PESQ results, there are higher amount of musical noise and residual noise remained unattenuated as depicted in larger KurtR and smaller NRR scores. Meanwhile, Figure 5.4 shows the results obtained by using modulated WGN, with similar patterns compared to the results obtained from pink noise. It can be seen that SSC_3 performs even better in non-stationary noise when compared SSC_1, with smaller KurtR, larger NRR and larger PESQ scores. Clearly, the depicted results from Figures 5.2 to 5.4 have shown that SSC_3 is a better choice than SSC_1 and SSC_2 in terms of better noise tracking and speech quality. Therefore, SSC_3 would be used for the following performance evaluation and is represented by SSC to avoid any confusion.

With the same speech-in-noise scenario as in Figure 5.2, Figure 5.5 shows noise tracking performance for all the evaluated noise PSD tracking algorithms: the reference methods (MS, IMCRA, HenMMSE and GerkSPP) and the proposed algorithms (SSC and SVAD). Two main issues are of interest from the figure: how each noise PSD tracking algorithm reacts to different amount of fluctuations in noise and to the step changes in noise level. Under low input SNR, both proposed methods track the noise PSD very well, which are highly comparable with MS and IMCRA methods. The SVAD approach follows the variation of the true noise PSD while the SSC algorithm represents a highly smoothed noise PSD estimator. Meanwhile, there are higher amount of fluctuations observed for both HenMMSE and GerkSPP approaches, which can lead to more overestimation and underestimation of the noise PSD, particularly the underestimation at 23 seconds. Such phenomenon is more obvious under high input SNR, where the IMCRA approach is more biased to overestimation, while the GerkSPP method has more frequent overestimation and underestimation at local time-frequency

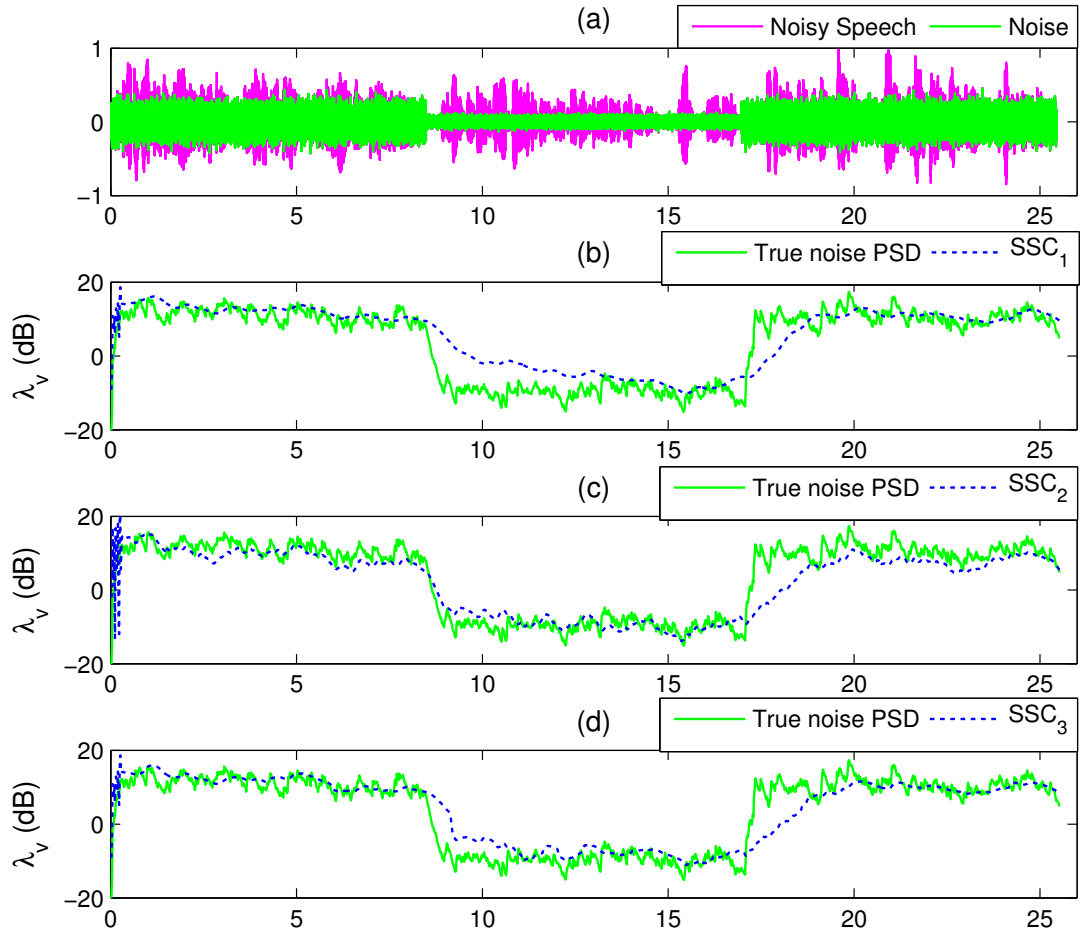

Figure 5.2: SSC tracking performance for pink noise at 0 dB SNR (0-8s), 10 dB SNR (8s-17s) and 0 dB SNR (17s-26s). (b)-(d) Comparison between true noise PSD (green line) and SSC method with different parameters; $\text{SSC}_1 : t_y = t_{y,1}, t_v = t_{v,1}$, $\text{SSC}_2 : t_y = t_{y,2}, t_v = t_{v,2}$, $\text{SSC}_3 : t_y$ and t_v computed with Eq. (5.11).

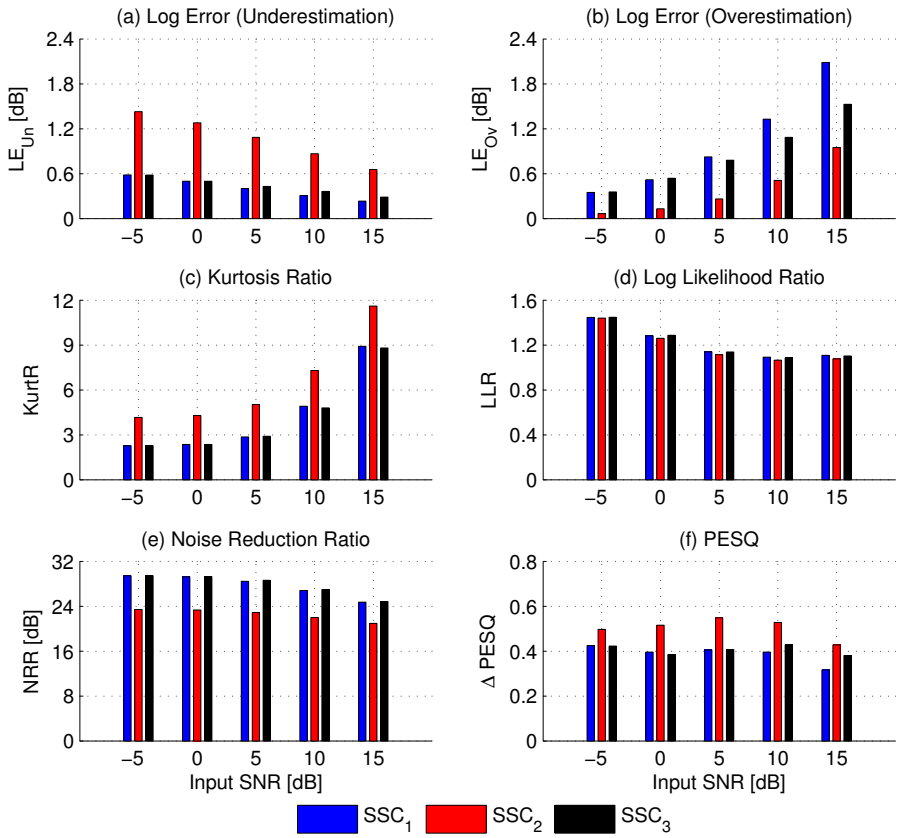

Figure 5.3: SSC performance for pink noise.

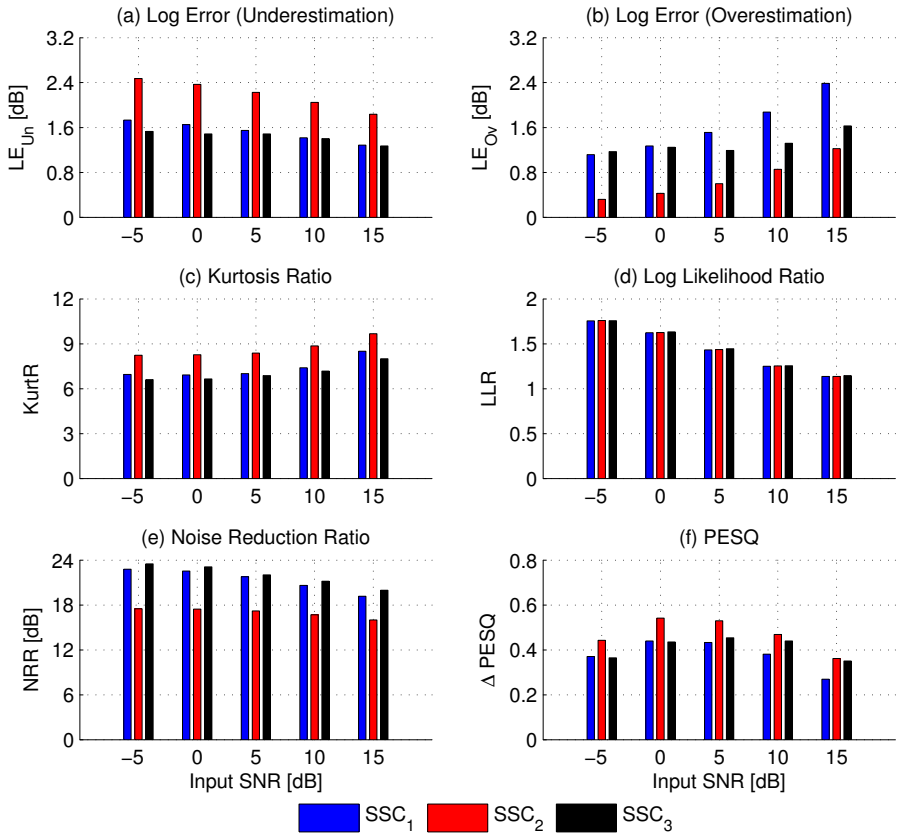

Figure 5.4: SSC performance for modulated WGN.

points compared to other evaluated methods. Improvement to the GerkSPP method has been shown in the SVAD approach, with significantly less errors in the noise power estimates, particularly when there are changes in noise level (9 seconds and 17 seconds). For other evaluated algorithms, when the noise level goes down, the SSC method tends to overestimate noise PSD, whereas both HenMMSE and GerkSPP tend to underestimate the noise power more often. It is worth to re-emphasise that the performance of SSC in tracking the noise PSD is very much depending on the time averaging of the speech estimate and the step-size used to update the noise estimate. When the noise volume goes up, the MS and IMCRA require longer time-frames to pick-up the noise PSD compared to the other methods.

Results from the objective evaluation of the noise tracking performance are shown in Figures 5.6 and 5.7. It can be seen that for more stationary noise such as pink noise, MS, IMCRA and SVAD perform better than the other evaluated methods. The proposed SVAD method has the lowest LogErr under low input SNR, but has slightly higher LogErr at 15 dB input SNR when compared to MS and IMCRA due to the higher overestimation in local time-frames as depicted in Figure 5.5. IMCRA has lower LogErr compared to MS (with significantly smaller $\mathrm{LE_{Un}}$ and slightly larger $\mathrm{LE_{Ov}}$ values), and records the lowest LogErr under high input SNR. For other evaluated methods, the GerkSPP approach has very large $\mathrm{LE_{Un}}$ but small $\mathrm{LE_{Ov}}$, particularly at low input SNR. The SSC algorithm has comparable results when compared to HenMMSE but has higher LogErr under 15 dB due to larger PSD overestimation. However, SSC would have lower delay considering that it has the lowest computational complexity among all the evaluated methods. For non-stationary noise such as the modulated WGN, HenMMSE, SPP and SVAD has comparable better noise tracking performance, with SVAD having the best results under low input SNR. MS and IMCRA perform relatively poor at this type of noise as they are slow in tracking the sudden rise in the noise PSD, with MS performing the poorest among all evaluated methods. The SSC approach has slightly better noise tracking performance compared to IMCRA, with different tracking nature that tends to overestimate instead of underestimating the fast changing noise PSD when compared to MS and IMCRA.

Figures 5.8 and 5.9 show the results from objective measurement after different noise PSD estimation methods applied to a speech enhancement system. It can again be seen from Figure 5.8 that IMCRA and SVAD perform well in terms of the trade-off measures (low KurtR with high NRR values) and PESQ measure for pink noise. However, IMCRA (and MS) cannot track non-stationary noise, which is inferred from the large KurtR and small NRR and PESQ results in Figure 5.9. Although HenMMSE and GerkSPP can track fast changes in non-stationary noise, both of these methods have larger KurtR values when compared to SVAD, which indicates more musical noise can be audible. As for the SSC method, it has poorer PESQ results among the evaluated algorithms due to the tendency to overestimate noise, particularly for modulated WGN in Figure 5.9. In general, either in more stationary or highly non-stationary noise types, SVAD performs the best among all evaluated algorithms in terms of noise tracking performance and the trade-off performance among the amount of musical noise, residual noise and speech distortion generated from the system.

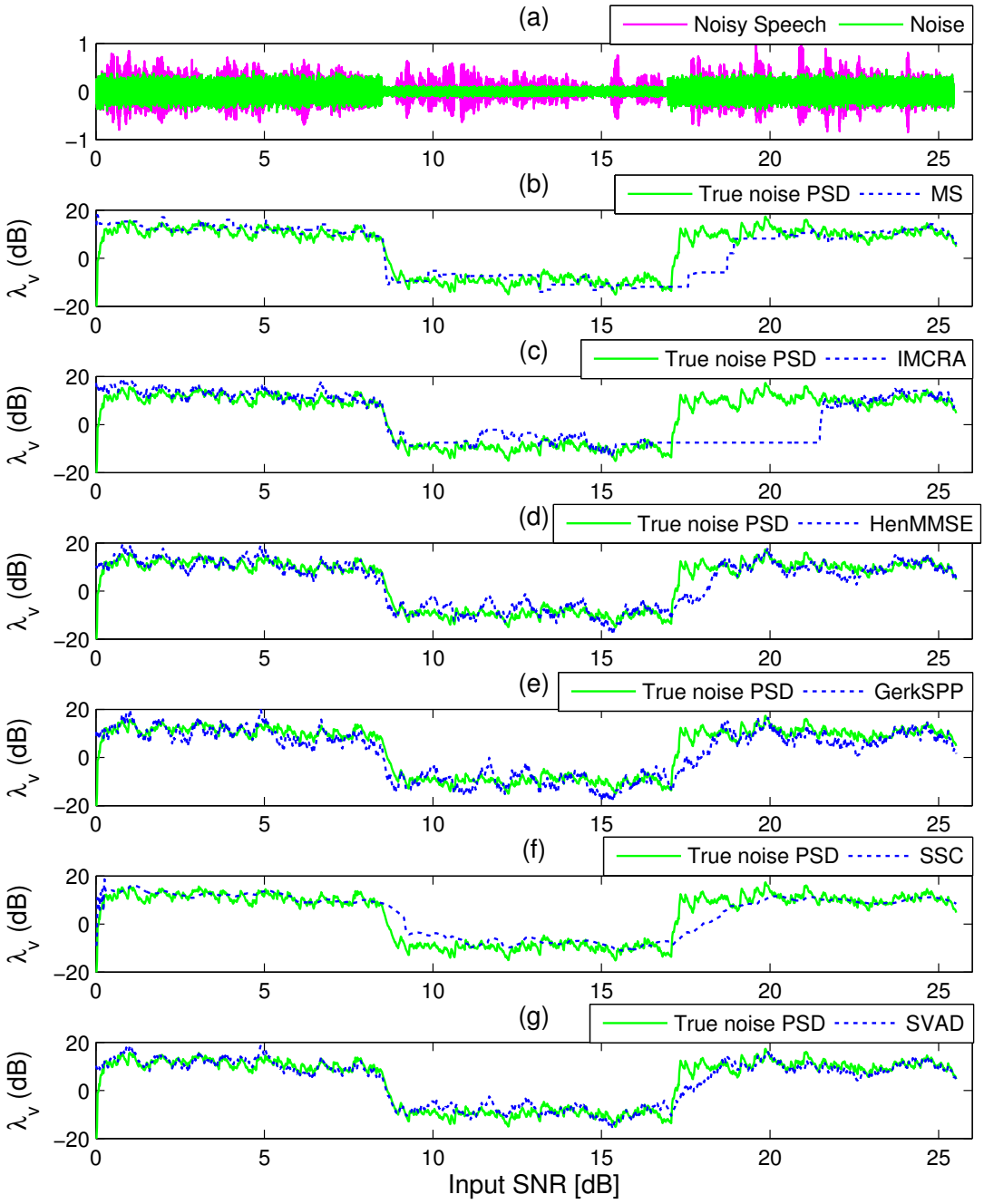

Figure 5.5: Noise tracking performance for pink noise. (a) Speech corrupted by pink noise at 0 dB SNR (0-8s), 10 dB SNR (8s-17s) and 0 dB SNR (17s-26s). (b)-(g) Comparison between true noise PSD (green line) and processed noise PSD at 937.5 Hz.

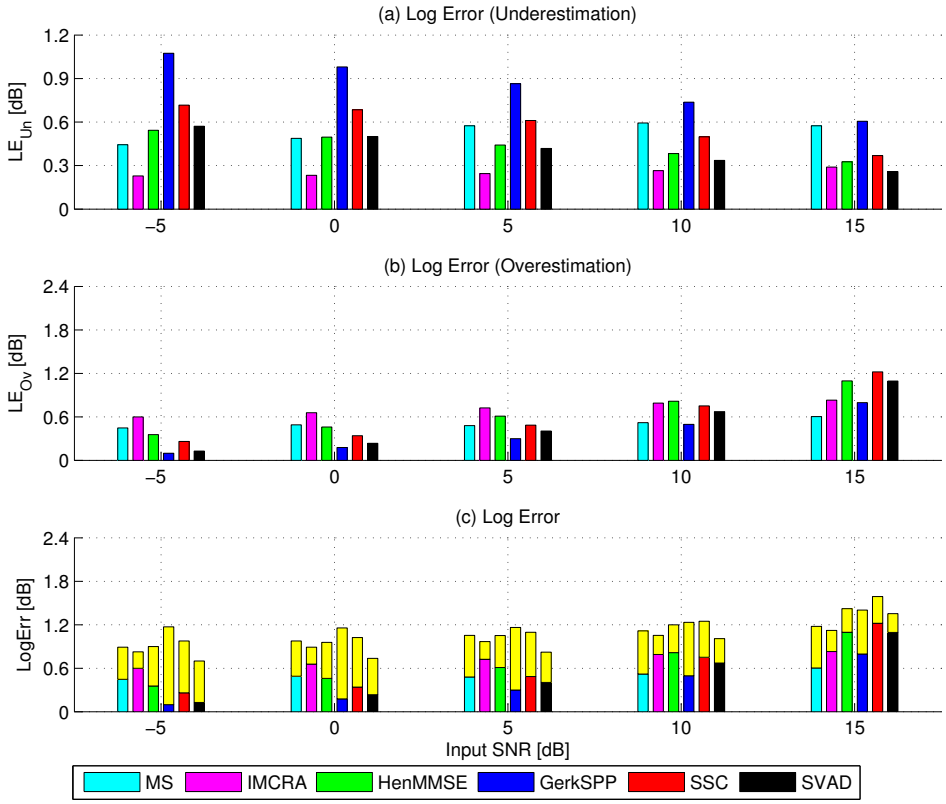

Figure 5.6: Log error performance for pink noise. The lower part of the bars in sub-plot (c) indicates $\mathrm{LE_{Ov}}$, while the upper part represents $\mathrm{LE_{Un}}$. The total height denotes the total LogErr.

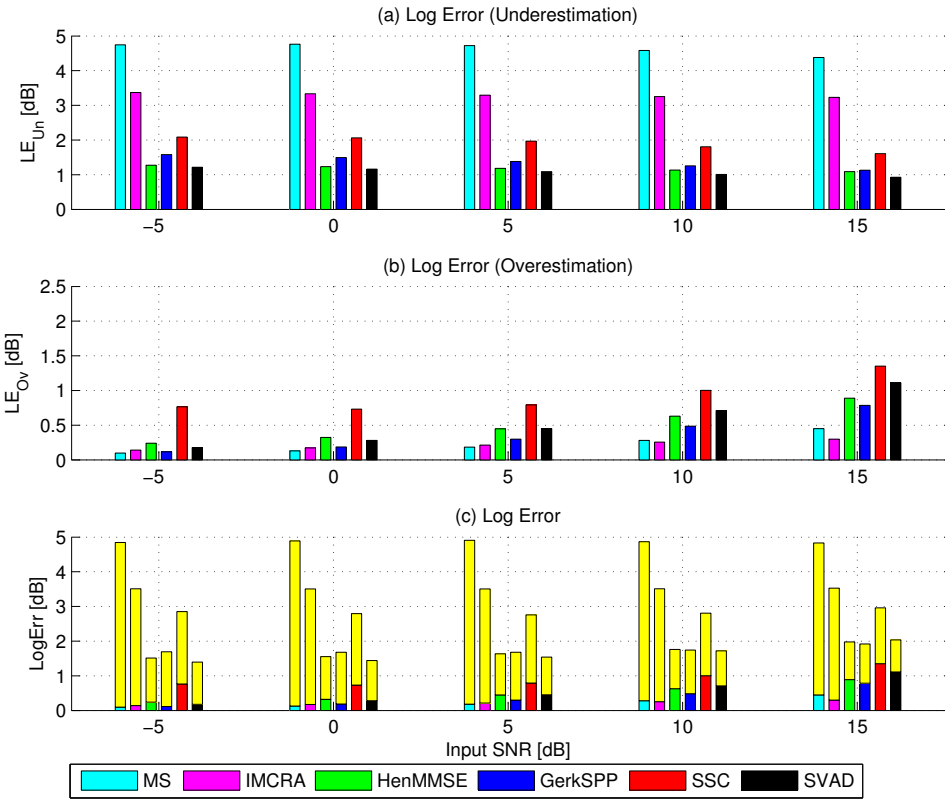

Figure 5.7: Log error performance for modulated WGN.

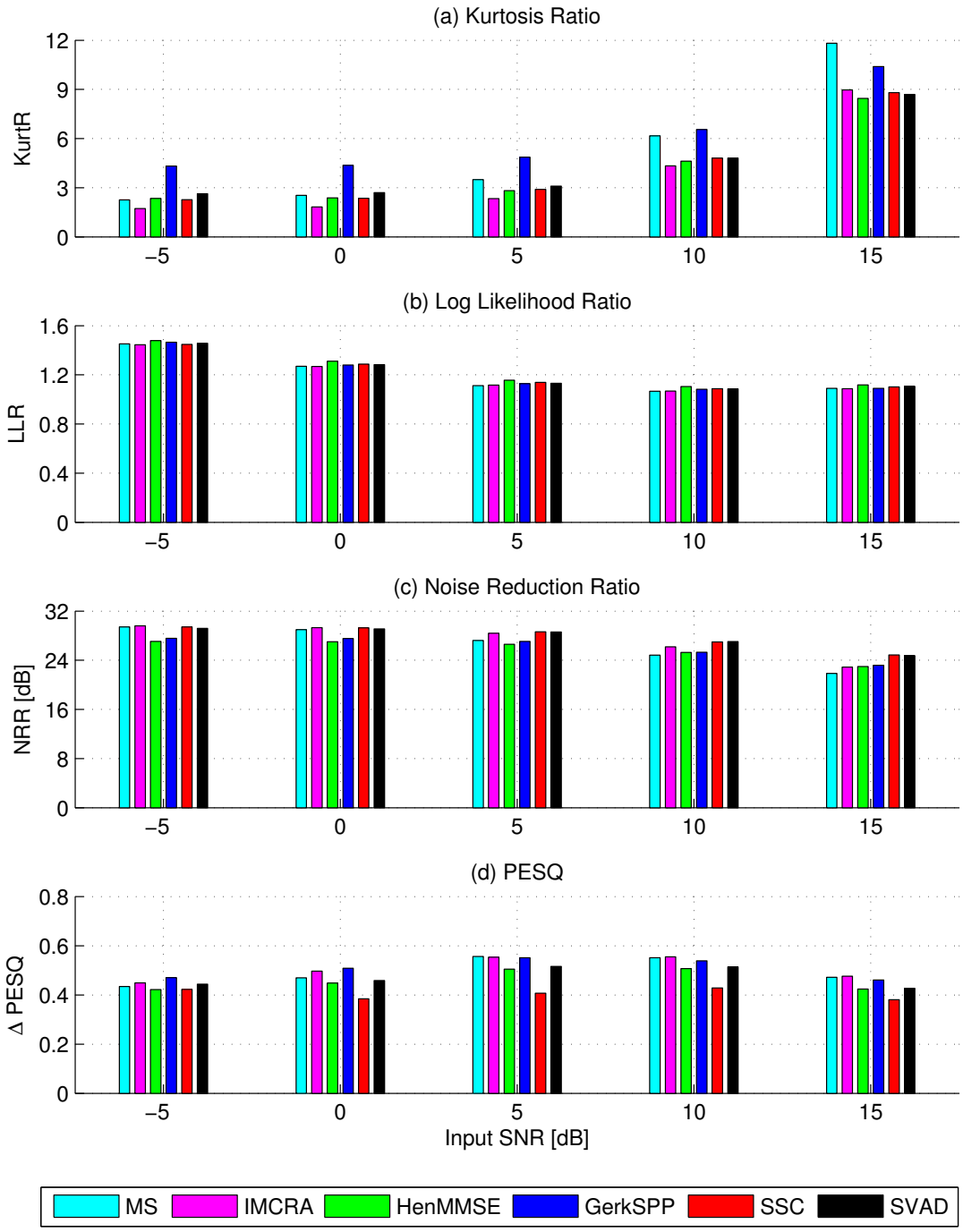

Figure 5.8: Mean performance for pink noise.

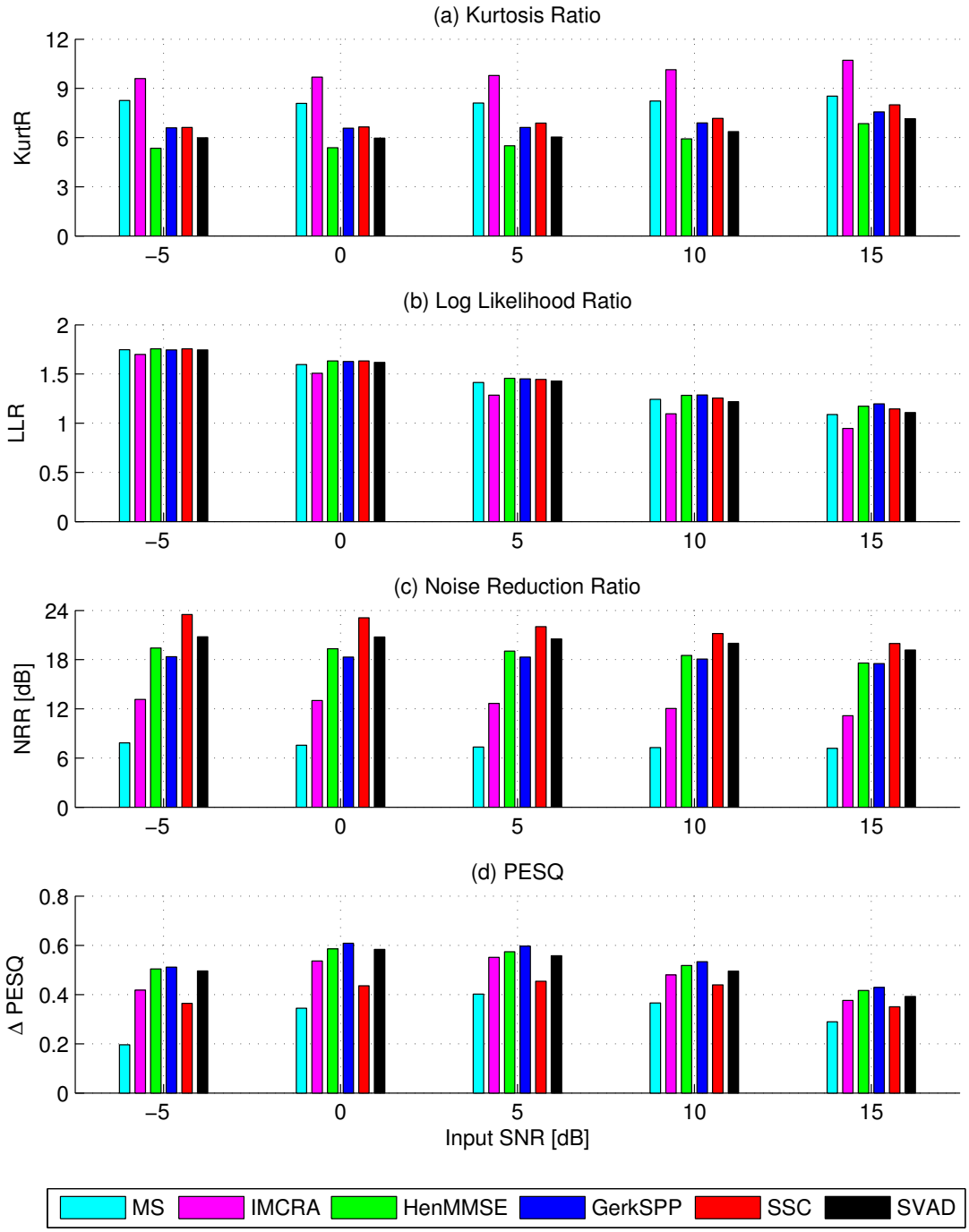

Figure 5.9: Mean performance for modulated WGN.

5.5 Summary

In this chapter, two noise estimators, namely the SSC noise PSD estimator and the SVAD noise PSD estimator have been proposed. The SSC algorithm employs a fixed step-size to track the variations in the noise spectrum based on the noisy speech spectrum in the preceding frame. The step-size is optimised such that it works over different noisy situations. Two noise environment were included for the optimisation procedure in this work, namely the pink noise and the factory noise. Despite having the results that show different values of step-size can be used for different noisy scenario, a fixed value is selected in performance comparison with other reference methods. The evaluation show that the proposed method performs relatively well at very noisy environment, but becomes poorer under higher input SNR conditions due the the tendency to overestimate noise power. This is improved by using speech and noise estimates with shorter averaging time constants. The second approach called the SVAD algorithm, employs a sigmoid function to represent the conditional SPP. The advantage is that the slope and mean of the sigmoid curve can be adjusted independently, thus the SPP can be more flexibly characterised for a compromised trade-off between noise overestimation and underestimation. Also, to cope with two issues: (i) since speech can also be present at very low SNR, and (ii) the transitions between speech absent and present states are slow, the soft decisions are modified by employing different exponential smoothing at different regions of the sigmoid function. This results in similar or better noise tracking and speech quality performance when compared to other evaluated algorithms.

Chapter 6

Multi-channel Wiener Filter

There's a way to do it better -
find it.
– Thomas A. Edison

6.1 Introduction

This chapter describes the limitations in speech distortion weighted MWF (SDW-MWF) and provides novel solutions to improve its performance and also have a practically useful solution. The SDW-MWF is promising as it does not require prior knowledge about the location of the desired speech signal and the microphone characteristics [13, 65]. As a result, it is more robust against microphone mismatch when compared to the well-known beamformer, the generalized sidelobe canceller (GSC) [51]. Similar to the GSC, SDW-MWF relies on a voice activity detection (VAD) algorithm to update the noise statistics in noise-only segments, and the signal statistics during voiced segments. As a VAD estimate is required in practice, wrong estimation often occurs under non-stationary and highly noisy environments, which leads to greater second order estimation errors and causes performance degradation in the SDW-MWF method [16, 52].

Alternatively, the SDW-MWF solution can be decomposed into a rank-one problem, namely the R1-MWF method that consists of a spatial filter and a single-channel postfilter [17, 150]. Although R1-MWF is more robust against the estimation errors, the single-channel postfilter may not be optimal in terms

of spectral tracking, since it is based on correlation matrices that are adapted slowly over time. This has been improved by using a multi-channel speech presence probability (MC-SPP) algorithm to adapt the noise statistics continuously over time [151]. Instead of using MC-SPP, a more direct speech presence probability (SPP) estimate can be obtained by taking one of the microphone inputs as reference, as used in [15] to adapt the parameter that trades off noise reduction and speech distortion. To increase the accuracy in speech detection, both MC-SPP and SPP require accurate estimates of *a priori* speech absence probability (SAP) and *a priori* signal-to-noise ratio (SNR), which in turn increase the processing delay. To avoid this, fixed prior estimates can be used not only to reduce the delay but also to maintain the accuracy in noise tracking in single-channel speech enhancement framework [9, 149].

This thesis proposes a soft-VAD based SDW-MWF solution that aims to reduce the second order statistics errors by avoiding the subtraction of noise-only correlation matrix from the speech-plus-noise correlation matrix. In that case, the desired signal is estimated from the reference microphone by using a single-channel speech enhancement framework from [7], which shows good performance in terms of trade-off among noise reduction, speech distortion and musical noise. The rank-one solution has also been investigated for such alternative formulation, which is faster in spectral tracking based on the single-channel noise reduction algorithm. Although it would require the estimation of the speech correlation matrix, the matrix inverse could be avoided making it more robust to the estimation errors [17]. In addition, the noise power spectral density (PSD) estimate in the reference channel is obtained by the modified SPP with fixed priors approach in [9], which is then employed to continuously update both speech plus noise and noise only second order statistics.

6.2 Multi-channel Wiener Filter

6.2.1 Signal Model and Notation

Let $Y_l(k,m)$, $l = 1, ..., L$, denote the microphone signals in time-frequency domain, where k is the frequency bin index, m is the frame index and L is the

number of microphones. The received signals are given by

$$Y_l\left(k,m\right) = X_l\left(k,m\right) + V_l\left(k,m\right) \tag{6.1}$$

where $X_l\left(k,m\right)$ and $V_l\left(k,m\right)$ are the short-time Fourier transform (STFT) representations of the target signal and the uncorrelated noise components of the l-th microphone, respectively. Here, speech enhancement is performed to remove the unwanted noise while preserving the target speech signal. This can be done by applying a set of filters $\mathbf{w}(k,m)$ to the observed signal, such that

$$Z\left(k,m\right) = \mathbf{w}^H\left(k,m\right)\mathbf{y}\left(k,m\right) \tag{6.2}$$

where Z is the output signal, and $\mathbf{y}\left(k,m\right) \in \mathbb{C}^{L\times 1}$ is a stacked vector given as[1]

$$\begin{aligned}\mathbf{y}\left(k,m\right) &= \left[Y_1\left(k,m\right)\ Y_2\left(k,m\right)\ ,...,\ Y_L\left(k,m\right)\right]^T \\ &= \mathbf{x}\left(k,m\right) + \mathbf{v}\left(k,m\right)\end{aligned} \tag{6.3}$$

with T indicating the transpose operator. The correlation matrices for the noisy speech $\mathbf{R}_y(k,m)$, the clean speech $\mathbf{R}_x(k,m)$, and the background noise $\mathbf{R}_v(k,m)$ are then defined, respectively, as

$$\begin{aligned}\mathbf{R}_y(k,m) &= E\left\{\mathbf{y}(k,m)\mathbf{y}^H(k,m)\right\}, \\ \mathbf{R}_x(k,m) &= E\left\{\mathbf{x}(k,m)\mathbf{x}^H(k,m)\right\}, \\ \mathbf{R}_v(k,m) &= E\left\{\mathbf{v}(k,m)\mathbf{v}^H(k,m)\right\},\end{aligned} \tag{6.4}$$

where E and H denote, respectively, the expected value and Hermitian transpose operators.

6.2.2 Formulation of Multi-channel Wiener Filter

The multi-channel Wiener filter (MWF) optimally estimates the speech signal, based on an MMSE criterion as

$$\mathbf{w}_{\mathrm{MWF}}(k,m) = \arg\min_{\mathbf{w}(k,m)}\ E\left\{\left|X_{\mathrm{ref}}(k,m) - \mathbf{w}^H(k,m)\mathbf{y}(k,m)\right|^2\right\} \tag{6.5}$$

where the desired signal in this case is the unknown speech component X_{ref} from the reference microphone. The drawback is that some residual noise will still

[1] Although the signal vectors contain complex-valued frequency-domain variables, they are denoted with lower-case letters throughout the thesis to distinguish them from matrices.

remain in the output signal, Z, which can be reduced by allowing a trade-off between noise reduction and speech distortion. This can be done by modifying the design criterion of the MWF as [65, 150]

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\mu}(k,m) = \arg \min_{\mathbf{w}(k,m)} \; & E\left\{\left|X_{\mathrm{ref}}(k,m) - \mathbf{w}^H(k,m)\mathbf{x}(k,m)\right|^2\right\} + \\ & \mu\, E\left\{\left|\mathbf{w}^H(k,m)\mathbf{v}(k,m)\right|^2\right\}\end{aligned} \tag{6.6}$$

where speech and noise are assumed to be uncorrelated, and μ is the trade-off parameter. A larger μ value here indicates more residual noise reduction at the expense of higher speech distortion. The solution of MWF_μ can then be obtained as

$$\mathbf{w}_{\mathrm{MWF}_\mu}(k,m) = \left[\mathbf{R}_x(k,m) + \mu\mathbf{R}_v(k,m)\right]^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{ref}} \tag{6.7}$$

where $\mathbf{e}_{\mathrm{ref}} = [0...0\ 1\ 0...0]^T$ is an L-element zero vector with the unity corresponds to the r^{th} element of the microphones. Here, the correlation matrices $\mathbf{R}_y(k,m)$ and $\mathbf{R}_v(k,m)$ can be recursively updated by using a VAD as

$$\mathcal{H}_0(k,m): \begin{cases}\hat{\mathbf{R}}_v(k,m) = (1-\alpha_{vv})\,\hat{\mathbf{R}}_v(k,m-1) + \alpha_{vv}\mathbf{y}(k,m)\mathbf{y}^H(k,m) \\ \hat{\mathbf{R}}_y(k,m) = \hat{\mathbf{R}}_y(k,m-1)\end{cases}$$

$$\mathcal{H}_1(k,m): \begin{cases}\hat{\mathbf{R}}_y(k,m) = (1-\alpha_{yy})\,\hat{\mathbf{R}}_y(k,m-1) + \alpha_{yy}\mathbf{y}(k,m)\mathbf{y}^H(k,m) \\ \hat{\mathbf{R}}_v(k,m) = \hat{\mathbf{R}}_v(k,m-1)\end{cases} \tag{6.8}$$

where $\mathcal{H}_0(k,m)$ and $\mathcal{H}_1(k,m)$ denote speech absence and speech presence in the k^{th} frequency bin of the m^{th} frame, respectively. Both smoothing factors α_{yy} and α_{vv} have to be chosen carefully to reflect the degree of stationarity of speech and noise signals.

From Eq. (6.7), it can be observed that an estimation of $\mathbf{R}_x$ is required, which is usually obtained by [65]

$$\mathbf{R}_x(k,m) = \mathbf{R}_y(k,m) - \mathbf{R}_v(k,m). \tag{6.9}$$

However, estimation errors in both of the complex-valued correlation matrices $\mathbf{R}_y(k,m)$ and $\mathbf{R}_v(k,m)$ can result in a very poor estimate of $\mathbf{R}_x$. Although this can be avoided by obtaining a pre-determined $\mathbf{R}_x(k,m)$ estimate either with a

calibration sequence [63], or by deriving a mathematical model [59, 152]. These methods rely on the *a priori* information, making them less attractive for on-line applications.

6.3 Proposed Method

6.3.1 Formulation of Proposed MWF and Estimation of Noisy and Noise Correlation Matrices

In order to avoid the aforementioned problems, a bi-criteria optimization problem for MWF is proposed. This consists of a criterion to minimise the error in Eq. (6.5) and another criterion to minimise the noise power. One way to formulate such problem is to use the weighted sum between the two criteria as given by

$$\mathbf{w}_{\mathrm{MWF}_\lambda}(k,m) = \arg \min_{\mathbf{w}(k,m)} \; (1-\lambda)\, E\left\{ \left| X_{\mathrm{ref}}(k,m) - \mathbf{w}^H(k,m)\mathbf{y}(k,m) \right|^2 \right\} + \lambda \left(E\left\{ \left| \mathbf{w}^H(k,m)\mathbf{v}(k,m) \right|^2 \right\} \right) \tag{6.10}$$

where λ is a weighting value between 0 and 1. The solution of the problem can then be found as

$$\mathbf{w}_{\mathrm{MWF}_\lambda}(k,m) = \left[(1-\lambda)\, \mathbf{R}_y(k,m) + \lambda \mathbf{R}_v(k,m) \right]^{-1} (1-\lambda)\, \mathbf{r}_{yx}(k,m) \tag{6.11}$$

where $\mathbf{r}_{yx}(k,m) = E\left\{ \mathbf{y}(k,m) X^*_{\mathrm{ref}}(k,m) \right\}$. It can be seen that by formulating the problem in this way, the estimation of the clean speech correlation matrix $\mathbf{R}_x(k,m)$ can be averted. Also, a set of pareto solutions can be found by varying λ, but this is not in the scope of this thesis.

Apart from that, instead of using a VAD to estimate the correlation matrices, the frame and frequency dependant modified SPP, $p(k,m)$ from [9] is employed.

This allows both $\hat{\mathbf{R}}_v(k,m)$ and $\hat{\mathbf{R}}_y(k,m)$ from Eq. (6.8) to be updated as

$$\begin{aligned}
\hat{\mathbf{R}}_v(k,m) &= (1-p(k,m)) \\
&\quad \times \left[(1-\alpha_{vv}(k,m))\,\hat{\mathbf{R}}_v(k,m-1) + \alpha_{vv}(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m)\right] \\
&\quad + p(k,m)\hat{\mathbf{R}}_v(k,m-1) \\
&= [p(k,m) + (1-\alpha_{vv}(k,m))\,(1-p(k,m))]\,\hat{\mathbf{R}}_v(k,m-1) \\
&\quad + \alpha_{vv}(k,m)\,(1-p(k,m))\,\mathbf{y}(k,m)\mathbf{y}^H(k,m) \\
&= (1-\tilde{\alpha}_v(k,m))\,\hat{\mathbf{R}}_v(k,m-1) + \tilde{\alpha}_v(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m)
\end{aligned} \tag{6.12}$$

$$\begin{aligned}
\hat{\mathbf{R}}_y(k,m) &= (1-p(k,m))\,\hat{\mathbf{R}}_y(k,m-1) + p(k,m) \\
&\quad \times \left[(1-\alpha_{yy}(k,m))\,\hat{\mathbf{R}}_y(k,m-1) + \alpha_{yy}(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m)\right] \\
&= [(1-p(k,m)) + p(k,m)\,(1-\alpha_{yy}(k,m))]\,\hat{\mathbf{R}}_y(k,m-1) \\
&\quad + p(k,m)\alpha_{yy}(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m) \\
&= (1-\tilde{\alpha}_y(k,m))\,\hat{\mathbf{R}}_y(k,m-1) + \tilde{\alpha}_y(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m)
\end{aligned} \tag{6.13}$$

where $\tilde{\alpha}_v(k,m)$ and $\tilde{\alpha}_y(k,m)$ denote, respectively, $\tilde{\alpha}_v(k,m) = \alpha_{vv}\,(1-p(k,m))$ and $\tilde{\alpha}_y(k,m) = \alpha_{yy}p(k,m)$. Here, α_{vv} and α_{yy} denote, respectively, the fixed smoothing factor for noise correlation matrix and speech plus noise correlation matrix.

6.3.2 Formulation of Rank-one MWF

In the case of a single target speech source, the speech signal vector can be defined as

$$\mathbf{x}(k,m) = \mathbf{a}(k,m)S(k,m) \tag{6.14}$$

where the L-dimensional steering vector $\mathbf{a}(k,m)$ contains the acoustic transfer functions from the speech source to the microphones and S is the speech signal. The speech correlation matrix is thus a rank-one matrix, such that

$$\mathbf{R}_x(k,m) = \Phi_{ss}\mathbf{a}(k,m)\mathbf{a}^H(k,m) \tag{6.15}$$

where $\Phi_{ss} = E\{|S|^2\}$ is the power of the speech signal.

Another way to write Eq. (6.11) is

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}}(k,m) &= \left[(1-\lambda)\,\mathbf{R}_x(k,m) + (1-\lambda)\,\mathbf{R}_v(k,m) + \lambda\mathbf{R}_v(k,m)\right]^{-1} \\ &\quad \times (1-\lambda)\,\mathbf{r}_{yx}(k,m) \\ &= \left[\mathbf{R}_x(k,m) + \psi\mathbf{R}_v(k,m)\right]^{-1}\mathbf{r}_{yx}(k,m)\end{aligned} \tag{6.16}$$

where $\psi = \frac{1}{1-\lambda}$. However, this expression contains $\mathbf{R}_x(k,m)$. To reduce the error, the inverse of $\mathbf{R}_x(k,m)$ needs to be avoided. By using matrix inversion lemma and rank-one source assumption from Eq. (6.15), Eq. (6.16) can be written as

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\lambda}(k,m) &= \left(\frac{1}{\psi}\mathbf{R}_v^{-1}(k,m) - \frac{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\mathbf{R}_v^{-1}(k,m)}{\psi\left(1+\psi^{-1}\Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\right)}\right) \\ &\quad \times \mathbf{r}_{yx}(k,m) \\ &= \frac{1}{\psi}\mathbf{R}_v^{-1}(k,m)\left(\mathbf{I} - \frac{\Phi_{ss}\mathbf{a}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)}{\psi+\Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)}\right) \\ &\quad \times \Phi_{ss}F\mathbf{a}(k,m) \\ &= \frac{1}{\psi}\mathbf{R}_v^{-1}(k,m)\left(\mathbf{a}(k,m) - \frac{\Phi_{ss}\mathbf{a}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)}{\psi+\Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)}\right) \\ &\quad \times \Phi_{ss}F(k,m) \\ &= \frac{1}{\psi}\mathbf{R}_v^{-1}(k,m)\left(1 - \frac{\Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)}{\psi+\Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)}\right) \\ &\quad \times \Phi_{ss}F(k,m)\mathbf{a}(k,m) \\ &= \frac{1}{\psi}\mathbf{R}_v^{-1}(k,m)\left(\frac{\psi}{\psi+\Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)}\right) \\ &\quad \times \Phi_{ss}F(k,m)\mathbf{a}(k,m).\end{aligned} \tag{6.17}$$

That can be written in a rank-one formulation, which does not involve speech power and steering vectors, as

$$\mathbf{w}_{\mathrm{MWF}_\lambda}\text{−rank1}(k,m) = \frac{\mathbf{R}_v^{-1}(k,m)\mathbf{r}_{yx}(k,m)}{\psi+\varrho} \tag{6.18}$$

with

$$\begin{aligned}\varrho(k,m) &= \Phi_{ss}\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m) \\ &= \Phi_{ss}\mathrm{Tr}\left\{\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\right\} \\ &= \mathrm{Tr}\left\{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\right\}.\end{aligned} \tag{6.19}$$

6.3.3 Employing Single-channel Algorithm

From Eq. (6.11), it can be seen that the proposed solution requires an estimate of the clean speech reference $X_{\text{ref}}(k,m)$. As opposed to previous methods [59, 63, 152], we propose to estimate $X_{\text{ref}}(k,m)$ by utilising a single-channel speech enhancement method and to use one microphone in the array as a reference. As such, the raw estimate of $\mathbf{r}_{yx}(k,m)$ can be defined as

$$\hat{\mathbf{r}}_{yx}(k,m) = \mathbf{y}(k,m)G(k,m)(X^*_{\text{ref}}(k,m) + V^*_r(k,m)) \tag{6.20}$$

where $X_{\text{ref}}(k,m) = A_{\text{ref}}(k,m)S(k,m)$ with $A_{\text{ref}}(k,m)$ denotes the acoustic transfer function (ATF) of the target speech signal, $S(k,m)$ at the reference channel. Here, $G(k,m)$ is a spectral weighting gain function, which involves the computation of the *a posteriori* and *a priori* SNR estimates. In contrast to $\mathbf{R}_x(k,m)\mathbf{e}_{\text{ref}} = (\mathbf{R}_y(k,m) - \mathbf{R}_v(k,m))\mathbf{e}_{\text{ref}}$ from Eq. (6.7) and Eq. (6.9), which takes the reference vector directly from the second order clean speech estimate, Eq. (6.20) uses an SNR based gain function to adapt the noisy stacked vectors to the desired clean speech signal. Such implementation is capable of generating a better clean speech estimate and improving the speech quality of the enhanced signal.

In this thesis, $G(k,m)$ in Eq. (6.20) is taken from the modified sigmoid (MSIG) gain function from Chapter 4 [7]. As the beamformer tries to adapt to the clean speech reference, an important aspect of the single-channel estimate is that the speech distortion has to be as small as possible. This can be done by setting smaller values for the SNR smoothing parameters from [7], i.e., $\beta \approx 0.9$ and $\alpha_y \approx 0$, such that the amount of speech distortion can be kept as low as possible while not having a large amount of musical noise. Apart from that, further reduction of musical noise is proposed by having $\hat{\mathbf{r}}_{yx}(k,m)$ updated recursively as

$$\hat{\mathbf{r}}_{yx}(k,m) = (1-\alpha_x)\hat{\mathbf{r}}_{yx}(k,m-1) + \alpha_x \mathbf{y}\,(k,m)\,\hat{X}^*_{\text{ref}}\,(k,m) \tag{6.21}$$

where α_x is the smoothing factor for target speech signal, and $\hat{X}_{\text{ref}}(k,m) = G(k,m)(X_{\text{ref}}(k,m) + V_r(k,m))$ indicates the clean speech estimate from the reference microphone.

The SNR estimates for $G(k,m)$ require the estimation of the noise PSD at the

reference channel. Here, the soft VAD (SVAD) noise PSD estimate in Chapter 5, which involves the calculation of the modified SPP [9], is used. This implies that the same SPP estimate can be used for estimating the noise PSD in the reference channel and also the correlation matrices in Eqs. (6.12) and (6.13).

6.4 Performance evaluation

Measurements were performed with 2 microphones (with inter-element space of 1 cm) embedded in the left side of a pair of earmuffs on a manikin so that the head-shadowing effect is included. The manikin was placed close to the center of a room with dimensions 3.05 m × 3.05 m, with a reverberation time T_{60} of approximately 0.2 s. The loudspeakers were positioned at 1 m from the center of the head, with the speech located at 0° and the non-stationary factory noise rendered at 45°, 90°, 135°, 180°, 225°, 270° and 315° to the left of the head. The speech signals consists of 5 (2 male and 3 female) sentences with length ranging from 11 s to 22 s. The signals were sampled at $f_s = 16$ kHz. An STFT length of $K = 512$ was used with a frame rate of $R = 256$ and a square-root Hann window.

Evaluation includes $\mathbf{w}_{\mathrm{MWF}_\mu}$ from Eq. (6.6) with $\mu = 5$, the output signal from reference microphone using MSIG function with a noise floor of -15 dB , $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ from Eq. (6.11) with $\lambda \approx (\mu - 1)/\mu = 0.8$ and $\mathbf{w}_{\mathrm{MWF}_{\lambda 2}}$ with $\lambda(k,m) = 1 - p(k,m)$. The smoothing constants are estimated by $\alpha = \exp(\frac{-2.2R}{tf_s})$, with $t_x = t_y = 0.02$ s and $t_v = 2$ s. The performance is measured by the speech intelligibility weighted segmental SNR in frequency domain (IFWSNRseg) [67, 153]

$$\mathrm{IFWSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{k=0}^{K-1} B_k \log_{10}\frac{\mathcal{A}^2(k,m)}{\mathcal{A}^2(k,m)-\hat{\mathcal{A}}^2(k,m)}}{\sum_{k=0}^{K-1} B_k} \tag{6.22}$$

where B_k is the ANSI SII weight placed on the k^{th} frequency bin [154], K is the number of bands, M is the number of frames, $\mathcal{A}(k,m)$ and $\hat{\mathcal{A}}(k,m)$ are spectrum amplitudes of the clean speech signal and enhanced speech signal, respectively. Each frame is threshold by a -10 dB lower bound and a 35 dB upper bound to discard non-speech frames.

In addition, segmental noise attenuation (NATTseg) and segmental speech

preservation (SPREseg) measures are utilised to study if a difference in IFWS-NRseg is due to more noise reduction or less speech distortion. Both are given, respectively, by [155]

$$\text{NATTseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{||\mathbf{v}_t(m)||^2}{||\tilde{\mathbf{v}}_t(m)||^2} \tag{6.23}$$

$$\text{SPREseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{||\mathbf{x}_t(m)||^2}{||\mathbf{x}_t(m)-\tilde{\mathbf{x}}_t(m)||^2}. \tag{6.24}$$

Here, $\mathbf{v}_t(m)$ and $\mathbf{x}_t(m)$ are m-th frame time-domain vectors for the noise and the clean speech signal, respectively. The signals $\tilde{\mathbf{v}}_t(m)$ and $\tilde{\mathbf{x}}_t(m)$ indicate both noise and the clean signals processed with the same corresponding filters as used to enhance the noisy signal. The widely-used perceptual evaluation of speech quality (PESQ) measure has also been included for performance comparison [67].

Figures 6.1-6.4 show the averaged results for SNRs of -5 dB, 0 dB, 5 dB, and 10 dB, respectively. It can be observed that $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ outperforms $\mathbf{w}_{\text{MWF}_{\mu}}$ for all objective measures in all scenarios, indicating that the proposed method allows more noise suppression, yet does not come with higher speech distortion. When $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ is compared to MSIG and $\mathbf{w}_{\text{MWF}_{\lambda 2}}$, it generally has larger noise reduction but larger speech distortion as well. This is the reason why $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ performs better relatively to other evaluated methods at low input SNR conditions but has a performance drop when the input SNR increases, as shown in IFWSNRseg and PESQ results. While $\mathbf{w}_{\text{MWF}_{\lambda 2}}$ improves the performance of $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ in terms of less speech distortion, more musical noise is audible since the NATTseg values are much lower than $\mathbf{w}_{\text{MWF}_{\lambda 1}}$. When compared to MSIG from the reference microphone, $\mathbf{w}_{\text{MWF}_{\lambda 2}}$ has higher noise reduction but also larger speech distortion. However, since MWF involves temporal averaging in the second order statistics estimation, the musical noise can be reduced, especially at low SNR conditions, as indicated by IFWSNRseg results from Figure 6.1.

Figures 6.5-6.8 depict the results for $\mathbf{w}_{\text{MWF}_{\mu}}$-rank1 and $\mathbf{w}_{\text{MWF}_{\lambda}}$-rank1, where $\lambda = \lambda_1$ has been used. The results of $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ have also been plotted for comparison. It can be clearly seen that $\mathbf{w}_{\text{MWF}_{\lambda}}$-rank1 performs better than the others in IFWSNRseg, SPREseg and PESQ scores at low input SNRs. However, it has lower noise reduction performance when compared to $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ as indicated by

the NATTseg results. This means that $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ might have over-attenuated the noise causing more speech distortion, thus resulting in poorer performance in the IFWSNRseg and PESQ scores. It is interesting to show that when the input SNR increases, the amount of noise reduction for $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ is dropping relatively faster than the rank-one methods. This phenomenon leads to the fact that $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ has the best performance recorded at high input SNR, while the performance of $\mathbf{w}_{\mathrm{MWF}_{\mu}}$-rank1 is merely between $\mathbf{w}_{\mathrm{MWF}_{\lambda}}$-rank1 and $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$.

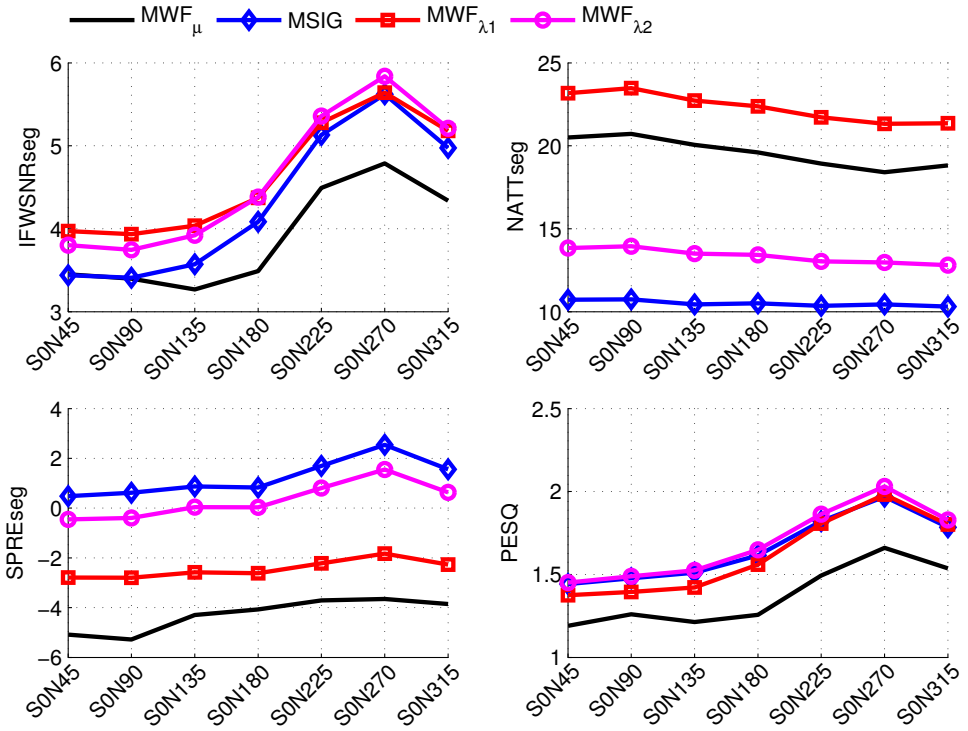

Figure 6.1: Comparison among $\mathbf{w}_{\mathrm{MWF}_\mu}$, MSIG, $\mathbf{w}_{\mathrm{MWF}_{\lambda1}}$, and $\mathbf{w}_{\mathrm{MWF}_{\lambda2}}$ for factory noise for input SNR -5 dB. Labels $SANB$ on x-axis indicate the directions of the target speech and noise, where S stands for speech and N stands for noise; A and B represent the directions of the target speech and noise, respectively.

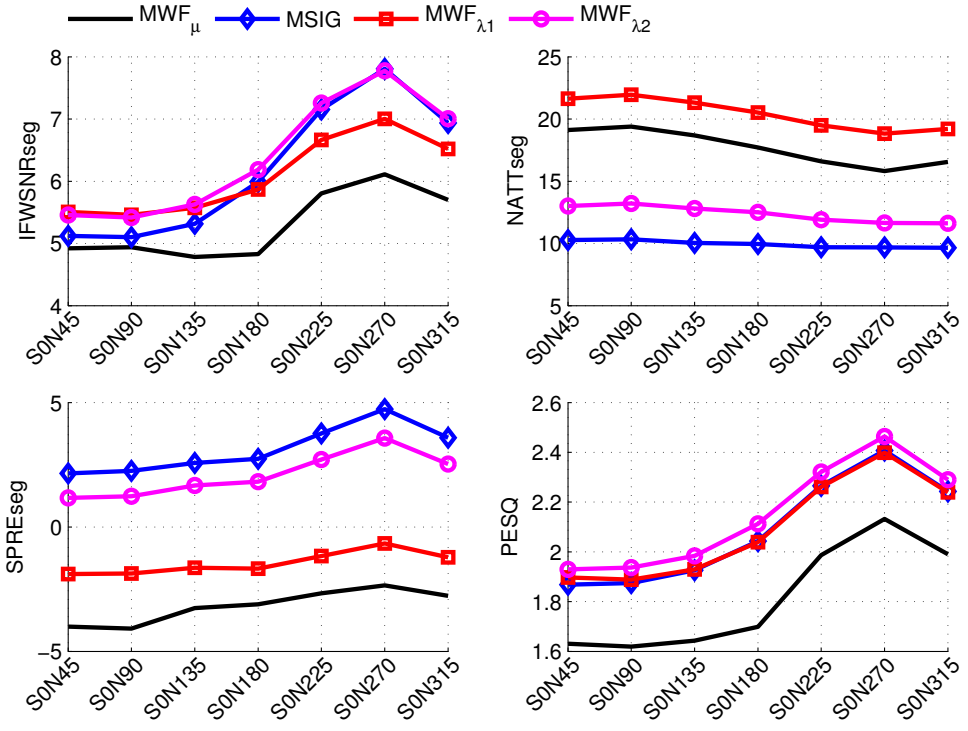

Figure 6.2: Comparison among $\mathbf{w}_{\mathrm{MWF}_\mu}$, MSIG, $\mathbf{w}_{\mathrm{MWF}_{\lambda1}}$, and $\mathbf{w}_{\mathrm{MWF}_{\lambda2}}$ for factory noise for input SNR 0 dB.

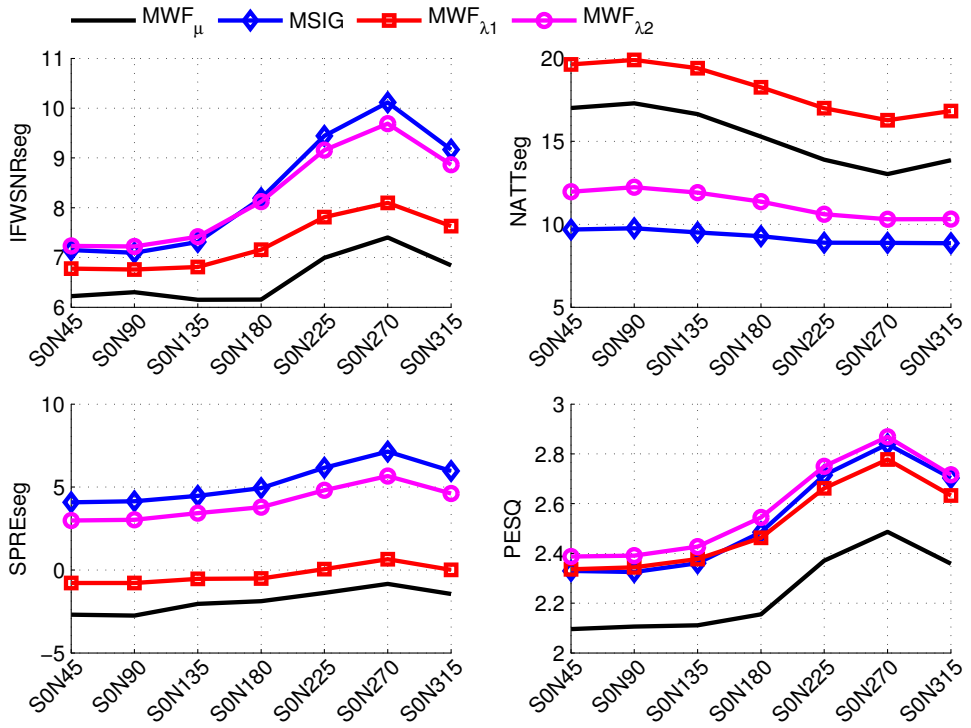

Figure 6.3: Comparison among $\mathbf{w}_{\mathrm{MWF}_\mu}$, MSIG, $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$, and $\mathbf{w}_{\mathrm{MWF}_{\lambda 2}}$ for factory noise for input SNR 5 dB.

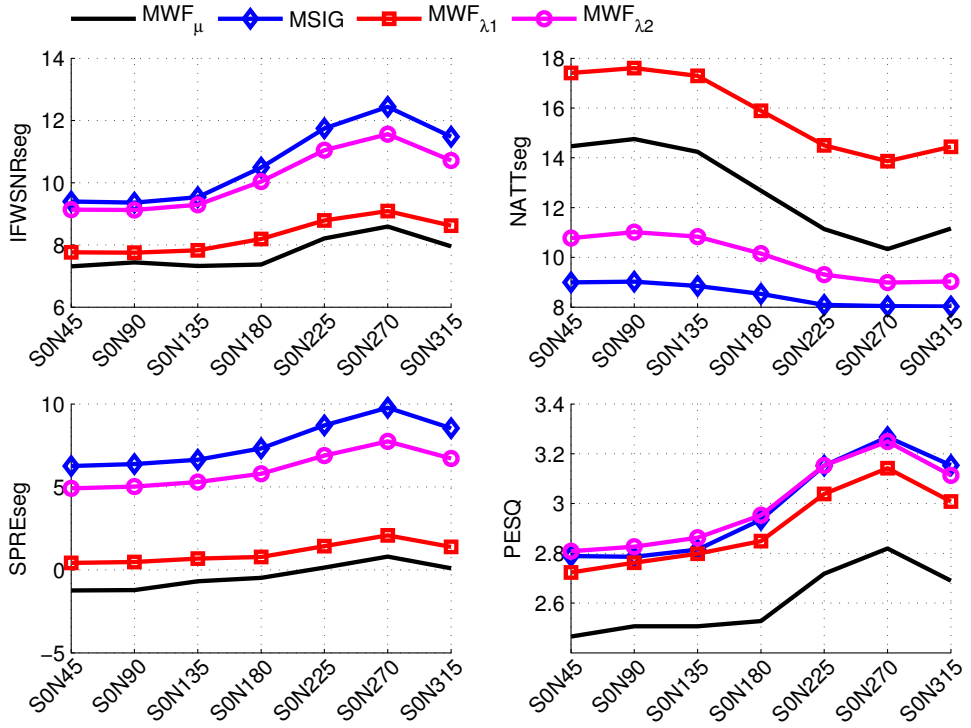

Figure 6.4: Comparison among $\mathbf{w}_{\mathrm{MWF}_\mu}$, MSIG, $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$, and $\mathbf{w}_{\mathrm{MWF}_{\lambda 2}}$ for factory noise for input SNR 10 dB.

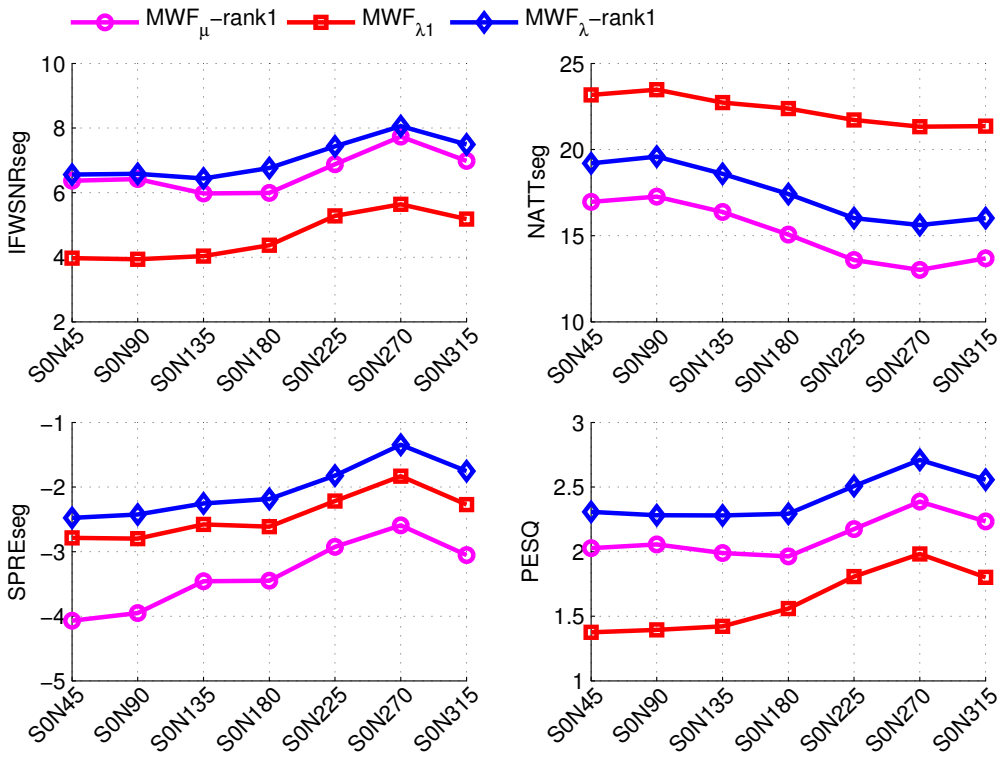

Figure 6.5: Comparison between rank-one and general formulations for factory noise for input SNR −5 dB.

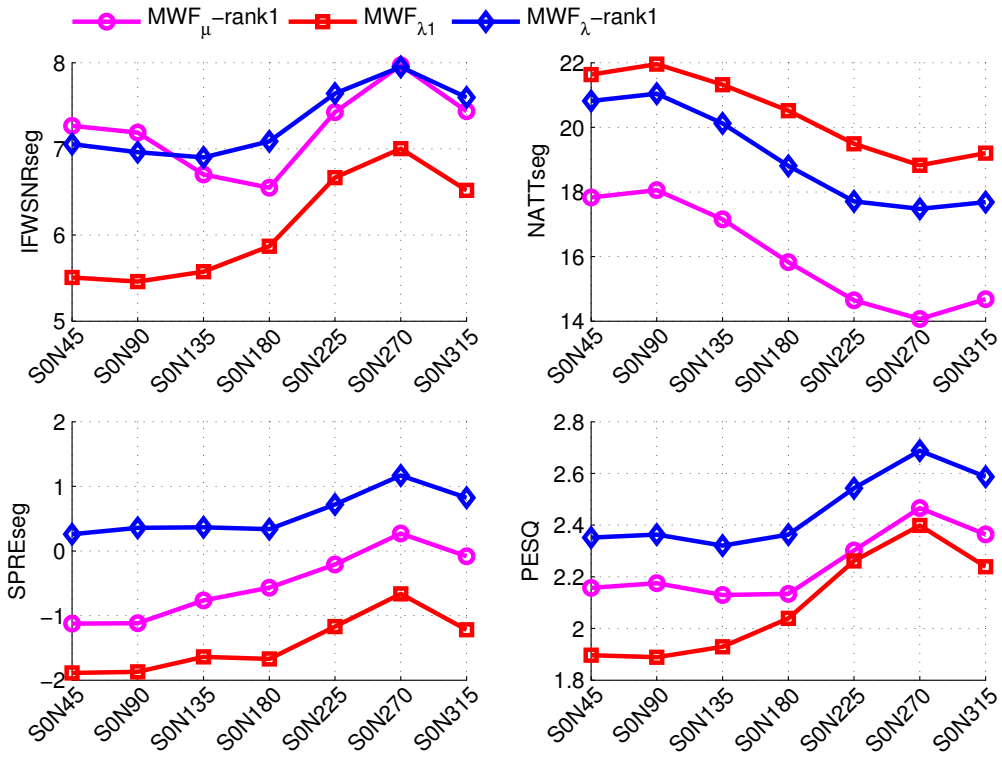

Figure 6.6: Comparison between rank-one and general formulations for factory noise for input SNR 0 dB.

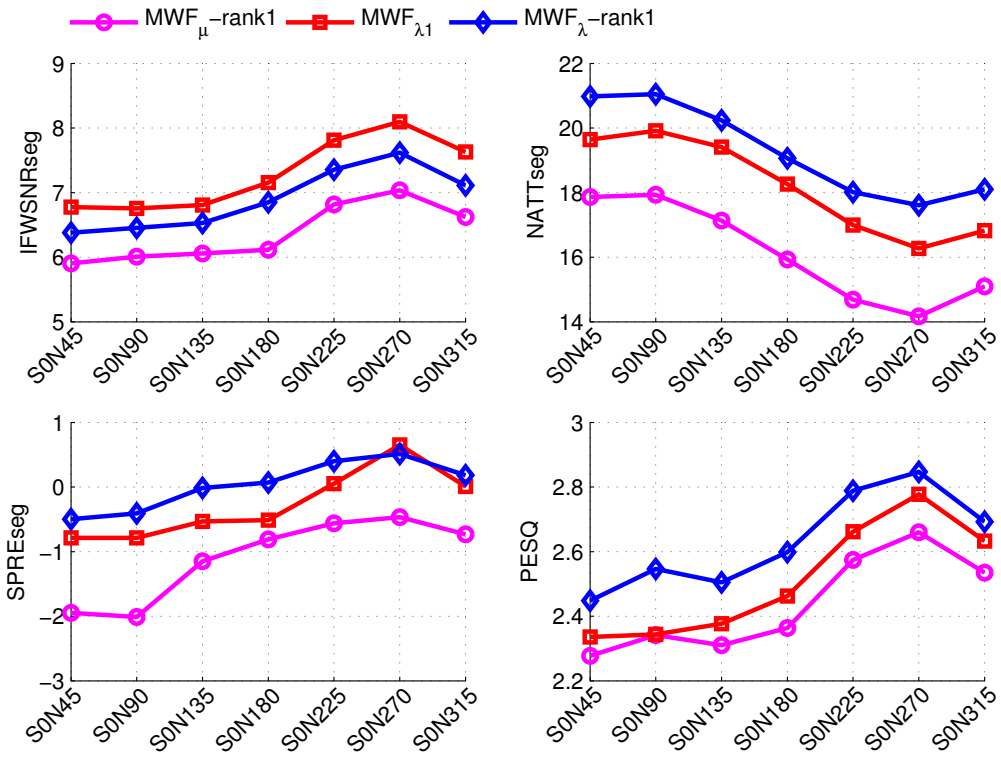

Figure 6.7: Comparison between rank-one and general formulations for factory noise for input SNR 5 dB.

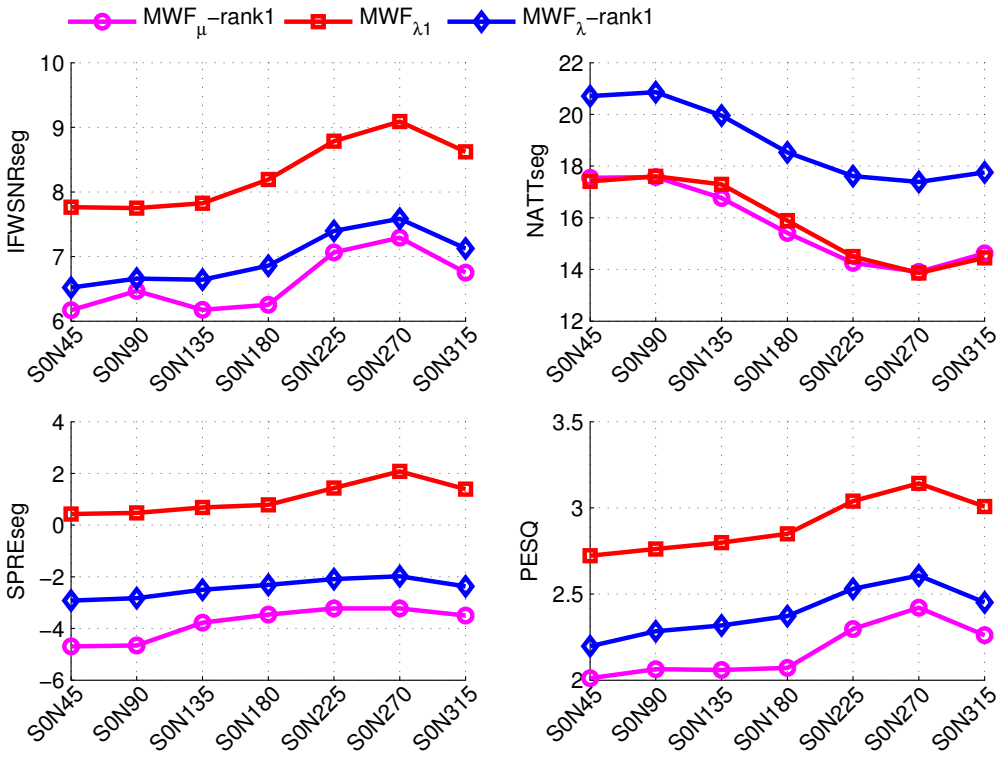

Figure 6.8: Comparison between rank-one and general formulations for factory noise for input SNR 10 dB.

6.5 Summary

This chapter is focused on an alternative SDW-MWF formulation that does not require the clean speech correlation matrix estimate, as opposed to previous formulations in the literature [17, 65, 150]. Furthermore, as in contrast to work that requires calibration [63] or pre-calculation using a mathematical model [59, 152], this work utilises single-channel noise reduction technique to estimate a reference channel, which as far as we are aware has not been considered before. In addition, unlike the previous approach, where SPP was only used to adapt the trade-off parameter [15], or only to estimate the noise correlation matrix [151], it is fully utilised in the proposed framework in estimating the noise PSD in the reference channel and also both noisy and noise correlation matrices. The rank-one formulation for the proposed SDW-MWF is also developed for comparison. Experimental results show that the proposed method outperforms the traditional method for all performance measures. The rank-one solution has recorded better performance than the general formulation in low SNR conditions. The incorporation of SPP in the trade-off parameter λ helps to reduce speech distortion, but there is a trade-off in terms of residual noise and "musical-type" artefacts in the enhanced signals. Such formulation will be extended to binaural configuration in Chapter 7.

Chapter 7

Binaural Noise Reduction Frameworks

The real voyage of discovery consists not
in seeking new landscapes but
in having new eyes.
– Marcel Proust

7.1 Introduction to Binaural Signal Processing

This chapter focuses on developing binaural noise reduction algorithms for speech enhancement in hearing protection devices (HPDs). Several binaural techniques have been proposed in recent years for future hearing aids, where the full-duplex exchange of microphone signals between the two devices is feasible. For hearing protectors where microphones are integrated into the hearing protector adjacent to each ear, the microphone can be connected by cables, binaural processing algorithms can be readily applied. The aim of binaural noise reduction techniques is to improve the signal-to-noise ratio (SNR) of the signal, while simultaneously preserving the binaural cues of both target speech and residual noise. The binaural noise reduction techniques can be roughly divided into two classes. In the first class, identical real-valued spectral gains are applied to one microphone signal on the left device and one microphone signal on the right device [156–160],

so that the binaural cues are indeed preserved. Although the outputs of a beamformer can be utilised to derive the spectral gain function (e.g., a superdirective beamformer [158]), in essence these techniques can be viewed as spectral filtering approaches, similar to techniques in single-channel noise reduction. As previously discussed, single-channel noise reduction usually introduces speech distortion and other artefacts, leading to limited or no speech intelligibility improvements. The perceptual speech intelligibility improvements obtained with this class of binaural techniques may thus also be limited, or not competitive with true beamforming techniques. Also, another drawback of these techniques is that the interfering sources located in the back direction cannot be suppressed due to the forward-backward ambiguity. In order to suppress (noise) signals originating from the back, differential microphone arrays (DMAs), which requires at least two microphones placed at each side of the ears, can be used.

The second class of binaural techniques combines all microphone signals from both ears to perform a true beamforming. Some techniques first construct a monaural output and then apply a postprocessing stage to reconstruct the binaural signals with correct binaural cues [161]. Other techniques apply fixed or adaptive beamformers which produce a binaural ouput, whereby the beamformers are designed or constrained so that the binaural cues are also preserved [11, 161–163]. For example, in [11, 162] (adaptive) beamforming is only applied in the higher frequencies, while for the lower frequencies the (low-pass filtered) original microphone signal with correct binaural cues is used. This approach is unsuitable for the context of this thesis as the noise components are mainly the industrial noise which consists of low frequency components. The binaural multi-channel Wiener filter (MWF) technique also performs a true beamforming with the microphone signals in order to produce binaural output signals, hence it belongs to this second class of techniques [65]. However, the conventional MWF method can only preserve binaural cues for speech but not for noise. Although the MWF cost function has been extended such that the binaural cues of both the target speech and the residual noise can be preserved, there is a trade-off between binaural cue preservation and noise reduction performance [14]. Such trade-off does not occur in the first class of binaural techniques, but these techniques may offer only

limited speech intelligibility improvements.

Therefore, we attempt to combine concepts from both classes of binaural techniques in this chapter, i.e., applying real-valued gain functions as well as applying beamforming techniques in order to achieve high noise suppression and less speech distortion with binaural cues maintained. From the starting point of human sound perception, two possible binaural noise suppression concepts are introduced differing in the following major aspect: The first approach applies DMAs and single-channel techniques to both sides of the devices, namely the DMA-BPF algorithm, while the second approach utilises a beamforming technique, namely a binaural MWF. The formulation of both frameworks will be given first, then proposed algorithms from previous chapters, which include the new MWF framework, the proposed modified sigmoid (MSIG) gain function and modified decision-directed (MDD) *a priori* SNR estimate for the single-channel speech enhancement algorithm, and the soft VAD (SVAD) noise power spectral density (PSD) estimation method, will be applied into the frameworks to improve the performance. There are four research questions discussed and investigated in this chapter: (i) What is the influence of the binaural noise reduction algorithm with DMAs in a dual monaural configuration and a binaural postfilter that combines two identical single-channel gain functions on the ability to localise sources? (ii) Does the single-channel noise estimation algorithm affect localisation and noise reduction performance compared to the Blocking Matrix algorithm and crossPSD algorithm? (iii) How does the proposed MWF perform in terms of combining localisation and noise reduction performance in comparison to the conventional speech distortion weighted MWF (SDW-MWF) approach? (iv) How does the performance of the MWF compare to the binaural technique with DMAs and single-channel gain functions?

7.1.1 Beamforming in Binaural Context

Consider a speech-in-noise scenario as in Figure 7.1. Each microphone l observes a signal $Y_l\left(\mathrm{e}^{j\Omega}\right)$, which consists of a target speech component $X_l(\mathrm{e}^{j\Omega})$ and an (unwanted) noise component $V_l\left(\mathrm{e}^{j\Omega}\right)$. The observed l-th microphone signal, $1 \leq$

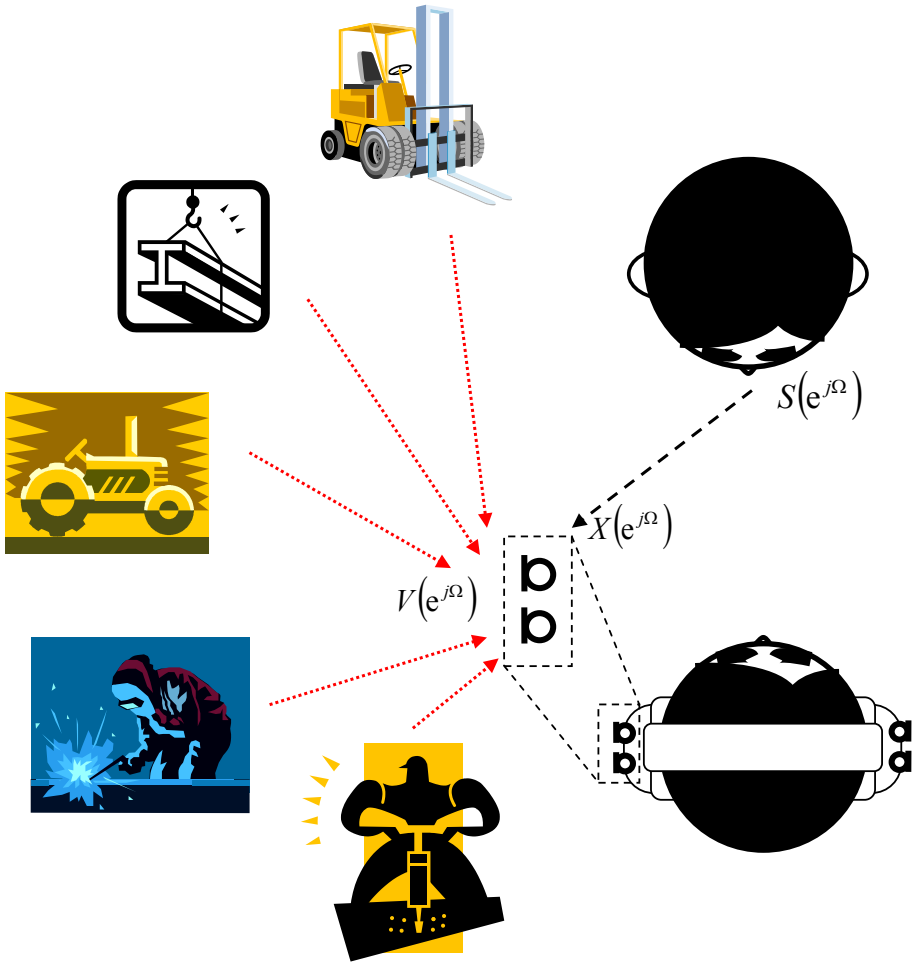

Figure 7.1: Signals arriving on the multi-microphone array of a hearing protection device, in a speech-in-noise scenario.

$l \leq L$ in the hearing aid $Y_l(e^{j\Omega})$ can be defined as

$$Y_l(e^{j\Omega}) = X_l(e^{j\Omega}) + V_l(e^{j\Omega}). \tag{7.1}$$

As in Figure 7.1, the target speech component can be a filtered version of a speech signal $S(e^{j\Omega})$ produced by a target speaker, such that

$$X_l(e^{j\Omega}) = A_l(e^{j\Omega})S(e^{j\Omega}) \tag{7.2}$$

where $A_l(e^{j\Omega})$ contains the complete acoustic transfer function (ATF) from the speech source to l-th microphone (which includes the head-related shadow effect, the reflections against walls and objects, and the microphone characteristics). A more general case, in which $X_l(e^{j\Omega})$ can be a superposition of multiple target

speech sources that are all of interest to the user, can also be considered. The noise component $V_l(\mathrm{e}^{j\Omega})$ contains all unwanted signals to be suppressed by the noise reduction. The noise can be a superposition of different localised noise sources as shown in Figure 7.1, but can also include diffuse noise or spatially uncorrelated noise.

In the context of binaural signal processing, the objective is to obtain one enhanced single-channel signal for each ear, where different types of spatial filtering can be applied. Ideally, these two beamformer output signals should contain the same amount of information about the target source (position) as the original signals at the ears, i.e., the interaural time and level differences should be preserved by the beamformer. To analyse the impact of beamforming on the binaural cues, a scenario with Q point sources $S_q(\mathrm{e}^{j\Omega})$, $q = 1, ..., Q$, and a total number of L microphones being attached to the ears, is considered. Here, $\mathrm{e}^{j\Omega}$ is the frequency-domain variable and Ω is the normalised frequency. The l-th microphone signal can then be expressed as

$$X_l(\mathrm{e}^{j\Omega}) = \sum_{q=1}^{Q} A_{ql}(\mathrm{e}^{j\Omega}, \phi_{s_q}) S_q(\mathrm{e}^{j\Omega}) \tag{7.3}$$

where $l = 1, ..., L$ and $A_{ql}(\mathrm{e}^{j\Omega}, \phi_{s_q})$ denotes the transfer function from the q-th source to the l-th microphone. Spatial filtering at the left side can be written in a general form as

$$\begin{aligned} Z_{\text{left}}(\mathrm{e}^{j\Omega}) &= \sum_{l=1}^{L} W_{l,\text{left}}(\mathrm{e}^{j\Omega}) X_l(\mathrm{e}^{j\Omega}) \\ &= \sum_{q=1}^{Q} \sum_{l=1}^{L} W_{l,\text{left}}(\mathrm{e}^{j\Omega}) A_{ql}(\mathrm{e}^{j\Omega}, \phi_{s_q}) S_q(\mathrm{e}^{j\Omega}) \end{aligned} \tag{7.4}$$

where $W_{l,\text{left}}(\mathrm{e}^{j\Omega})$ is the filter applied to the l-th microphone signal on the left side. The beamformer output at the right side $Z_{\text{right}}(\mathrm{e}^{j\Omega})$ can be written similarly to Eq. (7.4). If the spatial impression of an acoustic scene is to be preserved, the head-related transfer functions (HRTFs), which indicate transfer functions from the source to the respective ear in a free-field, should be maintained when processing the signals. This requires a free-field assumption, where each source should be maintained in free-field situation, based on the Green's function as

given by

$$g(r - r_0) = \frac{\exp(-j\frac{\omega}{c}r)}{4\pi|r - r_0|} \tag{7.5}$$

where ω is the angular frequency, c is the speed of sound, and $|r-r_0|$ is the distance between the location points r and r_0. Thus, for each individual source, the beamformer output must be identical to the actual signals at the ears themselves. For a single sound source located at angle ϕ_s, the signals at the respective ear can be modeled as

$$\begin{aligned} & X_{\text{left}}(\mathrm{e}^{j\Omega}) = H_{\text{left}}(\mathrm{e}^{j\Omega}, \phi_s)S(\mathrm{e}^{j\Omega}) = \alpha_{\text{left}}(\mathrm{e}^{j\Omega}, \phi_s)\mathrm{e}^{-j\Omega f_\mathrm{s}\tau_{\text{left}}(\phi_s)}S(\mathrm{e}^{j\Omega}), \\ \text{and} \quad & X_{\text{right}}(\mathrm{e}^{j\Omega}) = H_{\text{right}}(\mathrm{e}^{j\Omega}, \phi_s)S(\mathrm{e}^{j\Omega}) = \alpha_{\text{right}}(\mathrm{e}^{j\Omega}, \phi_s)\mathrm{e}^{-j\Omega f_\mathrm{s}\tau_{\text{right}}(\phi_s)}S(\mathrm{e}^{j\Omega}) \end{aligned} \tag{7.6}$$

where $H_{\text{left}}(\mathrm{e}^{j\Omega}, \phi_s)$ and $H_{\text{right}}(\mathrm{e}^{j\Omega}, \phi_s)$ denote the HRTFs. Here, f_s denotes the sampling frequency, $\alpha_{\text{left}}(\mathrm{e}^{j\Omega}, \phi_s)$, $\tau_{\text{left}}(\phi_s)$, $\alpha_{\text{right}}(\mathrm{e}^{j\Omega}, \phi_s)$, and $\tau_{\text{right}}(\phi_s)$ denote the frequency-dependent attenuation factor and delay, at the left side and the right side, respectively.

Comparing Eqs. (7.4) and (7.6) for the left side of the ear, it can easily be seen that for them to be identical, the condition

$$\sum_{l=1}^{L} W_{l,\text{left}}(\mathrm{e}^{j\Omega})H_{ql}(\mathrm{e}^{j\Omega}, \phi_{s_q}) = H_{\text{left}}^{(q)}(\mathrm{e}^{j\Omega}, \phi_{s_q}), \quad \forall q \tag{7.7}$$

must be fulfilled. The Q constraints define a system of linear equations, where the number of available degrees of freedom is determined by the number of microphones L. In general, a solution only exists for the case $L \geq Q$, i.e., the number of microphones needs to be greater or equal to the number of sources. Otherwise the system is underdetermined and cannot be solved. This implies that the spatial impression can in principle only be preserved for a certain number of point sources, which is limited by the number of utilised microphones. For complex noise fields, which may be modeled with an infinite number of point sources, the hearing impression cannot be maintained.

7.2 DMA-BPF

Figure 7.2 shows a binaural noise suppression concept based on DMAs and two single-channel speech enhancement algorithms. Both sides of the earmuff accommodate two microphones, such that a DMA can be realised on the left and right

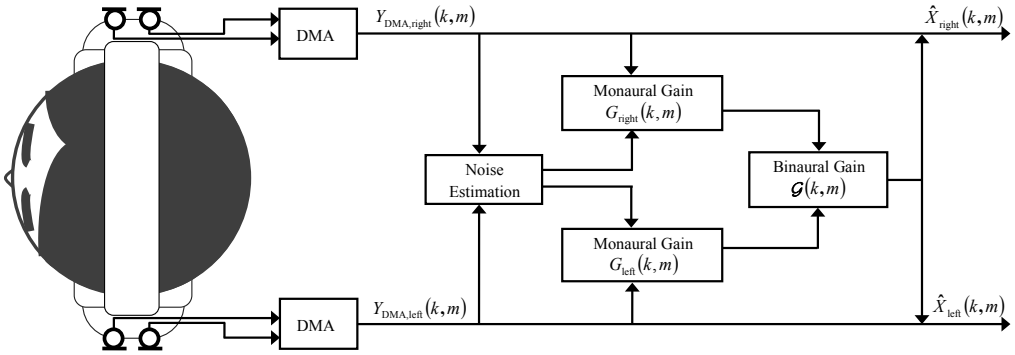

Figure 7.2: A binaural noise suppression approach with two DMAs, a noise PSD estimator, and two single-channel gain functions, which are combined to form a binaural gain function.

side. For each DMA output signal, the respective noise components are estimated and are then used to design two independent single-channel noise suppression filters. Since, in general, the left and right channels are not identical, the resulting filters will also be different. Processing both channels differently may corrupt the binaural cues required for localising sound sources. Therefore, both filters are merged to a single binaural gain function applied to both channels. The different realisations of the individual blocks are discussed in detail subsequently.

7.2.1 DMAs Incorporating HRTFs

First-order DMAs can be realised using two microphones, which are placed on each side of the head as shown in Figure 7.2. For sake of simplicity, we only consider a single sound source here. Since the distance between the microphones of a DMA is inherently small, it can be assumed that the corresponding HRTFs have an identical magnitude and only differ in phase. Hence, they can be related to each other at the left side as

$$H_{\text{left},2}(\mathrm{e}^{j\Omega}, \phi_s) = H_{\text{left},1}(\mathrm{e}^{j\Omega}, \phi_s)\mathrm{e}^{-j\Omega f_s \frac{d}{c}\cos\phi_s} \tag{7.8}$$

and at the right side as

$$H_{\text{right},2}(\mathrm{e}^{j\Omega}, \phi_s) = H_{\text{right},1}(\mathrm{e}^{j\Omega}, \phi_s)\mathrm{e}^{-j\Omega f_s \frac{d}{c}\cos\phi_s} \tag{7.9}$$

where d is the distance between microphones. The output signal of the DMA at the left side is then given by

$$\begin{aligned} Y_{\text{DMA,left}}(\mathrm{e}^{j\Omega}, \phi_s) &= S(\mathrm{e}^{j\Omega}) H_{\text{left},1}(\mathrm{e}^{j\Omega}, \phi_s) - S(\mathrm{e}^{j\Omega}) H_{\text{left},2}(\mathrm{e}^{j\Omega}, \phi_s) \mathrm{e}^{-j\Omega f_\mathrm{s} \tau} \\ &= S(\mathrm{e}^{j\Omega}) H_{\text{left},1}(\mathrm{e}^{j\Omega}, \phi_s) \left(1 - \mathrm{e}^{-j\Omega f_\mathrm{s}\left(\tau - \frac{d}{c}\cos\phi_s\right)}\right) \end{aligned} \tag{7.10}$$

where τ_{DMA} is the delay. The HRTFs corresponding to the front microphones appear at the output of the DMAs. The noisy output signal of the DMAs after applying the compensation filter is now expressed as

$$Y_{\text{DMA,left}}(\mathrm{e}^{j\Omega}) = S(\mathrm{e}^{j\Omega}) H_{\text{left}}(\mathrm{e}^{j\Omega}, \phi_s) + V_{\text{DMA,left}}(\mathrm{e}^{j\Omega}) \tag{7.11}$$

where $V_{\text{DMA,left}}(\mathrm{e}^{j\Omega})$ is the noise component contained in the left DMA output. Similarly, the noisy output signal at the right side is defined as $Y_{\text{DMA,right}}(\mathrm{e}^{j\Omega}) = S(\mathrm{e}^{j\Omega}) H_{\text{right}}(\mathrm{e}^{j\Omega}, \phi_s) + V_{\text{DMA,right}}(\mathrm{e}^{j\Omega})$. Note that $H_{\text{left}}(\mathrm{e}^{j\Omega}, \phi_s)$ and $H_{\text{right}}(\mathrm{e}^{j\Omega}, \phi_s)$ have been used as the HRTF of the front microphone at respective side, where the subscripts 1 are omitted for conciseness.

7.2.2 Estimation of Noise

An essential element to design reliable postfilters, which provide good noise suppression capability and keep the distortions of the target signal at minimum level, is a good estimate of all undesired signal components. Two different methods can be found in literature for two-channel noise reduction [164, 165]. Both schemes require differential microphone arrays with a fixed null at $\phi = 180°$ as a first processing step. This is necessary to suppress undesired signal components originating from the back, and cannot be achieved by subsequent postprocessing.

Blocking Matrix

The first approach to estimating the undesired signal components aims at suppressing the target signal by steering a null toward the desired source direction. This is achieved by subtracting the left and right microphone signal from each other, creating a spatial null for the front direction $\phi = 0°$, which is typically considered as the angular position of the desired speaker. Thereby, the speech signal is cancelled and only the interfering signal components remain, which then serve

as a noise estimate. However, signals arriving from $\phi = 180°$ result in the same relative delay between the two channels and, thus, cannot be distinguished from the front direction. This is known as forward-backward ambiguity and implies that undesired signals originating from the back are suppressed in the same way as the target signal. As a consequence, these components are not captured by the noise estimation process and cannot be suppressed by the subsequent postfilter. To tackle this, DMAs are required as a first processing step, and instead of subtracting the microphone signals from each other, the spatial null for the front direction is created by subtracting the two DMA output signals.

In order to allow for a cancellation of the target signal for lateral source positions, the DMA output signals must be time-aligned and the level also has to be equalised. Based on Eq. (7.11), an estimate of all undesired signals can be written as

$$\begin{aligned}\hat{V}(k,m) &= \frac{\alpha_{\text{right}}(k,\phi_s)\mathrm{e}^{-j2\pi\frac{k}{K}\tau_{\text{right}}(\phi_s)}}{\alpha_{\text{left}}(k,\phi_s)\mathrm{e}^{-j2\pi\frac{k}{K}\tau_{\text{left}}(\phi_s)}}Y_{\text{DMA,left}}(k,m) - Y_{\text{DMA,right}}(k,m)\\ &= \Upsilon(k)Y_{\text{DMA,left}}(k,m) - Y_{\text{DMA,right}}(k,m)\\ &= \Upsilon(k)V_{\text{DMA,left}}(k,m) - V_{\text{DMA,right}}(k,m)\end{aligned} \tag{7.12}$$

where $\alpha(k,\phi_s)$ and $\mathrm{e}^{-j2\pi\frac{k}{K}\tau(\phi_s)}$ model the interaural level differences (ILDs) and interaural time differences (ITDs), respectively. The real-value PSD of this combined noise estimate is given by

$$\begin{aligned}\Phi_{\hat{v}\hat{v}}(k,m) =&|\Upsilon(k)|^2\Phi_{v_{\text{left}}v_{\text{left}}}(k,m) + \Phi_{v_{\text{right}}v_{\text{right}}}(k,m)\\ &- 2\Re\left\{\Upsilon(k)\Phi_{v_{\text{left}}v_{\text{right}}}(k,m)\right\}\end{aligned} \tag{7.13}$$

where $\Re\{\cdot\}$ indicates the real part of the complex number. From this joint noise estimate, an estimate of the individual noise components for each of the two channels can be obtained by firstly expressing $\Phi_{v_{\text{left}}v_{\text{right}}}(k,m)$ in Eq. (7.13) in terms of the coherence, as

$$\Gamma_{v_{\text{left}}v_{\text{right}}}(k,m) = \frac{\Phi_{v_{\text{left}}v_{\text{right}}}(k,m)}{\sqrt{\Phi_{v_{\text{left}}v_{\text{left}}}(k,m)\Phi_{v_{\text{right}}v_{\text{right}}}(k,m)}}. \tag{7.14}$$

Then, by assuming that the PSDs of the noise components are identical in both channels, i.e., $\Phi_{v_{\text{left}}v_{\text{left}}}(k,m) = \Phi_{v_{\text{right}}v_{\text{right}}}(k,m) = \Phi_{v_{\text{i}}v_{\text{i}}}(k,m)$, and by substituting

Eq. (7.14) into Eq. (7.13) yields

$$\Phi_{\hat{v}\hat{v}}(k,m) = \left(|\Upsilon(k)|^2 + 1\right) \Phi_{\hat{v}_i\hat{v}_i}(k,m) - 2\Re\left\{\Upsilon(k)\Gamma_{v_{\text{left}}v_{\text{right}}}(k,m)\hat{\Phi}_{v_iv_i}(k,m)\right\}. \tag{7.15}$$

Finally, the noise PSD for the individual channels can be obtained by reordering Eq. (7.15) as

$$\hat{\Phi}_{v_iv_i}(k,m) = \frac{\Phi_{\hat{v}\hat{v}}(k,m)}{1 + |\Upsilon(k)|^2 - 2\Re\left\{\Upsilon(k)\Gamma_{v_{\text{left}}v_{\text{right}}}(k,m)\right\}}. \tag{7.16}$$

The drawback of such Blocking Matrix is that the HRTFs are difficult to obtain in real scenarios. Here, an assumption is made that the ILDs are independent of frequency, i.e., $\alpha_i(k,m) = \alpha(m)$ [38]. The ITDs are compensated by simply applying a scalar gain factor and estimate the ITDs based on the responses of the left and right DMA from a certain source angle ϕ_s measured in an anechoic chamber.

Noise Estimation Based on Cross PSD

A different approach to estimating the joint PSD of noise can be done by implicitly assuming that $H_{\text{left}}(k,\phi_s) = H_{\text{right}}(k,\phi_s) = 1$, which is only reasonable for a source coming from $\phi_s = 0°$ in the free-field. For this special case, the cross-PSD between the left and right channel is simply

$$\Phi_{y_{\text{left}}y_{\text{right}}}(k,m) = \Phi_{xx}(k,m) + \Phi_{v_{\text{left}}v_{\text{right}}}(k,m). \tag{7.17}$$

Assuming that the noise PSDs are identical for both sides, the geometric mean of the PSDs of the left and right channel is given by

$$\sqrt{\Phi_{y_{\text{left}}y_{\text{left}}}(k,m)\Phi_{y_{\text{right}}y_{\text{right}}}(k,m)} = \Phi_{xx}(k,m) + \Phi_{v_iv_i}(k,m). \tag{7.18}$$

Substituting $\Phi_{xx}(k,m)$ and $\Phi_{v_{\text{left}}v_{\text{right}}}(k,m)$ in Eq. (7.17) by Eqs. (7.14) and (7.18), respectively, yields

$$\begin{aligned}\Phi_{y_{\text{left}}y_{\text{right}}}(k,m) =& \sqrt{\Phi_{y_{\text{left}}y_{\text{left}}}(k,m)\Phi_{y_{\text{right}}y_{\text{right}}}(k,m)} - \Phi_{v_iv_i}(k,m) \\ &+ \Gamma_{v_{\text{left}}v_{\text{right}}}(k,m)\Phi_{v_iv_i}(k,m).\end{aligned} \tag{7.19}$$

By reordering the equation gives the real-values estimate of the noise PSD as

$$\hat{\Phi}_{v_iv_i}(k,m) = \frac{\sqrt{\Phi_{v_{\text{left}}v_{\text{left}}}(k,m)\Phi_{v_{\text{right}}v_{\text{right}}}(k,m)} - \Re\left\{\Phi_{v_{\text{left}}v_{\text{right}}}(k,m)\right\}}{1 - \Re\left\{\Gamma_{v_{\text{left}}v_{\text{right}}}(k,m)\right\}}. \tag{7.20}$$

7.2.3 Single-channel Postfilter

Once the noise components estimates are obtained, single-channel postfilters can be designed for the left and right channel. Such postfilters have been illustrated in Chapters 4 and 5. The basic principle of the single-channel postfilters is to evaluate the SNR for each frequency bin and to map these values to a gain function. An important aspect of designing a single-channel gain function is to take into account the trade-off among noise reduction, speech distortion and musical noise. This has been explicitly discussed in Chapter 5. Here, the last processing stage as shown in Figure 7.2, which merges the intermediate gains $G_{\text{left}}(k,m)$ and $G_{\text{right}}(k,m)$ for the left and right side to a single gain function $\mathcal{G}(k,m)$, is being discussed. It has been shown in [166] that an optimal combination of the spectral gain functions can be derived based on the cost function

$$\begin{aligned}\mathcal{J}(k,m) = E\{&(H_{\text{left}}(k,m)S(k,m) - \mathcal{G}(k,m)|Y_{\text{DMA,left}}(k,m)|)^2 \\ &+ (H_{\text{right}}(k,m)S(k,m) - \mathcal{G}(k,m)|Y_{\text{DMA,right}}(k,m)|)^2\}\end{aligned} \tag{7.21}$$

which indicates the sum of the power of the error signals between the true speech components and the estimated speech components for the respective side. Minimising Eq. (7.21) yields the optimal binaural gain function

$$\mathcal{G}_{\text{opt}}(k,m) = \frac{G_{\text{left}}(k,m)|Y_{\text{DMA,left}}(k,m)|^2 + G_{\text{right}}(k,m)|Y_{\text{DMA,right}}(k,m)|^2}{|Y_{\text{DMA,left}}(k,m)|^2 + |Y_{\text{DMA,right}}(k,m)|^2}. \tag{7.22}$$

This indicates that the power of true speech components is positively proportional to the weight applied on the corresponding gain function on each side. Note that for the special case of $|Y_{\text{DMA,left}}(k,m)|^2 \equiv |Y_{\text{DMA,right}}(k,m)|^2$, Equation (7.23) becomes

$$\mathcal{G}(k,m) = \frac{1}{2}\left(G_{\text{left}}(k,m) + G_{\text{right}}(k,m)\right). \tag{7.23}$$

This implies it is the average of the gain functions from respective side of the device.

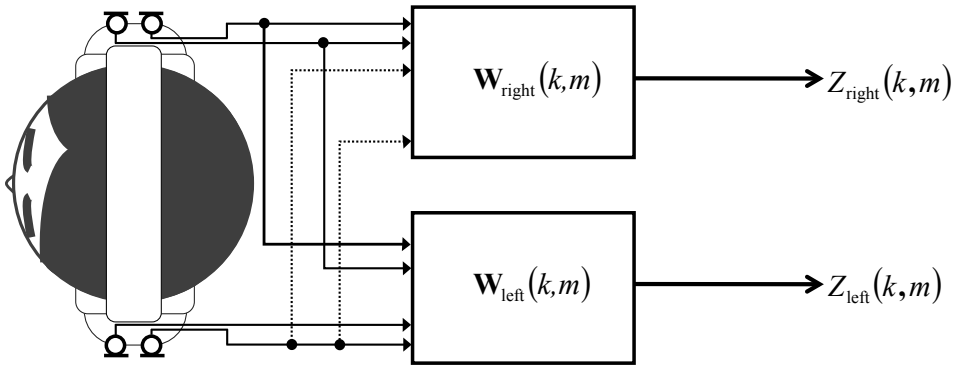

Figure 7.3: A binaural noise suppression approach with beamforming.

7.3 Binaural Multi-channel Wiener Filter

7.3.1 Configuration and Notation

The DMA is restricted to be used on each individual ear. Here we study a full form of beamforming which utilise all microphone elements. Now consider the l-th microphone signal on the left side of the hearing protection device defined as

$$Y_{\text{left},l}(k,m) = X_{\text{left},l}(k,m) + V_{\text{left},l}(k,m) \qquad l = 1, ..., L, \tag{7.24}$$

where $X_{\text{left},l}(k,m)$ and $V_{\text{left},l}(k,m)$ represent the speech and noise components in the microphone signal, respectively. Similarly, the observed signal at the right side is given by $Y_{\text{right},l}(k,m) = X_{\text{right},l}(k,m)+V_{\text{right},l}(k,m)$. The L-dimensional stacked microphone signal vectors $\mathbf{y}_{\text{left}}(k,m)$ and $\mathbf{y}_{\text{right}}(k,m)$, and the $2L$-dimensional signal vector $\mathbf{y}(k,m)$ are given as

$$\mathbf{y}(k,m) = \begin{bmatrix} \mathbf{y}_{\text{left}}(k,m) \\ \mathbf{y}_{\text{right}}(k,m) \end{bmatrix} \tag{7.25}$$

$$\begin{aligned} \mathbf{y}_{\text{left}}(k,m) &= [Y_{\text{left},1}(k,m) \quad Y_{\text{left},2}(k,m) \ \cdots \ Y_{\text{left},L}(k,m)]^T \\ \mathbf{y}_{\text{right}}(k,m) &= [Y_{\text{right},1}(k,m) \quad Y_{\text{right},2}(k,m) \ \cdots \ Y_{\text{right},L}(k,m)]^T . \end{aligned} \tag{7.26}$$

The correlation matrix of speech plus noise $\mathbf{R}_y(k,m)$, the clean speech correlation

matrix $\mathbf{R}_x(k,m)$, and the noise correlation matrix $\mathbf{R}_v(k,m)$ are defined as

$$\begin{aligned}\mathbf{R}_y(k,m) &= E\{\mathbf{y}(k,m)\mathbf{y}^H(k,m)\},\\ \mathbf{R}_x(k,m) &= E\{\mathbf{x}(k,m)\mathbf{x}^H(k,m)\},\\ \mathbf{R}_v(k,m) &= E\{\mathbf{v}(k,m)\mathbf{v}^H(k,m)\},\end{aligned} \tag{7.27}$$

where the $2L$-dimensional signal vectors $\mathbf{x}(k,m)$ and $\mathbf{v}(k,m)$ are similarly defined as $\mathbf{y}(k,m)$. Here, the speech and the noise components are assumed uncorrelated, such that $\mathbf{R}_y(k,m) = \mathbf{R}_x(k,m) + \mathbf{R}_v(k,m)$.

For speech enhancement algorithms, the r_{left}-th signal of the left device and the r_{right}-th signal of the right device will be used as the so-called reference signals. Typically, the front microphones are used as reference microphones. For conciseness, the reference microphone signals $Y_{\text{left},r_{\text{left}}}(k,m)$ and $Y_{\text{right},r_{\text{right}}}(k,m)$ at the left and the right hearing aid are denoted as $Y_{\text{left}}(k,m)$ and $Y_{\text{right}}(k,m)$, which are equivalent to

$$\begin{aligned}Y_{\text{left}}(k,m) &= \mathbf{e}_{\text{left}}^H\mathbf{y}(k,m) = X_{\text{left}}(k,m) + V_{\text{left}}(k,m)\\ Y_{\text{right}}(k,m) &= \mathbf{e}_{\text{right}}^H\mathbf{y}(k,m) = X_{\text{right}}(k,m) + V_{\text{right}}(k,m)\end{aligned} \tag{7.28}$$

where $\mathbf{e}_{\text{left}}$ and $\mathbf{e}_{\text{right}}$ are $2L$-dimensional vectors with only one element equal to 1 and the other elements equal to 0, i.e., $\mathbf{e}_{\text{left}}(r_{\text{left}}) = 1$ and $\mathbf{e}_{\text{right}}(L + r_{\text{right}}) = 1$.

The output signals at respective side are obtained by filtering and summing all microphone signals, i.e.,

$$Z_{\text{left}}(k,m) = \mathbf{w}_{\text{left}}^H(k,m)\mathbf{y}(k,m), \quad Z_{\text{right}}(k,m) = \mathbf{w}_{\text{right}}^H(k,m)\mathbf{y}(k,m) \tag{7.29}$$

where $\mathbf{w}_{\text{left}}(k,m)$ and $\mathbf{w}_{\text{right}}(k,m)$ are $2L$-dimensional complex weight vectors. The output signals can be written as

$$\begin{aligned}Z_{\text{left}}(k,m) &= \mathbf{w}_{\text{left}}^H(k,m)\mathbf{x}(k,m) + \mathbf{w}_{\text{left}}^H(k,m)\mathbf{v}(k,m)\\ &= Z_{x,\text{left}}(k,m) + Z_{v,\text{left}}(k,m),\\ Z_{\text{right}}(k,m) &= \mathbf{w}_{\text{right}}^H(k,m)\mathbf{x}(k,m) + \mathbf{w}_{\text{right}}^H(k,m)\mathbf{v}(k,m)\\ &= Z_{x,\text{right}}(k,m) + Z_{v,\text{right}}(k,m)\end{aligned} \tag{7.30}$$

where $Z_{x,\text{left}}$, $Z_{x,\text{right}}$ represent the speech component and $Z_{v,\text{left}}$, $Z_{v,\text{right}}$ represent the noise component of the output signals at respective side. Hereinafter, the

$4L$-dimensional complex stacked weight vector $\mathbf{w}(k,m)$ is denoted as

$$\mathbf{w}(k,m) = \begin{bmatrix} \mathbf{w}_{\text{left}}(k,m) \\ \mathbf{w}_{\text{right}}(k,m) \end{bmatrix}. \tag{7.31}$$

7.3.2 General Formulation

The MWF produces a minimum mean square error (MMSE) estimate of the speech component in the reference microphone at respective sides, simultaneously reducing noise and limiting speech distortion [20]. The mean square error (MSE) cost function for the filter $\mathbf{w}_{\text{left}}(k,m)$ and $\mathbf{w}_{\text{right}}(k,m)$ is equal to

$$\begin{aligned} \mathcal{J}_{\text{MWF}}\left(\mathbf{w}(k,m)\right) &= E\left\{ \left\| \begin{bmatrix} X_{\text{left}}(k,m) - Z_{\text{left}}(k,m) \\ X_{\text{right}}(k,m) - Z_{\text{right}}(k,m) \end{bmatrix} \right\|^2 \right\} \\ &= E\left\{ \left\| \begin{bmatrix} X_{\text{left}}(k,m) - \mathbf{w}_{\text{left}}^H(k,m)\mathbf{y}(k,m) \\ X_{\text{right}}(k,m) - \mathbf{w}_{\text{right}}^H(k,m)\mathbf{y}(k,m) \end{bmatrix} \right\|^2 \right\} \end{aligned} \tag{7.32}$$

where $X_{\text{left}}(k,m)$ and $X_{\text{right}}(k,m)$ are speech components at the reference microphones. To provide a more explicit trade-off between speech distortion and noise reduction, the SDW-MWF minimises a weighted sum of the residual noise energy and the speech distortion energy [14]. The binaural SDW-MWF[1] cost function is equal to

$$\begin{aligned} \mathcal{J}_{\text{MWF}_\mu}\left(\mathbf{w}(k,m)\right) =& E\{ \left\| \begin{bmatrix} X_{\text{left}}(k,m) - \mathbf{w}_{\text{left}}^H(k,m)\mathbf{x}(k,m) \\ X_{\text{right}}(k,m) - \mathbf{w}_{\text{right}}^H(k,m)\mathbf{x}(k,m) \end{bmatrix} \right\|^2 \\ & - \mu \left\| \begin{bmatrix} \mathbf{w}_{\text{left}}^H(k,m)\mathbf{v}(k,m) \\ \mathbf{w}_{\text{right}}^H(k,m)\mathbf{v}(k,m) \end{bmatrix} \right\|^2 \} \end{aligned} \tag{7.33}$$

where μ provides a trade-off between reduction and speech distortion. The optimal MWF_μ filters for the respective sides are equal to

$$\begin{aligned} \mathbf{w}_{\text{MWF}_\mu,\text{left}}(k,m) &= \left(\mathbf{R}_x(k,m) + \mu\mathbf{R}_v(k,m)\right)^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\text{left}}, \\ \mathbf{w}_{\text{MWF}_\mu,\text{right}}(k,m) &= \left(\mathbf{R}_x(k,m) + \mu\mathbf{R}_v(k,m)\right)^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\text{right}}. \end{aligned} \tag{7.34}$$

[1]For conciseness, SDW-MWF is abbreviated to MWF_μ in the equations.

7.3.3 Special Case with Single Target Source

In the case of a single target speech source, the speech signal vector can be modeled as

$$\mathbf{x}(k,m) = \mathbf{a}(k,m)S(k,m) \tag{7.35}$$

where the L-dimensional stacked vector $\mathbf{a}(k,m)$ is given by

$$\mathbf{a}(k,m) = \begin{bmatrix} \mathbf{a}_{\text{left}}(k,m) \\ \mathbf{a}_{\text{right}}(k,m) \end{bmatrix}, \tag{7.36}$$

with

$$\begin{aligned} \mathbf{a}_{\text{left}}(k,m) &= [A_{\text{left},1}(k,m) \quad A_{\text{left},2}(k,m) \ \cdots \ A_{\text{left},L}(k,m)]^T, \\ \mathbf{a}_{\text{right}}(k,m) &= [A_{\text{right},1}(k,m) \quad A_{\text{right},2}(k,m) \ \cdots \ A_{\text{right},L}(k,m)]^T. \end{aligned} \tag{7.37}$$

The speech correlation matrix is then a rank-one matrix, i.e.,

$$\mathbf{R}_x(k,m) = \Phi_{ss}(k,m)\mathbf{a}(k,m)\mathbf{a}^H(k,m) \tag{7.38}$$

with $\Phi_{ss}(k,m) = E\{|S(k,m)|^2\}$ representing the PSD of the speech signal. By assuming a single speech source and by applying the matrix inversion lemma, it has been shown in [167] that Eq. (7.34) can be reduced to the following optimal filter at each side:

$$\begin{aligned} \mathbf{w}_{\text{MWF}_\mu,\text{left}}(k,m) &= \mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m) \cdot \frac{\Phi_{ss}(k,m)A^*_{\text{left}}(k,m)}{\mu + \varrho(k,m)} \\ \mathbf{w}_{\text{MWF}_\mu,\text{right}}(k,m) &= \mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m) \cdot \frac{\Phi_{ss}(k,m)A^*_{\text{right}}(k,m)}{\mu + \varrho(k,m)} \end{aligned} \tag{7.39}$$

where $A^*_{\text{left}}(k,m) = \mathbf{a}^H(k,m)\mathbf{e}_{\text{left}}$, $A^*_{\text{right}}(k,m) = \mathbf{a}^H(k,m)\mathbf{e}_{\text{right}}$, and

$$\varrho(k,m) = \Phi_{ss}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m). \tag{7.40}$$

7.3.4 Rank-One Binaural MWF

It is shown in Eq. (7.39) that the solutions for special case with single target source requires *a priori* knowledge, or explicit estimation of the steering vector $\mathbf{a}(k,m)$ and the speech PSD $\Phi_{ss}(k,m)$. Also, due to the finite discrete Fourier transform (DFT) size in the short-time Fourier transform (STFT) analysis, the non-stationarity characteristic of the noise and the finite observation window

which leads to estimation errors, the rank of $\mathbf{R}_x(k,m)$ will be greater than one. Therefore, eigenvalue decomposition is often used for rank-1 approximation to extract the steering vector $\mathbf{a}(k,m)$. However, it is possible to derive an alternative expression which only uses the speech and noise second order statistics [150], similar to the general expression in Eq. (7.34). By rewriting $\varrho(k,m)$ as

$$\begin{aligned}\varrho(k,m) &= \Phi_{ss}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\\ &= \Phi_{ss}(k,m)\mathrm{Tr}\left\{\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\mathbf{a}^H(k,m)\right\}\\ &= \mathrm{Tr}\left\{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\right\}\end{aligned} \tag{7.41}$$

where $\mathrm{Tr}\{\cdot\}$ is the trace operator, Eq. (7.39) is equivalent to the following rank-one expression

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\mu\text{-rank1,left}} &= \frac{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{left}}}{\mu + \mathrm{Tr}\left\{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\right\}},\\ \mathbf{w}_{\mathrm{MWF}_\mu\text{-rank1,right}} &= \frac{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{right}}}{\mu + \mathrm{Tr}\left\{\mathbf{R}_v^{-1}(k,m)\mathbf{R}_x(k,m)\right\}}.\end{aligned} \tag{7.42}$$

Although Eq. (7.42) is derived for the special case of a single target speech source, it can be used when this assumption is not fulfilled. Otherwise, it is completely equivalent to Eq. (7.39) for a single target speech source case.

7.3.5 Cue Preservation and SNR Improvement

As an examination to the ability of binaural MWF_μ in maintaining binaural cues, the single target source optimal filter from Eq. (7.39) is utilised. From there, the binaural MWF_μ vectors for the left and right side of the devices are found to be parallel with respect to the ATF, such that

$$\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{left}}(k,m) = \mathrm{ITF}_x^{\mathrm{in},*}(k,m)\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{right}}(k,m) \tag{7.43}$$

where $\mathrm{ITF}_x^{\mathrm{in}}(k,m)$ is defined as

$$\mathrm{ITF}_x^{\mathrm{in}}(k,m) = \frac{X_{\mathrm{left}}(k,m)}{X_{\mathrm{right}}(k,m)} = \frac{A_{\mathrm{left}}(k,m)}{A_{\mathrm{right}}(k,m)}. \tag{7.44}$$

Hence, the ITFs of the output speech and noise components are both equal to $\mathrm{ITF}_x^{\mathrm{in}}$, implying that the binaural speech cues are perfectly preserved, but the binaural noise cues are distorted. As all output components are perceived as

coming from the speech direction, the auditory perception of the acoustic scene is therefore not preserved by the binaural MWF_μ.

Since the solutions of binaural MWF_μ are parallel, the narrowband output SNR for each side will be the same, such that

$$\begin{aligned}\text{SNR}^{\text{out}}_{\text{left}}(k,m) &= \text{SNR}^{\text{out}}_{\text{right}}(k,m)\\ &= \frac{E|Z_{x,\text{left}}(k,m)|^2}{E|Z_{v,\text{left}}(k,m)|^2}\\ &= \frac{|\mathbf{w}^H_{\text{MWF}_\mu,\text{left}}(k,m)\mathbf{a}(k,m)|^2\Phi_{ss}}{\mathbf{w}^H_{\text{MWF}_\mu,\text{left}}(k,m)\mathbf{R}_v\mathbf{w}_{\text{MWF}_\mu,\text{left}}(k,m)}\\ &= \Phi_{ss}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m) = \varrho(k,m).\end{aligned} \tag{7.45}$$

Given the input SNRs

$$\begin{aligned}\text{SNR}^{\text{in}}_{\text{left}}(k,m) &= \frac{E|X_{\text{left}}(k,m)|^2}{E|V_{\text{left}}(k,m)|^2} = \frac{|\mathbf{e}^H_{\text{left}}\mathbf{a}(k,m)|^2\Phi_{ss}}{\mathbf{e}^H_{\text{left}}\mathbf{R}_v(k,m)\mathbf{e}_{\text{left}}}\\ \text{SNR}^{\text{in}}_{\text{right}}(k,m) &= \frac{E|X_{\text{right}}(k,m)|^2}{E|V_{\text{right}}(k,m)|^2} = \frac{|\mathbf{e}^H_{\text{right}}\mathbf{a}(k,m)|^2\Phi_{ss}}{\mathbf{e}^H_{\text{right}}\mathbf{R}_v(k,m)\mathbf{e}_{\text{right}}},\end{aligned} \tag{7.46}$$

it follows that the SNR improvement at each respective side can be obtained as

$$\begin{aligned}\Delta\text{SNR}_{\text{left}}(k,m) &= \frac{\varrho(k,m)\mathbf{e}^H_{\text{left}}\mathbf{R}_v(k,m)\mathbf{e}_{\text{left}}}{\Phi_{ss}|A_{\text{left}}|^2}\\ \Delta\text{SNR}_{\text{right}}(k,m) &= \frac{\varrho(k,m)\mathbf{e}^H_{\text{right}}\mathbf{R}_v(k,m)\mathbf{e}_{\text{right}}}{\Phi_{ss}|A_{\text{right}}|^2}.\end{aligned} \tag{7.47}$$

This implies that the SNR improvements are directly related to the noise correlation matrix $\mathbf{R}_v(k,m)$ and speech correlation matrix $\mathbf{R}_x(k,m)$. As a matter of fact, they are related to their estimates and the accuracy of the model. Hence it is important that the estimates render an accurate reflection of the true noise correlation and speech power and the ATF of the target speech signal.

7.4 Proposed Methods

7.4.1 Improvement to DMA-BPF

As mentioned in previous sections, the binaural noise reduction techniques incorporating the DMA with a binaural gain function (DMA-BPF) have two major limitations. Firstly, recall that the noise estimation algorithm for the DMAs that

uses a target cancellation scheme for the noise PSD estimate, if the speech target estimation is not precise, speech will still be present in the noise estimate. This will lead to an overestimation of noise and the SNR estimate used in DMA-BPF will lower the volume of the speech. Secondly, the single-channel speech enhancement algorithms required for computing the binaural gain function has a trade-off among noise reduction, speech distortion and musical noise. This thesis gives possible solutions to the problems. We will mainly investigate methods to improve the noise reduction and distortion by using the single channel techniques that has been developed in this research. For noise estimation, we will adapt the SVAD method proposed in Chapter 5, while for the gain function we will employ the techniques from Chapter 4, which is capable of providing a good trade-off for a better speech quality performance. The SVAD noise estimation algorithm will be utilised in the DMA-BPF framework such that the gain functions at each side are computed with only the noise PSD estimate from the respective side. The added feature in this work is that since we have two output signals one from each DMA, investigation will be carried out to see if the binaural cues, i.e., the ITD and the ILD will be distorted.

7.4.2 Proposed Binaural MWF

In order to avoid the explicit assumptions or estimation of the location of the target speech source, a MWF can be used. Recently, an alternative SDW-MWF formulation that does not require the clean speech correlation matrix estimate, which has also been included in this thesis (see Chapter 6), has been proposed in [10]. In that work, the conditional speech presence probability (SPP) has been used in (i) estimating both speech plus noise correlation matrix $\mathbf{R}_y(k,m)$ and noise correlation matrix $\mathbf{R}_v(k,m)$, (ii) estimating the noise PSD in the reference channel for computing the cross-correlation vector $\mathbf{r}_{yx}(k,m)$, and also (iii) adapting the trade-off parameter $\lambda(k,m)$. Here, that formulation will be extended to binaural configuration, where a pair of contralateral microphones is added at the other side of the head. Evaluation will also be performed to test if such formulation can maintain binaural cues for both speech and noise.

Let the two-state model for speech events be defined in this context as

$$\begin{aligned}\mathcal{H}_0(k,m) &: Y_l(k,m) = V_l(k,m) \\ \mathcal{H}_1(k,m) &: Y_l(k,m) = X_l(k,m) + V_l(k,m).\end{aligned} \tag{7.48}$$

Assuming a complex Gaussian distribution of the STFT coefficients for both the speech and the noise, and by applying Bayes rule, the conditional SPP for each channel, $p_l(k,m)$ is given as in Eq. (5.14) for each frequency bin and each frame. The conditional SPP and the two-state model in Eq. (7.48) for speech events can be incorporated directly into the optimisation criterion of the SDW-MWF, leading to a weighted average where the first term is weighted by the probability that speech is present, while the second term is weighted by the probability that speech is absent. This can be defined as

$$\begin{aligned}\mathcal{J}_{\mathrm{MWF}_\lambda-\mathrm{SPP}}\left(\mathbf{w}(k,m)\right) =& p(k,m)\times \\ & E\{\left\| \begin{bmatrix} X_{\mathrm{left}}(k,m) - \mathbf{w}_{\mathrm{left}}^H(k,m)\mathbf{y}(k,m) \\ X_{\mathrm{right}}(k,m) - \mathbf{w}_{\mathrm{right}}^H(k,m)\mathbf{y}(k,m) \end{bmatrix} \right\|^2 |\mathcal{H}_1(k,m)\} \\ & + (1-p(k,m))\times \\ & E\{\left\| \begin{bmatrix} \mathbf{w}_{\mathrm{left}}^H(k,m)\mathbf{y}(k,m) \\ \mathbf{w}_{\mathrm{right}}^H(k,m)\mathbf{y}(k,m) \end{bmatrix} \right\|^2 |\mathcal{H}_0(k,m)\}\end{aligned} \tag{7.49}$$

where

$$p(k,m) = \frac{p_{\mathrm{left},r_{\mathrm{left}}}(k,m) + p_{\mathrm{right},r_{\mathrm{right}}}(k,m)}{2} \tag{7.50}$$

with $p_{\mathrm{left},r_{\mathrm{left}}}(k,m)$ and $p_{\mathrm{right},r_{\mathrm{right}}}(k,m)$ denote the conditional probability that speech is present obtained from the reference channels respectively at the left and the right, while $1-p(k,m)$ is the conditional probability that speech is absent. The solution is then given by

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,left}}(k,m) =& \left(p(k,m)\mathbf{R}_y(k,m) + (1-p(k,m))\mathbf{R}_v(k,m)\right)^{-1} \\ & p(k,m)\mathbf{r}_{yx,\mathrm{left}}(k,m) \\ \mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,right}}(k,m) =& \left(p(k,m)\mathbf{R}_y(k,m) + (1-p(k,m))\mathbf{R}_v(k,m)\right)^{-1} \\ & p(k,m)\mathbf{r}_{yx,\mathrm{right}}(k,m).\end{aligned} \tag{7.51}$$

where $\mathbf{R}_v(k,m)$ and $\mathbf{R}_y(k,m)$ are updated respectively by Eq. (6.12) and Eq. (6.13), both with the conditional SPP $p(k,m)$.

Compared to a fixed weighting factor λ, the conditional SPP $p(k,m)$ varies for each frequency bin k and for each frame m leading to faster dynamic changes in the MWF$_\lambda$-SPP approach. Although this results in less speech distortion, such added variations also cause more musical noise and more residual noise remaining in the enhanced signals [10]. The solution of MWF$_\lambda$-SPP is very similar to the formulation in [15], but we employ the cross-correlation vector $\mathbf{r}_{yx}(k,m)$ instead of using the speech correlation matrix, i.e., $\mathbf{R}_x(k,m)\mathbf{e}$.

The cross-correlation vector at both sides $\mathbf{r}_{yx,\text{left}}(k,m)$ and $\mathbf{r}_{yx,\text{right}}(k,m)$ are updated by employing single -channel speech enhancement algorithm, as given by

$$\begin{aligned}\hat{\mathbf{r}}_{yx,\text{left}}(k,m) =&(1-\alpha_x)\hat{\mathbf{r}}_{yx}(k,m-1)+\alpha_x\mathbf{y}(k,m)\mathcal{G}(k,m)Y^*_{\text{left}}(k,m)\\ \hat{\mathbf{r}}_{yx,\text{right}}(k,m) =&(1-\alpha_x)\hat{\mathbf{r}}_{yx}(k,m-1)+\alpha_x\mathbf{y}(k,m)\mathcal{G}(k,m)Y^*_{\text{right}}(k,m)\end{aligned} \tag{7.52}$$

where α_x is the smoothing factor. Here, $\mathcal{G}(k,m)$ is defined as

$$\mathcal{G}(k,m)=\frac{G_{\text{left}}(k,m)+G_{\text{right}}(k,m)}{2} \tag{7.53}$$

where $G_{\text{left}}(k,m)$ and $G_{\text{right}}(k,m)$ are single-channel weighting gain functions obtained for the corresponding reference channels $Y_{\text{left}}(k,m)$ and $Y_{\text{right}}(k,m)$. Their arithmetic mean $\mathcal{G}(k,m)$ is employed to ensure that the binaural cues can be preserved for the single-channel approach. For a fair comparison, the conditional SPP is also applied to the trade-off parameter of the conventional SDW-MWF function, such that $\mu(k,m)=1/p(k,m)$.

7.5 Performance Measures

Performance evaluation in this chapter includes the comparison of the binaural cues and the noise reduction performance. For the noise reduction performance, the measures from Chapter 6 were adapted to compare the noise reduction, speech distortion and the overall perceptual performance. Those include the speech intelligibility weighted segmental SNR in frequency domain (IFWSNRseg) measure, the segmental noise attenuation (NATTseg) measure, the segmental speech preservation (SPREseg) measure and the perceptual evaluation of speech quality (PESQ) measure. For all measurements, results of the individual channel (the left and right channels) were averaged to obtain a single value.

The binaural cues were evaluated using the ILD and the ITD measures. The ILD here are obtained by evaluating the logarithm of the power ratio between the respective signals of the left and right side. The ILDs of input clean speech and noise, and processed clean speech and noise are given as

$$\begin{aligned} \mathrm{ILD}_x^{\mathrm{in}} &= 10\log_{10}\left(\frac{\sum_{n=1}^{N} x_{\mathrm{left}}^2(n)}{\sum_{n=1}^{N} x_{\mathrm{right}}^2(n)}\right), \quad \mathrm{ILD}_v^{\mathrm{in}} = 10\log_{10}\left(\frac{\sum_{n=1}^{N} v_{\mathrm{left}}^2(n)}{\sum_{n=1}^{N} v_{\mathrm{right}}^2(n)}\right), \\ \mathrm{ILD}_x^{\mathrm{out}} &= 10\log_{10}\left(\frac{\sum_{n=1}^{N} \tilde{x}_{\mathrm{left}}^2(n)}{\sum_{n=1}^{N} \tilde{x}_{\mathrm{right}}^2(n)}\right), \quad \mathrm{ILD}_v^{\mathrm{out}} = 10\log_{10}\left(\frac{\sum_{n=1}^{N} \tilde{v}_{\mathrm{left}}^2(n)}{\sum_{n=1}^{N} \tilde{v}_{\mathrm{right}}^2(n)}\right) \end{aligned} \tag{7.54}$$

where N is the signal length in samples. The clean speech signals and the unprocessed reference signals at respective side were taken from the front microphones. This is the same for all the performance measures that require the access of the clean signals and/or the observed signals. The ITDs are computed using the cross-correlation, which is commonly used to estimate time delays, as defined by

$$\begin{aligned} R_{x_{\mathrm{left}}x_{\mathrm{right}}}(\eta) &= E\left(x_{\mathrm{left}}(n)x_{\mathrm{right}}(n-\eta)\right), \quad R_{v_{\mathrm{left}}v_{\mathrm{right}}}(\eta) = E\left(v_{\mathrm{left}}(n)v_{\mathrm{right}}(n-\eta)\right), \\ R_{\tilde{x}_{\mathrm{left}}\tilde{x}_{\mathrm{right}}}(\eta) &= E\left(\tilde{x}_{\mathrm{left}}(n)\tilde{x}_{\mathrm{right}}(n-\eta)\right), \quad R_{\tilde{v}_{\mathrm{left}}\tilde{v}_{\mathrm{right}}}(\eta) = E\left(\tilde{v}_{\mathrm{left}}(n)\tilde{v}_{\mathrm{right}}(n-\eta)\right). \end{aligned} \tag{7.55}$$

The delay is then given by the argument $\eta = \eta_0$, which yields the maximum absolute value of Eq. (7.55), as

$$\begin{aligned} \mathrm{ITD}_x^{\mathrm{in}} &= \arg\max_{\eta}\left[R_{x_{\mathrm{left}}x_{\mathrm{right}}}(\eta)\right], \quad \mathrm{ITD}_v^{\mathrm{in}} = \arg\max_{\eta}\left[R_{v_{\mathrm{left}}v_{\mathrm{right}}}(\eta)\right], \\ \mathrm{ITD}_x^{\mathrm{out}} &= \arg\max_{\eta}\left[R_{\tilde{x}_{\mathrm{left}}\tilde{x}_{\mathrm{right}}}(\eta)\right], \quad \mathrm{ITD}_v^{\mathrm{out}} = \arg\max_{\eta}\left[R_{\tilde{v}_{\mathrm{left}}\tilde{v}_{\mathrm{right}}}(\eta)\right]. \end{aligned} \tag{7.56}$$

However, since η_0 has integer values only and the delay is usually fractional, the cross-correlaiton function needs to be interpolated. After that, the delay τ_0 in seconds is obtained by dividing η_0 by the sampling frequency f_{s}.

7.6 Performance Evaluation

7.6.1 Test Setup

In this section, the overall performance for both binaural speech enhancement algorithms is evaluated. The setup for the underlying measurements is depicted

in Figure 7.4. A manikin with put-on earmuffs (with two-microphone array being mounted on each side) was placed close to the center of a room with dimensions 3.05 m × 3.05 m, with a reverberation time T_{60} of approximately 0.2 s. The loudspeakers were positioned at 1 m from the center of the head, with the speech and noise rendered at different position around the head to create point source sounds. Additional four loudspeakers were placed in each corner of the room facing the walls to create diffuse-like background noise.

Unless stated otherwise, we assume the *a priori* knowledge about the angular position of the desired source for all experiments. The speech signals consists of 5 (2 male and 3 female) sentences with length ranging from 11 s to 22 s, and the noise sources are industrial noises. For evaluation purpose, the speech and noise signals were recorded separately. The processing was done with a sampling frequency of 16 kHz using an STFT with the square root of a Hann window, both for analysis and synthesis, frame length $K = 512$, and 50% overlap, i.e., $R = 256$. The parameters of both frameworks are given in Table 7.1. For a fair comparison, MWF$_\mu$-SPP was chosen as a reference method for the MWF framework, whereas for the DMA-BPF framework, both two-channel noise PSD estimation algorithm, namely the Blocking Matrix (BM) approach and the CrossPSD (XPSD) approach were employed as references. The parameters of the algorithms were not adjusted to obtain the largest amount of noise suppression, but rather to achieve a good trade-off among the amount of noise suppression, speech distortion, and musical noise generated from the processing.

For evaluation of the ITD and ILD, two scenarios were considered. One is that speech source was from the front of the head with several noise configurations delivered at 45°, 90°, 135°, 180°, 225°, 270° and 315° with respect to the left of the head. The latter scenario had the noise source originated from behind the head with several speech configurations at 10°, 20°, 30°, 330°, 340°, and 350°. The reason of the choice of these speech directions is that the DMA processing cannot preserve the spatial impression from directions outside −40° to 40° [38]. This has been evaluated and again proven in the following results section.

For the noise reduction performance, only one scenario was concerned, with the speech source positioned at the head anterior and several point sources noise

<table>
<tr><td colspan="2">MWF</td><td colspan="3">DMA-BPF</td></tr>
<tr><td>MWF$_\mu$-SPP</td><td>MWF$_\lambda$-SPP</td><td>SVAD</td><td>BM</td><td>XPSD</td></tr>
<tr><td colspan="5">Smoothing constant for $\hat{\lambda}_v(k,m)$: $\alpha_v = 0.8$</td></tr>
<tr><td colspan="2">Smoothing constant for $\hat{\lambda}_y(k,m)$: $\alpha_y = 0$</td><td colspan="3">$\alpha_y = 0.17$</td></tr>
<tr><td colspan="2">Smoothing constant for $\xi_{\text{MDD}}(k,m)$: $\beta = 0.9$</td><td colspan="3">$\beta = 0.98$</td></tr>
<tr><td colspan="5">MSIG parameter: $a_1 = 3$, $a_2 = 1$, $c = 0.7$</td></tr>
<tr><td colspan="5">Spectral noise floor: $\epsilon = -15$ dB</td></tr>
<tr><td colspan="3">SVAD parameter: $P(\mathcal{H}_0(k,m)) = 0.3$</td><td colspan="2" rowspan="6"></td></tr>
<tr><td colspan="3">SVAD parameter: $P(\mathcal{H}_1(k,m)) = 0.7$</td></tr>
<tr><td colspan="3">SVAD parameter: $\xi_a = \xi_b = 12$ dB</td></tr>
<tr><td colspan="3">SVAD parameter: $t_1 = 0.05$</td></tr>
<tr><td colspan="3">SVAD parameter: $t_2 = 0.08$</td></tr>
<tr><td colspan="3">SVAD parameter: $t_3 = 240$</td></tr>
<tr><td colspan="3">Smoothing constant for $\hat{\mathbf{R}}_{\mathbf{y}}(k,m)$: $\alpha_{yy} = 0.17$</td><td colspan="2" rowspan="3"></td></tr>
<tr><td colspan="3">Smoothing constant for $\hat{\mathbf{r}}_{\mathbf{yx}}(k,m)$: $\alpha_x = 0.17$</td></tr>
<tr><td colspan="3">Smoothing constant for $\hat{\mathbf{R}}_{\mathbf{v}}(k,m)$: $\alpha_{vv} = 0.98$</td></tr>
</table>

Table 7.1: Parameter settings for DMA-BPF and MWF.

configurations rendered at 45°, 90°, 135°, 180°, 225°, 270° and 315° with respect to the left of the head. The results for different input SNRs will be plotted to show the robustness of the proposed methods from extremely noisy to more quiet environments. Note that for all the performance evaluations, only the average scores obtained from the evaluated five sentences will be shown rather than the measurement results from every speech sequence.

7.6.2 Results and Analysis for DMA-BPF

The improved DMA-BPF contains a different gain function algorithm and a different noise estimation method. In this section, only the noise estimation algorithms are evaluated since the gain function has already been evaluated in Chapters 3 and 4. Figures 7.5 and 7.6 depict the ILD and ILD results for DMA-BPF when different noise estimates were employed. Here, the results for both DMA and the observation from the reference channels (Ref.) were plotted for reference. More precisely, DMA takes the observed signals as reference, while the processed signals take DMA as reference for the evaluation of the binaural cues. It can be observed from Figure 7.5 that when noise came from behind the head, and the speech source was located at $-30°$ to 30°, the ITDs for both speech and noise were well maintained. For ILD results, the speech ILDs for DMA outputs

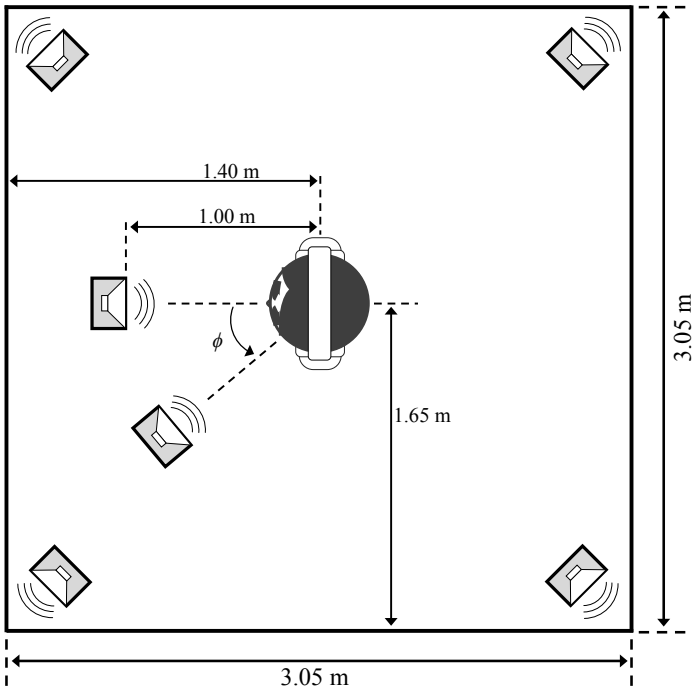

Figure 7.4: Measurement setup for the evaluation of the binaural noise reduction techniques.

and all the processed signals seem to be a scaled down version of the reference signals. The Blocking Matrix approach distorts the speech ILDs from the DMA outputs. It has also slightly altered the noise ILDs while the DMA outputs have the same noise ILDs recorded over different speech configurations. Meanwhile, the CrossPSD and SVAD manage to maintain both speech and noise ILDs with respect to the DMA's ILDs.

As for the results with speech fixed at 0° and different noise configurations in Figure 7.6, all algorithms have preserved the speech ILDs and ITDs. However, they fail to maintain the spatial impression of noise mainly because the DMAs cannot preserve the cues coming from $-45°$ to 320°, as observed in the figure. In spite of that, it can be seen that the noise ILDs and ITDs of DMA outputs still follow the patterns of the reference signals, which can be merely treated as a scaled version of the cues. This means that the users might still be able to localise the noise unless it is coming from the head posterior. For the processed signals, only the SVAD approach follows both the noise ILDs and ITDs of the DMA outputs. The Blocking Matrix and CrossPSD methods maintain the noise

ILDs with respect to the DMA outputs, but with a small degradation of the noise ITDs, which can be seen at configurations S0N225 and S0N270.

Figures 7.7 to 7.10 describe the noise reduction performance of DMA-BPF methods. It is obvious that DMA-BPF with SVAD noise estimator has the best performance among all evaluated algorithms in terms of IFWSNRseg, NATTseg and PESQ measures, at all input SNR levels. This implies that the two channels noise PSD estimation algorithm is capable of providing better SNR gains and better speech quality when compared to both the Blocking Matrix and CrossPSD algorithms, but with a slightly larger speech distortion as depicted in the lower SPREseg results. An advantage of this method is that the *a priori* information about speech location is not required in the noise estimate to obtain better overall performance in the enhanced speech quality. In addition, the two-channel SVAD method can also preserve the binaural cues, which makes it suitable for use in the DMA-BPF framework.

7.6.3 Results and Analysis for MWF

Figures 7.11 and 7.12 portray the ILD and ITD for two binaural MWF formulations with SPP, namely MWF_{μ}-SPP and MWF_{λ}-SPP. The results of binaural MSIG are also included in the figure for comparison. As predicted, all the evaluated methods have successfully preserved speech cues at all configurations. As for noise cues, an interesting finding is that ITDs of noise can be preserved with all algorithms, including binaural MWF_{μ}-SPP when noise source was fixed at 180° with speech source rendered from $-30°$ to 30°. However, when speech was fixed in front of the head with noise coming from the side (45° to 135°), the ITDs of noise were not preserved by binaural MWF_{μ}-SPP, with the ILDs also been distorted by approximately $2-4$ decibels (dBs). Under the same configurations, the binaural MWF_{λ}-SPP approach can however preserve both ILDs and ITDs of both speech and noise, which makes it a more preferable formulation compared to MWF_{μ}-SPP for a binaural assistive listening device.

As for the noise reduction performance, Figures 7.13 to 7.16 show that the results are consistent with the results obtained from monaural MWF formulations. The binaural MWF_{λ}-SPP approach has the best performance recorded among

all evaluated algorithms under all different input SNRs, which means having the largest IFWSNRseg, NATTseg and PESQ scores. However, it has consistently lower SPREseg compared to the binaural MSIG function. For 15 dB input SNR, the difference of SPREseg results between the binaural MSIG function and the binaural MWF$_\lambda$-SPP approach is the largest, whereby the latter produced slightly lower IFWSNRseg results. The binaural MWF$_\mu$-SPP approach was totally outperformed by the proposed method in terms of both the SNR gains and the overall perceptual speech quality. It is worth mentioning that when target speech is directed from the front, all algorithms show better performance when noise is coming from the left or the right side of the head, and contrary a poorer performance when noise is coming from behind due to the front-back ambiguity.

7.6.4 Comparison between DMA-BPF and MWF

In this part, the comparison between the MWF$_\lambda$-SPP method and the DMA-BPF with SVAD method is carried out. We started with the configuration with speech stays at 0° and noise comes from 45°, 90°, 135°, 180°, 225°, 270° and 315°. As depicted in Figures 7.17 to 7.20, the DMA-BPF method outperforms the MWF$_\lambda$-SPP in terms of the amount of noise reduction, particularly when noise source is further from the target speech. This is the advantage of employing DMAs at each side of the head to subtract more noise from behind the head. However, it can be seen from SPREseg results that the MWF$_\lambda$-SPP method can preserve more speech components, in other words less speech distortion compared to the DMA-BPF approach. This is shown in Figure 7.20 that the the MWF$_\lambda$-SPP method has comparable performance relative to the DMA-BPF approach at high input SNR.

The performance of both frameworks were again examined in realistic scenarios with diffuse background noise. Evaluation were done using two types of noise, namely the diffuse-like factory noise and diffuse-like jack-hammer noise. The results for the diffuse factory noise are depicted in Figures 7.21 to 7.24, while the results for the diffuse jack-hammer are shown in Figures 7.25 to 7.28. The results for both noise are consistent, which can be categorised into four main

points. First point is that the DMA-BPF approach always has higher noise reduction (NATTseg) in comparison to the MWF$_\lambda$-SPP method, while the latter can always preserve more speech components (SPREseg) when compared to the former. Secondly, in extremely noise scenarios, i.e., with low input SNRs, the DMA-BPF approach always performs better with higher SNR gains IFWSNRseg and better speech quality (PESQ). The third point is that at less noisy environments, the MWF$_\lambda$-SPP method always has better speech quality performance when compared to the DMA-BPF approach. Another interesting point, which influence the choice of the better framework, is that MWF$_\lambda$-SPP performs more consistently by having more consistent IFWSNRseg results over different configurations of target speech location when compared to the DMA-BPF approach. This is an important consideration given that the speaker will not always be located in a fixed position but move. However, as DMA-BPF is a fixed filtering structure for the left and right side of the head, it does not involve estimation of the second order statistics of speech and noise. This means that it has lower computational complexity when compared to the MWF$_\lambda$-SPP algorithm.

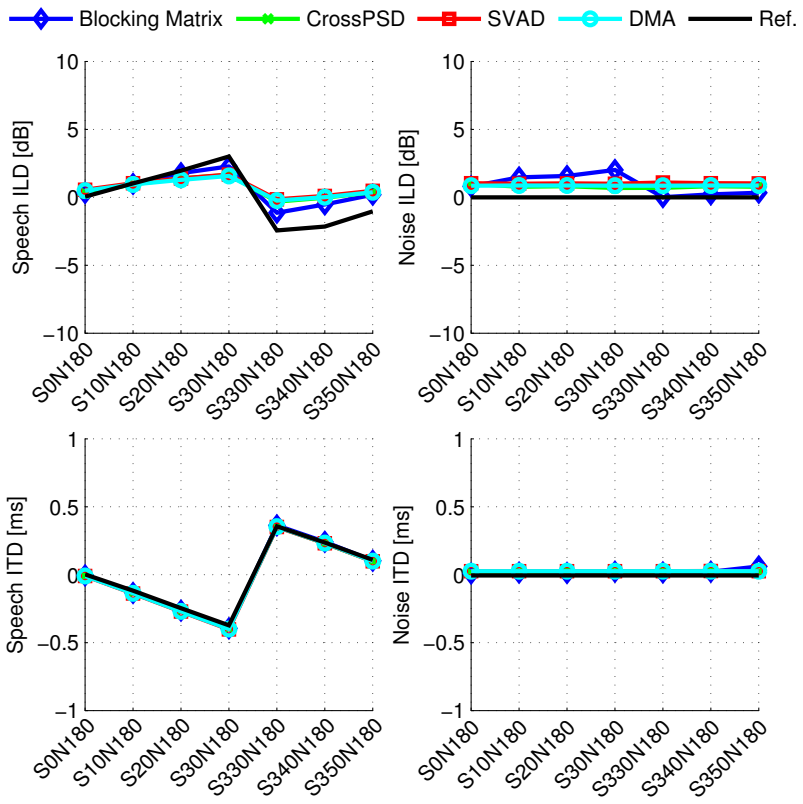

Figure 7.5: ITD and ILD results for DMA-BPF when direction of noise was fixed with speech source coming from different directions.

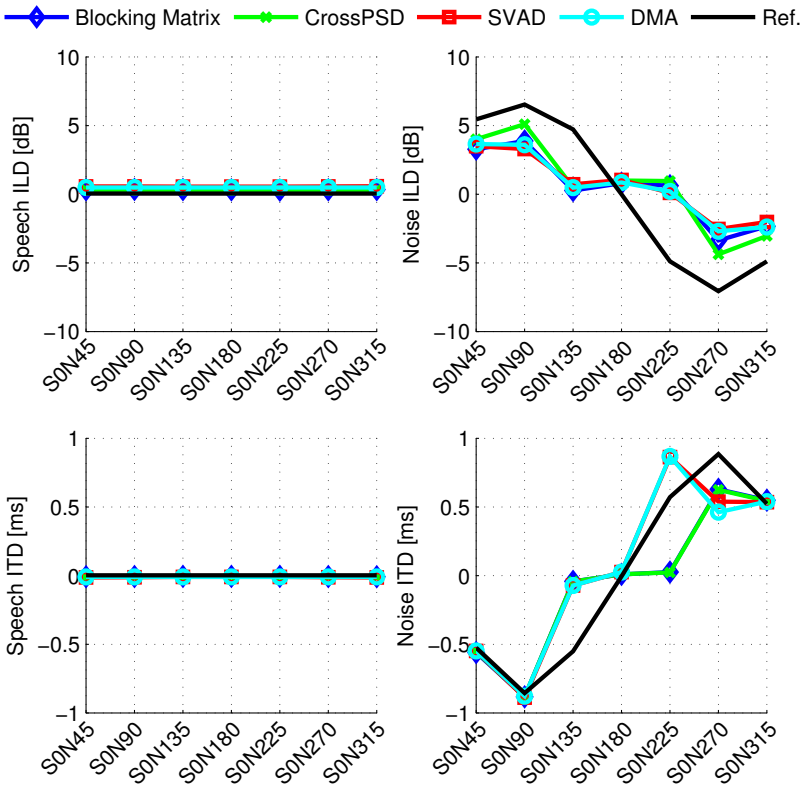

Figure 7.6: ITD and ILD results for DMA-BPF when direction of the target speech was fixed with noise source coming from different directions.

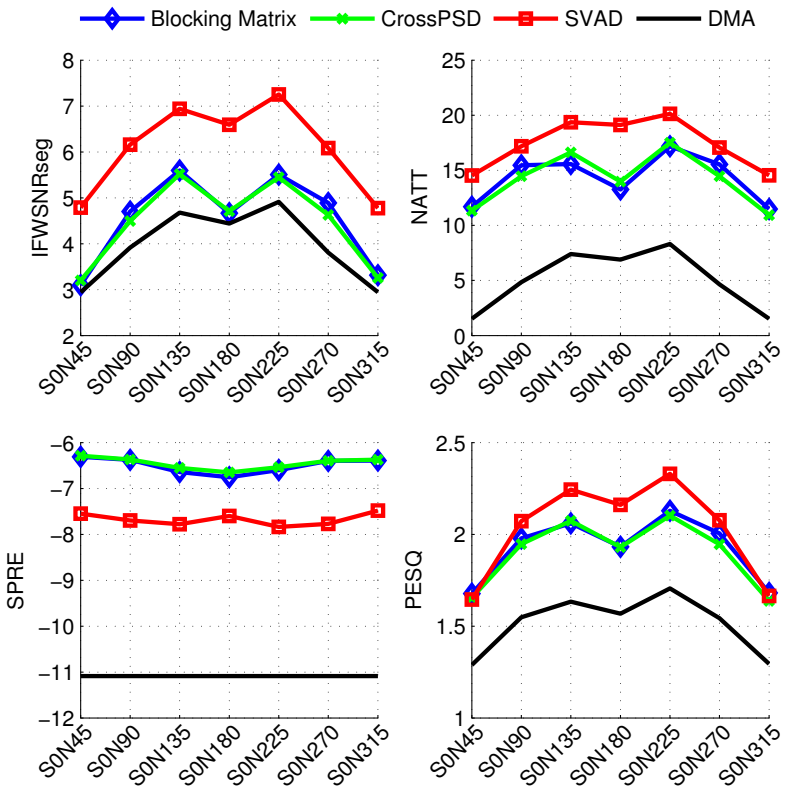

Figure 7.7: Noise reduction performance for difference noise PSD estimation algorithms employed in DMA-BPF at 0 dB SNR.

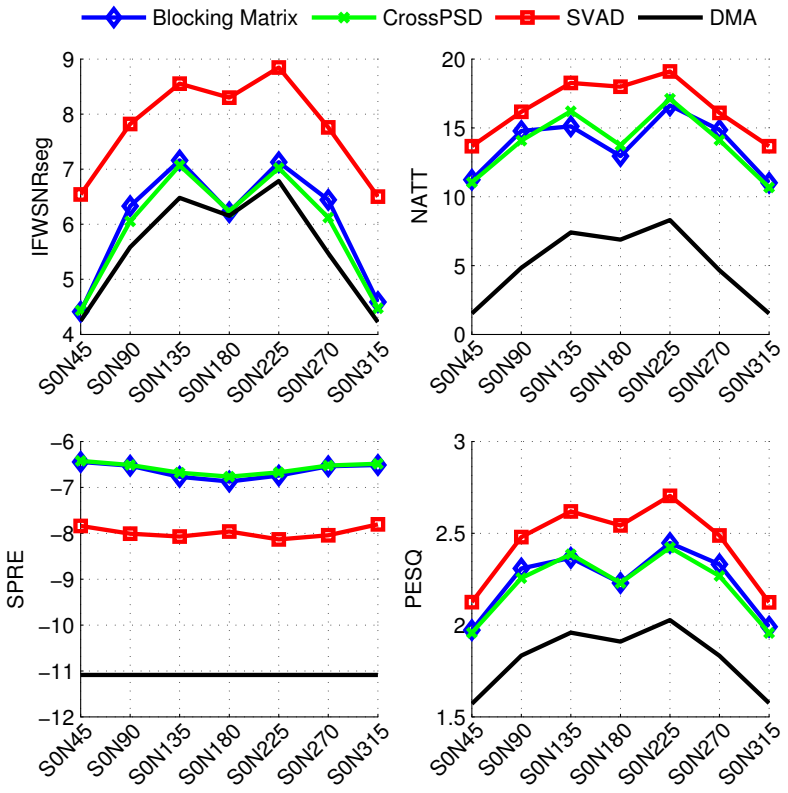

Figure 7.8: Noise reduction performance for difference noise PSD estimation algorithms employed in DMA-BPF at 5 dB SNR.

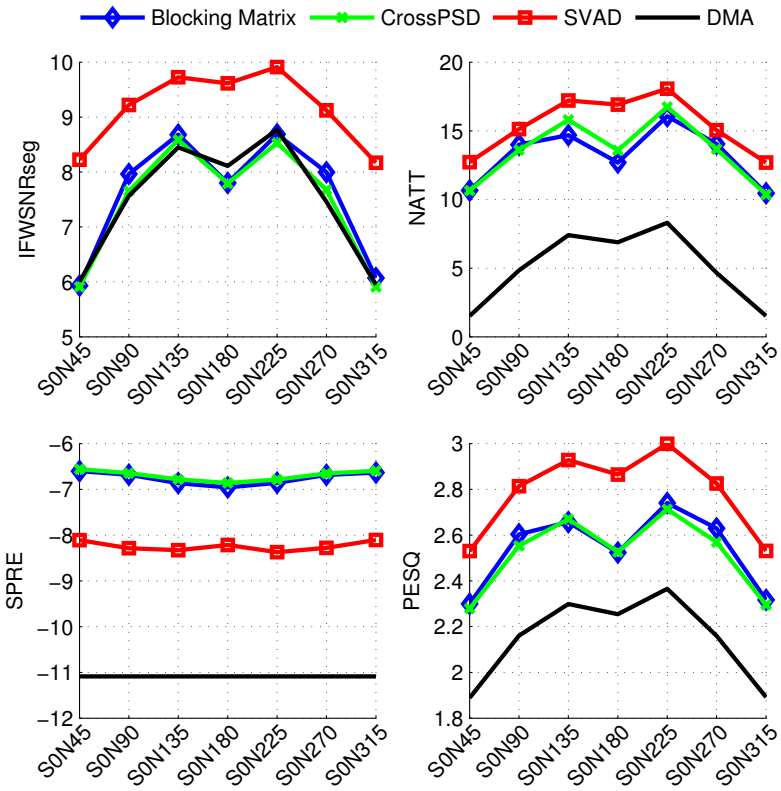

Figure 7.9: Noise reduction performance for difference noise PSD estimation algorithms employed in DMA-BPF at 10 dB SNR.

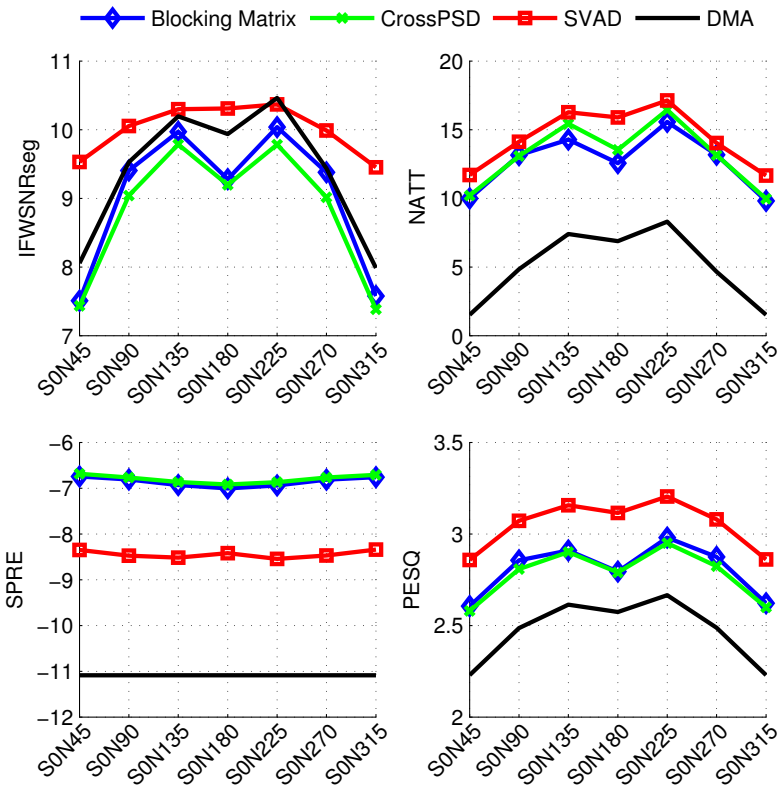

Figure 7.10: Noise reduction performance for difference noise PSD estimation algorithms employed in DMA-BPF at 15 dB SNR.

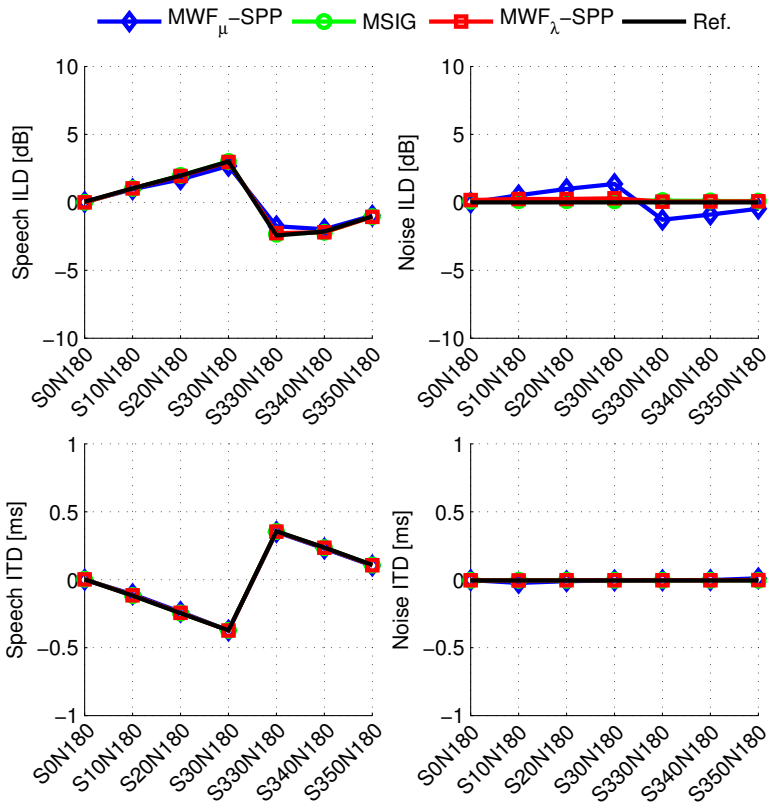

Figure 7.11: ITD and ILD results for MWF when direction of noise was fixed with speech source coming from different directions.

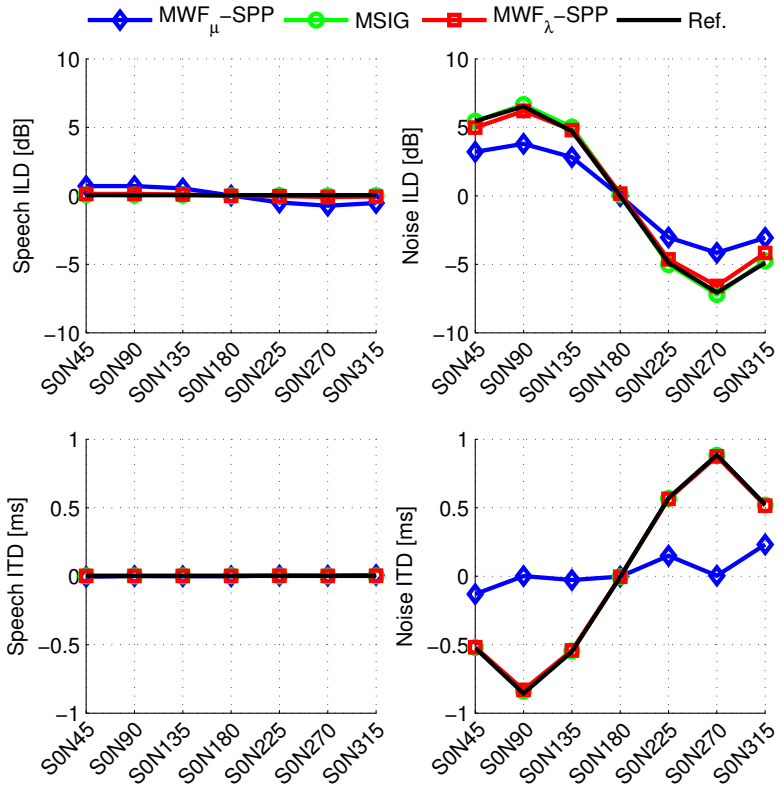

Figure 7.12: ITD and ILD results for MWF when direction of the target speech was fixed with noise source coming from different directions.

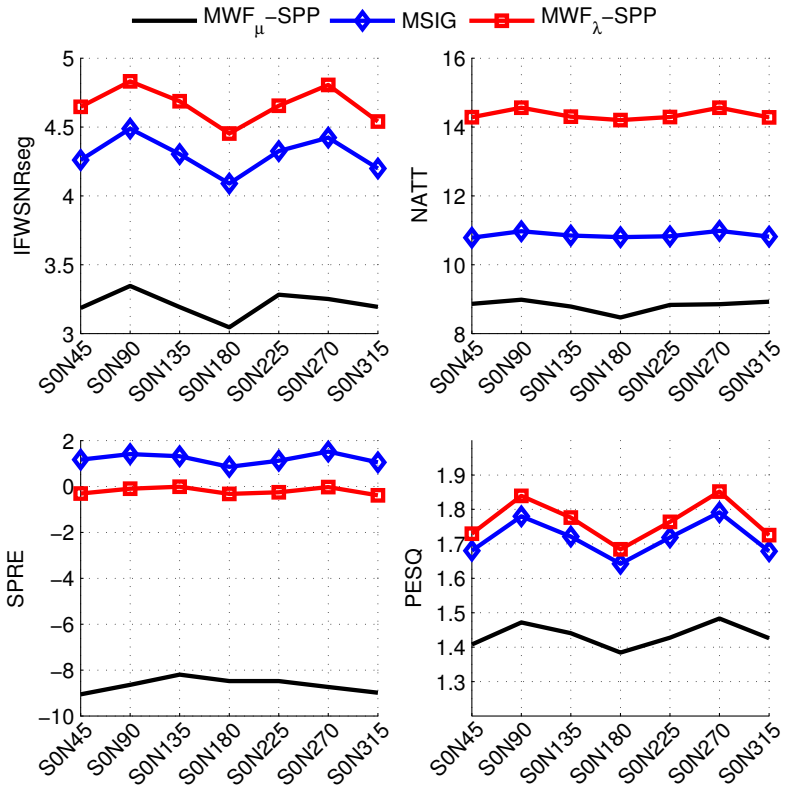

Figure 7.13: Noise reduction performance comparison among MWF$_\mu$-SPP, MSIG, and MWF$_\lambda$-SPP at 0 dB SNR.

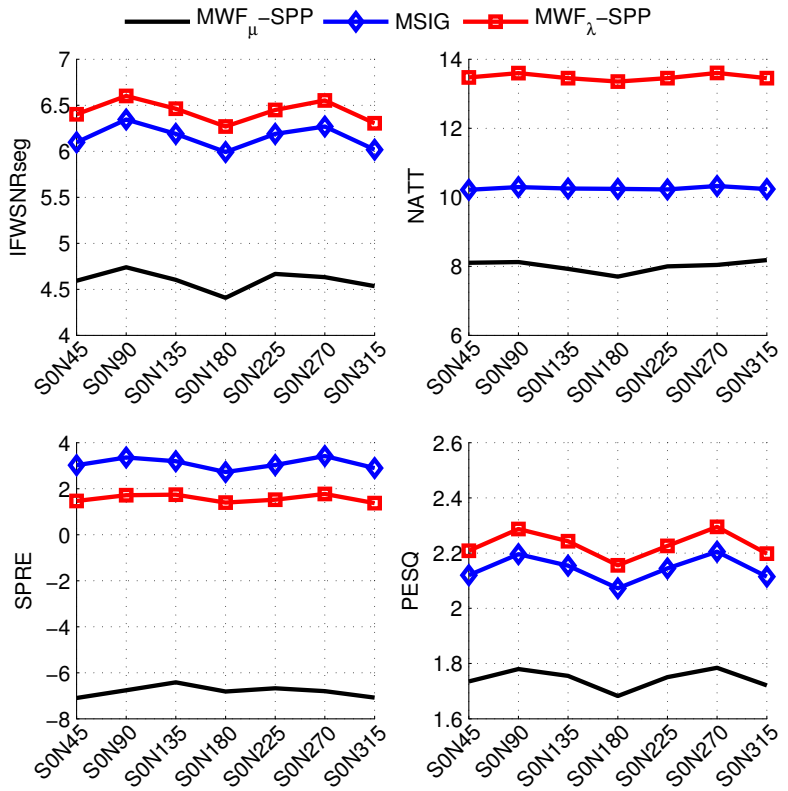

Figure 7.14: Noise reduction performance among MWF$_\mu$-SPP, MSIG, and MWF$_\lambda$-SPP at 5 dB SNR.

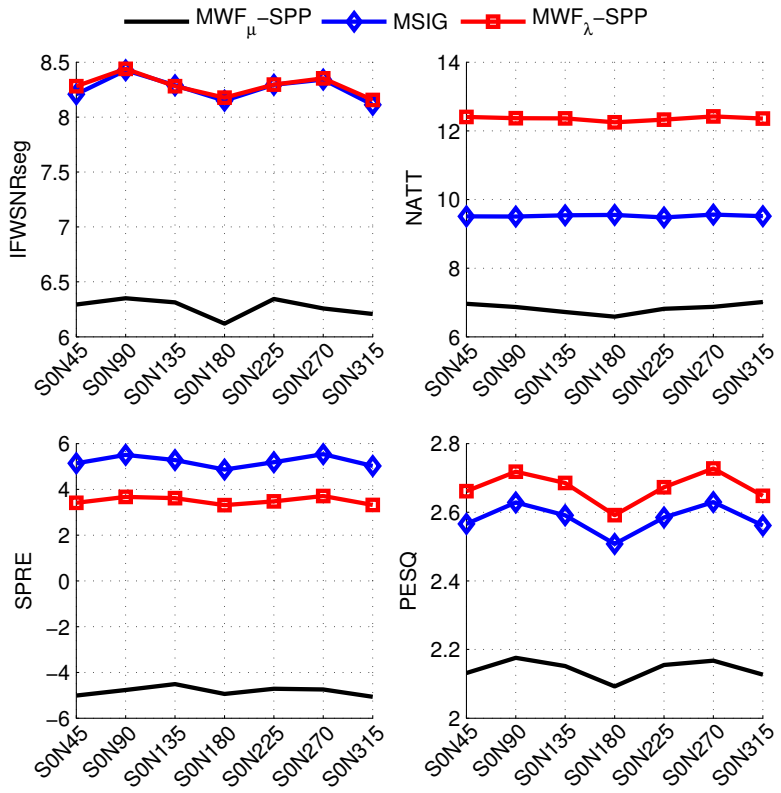

Figure 7.15: Noise reduction performance among MWF$_\mu$-SPP, MSIG, and MWF$_\lambda$-SPP at 10 dB SNR.

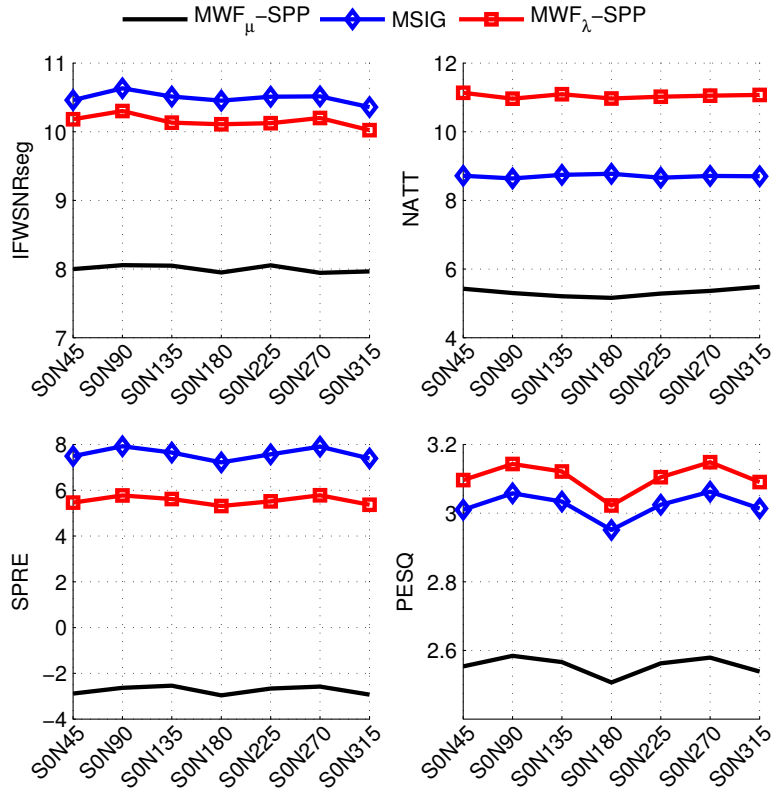

Figure 7.16: Noise reduction performance among MWF$_\mu$-SPP, MSIG, and MWF$_\lambda$-SPP at 15 dB SNR.

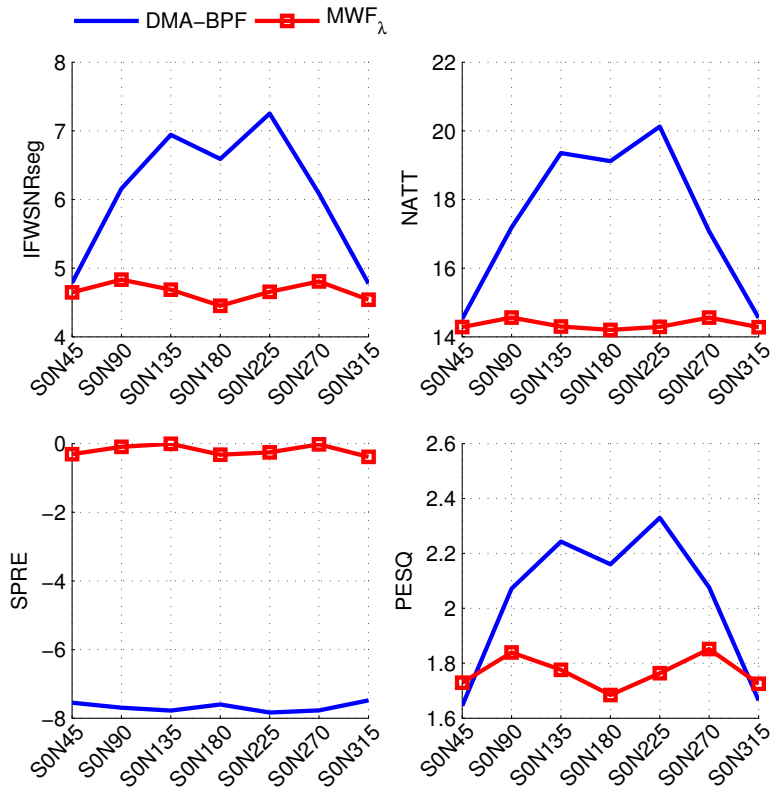

Figure 7.17: Noise reduction performance comparison between DMA-BPF and MWF at 0 dB SNR.

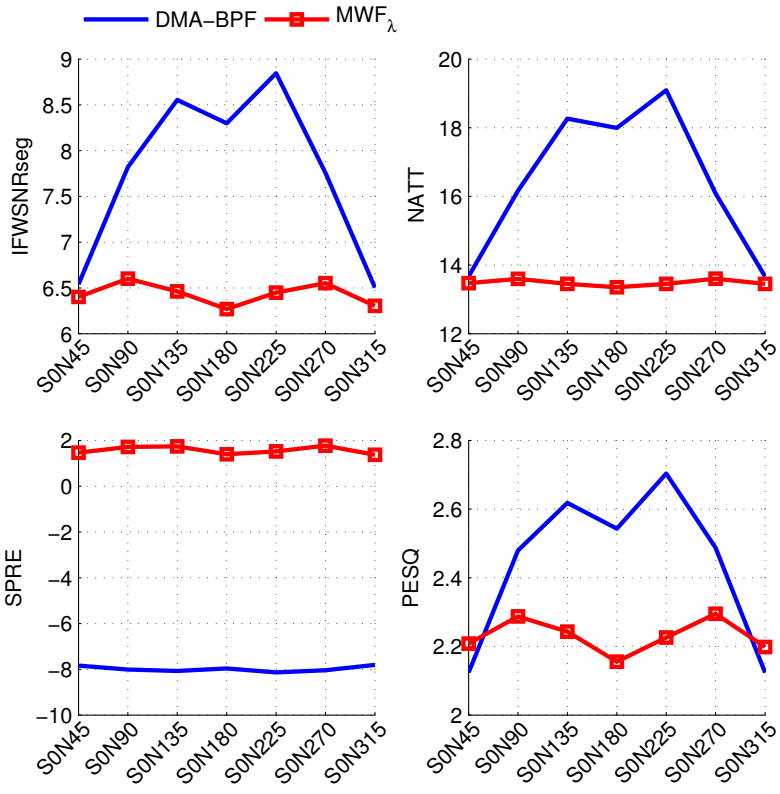

Figure 7.18: Noise reduction performance comparison between DMA-BPF and MWF at 5 dB SNR.

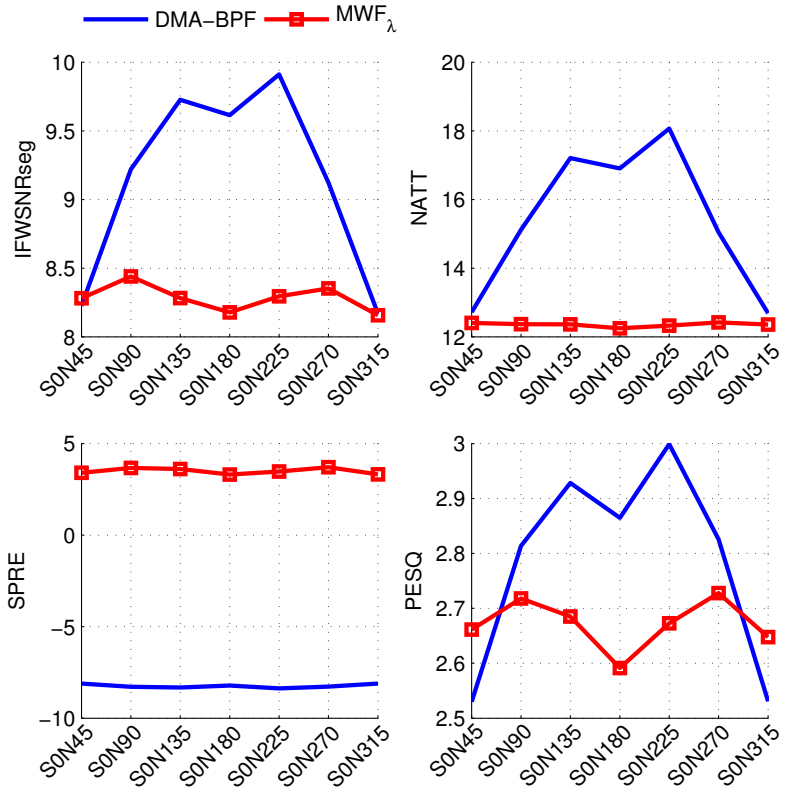

Figure 7.19: Noise reduction performance comparison between DMA-BPF and MWF at 10 dB SNR.

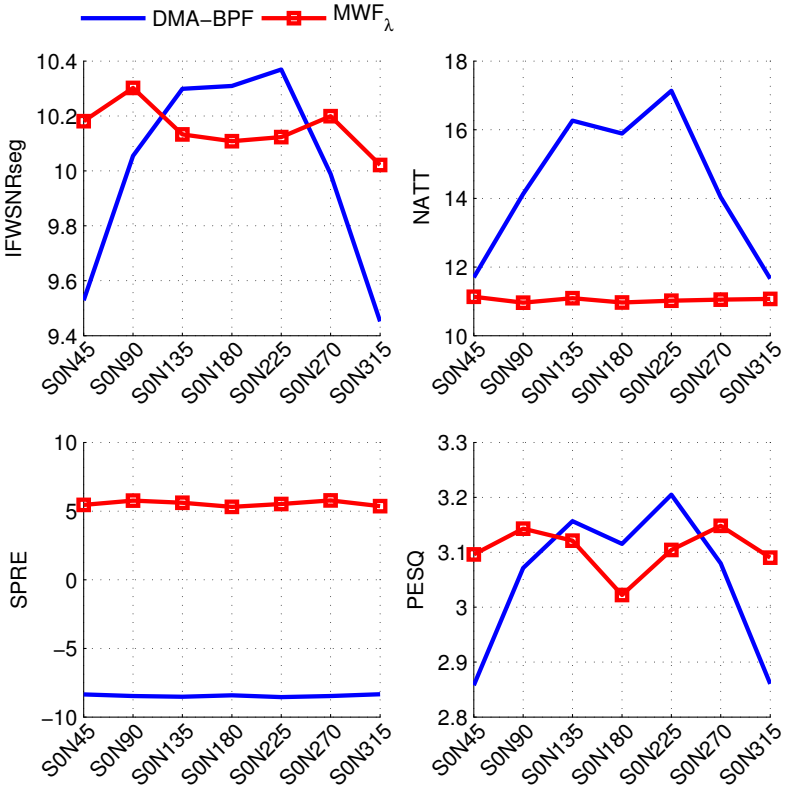

Figure 7.20: Noise reduction performance comparison between DMA-BPF and MWF at 15 dB SNR.

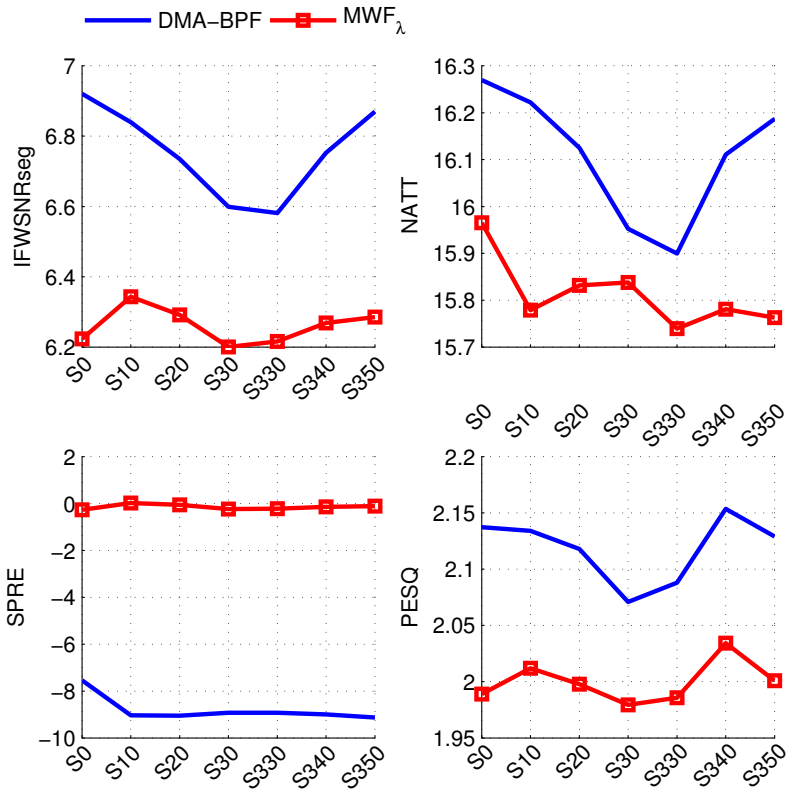

Figure 7.21: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like factory noise at 0 dB SNR.

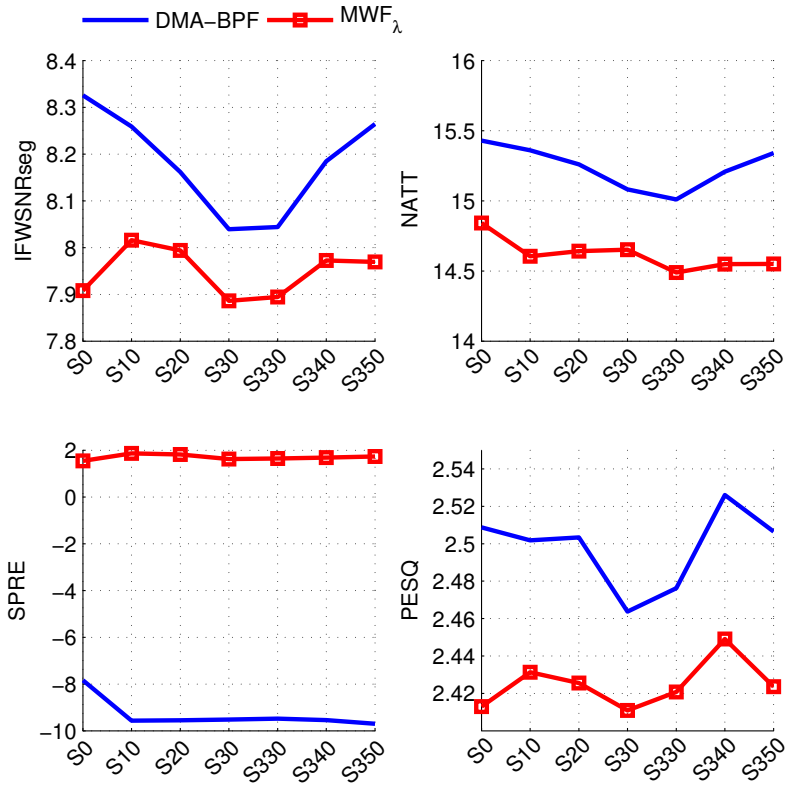

Figure 7.22: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like factory noise at 5 dB SNR.

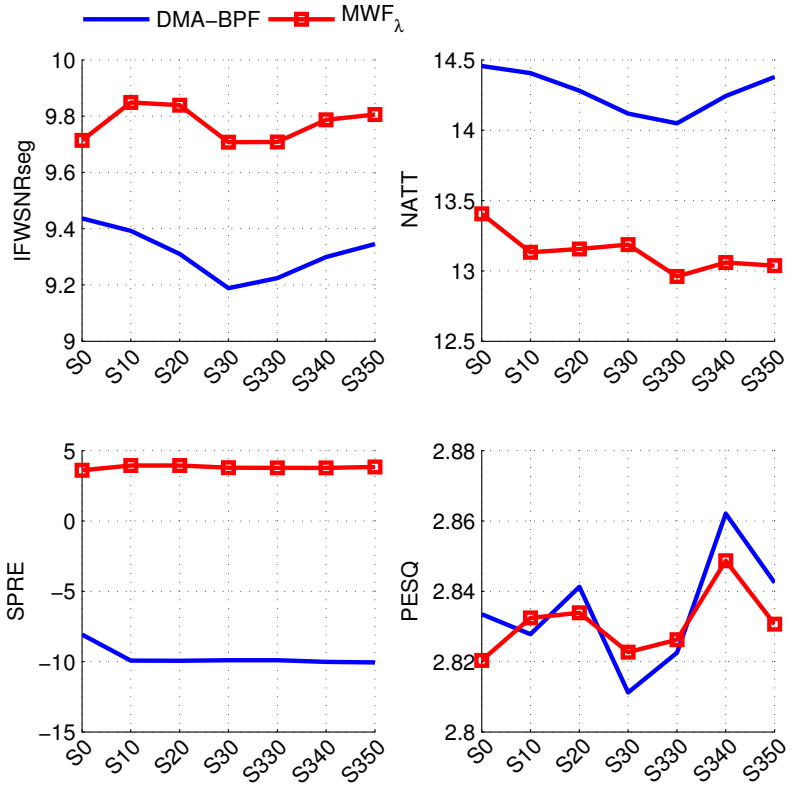

Figure 7.23: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like factory noise at 10 dB SNR.

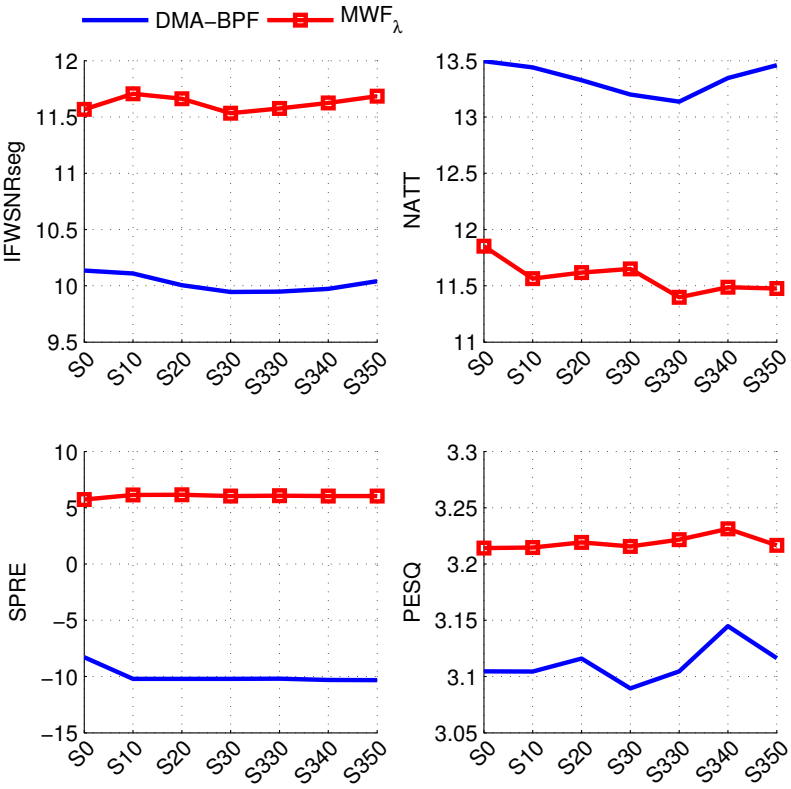

Figure 7.24: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like factory noise at 15 dB SNR.

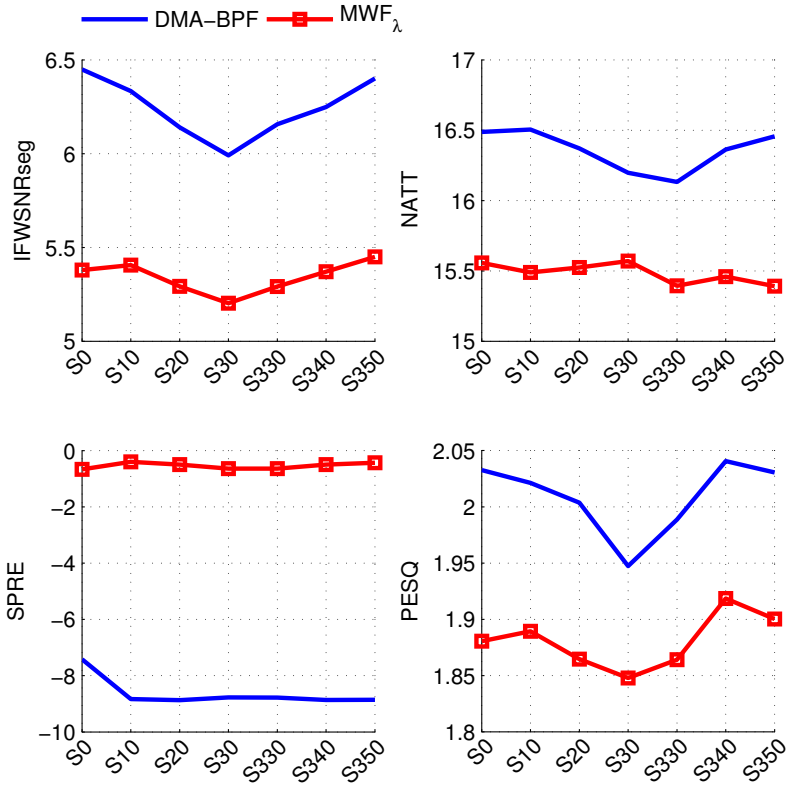

Figure 7.25: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like jack-hammer noise at 0 dB SNR.

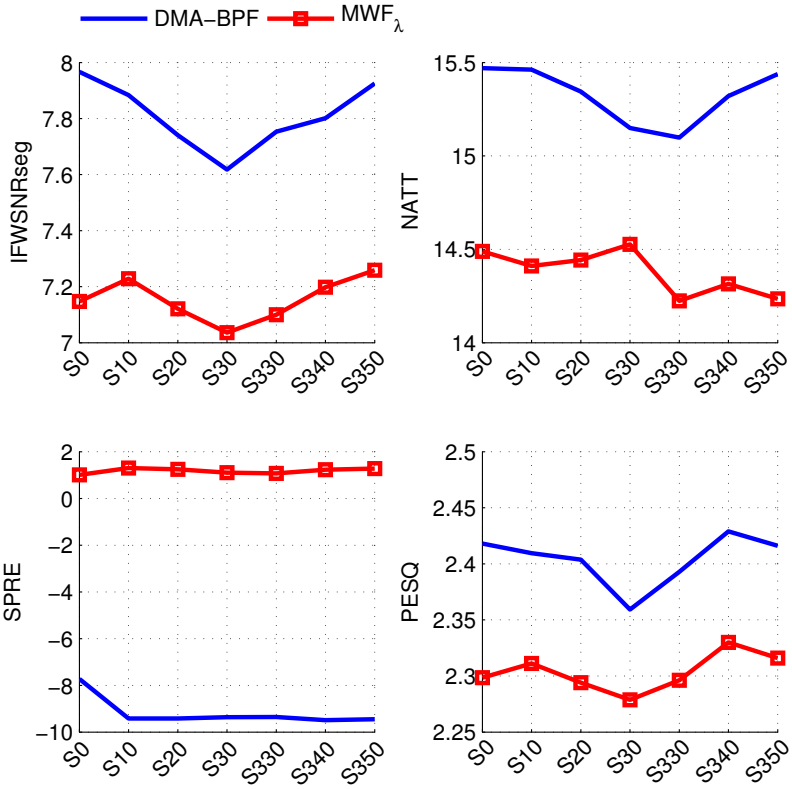

Figure 7.26: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like jack-hammer noise at 5 dB SNR.

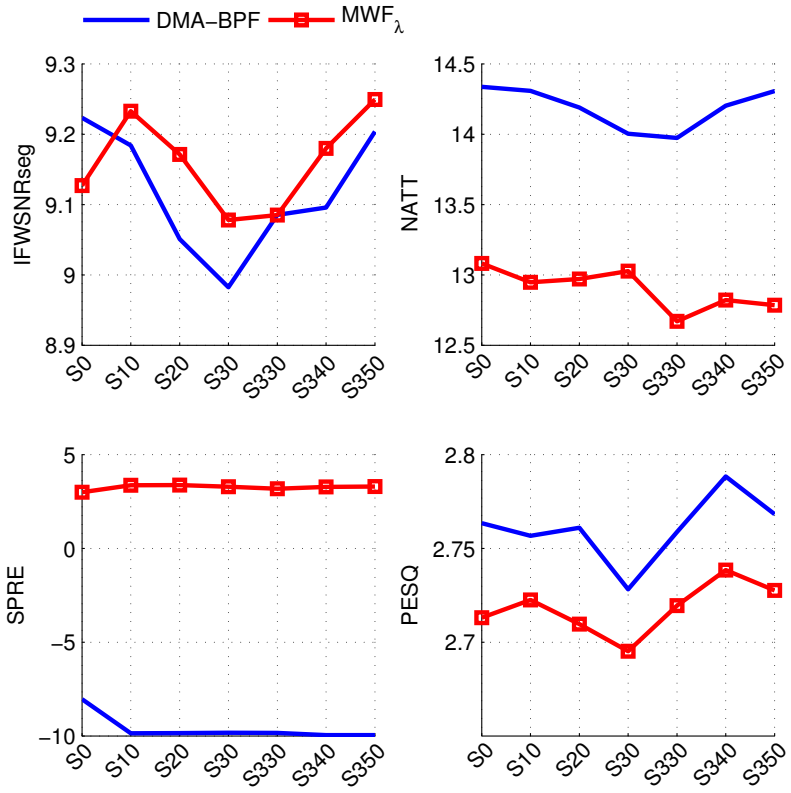

Figure 7.27: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like jack-hammer noise at 10 dB SNR.

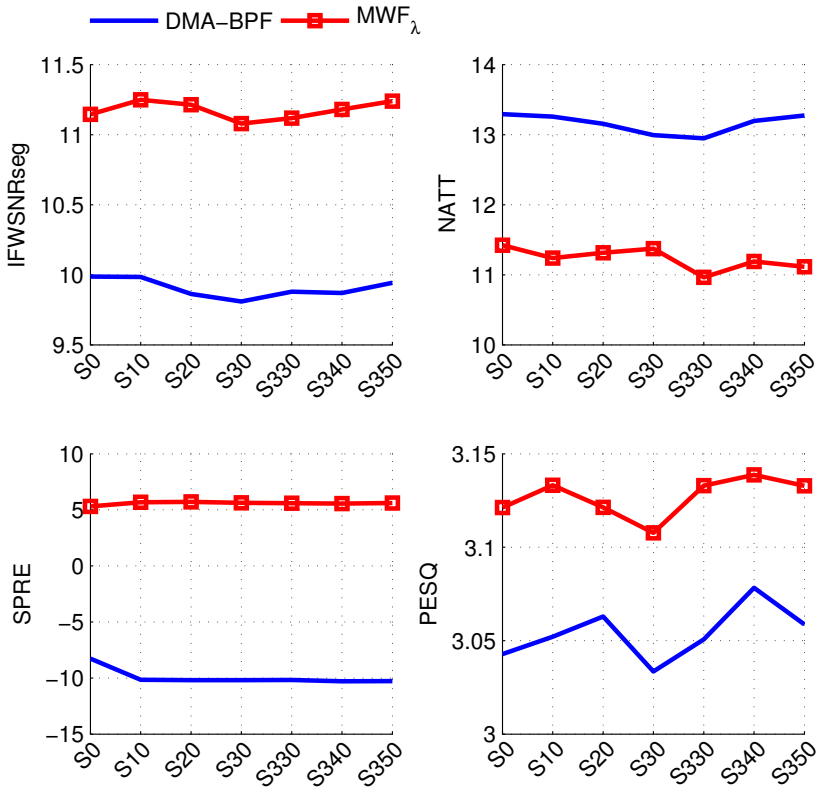

Figure 7.28: Noise reduction performance comparison between DMA-BPF and MWF for diffuse-like jack-hammer noise at 15 dB SNR.

7.7 Summary

This chapter presents two possible solutions for binaural speech enhancement for a hearing protection device, containing two microphones at each side of the earmuffs. The first one utilises a DMA at each side of the ear to suppress unwanted signals from the back, which leads to one single-channel DMA output at each side of the ear. After that, the remaining noise is estimated from the respective sides and then identical single-channel gain functions are applied to both channels to attenuate the noise signals. New gain function and new noise estimation method were employed in this framework to build a system with low computational complexity and without *a priori* information about noise required.

The second approach involves a true beamforming technique, termed binaural MWF. The drawback of this processing technique is that it does not preserve the spatial cues of noise, which is critical for industrial workers. Another issue with regards to the implementation of MWF is the flaw in the estimation of the second order statistics, which often requires the aid of a voice activity detection (VAD) to detect speech presence and absence. Therefore, the prospective solution presented in this thesis incorporates the binaural MWF with the single-channel noise reduction approach. As such, the speech and noise components in the framework were proposed to be continuously estimated by utilising a single-channel conditional SPP approach and a single-channel spectral weighing gain function.

Some conclusions of this chapter include:

(i) What is the influence of the binaural noise reduction algorithm with DMAs in a dual monaural configuration and a binaural postfilter that combines two identical single-channel gain functions on the ability to localise sources?

The DMAs has marginally distorted the ITDs and ILDs at different configurations, particularly when the sources are originated from behind the head. The degree of distortion is less with the sources are located from -40° to 40°.

(ii) Does the single-channel noise estimation algorithm affect the localisation and noise reduction performance compared to the Blocking Matrix algorithm and CrossPSD algorithm?

The SVAD does not affect the ITDs and ILDs at all different configurations of speech and noise sources. In contrast, the results showed that the Blocking

Matrix and CrossPSD algorithms can distort the binaural cues as they are also built based on the assumptions of the target speech source.

(iii) How do the proposed MWF perform in terms of combining localisation and noise reduction performance in comparison to the conventional SDW-MWF approach?

When compared to the SDW-MWF method, which cannot preserve the binaural cues of noise source, the proposed MWF formulation can preserve the ITDs and ILDs of both speech and noise. The proposed algorithm has also outperformed the traditional method in terms of the SNR improvement without introducing much speech distortion. The PESQ scores has indicated that the SDW-MWF is less preferred than the new approach.

(iv) How does the performance of the new MWF in comparison with the binaural technique with DMAs and single-channel gain functions?

The DMA-BPF algorithm has better SNR improvement and perceptual speech quality for most cases when compared to the MWF$_\lambda$-SPP approach, particularly when there is enough spatial distinctiveness between speech and noise. In contrast, MWF$_\lambda$-SPP has a higher consistency in performance for different speech and noise configurations. At diffused noise environment, the MWF$_\lambda$-SPP approach performs better than the DMA-BPF at high input SNR due to the consistently less speech distortion results for MWF$_\lambda$-SPP.

Chapter 8

Conclusions

Life is like riding a bicycle.
To keep your balance you must keep moving.
– Albert Einstein

8.1 Summary

This thesis dealt with speech enhancement in binaural hearing protection devices in a noisy environment, particularly with industrial noise. Four main criteria has been emphasised in the design (i) low computational complexity for long hour of usage at work, (ii) good noise reduction performance, (iii) continuous background noise tracking to cope with a changing environment, i.e., from noisy to quiet environment and vice versa, and (iv) spatial awareness preservation by having algorithms with short delays while being capable of maintaining binaural cues of speech and noise. New algorithms and approaches are proposed to achieve these criteria.

Chapter 2 studied the capability of speech enhancement algorithms in providing solutions to the problems in conventional hearing protection device (HPD), particularly in reducing the ambient noise while preserving the target speech signal and maintaining the spatial awareness. To build a binaural speech enhancement framework both the single-channel and multi-channel speech enhancement algorithms are applicable. The single-channel methods are suffice to decrease large amount of background noise as they are signal-to-noise ratio (SNR) based

algorithm. However, they often come with speech distortion and also musical noise. Multi-channel algorithms, on the other hand, exploit the spatial diversity of speech and noise and are able to decrease the background noise without introducing excessive artifacts.

Chapter 3 aimed at developing a spectral gain function for single-channel method with more flexibility and lower computational complexity.

- The SNR estimate and the gain function has been shown to impact the objective measures and provide varying subjective quality.
- A sigmoid function has been presented with a methodology to optimise the mean and the slope of the sigmoid function based on a proposed objective function.
- The sigmoid parameters were designed such that during the noise only periods, constant suppression is enforced thus avoiding annoying musical noise. This was achieved by mapping the function to the distribution of the SNR estimate.
- Optimisation of sigmoid function has been done based on two widely used objective measures: perceptual evaluation of speech quality (PESQ) and log-likelihood ratio (LLR).
- Experimental results have concluded that with a proper choice of parameters, the sigmoid function can be optimised to enhance the quality of the noisy speech while maintaining more energy of the speech components when compared to the spectral subtraction function.

Chapter 4 extended the idea in Chapter 3 in developing better gain function with the *a priori* SNR estimate.

- A new modified sigmoid (MSIG) function has been proposed to provide more flexibility to the gain function that can be optimised to match various criteria to achieve a compromised trade-off among speech distortion, noise reduction and musical noise.

- A new approach, namely the modified decision-directed (MDD) approach to estimate the *a priori* SNR has been proposed. This method is superior in that by reducing the one-frame delay, it reduces or to an extent, eliminates the speech transient distortion.

- The musical noise is further reduced by means of a recursive averaging algorithm which smoothens the *a posteriori* SNR. This level of smoothing is controlled by factors β and α_y.

In Chapter 5, two noise estimators have been presented - the step-size controlled (SSC) algorithm and the soft VAD (SVAD) algorithm.

- The first approach, namely the SSC algorithm has the advantage of low computational complexity and low immediate sensitivity to speech onsets. By using a fixed step-size, the developed estimator can efficiently track the variations in the noise spectrum. It is established that the optimal step-size for a specific noise scenario results in a noise estimation with higher precision. Alternatively, the step-size can be optimised such that one defined step-size works over a wide range of noisy situations. In this work, only pink noise and factory noise were chosen for that purpose. The performance of the proposed method is comparable to the conventional methods for a noisy environment. Under low background noise conditions, the performance of the SSC algorithm drops to its tendency to overestimate noise power. However, it is suitable for implementation in the HPD because of the lower computational complexity compared to the conventional methods.

- The SVAD algorithm, introduced as the second approach, is illustrated with a sigmoid function to represent the conditional speech presence probability (SPP). A distinctive feature of this function is that the slope and mean of the curve can be manipulated independently. As a result, the SPP can be more flexibly characterised for a compromised trade-off between noise overestimation and underestimation. Also, the soft decisions are made harder by employing different exponential smoothing at different regions of the sigmoid function. The SVAD produces the overall better noise tracking and speech quality performance when compared to the evaluated algorithms.

Chapter 6 presented an alternative speech distortion weighted MWF (SDW-MWF) formulation that deals with the nonstationarity of speech and noise.

- It utilises single-channel noise reduction technique to estimate a reference channel, and thus eliminating the requirement of the clean speech correlation matrix estimate.

- The single-channel algorithm employs a noise estimation method based on a modified conditional SPP, which also enables the regulating of the trade-off parameter. It is further used as the precursor for the estimation of the noise correlation matrix and the speech plus noise correlation matrix.

- The rank-one solution for the proposed formulation has also been developed and included for performance evaluation.

- Experimental results confirm that the proposed method performs better than the traditional method for all performance measures. A key finding from the experiment reveals that by incorporating SPP in the trade-off parameter λ, speech distortion is reduced, but more residual noise and musical noise are generated in the enhanced signals. The application of the proposed method in binaural configurations is showed in Chapter 7.

Chapter 7 presented two frameworks of binaural speech enhancement that simultaneously utilise both single-channel and multi-channel algorithms.

- Investigation has been done to identify Several problems in the frameworks, i.e., (i) there is room for improvement in the design of single-channel gain functions, (ii) noise tracking algorithms to replace voice activity detection (VAD) algorithms, and (iii) SDW-MWF preserves only binaural cues of target speech but not the noise cues.

- The proposed algorithms in Chapters 3 to 6 are integrated into the binaural speech enhancement frameworks, with performance outcomes analysed in terms of the SNR improvement and overall speech quality.

- The new formulation of SDW-MWF with conditional SPP preserves the binaural cues of both speech and noise.

- The SVAD provides higher noise reduction capability compared to the blocking matrix and coherence-based algorithm.
- Comparison of the proposed MWF$_\lambda$-SPP and the DMA-BPF indicated that the former has less speech distortion but lower SNR improvement.
- PESQ measurement showed that the DMA-BPF performed better at lower input SNR while MWF$_\lambda$-SPP generates better sound quality at higher input SNR conditions.
- The performance of MWF$_\lambda$-SPP is more consistent over different configurations of target speech location.

8.2 Future Research Directions

8.2.1 Parameters Selection for Modified Sigmoid Function

In this dissertation, a sigmoid function is employed as the single-channel gain function due to its flexibility in mapping with the SNR estimates. The parameters in the sigmoid functions are selected by using optimisation based on an objective-measures-constrained cost function (Chapter 3) and based on curve fitting with state-of-the-art gain functions (Chapter 4). To offer more flexibility in the gain functions to cope with different real-time scenarios, varying parameters can be considered under different motivation. One example is that recent research has demonstrated that speech quality in noise environment can be improved if the acoustic cues at low frequencies are well preserved [168, 169]. The reason behind this is that speech components are mainly located at low frequencies. A perceptually motivated frequency-specific Wiener filter (WF) has been proposed in [168], where a less aggressive gain function to used in place of the original WF gain function at low frequencies. Similarly, a recently proposed auditory-based minimum mean square error (MMSE) estimator also suggests a gain function with a decrease in the gain value at high frequencies compared to low frequencies [169]. This helps to increase the noise reduction, but also leads to more speech distortions at high frequencies. Although the distortions of the high frequency speech components, such as fricative consonants, are almost inaudible in low SNRs, they

could be more perceptible when there is less background noise [169]. Thus, it is desirable to have a larger gain at high instantaneous SNRs, while maintaining a rigid gain function at low instantaneous SNRs.

8.2.2 Incorporating Perceptual Criteria in Multi-channel Wiener Filter

In this dissertation, the conditional SPP has been utilised to adapt the trade-off parameters in the alternative formulation of SDW-MWF, which has been extended to binaural speech enhancement configurations. It was shown that this comes with higher amount of noise artefacts when compared to using an aggressive fixed value for more noise suppression. Future research can combine the conditional SPP with a perceptually motivated weighting factor. Inspiration can be taken from the characteristics of the human auditory system among which are the compressive non-linearities of the cochlea, the perceived loudness and the ears masking properties [169]. In this way the distortion could be kept low without compromising the amount of noise reduction.

8.2.3 Noise Estimation based on Structure of Noise

In this dissertation, a SVAD based noise power spectral density (PSD) estimation algorithm has been proposed, which employs soft decisions on top of a soft SPP with fixed priors. The derivation of the conditional SPP requires the assumptions of the distributions of noisy speech and noise signals, where complex Gaussian models are often used. However, the discrete Fourier transform (DFT) components of real noise hardly follow Gaussian distribution due to nonstationarity. The beauty of sigmoid function proposed in this thesis is that it offers flexibility to adjust the noise overestimation and underestimation. The parameters of the sigmoid function in this context were selected based on the bounds of fixed priors. Future improvement might include selecting the parameters based on the relationship of different assumptions of noise distribution, i.e., a generalised gamma model. However, this might indicate that the proposed sigmoid function

is no longer suitable to represent the conditional SPP equation, and thus different cumulative function, such as the Weibull function can be considered. Also, the SVAD algorithm can also be combined with the SSC algorithm which might provide improvement in the noise tracking performance.

8.2.4 Reduced Information Exchange for Binaural Speech Enhancement

In this thesis, the SDW-MWF frameworks utilises all channels at both sides of the ears to process the noisy speech signals and to obtain the enhanced speech signals. However, due to power and bandwidth limitations of the binaural link, it is typically not possible to transmit all microphone signals between each side. To limit the amount of transmitted information, either only the signals from the reference channels, or the filtered version of the contralateral signals are transmitted. The constraints of the signal transmission such as the bit rate has also to be considered. Such reduced bandwidth and rate constrained algorithms have recently received many attention and has opened up a pathway in developing more efficient binaural speech enhancement algorithms.

8.2.5 Evaluation of Speech Quality and Intelligibility

In this thesis, different objective evaluation measures have been used to predict the quality of the speech enhanced by noise reduction algorithms. However, most of them are not really consistent in performance over a wide range of non-stationary speech and noise scenarios. Thus, another pathway for future research directions is to design an objective evaluation metric that can better predict the performance in both speech quality and intelligibility. It is also desirable to conduct future evaluation on more speech and noise databases. In addition, one of the future works for the binaural speech enhancement algorithms are to conduct formal subjective listening tests to justify the results obtained from the objective evaluation measures.

Appendix A

Additional Results for Multi-Channel Wiener Filter

The greatest obstacle to discovery is not ignorance
-it is the illusion of knowledge.
– Daniel Boorstin

In Chapter 6, the proposed multi-channel Wiener filter (MWF) was compared with the traditional formulation under factory noise. In this appendix, some additional experimental results are included for completeness to examine the efficacy of the proposed algorithm across different types of noise. As such, two distinctive noise were included: a white Gaussian noise (WGN) to represent a stationary noise scenario, and a hammering noise indicating a nonstationary noisy environment. Parameters used in the algorithms are consistent with those used in Chapter 6.

Similar to Chapter 6, two sets of figures with input SNRs $-5\,\text{dB}$, $0\,\text{dB}$, $5\,\text{dB}$, and $10\,\text{dB}$ were plotted for each type of noise environment. The first set compares the performance of $\mathbf{w}_{\text{MWF}_{\lambda 1}}$ and $\mathbf{w}_{\text{MWF}_{\lambda 2}}$ with $\mathbf{w}_{\text{MWF}_{\mu}}$ and MSIG. The second set of results compares the performance of the rank-formulation $\mathbf{w}_{\text{MWF}_{\lambda}}$-rank1 with $\mathbf{w}_{\text{MWF}_{\mu}}$-rank1 and $\mathbf{w}_{\text{MWF}_{\lambda 1}}$. The figures are arranged such that the results for the same input SNR are plotted in the same page for a direct comparison between two types of noise environment.

Figures A.1, A.3, A.5, and A.7 depict the average first set's results for WGN, while Figures A.2, A.4, A.6, and A.8 show the results for hamming noise. It can

be seen that the results are more or less consistence across two types of noise. Observation shows that similar patterns have been obtained when compared to the results for factory noise in Chapter 6. $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ and $\mathbf{w}_{\mathrm{MWF}_{\lambda 2}}$ perform better than $\mathbf{w}_{\mathrm{MWF}_{\mu}}$ in terms of SNR improvement and overall speech quality for both types of noise and across different input SNRs. This merely means that the new MWF formulation is consistently better that the conventional approach, as it takes spectral tracking of nonstationary speech and noise into account. This motivates the extension of the algorithm to binaural speech enhancement configuration to examine its capability to preserve binaural cues, which has been discussed in Chapter 7.

Figures A.9 to A.16 show the results for the second set for both types of noise. Similar to the first set, they show consistency across different types of noise and different input SNRs. $\mathbf{w}_{\mathrm{MWF}_{\lambda}}$-rank1 still performs better than $\mathbf{w}_{\mathrm{MWF}_{\mu}}$-rank1 and $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ at low input SNRs, while $\mathbf{w}_{\mathrm{MWF}_{\lambda 1}}$ has better performance at high input SNR.

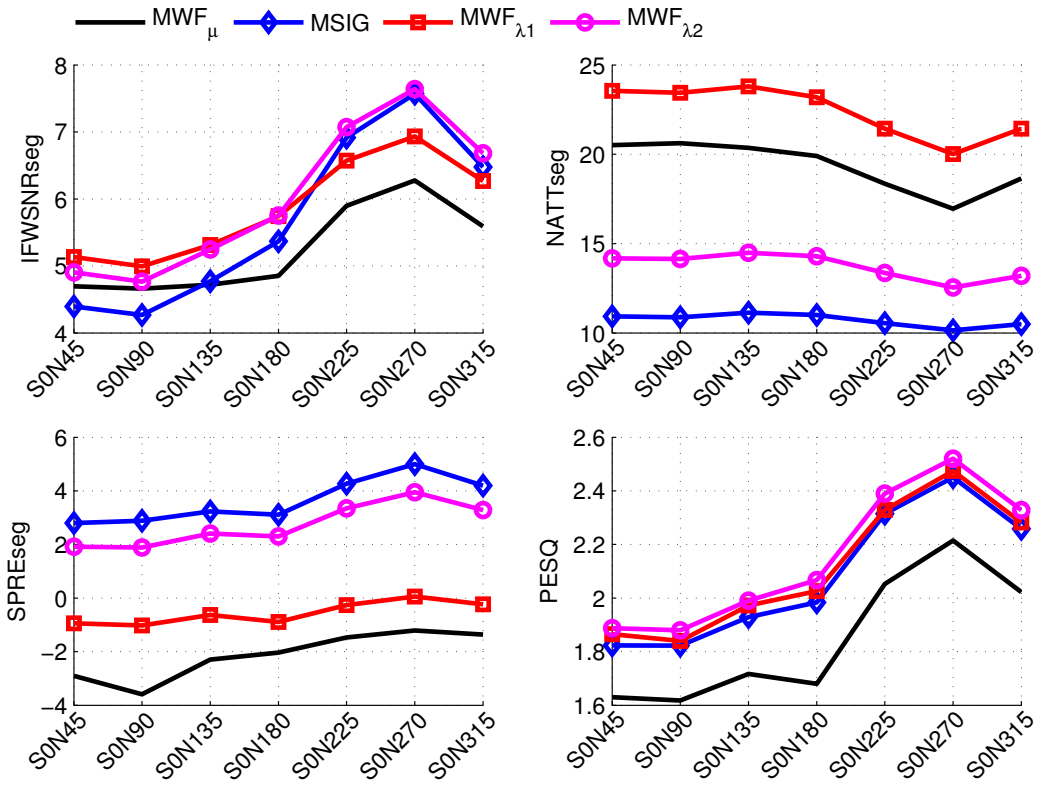

Figure A.1: Average results for WGN for input SNR −5 dB.

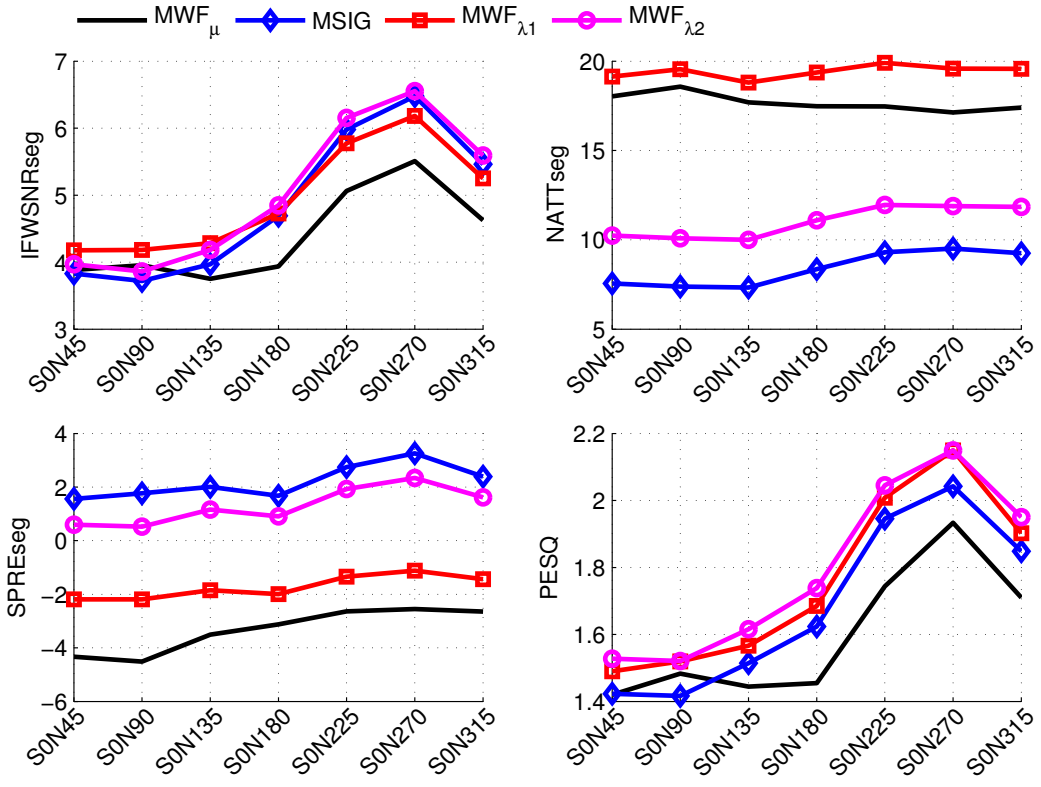

Figure A.2: Average results for hammering noise for input SNR −5 dB.

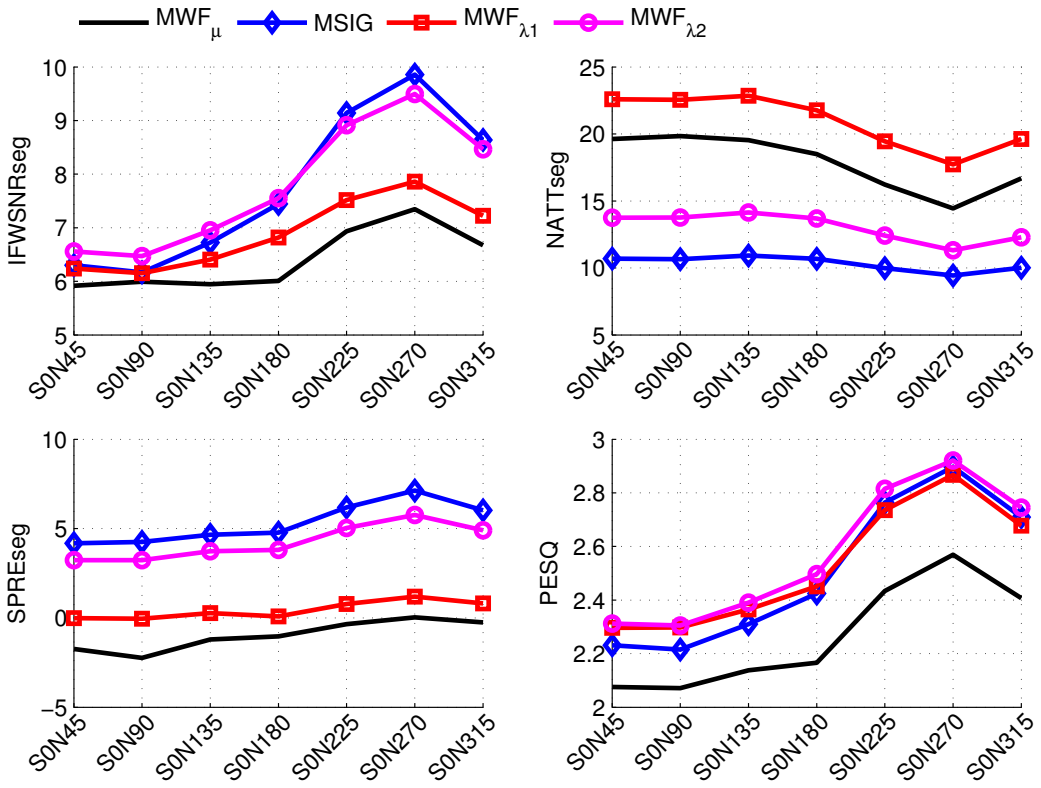

Figure A.3: Average results for WGN for input SNR 0 dB.

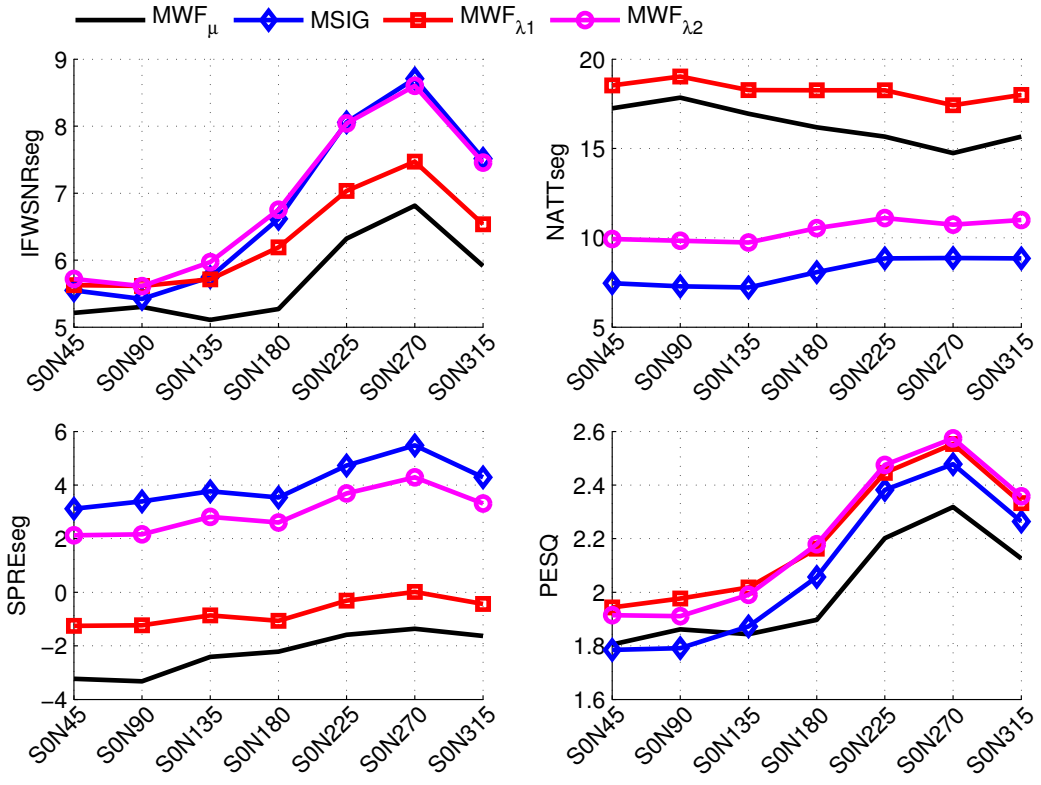

Figure A.4: Average results for hammering noise for input SNR 0 dB.

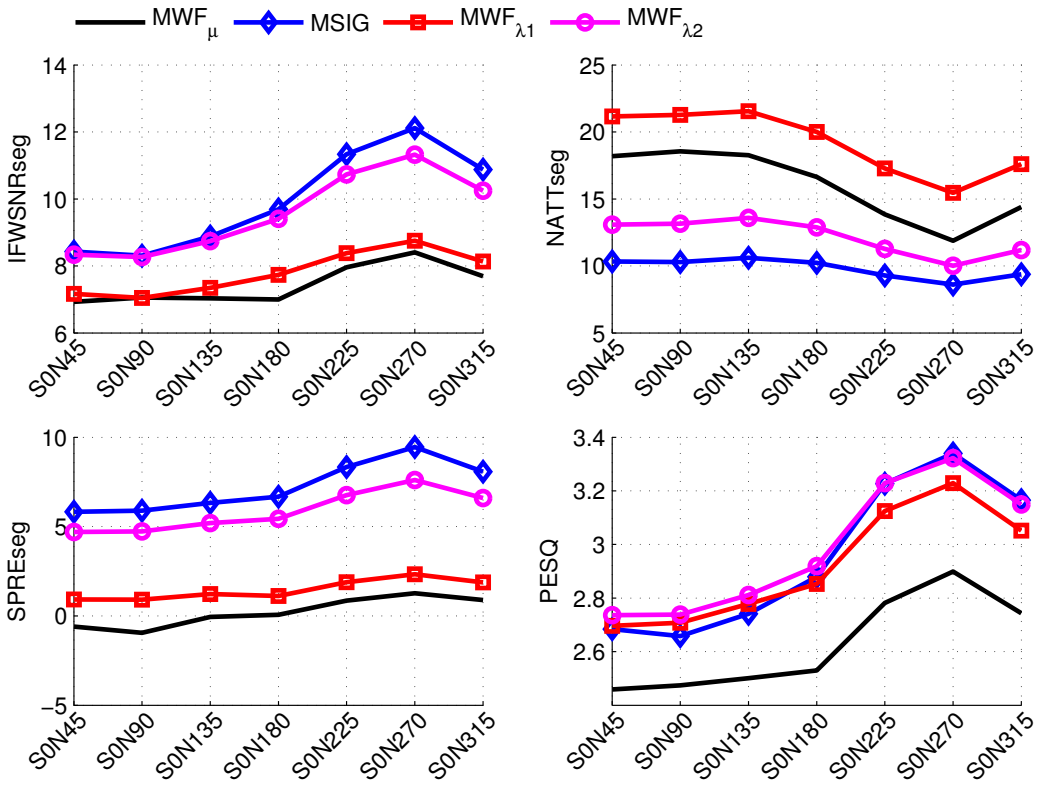

Figure A.5: Average results for WGN for input SNR 5 dB.

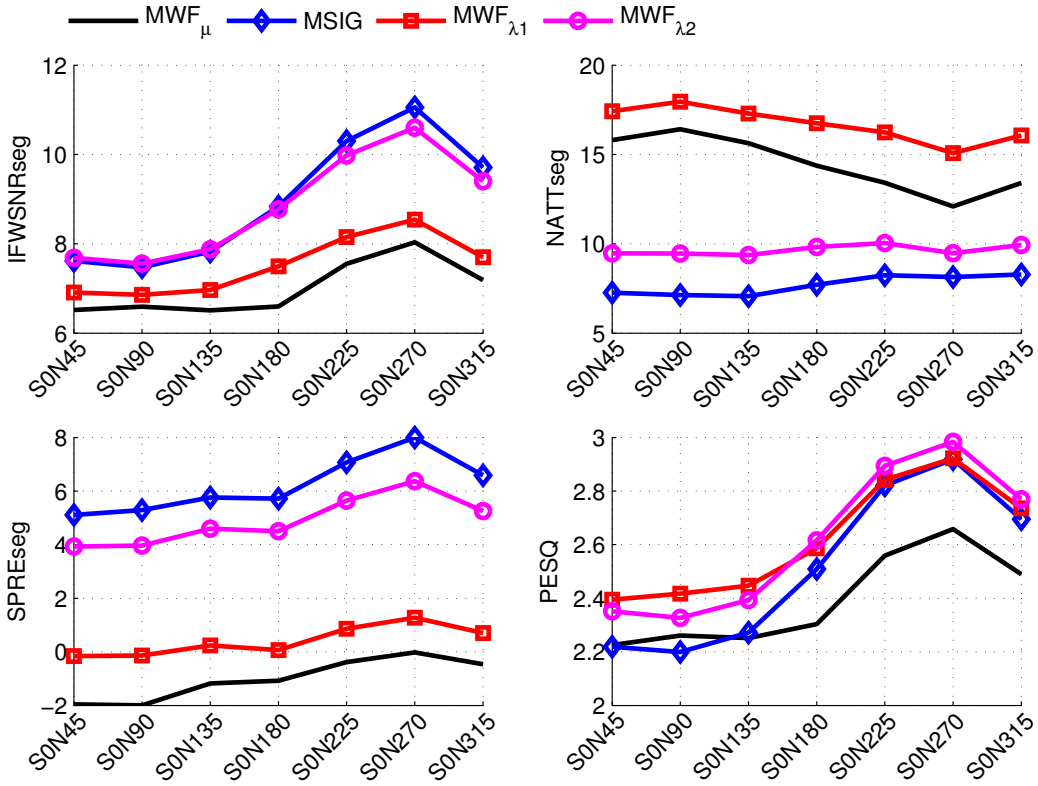

Figure A.6: Average results for hammering noise for input SNR 5 dB.

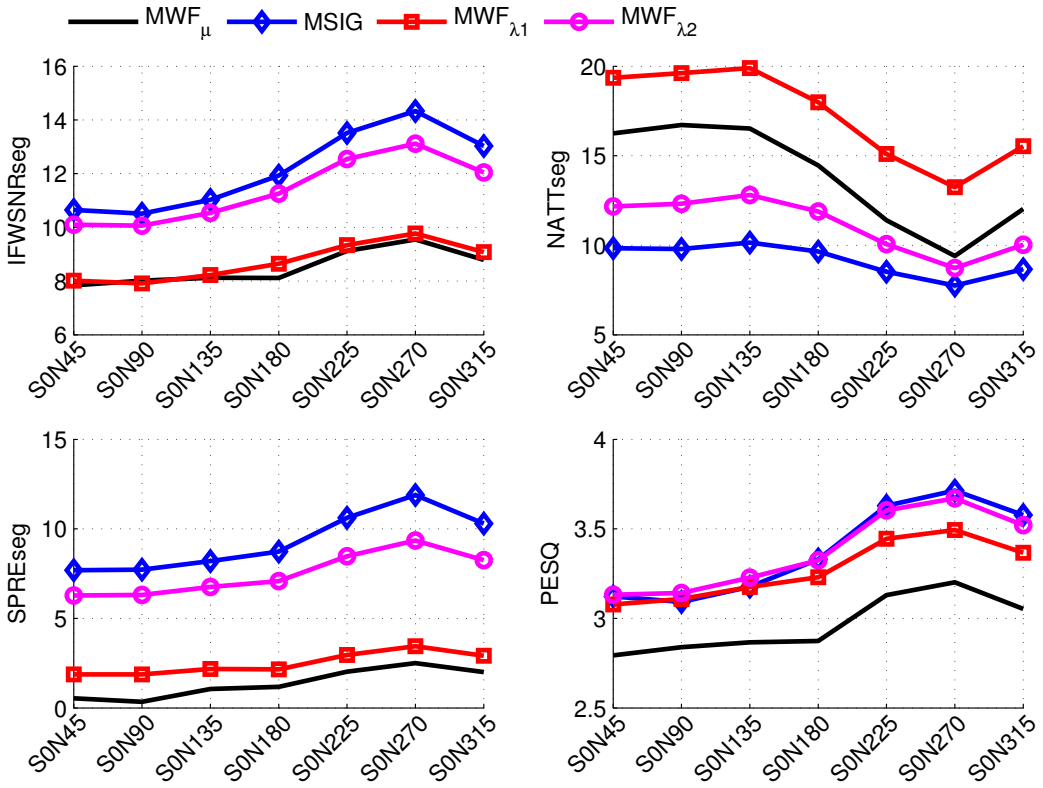

Figure A.7: Average results for WGN for input SNR 10 dB.

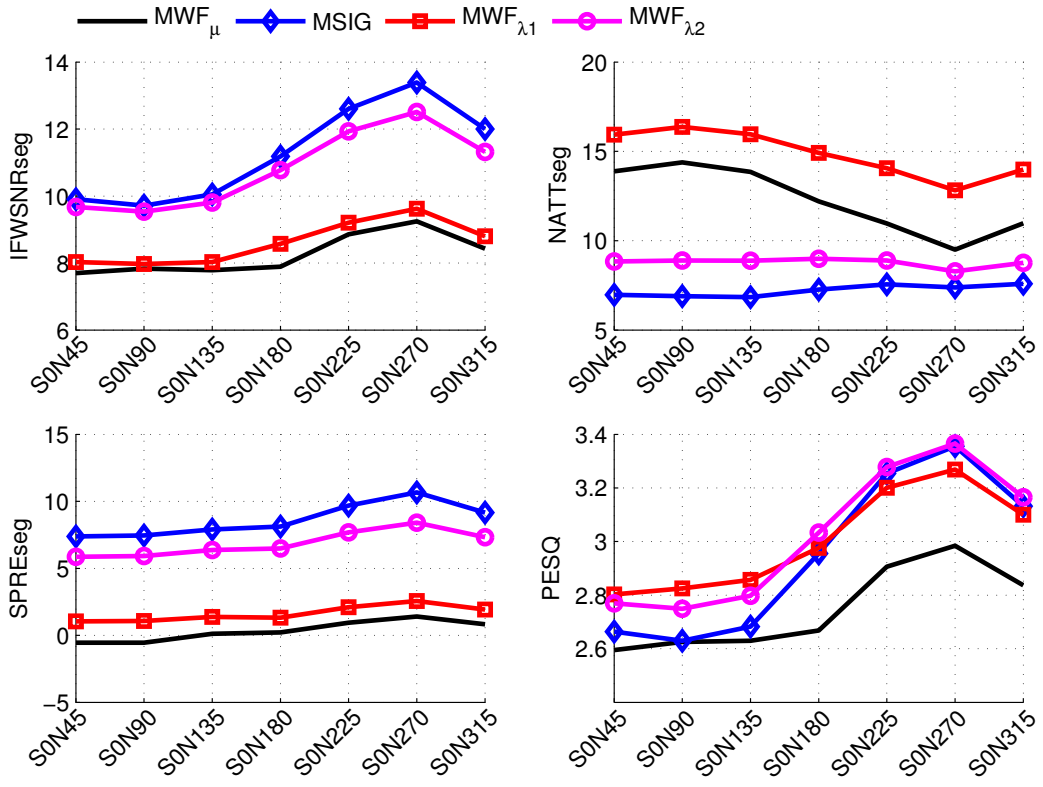

Figure A.8: Average results for hammering noise for input SNR 10 dB.

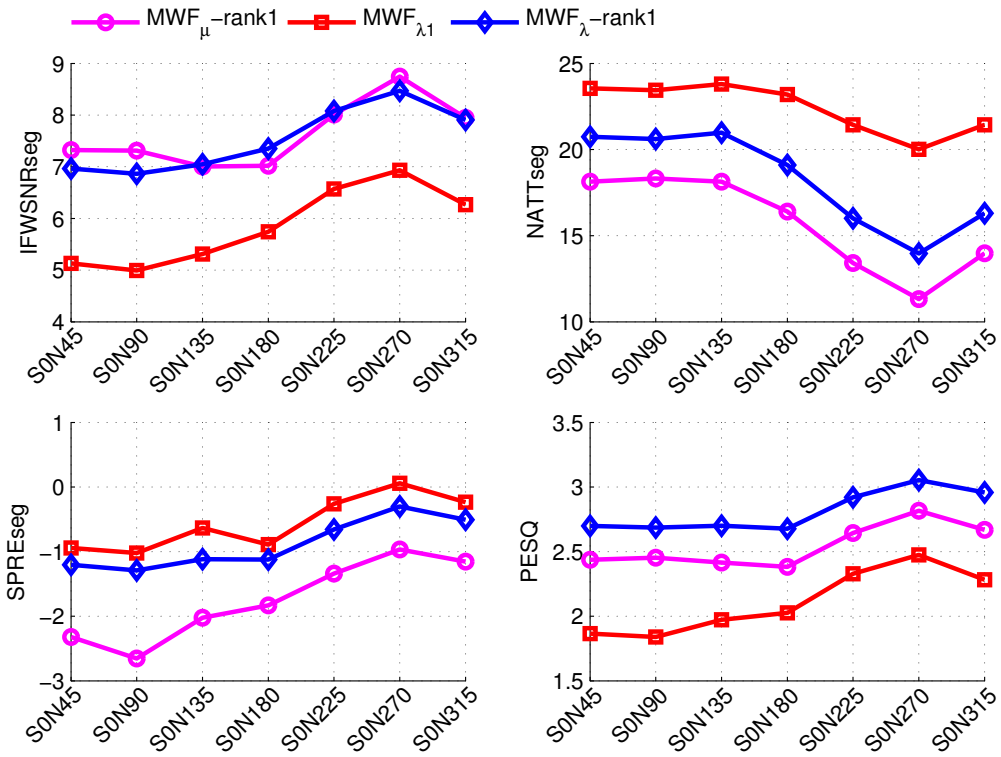

Figure A.9: Comparison between rank-one and general formulations for WGN for input SNR −5 dB.

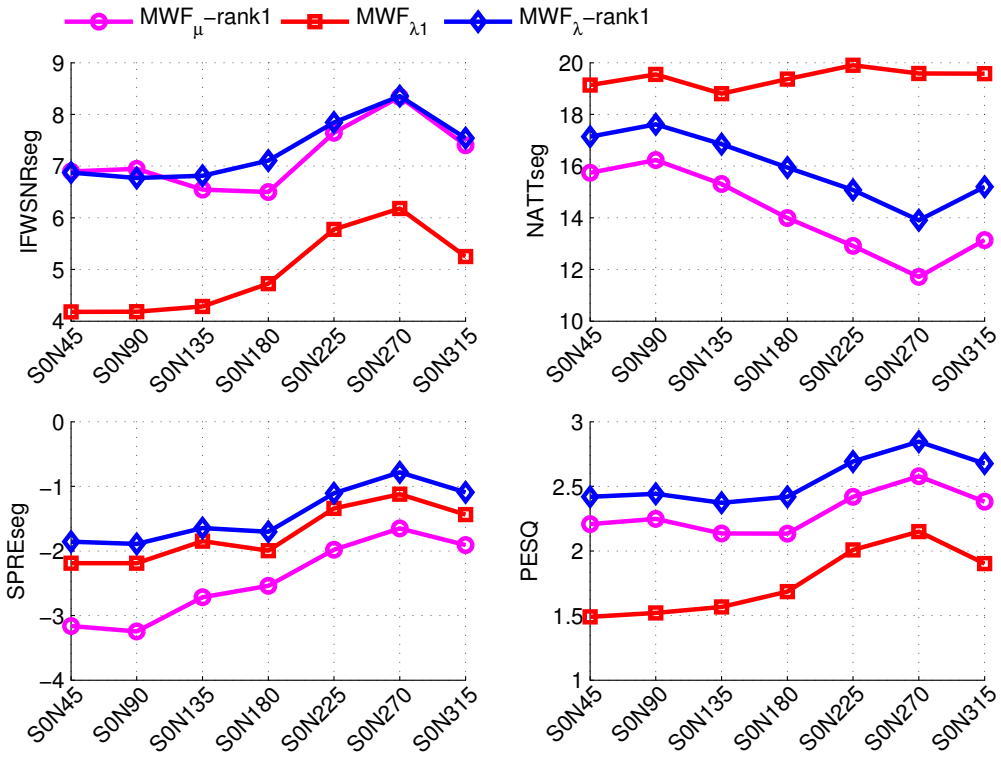

Figure A.10: Comparison between rank-one and general formulations for hammering noise for input SNR −5 dB.

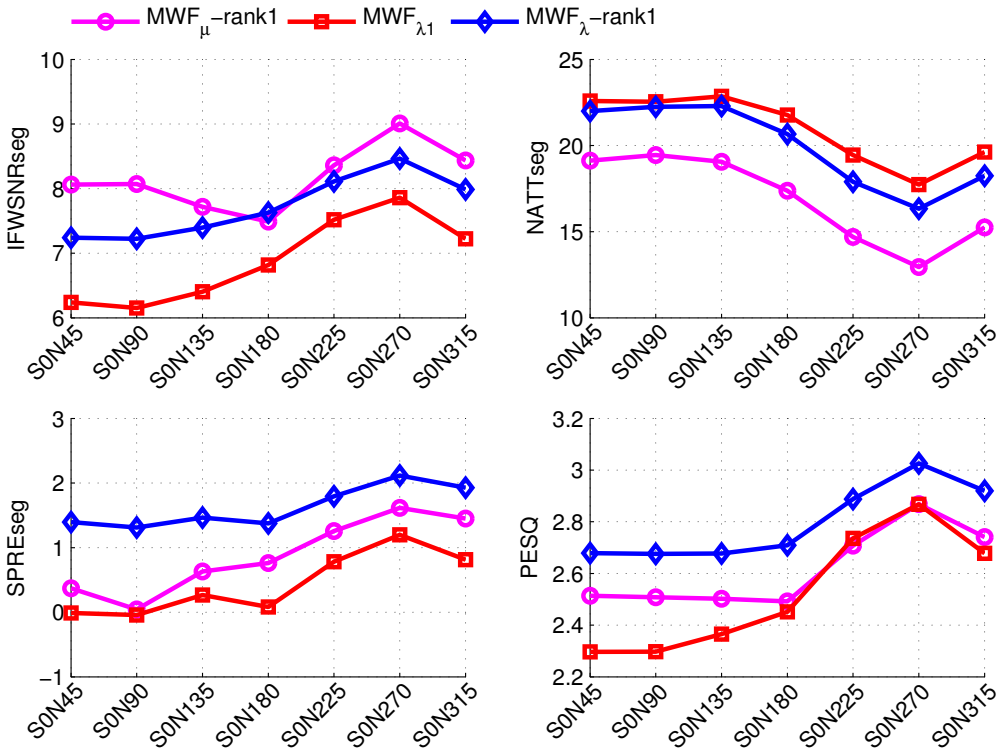

Figure A.11: Comparison between rank-one and general formulations for WGN for input SNR 0 dB.

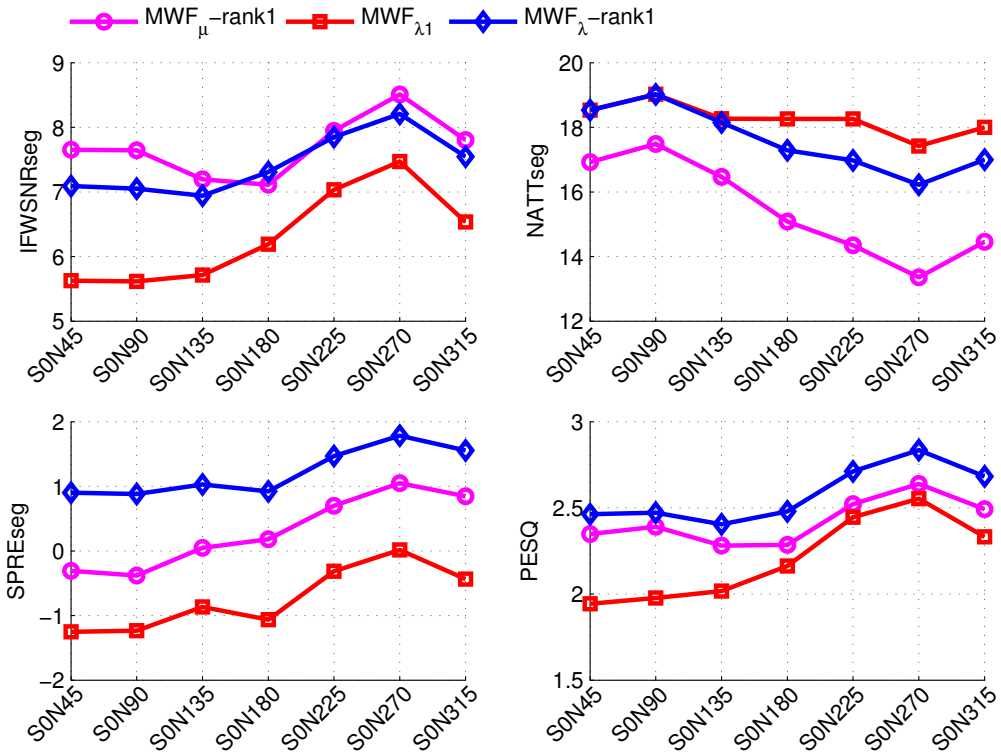

Figure A.12: Comparison between rank-one and general formulations for hammering noise for input SNR 0 dB.

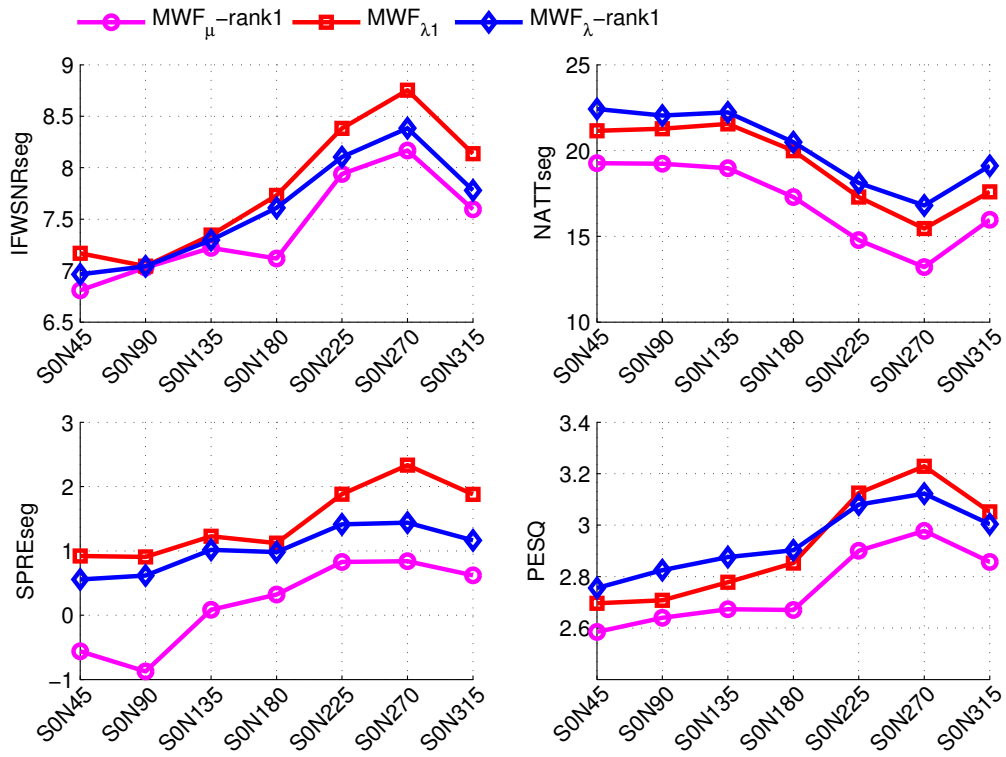

Figure A.13: Comparison between rank-one and general formulations for WGN for input SNR 5 dB.

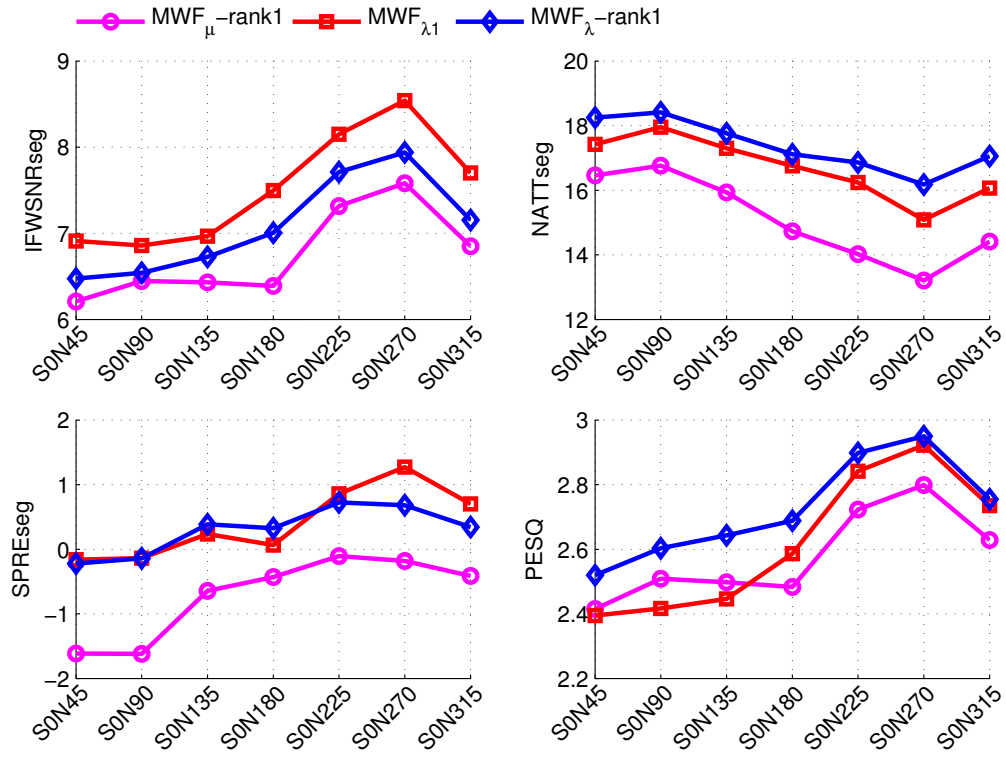

Figure A.14: Comparison between rank-one and general formulations for hammering noise for input SNR 5 dB.

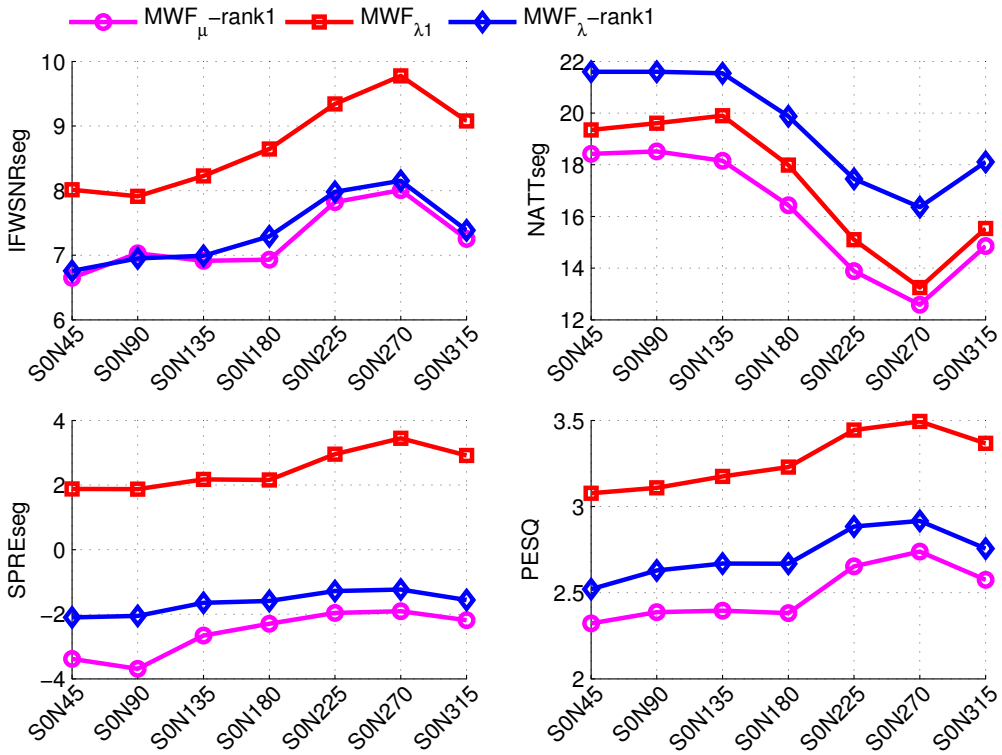

Figure A.15: Comparison between rank-one and general formulations for WGN for input SNR 10 dB.

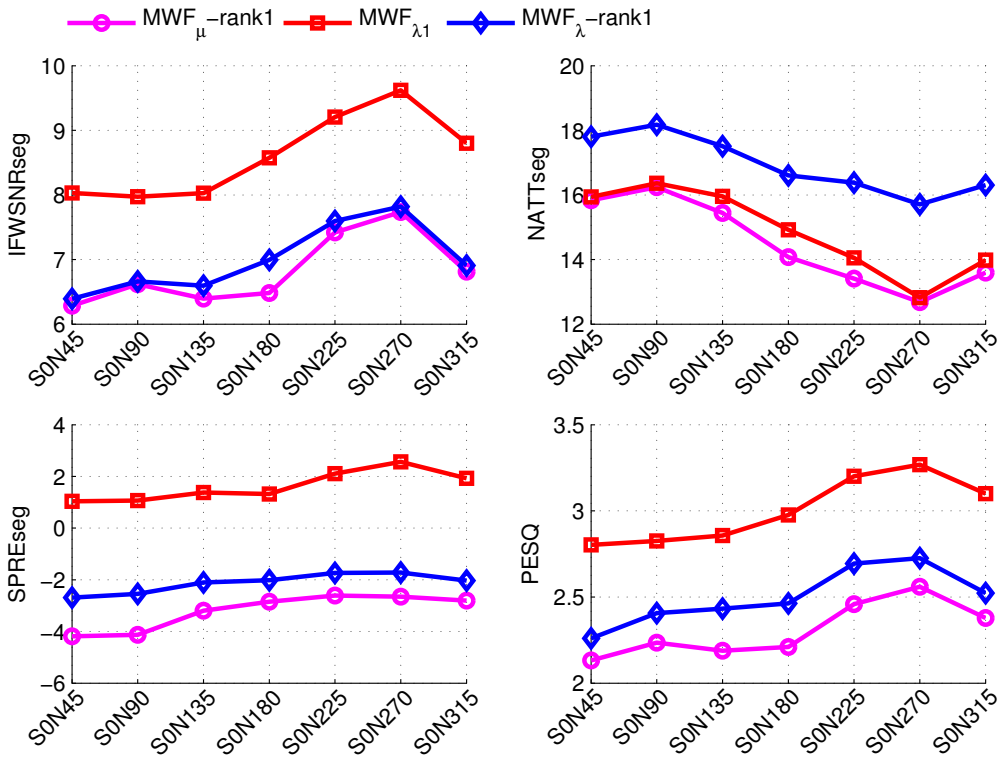

Figure A.16: Comparison between rank-one and general formulations for hammering noise for input SNR 10 dB.

References

[1] D. I. McBride, "Noise-induced hearing loss and hearing conservation in mining," *Occupational Medicine*, vol. 54, no. 5, pp. 290–296, 2004.

[2] H. Morris, "Work-related noise induced hearing loss in australia," Australia Safety and Compensation Council, Tech. Rep., 2006.

[3] E. R. Bauer and J. L. Kohler, "Cross-sectional survey of noise exposure in the mining industry," in *Proceedings of the 31st Annual Institute of Mining Health, Safety and Research. Blacksburg, VA: Virginia Polytechnic Institute and State University, Department of Mining and Minerals Engineering*, 2000, pp. 17–31.

[4] H. Puder, E. Fischer, and J. Hain, "Optimized directional processing in hearing aids with integrated spatial noise reduction," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC'12)*. Aachen, Germany: VDE, Sep. 2012, pp. 1–4.

[5] P. C. Yong, S. Nordholm, H. H. Dam, and S. Y. Low, "On the optimization of sigmoid function for speech enhancement," in *Proc. 19th European Signal Process. Conference (EUSIPCO'11)*, Barcelona, Spain, Aug. 2011, pp. 211–215.

[6] P. C. Yong, S. Nordholm, and H. H. Dam, "Trade-off evaluation for speech enhancement algorithms with respect to the a priori SNR estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'12)*. Kyoto, Japan: IEEE, Mar. 2012, pp. 4657–4660.

[7] ——, "Optimization and evaluation of sigmoid function with a priori

SNR estimate for real-time speech enhancement," *Speech Communication*, vol. 55, no. 2, pp. 358–376, Feb. 2013.

[8] ——, "Noise estimation with low complexity for speech enhancement," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*. New Paltz, USA: IEEE, Oct. 2011, pp. 109–112.

[9] ——, "Noise estimation based on soft decisions and conditional smoothing for speech enhancement," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC'12)*, Aachen, Germany, Sep. 2012, pp. 4640–4643.

[10] P. C. Yong, S. Nordholm, H. H. Dam, Y. H. Leung, and C. C. Lai, "Incorporating multi-channel Wiener filter with single-channel speech enhancement algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'13)*. Vancouver, Canada: IEEE, May 2013.

[11] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output. I. Fixed-processing systems," *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 6, pp. 529–542, 1997.

[12] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Process.*, vol. 2005, pp. 2915–2929, 2005.

[13] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids," *IEEE Trans. on Signal Process.*, vol. 53, no. 3, pp. 911–925, Mar. 2005.

[14] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. on Signal Process.*, vol. 55, no. 4, pp. 1579–1585, 2007.

[15] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter

based noise reduction in hearing aids," *EURASIP Journal on Advances in Signal Process.*, vol. 2009, no. 1, p. 930625, Jul. 2009.

[16] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 38–51, Jan. 2009.

[17] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 5, pp. 1368–1381, Jul. 2011.

[18] T. Barr, *Enquiry into the effects of loud sounds upon the hearing of boilermakers and others who work amid noisy surroundings.* Robert Anderson, 1886.

[19] P. Timmins, O. Granger, and S. W. Australia, *Occupational noise-induced hearing loss in Australia: overcoming barriers to effective noise control and hearing loss prevention.* Safe Work Australia, 2010.

[20] R. A. Dobie, "The burdens of age-related and occupational noise-induced hearing loss in the united states," *Ear and hearing*, vol. 29, no. 4, pp. 565–577, 2008.

[21] B. Shield, "Evaluation of the social and economic costs of hearing impairment," *Hear-it AISBL*, 2006.

[22] W. H. Organization *et al.*, "Report of the informal working group on prevention of deafness and hearing impairment programme planning," *Geneva: WHO*, pp. 18–21, 1991.

[23] ——, "Prevention of noise induced hearing loss. report of a WHO-PDH informal consultation geneva, 28–30 october 1997. no. 3 in the series. strategies for prevention of deafness and hearing impairment 1997," WHO-PDH-98.5, Tech. Rep.

[24] A. C. Davis, "The prevalence of hearing impairment and reported hearing disability among adults in great britain," *Int. Journal of Epidemiology*, vol. 18, no. 4, pp. 911–917, 1989.

[25] D. I. Nelson, R. Y. Nelson, M. Concha-Barrientos, and M. Fingerhut, "The global burden of occupational noise-induced hearing loss," *American journal of industrial medicine*, vol. 48, no. 6, pp. 446–458, 2005.

[26] G. Aniansson, K. Pettersson, and Y. Peterson, "Traffic noise annoyance and noise sensitivity in persons with normal and impaired hearing," *Journal of Sound and Vibration*, vol. 88, no. 1, pp. 85–97, 1983.

[27] J. G. Casali, "Passive augmentations in hearing protection technology circa 2010 including flat-attenuation, passive level-dependent, passive wave resonance, passive adjustable attenuation, and adjustable-fit devices: Review of design, testing, and research," *Int. Journal of Acoust. and Vibration*, vol. 15, no. 4, p. 187, 2010.

[28] C. Giguère, C. Laroche, A. Brammer, V. Vaillancourt, and G. Yu, "Advanced hearing protection and communication: Progress and challenges," in *Proceedings of the 10 th International Congress on Noise as a Public Health Problem, July*, vol. 24, 2011, pp. 225–33.

[29] J. G. Casali, W. A. Ahroon, J. A. Lancaster *et al.*, "A field investigation of hearing protection and hearing enhancement in one device: For soldiers whose ears and lives depend upon it," *Noise and Health*, vol. 11, no. 42, p. 69, 2009.

[30] S. M. Abel, S. Tsang, S. Boyne *et al.*, "Sound localization with communications headsets: Comparison of passive and active systems," *Noise and Health*, vol. 9, no. 37, p. 101, 2007.

[31] S. M. Abel, S. Boyne, H. Roesler-Mulroney *et al.*, "Sound localization with an army helmet worn in combination with an in-ear advanced communications system," *Noise and Health*, vol. 11, no. 45, p. 199, 2009.

[32] E. H. Berger, "Active noise reduction (ANR) in hearing protection: Does it make sense for industrial applications?" in *27th Conference of the National Hearing Conservation Association*, Dallas, USA, Feb. 2002.

[33] D. Gauger, "Should you consider ANR for hearing protection," *CAOHC Update*, vol. 15, no. 3, 2003.

[34] A. J. Brammer, C. Laroche *et al.*, "Noise and communication: A three-year update," *Noise and Health*, vol. 14, no. 61, p. 281, 2012.

[35] A. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, p. 1508, 1988.

[36] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Process. Magazine*, vol. 28, no. 1, pp. 33–42, 2011.

[37] D. R. Begault, "Auditory and non-auditory factors that potentially influence virtual acoustic imagery," in *Proc. 16th AES Int. Conf.: Spatial Sound Reproduction*. Rovaniemi, Finland: Audio Engineering Society, Apr. 1999.

[38] M. Bürger, "Binaural noise suppression techniques for assistive listening under extreme industrial noise conditions," Master's thesis, University of Erlangen-Nuremberg, 2012.

[39] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[40] L. Rayleigh, "XII. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[41] W. M. Hartmann, "How we localize sound," *Physics today*, vol. 52, p. 24, 1999.

[42] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[43] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, ser. Digital Signal Processing. Springer-Verlag, Berlin, 2001.

[44] H. Dillon, *Hearing aids*. Thieme, 2001.

[45] N. Grbic, *Optimal and adaptive subband beamforming*. Ph. D. dissertation, Blekinge Institute of Technology, 2001.

[46] T. A. Ricketts, "Directional hearing aids," *Trends in Amplification*, vol. 5, no. 4, pp. 139–176, 2001.

[47] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3, pp. 229–240, 1996.

[48] F. L. Luo, J. Yang, C. Pavlovic, and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Trans. on Signal Process.*, vol. 50, no. 7, pp. 1583–1590, 2002.

[49] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[50] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[51] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[52] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 4, pp. 487–503, 2005.

[53] S. Nordholm and I. Claesson, "Preventing target cancellation in adaptive broadband microphone arrays," *Recent Advances in Active Control of Sound and Vibration*, 1991.

[54] S. Nordholm, I. Claesson, and P. Eriksson, "The broad-band Wiener solution for Griffiths-Jim beamformers," *IEEE Trans. on Signal Process.*, vol. 40, no. 2, pp. 474–478, 1992.

[55] I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beaming," *IEEE Trans. on Antennas and Propagation*, vol. 40, no. 9, pp. 1093–1096, 1992.

[56] S. Nordebo, I. Claesson, and S. Nordholm, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE Journal of Oceanic Engineering*, vol. 19, no. 4, pp. 583–590, 1994.

[57] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on Array Processing and Sensor Networks*, pp. 269–302, 2008.

[58] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. on Vehicular Technology*, vol. 42, no. 4, pp. 514–518, 1993.

[59] N. Grbic and S. Nordholm, "Soft constrained subband beamforming for hands-free speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'02)*, vol. 1, Orlando, USA, May 2002, pp. I–885.

[60] S. Y. Low, N. Grbic, and S. Nordholm, "Speech enhancement using multiple soft constrained subband beamformers and non-coherent technique," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'03)*, vol. 5, Hong Kong, Apr. 2003, pp. V–489.

[61] J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *The Journal of the Acoustical Society of America*, vol. 91, p. 1662, 1992.

[62] S. Nordholm and I. Claesson, "Analytical evaluation of a self-calibrating microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal*

Process. (ICASSP'97), vol. 1. Munich, Germany: IEEE, Apr. 1997, pp. 243–246.

[63] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Trans. on Speech and Audio Process.*, vol. 7, no. 3, pp. 241–252, May 1999.

[64] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. on Signal Process.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[65] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7, pp. 636–656, 2007.

[66] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[67] P. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[68] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *American Institute of Physics Conference Series (AIP'86)*, vol. 151, Snowbird, USA, Apr. 1986, p. 206.

[69] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[70] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[71] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[72] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[73] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[74] D. W. Schobben and P. Sommen, "A frequency domain blind signal separation method based on decorrelation," *IEEE Trans. on Signal Process.*, vol. 50, no. 8, pp. 1855–1865, 2002.

[75] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 1, pp. 120–134, 2005.

[76] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE transactions on*, vol. 52, no. 7, pp. 1830–1847, 2004.

[77] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*. Springer, 2007, pp. 217–241.

[78] N. Grbic, X.-J. Tao, S. E. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 5, pp. 524–533, 2001.

[79] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 530–538, 2004.

[80] H. H. Dam, A. Cantoni, S. Nordholm, and K. L. Teo, "Second-order blind signal separation for convolutive mixtures using conjugate gradient," *IEEE Signal Process. Letters*, vol. 15, pp. 79–82, 2008.

[81] H. H. Dam, D. Rimantho, and S. Nordholm, "Second-order blind signal separation with optimal step size," *Speech Communication*, 2012.

[82] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, no. 6, pp. 1260–1277, 2006.

[83] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 539–548, 2004.

[84] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, p. 1777, 2007.

[85] ——, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[86] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[87] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Process.*, vol. 7, no. 2, pp. 126–137, 1999.

[88] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[89] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'79)*, vol. 4. Washington, USA: IEEE, Apr. 1979, pp. 208–211.

[90] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[91] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[92] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 412–424, 2006.

[93] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. on Speech and Audio Process.*, vol. 10, no. 5, pp. 245–256, 2002.

[94] M. J. Alam, D. O'Shaughnessy, and S.-A. Selouani, "Speech enhancement employing a sigmoid-type gain function with a modified a priori signal-to-noise ratio (SNR) estimator," in *Proc. IEEE Canadian Conf. on Electrical and Computer Engineering (CCECE'08).* Niagara Falls, Canada: IEEE, May 2008, pp. 000 631–000 636.

[95] Y. Hu, P. C. Loizou, N. Li, and K. Kasturi, "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. EL128–EL134, 2007.

[96] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[97] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications.* Technology Press of the Massachusetts Institute of Technology, 1950.

[98] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[99] R. Martin, "Speech enhancement based on minimum mean-square error

estimation and supergaussian priors," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[100] I. Andrianakis and P. White, "Speech spectral amplitude estimators using optimally shaped gamma and chi priors," *Speech Communication*, vol. 51, no. 1, pp. 1–14, 2009.

[101] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Process.*, vol. 2005, pp. 1110–1126, 2005.

[102] C. W. Therrien, *Discrete random signals and statistical signal processing.* Prentice Hall PTR, 1992.

[103] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013.

[104] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Process. Letters*, vol. 15, pp. 213–216, 2008.

[105] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 5, pp. 870–881, 2005.

[106] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'06)*, vol. 3. Toulouse, France: IEEE, May 2006, pp. III–III.

[107] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[108] R. Hendriks, J. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-square error amplitude estimators for speech enhancement under the generalized gamma distribution," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 2006.

[109] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Signal Process. Workshop on Statistical Signal Process. (SSP'01)*. Singapore: IEEE, Aug. 2001, pp. 496–499.

[110] J. S. Erkelens, J. Jensen, and R. Heusdens, "Improved speech spectral variance estimation under the generalized gamma distribution," in *Proc. 3rd Annual IEEE BENELUX/DSP Valley Signal Process. Symp. (SPS-DARTS'07)*, Metropolis, Belgium, Mar. 2007, pp. 43–46.

[111] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[112] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-gaussian densities," in *Proc. 10th Annual Conf. of the Int. Speech Communication Association (INTERSPEECH'09)*, Brighton, United Kingdom, Sep. 2009, pp. 1319–1322.

[113] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 4, pp. 475–486, 2005.

[114] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'08)*, Las Vegas, USA, Apr. 2008, pp. 4037–4040.

[115] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 2, pp. 277–289, 2011.

[116] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 5, pp. 857–869, 2005.

[117] E. Plourde and B. Champagne, "Perceptually based speech enhancement using the weighted β-SA estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'08)*. Las Vegas, USA: IEEE, Apr. 2008, pp. 4193–4196.

[118] R. C. Hendriks, J. S. Erkelens, and R. Heusdens, "Comparison of complex-dft estimators with and without the independence assumption of real and imaginary parts," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'08)*. Las Vegas, USA: IEEE, Apr. 2008, pp. 4033–4036.

[119] A. Papoulis and S. U. Pillai, "Probability, random variables and stochastic processes with errata sheet," *New York, NY, McGraw-Hill Education*, 2002.

[120] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.

[121] P. Vary *et al.*, "Noise suppression by spectral magnitude estimationmechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

[122] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[123] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 8, pp. 799–807, November 2001.

[124] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Adaptive β-order generalized spectral subtraction for speech enhancement," *Signal Processing*, vol. 88, no. 11, pp. 2764–2776, 2008.

[125] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'96)*, vol. 2, Atlanta, USA, May 1996, pp. 629–632.

[126] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech and Audio Process.*, vol. 2, no. 2, pp. 345–349, 1994.

[127] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, 2010.

[128] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, 2012.

[129] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Process. Letters*, vol. 11, no. 9, pp. 725–728, 2004.

[130] J. Chang, N. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. on Signal Process.*, vol. 54, no. 6, pp. 1965–1976, 2006.

[131] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2098–2108, 2006.

[132] Y. S. Park and J. H. Chang, "A novel approach to a robust a priori SNR estimator in speech enhancement," *IEICE Trans. on Communications*, vol. E90-B, no. 8, pp. 2182–2185, 2007.

[133] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'08)*, Las Vegas, USA, Apr. 2008, pp. 4897–4900.

[134] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 186–195, 2011.

[135] M. Alam, M. Chowdhury, and M. Alam, "Comparative study of a priori signal-to-noise ratio (SNR) estimation approaches for speech enhancement," *Journal of Electrical and Electronics Engineering*, vol. 9, no. 1, pp. 809–817, 2009.

[136] E. Plourde and B. Champagne, "Generalized bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Process. Letters*, vol. 16, no. 6, pp. 485–488, 2009.

[137] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'10)*. Dallas, USA: IEEE, Mar. 2010, pp. 4266–4269.

[138] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'01)*, vol. 2, Salt Lake City, USA, May 2001, pp. 749–752.

[139] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Process. (INTERSPEECH'08)*, Brisbane, Australia, Sep. 1998, pp. 2819–2822.

[140] A. Davis, S. Nordholm, S. Y. Low, and R. Togneri, "A multi-decision sub-band voice activity detector," in *Proc. 14th European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy, Sep. 2006.

[141] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC'08)*, Seattle, USA, Sep. 2008.

[142] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[143] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[144] K. W. Jang, D. K. Kim, and J.-H. Chang, "A uniformly most powerful test for statistical model-based voice activity detection," in *Proc. 8th Annual Conf. of the Int. Speech Communication Association (INTERSPEECH'07)*, Antwerp, Belgium, Aug. 2007.

[145] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

[146] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'98)*, vol. 1. Seattle, USA: IEEE, May 1998, pp. 365–368.

[147] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'09)*. Taipei, Taiwan: IEEE, Apr. 2009, pp. 4421–4424.

[148] J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'11)*. Prague, Czech: IEEE, May 2011, pp. 4640–4643.

[149] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383 – 1393, May 2012.

[150] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[151] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[152] H. Q. Dam, S. Y. Low, H. H. Dam, and S. Nordholm, "Space constrained beamforming with source PSD updates," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'04)*, vol. 4, Montreal, Canada, May 2004, pp. 93–96.

[153] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance." *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[154] A. S. of America, "ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index," Jun. 1997.

[155] R. C. Hendriks and R. Martin, "MAP estimators for speech enhancement under normal and rayleigh inverse gaussian distributions," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 918–927, Mar. 2007.

[156] V. Hamacher, "Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'02)*, vol. 4. Orlando, USA: IEEE, May 2002, pp. IV–4008.

[157] A. Kamkar-Parsi and M. Bouchard, "Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments," *IEEE Trans. on Instrumentation and Measurement*, vol. 60, no. 4, pp. 1141–1154, 2011.

[158] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Process.*, vol. 2006, pp. 175–175, 2006.

[159] K. Reindl, Y. Zheng, and W. Kellermann, "Analysis of two generic wiener filtering concepts for binaural speech enhancement in hearing aids," in *Proc. 18th European Signal Process. Conference (EUSIPCO'10)*, Aalborg, Denmark, Feb. 2010, pp. 988–993.

[160] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication*, vol. 39, no. 1, pp. 111–138, 2003.

[161] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann, "Multi-channel source separation preserving spatial information," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'07)*, vol. 1. Honolulu, USA: IEEE, Apr. 2007, pp. I–5.

[162] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system," *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 6, pp. 543–551, 1997.

[163] R. Nishimura, Y. Suzuki, and F. Asano, "A new adaptive binaural microphone array system using a weighted least squares algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'02)*, vol. 2. Orlando, USA: IEEE, May 2002, pp. II–1925.

[164] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME'11)*, Florence, Italy, Sep. 2011, pp. 41–46.

[165] M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary, "Robust dual-channel noise power spectral density estimation," in *Proc. 19th European*

Signal Process. Conference (EUSIPCO'11), Barcelona, Spain, Aug. 2011, pp. 2304–2308.

[166] H. Saruwatari, M. Go, R. Okamoto, and K. Shikano, "Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC'10)*, Tel Aviv, Israel, Aug. 2010.

[167] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 2006.

[168] F. Chen and P. Loizou, "Speech enhancement using a frequency-specific composite Wiener function," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'10)*. Dallas, USA: IEEE, Mar. 2010, pp. 4726–4729.

[169] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1614–1623, 2008.