

School of Electrical Engineering, Computing and
Mathematical Sciences

A Study into Speech Enhancement
Techniques in Adverse Environment

Lara Nahma

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

November 2018

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.



Lara Nahma
Faculty of Science and Engineering
Department of Electrical Engineering, Computing and Mathematical Sciences
Curtin University
November, 2018

Acknowledgements

I am thankful to many individuals for their help and guidance during my studies. My first and sincere appreciation goes to my supervisors Prof. Sven Nordholm and Dr. Hai Huyen Dam, for their continuous help and support in all stages of my Ph.D. study. Their guidance helped me in all the time of research and to write this thesis. I could not have imagined having better supervisors for my Ph.D study.

My greatest gratitude and acknowledgment go to my family, especially to my wonderful parents and parents in law who have given me tremendous encouragements and supports during these years. Their unending love and cheering have made the completion of this thesis a reality.

A lot of thanks also go to my friends for their support and encouragement to keep pushing me on my PhD studies. Further, many thanks to my colleagues for their wonderful times we shared, they have helped me in one or another way during my PhD, and I would like to thank them all. I also would like to thank Dr. Manora Caldera for her enormous effort in proof-reading and polish this thesis.

Last but definitely not least, I would like to thank my husband, Alan Mayahi, for all his love and tremendous support. He has always continued his encouragement during this journey. My special thanks go to my daughter Sally, for her love, patience and understanding during the course of this research work.

Lara Nahma

November, 2018

List of Publications

Much of the work in this thesis has been published. These papers are:

1. L. Khalid, S. Nordholm, and H. H. Dam, Design study on microphone arrays, IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015, pp. 1171-1175.

In this paper, I planned and carried out the matlab code for the design and simulations. I interpreted the results and wrote the manuscript. All the authors discussed the results and approved the final manuscript.

2. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Convex combination framework for a priori snr estimation in speech enhancement, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), New Orleans, USA, 2017, pp. 4975-4979.

My main contribution in this paper is carried out the main conception, design, evaluation and interpretation of results. I wrote the manuscript. All the authors discussed the results and approved the final manuscript.

3. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Improved a priori snr estimation in speech enhancement, 23rd Asia-Pacific Conference on Communications (APCC), Perth, Australia, 2017, pp. 15.

In this paper, I carried out the design problem, performed the calculation and wrote the matlab codes. Moreover, I wrote the manuscript. All the authors discussed the results and approved the final manuscript.

4. L. Nahma, H. H. Dam, S. Nordholm, "Robust beamformer design against mismatch in microphone characteristics and acoustic environments," International Workshop on Acoustic Echo and Noise Control (IWAENC), Tokyo, Japan, 2018, pp. 76-80.

In this paper, I developed the theoretical formulation, performed the numerical simulation and interpreted the results. I wrote the manuscript. All the authors discussed the results and approved the final manuscript.

5. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Cross evaluation of speech enhancement methods under different noise conditions, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Brighton, UK. 2019, pp.895-899.

In this paper, I carried out the study of conception, performed the experiments and interpreted the results. Furthermore, I wrote the manuscript. All the authors discussed the results and approved the final manuscript.

6. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, An adaptive a priori SNR estimator for perceptual speech enhancement, EURASIP journal on Audio, Speech and Music Processing, 2019, (1), p.7.

In this paper, I designed the proposed technique, the computational framework and analysed the data. Moreover, I planned and carried out the simulations and wrote the paper. All the authors discussed the results and approved the final manuscript.

7. L. Nahma, H. H. Dam, Cedric Ka Fai Yiu, and S. Nordholm, Robust Broadband Beamformer Design for Noise Reduction and Dereverberation, Multi-dimensional Systems and Signal Processing (MSSP), 2019, pp.1-21.

In this paper, I developed the theoretical formulation, performed the computation and the numerical simulations. Moreover, I carried out the experiments. In addition, I interpreted the results and wrote the paper. All the authors discussed the results and approved the final manuscript.

To Whom It May Concern,

I, Lara Nahma, contributed to the above listed publications as indicated therein.



Lara Nahma



Professor Sven Nordholm

June. 2019

June, 2019

Abstract

Speech enhancement algorithms are in high demand in practice for scenarios subject to noisy environments, in which the speech signal is degraded by different disturbances and noise such as room reverberation and background noise. This degradation can impact the quality and intelligibility of the speech and decrease the performance of other signal processing systems like automatic speech recognition and speech coding. The main goal of speech enhancement algorithms is to reduce or suppress the noise from the degraded speech signal while preserving the original speech components. The work presented in this dissertation focuses on providing speech enhancement solutions for voice communication applications such as hands-free mobile phones, hearing aids and teleconferencing systems by using different methods. In particular, two speech enhancement techniques are explored, namely single channel and multi-channel speech enhancement techniques.

For single channel speech enhancement techniques, the main purpose is to suppress the background noise and musical noise without distorting the speech components. The state of the art decision directed (DD) based a priori SNR approach is the most used approach due to its efficiency in reducing the musical noise. However, because of its slow adaptation towards abrupt changes in SNR, DD approach prone to distort the speech components. In order to overcome this problem, an a priori SNR estimation technique is introduced. Besides its ability to eliminate the frame delay generated by the DD approach, the proposed technique also increases the adaptation speed during abrupt changes in the SNR estimation. Results show that the proposed a priori SNR estimation approach outperforms the conventional approach in preserving weak speech components. In conjunction with that, we utilize a Critical Band (CB) mapping from Short Time Fourier Transform (STFT) analysis-resynthesis system in the speech enhancement framework for human perceptual processing and lower complexity.

Since most of the speech enhancement techniques are performed in STFT, this dissertation also studies the aforementioned proposed approach in STFT

domain and under different noise conditions. Furthermore, it presents a cross comparison between a priori SNR estimation methods integrated with different time-frequency analysis techniques (CB and STFT) using subjective listening test in order to test the efficiency of CB over STFT. The subjective results show that the listeners preferred the speech signals that were processed with CB over those processed with STFT in non-stationary noise conditions.

This dissertation also discusses the combination of noise reduction and dereverberation in indoor applications such as conference rooms. Since the single channel speech enhancement system exploits the temporal and spectral diversity of the microphone received signal, such method has a limited ability to suppress the reverberation. Thus, in this work, we employ a multi-channels speech enhancement method to exploit the spatial diversity that induces by the reverberation. A robust broadband beamformer is formulated by including the room reverberation as well as microphones characteristics (gain and phase) in the design procedure and by using mean performance optimization technique. Two different error models (multiplicative and additive) models are considered in this work. Compared to the non-robust beamformer design, robust indoor beamformer design shows an improved performance in terms of less sensitivity against mismatches in microphone characteristics (gain and phase). An extensive evaluation using simulated and measured Room Impulse Response (RIR) is presented in acoustically adverse environments to examine the efficiency of the investigated designs. Results demonstrate that robust direct design using additive error model can achieve almost the same results as including room response in the design in many reverberation environments. Furthermore, it provides robustness over larger variations in the reverberation environment. This means that the robust direct path based method which is based on mean variations in gain and phase can be used in low to medium ($T_{60} = 100 - 300$ ms) reverberant environments with good result.

Contents

Acknowledgments	v
Abstract	ix
1 Introduction	3
1.1 Objective	4
1.2 Contributions of the thesis	5
1.3 Thesis outlines	6
2 Literature Survey	9
2.1 Introduction	10
2.2 Applications of Speech enhancement technology	13
2.2.1 Aviation and military applications	14
2.2.2 Biomedical applications	15
2.2.3 Commercial applications	16
2.3 Speech Enhancement Techniques	17
2.3.1 Single channel speech enhancement techniques	17
2.3.2 Multi-channel speech enhancement technique	22
2.3.3 Beamforming	23
2.3.4 Beamforming classifications	25
2.3.5 Sensitivity of beamformer design	27
2.3.6 Dereverberation	29
2.4 Chapter Summary	31

3	Single Channel Speech Enhancement	33
3.1	Introduction	34
3.2	Critical band speech enhancement	37
3.3	Conventional a priori SNR estimation	40
3.4	Proposed a priori SNR estimation	41
3.5	Objective and subjective quality measurements	47
3.6	Experimental results and discussion	50
3.6.1	Experimental setup	50
3.6.2	Evaluation the effect of the bark scale frequency resolution on the noise characteristics	51
3.7	Summary	60
4	Single Channel Speech Enhancement in STFT Domain	71
4.1	Introduction	72
4.2	Single channel speech enhancement	74
4.3	Proposed a priori SNR estimation	75
4.4	Cross subjective evaluation	77
4.5	Experimental evaluation	79
4.5.1	Evaluation of a priori SNR estimation	79
4.5.2	Objective results	80
4.5.3	Spectrograms	81
4.5.4	Evaluation of listening test	81
4.5.5	Statistical analysis	84
4.6	Summary	86
5	Robust Broadband Beamformer in Reverberant Environment	101
5.1	Introduction	102
5.2	Problem formulation	104
5.3	Indoor Broadband Beamformer Design	106
5.4	Robust Beamformer Design	107
5.4.1	Additive error model	107
5.4.2	Multiplicative error model	110

5.5	Aperture Size Optimization	112
5.6	Objective Measurements	112
5.7	Design Examples	113
5.7.1	Overall performance and cost function evaluation for dif- ferent reverberation time	114
5.7.2	Dereverberation performance	117
5.7.3	Suppression performance in stopband region	118
5.7.4	Joint dereverberation and noise suppression performance .	119
5.7.5	Sensitivity test of beamformer designs	122
5.7.6	Evaluation of calculation time for different beamformer de- signs	124
5.7.7	Results of aperture size optimization	125
5.8	Summary	126
6	Conclusions and Future work	129
6.1	Summary	130
6.2	Future work	131
6.2.1	Single channel speech enhancement techniques	132
6.2.2	Multi-channel speech enhancement techniques	132

List of Figures

2.1	Speech degradation scenario in noisy environment.	12
2.2	Spectrograms of clean speech signal and noisy speech signal consisting of clean speech and pink noise at 0 dB SNR.	13
2.3	Framework of single channel speech enhancement.	18
2.4	Framework of multiple channel speech enhancement.	23
2.5	Beam-pattern comparison between narrowband beamformer and broadband beamformer for a linear microphone array.	24
2.6	Structure of fixed beamformer.	26
2.7	Structure of Generalized Sidelobe Canceler as an example of Adaptive beamformer.	28
2.8	Simulated room impulse response using Image Source Method (ISM) at reverberation time $T_{60} = 0.3 s$	29
3.1	Block diagram for the critical band processing.	37
3.2	Block diagram of the spectral gain function computation using the proposed a priori SNR estimation method.	43
3.3	Histogram of a posteriori SNR estimate for different background noise (1 st row) for pink noise, (2 nd row) for white noise at 9 th critical band mapped with adaptive smoothing factor calculated with different sets of parameters (adaptive smoothing factor calculated at (i) $\gamma_u=5$ dB, (ii) $\gamma_u=7$ dB, (iii) $\gamma_u=9$ dB and (iv) $\gamma_u=15$ dB). Left figure for noise period only and right figure for speech and noise period.	45

3.4	Histogram of a posteriori SNR estimate for different background noise (1 st row) for factory noise and (2 nd row) for babble noise at 9 th critical band mapped with adaptive smoothing factor calculated with different sets of parameters (adaptive smoothing factor calculated at (i) $\gamma_u=5$ dB, (ii) $\gamma_u=7$ dB, (iii) $\gamma_u=9$ dB and (iv) $\gamma_u=15$ dB). Left figure for noise period only and right figure for speech and noise period.	46
3.5	Evaluation of bark scale based frequency analysis at 6 th critical band under different background noise: 1 st row for pink noise and 2 nd row for white noise.	53
3.6	Evaluation of bark scale based frequency analysis at 6 th critical band under different background noise: 1 st row for factory noise and 2 nd row for babble noise.	54
3.7	Comparison of the a priori SNR estimation over a short time period between the true a priori SNR ξ (black solid line with a marker), ML a priori SNR estimate (green dashed line), $\hat{\xi}_{DD}$ (blue solid line), $\hat{\xi}_{MDD}$ (cyan dot solid line), and $\hat{\xi}_{prop}$ (red solid line with a marker), at 9 th critical band and 10 dB SNR under different background noise: 1 st row for pink noise and 2 nd row for white noise.	56
3.8	Comparison of the a priori SNR estimation over a short time period between the true a priori SNR ξ (black solid line with a marker), ML a priori SNR estimate (green dashed line), $\hat{\xi}_{DD}$ (blue solid line), $\hat{\xi}_{MDD}$ (cyan dot solid line), and $\hat{\xi}_{prop}$ (red solid line with a marker), at 9 th critical band and 10 dB SNR under different background noise: 1 st row for factory noise and 2 nd row for babble noise.	57
3.9	Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by Wiener filter speech estimation technique.	62
3.10	Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by Wiener filter speech estimation technique.	63
3.11	Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by Wiener filter speech estimation technique.	64
3.12	Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by Wiener filter speech estimation technique.	65
3.13	Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by LSA speech estimation technique.	66

3.14	Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by LSA speech estimation technique.	67
3.15	Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by LSA speech estimation technique.	68
3.16	Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by LSA speech estimation technique.	69
4.1	Adaptive mean factor as a function of frequency bins.	76
4.2	Adaptive smoothing factor as a function of instantaneous SNR at different frequency bins.	77
4.3	Comparison of a priori SNR estimation over short time between Υ (green solid line) , $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB pink background noise.	81
4.4	Comparison of a priori SNR estimation over short time between Υ (green solid line) , $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB white background noise.	83
4.5	Comparison of a priori SNR estimation over short time between Υ (green solid line) , $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB factory background noise.	84
4.6	Comparison of a priori SNR estimation over short time between Υ (green solid line) , $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB babble background noise.	85
4.7	Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by Wiener filter speech estimation technique. . .	89
4.8	Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by Wiener filter speech estimation technique. .	90
4.9	Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by Wiener filter speech estimation technique. .	91
4.10	Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by Wiener filter speech estimation technique. .	92
4.11	Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by LSA speech estimation technique.	93
4.12	Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by LSA speech estimation technique.	94

4.13	Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by LSA speech estimation technique.	95
4.14	Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by LSA speech estimation technique.	96
4.15	Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with WF gain function and evaluated in pink background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.	97
4.16	Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with WF gain function and evaluated in babble background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.	97
4.17	Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with MMSE-LSA gain function and evaluated in pink background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.	98
4.18	Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with MMSE-LSA gain function and evaluated in babble background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.	98
5.1	Block diagram of the broadband beamformer.	105
5.2	Magnitude response of direct beamformer and robust direct beamformer using additive model under different reverberation time. .	115
5.3	Magnitude response of indoor beamformer and robust indoor beamformer using additive model under different reverberation time. .	116
5.4	Direct to reverberant ration performance under (a) different source-microphone array distance, (b) different number of microphones. Robust beamformers designed using additive error model.	118
5.5	Room setup with a linear microphone array.	120

5.6	SSNRR results obtained for different beamformer designs for varying SJR's and reverberation times. The reverberant signals were generated using simulated RIR (left side) and measured RIR (right side) with different reverberation time, $T_{60} = 0.16 s$ (top) and $T_{60} = 0.36 s$ (bottom). Robust beamformers designed using additive error model.	121
5.7	Histogram of cost function values distribution direct path based beamformer and its robust design using multiplicative error model ($T_{60} = 0 s$), indoor beamformer and its robust design using multiplicative error model ($T_{60} = 0.2 s$).	123
5.8	Histogram of cost function values distribution direct path based beamformer and its robust design using additive error model ($T_{60} = 0 s$), indoor beamformer and its robust design using additive error model ($T_{60} = 0.2 s$).	124
5.9	Cost function comparison for different inter-element spacing for direct path based beamformer and its robust design using multiplicative error model ($T_{60} = 0 s$), indoor beamformer and its robust design using multiplicative error model ($T_{60} = 0.2 s$).	126

List of Tables

3.1	Scale description of the listening test criteria	50
3.2	Noise variance comparison before and after frequency analysis and for different noise types.	52
3.3	Mean objective results for pink noise.	55
3.4	Mean objective results for white noise.	55
3.5	Mean objective results for factory noise.	55
3.6	Mean objective results for babble noise.	58
3.7	Listening test results for pink noise at 10 dB input SNR.	59
3.8	Listening test results for babble noise at 10 dB input SNR.	60
4.1	Mean objective results for pink noise.	82
4.2	Mean objective results for white noise.	82
4.3	Mean objective results for factory noise.	82
4.4	Mean objective results for babble noise.	82
4.5	One way ANOVA test results to verify the statistically significant difference between different frequency warping scales used in the listening test under different noise conditions.	87
4.6	Tukey's HSD Comparison between the enhanced speech signal using WF gain function and the unprocessed speech signal under different conditions.	88
4.7	Tukey's HSD Comparison between the enhanced speech signal using MMSE-LSA gain function and the unprocessed speech signal under different conditions.	99
5.1	Comparison of the cost function for different reverberation times for the direct design and the robust direct design ($T_{60} = 0 s$) using additive model, the indoor design and the indoor robust design ($T_{60} = 0.2 s$) using additive model.	117

5.2	Comparison results among direct path based beamformer and its robust design ($T_{60} = 0 s$) using additive model, indoor beamformer and its robust design ($T_{60} = 0.2 s$) using additive model on the interference suppression at different reverberation time.	119
5.3	Comparison of condition number of correlation matrix among direct path based beamformer and its robust design ($T_{60} = 0 s$), indoor beamformer and its robust design ($T_{60} = 0.2 s$).	123
5.4	Calculation time for different beamformer designs.	125
5.5	Array aperture size optimization for different beamformer designs.	126
5.6	Parameters for the evaluation of the microphone array.	127

List of Acronyms

SNR	Signal to Noise Ratio
CB	Critical Band
STFT	Short-Time Fourier Transform
ATC	Air Traffic Controller
ASR	Automatic Speech Recognition
SEU	Speech Enhancement Unit
HPD	Hearing Protection Devices
WHO	World Health Organizer
HAD	Hearing Aid Devices
BTE	Behind The Ear
VCD	Voice Controlled Devices
SS	Spectral Subtraction
IFFT	Inverse Fast Fourier Transform
FFT	Fast Fourier Transform
ERB	Equivelant Rectangular Bandwidth
WPT	Wavelet Packet Transform
PSD	Power Spectral Density
VAD	Voice Activity Detection
DFT	Discrete Fourier Transform
MS	Minimum Statistics
PESQ	Perceptual Evaluation Speech Quality
WF	Wiener Filter
MMSE	Minimum Mean Square Error
ROI	Region Of Intrest
DAS	Delay And Sum
FIR	Finite ImpulseResponse
LS	Least Square
WLS	Weighted Least Square
LCMV	Linearly Constrained Minimum Variance

GSC	Generalized Sidelobe Canceler
NLMS	Normalized Least Mean Square
WNG	White Noise Gain
LPC	Linear Predictive Coding
RIR	Room Impulse Response
ISM	Image Source Method
DRR	Direct to Reverberant Ratio

Chapter 1

Introduction

The work presented in this thesis is motivated by the fast expanding market of voice communication applications such as hearing aids, voice controlled devices (VCDs) and teleconferencing systems. In many countries, using a mobile phone while driving is illegal since it is considered fatal because of its potential for causing distracted driving and accidents. In Australia, all states have banned the use of a mobile phone while driving. That's why it's common now for cars to offer VCDs integration. The main motivation of using such applications in cars is to provide more convenience and safety where the driver can use the phone to initiate phone calls or use the GPS while still keeping both hands on the wheel. Automatic Speech Recognition (ASR) is the main signal processing technique that is used to achieve the main task of voice commands. In such applications, since the microphone is located at a distance from the speaker, this means that the received speech signal is corrupted by different sources of noise such as background noise, reverberation and other interference which provides substantial speech distortion and poor speech quality. Thus, speech enhancement procedures are required to attenuate the background noise while maintain the speech contents in order to improve the human-machine interfacing.

Speech enhancement has long been an attractive solution to problems of VCDs. The main objective of speech enhancement is to extract the desired speech signal while suppressing the background noise in order to improve the speech perception by human or machines. Many studies focused mainly on single channel speech enhancement approaches due to their simple implementation and efficient noise reduction ability. However, the drawback of single channel techniques is the limited capability to suppress the background noise without attenuating the desired speech especially in adverse environments. Apart from the speech distortion and residual noise, another drawback of many single channel speech enhancement techniques is the appearance of unnatural noise artifact known as musical noise which is unpleasant to the listeners. Nevertheless, single channel speech enhancement techniques are still preferable in many applications such as hearing aids and mobile phones and hence, motivates us to propose a novel single channel speech enhancement technique that can improve the trade off between the speech transient distortion and musical noise in non-stationary environments.

In addition to noise reduction, another challenge in indoor applications is the presence of room reverberation. A reverberation scenario in closed spaces has to be considered through speech processing and then has to be suppressed. Since the room reverberation includes the spatial diversity, dereverberation with a microphone array has to be observed in particular due to the ability of beamforming to provide spatial selectivity through speech enhancement capabilities. Many research works have focused on the dereverberation and noise reduction combination using multi-channels speech enhancement techniques. It has to be remarked that beamformer design is usually sensitive to mismatches in microphone characteristics such as gain, phase and element position. Hence it is important to consider those deviations in the beamformer design formulation. To the best of our knowledge no research work has been presented to include reverberation and robustness in the design formulation which motivates us to present a simple and yet efficient robust beamformer design in reverberant environment and against mismatches in microphones characteristics.

1.1 Objective

Since the presence of background noise and reverberation can significantly deteriorate the overall performance of the speech enhancement techniques in speech communication devices, we focus in this work on noise reduction and dereverberation. The objective of this dissertation is twofold; first, to propose an improved

single channel speech enhancement solution for voice communication devices in adverse environments in order to improve the speech quality and suppress the background noise. The proposed technique should be able to control the trade-off between the speech transient distortion and the musical noise. Second, to investigate and design a robust broadband beamformer for indoor applications. The proposed design should be able to pick up the signals that are originating from the region of interest while suppressing reverberation and interference that are originating from undesirable directions. Namely, we focused on the following aspects:

- To control the adaptation speed of the a priori SNR estimation in order to preserve more weak speech components while maintaining the advantage of decision directed based a priori SNR estimation methods in reducing the musical noise.
- Perform a combination of noise reduction and dereverberation by introducing a robust broadband beamformer design formulation that includes the room reverberation as well as robustness to mismatches in microphone characteristics (gain and phase).

1.2 Contributions of the thesis

In this dissertation, the original contributions are divided into two parts, single channel and multi-channels speech enhancement techniques to provide better speech quality and less noise. This section highlights the main contributions of this thesis as follows:

- It utilizes a critical band mapping from STFT analysis-resynthesis system in the speech enhancement framework for human perceptual processing and lower complexity.
- It emphasizes the efficiency of the critical band processing method in reducing the musical noise due to its ability to lower the noise variance.
- It proposes an improved a priori SNR estimation approach by utilizing a fusion function based on a sigmoidal shape in order to control the adaptation speed of the a priori SNR estimation.
- It introduces a new evaluation technique which is called the modified Hamming distance to measure the weak speech components.

- It presents a cross comparison between STFT and CB spectral analysis methods by using a subjective listening test.
- It proposes a beamformer design formulation as a minimization of cost function with respect to filter coefficients and inter-element space between adjacent microphones.
- It introduces robust design methods for broadband beamformers in reverberant environments. In the design formulation room reverberation as well as robustness to amplitude and phase mismatches in the microphones has been included.
- It provides an extensive evaluation to assess the efficiency of the robust direct beamformer design in reverberant environments.

1.3 Thesis outlines

The remainder of this thesis is structured as follow.

- Chapter 2: “*Literature Survey*” gives a comprehensive overview of the speech enhancement system and its applications. Preliminary studies on single channel and multi-channels speech enhancement techniques are presented. These include discussions on the framework of single channel speech enhancement techniques. Moreover, background information related to the different beamformer designs are presented, i.e. narrowband , broadband , fixed and adaptive beamformers. In addition, outlines the important design considerations that have to be included in the beamformer problem formulation which can affect the overall performance.
- Chapter 3: “*Single Channel Speech Enhancement* ” proposes a modified a priori SNR estimator by employing a model of speech absence probability based on a sigmoid function to improve the adaptation speed of the a priori SNR estimation and hence yields in preserving weak speech components. Moreover, a critical band mapping for STFT analysis-synthesizes system is used in the speech enhancement framework to reduce the musical noise and computational complexity.
- Chapter 4: “*Single Channel Speech Enhancement in STFT domain*” examines the efficiency of the proposed a priori SNR estimation method in

Chapter 3 but in STFT domain and under different noise conditions. Moreover, a cross comparison between STFT and CB spectral analysis methods has been presented using a subjective listening test.

- Chapter 5: “*Robust Broadband Beamformer in reverberant environment*” discusses several broadband beamformer designs for combined dereverberation and noise reduction. Also it includes stochastic error models that are representing mismatches in microphone characteristics (gain and phase). The robust design procedure is formulated using mean performance optimization method. Two different stochastic error models are presented in this chapter; multiplicative error and additive error. These design methods achieved robust beamformer design against deviations in acoustically adverse environments and mismatches in microphone characteristics (gain and phase).
- Chapter 6: “*Conclusion*” concludes the findings and outlines the suggestions for further research.

Chapter 2

Literature Survey

This chapter gives a comprehensive overview of the speech enhancement, and signifies its necessity in numerous speech communication applications. Based on the number of microphones, we outlined the main concepts of single channel and multiple channels solutions to address the noise reduction and dereverberation under different adverse environments.

2.1 Introduction

For many speech communication applications, such as hands free communications, hearing aids and teleconferencing systems, speech quality and intelligibility have a direct effect on the ease and accuracy of information exchange. In acoustically adverse environments, due to the presence of different sources of disturbances such as background noise, reverberation or interference, the desired speech signal captured by the microphones is contaminated as shown in Figure 2.1, where a person (desired source) is talking in a busy conference room with different sources of noise. Consequently, this leads to a significant degradation in the quality of the picked up speech. Figure 2.2 shows the corrupted speech signal received by the microphone array (bottom figure) and the clean speech signal (top figure). It can be clearly noticed how the corrupted speech signal is different from the clean speech, in which great portions of the speech spectra are masked and less distinct. This can significantly impair the speech quality and intelligibility. Moreover, it can increase listener's fatigue and lower information exchange ability. Therefore, speech enhancement systems are useful to clean up the desired speech signal and mitigate the effect of the corruption in order to improve the performance of the aforementioned applications [1].

Noise suppression or speech enhancement has attracted considerable research effort in the last decades due to its uses in widely spread devices like for instance, mobile phones, hearing aids, assistive listening devices and voice communication. Particularly, hearable devices have been poised to assist people with hearing difficulties in social environments. For noise suppression and speech enhancement to work in those environments where acoustic noise becomes more intrusive, it is important to preserve weak speech components while still balance the amount of noise reduction. Accordingly, techniques that can enhance speech signals while preserving weak speech components under a large variety of acoustic scenarios are underpinning the success of different applications such as speech coding, speech recognition and hand free telephony [2, 3, 4, 5].

The type of speech enhancement or noise reduction algorithm to be selected depends on many factors that need to be considered such as, the noise type and its characteristics, applications at hand and the number of microphones available. These factors have significant impact on the quality of the estimated speech and the overall performance of the speech processing systems.

Generally speaking, it is important to consider the characteristics of the noise to achieve a speech enhancement/noise reduction method that works effectively

in different conditions. Different types of noise based on characteristics of the noise source, such as stationary noise (pink noise) which statistically does not change over time, or non stationary noise with high variability and similar characteristics as a speech signal (babble noise) [6]. Moreover, based on the way the noise contaminates the speech signal, the noise can be categorized into additive noise originating from different noise sources such as fans, air-conditions, traffic, babble etc, and a noise caused by multi path propagation due to the room acoustics (reverberation). Hence, according to the type of noise, different signal processing techniques can be implemented, such as noise reduction and speech dereverberation or a combination of both [7].

The goal of the speech enhancement algorithm is to remove noise and recover the original signal with as little distortion and residual noise as possible. The procedure of denoising is being complicated by the fact that, no transform domain, e.g. time, frequency or others, exists where signal and noise have non-overlapping supports; hence aggressive removal of noise is always accompanied by signal distortion and efforts to reduce distortion are in conflict with the amount of reduced noise [8].

For a computationally efficient implementation, most of the speech enhancement techniques are utilizing the short time Fourier transform (STFT), where the desired signal is estimated from the degraded speech by applying noise reduction algorithm to the complex STFT coefficients [3, 5, 9, 10]. The main advantage of using STFT is the flexibility in exploiting the noise statistics to optimize the noise reduction performance since this type of decomposition helps to handle different frequencies independently. However, deploying STFT results in uniform resolution for the whole band of frequencies, which is not the case for the human auditory system with a non uniform resolution. This fact has motivated many researchers to propose alternative speech enhancement methods based on the human auditory system in order to improve the speech quality and reduce the annoying residual musical artifacts that known as musical noise [11, 12, 13, 14].

Human auditory spectrum model consists of a bank of bandpass filters which follows a spectral bark scale or so-called critical bands [14, 15]. In [14], a standard subtractive speech enhancement method is presented to eliminate the musical artifacts in very noisy situations. The masking properties of the auditory system are utilized to compute the subtraction parameter. In [16], a spectral subtraction noise reduction method was proposed using a spatial weighting technique based on the inhibitory property of the auditory system, which results in improving the estimated speech while reducing the musical noise.

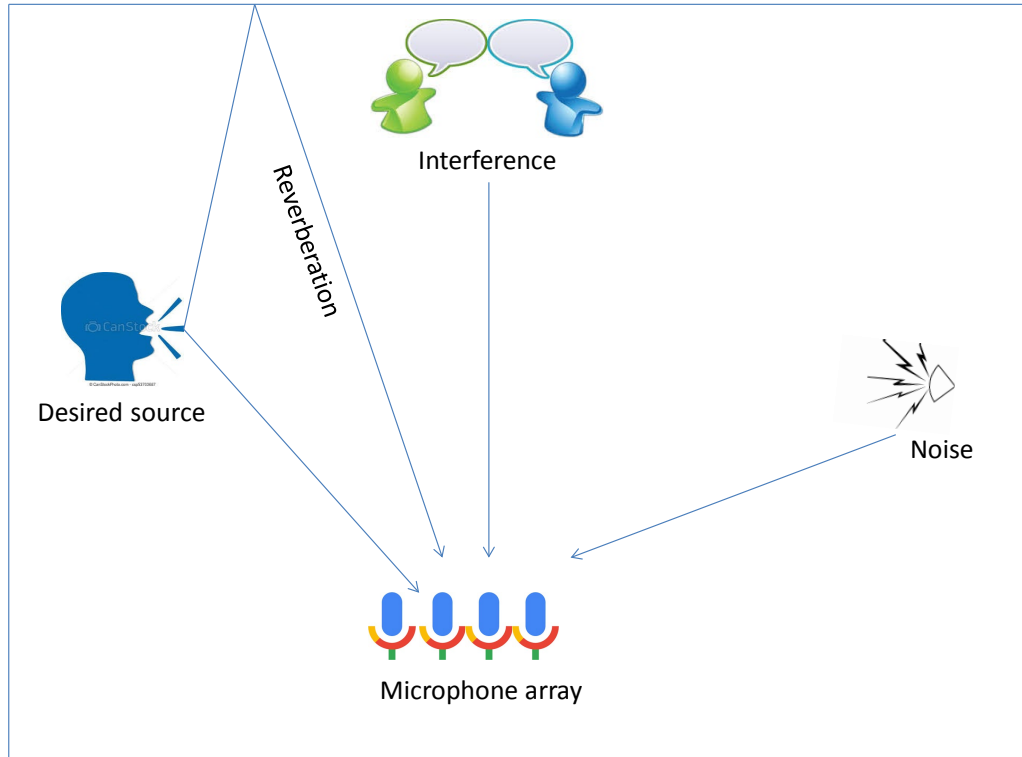


Figure 2.1: Speech degradation scenario in noisy environment.

Besides noise reduction algorithm, dereverberation algorithm in adverse environments has been discussed in this thesis. In such environment, the microphone picked up the direct path signal as well as a multiple attenuated and delayed replicas of the signal due to the room acoustics. This results in a severe degradation in the observed signal. As such for more than four decades, researchers were trying to improve the performance of speech applications in adverse environments, and recent research is interested in a robust combination of noise reduction and dereverberation technique under different noise conditions.

This chapter presents the main problem of speech communication in noisy environments and reviews recent speech enhancement techniques depending on the number of microphones available. First, some applications where speech enhancement is used are presented. Consequently, different speech enhancement techniques for additive noise are described for both single microphone and multiple microphone techniques. Thereafter, convolutive distortion is discussed and dereverberation processing is described using some recent efficient methods.

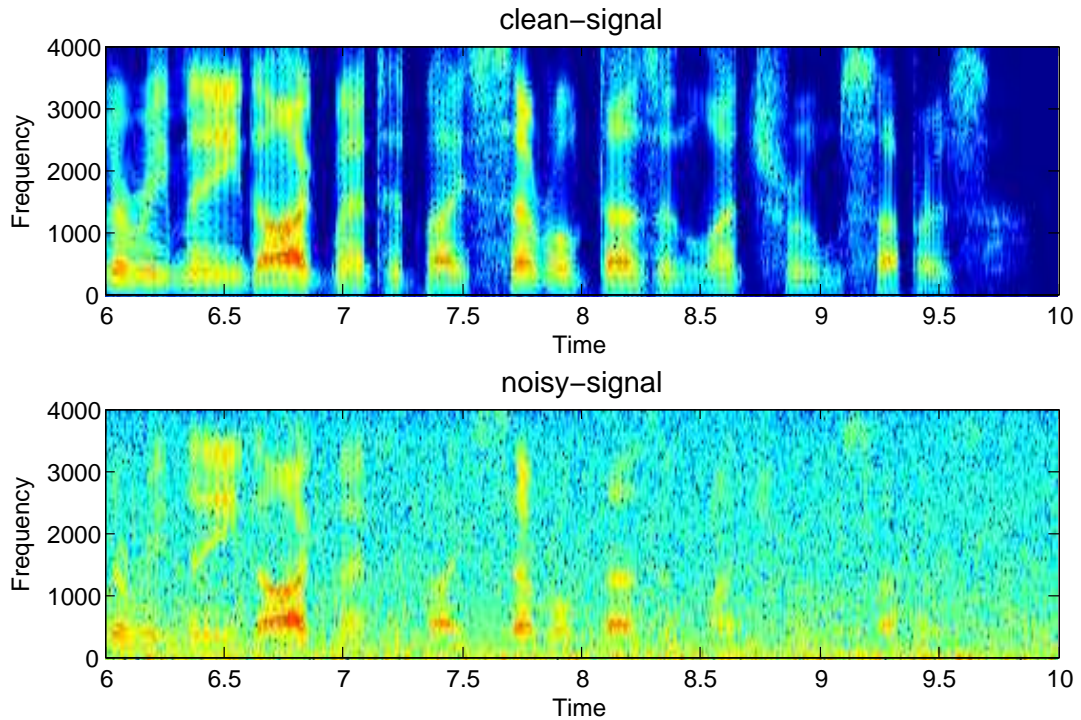


Figure 2.2: Spectrograms of clean speech signal and noisy speech signal consisting of clean speech and pink noise at 0 dB SNR.

2.2 Applications of Speech enhancement technology

Prior to design speech enhancement algorithm, it is important to consider the target application. What the industry is looking for is a way to provide a clean voice signal (free of all ambient noise) to the desired application such as Automatic Speech Recognition (ASR) engine. The key concept here is extracting a clean voice signal in very noisy environments while mitigating the impact of unwanted signals such as background noise, echo, other speech sources and reverberation by using speech enhancement techniques [17, 18]. However, different speech communication applications have different preferences and approaches that need to be tailored to the specified application [19]. In many applications, distance from the speaker, noise characteristics and levels are important properties to consider not only in single channel techniques but also in multi-channel techniques. In this section, three types of voice communication systems will be discussed.

2.2.1 Aviation and military applications

Speech enhancement algorithms play an important role for effective aviation and military operations. They can be used either as a stand alone component to improve the speech quality or as a pre-processor in a larger speech communication system to enhance the input signal prior to further processing. Many research works have been presented in the literature that develop speech enhancement technologies. These technologies have been employed in different military and civic applications, such as for communication between pilot and civil air traffic controller (ATC) systems [20] or advanced air traffic control training system in military applications [21].

Air travel is an important part of our lives. Ensuring the highest level of passenger safety is a key goal for ATC authorities. The ability to capture the information from the verbal messages which are used to predict the current state of the airspace, or providing an early warning to avoid hazard helps to ensure passenger safety [22]. The main problem for ATC is the presence of background noise, which can dramatically degrade the intelligibility of the verbal communications since low quality messages delivered to the ATC may have fatal effects [23]. Thus, speech enhancement algorithm can be used as a pre-processor of the noisy speech signal before being fed to an automatic speech recognizer (ASR) in order to increase the robustness to background noise. This helps to enhance the recognition accuracy in adverse environments [24].

Speech enhancement plays an important role for the voice communication systems in military applications, since such systems can improve communication ability among people or between humans and computers in airborne environments. In most of these applications, improving quality and intelligibility of the desired speech signal that has been degraded and interfered due to the harsh noisy environments are highly desirable. In [25], an advanced headphone technology was presented for speech communication application in military aircraft environment such as an air fighter cockpit. The main task of such a system is to reduce the noise in the listener's ears in very noise environments. In [26], a speech enhancement system is proposed as a pre-processor in order to improve the overall performance of the automatic speech recognition (ASR) in high performance fighter aircraft. Moreover, a developed speech enhancement system called speech enhancement unit (SEU) sponsored by the Rome Air Development Center [27] was proposed to improve the speech readability and reduce the operator fatigue.

2.2.2 Biomedical applications

In industrial and heavy manufacturing workplaces such as mining, and construction sites, the workers are exposed to high level of noise for prolonged time which may result in a temporary or permanent hearing impairment. Hearing protection device (HPD) is a good solution to avoid the hearing impairment in very loud noise environments.

Noise control is the most used noise reduction technique for HPD. It can be classified into passive and active techniques. Passive noise control is used to reduce the ambient noise in noisy environment such as areas surrounding airports in order to sleep conveniently or listen to music without disturbance [28]. This kind of protection devices is sound reduction by noise-isolating materials such as insulation, sound-absorbing tiles, or a muffler. The main drawback of such a technique is the reduction of the background noise as well as the wearer ability to hear speech. This affect the verbally communication between workers and reduce the awareness of their surroundings and any kind of hazard [29].

In contrast, active noise control (ANC) reduces offensive (especially low frequency portions) of the noise by using cancellation techniques. ANC systems use microphones, speakers and digital signal processor. The main advantage of such devices is their ability to increase the efficiency of verbal communication and reducing the background noise while achieving a hearing protection [30]. However, they also result in speech distortion while reducing background noise. Hence, the essential crucial point in HPD is to develop a speech enhancement algorithm in order to isolate and enhance the signals of interest while reducing the harmful background noise to a safe level. This allows the wearer to remain aware of their surroundings while protecting their hearing [31].

Over the last three decades, the number of people with hearing problems have been increased around the world. According to World Health Organization (WHO) research, more than 200 million people around the world have hearing impairment issues [32]. It is a fact that hearing impaired people struggle to recognize necessary content from the original speech that degraded by different sources of noise, like interference, background noise and echo sounds. Hence, Assistive Listening Devices (ALD) such as Hearing Aids (HA) and Choclear Implant (CI) have been widely employed to provide assistance for people with hearing loss [33]. The main aim of ALD is to reduce the background noise while enhancing the useful signals. The efficiency of such devices is strongly depend on the performance of the signal processing for the speech enhancement. Current research has focused

on the speech enhancement algorithms in order to improve the speech quality and intelligibility of the hearing aid devices. This is achieved by contrasting speech and noise through a masking function or a gain function which localizes and preserves the speech components while attenuates the undesired noise [34]. This assist people with hearing loss to have a better verbal communication in noisy environments [35]

2.2.3 Commercial applications

A part from the biomedical or civic and military applications, speech enhancement plays an important role to most of the commercial voice applications. One of the most popular applications of voice communication systems is the use of voice controlled devices (VCD). It can be defined as a device that is controlled by the human voice which can be found in mobile phones, cars, internet search engine and home appliances.

In the automotive industry, an increasing number of new models feature voice-activated controlled system. Hyundai Motors has been one of the major manufacturers at the forefront of bringing this new technology to the market. The VCD allows the driver to issue voice commands in order to control the mobile phone, play music, send messages, give GPS navigation addresses or coordinates all the above via the cars inbuilt microphone and without being distracted. Automatic Speech recognition (ASR) system is the main speech processing technique employed to achieve this task. For a robust ASR system in highly noisy environments, the speech enhancement technique is utilized as a pre-processing operation which helps to mitigate unwanted signals such as background noise, echo or reverberation while preserving the desired speech signal for further processing. This improves the robustness of ASR system and enables the human-machine communication.

Another important commercial application for speech enhancement system is the teleconferencing application. This growing area of communication have attracted much attention in the last decades as it provides convenience and flexibility. The ultimate intention of any conference system is to facilitate communication between remote participants with high speech quality and low latency technique. High quality teleconferencing also saves the environment by providing the means to have effective remote meetings without the need to meet face to face. Speech enhancement is an enabling technology for this application. Depending on the size of the conference room, or type of the activities such as, formal presentation or distance learning, different speech enhancement techniques with

different number of microphones and configuration need to be used [36].

2.3 Speech Enhancement Techniques

The speech enhancement system based on the number of microphones can be classified into single and multi-channel speech enhancement techniques. Multi-channel techniques can provide an improved dereverberation, strong noise suppression and interference rejection as compared to the single channel techniques. Single channel speech enhancement technique however, is still useful because of its simple implementation. Accordingly, we provide an overview of single and multi-channel speech enhancement techniques.

2.3.1 Single channel speech enhancement techniques

In real-time applications like mobile communications and hearing aids, single channel speech enhancement is usually preferred. The main formulation of such technique is to estimate the speech signal that is degraded by uncorrelated additive noise. It often consists of one microphone used to estimate the clean speech using the temporal and spectral information of the degraded speech signal.

Background noise reduction in such applications is a hard challenge since there is no second channel used as a reference for the background noise. This makes the noise reduction without distorting the desired speech a very difficult task. The main advantage of single channel speech enhancement is having the potential to provide a very economic solution to the noise reduction problem because these systems are easy to build with less power usage and less computational complexity than multi channel systems [31]. In addition to their usage as a stand alone noise reduction system, single microphone system can be used in multi microphone systems as post filter to the beamformer algorithms. Over the last 30 years, the single channel speech enhancement techniques have attracted a significant amount of attention. One of the popular methods has been introduced by Boll is the spectral subtraction (SS) approach [37]. This non-parametric based approach is formulated to estimate the clean speech spectrum by subtracting an estimate of the noise spectrum from the observation spectrum. The main drawback of this approach is the appearance of unnatural sounding artifacts known as musical noise which is annoying and unpleasant to the listeners. Several methods have been presented to overcome this problem such as [38], which used a low variance spectrum estimator based on wavelet denoising that is thresholding the multitaper spectrum of the speech. Combining multitaper spectrum estimation with wavelet thresholding suppressed the musical noise and improved the speech quality.

Many important factors have to be taken into account when design a single

channel speech enhancement, such as the observed noisy spectrum estimate, noise estimate and the spectral gain function.

Literature review for framework of single channel speech enhancement technique

The overall architecture of the single channel speech enhancement framework consists of four blocks as shown in Figure 2.3. Each block is fulfilling a specific task to achieve the main goal of speech enhancement system. In the real time applications, only the contaminated observation signal is known, while the noise characteristics are unknown. As a consequence, the noise, a priori and/or a posteriori signal to noise ratio (SNR) have to be estimated. In many implementations where an efficient real-time performance is required, the observed signal is first transformed into a transform domain, then passed through an adaptive filter known as spectral gain function in order to estimate the clean speech spectrum. Thereafter, the reconstructed estimated signal is synthesized by using Inverse Fast Fourier Transform (IFFT) and overlap add method.

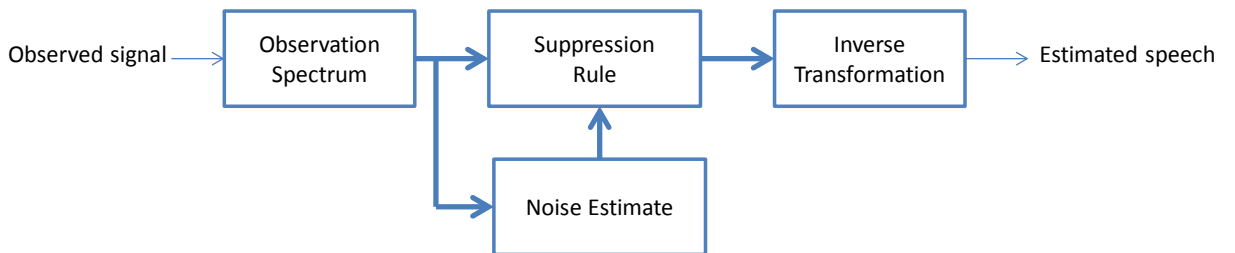


Figure 2.3: Framework of single channel speech enhancement.

Observation Spectrum

In the speech enhancement system design, it is important to consider the fact that speech is non-stationary. A common way to process the signal in order to reflect the non-stationarity of speech is to decompose the input signal into short frames and transform the signal into subbands. The choice of the spectrum transformation has a significant impact on the performance of speech enhancement systems in terms of speech quality and computational complexity [39]. The most popular choice for the spectral analysis model is the Short Time Fourier Transform (STFT). This can be achieved by dividing the observed time domain speech signal into short time frames, particularly, about (20-30) ms by using an

analysis window. Thereafter, transform each time frame into the frequency domain by implementing Fast Fourier transform (FFT). The significant benefits of using STFT is that it enables us to distinguish between speech and noise. This is especially the case for voiced speech components which can be clearly detected from the time varying STFT coefficients that known as frequency bins [40].

The main drawback of STFT is that its analysis results in frequency bands with uniform resolution which is not well adapted to the non uniform resolution of the human auditory system [41]. Given that the human auditory system performs some form of frequency signal analysis under adverse listening conditions, using human auditory models as a pre-processor in speech enhancement system can improve the subjective quality and/or intelligibility for the enhanced speech[11].

Many proposals in the field of speech enhancement have been represented the speech signals according to the human auditory system. For instance, in [42] a noise reduction technique is investigated based on an auditory filter bank with equivalent rectangular bandwidth (ERB) rate scale. The denoising technique is simply working by multiplying the output of each auditory filter by a weighting factor calculated from the updated signal statistics. The proposed technique improved the speech quality while reducing the musical noise level.

Wavelet Packet Transform (WPT) is another non uniform decomposition method that has been incorporated in many speech enhancement systems. The main advantage of such a method is its ability to simplify the mapping of audio signals into a scale that helps to preserve the time frequency related information. In [43], a speech enhancement method proposed that employed a bark-scaled WPT for higher time and frequency resolution than conventional WPT and critical band decomposition, respectively. The proposed technique improved the noise reduction while preserving the quality and intelligibility of the speech components. In this dissertation, the STFT coefficients are transformed into critical bands when calculating the gain function to study its impact on the overall performance.

Noise Estimation

Generally, all the speech enhancement methods are heavily dependent on the noise power spectral density (PSD). Since the spectral power of noise is unknown in advance, it has to be estimated from the noisy data. The quality of the noise PSD estimate can have a major impact on the quality of the enhanced signal. An inaccurate tracking of noise power changes can lead to an underestimation or overestimation of the noise PSD. The result of the underestimation and overestimation is speech distortion and/or low noise suppression. Therefore, an accurate

noise PSD tracking is essential to obtain a speech signal estimate that is close to the true speech signal. Characterization of noise PSD estimate must be performed during periods of silence between utterances, thus requiring a stationarity assumption of the background noise [3]. Accurate tracking and low computational complexity are the most important properties of noise estimators and are very challenging especially in adverse noise conditions such as traffic noise or babble noise.

A variety of different methods for estimation have been proposed over the years. The most established estimators are those based on voice activity detection (VAD). For relatively stationary noise, this estimator can be used to exploit the speech pauses to estimate and update the spectral noise [44]. However, tuning off VADs are not an easy task. Furthermore, VADs are not accurate under low SNRs which means that often low speech energy segments are not detected. Moreover, in non stationary case, VAD fails to track the fast power variations of the noise source, which might lead to high under-estimation or over-estimation over a relatively long time frame i.e. 100ms or so. Many other approaches have been proposed to estimate the noise PSD in non stationary noise environments. In [45], a minimum statistics (MS) based method was presented that tracks the minimum of the observed signal over a time span of about 1-3 seconds. This method provides an alternative of using a VAD for the noise estimation. The main advantage of MS based method is its ability to track the low variance noise. Moreover, it is capable of updating the noise PSD during speech activity. However, if the noise power rises during the time span, it will result in either underestimate of the noise power or overestimate depending on the time span length. As a consequence, these bias could result in musical noise or speech distortion when the noise estimation is applied to the speech enhancement framework. In order to reduce the tracking delay, a noise estimation based on DFT domain subspace is proposed [46], which exploits the STFT data over a few frames and estimates a correlation matrix over those frames and decomposes it using eigenvalue decomposition. The largest eigenvalues are assumed to belong to the speech signal subspace and the smaller eigenvalues are corresponding to the noise subspace. The main advantage of this approach is its ability to track slowly time varying power in the noise source at the cost of increased computational complexity.

For applications with complexity constraints such as mobile phone and hearing aids, low complexity noise PSD estimator is a crucial part in the noise reduction algorithm. Thus, [47] proposed high resolution discrete Fourier transform subspace based method, which achieves a reliable fast tracking noise estima-

tion and provides the same performance as DFT method when combined with the speech enhancement framework but with less complexity. Compared to MS based method, this approach improves the speech enhancement performance such as Signal-to-Noise Ratio (SNR) and Perceptually Evaluation of Speech Quality (PESQ) for non-stationary noise source with low computational complexity.

Suppression Rule

Noise reduction can be defined as a suppression rule or a non-negative real-valued spectral weighting gain to the observed signal spectrum. The main aim of most of the suppression rules is suppressing the spectral components of low SNR which are dominated by noise while spectral components of speech not affected. As such, the average SNR of the speech can be increased. As mentioned before, SS method is the most popularly used method due to its simplicity. But the enhanced signal derived by the SS method is not optimal since the enhanced speech is accompanied by an annoying perceptible tonal characteristic known as musical noise, which affects the human listening. This noise is sometimes more disturbing not only for the human ear, but also for speaker recognition systems [48]. Many solutions have been proposed to overcome this shortcoming, such as using the noise overestimation method which could help to eliminate the musical noise but at the cost of increasing the speech distortion. Another possible way is to introduce a gain floor parameter in order to limit variance of the gain function at low SNR. This solution can help to reduce the musical noise but not eliminate it [49]. In [14] this problem addressed by introducing knowledge on human perception in the subtraction style speech enhancement process. This helps to reduce the musical noise and improve the speech quality at low SNRs (< 10 dB).

Since the computation of spectral weighting rules in speech enhancement are often driven by the a posteriori and the a priori SNR [50], many gain functions were thus proposed, including the Wiener Filter (WF) [51] and the Minimum Mean Square Error short time Log Spectral Amplitude estimation (MMSE-LSA) gain functions [52]. However, the performance of most weighting rules is dominantly determined by the a priori SNR, while the a posteriori SNR acts merely as a correction parameter in case of low a priori SNR [49]. Hence, the suppression rule can be improved by modifying the a priori SNR estimation, but that is not enough since the joint temporal dynamics between the weighting function and the a priori SNR estimate have to be taken into account due to its significant impact on the performance of the speech enhancement system.

In many applications, improving the gain function and a priori SNR estima-

tion approach are not enough to improve the limited performance of the single channel speech enhancement. More effective solutions are in high demand to add more degree of freedom in the speech enhancement system design. One possible solution is to increase the number of microphones in order to exploit the spatial information in addition to the temporal and spectral properties.

2.3.2 Multi-channel speech enhancement technique

Multi-channel speech enhancement technique consists of microphone array with elements located at diverse spatial positions. The main advantage of using a microphone array is its ability to exploit the spatial information of the received signal in addition to the temporal and spectral information. Since the speech signal and noise are located in different positions in the room, the desired signal can be spatially separated from the noise [53]. This provides extra information about the desired signal characteristics and the noise properties [54].

In real time applications, speech and noise sources are usually generated from different locations while occupying overlapping frequency bands. As such employing microphone array as a speech enhancement technique can solve this problem by exploiting the spatial diversity. The main concept behind using a microphone array is the extraction of the desired speech signal that originates from a region of interest (ROI), and at the same time reducing incoming signals that originate from different locations other than ROI.

Many considerations have to be taken into account to design a microphone array system. One of these considerations is the microphone placement. Depending on the target applications, the geometry of the microphone array plays an important role in the design process and should not be neglected. In general, different array configurations have different overall performance. For example, linear uniform microphone arrays are the most common array configurations that are used due to their simplicity and ability to improve the spatial resolution. Another important consideration is the aperture (array) size. In order to improve the overall performance at a given frequency, it is necessary to increase the array size [55]. However, the aperture size is limited by the inter element spacing and the number of microphones. According to the Nyquist sampling theorem, the limit requires for the inter element spacing should be less than half a wavelength to avoid the spatial aliasing due to the under sampling of the received signal. The main drawback of under sampling is the appearance of the side lobes which reduces the attenuation of background noise and the undesired signals. Moreover, increasing the number of microphones is not an efficient solution since in

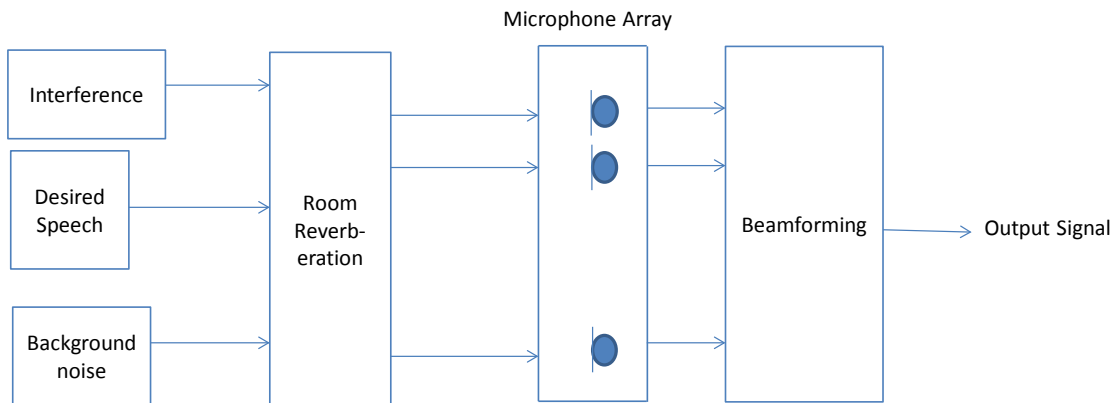


Figure 2.4: Framework of multiple channel speech enhancement.

many speech applications such as mobile devices impose a constraint on the array size. Therefore, new techniques have been attracted some attention to solve these problems such as superdirectivity [56] and sparse array design [57].

2.3.3 Beamforming

Beamforming is a temporal and spatial processor used in conjunction with the microphone array to perform the spatial filtering. The fundamental concept of the beamforming relies on the spatial and spectrotemporal discrimination of the desired components in the presence of background noise, reverberation and interfering signals. That means the main task of the beamformer is to extract the signal that originated from the region of interest while attenuating all other signals coming from different locations. Figure 2.4 illustrates a general framework for the microphone array system.

Beamforming has been applied in wide variety of application fields such as communication, radar, sonar, and biomedical. Generally, beamformers can be categorized into two types depending on the bandwidth of the signals received by the array: narrowband and broadband beamformers. Narrowband beamformer is basically based on the spatial selectivity to filter out the undesired signal that originated outside ROI. Delay and sum beamformer (DAS) is an example of the narrowband beamformer. The main concept of this kind of beamformer is synchronizing and adding. It is simply applying a complex weight to the received signals at each microphone and adding them together [58]. If the received signals are originating from ROI, they will be constructively summed and hence reinforced, and destructively summed otherwise. Radar is one of the most common applications of narrowband beamformer. However, in many speech applications,

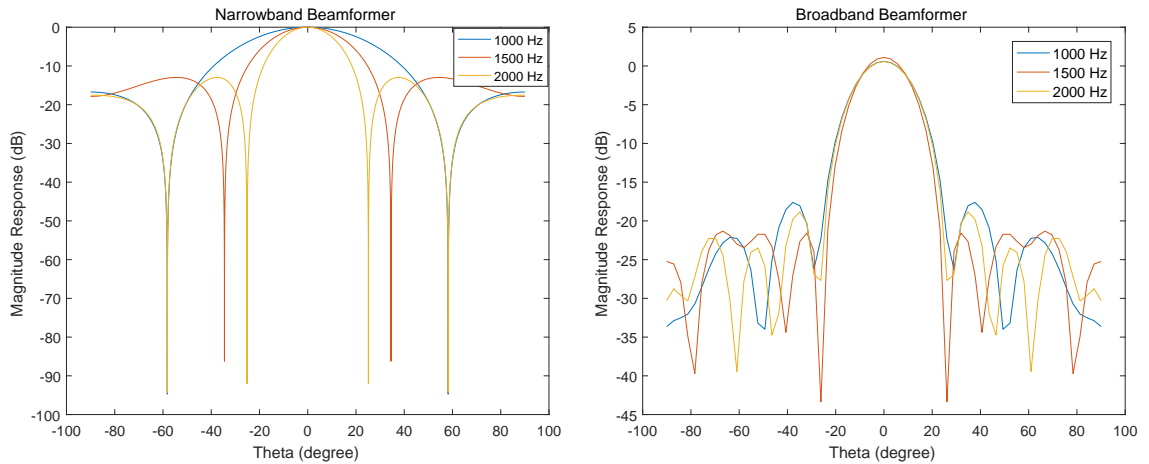


Figure 2.5: Beam-pattern comparison between narrowband beamformer and broadband beamformer for a linear microphone array.

it is required that the beamformer performance should be adequately constant over the entire frequency band of the speech signal [54]. Filter and sum beamformer is an example of the broadband beamformer. The main conception for such beamformer is that the signal at each microphone is processed by a finite impulse response (FIR) filter before they are summed together. Conceptually, narrow band beamformers are simpler than the broadband beamformers since they can exclude the temporal frequency variable and are based on the spatial separation only [59]. However, for speech communication applications, narrow band beamformers are ineffective since they cannot perform the same directivity pattern for a wide range of frequencies which degrades the performance of the beamformer. In addition, the unwanted signals will not be uniformly attenuated over its entire spectrum. Figure 2.5 shows beam pattern comparison between the narrowband and broadband beamformers for different frequencies. It illustrates the ability of the broadband beamformer in providing frequency invariant beam pattern while the narrowband beamformer gives a spatial resolution varies with frequency which is not useful for broadband applications.

For indoor applications such as conference rooms, nearfield broadband beamformers are required as the speaker would be in the near-field region. Far-field assumption in the design of near-field broadband beamformer will result in a

severe degradation in the overall performance [60].

Generally, the broadband beamformer design problem is to calculate the filter coefficients such that the actual response of the beamformer optimally fits the desired response, which is specified depending on the target application. In the literature, there are wide variety of optimization techniques dedicated to the design of broadband beamformers such as Least Square (LS), Weighted Least Squares (WLS) [61], [62] and Minimax [63], [64] criteria. One of the major difficulties in the design of broadband beamformer especially in the near-field region is that it becomes a large scale optimization problem when the spatial domain is a two or three dimensional region: the solution for such a problem may not be available due to the sheer size of the problem [65]. In [65] an interesting approach is presented to address this difficulty and a solution to the problem is suggested by decomposing the problem into two stages resulting in a significant reduction in the memory usage and computational complexity.

2.3.4 Beamforming classifications

The design of beamformers can be divided into two types: data independent and data dependent beamformers. The filter coefficients of the data independent beamformer (also known as fixed beamformer) do not depend on the target source or environment conditions and are chosen based on a pre-specified beam-pattern. In contrast, the filter coefficients of the data dependent beamformer (known as adaptive beamformer) are chosen based on the statistics of the received data to optimize the beamformer output. In the following subsection, more details about both types are discussed.

2.3.4.1 Fixed beamformer

In data independent beamformer design, the filter coefficients are pre-determined to extract the desired signal regardless of the statistical properties of the source signals. The main goal of such design is to obtain the spatial focusing on the desired source that originated from the ROI, which yields suppression to unwanted signals such as interference, or background noise. Different types of fixed beamformers include filter and sum [59], differential [66] and superdirectivity beamformers [67]. Figure 2.6 shows a fixed beamformer (filter and sum) structure. The main advantages of fixed beamformers are their ability to avoid signal distortion with no requirement of control algorithms and relatively simple numerical complexity with ease of implementation [68]. However, fixed beamformers have limited noise suppression since it is not well adapted to the changing acoustic environments. Therefore, many recent research have been interested in investi-

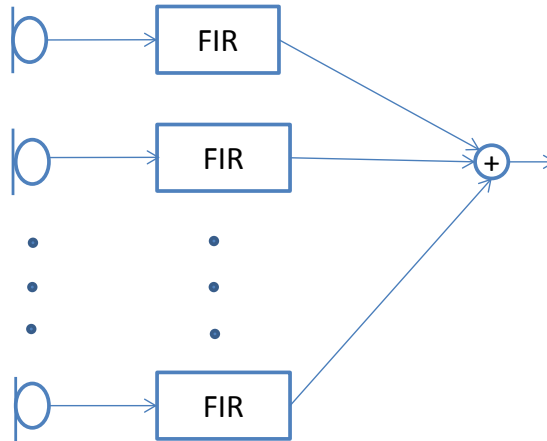


Figure 2.6: Structure of fixed beamformer.

gating fixed beamformer designs that are robust for adverse conditions such as [69], which discussed the beamformer design for indoor applications. It formulated the fixed indoor beamformer design as a minimax optimization problem and investigated its sensitivity against adverse noisy environments. The numerical experiments showed that the proposed designs are more effective in adverse environments than classic fixed beamformers.

2.3.4.2 Adaptive beamformer

Data dependent beamformer design is based on the statistical properties of the received signals. Compared to the data independent beamformers, adaptive beamformer exploits spectrotemporal signal properties and spatial filtering with an adaptive noise suppression algorithm. This typically leads to better noise suppression performance than fixed beamformer. Moreover, such design is continuously updated in order to track current statistics of the room propagation conditions. One of the most popular examples of adaptive beamformer is linearly constrained minimum variance (LCMV) beamformer [70]. The basic idea of the LCMV beamformer is to apply a linear constraint to the weight vector in order to control the beamformer output and maintain a constant gain and phase to the desired signals. This helps preserving the desired signal and minimizing the contribution of the noise signals and interference [59]. Another example of adaptive beamformer is the generalized sidelobes canceller (GSC) [71], which is an alternative unconstrained design problem of LCMV beamformer.

Figure 2.7 shows the general structure of generalized side lobe canceler. It

consists of three parts, the first part uses a fixed beamformer to form a beam towards the ROI in order to pass the target signal that originated from the same region while the other signals are attenuated. Second part uses a blocking matrix to form a null in the look direction in order to suppress the target signal and pass only the noise signals and interference. The third part uses multi channel adaptive filter to eliminate the noise signals that leaks through the sidelobes of the fixed beamformer. The main drawback of GSC is being sensitive to modeling errors caused by widespread speech source, steering delay errors, microphone characteristics and room reverberation. This might lead to a cancellation of the target speech signal since the block matrix fails to block the source signal which causes a leakage to the adaptive noise canceller. Different methods have been proposed to overcome this drawback, for instance, [72] proposed a method to design the blocking structure using a spatial filtering technique in order to broaden the look direction and therefore prevent the adaptive filter from cancelling the target signal.

As mentioned earlier, modeling errors can severely degrade the performance of the broadband beamformers. Any violation in assuming model can significantly degrade the overall performance. As such, it is important to model these errors and take them into account in the design procedure to reduce the sensitivity of the designed beamformer.

2.3.5 Sensitivity of beamformer design

Most of the beamformer design approaches assume the ideal microphone characteristics such as gain, phase and element position. This is not applicable in practice, and can lead to a severe degradation in the overall performance with respect to any mismatches in sensor characteristics, imperfect array calibration, mismatches in element position, and local scattering [73]. Hence, the development of an effective broadband beamformer design robust against microphone array imperfections has long been an important topic of research.

Numerous methods have been proposed to introduce robustness to such errors. One of these methods is sensor calibration [74]. This useful measure can be accomplished by using adaptive noise canceler. The main advantage of such method is the ability to improve directivity at low frequencies. However, this solution is less practical since it is kept fixed during the beamformer operation while in practice the microphone parameters change with time which necessitate re-calibration.

One of the most common methods used to improve the robustness of the

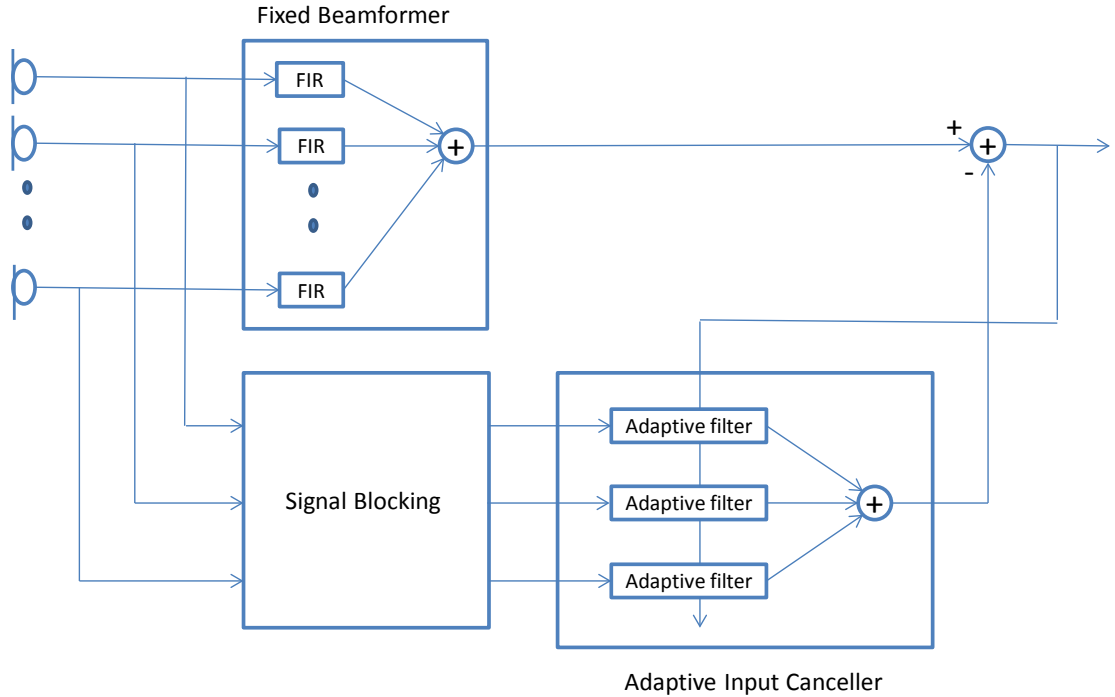


Figure 2.7: Structure of Generalized Sidelobe Canceller as an example of Adaptive beamformer.

beamformer design is White Noise Gain (WNG) constraint. In such technique, the filter coefficients are obtained by minimizing the beamformer output power subject to this constraint. The main reason for imposing a constraint into the optimization design is to limit the norm of the beamformer coefficients which results in lower beamformer sensitivity against mismatches in microphone characteristics. Moreover, this kind of constrained optimal solution represents a good trade off between increased array gain and sensitivity to errors [75] [76]. Although WNG constraint method provides a quick and simple method to develop a robust beamformer design, in practice it is hard to choose a suitable value for the minimum desired level of WNG for a range of microphone mismatches [77].

Another method to develop the robustness of the beamformer design is by exploiting the statistical knowledge about the error in the design procedure, using the probability of the microphone characteristics as weights. The beamformer filter coefficients are then calculated by optimizing either the weighted sum of the cost function, which is known as mean performance optimization technique, or the maximum cost function for all the feasible characteristics, which is known as the worst case optimization technique. When comparing the mean performance and the worst case optimization techniques, the main advantage of the worst case technique is no explicit knowledge of probability density function (PDF) of the

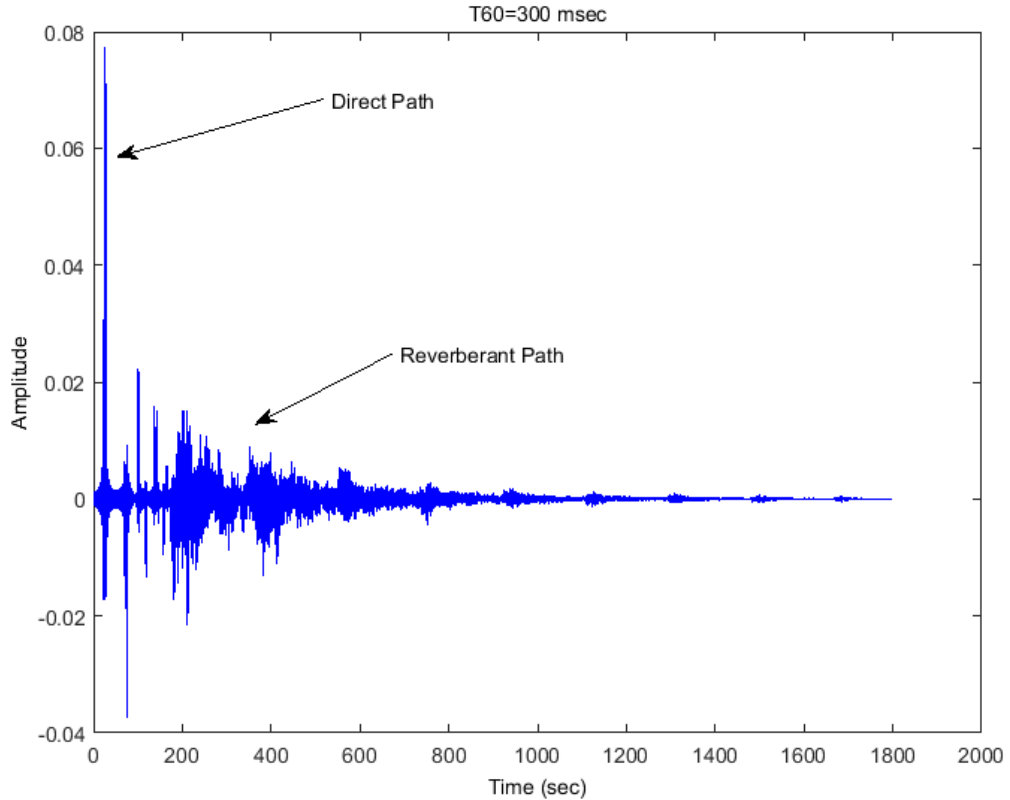


Figure 2.8: Simulated room impulse response using Image Source Method (ISM) at reverberation time $T_{60} = 0.3 s$

element characteristics are required in the design process. In addition, beamformer designed using the worst case optimization technique is robust under the worst case scenarios. On the contrary, such scenarios might not occur frequently, that makes the beamformer design more pessimistic [78]. On the other hand, mean performance optimization technique provides a robust beamformer design operates for mean conditions. However, if a sudden variance happens in the microphone characteristics, the overall performance may deteriorate.

2.3.6 Dereverberation

More challenging than a classical noise reduction problem is the dereverberation problem. In indoor applications, the microphones are picking up the desired signal as well as attenuated and delayed replicas of the desired signal known as reverberant parts due to the reflection from the walls, ceiling and furniture. This distorts the observed signal and degrade the intelligibility of the speech signal. Figure 2.8 shows the simulated Room Impulse Response (RIR) using Image Source Method (ISM) at $T_{60} = 0.3 s$, It can be seen that it consists of the direct path in addition to the attenuated and delayed reverberant path.

For many decades, acoustics researchers have been focusing on derreverberation

in speech communication in addition to noise reduction. A combined algorithm for both techniques is a very difficult task since the speech signal and noise are non stationary, the impulse response of the acoustic channel is very long and not a minimum phase style [79]. Therefore, an effective algorithm that is applicable in real time implementation is a very challenging task. Numerous techniques were developed using single channel noise reduction algorithm. Among them is Wu and Wang [80] who proposed a dereverberation technique combined of inverse filtering method to reduce the early reverberation and spectral subtraction to reduce the late reverberation. However, this technique did not take into account the noisy environments, which is not applicable in real time implementation where the speech signal is contaminated by the additive background noise and reverberation. Moreover, it is not effective in high reverberation conditions with low direct to reverberant ratio (DRR). To overcome these drawbacks, Saeed in [81] provides an accurate dereverberation technique that is robust to non stationary noise and high reverberation condition. It consists of two stages, in the first stage, it blindly estimates the inverse filter of the room impulse response in order to attenuate the early reverberation. In the second stage, it reduces the effect of the background noise and the residual reverberation by introducing an effective two step spectral subtraction method.

Dereverberation using multi-microphone speech enhancement have attracted more attention as a topic for further research for dereverberation and noise reduction since the spatial filtering facility of the beamforming process can separate the reverberation part from the direct part. Habets in [82] proposed a two stage beamforming approach by combining two different types of beamformer, DAS and MVDR beamformers. The first stage exploits DAS beamformer to spatially filtering the observed signal and dereverberate it, while the second stage is represented as multichannel noise reduction method to estimate the desired speech. The main advantage of this approach is its ability to trade off between the noise reduction and dereverberation with low speech distortion. [83] proposed an improved dereverberation method based on generalized side lobe canceler scheme combined with prewhiting approach for speech signal. The main concept is to design a blocking matrix to additionally block the early reverberant part under the assumption that the late reverberant part can be modeled as a diffuse noise. A hybrid dereverberation method is proposed in [84] to provide superior speech quality. In this method, correlation based blind deconvolution and modified spectral subtraction are combined in order to suppress the tail of the inverse filter reverberation and improve the quality of the inverse filtered speech signal. In this

thesis, we address the joint dereverberation and noise reduction issue using multi channel speech enhancement techniques.

2.4 Chapter Summary

In hands free communication, the speech quality is degraded due to the presence of several types of signal degradation such as background noise, echo, reverberation, which prevents the listener to have a relaxed communication. This implies a growing demand for speech enhancement or noise reduction in order to improve the speech quality and to reduce the listener's fatigue. This chapter presented various speech communication applications in different fields that exploited speech enhancement techniques to improve their performance.

We have discussed two basic speech enhancement techniques as the state of the art. The first technique is the single channel speech enhancement which is the most popular technique used in the hearing aids and mobile phones. The main advantages of single channel speech enhancement techniques consist of being simple, easy to implement in hardware and cost effective in practice. A detailed framework of single channel speech enhancement is presented to discuss the design decision considerations that have to be taken into account in the design procedure in order to preserve the speech components while reducing the musical noise.

The other speech enhancement techniques are multi channel speech enhancement techniques, which employ both the spatial and temporal filtering processes using a microphone array. Depending on the source signal and the target applications, there are different types and structure of beamformers. There is no optimal beamformer design that fits the desired response under all the conditions. It has to consider different parameters to optimize the best solution for the application at hand such as the array geometry, number of microphones and design specifications. Due to the sensitivity of most beamformer designs towards mismatches in microphone characteristics, there are high demands for developing beamformer design methods which are robust against deviations in microphone characteristics.

Furthermore, in the last section of this chapter, a brief overview was given of recent single and multi channel dereverberation techniques with their main advantages and disadvantages.

Chapter 3

Single Channel Speech Enhancement

This chapter addresses two problems of single-channel speech enhancement mainly, preservation of weak speech components, and musical noise. A modified a priori SNR estimation technique is proposed to improve faster tracking when SNR changes. This improvement of onset tracking is achieved by developing an adaptive weighting factor. As a consequence, better preservation of speech components is achieved. Moreover, we utilize a critical band mapping for STFT analysis-synthesis system in the speech enhancement framework to reduce the noise variance which results in a significant reduction in musical noise and computational complexity.

The main work in this chapter have previously appeared in the following publications:

1. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, "Convex combination framework for a priori snr estimation in speech enhancement," in IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), March 2017, pp. 4975-4979.
2. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, An adaptive a priori SNR estimator for perceptual speech enhancement, submitted to EURASIP journal on Audio, Speech and Music Processing 2019.

3.1 Introduction

Noise suppression and speech enhancement are important techniques employed in many products, for instance, mobile phones, hearing aids, hearables and assistive listening devices. Particularly, hearable devices have been poised to assist people with difficulties in hearing in social environments. For noise suppression and speech enhancement to work in those environments where acoustic noise becomes more intrusive, it is important to maintain weak speech components while still balance the amount of noise reduction. Accordingly, techniques that can enhance speech signals while preserving weak speech components under a large variety of acoustic scenarios are key to successful products [2, 3, 4]. In this context, it is important to not only consider the speech but also the quality of noise after suppression. Unnatural sounding background noise is very disturbing for users of hearable devices or hearing aids.

Traditionally, speech enhancement techniques have been utilizing the frequency domain for the processing where short-time Fourier transform (STFT) have been used as a tool to process the input data using frame-based oversampling techniques [3, 5, 9, 10]. When deploying STFT the bandwidth is constant for each frequency bin which is not the case for the human auditory system. Thus a natural extension has been to use human auditory models in the speech enhancement in order to improve the speech quality and intelligibility [13, 12, 11, 14].

Human auditory spectrum model consists of a bank of bandpass filters which follows a spectral bark scale or so-called critical bands [14, 15]. In [14], a standard subtractive speech enhancement method is presented to eliminate the musical artifacts in very noisy situations. The masking properties of the auditory system are utilized to compute the subtraction parameter. In [16], a spectral subtraction noise reduction method is proposed using a spatial weighting technique based on the inhibitory property of the auditory system, which results in improving the estimated speech while reducing the musical noise.

Speech enhancement algorithms calculate a gain function which is in most cases a function of a posteriori signal to noise ratio (SNR) or a combination of a posteriori and a priori SNR [85]. One classic speech enhancement algorithm is the spectral subtraction (SS) method proposed by Boll [37]. This algorithm is the most commonly used mainly due to its straightforward implementation and low computational complexity. In this method, a clean speech estimate is obtained by subtracting an estimated noise power spectrum from the noisy speech power

spectrum while keeping the phase of the degraded speech signal. The rationale for using the noisy phase is based on the assumption that the phase distortion is not perceived by the human ear. Even so, the spectral subtraction method embeds erroneous estimation of noise statistics resulting in an annoying artifact in the estimated speech signal commonly known as musical noise which can be masked using perceptual thresholds [86, 14].

In contrast, the log spectral amplitude minimum mean square error (LSA) estimator proposed by Ephraim [52], does not directly inherit the musical noise artifact. This estimator uses a priori SNR estimation based on a decision directed estimation which involves a weighted sum of two terms, the a priori SNR estimate from the previous frame and the maximum likelihood (ML) SNR estimate from the current frame. This estimation technique reduces the variance of the a priori SNR estimates particularly during noise frames and as a result, the musical noise artifact is eliminated [87]. However, the emphasis of the previous frame in the decision directed estimation has as a consequence that it leads to a slow adaptation towards speech onsets and offsets. Moreover, as the decision directed (DD) approach depends on the a priori SNR estimation in the previous frame, an extra one frame delay is included during speech transient and results in a degradation of the speech quality [10].

The a priori SNR estimation algorithm has been improved in many ways, e.g. Breithaupt et al. [88] proposed the temporal Cepstrum smoothing (TCS) technique for speech enhancement. This technique improves the accuracy of the a priori SNR estimation by exploiting the a priori knowledge of speech and noise signal and selectively smooth the maximum likelihood estimate in the Cepstral domain. This allows the preservation of speech components while simultaneously achieving high noise attenuation. However, this method has limitations under low SNR conditions where the noise components cannot be separated from the speech components. Suhadi [49] suggested a data-driven technique employing two trained neural networks to estimate the a priori SNR with one for speech and another for noise. The use of neural networks requires a substantial training process for estimating the a priori SNR since the proposed method is not robust estimator under different noise environment, which results in a degradation of the estimated speech quality under non-stationary noise conditions. In Plapous [89], a two-step noise reduction technique (TSNR) was presented in order to refine the estimation of the a priori SNR and increase the estimator adaptation speed. The main disadvantage when using this TSNR method is its sensitivity to the selection of the gain function. Different choices of the gain function give

very different estimation results. A modified decision directed approach (MDD) proposed by Yong et.al. [10] matches the current noisy speech spectrum with the current a priori SNR estimate rather than the delayed one. This reduces the one frame delay for speech onsets but the tracking speed of the a priori SNR estimation is still too slow compared to the true SNR change since the recursive smoothing factor is constant and close to one.

In this chapter, we presented an improved a priori SNR estimation based on modeling the speech absence probability with a sigmoid function. This sigmoid function was used to control the adaptation speed of the a priori SNR estimation. The sigmoid function operates as an adaptive weighting function that emphasizes either the DD term or the ML estimate in the a priori SNR estimate update. The rationale used when developing the weighting function was that for positive SNR values the a priori and the a posteriori SNR estimates are almost the same. Accordingly, by adding flexibility to select either of the two terms for SNR values below or above a certain threshold we provide an effective way to achieve the advantage of both estimates. By utilizing a threshold and the sigmoid shape, an improved adaptation of the a priori SNR estimate is obtained, which results in better preservation of weak speech components. The contributions we make in this chapter are fourfold:

- The robustness of the proposed method by employing different gain functions has been evaluated.
- An analysis of the effect for the key parameters to control the shape and the slope of the adaptive weighting function has been added.
- A new evaluation technique is introduced which is called the modified Hamming distance.
- More extensive listening tests and a larger evaluation test set have been investigated.

The choice of gain function plays a role since it is included in the DD estimation resulting in different performance. In this work, we propose an improved a priori SNR estimation using different gain functions namely Wiener filter [90] and MMSE-LSA gain function [52]. Since we are particularly interested in weak speech components, a new evaluation technique referred to as the modified Hamming distance has been proposed. In normal objective measures, weak speech components are not emphasized since they have small amplitudes or small energy.

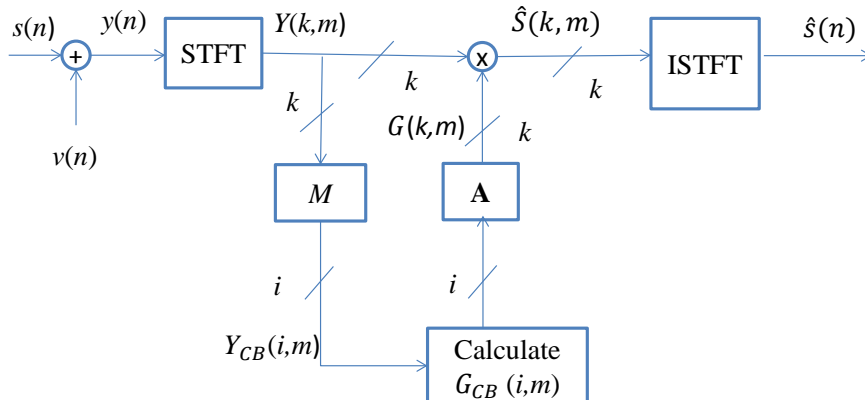


Figure 3.1: Block diagram for the critical band processing.

The proposed modified Hamming distance is based on VAD decision information in each time-frequency bin. Since it is binary, data scaling that depends on amplitude is avoided. Also in this work, we utilise a critical band mapping for STFT analysis-resynthesis system in the speech enhancement framework to better fit auditory listening experience. The proposed technique helps to reduce computational complexity since critical band processing combines K FFT frequency bins into I critical bands instead ($I \ll K$).

This chapter is organized as follows. In Section 3.2, a single channel speech enhancement framework with critical band processing is developed. Section 3.3 shows the decision-directed based a priori SNR estimators. In Section 3.4 the proposed a priori SNR estimation approach is developed, and an investigation into the effect of the key parameters of the sigmoid function is presented. In Section 3.5 the objective and subjective quality measurements used for evaluation are outlined. Section 3.6 presents the results of the experimental evaluation and Section 4.6 concludes the chapter.

3.2 Critical band speech enhancement

A natural way to process speech signals is to use a perceptual filter bank [39] since this would possibly improve the performance of the speech processing system by employing the inhibitory property of the human auditory system and combined with the speech enhancement algorithms [14]. There are many perceptual frequency warping scales used for the speech processing [91] and [92]. In this work, we employed a bark scale filter bank with a non-uniform resolution and incorporated it into a speech enhancement framework with the proposed a priori

SNR estimation method. We assume that the speech and noise are additive and uncorrelated, thus the noisy speech signal is given by

$$y(n) = s(n) + v(n) \quad (3.1)$$

where $s(n)$ and $v(n)$ denote the clean speech signal and noise, respectively. The block diagram for critical band speech processing is described in Figure 3.1.

In the sequel, we will outline the details of the processing. In the first step the noisy signal is transformed to the time-frequency domain by applying STFT with K frequency bins

$$Y(k, m) = S(k, m) + V(k, m) \quad (3.2)$$

where k is the frequency bin index and m is the time frame index. Then, in order to transform the output from the STFT domain $Y(k, m)$ into the critical band, an approximate analytical function is used to express the transformation between frequency f (in Hz) and critical band z (in bark scale), which is defined by [93]

$$f = 600 \sinh\left(\frac{z}{6}\right). \quad (3.3)$$

The noisy spectrum is expressed in terms of the critical band numbers i and frame index m by combining the FFT frequency bins into I critical bands as follows

$$Y_{\text{CB}}(i, m) = \sum_{k=1}^{K/2+1} M(i, k) |Y(k, m)| \quad (3.4)$$

where $i = [1, 2, \dots, I]$. The number of critical bands I is chosen with respect to the bark scale [93]. Here, $M(i, k)$ is the critical bandpass filter coefficients which is defined as [94]

$$M(i, k) = \begin{cases} 10^{(z(k)-z_c(i)+0.5)} & z(k) < z_c(i) - 0.5 \\ 1 & z_c(i) - 0.5 < z(k) < z_c(i) + 0.5 \\ 10^{-2.5(z(k)-z_c(i)-0.5)} & z(k) > z_c(i) + 0.5 \end{cases} \quad (3.5)$$

where $z_c(i)$ represents the center frequency of the i^{th} critical band. The main task of the speech enhancement scheme is to enhance the speech signal by applying a specific spectral gain function to the noisy spectrum. $\mathbf{G}_{\text{CB}}(m)$ denotes the gain vector in critical band for the m^{th} frame

$$\mathbf{G}_{\text{CB}}(m) = [G_{\text{CB}}(1, m), G_{\text{CB}}(2, m), \dots, G_{\text{CB}}(I, m)]^T$$

There are many different gain functions proposed in the literature. Common gain function often can be expressed as a function of the a priori SNR $\xi(i, m)$, such as WF method which can be defined as [90]

$$G_{\text{WF,CB}}(i, m) = \frac{\xi(i, m)}{1 + \xi(i, m)} \quad (3.6)$$

with $\xi(i, m)$ denoting a priori signal to noise ratio SNR, which is defined as

$$\xi(i, m) = \frac{\lambda_s(i, m)}{\lambda_v(i, m)} \quad (3.7)$$

where $\lambda_v(i, m) = E[|V(i, m)|^2]$ and $\lambda_s(i, m) = E[|S(i, m)|^2]$ are the power spectral density of noise and clean speech, respectively.

MMSE-LSA [52] is another widely used speech estimator, which is obtained by minimizing the mean square error of the logarithm of original and enhanced speech spectra, and can be defined as a function of the priori SNR and the posteriori SNR, given by

$$G_{\text{LSA,CB}}(i, m) = \frac{\xi(i, m)}{1 + \xi(i, m)} \exp \left\{ \frac{1}{2} \int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt \right\} \quad (3.8)$$

where the lower limit ν_k of the integral is given by

$$\nu_k = \frac{\xi(i, m)}{1 + \xi(i, m)} \gamma(i, m) \quad (3.9)$$

and $\gamma(i, m)$ denotes a posteriori SNR defined as

$$\gamma(i, m) = \frac{|Y_{\text{CB}}(i, m)|^2}{\lambda_v(i, m)}. \quad (3.10)$$

Once the gain vector $\mathbf{G}_{\text{CB}}(m)$ in critical band is calculated, it is interpolated back to the STFT resolution $\mathbf{G}(m)$ through an interpolation matrix \mathbf{A} ,

$$\mathbf{G}(m) = \mathbf{A} \mathbf{G}_{\text{CB}}(m) \quad (3.11)$$

where the \mathbf{A} matrix can be defined by least square approximation as $\mathbf{A} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T$ and \mathbf{M} denotes the matrix with elements $M(i, k)$. From empirical findings, better results are obtained by simplifying the reconstruction matrix as

$$\mathbf{A} = \text{diag} \left(\frac{1}{\mathbf{1M}} \right) \mathbf{M}^T$$

where $\mathbf{1}$ is $1 \times I$ row vector. The estimated speech in the STFT domain is then reconstructed by applying the interpolated gain function $G(k, m)$ on the noisy signal in Eq. (3.2)

$$\hat{S}(k, m) = G(k, m)Y(k, m). \quad (3.12)$$

Finally, the speech estimate is obtained by taking the inverse STFT of the enhanced speech and using the overlap-add method

$$\hat{s}(n) = \text{ISTFT} \left(\hat{S}(k, m) \right). \quad (3.13)$$

3.3 Conventional a priori SNR estimation

In many speech enhancement algorithms, a priori SNR estimation is a dominant part of the gain function calculation as in Eq. (3.6) and Eq. (3.8). Inaccuracies in the estimation of the a priori SNR can lead to audible speech distortion and musical noise. The state-of-the-art method to estimate the a priori SNR from noisy speech while avoiding the musical noise is the decision directed (DD) approach [50]. In this method, the a priori SNR estimation is expressed as a weighting average of the amplitude estimate at the previous frame and the maximum likelihood estimate of the a priori SNR at the current frame. This method is defined by

$$\hat{\xi}_{\text{DD}}(i, m) = \beta \frac{|\hat{S}(i, m-1)|^2}{\hat{\lambda}_v(i, m-1)} + (1 - \beta)P[\hat{\gamma}(i, m) - 1] \quad (3.14)$$

where $\hat{S}(i, m-1)$ and $\hat{\lambda}_v(i, m-1)$ denote the amplitude estimate and the noise estimate at the previous frame, respectively. P is the half wave rectification to keep the a priori SNR value positive, and $0 < \beta < 1$ denotes a weighting factor that controls the trade-off between the a priori SNR from previous frame and the posteriori SNR at current frame.

In this method, by setting the weighting factor close to 1, two different behaviors of the a priori SNR estimation can be observed as explained in [87]. In the noise frames where the estimate a posteriori SNR is lower than or close to 0 dB, the a priori SNR estimate corresponds to a scaled version of the a posteriori SNR since the second term of the DD approach is equal to zero. By substituting Eq.(3.10) and Eq.(3.12) into Eq.(3.14), a priori SNR estimation can be expressed by

$$\hat{\xi}_{\text{DD}}^\downarrow(i, m) \approx \beta G_{\text{CB}}^2(i, m-1)\hat{\gamma}(i, m-1).$$

This behavior reduces the variations in the a priori SNR estimate and thus reduces the amount of musical noise produced. In the frames with speech onsets, the a priori SNR follows the a posteriori SNR from the preceding frame as given by

$$\begin{aligned}\hat{\xi}_{\text{DD}}^{\uparrow\uparrow}(i, m) &= \beta \frac{G_{\text{CB}}^2(i, m-1) |Y_{\text{CB}}(i, m-1)|^2}{\hat{\lambda}_v(i, m)} + (1-\beta)P[\hat{\gamma}(i, m) - 1] \\ &\approx \beta G_{\text{CB}}^2(i, m-1)\hat{\gamma}(i, m-1) + (1-\beta)P[\hat{\gamma}(i, m) - 1]\end{aligned}$$

where the second term that indicates the ML estimate would only have little impact on the estimation process since β is very close to 1. In this case, the tracking of change in the a priori SNR estimate is slow since the a priori SNR estimation mainly depends on the posteriori SNR estimation in the previous frame. This behavior can lead to speech transient distortion. In order to overcome this problem, the authors in [10] proposed a modified decision directed (MDD) approach. In that method, the a priori SNR estimate at the current frame is matched with the a posteriori SNR in the current frame instead of the previous one. Thus the one frame delay is reduced which results in less speech distortion comparing to the conventional DD approach. The MDD a priori SNR estimate is given by

$$\hat{\xi}_{\text{MDD}}(i, m) = \beta \frac{G_{\text{CB}}^2(i, m-1) |Y_{\text{CB}}(i, m)|^2}{\hat{\lambda}_v(i, m)} + (1-\beta)P[\hat{\gamma}(i, m) - 1]. \quad (3.15)$$

In addition, to maintain the advantage of the DD approach in eliminating the musical noise, the magnitude square of noisy signal has been smoothed by using first order recursive smoothing procedure as given by [10] to reduce the variance of the a priori SNR estimate. The first order recursive averaging of the noisy signal is given by

$$\lambda_y(i, m) = \alpha_y \lambda_y(i, m-1) + (1-\alpha_y) |Y_{\text{CB}}(i, m)|^2 \quad (3.16)$$

where α_y is a smoothing constant. The smoothed $|Y_{\text{CB}}(i, m)|^2$ is replacing the instantaneous power estimate in the a posteriori SNR in Eq. (3.10).

3.4 Proposed a priori SNR estimation

The drawback of the MDD approach is that the fix weighting factor β , e.g. $\beta = 0.98$, reduces the influence from the second term towards the a priori SNR update resulting in a scaled down a priori SNR estimate when compared to the true a priori SNR. In the light of this, we can conclude that the fix weighting

factor β gives low variability of the gain function during noise only periods but does not provide a fast change of the gain function when a speech utterance comes. Thus it is desirable to replace the fix weighting factor β with an adaptive weighting factor $\beta(i, m)$.

Recognizing that the speech absence probability is a key for the weighting according to Eq. (3.15), we model the speech absence probability based on a sigmoid function. As a remark, if the CDF is a sigmoid function, the pdf is similar to a Gaussian pdf but with larger tails which is plausible for speech applications. The sigmoid consists of two parameters, σ to control transition speed and ρ to determine the threshold for active speech signal versus noise only. The selection of these parameter values are based on the observation that the a priori SNR equals the posterior SNR for high SNRs. An adaptive weighting function $\hat{\beta}(i, m)$ is proposed based on the a posteriori SNR and is given by

$$\hat{\beta}(i, m) = \frac{\beta_0}{1 + \exp[-\sigma(\tilde{\gamma}(i, m) - \rho)]} \quad (3.17)$$

where β_0 is a constant. The modified a priori SNR estimation approach is then defined by

$$\hat{\xi}_{\text{prop}}(i, m) = \hat{\beta}(i, m) \frac{G_{\text{CB}}^2(i, m-1) |Y_{\text{CB}}(i, m)|^2}{\hat{\lambda}_v(i, m)} + (1 - \hat{\beta}(i, m)) P[\tilde{\gamma}(i, m) - 1] \quad (3.18)$$

where $\tilde{\gamma}(i, m)$ is the a posteriori SNR estimate employing the smoothed estimate of the noisy speech from Eq. (3.16). Figure 3.2 describes the computation of the gain function by using the proposed method with an adaptive weighting function. In the following, we investigate the effect of two parameters σ and ρ on the proposed adaptive weighting function $\hat{\beta}(i, m)$.

To retain similar property as a constant weighting factor β for speech only and the noise only frames, we impose constraints on $\hat{\beta}(i, m)$ as:

$$\hat{\beta}(i, m) = \begin{cases} \beta, & \text{for noise only frames or when } \tilde{\gamma}(i, m) = 1 \\ 1 - \beta, & \text{for speech only frames or when } \tilde{\gamma}(i, m) = \gamma_u, \gamma_u \gg 1. \end{cases} \quad (3.19)$$

For $\beta = 0.98$, the constraints in Eq. (3.19) lead to

$$\begin{cases} \frac{\beta_0}{1 + \exp(-\sigma(1 - \rho))} = 0.98 \\ \frac{\beta_0}{1 + \exp(-\sigma(\gamma_u - \rho))} = 0.02 \end{cases} \quad (3.20)$$

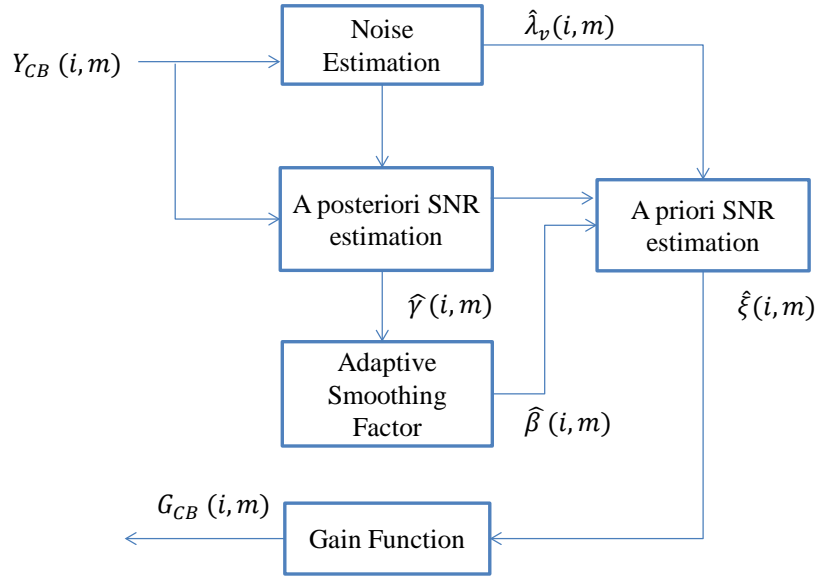


Figure 3.2: Block diagram of the spectral gain function computation using the proposed a priori SNR estimation method.

or

$$\begin{cases} \sigma(1 - \rho) &= -\ln\left(\frac{\beta_0}{0.98} - 1\right) \\ \sigma(\gamma_u - \rho) &= -\ln\left(\frac{\beta_0}{0.02} - 1\right). \end{cases} \quad (3.21)$$

We now calculate the parameters σ and ρ directly for different levels of γ_u . From Eq. (3.21), we have

$$\frac{1 - \rho}{\gamma_u - \rho} = \frac{\ln\left(\frac{\beta_0}{0.98} - 1\right)}{\ln\left(\frac{\beta_0}{0.02} - 1\right)}. \quad (3.22)$$

As the right terms in Eq.3.21 $\frac{\beta_0}{0.98}$ and $\frac{\beta_0}{0.02}$ have to be larger than 1, β_0 has to be slightly larger than 0.98. Through speech enhancement experiments, good results were obtained with $\beta_0=0.983$. As such, the parameter ρ can be obtained from γ_u as

$$\rho = \frac{1 - \gamma_u \frac{\ln\left(\frac{0.983}{0.98} - 1\right)}{\ln\left(\frac{0.983}{0.02} - 1\right)}}{1 - \frac{\ln\left(\frac{0.983}{0.98} - 1\right)}{\ln\left(\frac{0.983}{0.02} - 1\right)}}. \quad (3.23)$$

The parameter σ can be calculated as

$$\sigma = \frac{-\ln\left(\frac{0.983}{0.98} - 1\right)}{1 - \rho}. \quad (3.24)$$

Figures 3.3 and 3.4 show the pdf of a posteriori SNR for different noise types, mapped with a different adaptive smoothing factor calculated at several posteriori SNR values γ_u : (i) at $\gamma_u=5$ dB SNR with $\sigma = -4.469$, $\rho = 2.295$, (ii) at $\gamma_u=7$ dB SNR with $\sigma = -2.408$, $\rho = 3.402$, (iii) at $\gamma_u=9$ dB SNR with $\sigma = -1.391$, $\rho = 5.159$ and (iv) at $\gamma_u=15$ dB SNR with $\sigma = -0.315$, $\rho = 19.344$. Adaptive smoothing factor with different parameters (slopes and means) can control the trade off between the musical noise and the ability to preserve weak speech components. In pink and white noise cases, SNR estimate in noise only case is distributed approximately between 0 and 1. According to Eq. (3.19), adaptive smoothing factor should be almost β during this period to reduce the SNR variance.

It can be noted from Figure 3.3 (first two plots on the left), where the adaptive smoothing factor is almost (0.983), which explains the ability of the proposed method to maintain the advantage of the conventional decision directed and modified decision directed method in reducing the musical noise at low SNRs. Moreover, in the factory noise case where the SNR estimate is distributed between 0 and 2 during noise only period, the proposed smoothing factors designed at $\beta_u = 9$ and $\beta_u = 15$ reached the imposed constraint (0.983) during the noise variance, whereas adaptive factors designed at $\beta_u = 5$ and $\beta_u = 7$ starts reducing during noise period which leads to increase the musical noise. However, for the babble noise scenario as shown in Figure 3.4, figure on left shows the PDF of a posteriori SNR estimate during noise only period. It can be observed that the PDF has a large spread because of the non-stationary character of the babble noise, which means that an adaptive smoothing factor designed at higher a posteriori SNR γ_u is required to reduce the SNR variance during noise only frame and reducing the effect of the musical noise. From the figure, it can be clearly noted that adaptive smoothing factor designed at $\gamma_u = 15dB$ is the best among the other designed factors since it attained a higher value over the a posteriori SNR distribution during the noise only frame.

In addition, it can be noted that the weighting factor is inversely proportional to the a posteriori SNR γ . Thus during the noise frames, γ takes small values. Consequently, the resulting weighting factor $\hat{\beta}(i, m)$ is close to 1, which means that the proposed method will have the identical behaviour as the DD and the MDD methods. This explains the ability of the proposed method to maintain the advantage of DD method in reducing the musical noise in the low SNRs. Since the second term is zero, the a priori SNR estimate in noise frames will be given by

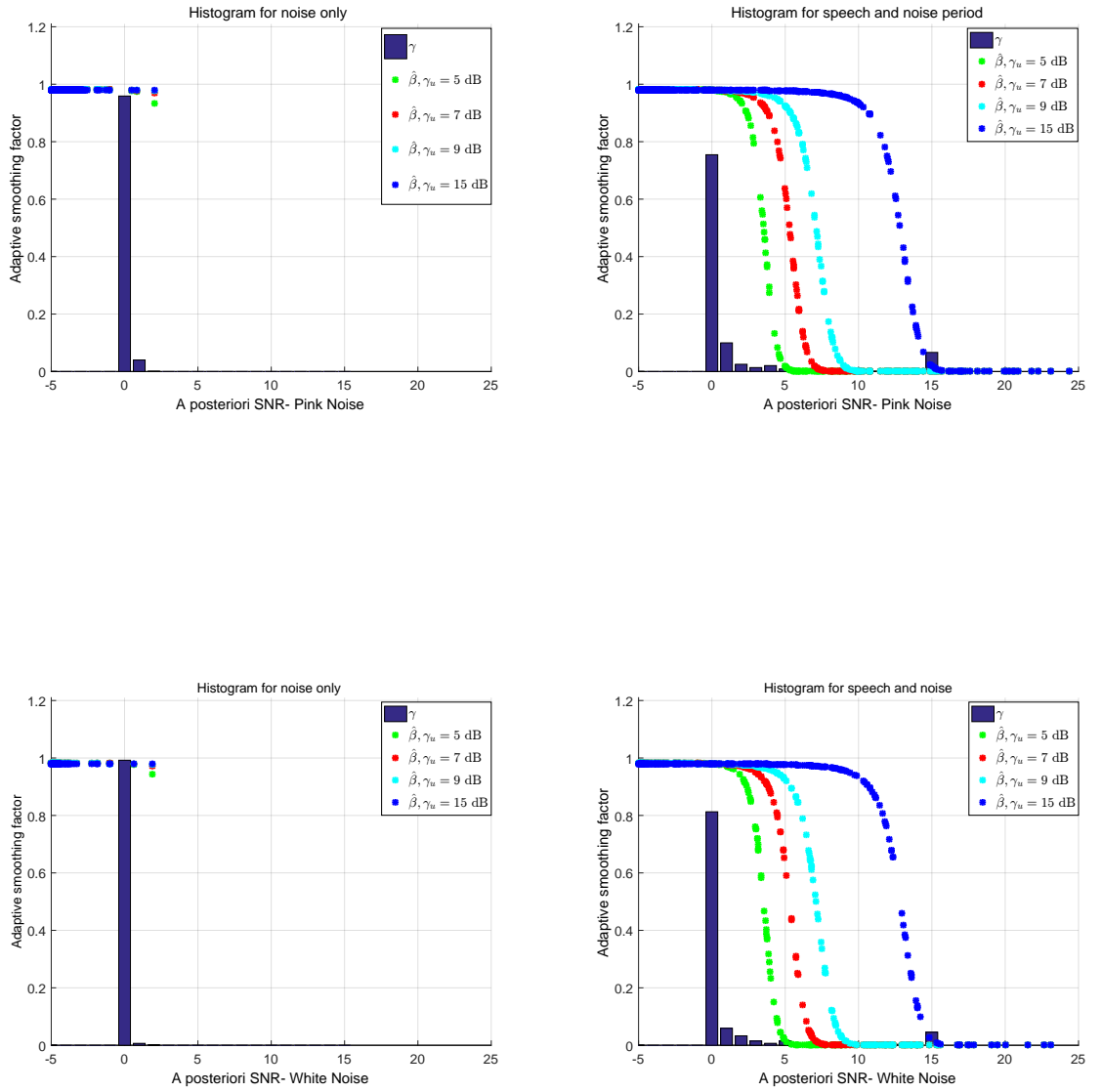


Figure 3.3: Histogram of a posteriori SNR estimate for different background noise (1^{st} row) for pink noise, (2^{nd} row) for white noise at 9^{th} critical band mapped with adaptive smoothing factor calculated with different sets of parameters (adaptive smoothing factor calculated at (i) $\gamma_u=5$ dB, (ii) $\gamma_u=7$ dB, (iii) $\gamma_u=9$ dB and (iv) $\gamma_u=15$ dB). Left figure for noise period only and right figure for speech and noise period.

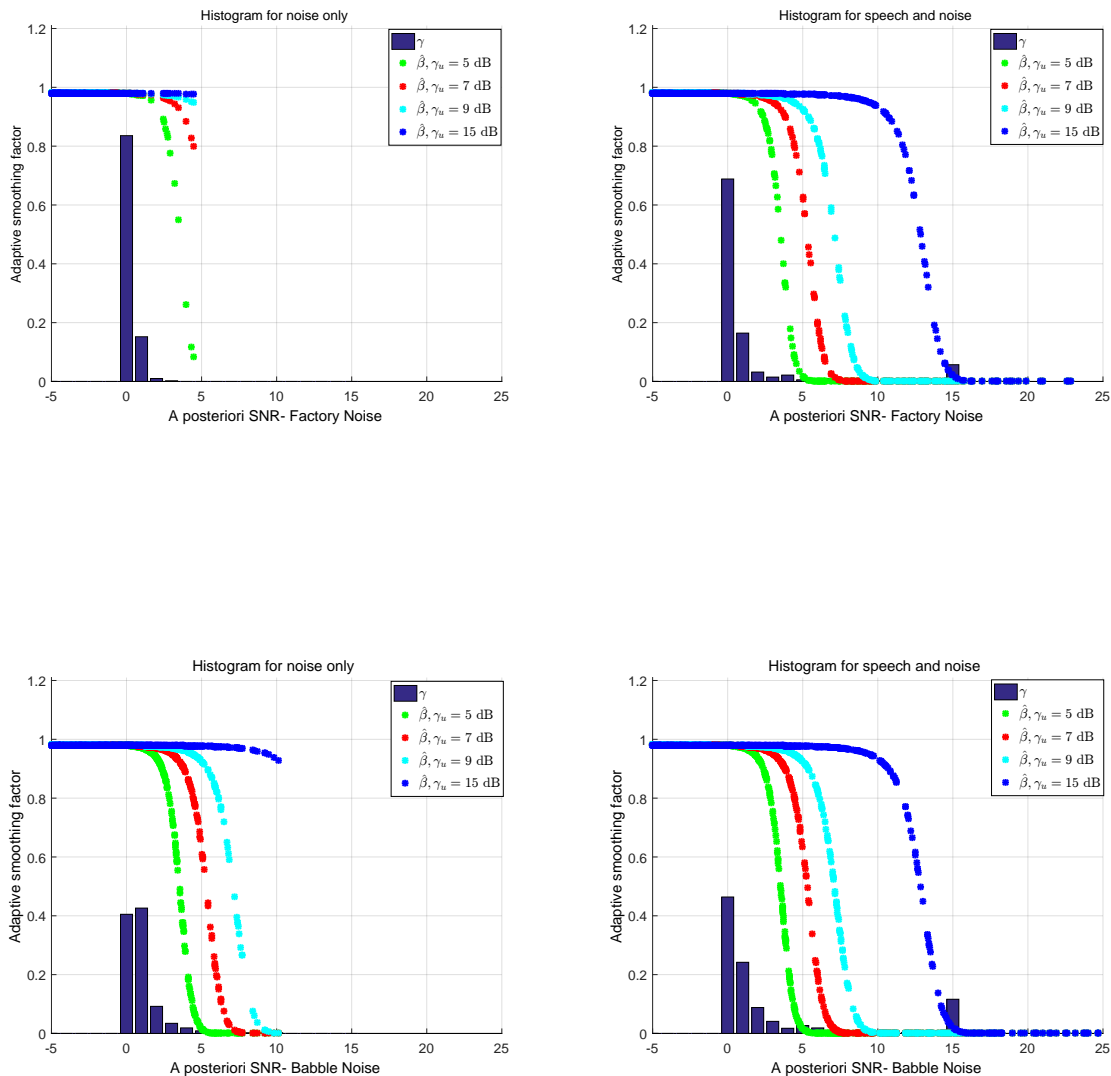


Figure 3.4: Histogram of a posteriori SNR estimate for different background noise (1st row) for factory noise and (2nd row) for babble noise at 9th critical band mapped with adaptive smoothing factor calculated with different sets of parameters (adaptive smoothing factor calculated at (i) $\gamma_u=5$ dB, (ii) $\gamma_u=7$ dB, (iii) $\gamma_u=9$ dB and (iv) $\gamma_u=15$ dB). Left figure for noise period only and right figure for speech and noise period.

$$\hat{\xi}_{\text{prop}}^{\downarrow}(i, m) = \hat{\beta}(i, m)G_{\text{CB}}^2(i, m - 1)\hat{\gamma}(i, m).$$

During speech activity frames, the resulting weighting factor takes values close to 0. In that scenario the first term of Eq. (3.18) is almost negligible, the a priori SNR estimate in speech activity frames will correspond to a smoothed version of maximum likelihood estimate as given by

$$\hat{\xi}_{\text{prop}}^{\uparrow\uparrow}(i, m) = (1 - \hat{\beta}(i, m))P[\tilde{\gamma}(i, m) - 1].$$

During the speech transition, the weighting factor decreases with the increment of the instantaneous SNR. As a consequence, the a priori SNR estimation corresponds to a combination of the first and second terms in Eq. (3.18) as given by

$$\hat{\xi}_{\text{prop}}^{\uparrow}(i, m) = \hat{\beta}(i, m)G_{\text{CB}}^2(i, m - 1)\hat{\gamma}(i, m) + (1 - \hat{\beta}(i, m))P[\tilde{\gamma}(i, m) - 1].$$

From Eq. (3.18), it can be noticed that the second term will have a varying impact on the a priori SNR updating process depending on the instantaneous SNR estimate. It is here the method makes a difference in tracking any abrupt SNR changes. The apparent result is that more speech components are preserved as well as a reduction in the speech transient distortion.

3.5 Objective and subjective quality measurements

Speech quality evaluation can be classified into two measurements categories: objective measurement and subjective measurement [3]. The first category is based on the comparison between the original and the enhanced speech signals. Many objective measurements have been proposed in the literature such as perceptual evaluation of speech quality measure (PESQ) [95, 96], segmental SNR measure SNR_{seg} [97, 98] and kurtosis ratio measure (KurtR) [99]. In addition, we propose a new evaluation method based on Hamming distance to measure the weak speak components. The Hamming distance is a GF(2) measure that takes into account speech presence or not for each time frequency point. By measuring the difference between clean speech binary mask and processed speech binary mask the measure takes into account the presence of speech in each time frequency bin without amplitude weighting.

Perceptual evaluation of speech quality measure (PESQ) is the speech quality assessment recommended by ITU-T P.862 for its ability to predict the speech

quality with a high correlation versus subjective listening tests [100]. PESQ implementation consists of first, estimating the bark spectrum of the input and the degraded signals by using a perceptual model in order to compute the loudness spectra, and then compare between them to predict the perceived quality of the degraded signal. This objective means of quality assessment is expressed in terms of the mean opinion scores (MOS), measured from 1 to 5, where higher scores indicate higher quality. Here, we are using the implementation provided by Loizou [3].

Time domain based segmental SNR is one of the widely used objective measures to evaluate the performance of speech enhancement algorithms, which is formed by averaging the frame level of SNR estimate [97] as given by

$$\text{SNR}_{\text{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\|\mathbf{s}(m)\|^2}{\|\mathbf{s}(m) - \hat{\mathbf{s}}(m)\|^2} \quad (3.25)$$

where M denotes the number of frames, while $\hat{\mathbf{s}}(m)$ and $\mathbf{s}(m)$ are the estimated and original speech vectors, respectively in time domain. The segmental SNR values are limited in the range of $[-10, 35]$ dB in order to exclude frames with no speech.

Kurtosis ratio measure is a mathematical measure used to calculate the musical noise, which is defined by the estimated speech signal and the noisy speech signal during noise frames only [99]. In order to detect the speech silence and presence, a multi decision sub band VAD (MDSVAD) is employed [101], given two hypothesis $\mathcal{H}_0(k, m)$ and $\mathcal{H}_1(k, m)$ indicate the speech absence and presence, respectively. MDSVAD is given by

$$D(k, m) = \begin{cases} 1 & \mathcal{H}_1(k, m) \\ 0 & \mathcal{H}_0(k, m) \end{cases} \quad (3.26)$$

and $V(k, m) = 1 - D(k, m)$ denotes the activity detection of the noise periods. Kurtosis ratio can be defined by

$$\text{KurtR} = E \left\{ \frac{\kappa_{\hat{s}}(k)}{\kappa_y(k)} \right\} \quad (3.27)$$

where $\kappa_{\hat{s}}(k)$ and $\kappa_y(k)$ indicate the kurtosis of the enhanced signal and the noisy signal at the k^{th} frequency bin, respectively. They are defined as follows

$$\kappa_{\hat{s}}(k) = \frac{\sum_{m=1}^M \left| \hat{S}_s(k, m)V(k, m) \right|^4}{\left\{ \sum_{m=1}^M \left| \hat{S}_s(k, m)V(k, m) \right|^2 \right\}^2} - 2 \quad (3.28)$$

and

$$\kappa_y(k) = \frac{\sum_{m=1}^M |Y(k, m)V(k, m)|^4}{\left\{ \sum_{m=1}^M |Y(k, m)V(k, m)|^2 \right\}^2} - 2. \quad (3.29)$$

Based on the MDSVAD [102], we propose an evaluation method to measure the capability of the speech enhancement technique for preserving more weak speech components, referred to as the modified Hamming distance. It is determined by the difference of the time-frequency points detected using the MDSVAD [102] applied on the clean speech signal and the estimated speech signal. The detection of the MDSVAD decisions for the noisy speech signal and estimated speech signal was performed only based on full band VAD decisions for clean speech frames. The rationale for developing this new measure is that the result is amplitude invariant which is important when measuring weak speech components. Those weak speech components would otherwise be overshadowed by strong amplitude components. The modified Hamming distance measure is calculated as

$$\text{HD} = \frac{2}{KM} \sum_{m=1}^M \sum_{k=1}^{K/2} \left(\hat{D}(k, m) \oplus D(k, m) \right). \quad (3.30)$$

where \oplus performs a logical XOR operation that returns output containing elements set to either logical 1 (true) or logical 0 (false). Since the used signals are real, the FFT will be symmetrical, thus we used half of the FFT length and then multiply it by 2. Here, $D(k, m)$ denotes the voice activity detection of the clean signal and $\hat{D}(k, m)$ denotes the MDSVAD of the estimated speech signal conditioned on clean speech detected which is computed initially by testing each sub-band independently for speech activity using the decision device, then analyzed by further logic to reduce false-alarm. A lower HD score indicates more weak speech components are preserved.

The second category of evaluations is based on the subjective listening tests, which are considered more accurate and reliable [103]. For the subjective listening test, a total of 10 subjects (5 males and 5 females) were recruited to compare and rate between the estimated speech signals, the noisy signals and the clean speech signals under different SNR conditions. Three different utterances from 1 female and 2 male speakers have been concatenated and used for this test. They were corrupted with pink noise or babble noise at 10 dB SNR. The listening test was performed in a quiet office room using a DT-880 Beyerdynamic open air headphones. A laptop was connected through the USB interface to the headphones

Rating	Description
Speech	
5	very natural, no degradation
4	fairly natural, little degradation
3	somewhat natural, somewhat degraded
2	fairly unnatural, fairly degraded
1	very unnatural, very degraded
Background Noise	
5	not noticeable
4	somewhat noticeable
3	noticeable but not intrusive
2	fairly conspicuous, somewhat intrusive
1	very conspicuous, very intrusive
Musical noise	
5	not noticeable
4	somewhat noticeable
3	noticeable but not intrusive
2	fairly conspicuous, somewhat intrusive
1	very conspicuous, very intrusive

Table 3.1: Scale description of the listening test criteria

via a Topping VX-1 amplifier to provide good quality audio and consistent sound level. The sound clips were embedded in a PowerPoint document which was also used for recording the results. The listeners were required to listen to the sentences enhanced by the different methods (DD, MDD and the proposed method) and rate them on a scale goes from 1 to 5 by steps of 1. This rating takes into account three criteria: speech quality, background noise, and musical noise [3]. The ranking instruction can be found in Table 3.1, which describes the scale of the criteria used in the listening test. The clean speech signals and the noisy signals were included in the listening tests as references.

3.6 Experimental results and discussion

3.6.1 Experimental setup

In this section, extensive experiments are conducted to evaluate the performance of the proposed approach in different scenarios. First, the performance of the proposed method is compared to the performances of the DD approach [50]

and the MDD approach [104]. Second, we demonstrate the robustness of the proposed a priori SNR estimator by employing different gain functions. The speech sequences and noise are extracted from the NOISEUS and NOISEX database, respectively [3]. In this work, 30 speech sentences are used (15 male speakers and 15 female speakers). Four different background noise types are employed which include pink noise, white noise, factory noise and babble noise. The noisy signal is obtained by combining the speech sequences with background noise at input SNRs of 0, 5 dB and 10 dB. All the sequences have been re-sampled to $f_s = 8000$ Hz. An STFT analysis with a length of $K = 512$ is used with a frame rate of $R = 256$ and square-root Hanning window. Based on these values, the frequency bins of the noisy spectrum are then grouped into $I = 17$ critical bands as shown in Eq. (3.4).

Minimum mean square error (MMSE) noise power estimator based on the speech presence probability [105] was employed to estimate the noise PSD for all the a priori SNR estimators. The value of the smoothing constant in Eq. (3.16) was chosen as $\alpha_y = 0.3$. The fixed weighting constants for DD and MDD approaches were chosen as $\beta = 0.98$. As discussed in Section 4, the level γ_u in Eq. (3.19) for the adaptive smoothing factor should be chosen lower for stationary noise when compared with non-stationary noise. As such, for pink noise, white noise, and factory noise, an adaptive smoothing factor is obtained with $\gamma_u = 9$ dB, resulting in $\sigma = -1.391$ and $\rho = 5.159$. For highly variance background noise such as babble noise, the adaptive smoothing factor is obtained with $\gamma_u = 15$ dB resulting in $\sigma = -0.315$ and $\rho = 19.344$ to keep the weighting factor close to 1 during noise frames, which helps to increase the robustness of the a priori SNR estimation against the SNR fluctuations.

3.6.2 Evaluation the effect of the bark scale frequency resolution on the noise characteristics

An extensive experiment is conducted to prove the efficiency of the proposed bark scale based frequency method in eliminating the musical noise. For this experiment, we have fitted Normal (Gaussian) distribution [106] and Weibull distribution [107] to histograms of the noise data before and after the frequency analysis. Figure 3.6 shows comparisons between the histogram and the fitted distributions for different types of noise at frequency 546.87 Hz and the 6th critical band. In the pink and white noise cases, it can be clearly seen that the noise histograms fit well to a Gaussian distribution which is the common assumption in most noise estimation methods. This can help to reduce the musical noise by

Noise Type	Variance before frequency analysis	Variance after frequency analysis
Pink	0.1077	0.0387
White	0.0967	0.0080
Factory	0.1418	0.0789
Babble	0.2535	0.1590

Table 3.2: Noise variance comparison before and after frequency analysis and for different noise types.

reducing the bias and provide a more precise estimate. Whereas in factory noise case and babble noise case, although Gaussian distribution does not really fit the noise distribution after the frequency analysis, it can be seen that the distribution becomes more concentrated with shorter tail compared to the noise distribution before the bark scale transformation.

In order to highlight the ability of the bark scale based processing in reducing the effect of the musical noise, a variance comparison of the noise PDF before and after the bark scale transformation is presented for different noise types as shown in Table 3.2. It can be clearly observed the ability of the bark scale based processing to significantly reduce the noise variance. This helps to reduce the musical noise effect and make it unnoticeable.

3.6.2.1 Evaluation of a priori SNR estimation

Figure 3.8 demonstrates the behaviours of the decision directed (DD), modified decision directed (MDD) and the proposed a priori SNR estimators for 10 dB input SNR and under pink, white, factory and babble background noise conditions, respectively. Speech enhancement is performed by using Wiener filter [90] as shown in the sub-figures on the left side, and MMSE-LSA [52] as shown in the sub-figures on the right side. It is clearly observed that during noise only period, the conventional decision directed and modified decision directed methods provide a smoothed version of the a posteriori SNR. The proposed method has identical behavior as DD and MDD since $\hat{\beta}$ is very close to 1, which is aligned with Eq. (3.19). This explains the ability of the proposed method to eliminate the musical noise. During speech onset, the proposed a priori SNR estimation with different gain functions is responding more quickly to abrupt changes in the a posteriori SNR when compared to DD and MDD a priori SNR estimators. Moreover, it can be observed that both DD and MDD a priori SNR estimation follow the a posteriori SNR with a delay in the speech onset frames which results in a speech distortion, whereas the proposed a priori SNR estimation reduces the

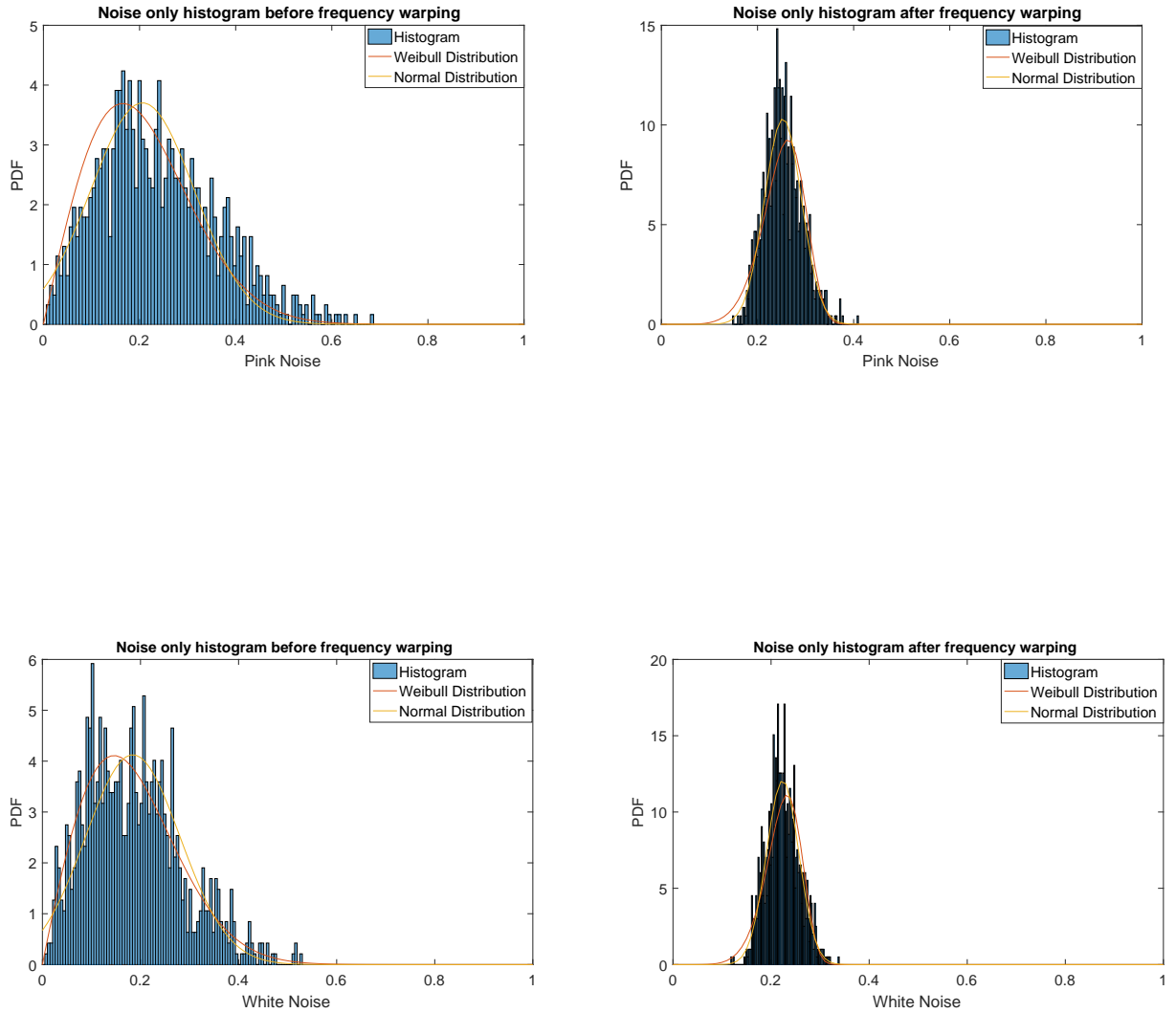


Figure 3.5: Evaluation of bark scale based frequency analysis at 6th critical band under different background noise: 1st row for pink noise and 2nd row for white noise.

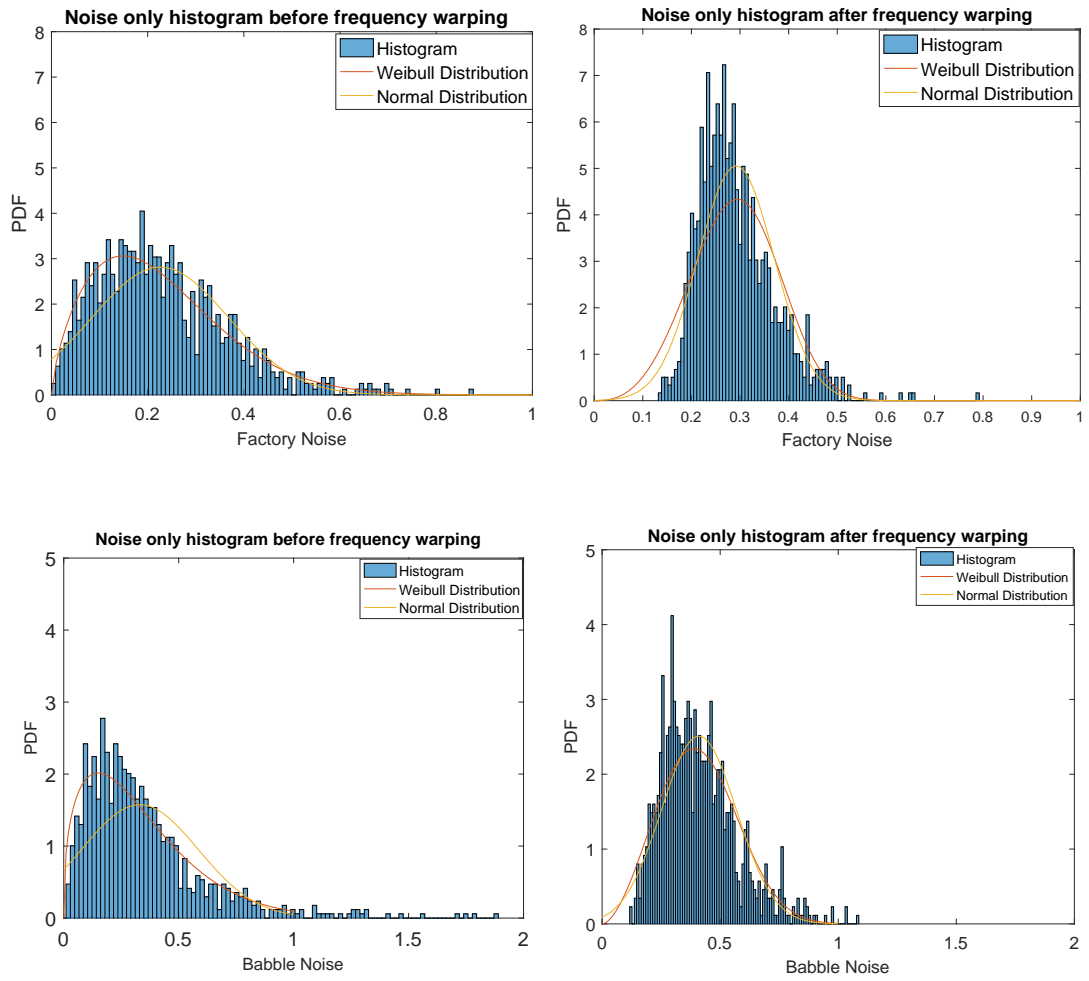


Figure 3.6: Evaluation of bark scale based frequency analysis at 6th critical band under different background noise: 1st row for factory noise and 2nd row for babble noise.

Gain	SNR	PESQ			SNR _{seg}			HD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.8228	1.8174	1.9736	-0.9732	-0.8229	0.2273	0.7679	0.7737	0.7270	1.0430	1.0043	1.0227
	5	2.2936	2.3124	2.3994	1.6672	1.9328	3.0328	0.6633	0.6768	0.6085	1.2102	1.0450	1.1471
	10	2.6543	2.6886	2.7450	4.5326	4.8639	5.9542	0.5192	0.5396	0.4623	1.5797	1.1759	1.4099
LSA	0	1.8461	1.8796	2.0226	-1.0655	-0.6178	0.3294	0.7334	0.7579	0.7082	1.2838	1.0150	1.0604
	5	2.2692	2.3679	2.4450	1.4332	2.1557	3.1458	0.5960	0.6511	0.5803	1.7229	1.0795	1.2286
	10	2.6157	2.7528	2.7999	4.3379	5.1146	6.0884	0.4318	0.5056	0.4312	2.0138	1.2402	1.6022

Table 3.3: Mean objective results for pink noise.

Gain	SNR	PESQ			SNR _{seg}			HD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.4678	1.4472	1.6641	-0.9955	-0.8243	0.4638	0.7912	0.7954	0.7612	1.0313	1.0072	1.0245
	5	2.0679	2.0687	2.2245	2.0611	2.4195	3.7346	0.7066	0.7166	0.6644	1.1361	1.0438	1.1115
	10	2.5211	2.5368	2.6171	5.3278	5.8086	7.2629	0.5940	0.6075	0.5513	1.3542	1.1430	1.3116
LSA	0	1.5311	1.5254	1.7159	-1.0150	-0.5611	0.6021	0.7674	0.7836	0.7478	1.1554	1.0160	1.0480
	5	2.0948	2.1295	2.2749	1.8198	2.6550	3.8684	0.6598	0.6991	0.6442	1.4148	1.0674	1.1583
	10	2.5062	2.5913	2.6604	5.0773	6.0663	7.3952	0.5341	0.5873	0.5290	1.7485	1.1754	1.3757

Table 3.4: Mean objective results for white noise.

delay and preserves more weak speech components.

3.6.2.2 Objective results

The performance of the proposed a priori SNR estimation method is evaluated and compared to the performance of the conventional decision directed DD and modified decision directed MDD methods for different noise types and under various SNR conditions. The clean speech is corrupted by pink, white, factory and babble noise at 0, 5 and 10 dB input SNRs.

Tables 3.3-3.6 show the mean objective results for the stationary background noise cases (pink and white), and non-stationary background noise cases (factory and babble), respectively, with DD, MDD and the proposed a priori SNR estimation methods combined with WF or MMSE-LSA gain functions.

The improvement in terms of speech quality is affirmed by the perceptual

Gain	SNR	PESQ			SNR _{seg}			HD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.8675	1.8796	1.9765	-1.0437	-0.9059	0.1235	0.7577	0.7632	0.7043	1.0631	1.0241	1.2544
	5	2.2878	2.3027	2.3923	1.5133	1.7543	2.6996	0.6590	0.6704	0.5926	1.2117	1.0621	1.3116
	10	2.6551	2.6931	2.7510	4.2072	4.5262	5.5700	0.5109	0.5287	0.4443	1.6213	1.2190	1.6007
LSA	0	1.8678	1.9167	2.0119	-1.1575	-0.7032	0.2029	0.7206	0.7430	0.6885	1.3199	1.0828	1.3544
	5	2.2441	2.3559	2.4268	1.2664	1.9554	2.8143	0.5992	0.6431	0.5753	1.7276	1.1429	1.4545
	10	2.6042	2.7541	2.7886	3.9769	4.7545	5.7060	0.4421	0.4986	0.4303	2.1698	1.3330	1.7427

Table 3.5: Mean objective results for factory noise.

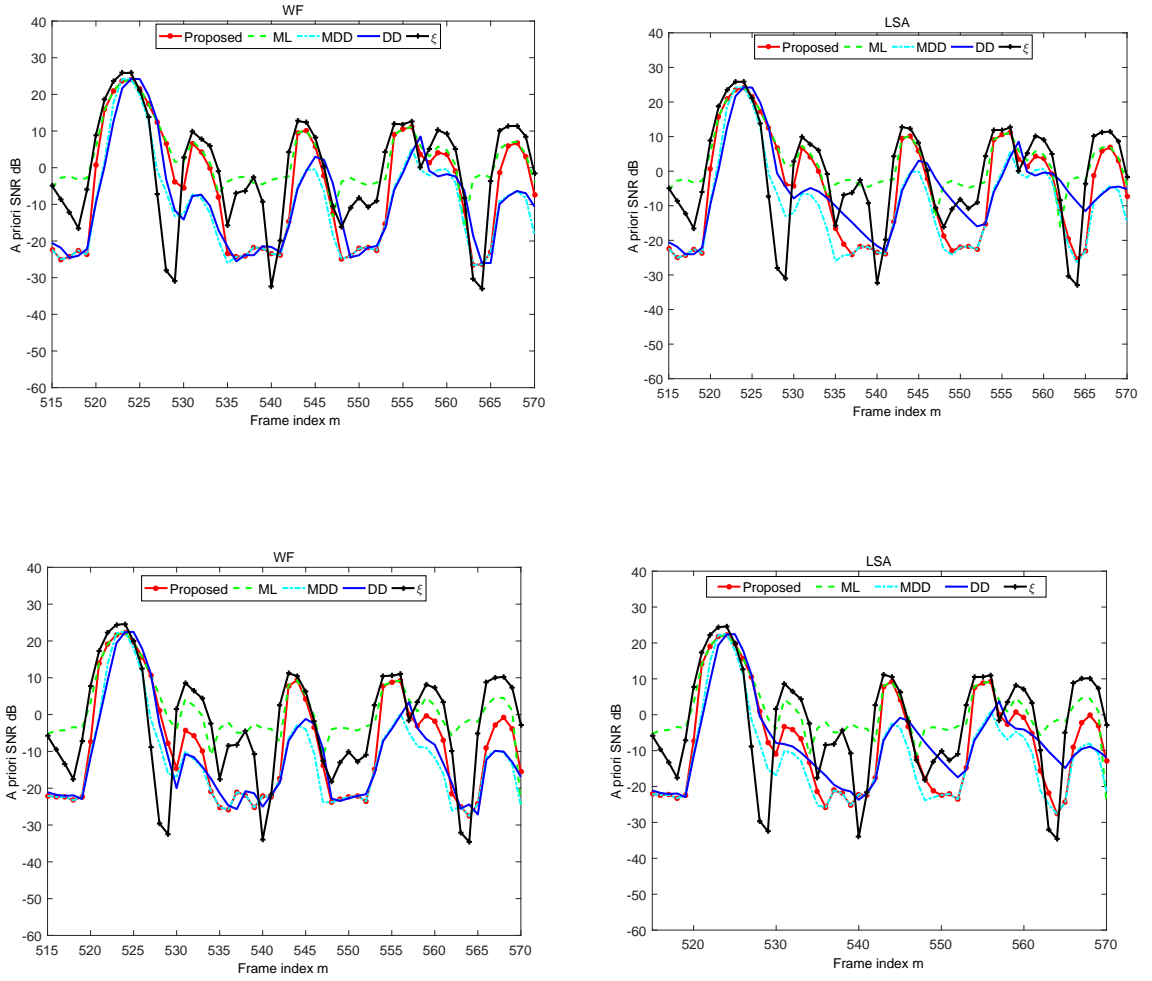


Figure 3.7: Comparison of the a priori SNR estimation over a short time period between the true a priori SNR ξ (black solid line with a marker), ML a priori SNR estimate (green dashed line), $\hat{\xi}_{DD}$ (blue solid line), $\hat{\xi}_{MDD}$ (cyan dot solid line), and $\hat{\xi}_{prop}$ (red solid line with a marker), at 9th critical band and 10 dB SNR under different background noise: 1st row for pink noise and 2nd row for white noise.

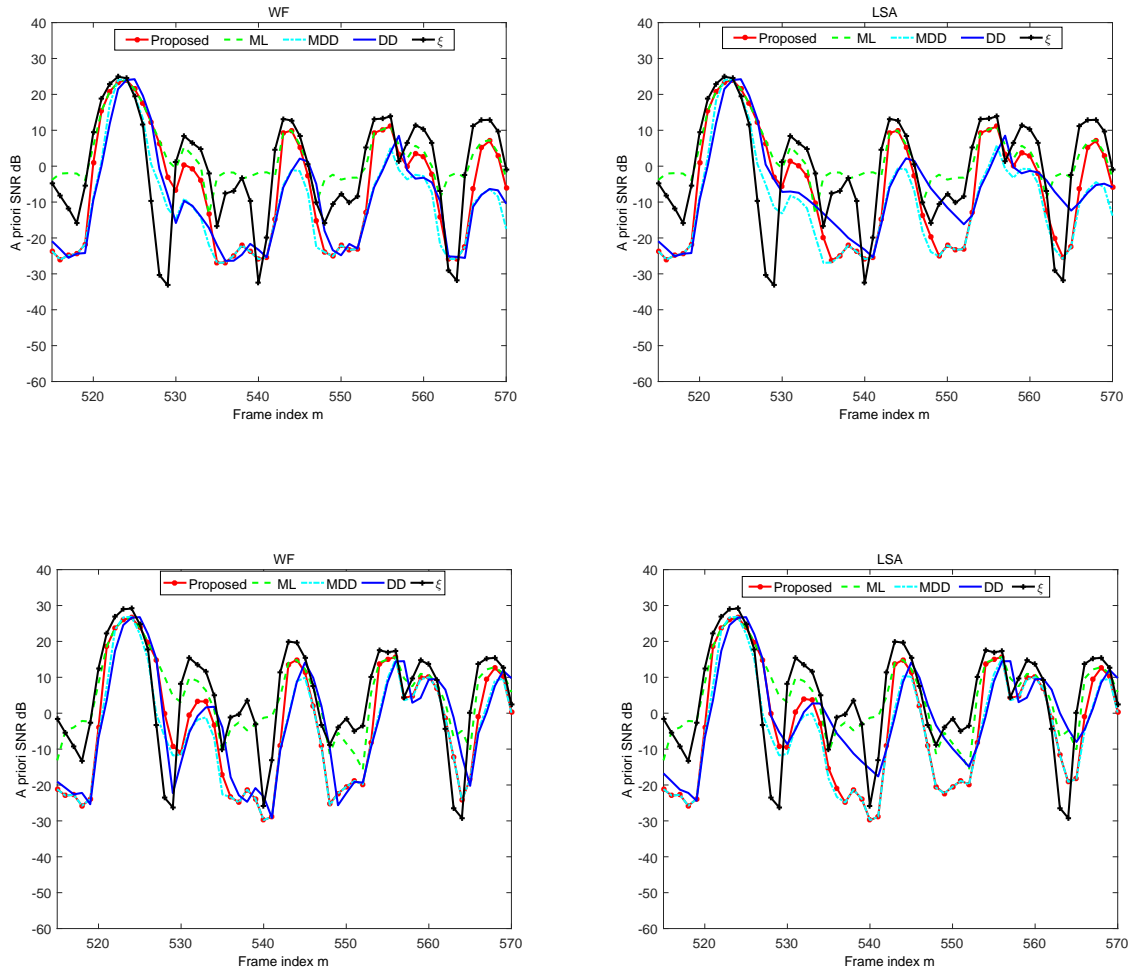


Figure 3.8: Comparison of the a priori SNR estimation over a short time period between the true a priori SNR ξ (black solid line with a marker), ML a priori SNR estimate (green dashed line), $\hat{\xi}_{DD}$ (blue solid line), $\hat{\xi}_{MDD}$ (cyan dot solid line), and $\hat{\xi}_{prop}$ (red solid line with a marker), at 9th critical band and 10 dB SNR under different background noise: 1st row for factory noise and 2nd row for babble noise.

Gain	SNR	PESQ			SNR _{seg}			HD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.9013	1.9099	1.9324	-1.3463	-1.1911	-0.9014	0.6871	0.6976	0.6872	1.1652	1.0721	1.0935
	5	2.2140	2.2498	2.2612	0.9063	1.1306	1.4094	0.5507	0.5684	0.5528	1.4270	1.1932	1.2441
	10	2.5338	2.5810	2.5842	3.4011	3.7187	4.1677	0.3872	0.4065	0.3898	1.8242	1.4378	1.5469
LSA	0	1.8844	1.9413	1.9609	-1.4671	-1.0636	-0.8349	0.6486	0.6811	0.6708	1.3932	1.1319	1.1600
	5	2.1799	2.2825	2.2944	0.6878	1.2848	1.5080	0.4867	0.5491	0.5351	1.7315	1.2674	1.3205
	10	2.4923	2.6254	2.6264	3.1910	3.9562	4.3286	0.3129	0.3901	0.3757	2.0249	1.5164	1.6185

Table 3.6: Mean objective results for babble noise.

evaluation of speech quality PESQ measures. The proposed a priori SNR estimator always results in better speech quality than the conventional decision directed and modified decision directed approaches, indicated by higher PESQ measures. However, in babble noise case, PESQ measures reveal that the proposed estimator achieves approximately same speech quality improvement as MDD approach, while better than the conventional DD approach.

Moreover, the results indicate that the proposed method achieves better noise reduction as it outperforms the conventional DD and MDD approaches in terms of segmental SNR for all noise types and SNR conditions.

Besides speech quality and noise reduction, we also evaluate the weak speech preservation performance of the proposed method. HD measure results indicate that the proposed method delivers better speech preservation in terms of lower HD scores than the conventional DD and MDD approaches, as the proposed technique improved the tracking of onset changes in speech signal. In babble noise case, although it achieves better results than MDD approach, it has slightly higher HD measures than DD approach when combined with LSA gain function at low SNR (<5 dB).

Kurtosis ratio results show the ability of the proposed method to maintain the advantage of DD and MDD methods in reducing the musical noise. Under different types of noise and SNR conditions, the proposed a priori SNR estimation method delivers lower Kurtosis ratio scores than the conventional DD approach. In factory noise case, although the proposed method has slightly higher Kurtosis ratio measures than DD method, it achieves significant improvements in speech quality, noise reduction and speech preservation in terms of better PESQ, segmental SNR and HD measures.

3.6.2.3 Evaluation of listening test

Tables 3.7 and 3.8 demonstrate the average results of the listening test. Ten normal hearing participants in the age of (20-35) took part in this test. They

Gain	Categories	Pink noise		
		DD	MDD	Prop
WF	Speech	3.6	3.9	4.3
	Background noise	3.7	3.8	4.1
	Musical noise	4.0	4.6	4.6
	Over all	3.8	4.1	4.3
LSA	Speech	3.8	4.2	4.3
	Background noise	3.7	4.0	4.1
	Musical noise	3.8	4.5	4.3
	Over all	3.8	4.2	4.3

Table 3.7: Listening test results for pink noise at 10 dB input SNR.

were asked to rate speech signals estimated by three different a priori SNR estimators in terms of speech, background noise and musical noise as explained in the previous section. The speech and background results show that the participants preferred the proposed method compared to either of the DD method or the MDD method and that aligned with the objective results of PESQ and segmental SNR. Moreover, for the musical noise ratings, the proposed method combined with different gain functions and different background noise scored approximately the same as MDD method which is slightly higher than DD method. This means that the proposed method maintains the advantage of the DD approach in generating less musical noise which aligned with the objective measurement Kurtosis ratio.

Furthermore, the overall results of the 10 participants have been evaluated using a statistical analysis to assess the differences between the ratings obtained for each a priori SNR estimation method in term of overall quality. For this purpose, we used analysis of variance (ANOVA) to indicate a significant difference between scores if the level of significance is smaller than 0.05. A significant difference between scores has been noted when LSA was combined with all the different a priori SNR estimation methods with obtained $p=0.03$ and 0.005 for pink noise and babble noise, respectively. Moreover, a significant difference noted when WF was employed with all the different a priori SNR estimation methods in pink noise with obtained p value= 0.01 . However, for the babble noise case, the difference in scores was not found to be statistically significant.

Gain	Categories	Babble noise		
		DD	MDD	Prop
WF	Speech	3.7	4.1	4.1
	Background noise	3.0	3.4	3.7
	Musical noise	3.7	4.2	4.2
	Over all	3.5	3.9	4.0
LSA	Speech	3.8	4.3	4.4
	Background noise	3.0	3.6	3.7
	Musical noise	3.5	4.3	4.2
	Over all	3.4	4.1	4.1

Table 3.8: Listening test results for babble noise at 10 dB input SNR.

3.6.2.4 Spectrograms

Figures 3.9 and 3.16 highlight the ability of the proposed a priori SNR estimator in preserving more weak speech components than the decision directed (DD) and modified decision directed (MDD) a priori SNR estimators under different gain functions. The clean speech signal is corrupted by different background noise (pink noise, white noise, factory noise and babble noise, respectively) with 10 dB SNR. To compare the performance for different gain functions, we employed WF and MMSE-LSA. It can be clearly seen that the proposed a priori SNR estimator preserves more weak speech components than DD and MDD a priori SNR estimators which leads to less speech transient distortion.

3.7 Summary

In this chapter, an adaptive a priori SNR estimator has been developed and evaluated for different speech enhancement gain functions. As a basis for the adaptation, the a priori SNR estimation employs a model of speech absence probability based on a sigmoid function. The sigmoid function can be tuned to provide a trade-off between the speech onset sensitivity and the annoying noise artifacts also known as musical noise. In combination with different gain functions include the Wiener Filter (WF) and the log spectral amplitude (LSA) minimum mean square error estimator, the objective results show that the proposed method outperforms the conventional DD and MDD approaches with higher scores in PESQ, SNR_{seg} and lower scores in HD measures. One of the important findings in these results is that the weaker speech components become more prominent. Moreover, the proposed bark scale based frequency processing helps to reduce the effect of

the musical noise and make it unnoticeable because of the significant reduction in the noise variance which helps in the noise estimation needed for the SNR estimation. The obtained objective evaluation results are supported by the averaged results from the subjective listening tests, as the proposed method was preferred by the listeners.

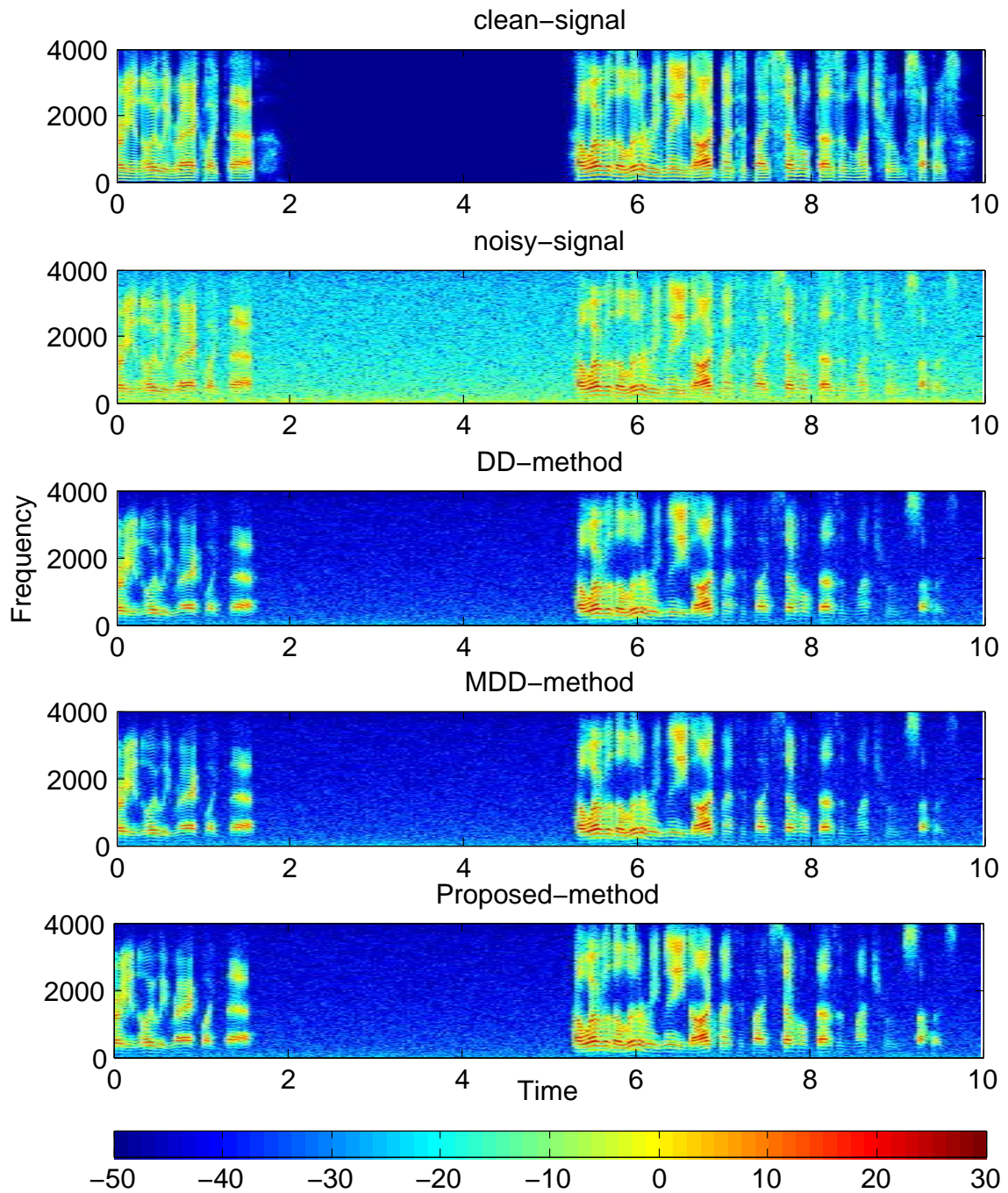


Figure 3.9: Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by Wiener filter speech estimation technique.

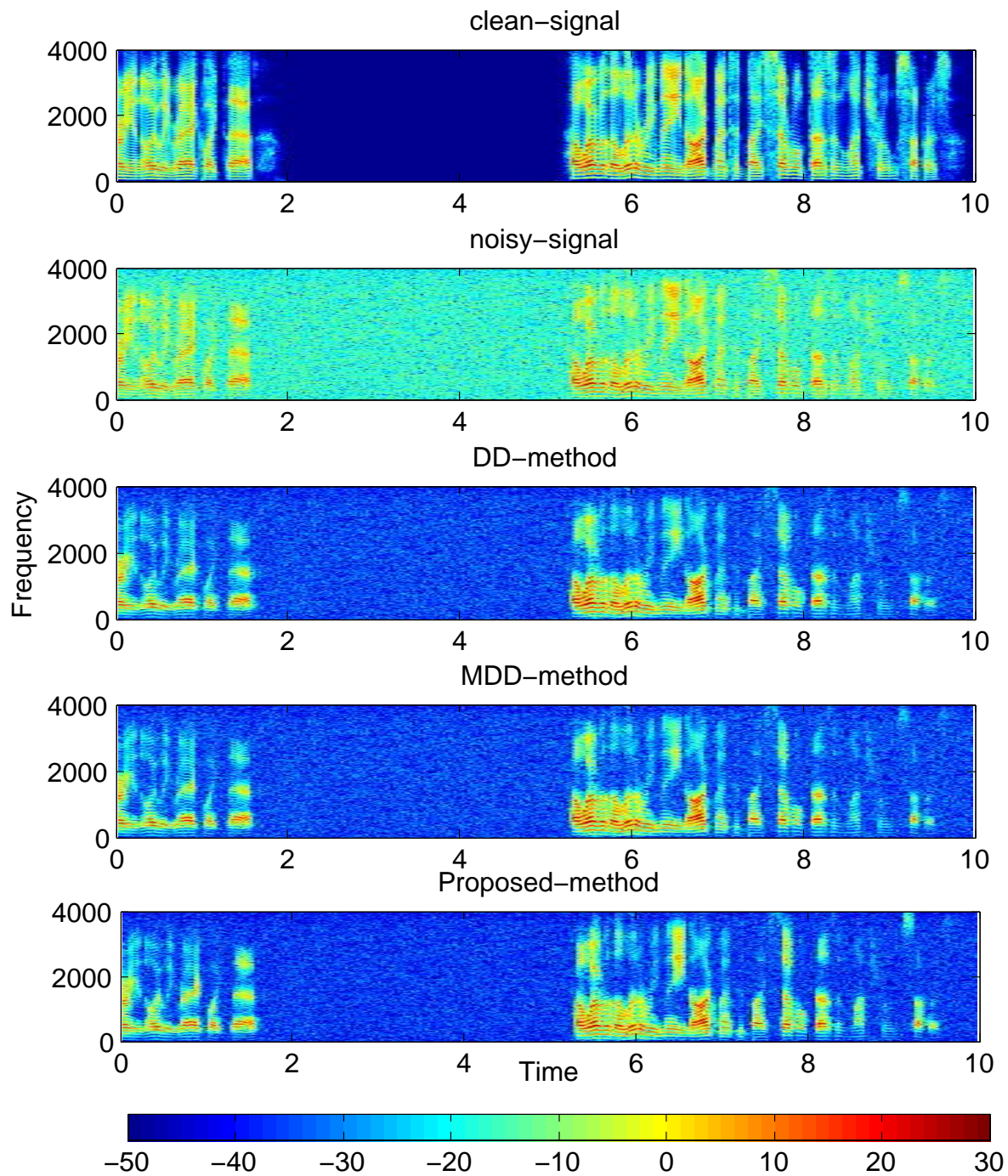


Figure 3.10: Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by Wiener filter speech estimation technique.

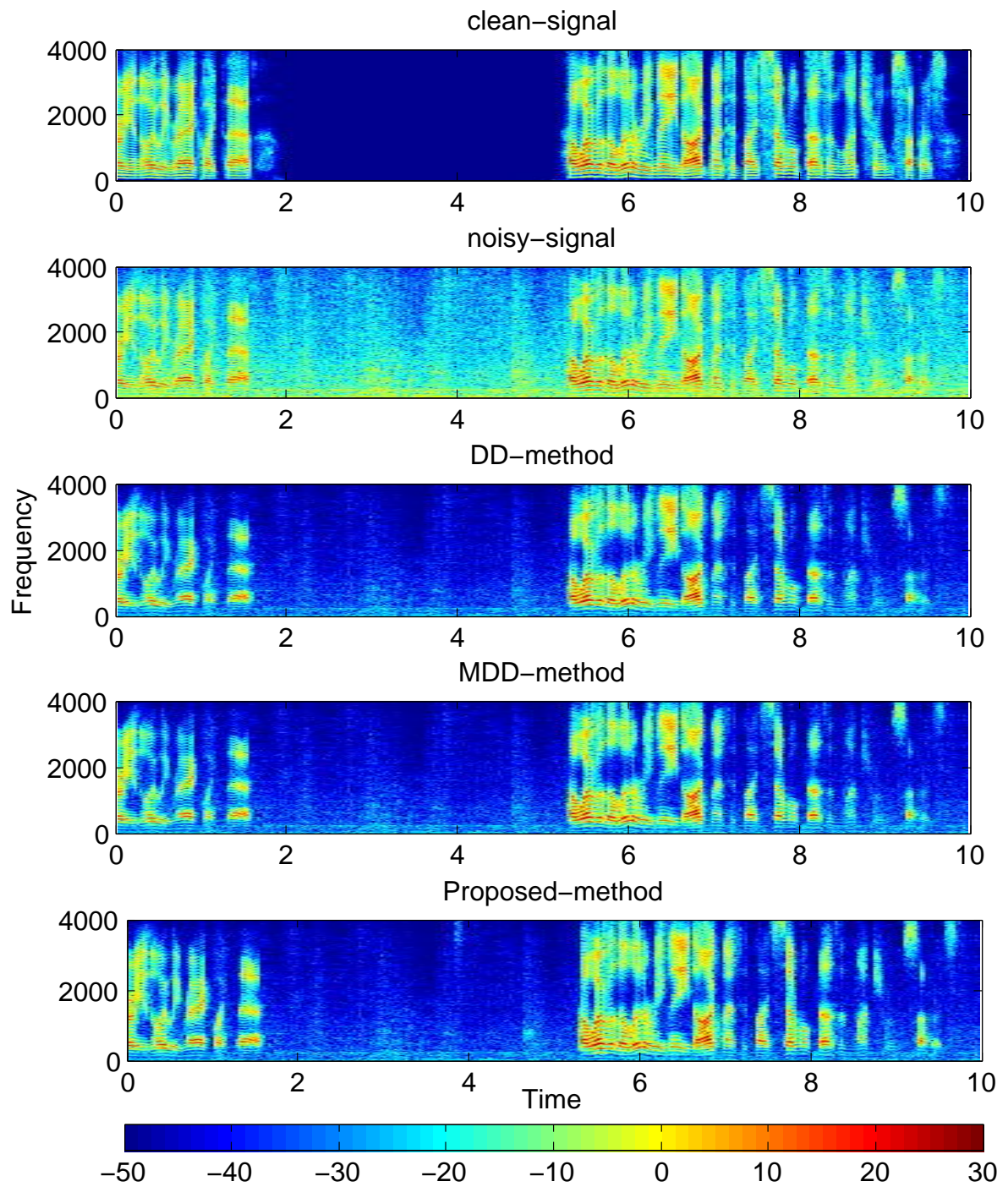


Figure 3.11: Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by Wiener filter speech estimation technique.

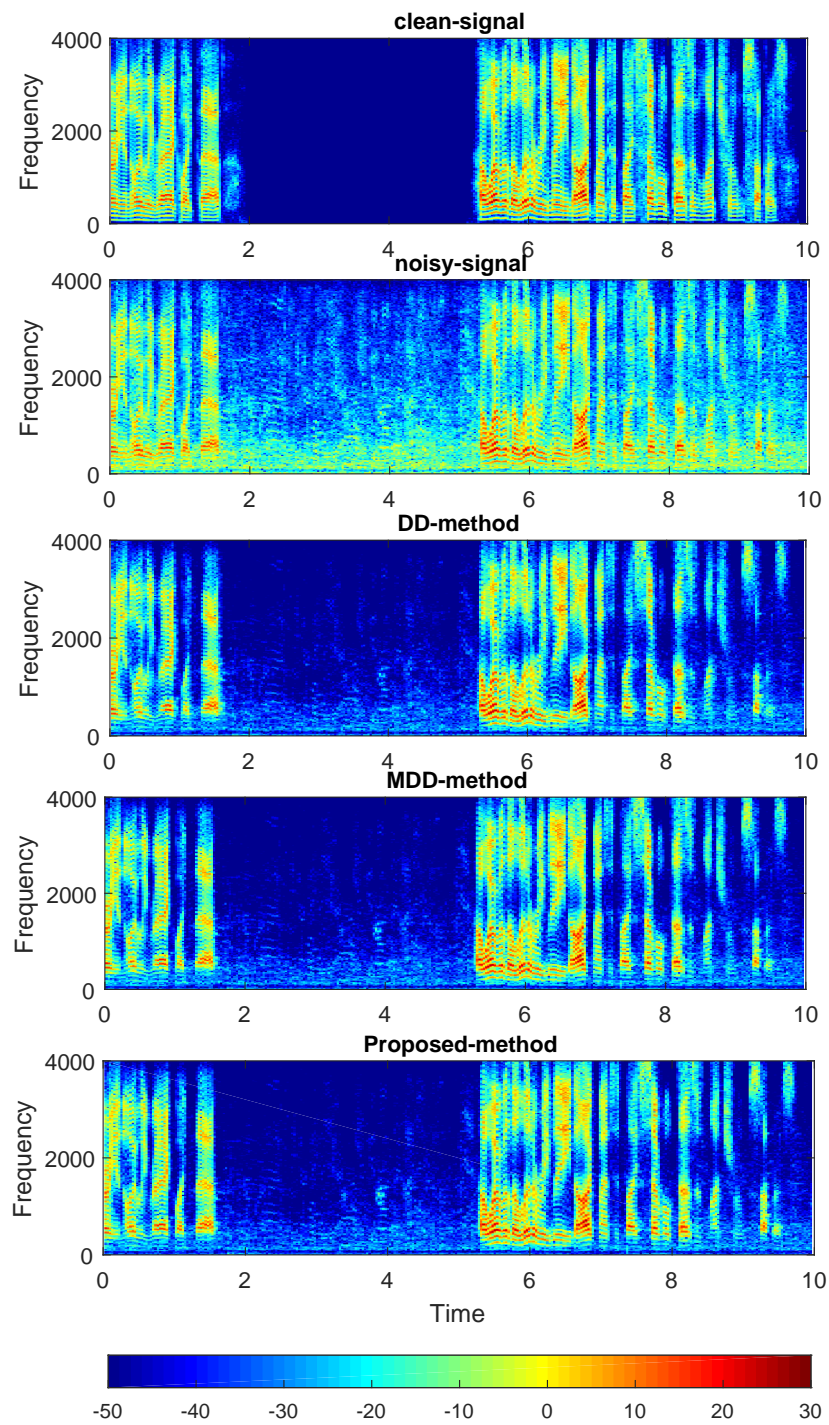


Figure 3.12: Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by Wiener filter speech estimation technique.

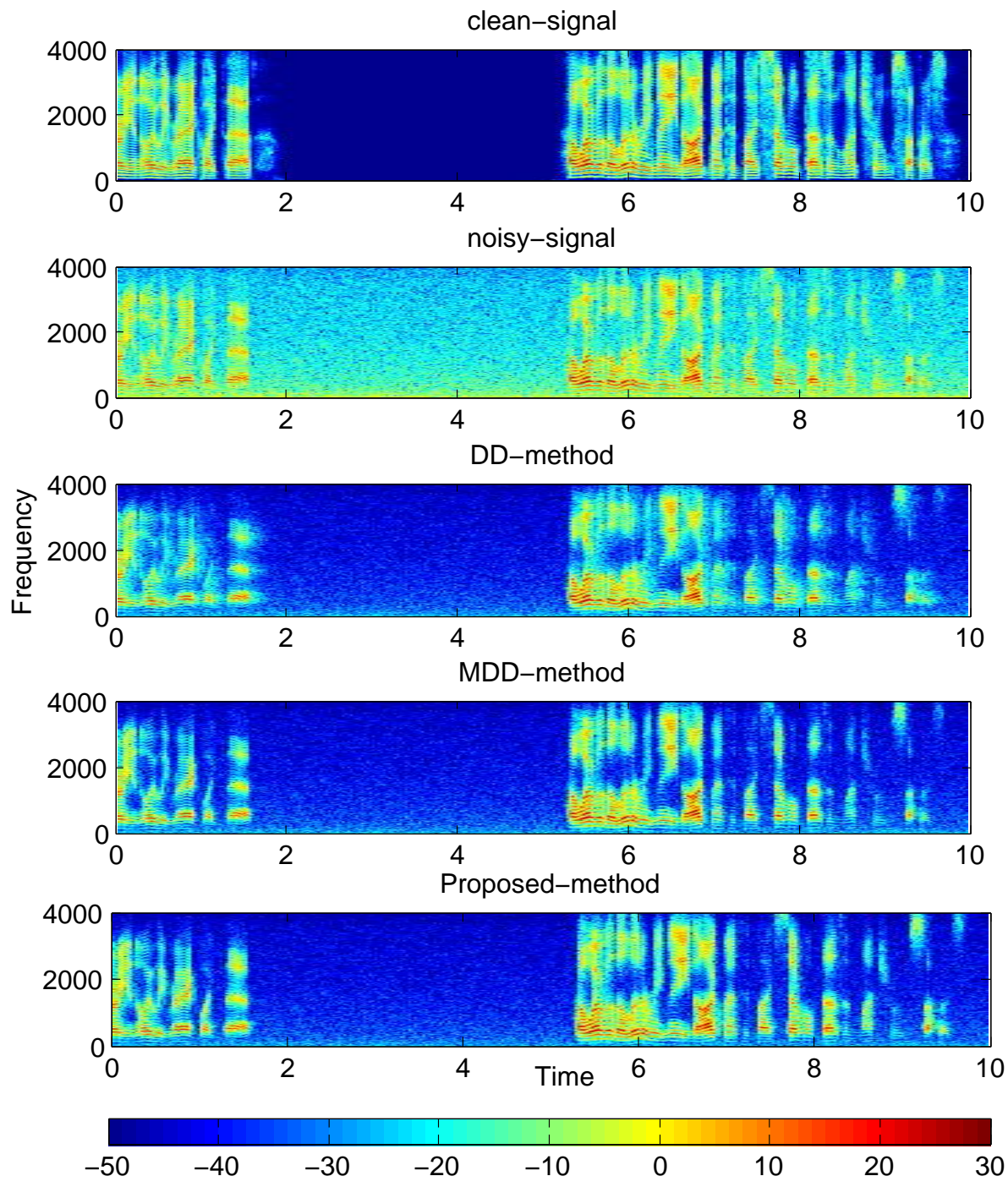


Figure 3.13: Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by LSA speech estimation technique.

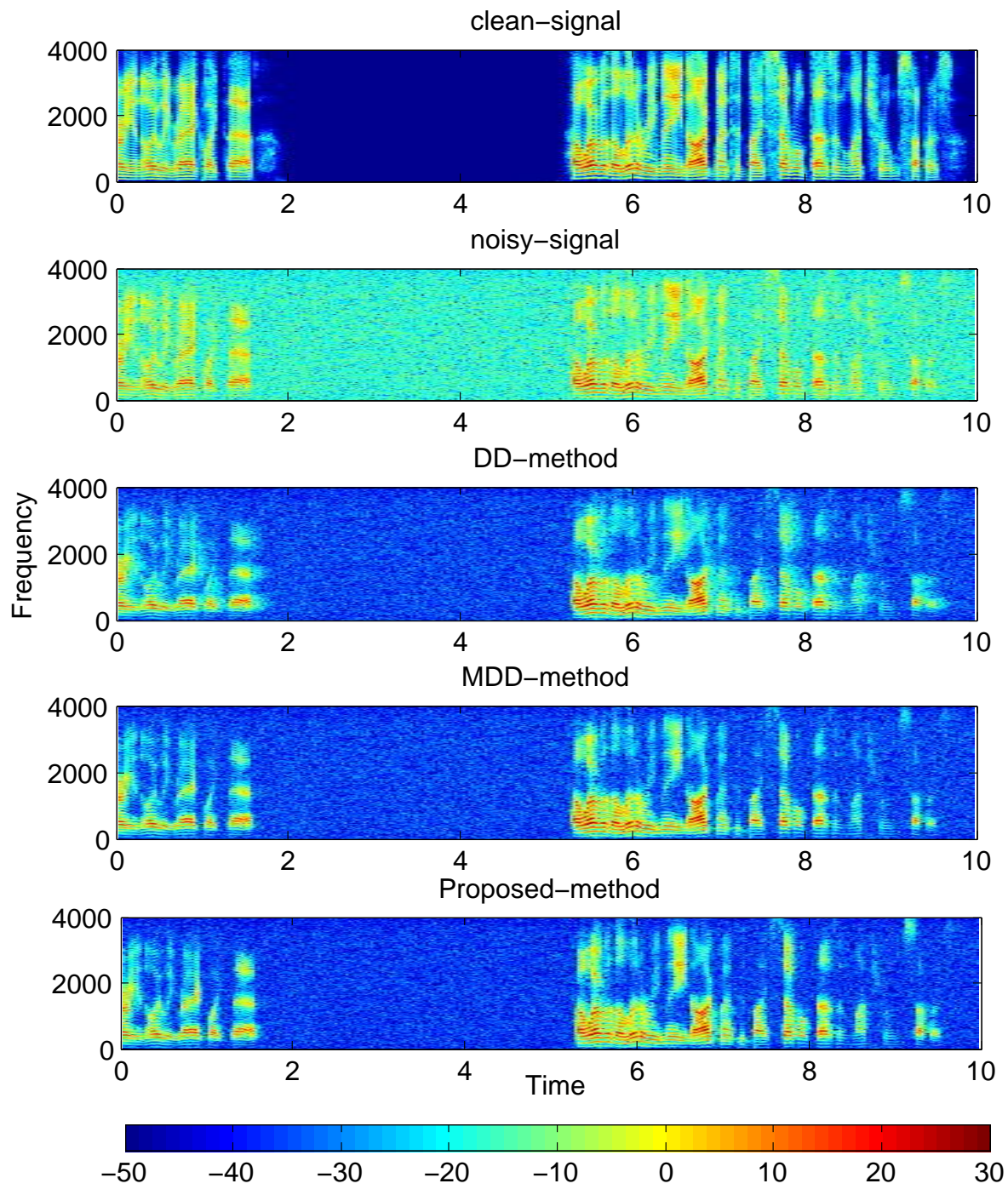


Figure 3.14: Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by LSA speech estimation technique.

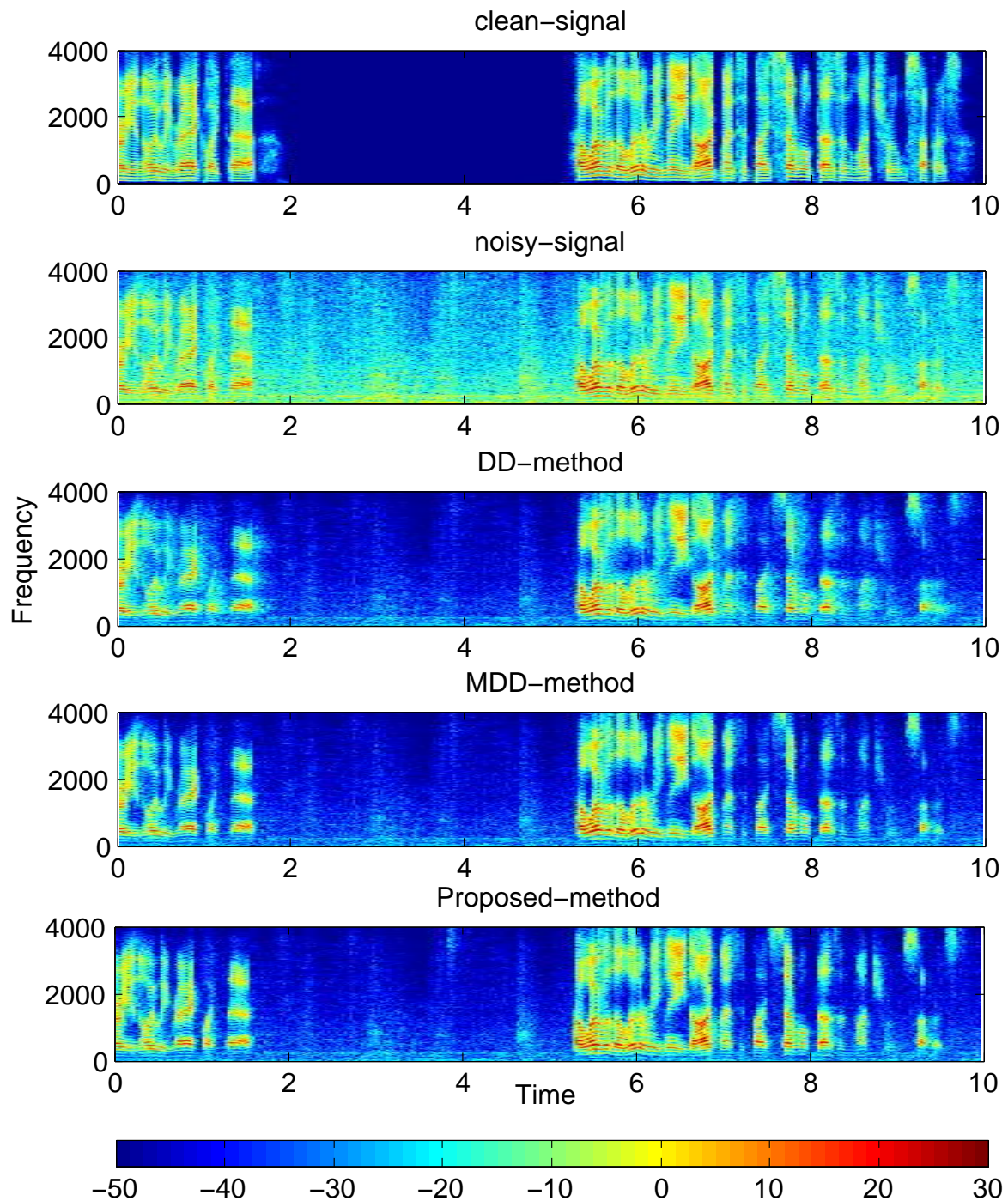


Figure 3.15: Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by LSA speech estimation technique.

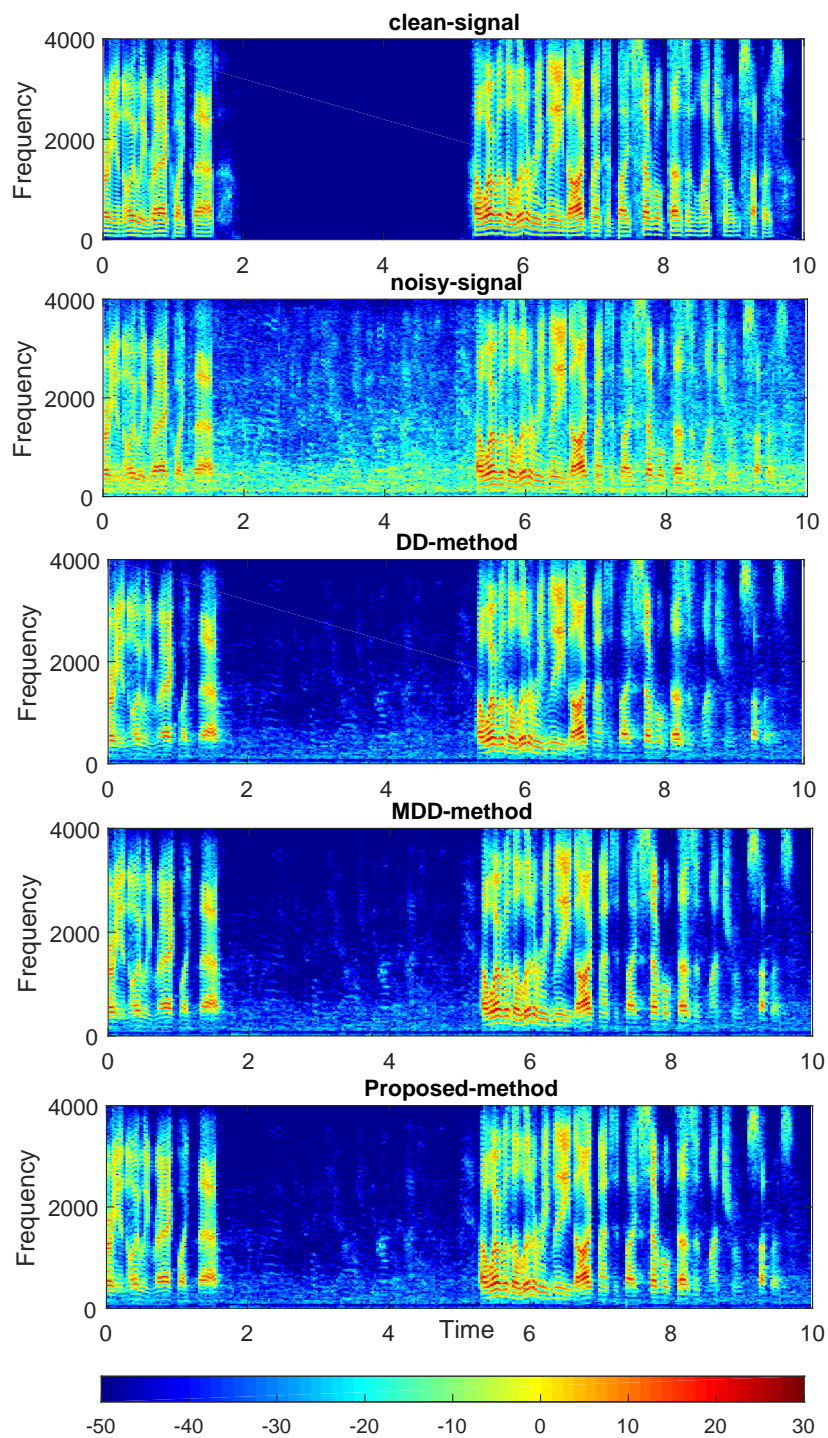


Figure 3.16: Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by LSA speech estimation technique.

Chapter 4

Single Channel Speech Enhancement in STFT Domain

In the previous chapter, we have addressed the problem of single channel speech enhancement in CB domain. This chapter examines the efficiency of the proposed a priori SNR estimation method in STFT domain. The estimation method needs to be modified in order to control the high variability of the a priori SNR estimate. Thus, a modified adaptive smoothing factor with frequency dependent mean parameter is introduced to the MDD method. The advantage of the approach is that weak speech components especially those at higher frequency range can be preserved. Furthermore, a cross comparison between STFT and CB processing has been investigated using a subjective listening test under comprehensive selection of noise conditions.

The main work in this chapter have previously appeared in the following publications:

1. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Improved a priori snr estimation in speech enhancement, 23rd Asia-Pacific Conference on Communications (APCC), 2017, pp. 15.
2. L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Convex combination framework for a priori snr estimation in speech enhancement, submitted to IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) 2019.

4.1 Introduction

The main challenge in single channel speech enhancement is to find the optimal denoising filter to reduce the background noise while preserving the speech components. In other words, the designed filter has to control the trade off between the noise reduction and speech distortion. One of the most important considerations that needs to be taken into account in the filter design is the speech characteristics. It is well known that speech is highly non stationary; thus dividing the degraded noisy signal into short frames is necessary in order to be able to treat the speech signal in each frame as approximately stationary.

Depending on the domain of processing, the denoising filter can be designed either in the time domain or the frequency domain. Although time domain based speech enhancement techniques do not generate notorious artifacts known as musical noise, which are unpleasant to listen to and increase the listener's fatigue, they are more computationally complex since they involve the computation of a matrix inversion. On contrary, frequency domain based speech enhancement overcomes the complexity issue by utilizing Discrete Fourier transform (DFT). STFT is a widely used tool for speech spectral analysis and synthesis due to great efficiency from arithmetic computational complexity point of view and that because it delivers approximately uncorrelated transform coefficients. Moreover, the resulting statistically independent superposition coefficients lead to a straightforward interpretation in terms of spectral signal content.

The main task of spectral speech enhancement is to apply a denoising spectral gain function to the noisy speech spectrum. The major issue in such approach is the large variations of the spectral coefficients in the noisy frames. These variations can lead to spectral outliers in the filter gain adaptation and result in what is known as musical noise. The a priori SNR is the main quantity estimate when determining the spectral gain function. Thus, improving the a priori SNR estimate is one of the most important tasks for practical solutions. Ephraim and Malah in [52] proposed decision directed (DD) based a priori SNR estimation which is defined as a weighted combination of the a priori SNR estimation in the previous frame and the maximum likelihood (ML) estimation of the a priori SNR in the current frame [50]. This approach helps to reduce the variance of the a priori SNR estimate and as a result, the musical noise is significantly reduced [87]. Since this estimation uses the a priori SNR estimate from the preceding frame, there is a one frame delay between the estimated and the true a priori SNR. For this reason, speech transient distortion occurs which degrades the quality of the

estimated speech signal. Also, the weight in the DD estimator is controlled by a smoothing factor that is usually a value close to 1 in order to avoid the musical noise, which results in slow tracking of abrupt changes in the instantaneous SNR.

Many techniques have been proposed to overcome the drawbacks in the DD approach [88], [108], [109], [49], [10] and [110]. Hasan et. al. [108] employs a self-adaptive smoothing factor to estimate the a priori SNR. As a consequence, the adaptation speed of the a priori SNR estimation is improved, but with the one frame delay problem. Yong et. al. [10] proposed a modified DD approach (MDD) by matching the ML a priori SNR estimate with the noisy speech spectrum in the current frame. The advantage of this approach is its ability to reduce the one frame delay in the a priori SNR estimate. However, the slow tracking speed for speech onsets and offsets remain since it still uses a smoothing factor close to 1.

In order to overcome this problem, in the preceding chapter we proposed an adaptive smoothing factor to the modified a priori SNR estimation method. A time-frequency dependent weighting factor with a sigmoid shape is proposed to control the smoothing factor. Accordingly, flexible smoothing of the a priori SNR estimation is obtained to improve the tracking mechanism to abrupt changes in instantaneous SNR. In conjunction with that, we utilize a critical band mapping from STFT analysis-resynthesis system in the speech enhancement framework for human perceptual processing and lower complexity. The proposed method does not only eliminate the one frame delay generated by the well-known decision directed approach but also increases the adaptation speed during abrupt changes in the SNR estimation. This helps to preserve the weak speech components while the advantage of low musical noise has been maintained.

Since most speech enhancement methods are performed in STFT domain, we are investigating the proposed method in STFT domain. When processing the noisy speech in STFT domain, we have a higher variability in the SNR estimation. To improve that aspect, a modified adaptive smoothing factor with a frequency dependent mean parameter is proposed to the MDD method. Objective results show the ability of the proposed method to preserve weak speech components in the high frequency region while maintain the same overall speech quality as the MDD method. Furthermore, we present a cross evaluation for STFT and CB processings using listening test and under different noise conditions. This chapter is organized as follows. In section 4.2, a review of the single channel speech enhancement is presented. Section 4.3 presents the proposed a priori SNR estimation approach. Section 4.5 demonstrates the results of the experimental evaluation. Section 4.4 presents a cross evaluation for STFT and CB processings.

Section 4.6 concludes the chapter.

4.2 Single channel speech enhancement

Let $y(t)$ be denoting the noisy signal in the discrete time domain, which consists of the clean speech signal $s(t)$ and additive noise signal $v(t)$, as given by

$$y(t) = s(t) + v(t) \quad (4.1)$$

where speech and noise signals are assumed to be uncorrelated. The noisy signal is then sampled and transformed into the frequency domain by using short time Fourier transform (STFT). The noisy spectrum is defined by

$$Y(k, m) = S(k, m) + V(k, m) \quad (4.2)$$

where $S(k, m)$ and $V(k, m)$ represent the spectral components of speech signal and noise at the frequency bin k and the time frame index m , respectively.

The estimate of the clean speech is obtained by applying a spectral gain function $G(k, m)$ to the noisy spectrum. In general, the spectral gain function depends on the a priori SNR $\xi(k, m)$ and/or the a posteriori SNR $\gamma(k, m)$ which are defined as

$$\xi(k, m) = \frac{\lambda_s(k, m)}{\lambda_v(k, m)} \quad (4.3)$$

and

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

where $\lambda_s(k, m)$ and $\lambda_v(k, m)$ denote the clean speech PSD and noise PSD, respectively.

In this chapter and for comparison reason, we have chosen the same gain functions as in the previous chapter, i.e., Wiener filter (WF) gain function [90], which is defined by

$$G_{\text{WF}}(k, m) = \frac{\xi(k, m)}{1 + \xi(k, m)} \quad (4.5)$$

and MMSE-LSA [52], which is obtained by minimizing the mean square error of the logarithm of original and enhanced speech spectra, and can be defined as a function of the priori SNR and the posteriori SNR, given by

$$G_{\text{LSA}}(k, m) = \frac{\xi(k, m)}{1 + \xi(k, m)} \exp \left\{ \frac{1}{2} \int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt \right\} \quad (4.6)$$

where the lower limit ν_k of the integral is given by

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

and $\gamma(k, m)$ denotes a posteriori SNR defined as

$$\gamma(k, m) = \frac{|Y(k, m)|^2}{\lambda_v(k, m)}. \quad (4.8)$$

Then the estimated speech signal is obtained as

$$\hat{S}(k, m) = G(k, m)Y(k, m). \quad (4.9)$$

Since both clean speech and noise are unknown and only a noisy signal is accessible, both a priori SNR and a posteriori SNR have to be estimated.

4.3 Proposed a priori SNR estimation

As mentioned in the previous chapter, the smoothing factor for the DD based a priori SNR estimation plays an important role in controlling the trade-off between the musical noise, speech distortion and noise reduction in the enhanced speech signals. Hence, using a constant smoothing factor is not ideal, since using a factor close to one which will reduce the musical noise, but will lead to speech transient distortion during the speech frames [3].

In order to reduce the transient distortion as well as preserve weak speech components, an adaptive smoothing factor is proposed for the MDD approach to increase the speed of tracking during speech onsets. This is done by utilizing a time-frequency varying smoothing factor based on a sigmoid function [111]. The proposed adaptive smoothing factor is defined as given below

$$\hat{\beta}(k, m) = \frac{\rho}{1 + \exp[-a(\hat{\gamma}(k, m) - \sigma(k))]} \quad (4.10)$$

where $0 \leq \hat{\beta}(k, m) \leq 0.98$. a and $\sigma(k)$ denote the slope and the mean of the sigmoid function, respectively. ρ represents the parameter that control the upper limit of the weighting factor in order to retain similar property as a constant weighting factor

for the noise only frames.

In order to preserve speech components especially those at higher frequency range, an adaptive mean factor $\sigma(k)$ is proposed instead of a constant value. A quadratic function in vertex form is used to control the sensitivity of the mean

factor with respect to the increment of the frequency bins. $\sigma(k)$ can be defined as

$$\sigma(k) = \nu * (k - \delta)^2 + \kappa \quad (4.11)$$

where δ and κ are the vertex of the parabola which control the x-axis and y-axis shift, respectively. ν denotes a scale that controls the maximum limit for the varying mean. The values of $\sigma(k)$ for frequency bins after δ are set to κ to preserve weak speech components at high frequency range. Fig. 4.1 demonstrates the behavior of the adaptive mean parameter. It shows that $\sigma(k)$ is inversely proportional to the increment of the frequency bin, and becomes constant after δ . It indicates that the sensitivity of the smoothing factor will be increased at higher frequencies. The modified a priori SNR estimation approach is then given by

$$\begin{aligned} \hat{\xi}_{\text{prop}}(k, m) = & \hat{\beta}(k, m) \frac{|G(k, m - 1)Y(k, m)|^2}{\hat{\lambda}_v(k, m)} \\ & + (1 - \hat{\beta}(k, m))P [\tilde{\gamma}(k, m) - 1]. \end{aligned} \quad (4.12)$$

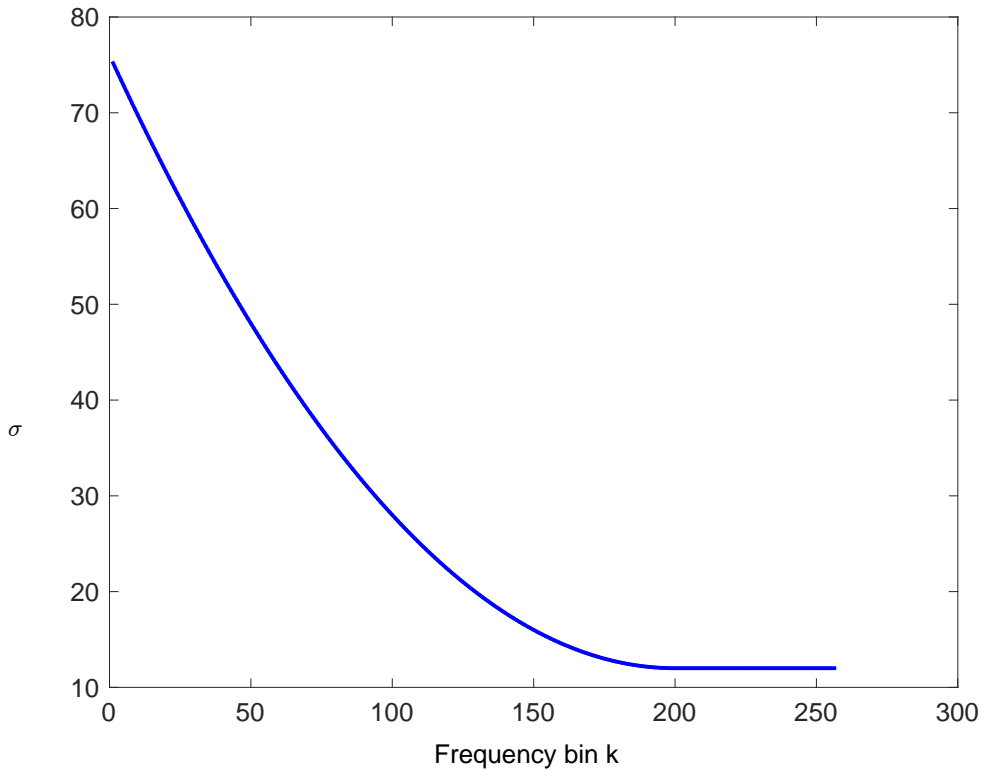


Figure 4.1: Adaptive mean factor as a function of frequency bins.

Fig. 4.2 shows the proposed adaptive smoothing factor $\hat{\beta}(k, m)$ as a function of a posteriori SNR. It can be clearly seen that the smoothing factor attains different values depending on the a posteriori SNR instead of a constant close to one. This means that the a priori SNR estimation has become more flexible. For instance, when $\hat{\gamma}$ is less than 0 dB, a priori SNR estimation corresponds to a smoothed version of the a priori SNR in the same way as the MDD does. In this case, the proposed method has similar noise suppression and musical noise level as MDD for noise-only frames. Meanwhile, when $\hat{\gamma}$ is larger than 0 dB, a priori SNR estimation is expressed as a weighted sum of the amplitude speech estimation and the a posteriori SNR. Moreover, the frequency varying mean parameter plays an important role in controlling the sensitivity of the smoothing factor which improves the adaptation speed of the proposed estimation. This yields better tracking of speech onsets especially at higher frequencies, which helps to avoid unwanted attenuation of weak speech components.

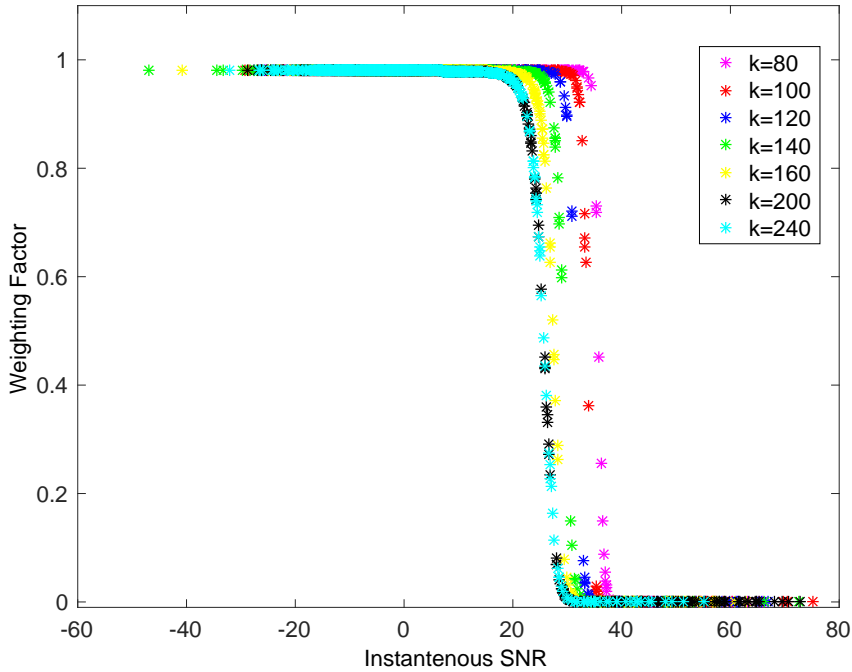


Figure 4.2: Adaptive smoothing factor as a function of instantaneous SNR at different frequency bins.

4.4 Cross subjective evaluation

In order to test the efficiency of the critical band based frequency scale compared to the conventional uniform scale (STFT), we present a cross evaluation of

a priori SNR estimators integrated with different time-frequency analysis techniques (STFT and CB) by a subjective listening test according to ITU-T recommendation P.835 and conducted by ten participants. Three speech sentences consisting of 1 female speaker and 2 male speakers from the NOISEUS database [3] have been concatenated and corrupted with two different background noise (babble or pink) from NOISEX-92 database [112] for two levels of input SNR (0 dB and 10 dB). The noisy speech signals have been processed using three decision directed based a priori SNR estimation methods. The evaluated estimation methods include DD method [113] processed either with STFT, denoted by (DD-STFT) or processed with CB, denoted by (DD-CB). MDD method [10] processed either with STFT, denoted by (MDD-STFT) or processed with CB, denoted by (MDD-CB), critical band based a priori SNR estimation method (Proposed-CB), which is presented in the previous chapter [111] and STFT based a priori SNR estimation method (Proposed-STFT) [114] that is presented in this chapter. In order to estimate the enhanced speech signals, the aforementioned a priori SNR estimation methods are combined with either WF or MMSE-LSA gain function.

The listening test was performed in a tranquil office room utilizing DT- 880 Beyerdynamic open air headphone. In order to reduce the time consuming, the subjects were divided into two groups, one group to rate the enhanced speech signals using WF gain function, while the other group rated the enhanced speech signals using MMSE-LSA gain function. The test was lasting around 25 minutes for each participant. Prior to giving their scores on the processed speech signals, the listeners were presented with the clean speech signal and the unprocessed speech signal as a kind of perspective for the best case and the worst case, individually. After that the participants were asked to listen and rate the enhanced signals according to ITU-T recommendation P.835. This methodology guides the participants to form the basis of their ratings regarding speech signal alone, background noise, musical noise and overall quality as shown in Table 3.1 in the previous chapter.

Furthermore, to assess the difference between the listening test ratings, a statistical analysis of variance (ANOVA) is conducted to present a comparative analysis in reference to the unprocessed speech signal. A significant difference between scores was recognized depending on the obtained significance level (p -value). \mathcal{H} represents the equality hypothesis and is defined as follows

$$\mathcal{H} = \begin{cases} \text{No significant difference is recognized, } p > 0.05 \\ \text{Significant difference is recognized, } p < 0.05 \end{cases} \quad (4.13)$$

which means if $p > 0.05$, equality hypothesis is accepted. Otherwise, the equality hypothesis is rejected.

4.5 Experimental evaluation

In this section, the performance of the proposed a priori SNR estimation approach was evaluated and compared with DD and MDD approaches. Based on the equations (3.23) and (3.24) in the previous chapter, the slope and the lower limit of the mean are calculated at $\gamma_u = 15$ dB for pink noise case and $\gamma_u = 18$ dB for babble noise case, which result in $\alpha = -0.4$ and $\kappa = 19.344$ and $\alpha = -0.11$ and $\kappa = 38.194$, respectively. ρ is chosen to be 0.983. From the speech enhancement experiments, the parameters for the adaptive mean factor in Eq. (4.11) were set to the following values: $\nu = 0.0016$, $\delta = 200$. The value of β in DD and MDD approaches was set to 0.98. The noise PSD estimation was obtained by using the minimum mean square error (MMSE) noise power estimator based on the speech presence probability [105] for all the a priori SNR estimators. A sampling frequency of $f_s = 8000$ Hz with $K = 512$ was used, and a square root Hanning window with 50% overlapping was applied. The estimated signal is reconstructed using overlap-add method.

4.5.1 Evaluation of a priori SNR estimation

Figures 4.3- 4.6 show comparisons of different a priori SNR estimators evaluated in connection with different gain functions and under different noise conditions. Figures on the left side depict the a priori SNR estimation when combined with Wiener filter gain function, whereas figures on right side depict the a priori SNR estimation when combined with MMSE-LSA gain function. It can be clearly observed that the smoothing property of the decision directed based a priori SNR estimators are depending on the gain function as well as the smoothing factor. For MMSE-LSA gain function case, it can be demonstrated that the decision directed (DD) a priori SNR estimate exhibit significant smoothing characteristics during noise frames compared to the MDD and the proposed methods. That explains the decision directed methods ability to eliminate musical noise better than the other two methods. However, a priori SNR estimation methods in connection with Wiener filter gain function show an improved smoothing characteristics for the proposed method. It can be clearly observed that during noise-only frames the proposed method follows the a posteriori SNR with lower values similar to the patterns from both DD and MDD approaches which helps to reduce the musical noise. Whilst during speech onsets, as the value of the adaptive smoothing factor decreases, the proposed method follows the a posteriori SNR with less delay

compared to DD and MDD methods. As a consequence, less speech transient distortion can be achieved.

4.5.2 Objective results

The proposed method was also evaluated by using four different objective measures, namely the kurtosis ratio that evaluates the amount of musical noise generated [99], the segmental SNR measures [97], the cepstral distance based on Linear predictive coding (LPC) for speech transient distortion measure and the perceptual evaluation of speech quality (PESQ) measure [3]. A better enhanced speech is indicated by lower scores in kurtosis ratio and cepstral distance, and higher scores in segmental SNR and PESQ. The evaluation was performed by using 30 different utterances from six different speakers (3 males and 3 females) corrupted by pink noise, white noise, factory noise and babble noise under different SNR conditions. Both speech signals and noise are taken from NOISEUS and NOISEX database, respectively. The proposed method were compared with DD and MDD methods.

Tables 4.1-4.4 show the mean performances of DD, MDD and the proposed methods in terms of the four objective measures. It can be seen that at low SNRs, the performance of the proposed method is similar to the MDD approach, since the value of the adaptive smoothing factor is almost 0.98. As a result, the proposed method generates less musical noise when compared to the DD method due to its ability to reduce the sensitivity of the a priori SNR estimation during noise frames. During speech onsets, the adaptive smoothing factor attains smaller values, which improves the tracking speed of the a priori SNR estimation as already shown in Figures 4.3-4.6. This also helps to preserve the weak speech components better than the DD and MDD methods as can be found in the cepstral distance results. However, due to the increased sensitivity of the a priori SNR estimation towards the abrupt changes in the instantaneous SNRs, a slight increase in the musical noise is obtained when compared to the MDD approach. This can also be reflected in terms of the segmental SNR results, where the proposed method records a consistent improvement in SNR compared to both the DD and the MDD approaches under different noise conditions. The PESQ results show that the overall speech quality produced by the proposed method is slightly higher than the DD approach and approximately similar to the MDD approach. However, in babble noise case as shown in Table 4.4, although the proposed method outperforms the other methods in terms of better segmental SNR and higher speech quality, it can be observed that it prone to generate higher musical noise

due to its higher sensitivity to abrupt changes in SNR. Moreover, comparing between the performance of WF and MMSE-LSA gain functions, it can be clearly noted that less musical noise is obtained for the DD method combined with the MMSE-LSA gain function since it exhibits higher smoothing ability. However, this comes at cost on speech quality and noise reduction.

4.5.3 Spectrograms

Figures 4.7-4.14 depict the spectrogram of the enhanced speech signal by using different a priori SNR estimators in different noise scenarios at 10 dB. It can be observed that the ability of the proposed method to preserve speech components is better than the DD and MDD methods, especially in the high frequencies. At the same time, the background noise is effectively reduced.

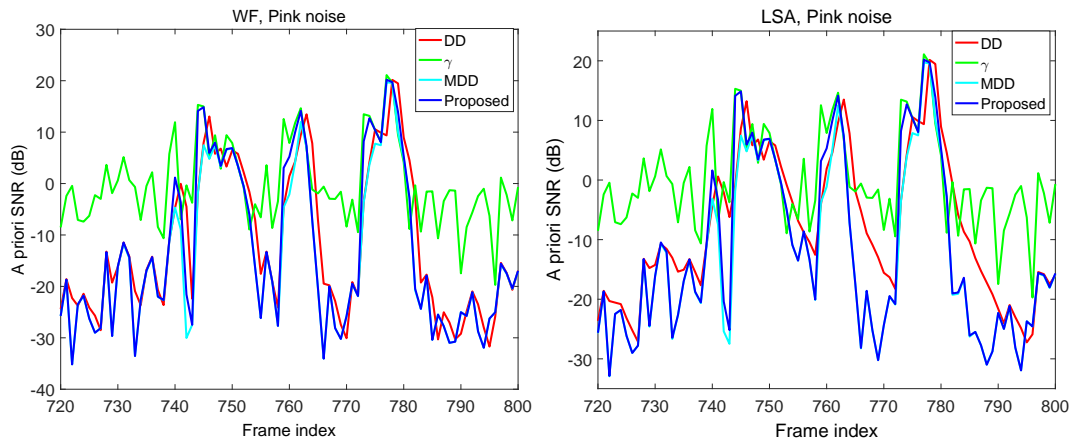


Figure 4.3: Comparison of a priori SNR estimation over short time between Υ (green solid line), $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB pink background noise.

4.5.4 Evaluation of listening test

Figures 4.15 and 4.17 depict the results of the listening test in pink background noise. Figure 4.15 shows the results when WF gain function was combined to different a priori SNR estimation methods for different levels of input SNR, whereas Figure 4.17 shows the results when MMSE-LSA gain function was employed instead. In terms of speech signal, listening test results for WF gain function case show that the participants preferred STFT enhanced speech signals over CB in

Gain function	Input SNR	PESQ			SNR _{seg}			CD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	2.3658	2.3623	2.3449	2.0960	2.3660	2.5946	5.1139	5.1322	5.1182	2.1820	1.9584	2.1741
	5	2.6711	2.6843	2.6563	4.7877	5.0471	5.3099	4.8057	4.8452	4.7251	2.6031	2.1411	2.3904
	10	2.9341	2.9604	2.9351	7.5057	7.7525	8.1430	4.4512	4.5024	4.3050	3.3627	2.5499	2.8891
LSA	0	2.2676	2.3650	2.3671	1.2054	2.0642	2.2619	4.9393	4.9296	4.9142	1.7315	2.2623	2.3911
	5	2.5961	2.7116	2.7014	4.1715	5.0503	5.2937	4.5245	4.5082	4.4184	2.0758	2.4125	2.5623
	10	2.8606	2.9961	2.9824	7.2672	8.0658	8.3820	4.1022	4.0898	3.9435	2.4719	2.6884	2.8849

Table 4.1: Mean objective results for pink noise.

Gain function	Input SNR	PESQ			SNR _{seg}			CD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	2.1961	2.1898	2.1924	2.4145	2.7077	2.9573	6.0770	6.1530	6.1862	1.8689	1.7216	1.8834
	5	2.5201	2.5279	2.5116	4.9180	5.1936	5.4888	5.6564	5.7395	5.6934	2.0496	1.7835	1.9607
	10	2.7929	2.8013	2.7997	7.4647	7.7047	8.1100	5.2932	5.3883	5.2507	2.4550	1.9737	2.1917
LSA	0	2.0545	2.1442	2.1636	1.4098	2.2833	2.5019	6.0655	6.1303	6.1499	1.4413	2.1291	2.2153
	5	2.4167	2.5250	2.5288	4.1700	5.0507	5.3159	5.4737	5.5367	5.4976	1.6439	2.1970	2.2953
	10	2.7050	2.8346	2.8283	7.1005	7.9159	8.2433	4.9775	5.0296	4.9261	1.9815	2.3400	2.4648

Table 4.2: Mean objective results for white noise.

Gain function	Input SNR	PESQ			SNR _{seg}			CD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	2.2566	2.2712	2.2259	1.8200	2.1551	2.2838	5.4677	5.4662	5.5764	2.8372	2.5646	2.8282
	5	2.6011	2.6349	2.5994	4.8539	5.1756	5.3341	4.9038	4.9263	4.9007	3.1100	2.6999	2.9758
	10	2.9068	2.9439	2.9301	7.8559	8.1593	8.4900	4.3538	4.3883	4.2506	3.5574	3.0309	3.3116
LSA	0	2.1774	2.2728	2.2534	1.0113	1.9133	2.0342	5.2757	5.2866	5.3589	2.0747	2.6813	2.8683
	5	2.5387	2.6492	2.6263	4.1421	5.1460	5.2755	4.6522	4.6688	4.6506	2.2000	2.7740	2.9510
	10	2.8437	2.9678	2.9466	7.4393	8.3790	8.6033	4.0440	4.0664	3.9745	2.3249	2.9665	3.1425

Table 4.3: Mean objective results for factory noise.

Gain function	Input SNR	PESQ			SNR _{seg}			CD			KurtR		
		DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop	DD	MDD	Prop
WF	0	1.9248	1.9155	1.9014	-0.6851	-0.5569	-0.5092	6.0480	6.1057	6.0693	2.8492	3.2080	3.2084
	5	2.3076	2.3050	2.2970	2.3144	2.4745	2.5305	5.1040	5.1526	5.0575	2.8528	3.2428	3.2072
	10	2.6588	2.6557	2.6483	5.7321	5.8815	5.9979	4.1042	4.1475	4.0454	2.7789	3.2029	3.1140
LSA	0	1.9301	1.9333	1.9226	-0.9637	-0.7017	-0.6500	5.7345	5.9067	5.8714	1.8794	2.7420	2.7117
	5	2.2927	2.3055	2.3004	2.0138	2.3624	2.4111	4.8300	4.9749	4.9038	1.8271	2.7365	2.6922
	10	2.6322	2.6530	2.6491	5.4629	5.8668	5.9522	3.8864	3.9960	3.9217	1.7471	2.6843	2.6166

Table 4.4: Mean objective results for babble noise.

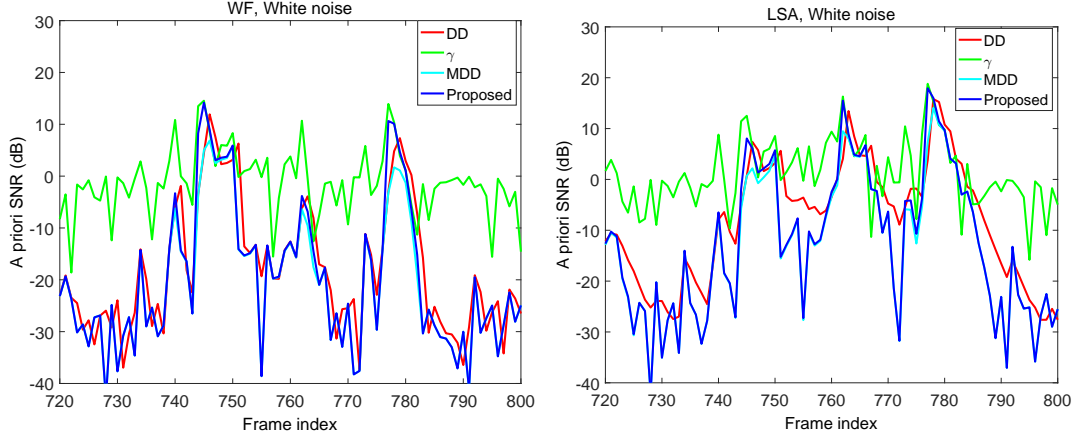


Figure 4.4: Comparison of a priori SNR estimation over short time between Υ (green solid line), $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB white background noise.

pink noise condition at low SNR values, whereas in high SNR (10 dB) the participants could not recognize any difference between STFT and CB speech signals. Moreover, in the MMSE-LSA scenarios, participants found STFT speech signals sound better than CB speech signals for the tested input SNR levels. Background noise results show that speech signals with STFT have less background noise than CB methods for different gain function cases and varying levels of SNR.

However, musical noise results show that CB methods have recorded the least amount of musical noise for different gain function cases and under different levels of input SNR values. From the overall scores it can be observed that the participants preferred the signals processed by STFT combined with the WF gain function for low SNR values. In contrast, they preferred signals with CB at high SNR. In the MMSE-LSA case, the participants preferred STFT methods over CB for the evaluated SNR levels.

Figures 4.16 and 4.18 depict the results of the listening test in babble background noise. Figure 4.16 shows the WF gain function results for the different a priori SNR estimation methods and varying levels of input SNR, whereas Figure 4.18 shows the corresponding results when MMSE-LSA gain function is used instead. In terms of speech quality, the subjective listening results of the WF gain function scenario show that speech signals processed with CB have recorded higher scores than signals processed with STFT at low SNR levels, while achieve approximately same scores at high SNR. This means that CB methods combined with WF gain function achieve significant improvement in speech quality com-

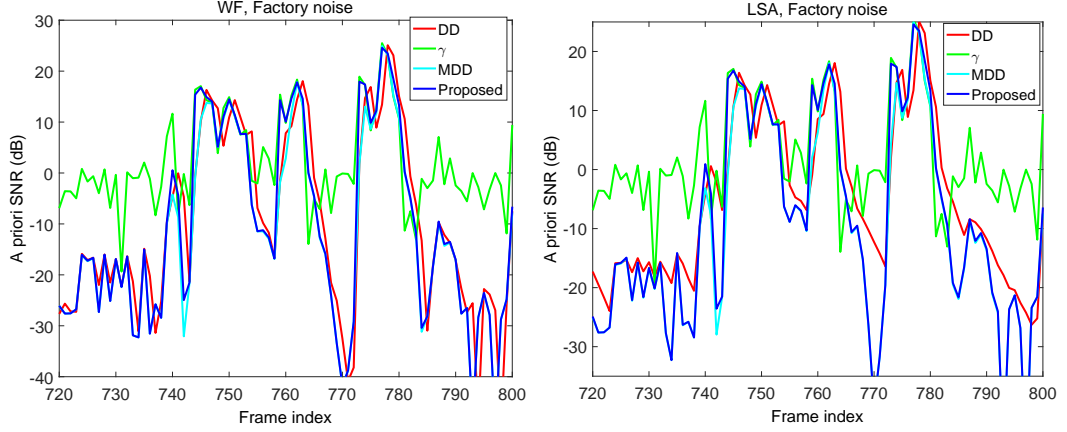


Figure 4.5: Comparison of a priori SNR estimation over short time between Υ (green solid line), $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB factory background noise.

pared to the STFT methods in adverse noise conditions. For MMSE-LSA case, the participants could not recognize any difference between the speech signals at different SNR levels. In terms of background noise, speech signals processed by CB have recorded higher scores (better noise reduction) than speech signals processed by STFT at low SNRs when the WF gain function is used, while in high SNR the background noise scores between CB and STFT are almost same which means all methods achieve same amount of noise suppression. For the MMSE-LSA gain function scenario, although all methods achieve same amount of noise suppression at low SNR, the participants preferred speech signals processed by STFT at high SNR levels. Musical noise results show that speech signals processed by CB achieved better results (less musical noise) than speech signals processed by STFT for the different gain functions as well as for different input SNR levels.

From the overall scores, it can be clearly seen that the participants preferred speech signals with CB when WF gain function was used for low SNR levels. While at high SNR all methods achieves almost the same results. Furthermore, when MMSE-LSA gain function used, there was no significant difference between speech signals in terms of overall scores for low SNR levels, but the participants preferred speech signals with STFT for high SNR levels.

4.5.5 Statistical analysis

Table 4.5 reports the obtained p -values of ANOVA test under different noise conditions. In terms of speech quality, The test shows that all obtained p -values

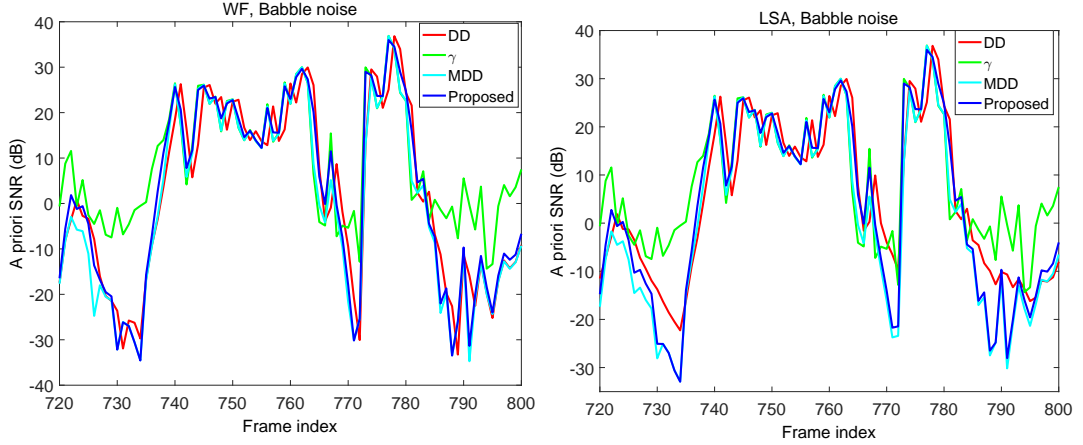


Figure 4.6: Comparison of a priori SNR estimation over short time between Υ (green solid line), $\hat{\xi}_{DD}$ (red solid line), $\hat{\xi}_{MDD}$ (cyan solid line), and $\hat{\xi}_{prop}$ (blue solid line) at 10 dB babble background noise.

are higher than 0.05, i.e., there was no statistically significant difference in speech quality between the obtained scores of the examined algorithms. This means that the enhanced speech signals did not contain a detectable speech distortion compared to the unprocessed speech signals. From background noise results, it can be clearly observed that speech signals enhanced using WF gain function provided significant differences when compared to the noisy speech signals in pink noise. In contrast, although there was no significant difference deemed in the babble noise case for low SNR, a significant improvement was achieved for high SNRs. In the MMSE-LSA gain function case, results show that most of the SNR estimation methods did not provide a significant noise suppression compared to the noisy signal except for pink noise and low SNR level.

In terms of musical noise, there was no significant difference between the enhanced speech signals that were estimated by the evaluated methods and the unprocessed speech signals detected in the different noise conditions and for varying levels of SNR. From the overall results, we observed that WF method provides significant difference than the unprocessed speech signal in pink noise condition and high SNR level. In the MMSE-LSA scenarios, the estimation methods provide significant difference in pink noise case for varying SNR levels.

However, the above mentioned statistical analysis can not provide the answer as to which method performed better than the unprocessed speech signal. As such, along with ANOVA results, a post hoc comparison test according to Tukey's HSD test was also conducted to identify which method significantly improved the

quality of the unprocessed speech signal. By comparing the scores obtained from the unprocessed speech signals and the scores obtained with speech signals enhanced by the various methods, the results of Tukey's HSD test are tabulated in Table 4.6 for WF gain function case and Table 4.7 for MMSE-LSA gain function. In these tables asterisk indicate significant differences between enhanced speech signals and noisy signal. From Table 4.6 for WF gain function case, it can be observed that some methods only provided significant differences when compared to the unprocessed speech signal in terms of background noise and overall quality. In pink noise case, (DD-STFT) and (Proposed-STFT) methods achieved significant noise suppression compared to noisy signal for different levels of SNR. In contrast, the rest of the methods achieved better noise suppression performance than noisy signals in higher SNR level only. In babble noise case, most of priori SNR estimators for STFT and CB achieved significant noise suppression at high SNR level. In terms of overall quality, the methods (DD-STFT) and (Proposed-CB) significantly improved the overall quality when compared to the unprocessed speech signal in pink noise case for high SNR levels.

From Table 4.7 for MMSE-LSA gain function case, only two methods (Proposed-STFT) and (DD-CB) achieved significant difference in noise suppression performance compared to the noisy speech signal in pink noise case and low SNR levels. Moreover, in babble noise case (DD-STFT) and (Proposed-STFT) achieved significant noise suppression only for low SNR levels. In addition, overall results show that the methods (DD-STFT), (MDD-STFT) and (Proposed-STFT) achieved significant improvement when compared to noisy signal in pink noise for low SNR levels, while at high SNR, only (Proposed-STFT) method achieved a significant improvement over the noisy signal.

4.6 Summary

In this chapter, a modified a priori SNR estimation method with an adaptive smoothing factor is presented. The proposed approach helps to improve the tracking speed of the a priori SNR estimation by employing a time-frequency varying smoothing factor based on a sigmoid function with a frequency dependent mean parameter. As a consequence, the adaptation speed of the a priori SNR estimation is improved during speech transitions while maintaining the advantage of DD approach in reducing the musical noise during noise frame. Experimental results show that more speech components are preserved compared to DD and MDD methods, especially in higher frequency region. Moreover, an improvement in the segmental SNR is achieved at different input SNRs.

Gain function	Noise	Input SNR	p-value			
			Speech	Background noise	Musical noise	Overall
WF	Pink	0dB	0.806	0.013	0.723	0.312
		10dB	0.235	0.0001	0.540	0.002
	Babble	0dB	0.938	0.119	0.215	0.794
		10dB	0.984	0.014	0.460	0.416
LSA	Pink	0dB	0.075	0.021	0.955	0.0004
		10dB	0.275	0.901	0.967	0.019
	Babble	0dB	0.993	0.842	0.997	0.530
		10dB	0.332	0.807	0.963	0.504

Table 4.5: One way ANOVA test results to verify the statistically significant difference between different frequency warping scales used in the listening test under different noise conditions.

Furthermore, a cross evaluation for STFT and CB processing was conducted by using subjective listening test. From the test results, it can be clearly noted that although STFT method achieves better results in stationary background noise in terms of better noise suppression and speech quality, its performance degraded in non-stationary background noise and is generating more musical noise. On the other hand, CB processing provides significant benefit in terms of less musical noise under different noise conditions and different levels of input SNR. In addition, it achieves better performance compared to STFT in non-stationary noise (babble noise) especially for low SNR levels. This means that the proposed critical band based frequency warping scale is more useful in adverse noisy conditions such as low SNR and non-stationary background noise.

Noise	Input SNR	Rate	STFT			CB		
			DD	MDD	Proposed	DD	MDD	Proposed
Pink	0 dB	Speech						
		Noise	*		*			
		Musical noise						
		Overall						
	10 dB	Speech						
		Noise	*	*	*	*	*	*
		Musical noise						
		Overall	*					*
Babble	0 dB	Speech						
		Noise						
		Musical noise						
		Overall						
	10 dB	Speech						
		Noise	*	*	*	*	*	
		Musical noise						
		Overall						

Table 4.6: Tukey's HSD Comparison between the enhanced speech signal using WF gain function and the unprocessed speech signal under different conditions.

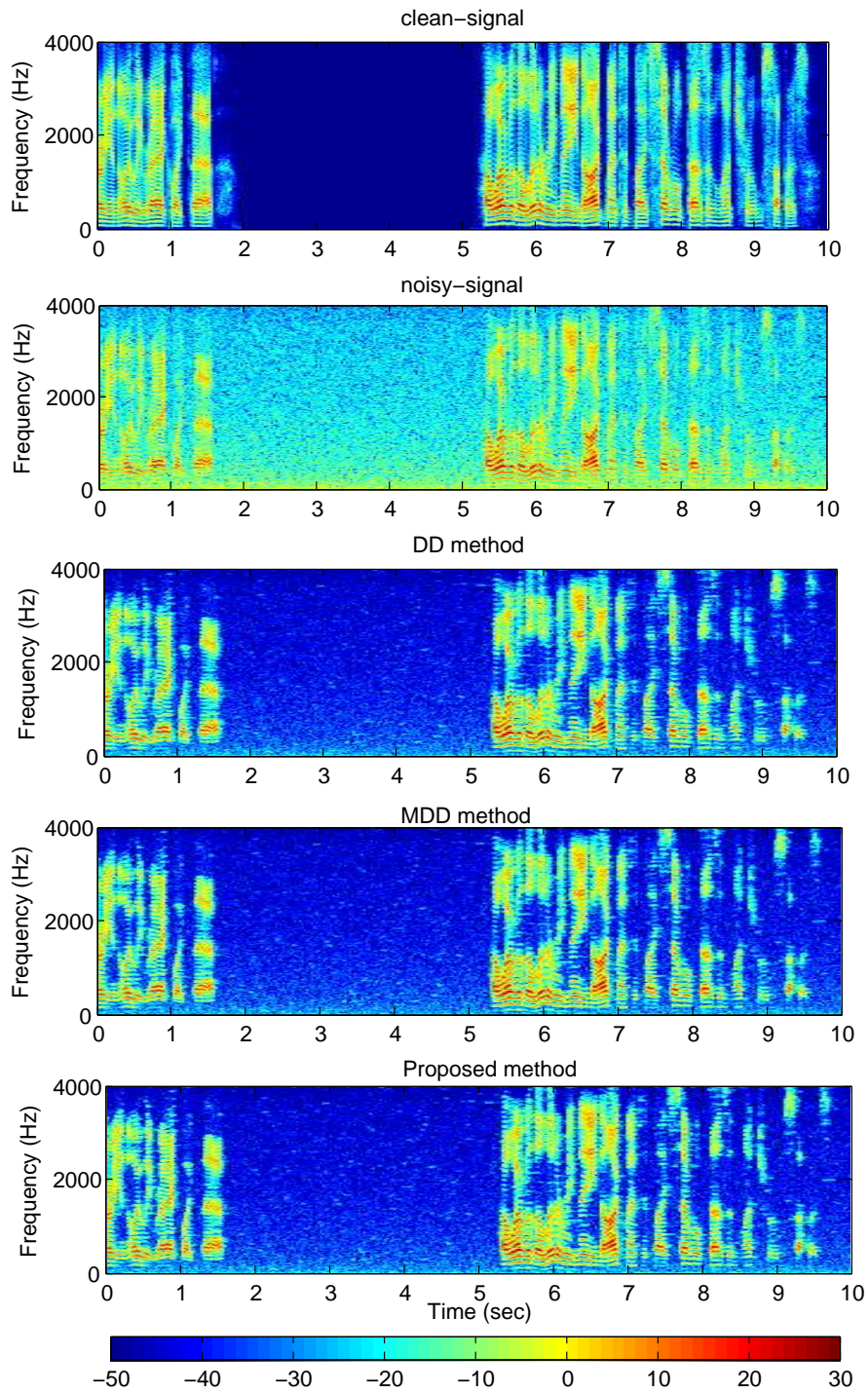


Figure 4.7: Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by Wiener filter speech estimation technique.

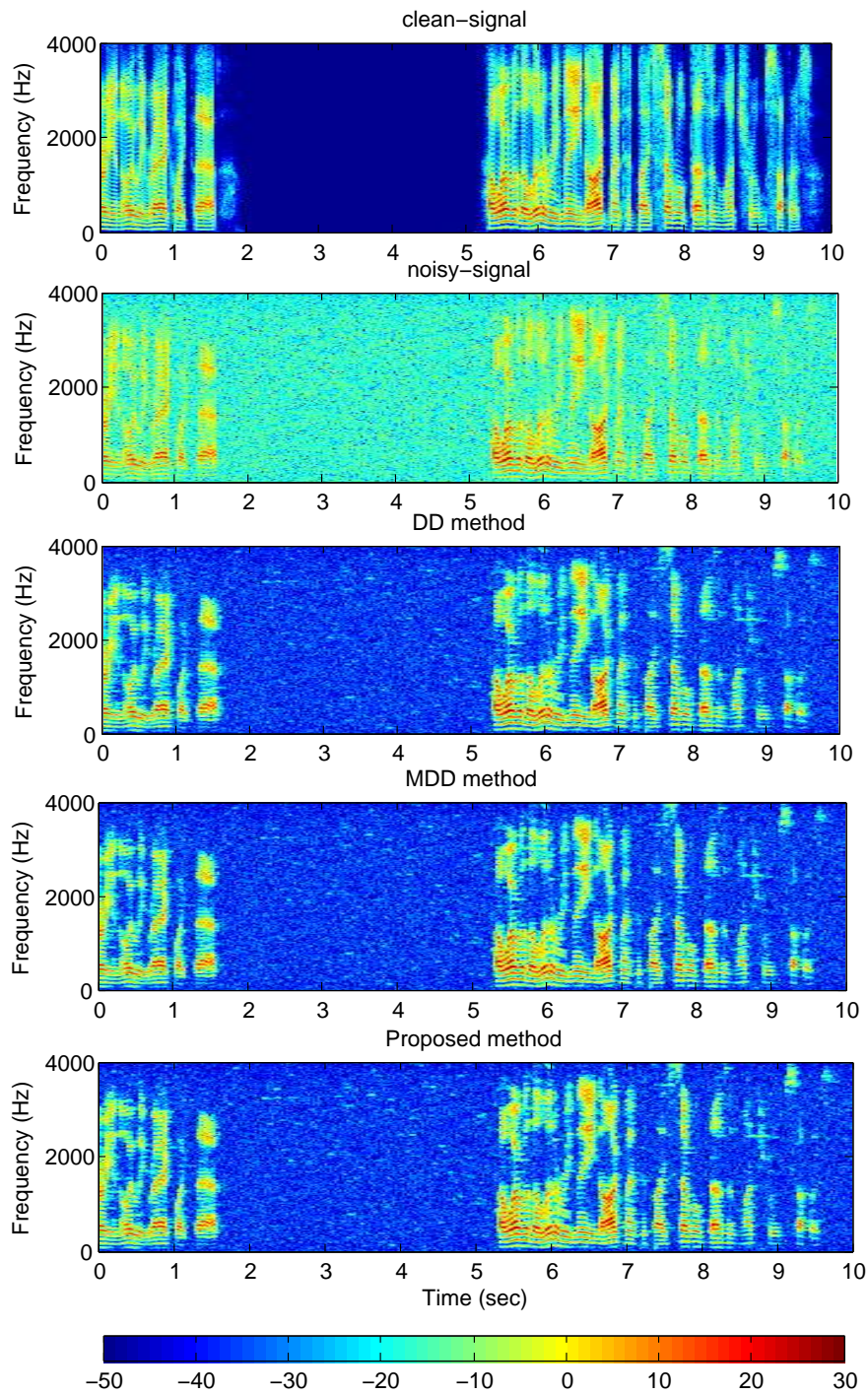


Figure 4.8: Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by Wiener filter speech estimation technique.

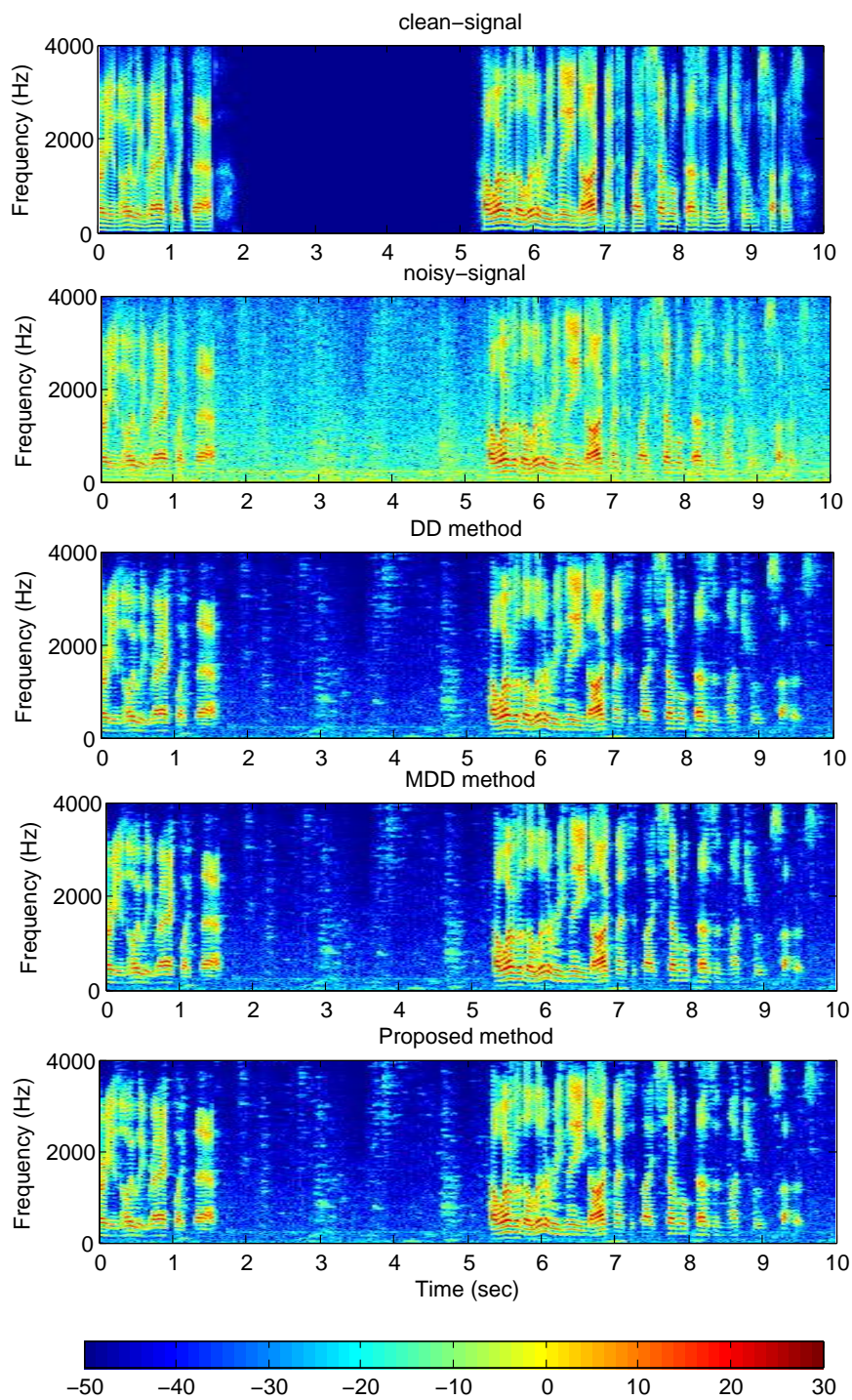


Figure 4.9: Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by Wiener filter speech estimation technique.

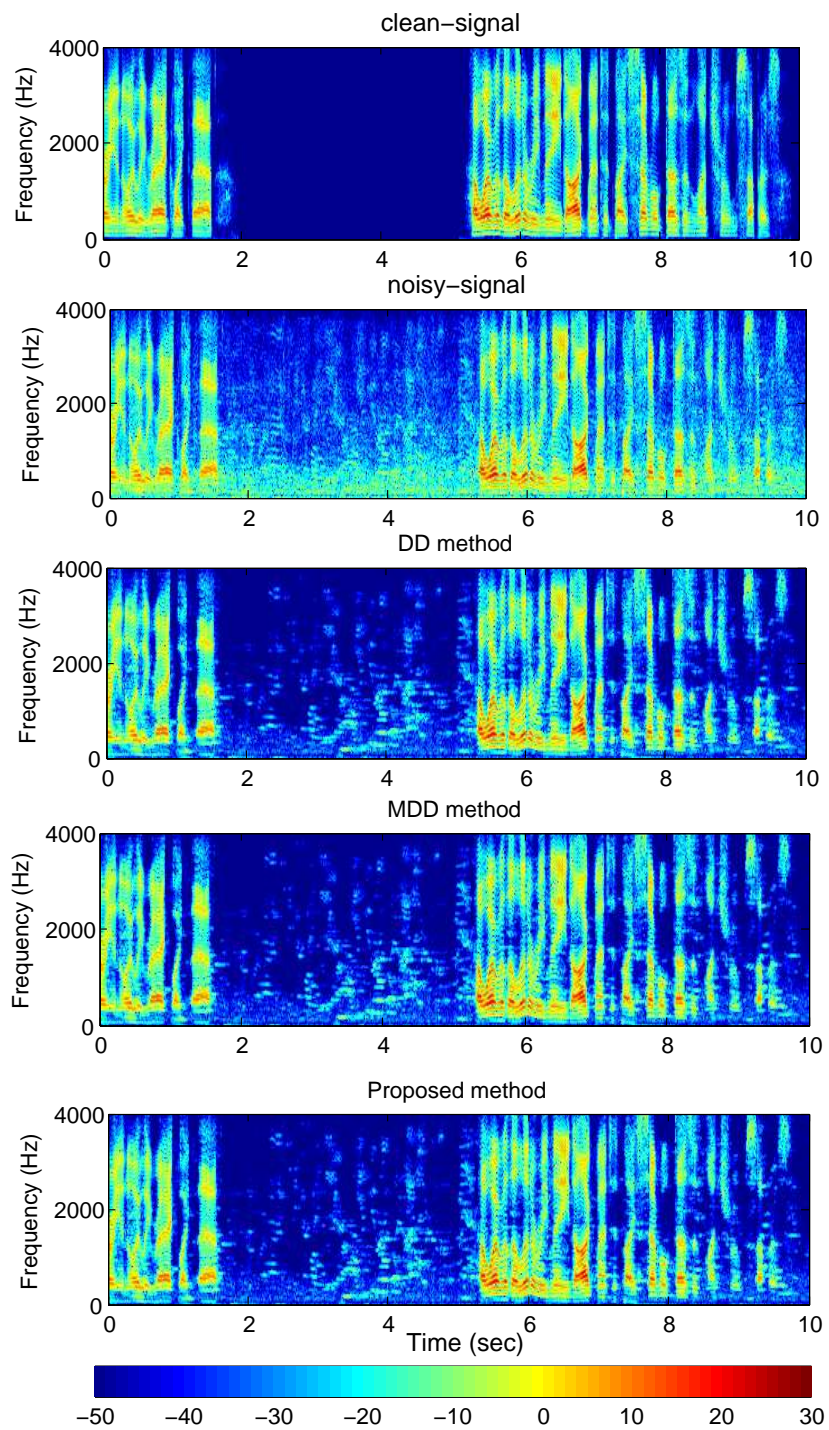


Figure 4.10: Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by Wiener filter speech estimation technique.

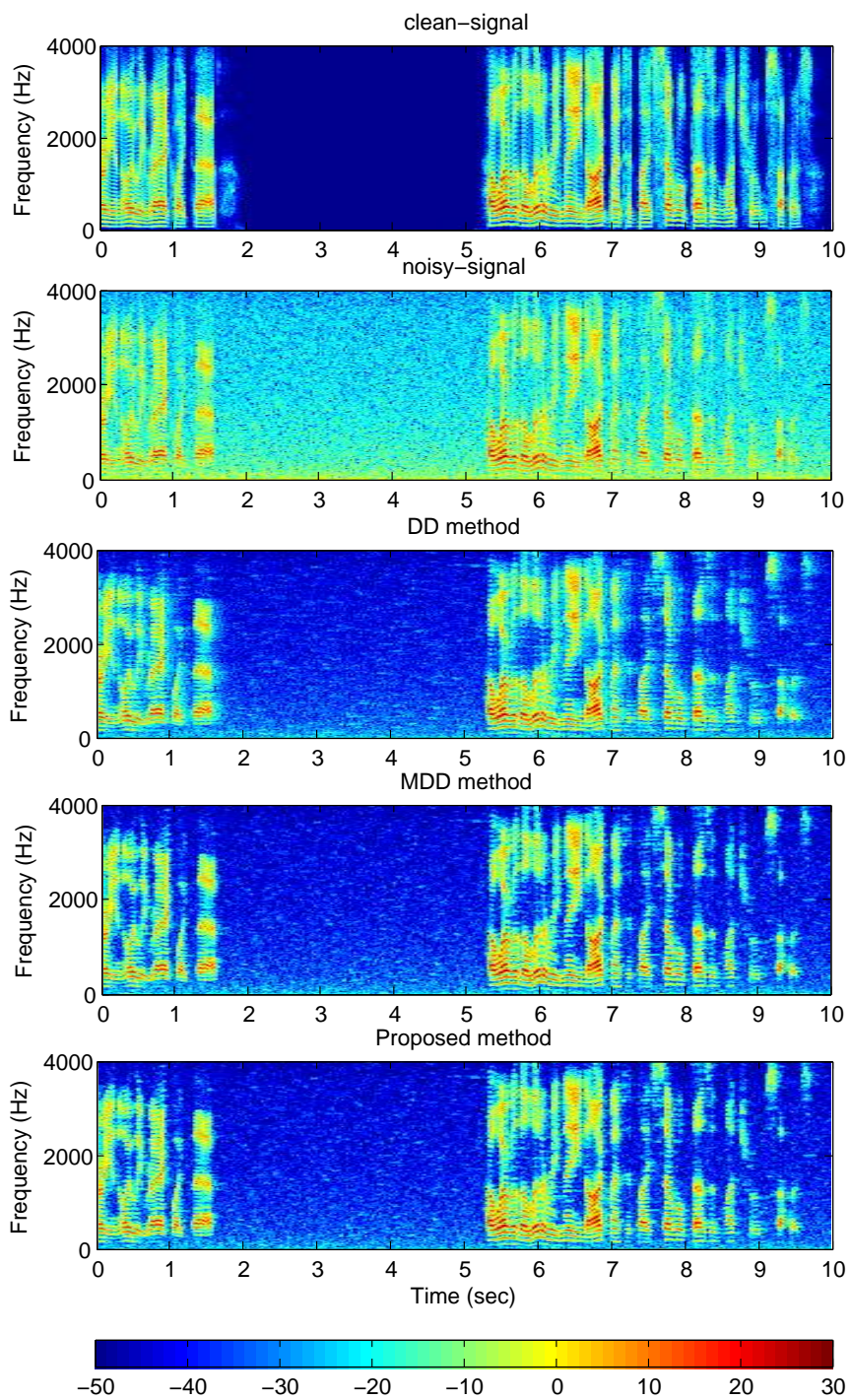


Figure 4.11: Speech spectrograms for noisy speech corrupted with pink noise at 10 dB enhanced by LSA speech estimation technique.

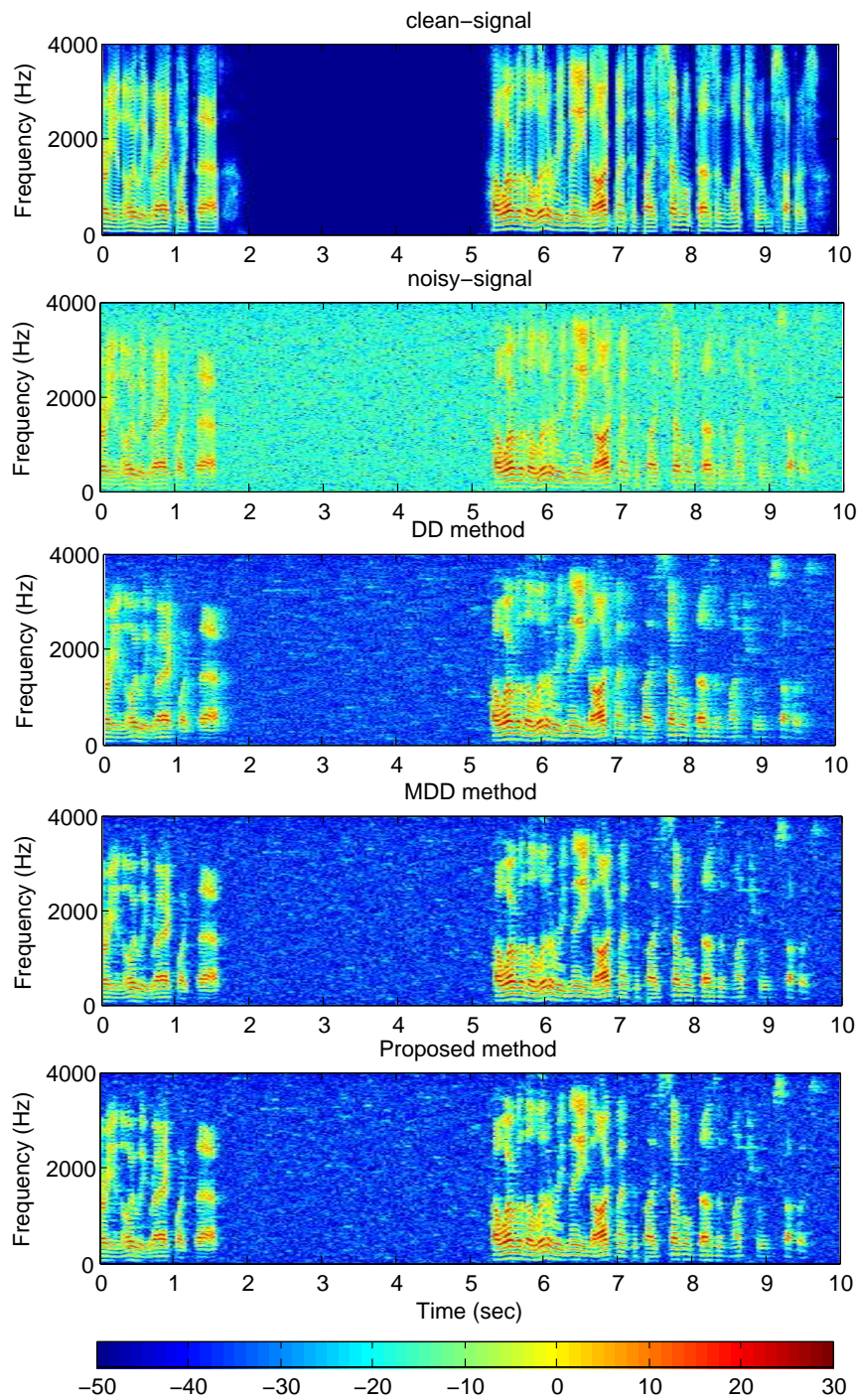


Figure 4.12: Speech spectrograms for noisy speech corrupted with white noise at 10 dB enhanced by LSA speech estimation technique.

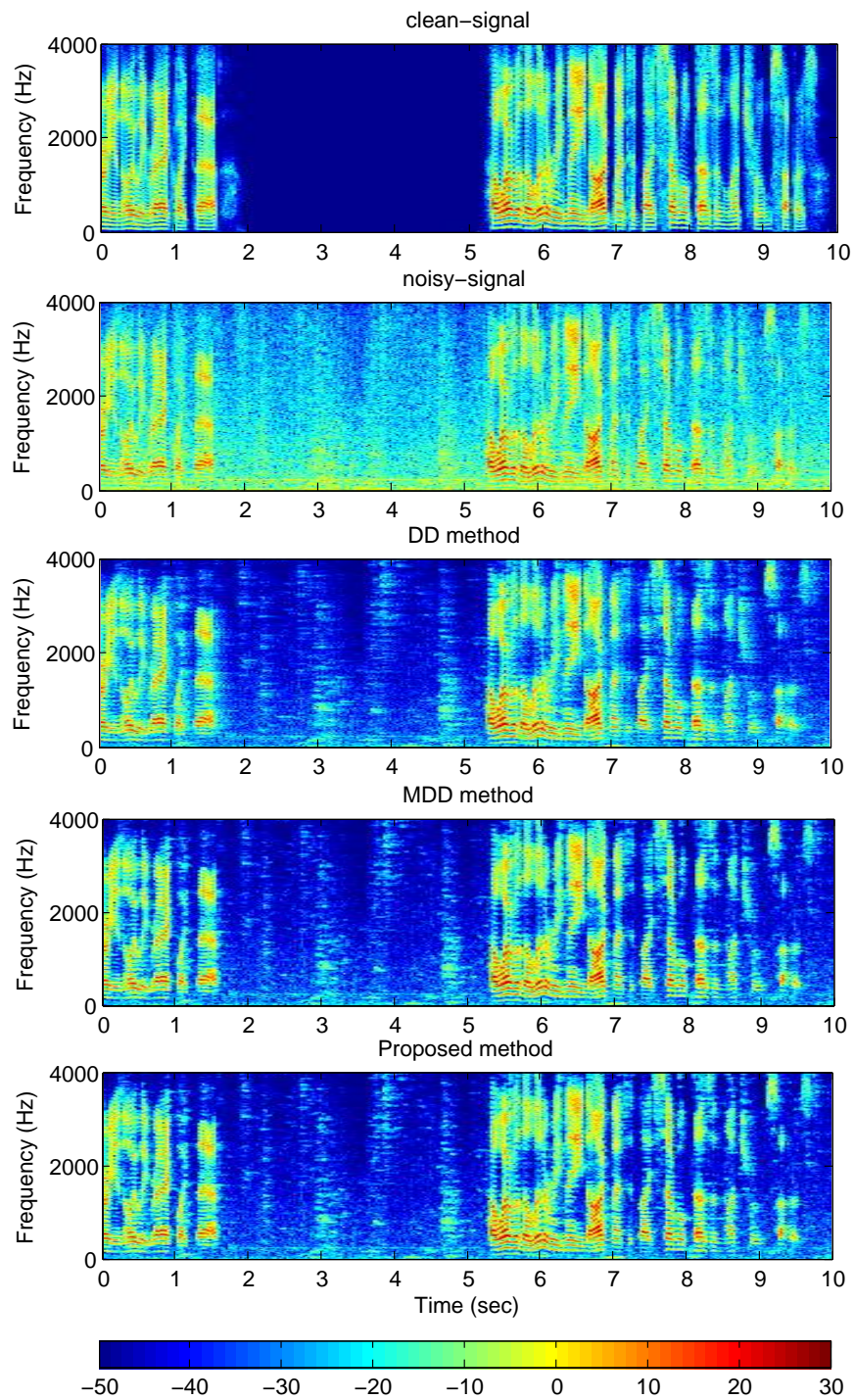


Figure 4.13: Speech spectrograms for noisy speech corrupted with factory noise at 10 dB enhanced by LSA speech estimation technique.

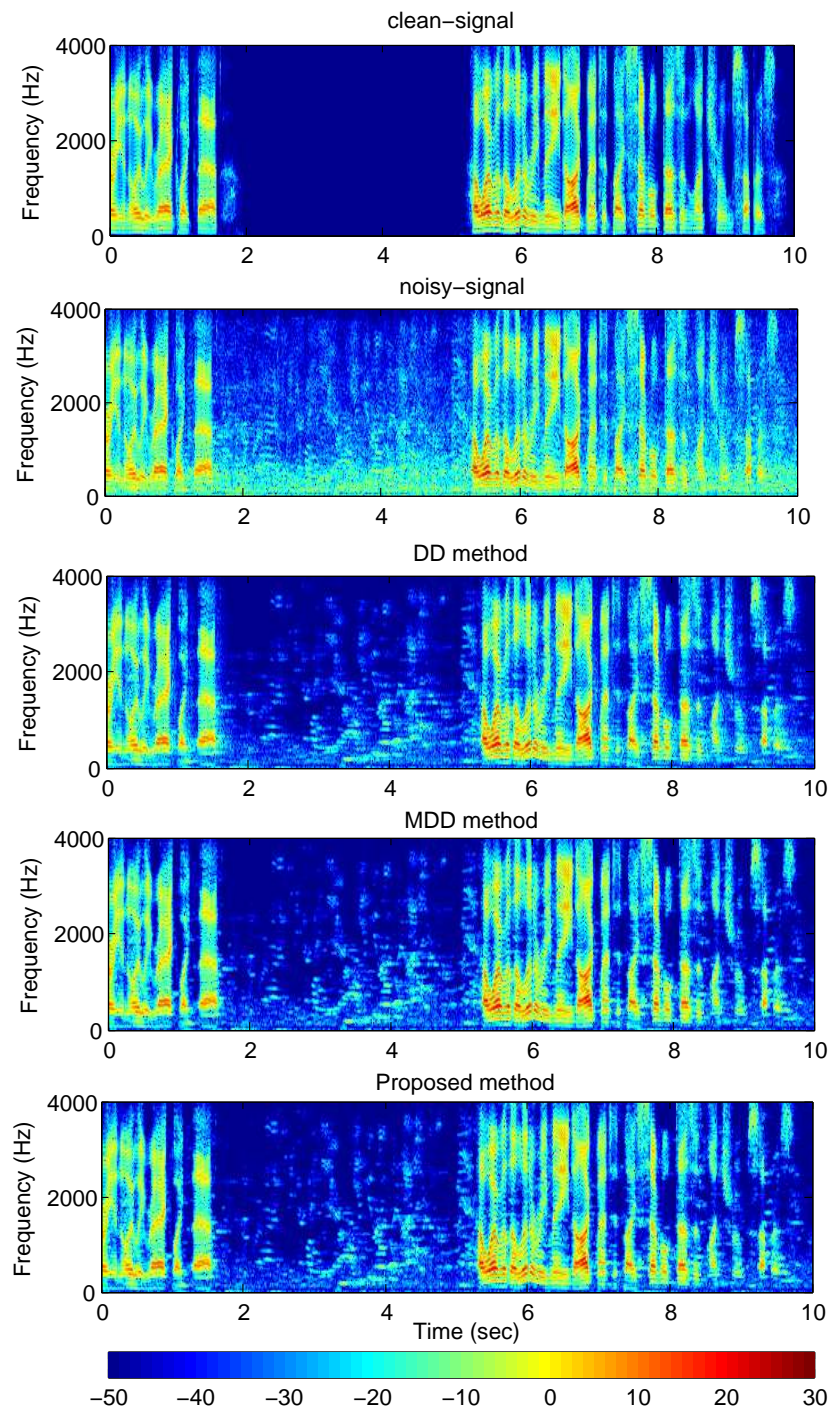


Figure 4.14: Speech spectrograms for noisy speech corrupted with babble noise at 10 dB enhanced by LSA speech estimation technique.

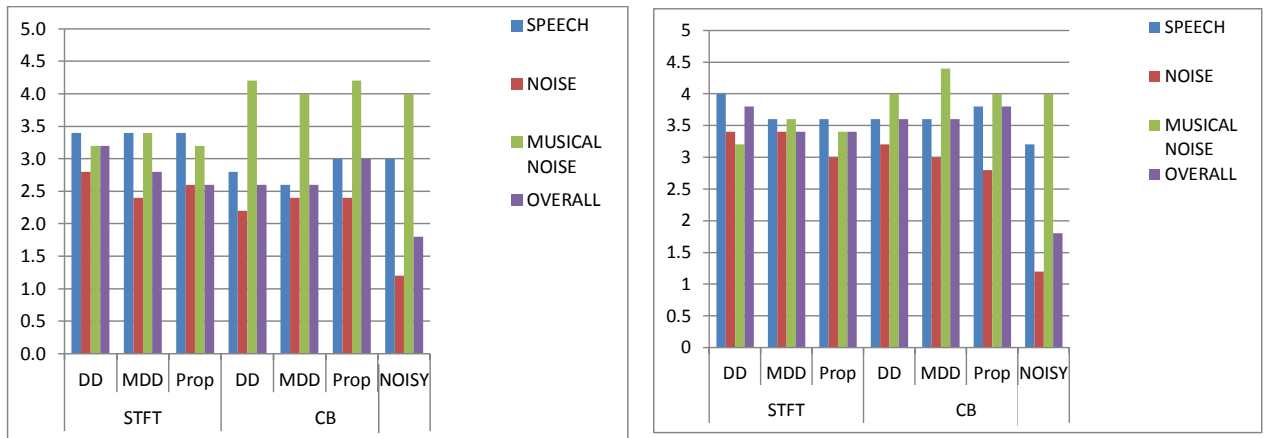


Figure 4.15: Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with WF gain function and evaluated in pink background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.

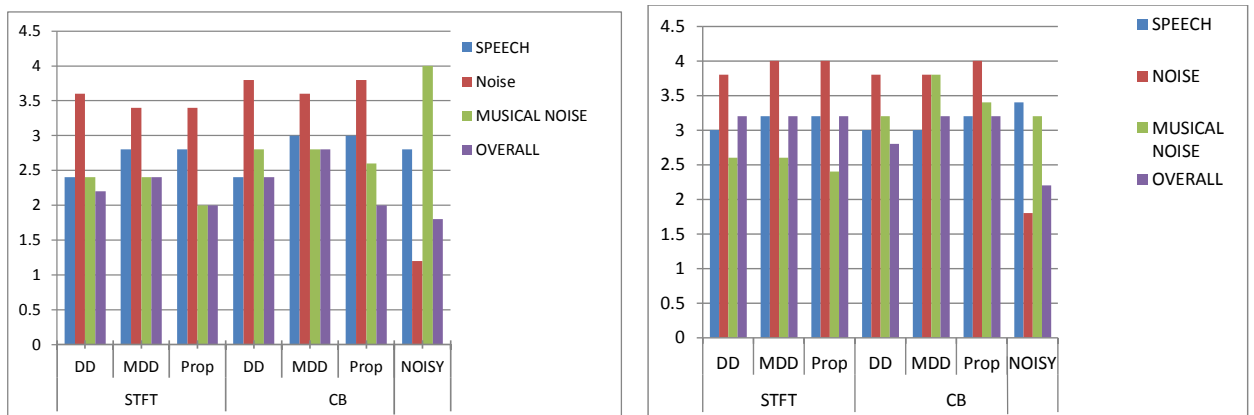


Figure 4.16: Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with WF gain function and evaluated in babble background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.

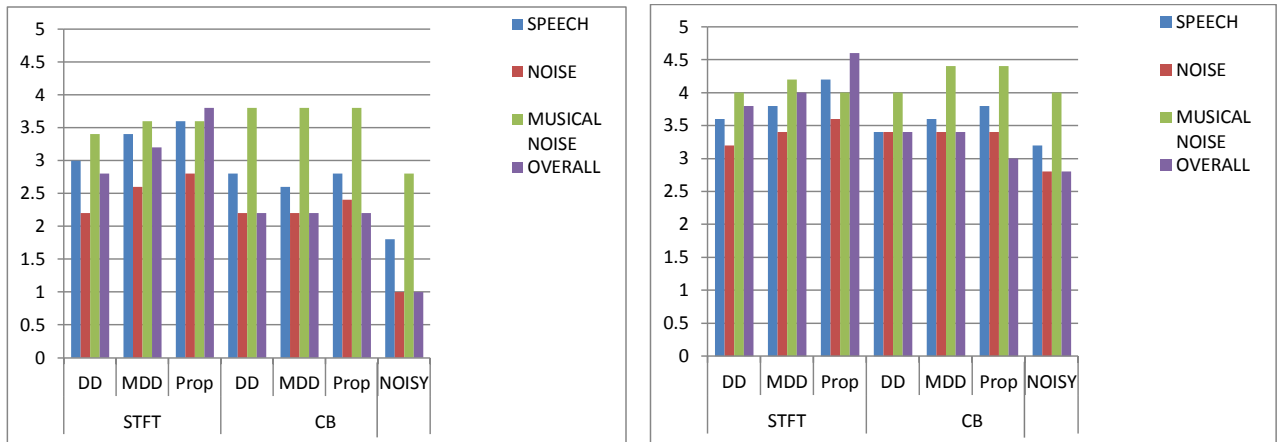


Figure 4.17: Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with MMSE-LSA gain function and evaluated in pink background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.

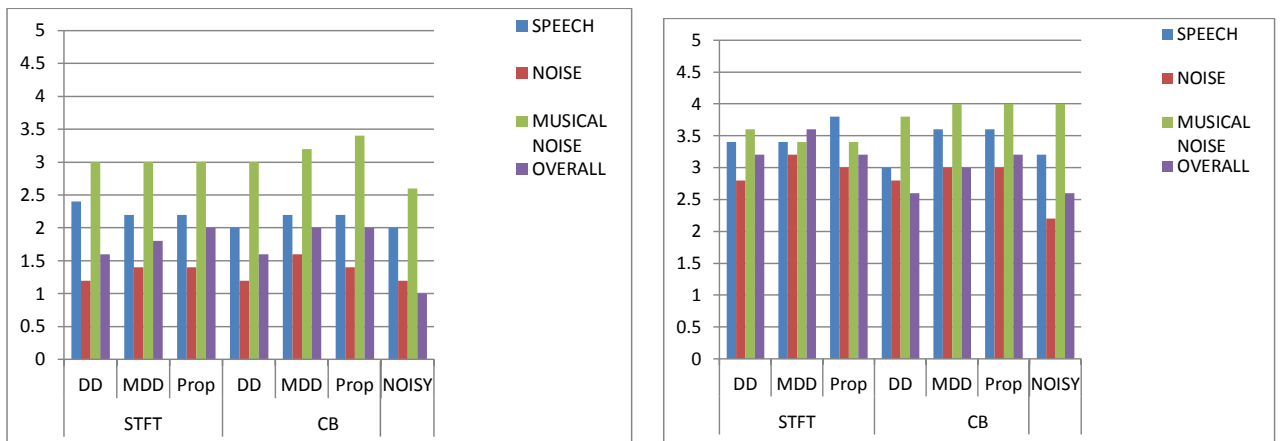


Figure 4.18: Mean subjective listening test scores for speech processed by different a priori SNR estimation methods combined with MMSE-LSA gain function and evaluated in babble background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.

Noise	Input SNR	Rate	STFT			CB		
			DD	MDD	Proposed	DD	MDD	Proposed
Pink	0 dB	Speech						
		Noise			*	*		
		Musical noise						
		Overall	*	*	*			
	10 dB	Speech						
		Noise						
		Musical noise						
		Overall			*			
Babble	0 dB	Speech						
		Noise	*		*			
		Musical noise						
		Overall						
	10 dB	Speech						
		Noise						
		Musical noise						
		Overall						

Table 4.7: Tukey’s HSD Comparison between the enhanced speech signal using MMSE-LSA gain function and the unprocessed speech signal under different conditions.

Chapter 5

Robust Broadband Beamformer in Reverberant Environment

In indoor applications with reverberation where spatial diversity is also included, single channel speech enhancement techniques cannot suppress the reverberation since it exploits only the time-frequency information of the received signal. In order to exploit the spatial diversity, multi-channel speech enhancement techniques are required. In this chapter, we will investigate different designs of broadband beamformers in reverberant environments. In addition to room reverberation, robustness to amplitude and phase mismatches in the microphones will be included.

The main work in this chapter have previously appeared in the following publications:

1. L. Khalid, S. Nordholm, and H. H. Dam, Design study on microphone arrays, IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015.
2. Nahma, L., Dam, H.H.D., Nordholm, S., "Robust beamformer design against mismatch in microphone characteristics and acoustic environments," International Workshop on Acoustic Echo and Noise Control (IWAENC), Tokyo, Japan 2018.
3. L. Nahma, H. H. Dam, Cedric Ka Fai Yiu, and S. Nordholm, Robust Broadband Beamformer Design for Noise Reduction and Dereverberation, submitted to Multidimensional Systems and Signal Processing (MSSP).

5.1 Introduction

Microphone arrays are used to extract signals of interest while reducing or canceling undesired signals by employing an array of spatially separated microphones. The microphone array measures the wave field in space and beamformer filters provide a spatial and temporal filtering of the signals from each microphone [54, 79, 115].

In open space applications where the sound propagates unencumbered a free field Green function describes the transmission between a sound source and each microphone [116]. In contrast to this scenario, we have indoor applications, where the sound wave propagates inside an enclosure. In this situation, the microphone signals contain not only the direct path source signal but also delayed and attenuated duplicate signals created by reflections from the enclosure and objects inside it. For this scenario, RIR becomes more complex [53, 54, 69, 82].

In order to describe the wave propagation from a source to each microphone element inside an enclosure, room acoustic simulators are useful tools. There are different methods for room acoustic model depending on application and range of frequencies. They can be classified into three types for different ranges of frequencies: firstly, wave-based model, which is more suitable for low frequencies and small enclosures. Secondly, a statistical model which is valid for many practical situations under different ranges of frequencies, [117], and thirdly, ray-based model, which is a geometrical acoustics modeling technique. One implementation of such a geometric model is the image source method (ISM) [118]. This method is one of the most commonly used techniques for simulating room acoustics. It is simple and yet efficient and provides a good model correspondence over the audible frequency range. The main drawback of this method is the high computational cost as the computation of energy time-curves is very demanding [119]. Accordingly in [120] Lehmann et. al. proposed the diffuse reverberation model (DRM), which is based on modeling the diffuse reverberation part as decaying random noise by decomposing room impulse response into three parts; direct path, early reflections and late (diffuse) reverberation. In this method, the first two parts are simulated by ISM method while the last diffuse part is simulated by using a DRM approach.

Joint dereverberation and noise reduction algorithms have become a major research subject in the last decade since reverberation and noise typically result in a degradation of speech quality and intelligibility as well as reduced listening comfort. Recently renewed interest in such algorithms has been driven by com-

mercial speech recognition applications [121]. Many different studies have been done in reverberant environments while considering different aspects of processing [79]. In the paper by Li et. al. [69] several multi-criteria optimization models were formulated based on L-1 norm for the fixed indoor beamformer design. The proposed method separates the early and late reverberation in the design process.

A two stage beamforming approach was proposed for dereverberation and noise reduction [82]. A combination of fixed and adaptive beamformers have been employed in two stage approach to achieve a joint dereverberation and noise reduction. In [122] a combination of MVDR beamformer and signal channel spectral enhancement scheme was presented for a joint dereverberation and noise reduction. The proposed system aims to suppress noise and reverberation by first employing a minimum variance distortion-less response beamformer, then the beamformer output is processed by a single channel speech enhancement method to suppress the residual noise and reverberation.

Due to the sensitivity of beamformer designs to mismatches in microphone characteristics such as gain, phase and position or source spreading and local scattering, any violation in these characteristics can lead to a significant degradation in the overall performance [62, 123]. Hence, developing a robust beamformer design techniques which accounts for arbitrary unknown model mismatch is desirable. In principle it would be possible to calibrate each microphone as well as the combined array. However, the drawbacks of calibration are: Firstly, microphone characteristics change over time which means that calibration does not provide a long term solution. Secondly, they are time consuming as every individual microphone as well as the combined array is required to be calibrated. Another robustness technique is achieved by considering the array characteristics in the beamformer design procedure, either by using the mean performance optimization [69, 124] or the worst case optimization method [68].

To the best of our knowledge, no extended indoor design which includes the robustness against mismatch in element characteristics has previously been carried out. To bridge this literature gap, in this chapter we extend the indoor beamforming design [69] by including robustness towards the microphone characteristics (gain and phase) into the design. Specifically, the mean performance of the designed beamformer for all possible microphone characteristics according to a given distribution and uncertainty has been developed. We have investigated the robust design using multiplicative and additive error models. The beamformer design methods that have been considered in this study are:

- i Design using direct path only of the RIR.
- ii Robust design using direct path.
- iii Using RIR based on the Image Source Method (ISM) with a specific reverberation time.
- iv Robust Indoor beamformer design which combines steps ii and iii.

The aforementioned beamformer designs are examined for different acoustically adverse environments using simulated and measured room impulse responses. By comparing the sensitivity performance of non-robust and robust designs, an improved performance is pointed out in terms of significant reduction in error sensitivity for the robust beamformer designs. Moreover, evaluation results from the four designs show that the robust direct path based beamformer can achieve almost the same performance as the indoor beamformer design under different reverberant environments despite being a much simpler and faster design method. Moreover, robust direct design shows robustness in the beamformer response in presence of local scattering perturbation. In addition, the robust indoor design provides stronger robustness towards combinations of reverberation and microphone perturbations in amplitude and phase. This chapter is organized as follows: In Section 5.2 the problem is formulated. Section 5.3 describes indoor broadband beamformer design problem as WLS problem where RIR is simulated using ISM room simulator. Section 5.4 demonstrates the robust broadband beamformer design using mean performance optimization method and by using two different error models: additive and multiplicative. Section 5.5 discusses the aperture size optimization problem while Section 5.6 outlines the objective measurements used for performance evaluation. Section 5.7 presents evaluation results and Section 5.8 concludes the paper.

5.2 Problem formulation

Consider a microphone array with M elements in positions \mathbf{r}_m , $m = 1, 2, \dots, M$, and an L taps FIR filter behind each microphone as depicted in Figure 5.1. Assume $S(f)$ is the spectral density of source signal at position vector \mathbf{r} traveling in a homogeneous non-dispersive free field. The received signals are sampled synchronously at a rate of f_s . The transfer function (Green function) between the source signal and each microphone array element can be written as

$$R_m(\mathbf{r}, f) = \frac{1}{\|\mathbf{r} - \mathbf{r}_m\|} \exp(-j2\pi f \frac{\|\mathbf{r} - \mathbf{r}_m\|}{c}), \quad 1 \leq m \leq M \quad (5.1)$$

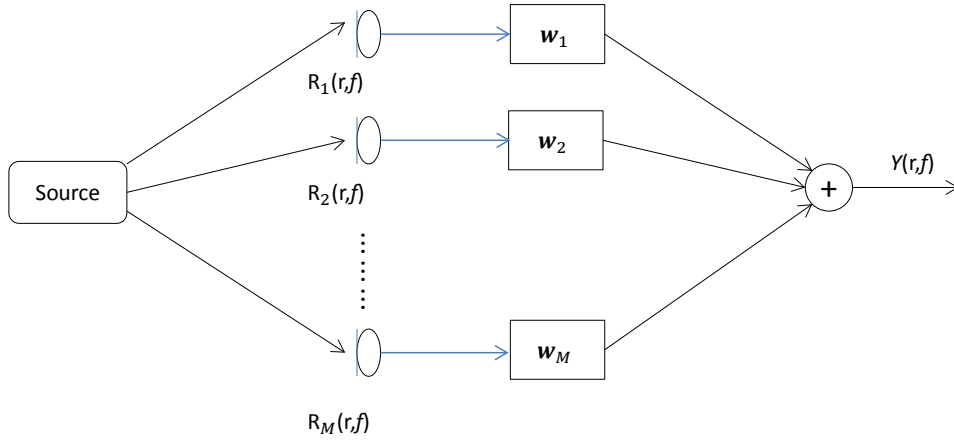


Figure 5.1: Block diagram of the broadband beamformer.

where f is the frequency and c is the speed of the sound. The array response vector can be obtained by combining the transfer function from the source to each microphone element with FIR filter response

$$\mathbf{d}(\mathbf{r}, f) = \mathbf{R}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) \quad (5.2)$$

where \otimes denotes for the Kronecker product,

$$\mathbf{R}(\mathbf{r}, f) = [R_1(\mathbf{r}, f), \dots, R_M(\mathbf{r}, f)]^T$$

$\mathbf{d}_0(f)$ is the FIR filter response vector,

$$\mathbf{d}_0(f) = [1, e^{-j2\pi f/f_s}, \dots, e^{-j2\pi f(L-1)/f_s}]^T \quad (5.3)$$

and the subscript $[\cdot]^T$ represents the vector transpose. The beamformer response can be given by

$$G(\mathbf{r}, f) = \mathbf{w}^T \mathbf{d}(\mathbf{r}, f) \quad (5.4)$$

where $\mathbf{w} \in \mathbb{R}^{ML \times 1}$ is the FIR filter coefficients vector and $\mathbf{d}(\mathbf{r}, f)$ is a column vector of length ML . The beamformer frequency domain output is given by

$$Y(r, f) = G(\mathbf{r}, f)S(f). \quad (5.5)$$

5.3 Indoor Broadband Beamformer Design

Consider a speech source and a microphone array in an indoor room environment according to Figure 5.1. Now the source signal will be convolved by an individual RIR between the source to each microphone element. To model the room, an acoustic room simulator using the image source method (ISM) was used to obtain RIRs. The frequency domain room response from the speech source to the microphone array can be written as

$$\mathbf{R}(\mathbf{r}, f) = \mathbf{R}_{dir}(\mathbf{r}, f) + \mathbf{R}_{rev}(\mathbf{r}, f). \quad (5.6)$$

where $\mathbf{R}_{dir}(\mathbf{r}, f)$ denotes the direct path frequency response and $\mathbf{R}_{rev}(\mathbf{r}, f)$ denotes the frequency response of the reverberation path i.e. RIR with direct path excluded.

In general, the broadband beamformer design problem is to calculate the filter coefficients \mathbf{w} such that the actual response $G(\mathbf{r}, f)$ fits the desired response $G_d(\mathbf{r}, f)$, which is specified depending on the application with

$$G_d(\mathbf{r}, f) = \begin{cases} e^{-j2\pi f(\frac{\|\mathbf{r}-\mathbf{r}_c\|}{c} + \frac{L-1}{2}T)}, & \forall(\mathbf{r}, f) \in \mathcal{P} \\ 0, & \forall(\mathbf{r}, f) \in \mathcal{S} \end{cases} \quad (5.7)$$

where \mathcal{P} and \mathcal{S} denote the passband and the stopband regions, respectively. \mathbf{r}_c is the location of the reference point, and $T = 1/f_s$. The problem is to minimize the Weighted Least Square (WLS) error $J_{WLS}(\mathbf{w})$ as

$$J_{WLS}(\mathbf{w}) = \int_{\Omega} \int_{\mathcal{R}} V(\mathbf{r}, f) |G(\mathbf{r}, f) - G_d(\mathbf{r}, f)|^2 d\mathbf{r}df \quad (5.8)$$

where $G(\mathbf{r}, f)$ is the beamformer response as defined in Eq. (5.4), Ω is the frequency domain and \mathcal{R} is the spatial domain, and $V(\mathbf{r}, f)$ is a positive weighting function. According to Eq. (5.8) the reverberation path cannot be controlled directly as it is part of the whole room impulse response as in Eq.(5.6) [69]. Therefore, the design problem has been modified in order to include deviation from the direct path with the desired frequency response and the error due to the reverberation path as follows

$$\begin{aligned} J_{mod,WLS}(\mathbf{w}) &= \int_{\Omega} \int_{\mathcal{R}} (V_1(\mathbf{r}, f) |G_{dir}(\mathbf{r}, f) - G_d(\mathbf{r}, f)|^2 + V_2(\mathbf{r}, f) |G_{rev}(\mathbf{r}, f)|^2) d\mathbf{r}df \\ &= \int_{\Omega} \int_{\mathcal{R}} (V_1(\mathbf{r}, f) |\mathbf{w}^T \mathbf{d}_{dir}(\mathbf{r}, f) - G_d(\mathbf{r}, f)|^2 \\ &\quad + V_2(\mathbf{r}, f) |\mathbf{w}^T \mathbf{d}_{rev}(\mathbf{r}, f)|^2) d\mathbf{r}df, \end{aligned}$$

where $V_1(\mathbf{r}, f)$ and $V_2(\mathbf{r}, f)$ are positive weighting functions. The above cost function can be simplified to a quadratic cost function,

$$J_{mod,WLS}(\mathbf{w}) = \mathbf{w}^T \mathbf{Q}_{mod,WLS} \mathbf{w} - 2\mathbf{p}_{mod,WLS}^T \mathbf{w} + const \quad (5.9)$$

where

$$\begin{aligned} \mathbf{Q}_{mod,WLS} &= \int_{\Omega} \int_{\mathcal{R}} (V_1(\mathbf{r}, f) \Re \{ \mathbf{d}_{dir}(\mathbf{r}, f) \mathbf{d}_{dir}^H(\mathbf{r}, f) \} \\ &\quad + V_2(\mathbf{r}, f) \Re \{ \mathbf{d}_{rev}(\mathbf{r}, f) \mathbf{d}_{rev}^H(\mathbf{r}, f) \}) d\mathbf{r}df \\ \mathbf{p}_{mod,WLS} &= \int_{\Omega} \int_{\mathcal{R}} V_1(\mathbf{r}, f) \Re \{ \mathbf{d}_{dir}(\mathbf{r}, f) G_d^H(\mathbf{r}, f) \} d\mathbf{r}df \\ const &= \int_{\Omega} \int_{\mathcal{R}} V_1(\mathbf{r}, f) |G_d(\mathbf{r}, f)|^2 d\mathbf{r}df \end{aligned}$$

and $\mathbf{d}_{dir}(\mathbf{r}, f) = \mathbf{R}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f)$, $\mathbf{d}_{rev}(\mathbf{r}, f) = \mathbf{R}_{rev}(\mathbf{r}, f) \otimes \mathbf{d}_0(f)$. The optimal filter coefficients vector that minimizes $J_{mod,WLS}(\mathbf{w})$ is obtained by

$$\mathbf{w} = \mathbf{Q}_{mod,WLS}^{-1} \mathbf{p}_{mod,WLS}. \quad (5.10)$$

5.4 Robust Beamformer Design

Broadband beamformers designed with Minimax or WLS techniques are highly sensitive to errors in microphone characteristics such as gain, phase and position [123]. Even small changes can lead to a severe degradation in beamformer performance. Thus, to design beamformers for practical applications it is important to consider robustness in the beamformer design procedure. Generally, errors can be formulated as multiplicative or additive models. In this chapter, we considered both models to microphone characteristics error (gain and phase) [54].

5.4.1 Additive error model

Denote by $\mathbf{a}_{dir}(\mathbf{r}, f)$ and $\mathbf{a}_{rev}(\mathbf{r}, f)$ the complex model error random vectors for the direct and the reverberation parts, respectively, where m^{th} elements $\mathbf{a}_{dir}(\mathbf{r}, f)$ and $\mathbf{a}_{rev}(\mathbf{r}, f)$ can be characterized by the gain errors $\mathbf{a}_{\rho,dir}(\mathbf{r}, f) = |\mathbf{a}_{dir}(\mathbf{r}, f)|$, $\mathbf{a}_{\rho,rev}(\mathbf{r}, f) = |\mathbf{a}_{rev}(\mathbf{r}, f)|$ and the phase errors $\mathbf{a}_{\gamma,dir}(\mathbf{r}, f) = \arg(\mathbf{a}_{dir}(\mathbf{r}, f))$, $\mathbf{a}_{\gamma,rev}(\mathbf{r}, f) = \arg(\mathbf{a}_{rev}(\mathbf{r}, f))$. The perturbed response vectors are given by

$$\begin{aligned} \tilde{\mathbf{R}}_{dir}(\mathbf{r}, f) &= \mathbf{R}_{dir}(\mathbf{r}, f) + \mathbf{a}_{dir}(\mathbf{r}, f) \\ \tilde{\mathbf{R}}_{rev}(\mathbf{r}, f) &= \mathbf{R}_{rev}(\mathbf{r}, f) + \mathbf{a}_{rev}(\mathbf{r}, f). \end{aligned} \quad (5.11)$$

Following from Eq. (5.2), the perturbed array response vector is given by

$$\begin{aligned}
\tilde{\mathbf{d}}_{dir}(\mathbf{r}, f) &= \tilde{\mathbf{R}}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) \\
&= (\mathbf{a}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f)) + \mathbf{d}_{dir}(\mathbf{r}, f)
\end{aligned} \tag{5.12}$$

and

$$\tilde{\mathbf{d}}_{rev}(\mathbf{r}, f) = (\mathbf{a}_{rev}(\mathbf{r}, f) \otimes \mathbf{d}_0(f)) + \mathbf{d}_{rev}(\mathbf{r}, f)$$

where $\mathbf{1}_L$ is an $L \times 1$ vector with all unity elements. We have

$$\begin{aligned}
\tilde{\mathbf{Q}}_{dir}(\mathbf{r}, f) &= \tilde{\mathbf{d}}_{dir}(\mathbf{r}, f) \tilde{\mathbf{d}}_{dir}^H(\mathbf{r}, f) \\
&= (\mathbf{a}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) + \mathbf{d}_{dir}(\mathbf{r}, f)) \\
&\quad \times (\mathbf{a}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) + \mathbf{d}_{dir}(\mathbf{r}, f))^H \\
&= \mathbf{Q}_{dir}(\mathbf{r}, f) + (\mathbf{a}_{dir}(\mathbf{r}, f) \mathbf{a}_{dir}^H(\mathbf{r}, f) + \mathbf{a}_{dir}(\mathbf{r}, f) \mathbf{R}_{dir}^H(\mathbf{r}, f) \\
&\quad + \mathbf{R}_{dir}(\mathbf{r}, f) \mathbf{a}_{dir}^H(\mathbf{r}, f)) \otimes \mathbf{d}_0(f) \mathbf{d}_0^H(f)
\end{aligned} \tag{5.13}$$

$$\begin{aligned}
\tilde{\mathbf{Q}}_{rev}(\mathbf{r}, f) &= \tilde{\mathbf{d}}_{rev}(\mathbf{r}, f) \tilde{\mathbf{d}}_{rev}^H(\mathbf{r}, f) \\
&= (\mathbf{a}_{rev}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) + \mathbf{d}_{rev}(\mathbf{r}, f)) \\
&\quad \times (\mathbf{a}_{rev}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) + \mathbf{d}_{rev}(\mathbf{r}, f))^H \\
&= \mathbf{Q}_{rev}(\mathbf{r}, f) + (\mathbf{a}_{rev}(\mathbf{r}, f) \mathbf{a}_{rev}^H(\mathbf{r}, f) + \mathbf{a}_{rev}(\mathbf{r}, f) \mathbf{R}_{rev}^H(\mathbf{r}, f) \\
&\quad + \mathbf{R}_{rev}(\mathbf{r}, f) \mathbf{a}_{rev}^H(\mathbf{r}, f)) \otimes \mathbf{d}_0(f) \mathbf{d}_0^H(f)
\end{aligned} \tag{5.14}$$

where $\mathbf{Q}_{dir}(\mathbf{r}, f) = \mathbf{d}_{dir}(\mathbf{r}, f) \mathbf{d}_{dir}^H(\mathbf{r}, f)$ and $\mathbf{Q}_{rev}(\mathbf{r}, f) = \mathbf{d}_{rev}(\mathbf{r}, f) \mathbf{d}_{rev}^H(\mathbf{r}, f)$. Also,

$$\begin{aligned}
\tilde{\mathbf{p}}_{dir} &= \tilde{\mathbf{d}}_{dir}(\mathbf{r}, f) G_d^H(\mathbf{r}, f) \\
&= \mathbf{p}_{dir}(\mathbf{r}, f) + (\mathbf{a}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f)) G_d^H(\mathbf{r}, f)
\end{aligned} \tag{5.15}$$

where $\mathbf{p}_{dir}(\mathbf{r}, f) = \mathbf{d}_{dir}(\mathbf{r}, f) G_d^H(\mathbf{r}, f)$. We now consider a random matrix that contains the perturbation elements for the direct path,

$$\mathbf{\Xi}_{dir} = \mathbf{a}_{dir} \mathbf{a}_{dir}^H. \tag{5.16}$$

Here, the term (\mathbf{r}, f) is dropped from \mathbf{a}_{dir} for convenience. Now as we aim to use mean performance optimization technique by using probability density function of the gain and phase

$$\begin{aligned}
\bar{\Xi}_{dir} &= E[\Xi_{dir}] = E[\mathbf{a}_{dir}\mathbf{a}_{dir}^H] \\
&= \int_{a_1} \cdots \int_{a_M} \Xi_{dir} f_{\Xi}(a_1) \cdots f_{\Xi}(a_M) da_1 \cdots da_M
\end{aligned} \tag{5.17}$$

and

$$\begin{aligned}
\bar{\mathbf{a}}_{dir} &= E[\mathbf{a}_{dir}] \\
&= \int_{a_1} \cdots \int_{a_M} \mathbf{a}_{dir} f_{\Xi}(a_1) \cdots f_{\Xi}(a_M) da_1 \cdots da_M.
\end{aligned} \tag{5.18}$$

where $f_{\Xi}(a_m)$, $1 \leq m \leq M$, is the PDF for m^{th} sensor's errors. In order to simplify the design problem we assume that each sensor's error is independent of frequency and space. Then, using expectation on Eq. 5.13 to Eq. 5.15

$$\begin{aligned}
\bar{\mathbf{Q}}_{dir}(\mathbf{r}, f) &= E[\tilde{\mathbf{Q}}_{dir}(\mathbf{r}, f)] \\
&= (\bar{\Xi}_{dir} + \bar{\mathbf{a}}_{dir}\mathbf{R}_{dir}^H(\mathbf{r}, f) + \mathbf{R}_{dir}(\mathbf{r}, f)\bar{\mathbf{a}}_{dir}^H) \otimes \mathbf{d}_0(f)\mathbf{d}_0^H(f) + \mathbf{Q}_{dir}(\mathbf{r}, f)
\end{aligned} \tag{5.19}$$

$$\begin{aligned}
\bar{\mathbf{Q}}_{rev}(\mathbf{r}, f) &= E[\tilde{\mathbf{Q}}_{rev}(\mathbf{r}, f)] \\
&= ((\bar{\Xi}_{rev} + \bar{\mathbf{a}}_{rev}\mathbf{R}_{rev}^H(\mathbf{r}, f) + \mathbf{R}_{rev}(\mathbf{r}, f)\bar{\mathbf{a}}_{rev}^H) \otimes \mathbf{d}_0(f)\mathbf{d}_0^H(f)) + \mathbf{Q}_{rev}(\mathbf{r}, f)
\end{aligned} \tag{5.20}$$

and

$$\begin{aligned}
\bar{\mathbf{p}}_{dir}(\mathbf{r}, f) &= E[\tilde{\mathbf{p}}_{dir}(\mathbf{r}, f)] \\
&= (\bar{\mathbf{a}}_{dir} \otimes \mathbf{d}_0(f))G_d^H(\mathbf{r}, f) + \mathbf{p}_{dir}(\mathbf{r}, f)
\end{aligned} \tag{5.21}$$

For simplicity, we assume that the error model for the direct path and the reverberation part are the same. As such,

$$\bar{\Xi}_{dir} = \bar{\Xi}_{rev} = \bar{\Xi}. \tag{5.22}$$

and

$$\bar{\mathbf{a}}_{dir} = \bar{\mathbf{a}}_{rev} = \bar{\mathbf{a}}. \tag{5.23}$$

The matrix $\bar{\Xi}$ can be written as

$$\bar{\Xi} = (\bar{\mathbf{a}}\bar{\mathbf{a}}^H) \odot (\mathbf{1}_M\mathbf{1}_M^T - \mathbf{I}_M) + \sigma \tag{5.24}$$

where σ denotes the variance of the gain pdf.

By assuming the gain and phase errors are independent, the expectation of the error vector $\bar{\mathbf{a}}$ can be simplified into

$$\bar{\mathbf{a}} = \bar{\mathbf{a}}_\rho \odot (\bar{\mathbf{a}}_\gamma^c + j\bar{\mathbf{a}}_\gamma^s) \quad (5.25)$$

where

$$\bar{\mathbf{a}}_\rho = \int \mathbf{a}_\rho f_{\Xi_\rho}(\mathbf{a}_\rho) d(\mathbf{a}_\rho) \quad (5.26)$$

$$\bar{\mathbf{a}}_\gamma^c = \int \cos(\mathbf{a}_\gamma) f_{\Xi_\gamma}(\mathbf{a}_\gamma) d(\mathbf{a}_\gamma) \quad (5.27)$$

$$\bar{\mathbf{a}}_\gamma^s = \int \sin(\mathbf{a}_\gamma) f_{\Xi_\gamma}(\mathbf{a}_\gamma) d(\mathbf{a}_\gamma) \quad (5.28)$$

The robust weighted least square error can be given as

$$J_{mod,WLS,rb}(\mathbf{w}_{rb}) = \mathbf{w}_{rb}^T \bar{\mathbf{Q}}_{rb} \mathbf{w}_{rb} - 2\bar{\mathbf{p}}_{rb}^T \mathbf{w}_{rb} + const \quad (5.29)$$

where

$$\begin{aligned} \bar{\mathbf{Q}}_{rb} &= \int_{\Omega} \int_{\mathcal{R}} (V_1(\mathbf{r}, f) \Re \{ \bar{\mathbf{Q}}_{dir}(\mathbf{r}, f) \} + V_2(\mathbf{r}, f) \Re \{ \bar{\mathbf{Q}}_{rev}(\mathbf{r}, f) \}) d\mathbf{r} df \\ \bar{\mathbf{p}}_{rb} &= \int_{\Omega} \int_{\mathcal{R}} V_1(\mathbf{r}, f) \Re \{ \bar{\mathbf{p}}_{dir}(\mathbf{r}, f) \} d\mathbf{r} df. \end{aligned}$$

The design of robust beamformer can be obtained by

$$\mathbf{w}_{rb} = \bar{\mathbf{Q}}_{rb}^{-1} \bar{\mathbf{p}}_{rb}. \quad (5.30)$$

5.4.2 Multiplicative error model

We now develop a robust beamformer design using a stochastic multiplicative error model to microphone characteristics (gain and phase) instead of the additive error model, i.e

$$\begin{aligned} \tilde{\mathbf{R}}_{dir}(\mathbf{r}, f) &= \mathbf{R}_{dir}(\mathbf{r}, f) \odot \mathbf{a}(\mathbf{r}, f) \\ \tilde{\mathbf{R}}_{rev}(\mathbf{r}, f) &= \mathbf{R}_{rev}(\mathbf{r}, f) \odot \mathbf{a}(\mathbf{r}, f) \end{aligned} \quad (5.31)$$

where \odot denotes the element-by-element product. Following from (5.2), the perturbed array response vector is given by

$$\begin{aligned} \tilde{\mathbf{d}}_{dir}(\mathbf{r}, f) &= \tilde{\mathbf{R}}_{dir}(\mathbf{r}, f) \otimes \mathbf{d}_0(f) \\ &= (\mathbf{a}(\mathbf{r}, f) \otimes \mathbf{1}_L) \odot \mathbf{d}_{dir}(\mathbf{r}, f) \end{aligned} \quad (5.32)$$

and

$$\tilde{\mathbf{d}}_{rev}(\mathbf{r}, f) = (\mathbf{a}(\mathbf{r}, f) \otimes \mathbf{1}_L) \odot \mathbf{d}_{rev}(\mathbf{r}, f)$$

where $\mathbf{1}_L$ is an $L \times 1$ vector with all unity elements. We have

$$\begin{aligned} \tilde{\mathbf{Q}}_{dir}(\mathbf{r}, f) &= \tilde{\mathbf{d}}_{dir}(\mathbf{r}, f) \tilde{\mathbf{d}}_{dir}^H(\mathbf{r}, f) \\ &= (\mathbf{a}(\mathbf{r}, f) \mathbf{a}^H(\mathbf{r}, f) \otimes \mathbf{1}_L \mathbf{1}_L^T) \\ &\quad \odot \mathbf{Q}_{dir}(\mathbf{r}, f) \end{aligned} \quad (5.33)$$

$$\begin{aligned} \tilde{\mathbf{Q}}_{rev}(\mathbf{r}, f) &= (\mathbf{a}(\mathbf{r}, f) \mathbf{a}^H(\mathbf{r}, f) \otimes \mathbf{1}_L \mathbf{1}_L^T) \\ &\quad \odot \mathbf{Q}_{rev}(\mathbf{r}, f) \end{aligned} \quad (5.34)$$

where $\mathbf{Q}_{dir}(\mathbf{r}, f) = \mathbf{d}_{dir}(\mathbf{r}, f) \mathbf{d}_{dir}^H(\mathbf{r}, f)$ and $\mathbf{Q}_{rev}(\mathbf{r}, f) = \mathbf{d}_{rev}(\mathbf{r}, f) \mathbf{d}_{rev}^H(\mathbf{r}, f)$. Also,

$$\begin{aligned} \tilde{\mathbf{p}}_{dir} &= \tilde{\mathbf{d}}_{dir}(\mathbf{r}, f) G_d^H(\mathbf{r}, f) \\ &= (\mathbf{a}(\mathbf{r}, f) \otimes \mathbf{1}_L) \odot \mathbf{p}_{dir}(\mathbf{r}, f) \end{aligned} \quad (5.35)$$

where $\mathbf{p}_{dir}(\mathbf{r}, f) = \mathbf{d}_{dir}(\mathbf{r}, f) G_d^H(\mathbf{r}, f)$. Then, by following the same procedure as in the additive error model in Section 5.4.1

$$\begin{aligned} \bar{\mathbf{Q}}_{dir}(\mathbf{r}, f) &= E \left[\tilde{\mathbf{Q}}_{dir}(\mathbf{r}, f) \right] \\ &= (\bar{\boldsymbol{\Xi}} \otimes \mathbf{1}_L \mathbf{1}_L^T) \odot \mathbf{Q}_{dir}(\mathbf{r}, f) \end{aligned} \quad (5.36)$$

and

$$\begin{aligned} \bar{\mathbf{p}}_{dir}(\mathbf{r}, f) &= E [\tilde{\mathbf{p}}_{dir}(\mathbf{r}, f)] \\ &= (\bar{\mathbf{a}} \otimes \mathbf{1}_L) \odot \mathbf{p}_{dir}(\mathbf{r}, f) \end{aligned} \quad (5.37)$$

The robust weighted least square error can be given as

$$J_{mod,WLS,rb}(\mathbf{w}) = \mathbf{w}^T \bar{\mathbf{Q}}_{rb} \mathbf{w}_{rb} - 2\Re \{ \bar{\mathbf{p}}_{rb}^H \mathbf{w}_{rb} \} + const \quad (5.38)$$

where

$$\begin{aligned} \bar{\mathbf{Q}}_{rb} &= (\bar{\boldsymbol{\Xi}} \otimes \mathbf{1}_L \mathbf{1}_L^T) \odot \mathbf{Q}_{mod,WLS} \\ \bar{\mathbf{p}}_{rb} &= (\bar{\mathbf{a}} \otimes \mathbf{1}_L) \odot \mathbf{p}_{mod,WLS}. \end{aligned}$$

The design of robust beamformer can be obtained by

$$\mathbf{w}_{rb} = \bar{\mathbf{Q}}_{rb}^{-1} \bar{\mathbf{p}}_{rb}. \quad (5.39)$$

5.5 Aperture Size Optimization

So far the formulation of the beamformer design problem has only considered one specific array size and configuration as in Eq. (5.8). But it is well established that there is an impact of array aperture size on the design performance. Beamformer design problem in this case can be formulated as a minimization of cost function with respect to filter coefficients \mathbf{w} and interelement space between adjacent microphones (d) which can be written as [124]

$$J_{\text{WLS,opt}}(\mathbf{w}, d) = \int_{\Omega} \int_{\mathcal{R}} V(\mathbf{r}, f) |G(\mathbf{r}, f, d) - G_d(\mathbf{r}, f)|^2 d\mathbf{r}df \quad (5.40)$$

This problem can be solved by combining Weighted Least Square and Golden Section Search optimization [125, 126] techniques, by first optimizing the cost function with respect to \mathbf{w} while searching for the optimal interelement space (d). Algorithm 1 shows how this combined optimization has been performed.

Algorithm 1 Array aperture size optimization algorithm

Step 1: Initialize an interval $[d_l, d_u]$ and tol sufficiently small

Step 2: Set the golden ratio $\varphi = (\sqrt{5} - 1)/2$

Step 3: Set intermediate points, $a = d_u - \varphi * (d_u - d_l)$ and $b = d_l + \varphi * (d_u - d_l)$

Step 4: Evaluate the function at the intermediate points $f(a) = J_{\text{WLS}}(a)$ and $f(b) = J_{\text{WLS}}(b)$

Step 5: **While** $((a - b) > tol)$, number

If $f(a) < f(b)$ then update the intermediate points

$[d_u = b], [b = a]$ and $a = d_u - \varphi * (d_u - d_l)$

else

$[d_l = a], [a = b]$ and $b = d_l + \varphi * (d_u - d_l)$

end

Step 6: Evaluate the functions in the updated points

end

Step 7: The minimum occurs at $d = (d_u + d_l)/2$

5.6 Objective Measurements

There are different objective measurements in the literature to evaluate the performance of the beamformer designs. For dereverberation performance, objective measurements can be classified into two categories: channel based measure-

ment [116] and signal based measurement [79]. In this chapter, for the channel based measurement, we used direct to reverberant ratio measurement to evaluate the dereverberation ability of the designed beamformers. The direct to reverberant ratio, DRR, is defined as follows [127]

$$DRR = 20 \log_{10} \frac{DRR_{out}}{DRR_{in}} \text{ [dB]}. \quad (5.41)$$

Where

$$DRR_{out} = \frac{\|\mathbf{w}^T \mathbf{d}_{dir}(\mathbf{r}, f)\|_2^2}{\|\mathbf{w}^T \mathbf{d}(\mathbf{r}, f) - \mathbf{w}^T \mathbf{d}_{dir}(\mathbf{r}, f)\|_2^2} \quad (5.42)$$

$$DRR_{in} = \frac{\|\mathbf{1}^T \mathbf{R}_{dir}(\mathbf{r}, f)\|_2^2}{\|\mathbf{1}^T (\mathbf{R}(\mathbf{r}, f) - \mathbf{R}_{dir}(\mathbf{r}, f))\|_2^2} \quad (5.43)$$

where $\mathbf{1}$ is an M element vector with ones and $\|(\cdot)\|_2^2$ denotes $\int \int_{\mathcal{P}} |(\cdot)|^2 d\mathbf{r}df$, $\forall(\mathbf{r}, f) \in \mathcal{P}$ where \mathcal{P} is the passband region.

For the signal based measurement, segmental signal to noise and reverberation ratio SSNRR is used to measure the speech distortion because of noise and reverberation. [128]. It can be formulated as

$$SSNRR_{seg} = \frac{1}{N_{seg}} \sum_{l=0}^{N_{seg}-1} 10 \log_{10} \left(\frac{\|\mathbf{s}_d(l)\|_2^2}{\|\mathbf{s}_d(l) - \mathbf{y}(l)\|_2^2} \right) \text{ [dB]} \quad (5.44)$$

where $\mathbf{s}_d(l)$ represents desired signal, $\mathbf{y}(l)$ represents the estimated speech from the beamformer output, and N_{seg} denotes the number of signal segments. This is obtained by computing desired and estimated signals as short overlapping signal segments and then an average of SSNRR values in dB is taken over all segments. Moreover, to test the overall suppression performance in the stopband region, signal suppression measurement is used as follows [69]

$$SUPP = 10 \log_{10} \frac{\|S(f)\|_2^2}{\|Y(f)\|_2^2} \quad (5.45)$$

where $S(f)$ and $Y(f)$ denote the frequency spectrum of the input signal and the output signal, respectively. Furthermore, $\|(\cdot)\|_2^2$ denotes $\int_{\mathcal{F}} |(\cdot)|^2 df$, $\forall(f) \in \mathcal{F}$ where \mathcal{F} is the passband region.

5.7 Design Examples

This section presents a number of design examples with the aim to verify the beamformer design formulations in Section 3 (indoor beamformer design) and

Section 4 (robust beamformer design) using simulated data (room impulse response) and real data. The parameters used in the simulation is given in Table 5.6. Those are the parameter values used unless otherwise specified. The frequency domain expression of the room impulse response is computed using Eq. (5.6). As a special case, direct path based beamformer is designed by using Eq. (5.6) with a reverberation time $T_{60} = 0 s$, i.e. room response consists of direct path response only. Whereas, indoor beamformers are designed with reverberation time $T_{60} = 0.2 s$

Eq. (5.10) is used for direct path and indoor beamformer designs. For robust direct path and indoor beamformer design examples, mean performance optimization method is used with amplitude and phase variation both following a uniform distribution with intervals $[\pm 10\% \mathbf{R}(\mathbf{r}, f)]$ and $[-0.1 \text{ rad}, 0.1 \text{ rad}]$, respectively. Eq. (5.11) is used as the perturbed response and Eq. (5.30) is used for the beamformer design. The designed beamformers were tested using simulated room impulse response from room acoustic simulator based on the ISM method [118, 119]. We define a simple rectangular room with dimensions $4m \times 8m \times 3m$ and uniform absorption coefficients characterizing the room surfaces. The pass-band region is given as

$$\mathcal{P} = \{x = 1m, 3.5m \leq y \leq 4.5m, z = 1m, 200kHz \leq f \leq 3800Hz\}$$

while the stop band region is

$$\mathcal{S} = \{x = 1m, 3.5m \leq y \leq 4.5m, z = 1m, 3850Hz \leq f \leq 4kHz\} \\ \cup \{x = 1m, 1m \leq y \leq 2.5m \cup 5.5m \leq y \leq 7m, z = 1m, 100Hz \leq f \leq 4000Hz\}.$$

Different scenarios are presented to evaluate the designed beamformers. First, the cost function and the beampattern performances in varying reverberant environments are evaluated, then the suppression performance in stopband region for different reverberation conditions are evaluated, finally, the joint de-reverberation and noise suppression performance in environments which included both noise and reverberation are evaluated using estimated Room Impulse Response (RIR) or measured (RIR) [129].

5.7.1 Overall performance and cost function evaluation for different reverberation time

The four design methods are evaluated by calculating the amplitude response of the overall beamformer including the room response according to Eq. (5.4) as a function of spatial coordinate and frequency. The designs have been evaluated for ($T_{60} = 0.1 s$) and ($T_{60} = 0.2 s$).

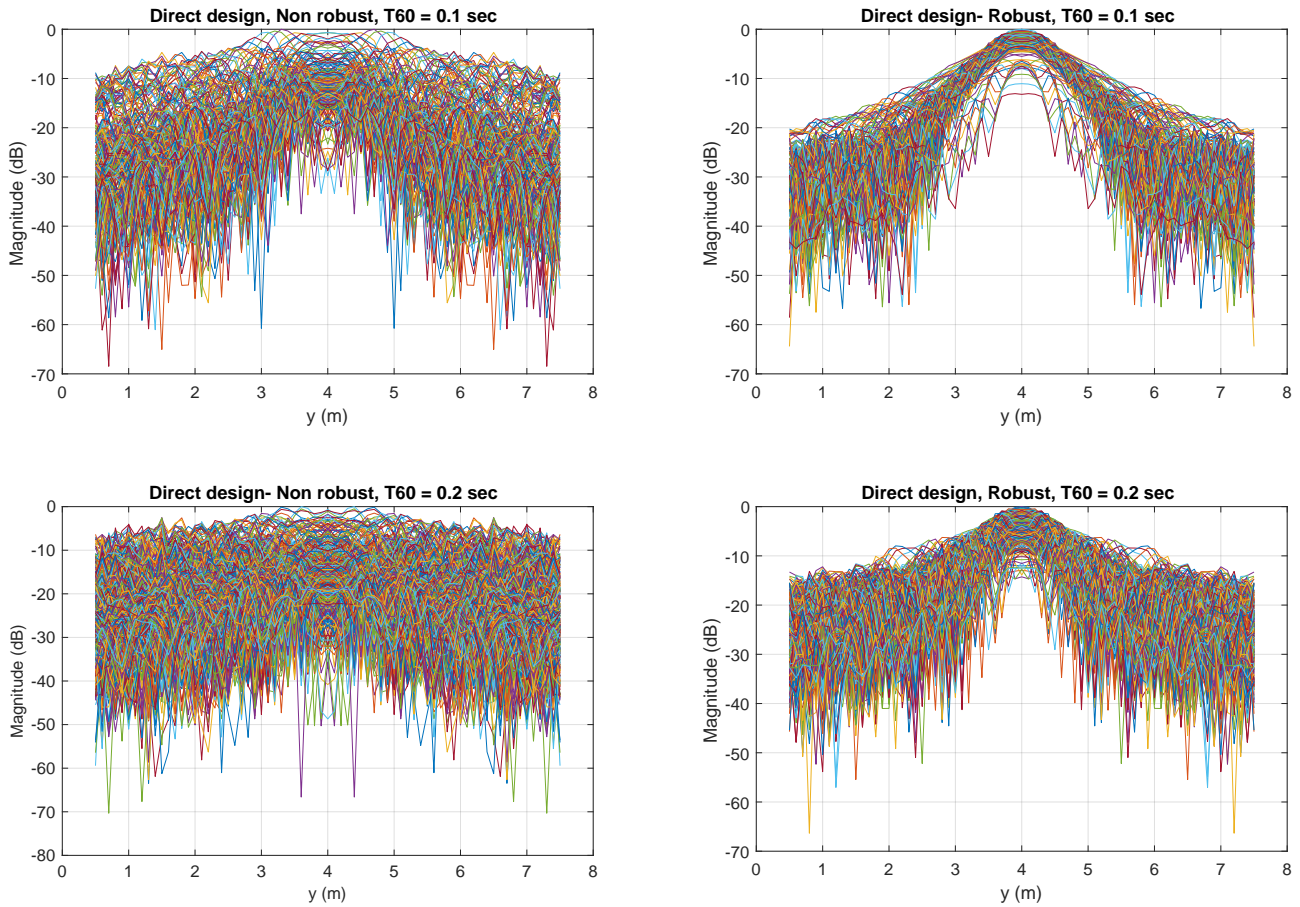


Figure 5.2: Magnitude response of direct beamformer and robust direct beamformer using additive model under different reverberation time.

Figure 5.2 shows the magnitude frequency response of direct path and robust direct beamformer designs applied for different reverberation time. Similarly, Figure 5.3 shows the magnitude frequency response of indoor design and robust indoor design applied for different reverberation time. It can be seen from the figures that the robust direct path beamformer design has a similar performance as the indoor design response while the direct path beamformer response performance deteriorates in the presence of the reverberation. As such a simple robust direct path beamformer can be employed to the indoor applications as it can achieve approximately the same performance as the indoor beamformer design while having a significantly lower computational complexity as the reverberation part is not included.

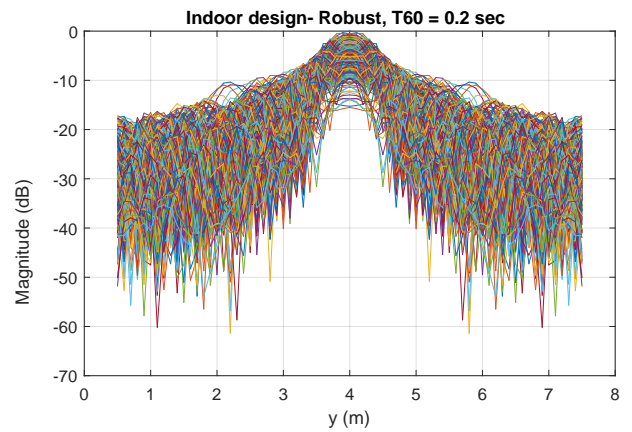
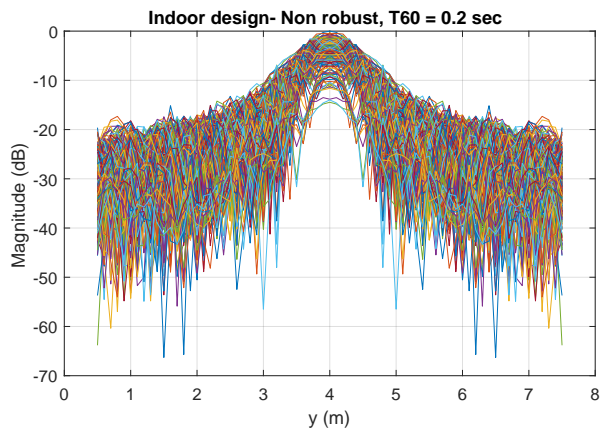
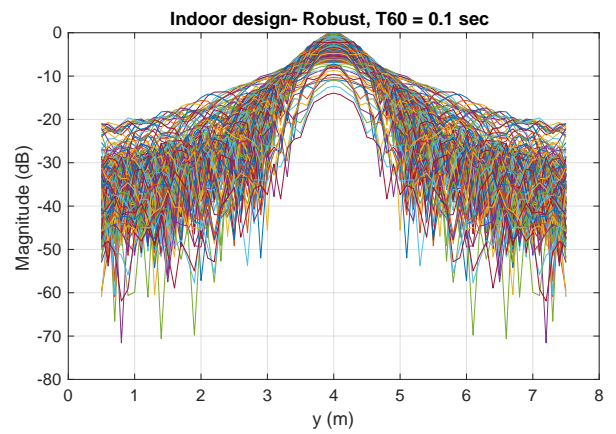
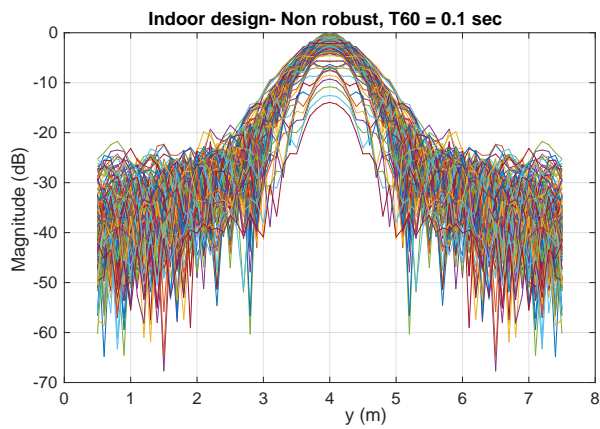


Figure 5.3: Magnitude response of indoor beamformer and robust indoor beamformer using additive model under different reverberation time.

Table 5.1 shows the values of cost function in Eq. (5.8) for different design methods. We evaluate the cost function for different reverberation times using the optimal weights with $T_{60} = 0 s$ and $T_{60} = 0.2 s$ for different design methods to get an impression of the sensitivity of the cost function for changing reverberation times. It can be seen from Table 5.1 that the cost function of the robust direct design follows a similar trend as the indoor beamformer, whereas the cost function of the direct design increased significantly with increasing reverberation time.

$T_{60}(\text{sec})$	Cost function of Direct Path beamformer design (dB)	Cost function of Robust Direct Path beamformer design (dB)	Cost function of Indoor beamformer design (dB)	Cost function of Robust Indoor beamformer design (dB)
0.1	-7.53	-23.22	-23.29	-20.98
0.15	-2.13	-20.56	-22.02	-20.24
0.2	1.07	-18.50	-20.80	-19.43
0.25	3.35	-16.98	-19.76	-18.68
0.3	5.08	-15.79	-18.22	-17.98
0.35	7.53	-14.69	-17.32	-16.76

Table 5.1: Comparison of the cost function for different reverberation times for the direct design and the robust direct design ($T_{60} = 0 s$) using additive model, the indoor design and the indoor robust design ($T_{60} = 0.2 s$) using additive model.

5.7.2 Dereverberation performance

In this section, we evaluated the performance of the designed beamformers in terms of DRR for different source distance and number of microphones while the distance between the microphones remained constant. We assume a noise free reverberant environment ($T_{60}=0.2$ sec). The direct to reverberant ratio has been studied as a function of the distance between the desired source and the microphone array as depicted in Figure 5.4 (left side). For the indoor beamformer design case, it can be clearly noticed that for 1.5 m array-source distance, designed beamformer achieves significantly better DRR scores than other designs. Robust direct design and robust indoor design show less sensitive reaction to the increasing in the distance between source and microphone array. Moreover, DRR has been studied as a function of the number of microphones as shown in Figure 5.4 (right side), significant improvements in DRR with growing number of microphone elements are obtained by the indoor designs and the robust direct design.

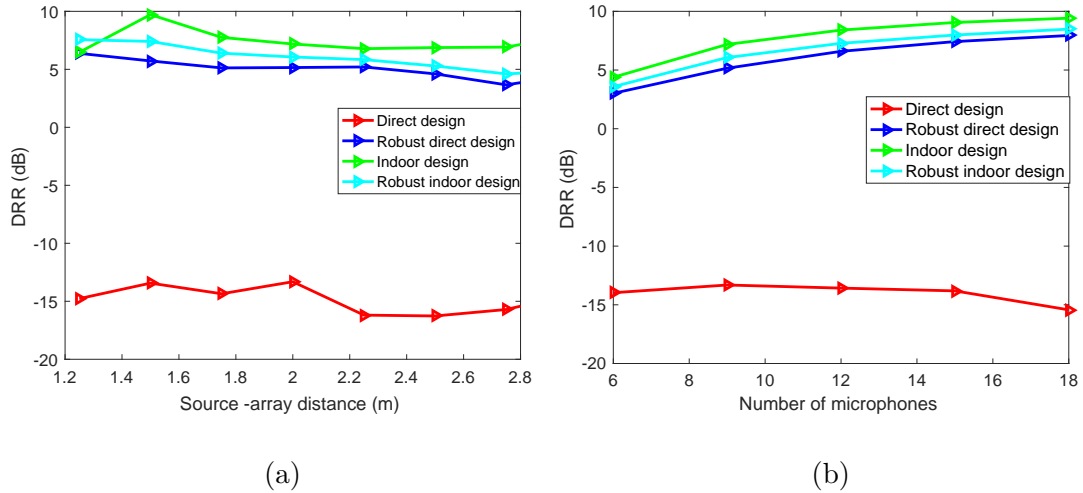


Figure 5.4: Direct to reverberant ratio performance under (a) different source-microphone array distance, (b) different number of microphones. Robust beamformers designed using additive error model.

Whereas, the direct design does not show any improvement.

5.7.3 Suppression performance in stopband region

In this section, we present a comparison of the interference suppression capabilities of the designed beamformers under different reverberation conditions. We use a female speech signal as an interference in the stopband region from position = (1, 6, 1)m. Table 5.2 shows the amplitude suppression results obtained from the different designed beamformers under varying reverberation times. It can be clearly observed that the indoor designs perform better than the direct design. Moreover, the robust direct design follows the same trend as the indoor design under different reverberation conditions. This demonstrates the suppression capability of the robust direct design in reverberant conditions.

$T_{60}(\text{sec})$	SUPP of Direct Path beamformer design (dB)	SUPP of Robust Direct Path beamformer design (dB)	SUPP of Indoor beamformer design (dB)	SUPP Robust Indoor beamformer design (dB)
0.1	-4.952	12.184	14.244	14.648
0.15	-5.989	10.247	12.579	13.323
0.2	-6.976	9.031	11.323	12.224
0.25	-7.271	8.210	10.401	11.389
0.3	-8.292	7.600	9.685	10.737
0.35	-8.750	7.113	9.112	10.206

Table 5.2: Comparison results among direct path based beamformer and its robust design ($T_{60} = 0 s$) using additive model, indoor beamformer and its robust design ($T_{60} = 0.2 s$) using additive model on the interference suppression at different reverberation time.

5.7.4 Joint dereverberation and noise suppression performance

Now the combined dereverberation and noise reduction performance for the designed beamformers are evaluated in terms of segmental signal to noise and reverberation ratio (SSNRR) [128], which is a measure of the distortion occurs due to the interference (noise and reverberation). The reverberant signals are generated using simulated room impulse response and measured room impulse response.

In this example, a linear microphone array with 8 elements with inter-element space of 0.08 m is placed in a reverberant room of size ($6m \times 6m \times 2.4m$) with variable reverberation time $T_{60} = 0.16 s$ and $T_{60} = 0.36 s$. The desired speaker is 1 m from the microphone array at angle 0° and the noise source is 1 m from the microphone array at angle 90° . The room setup is depicted in Figure 5.5. Different beamformer designs are tested both in simulated and real room environments. The noisy environment consists of reverberation and directional white noise source (jammer) with varied SJR levels (10-30 dB).

In the simulated room scenario, RIR is generated using image source method (ISM) [118, 119]. The reverberant signals received by the microphone array are obtained by convolving the simulated RIR with the source signal. For the real room environment evaluation, we used measured RIR [129]. The reverberant

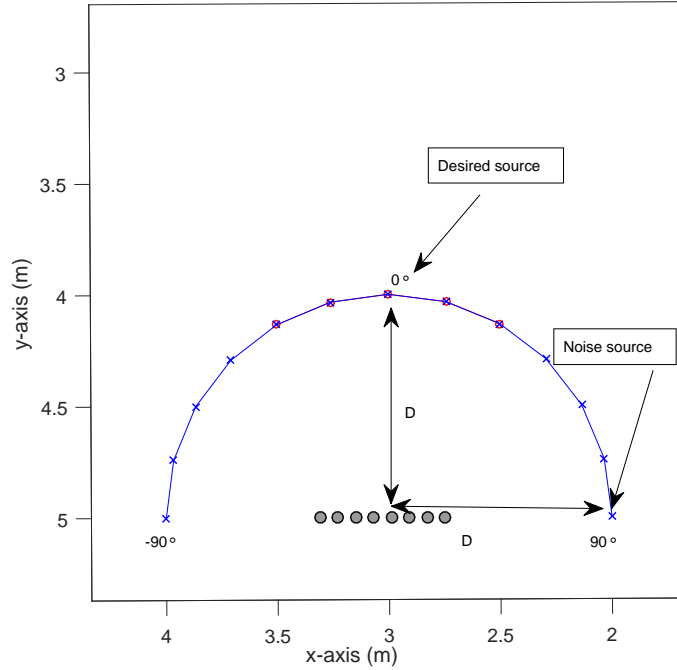


Figure 5.5: Room setup with a linear microphone array.

signals received by the microphone array are generated by convolving the speech signals with the measured room impulse response.

Figure 5.6 shows the results for the SSNRR using simulated RIR (left side) and measured RIR (right side) under different reverberation time values $T_{60} = 0.16 s$ and $T_{60} = 0.36 s$. From the simulated results, it can be clearly observed that the SSNRR results that are obtained by the indoor design and the direct designs are almost identical at $T_{60}=0.16$ sec. In higher reverberation time $T_{60} = 0.36 s$ the SSNRR results that are obtained by the indoor designs and the robust direct design are much higher than those obtained by the direct design. The SSNRR results that are obtained by using measured RIR show that robust direct design performs almost identical as the indoor design and better than the direct design under low reverberation time $T_{60} = 0.16 s$. Moreover, robust indoor beamformer design shows significantly better results compared to direct designs and indoor design. However, for higher reverberation time $T_{60} = 0.36 s$ indoor designs perform clearly better than direct designs. Although robust direct design shows similar results to direct design in SJR level < 10 dB, SSNRR starts to increase at SJR level > 15 dB.

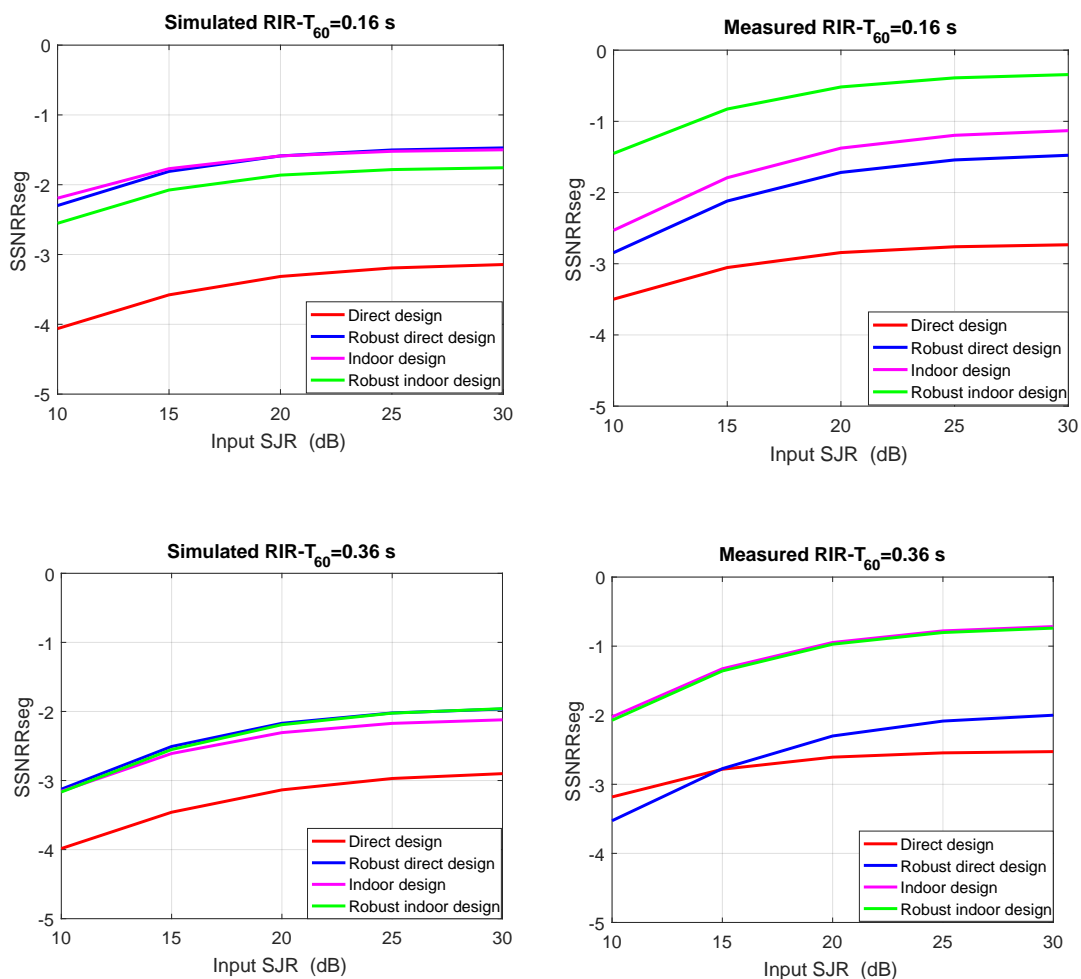


Figure 5.6: SSNRR results obtained for different beamformer designs for varying SJR's and reverberation times. The reverberant signals were generated using simulated RIR (left side) and measured RIR (right side) with different reverberation time, $T_{60} = 0.16$ s (top) and $T_{60} = 0.36$ s (bottom). Robust beamformers designed using additive error model.

5.7.5 Sensitivity test of beamformer designs

5.7.5.1 Perturbation in microphone characteristics

The next evaluation is on the sensitivity of the designed beamformers against gain and phase mismatches in microphone characteristics. This evaluation is done by performing a Monte-Carlo simulation of the gain and the phase mismatches and evaluate the cost function with fix beamformer coefficients for each simulation round. In Figure 5.7 the cost function distribution for the different beamformer designs in form of histograms are presented: (i) non-robust and robust direct path based beamformer with $T_{60} = 0 s$ and (ii) non-robust and robust indoor beamformer with $T_{60} = 0.2 s$. Robust broadband beamformer is designed using mean performance optimization method with uniform gain and phase distributions of $[0.997, 1.007]$ and $[-0.1, 0.1]$ rad, respectively. The cost function have been evaluated by using 100 Monte Carlo simulations with amplitude and phase errors of the array response vector $\mathbf{R}(\mathbf{r}, f)$. It can be seen from Figure 5.7 that non-robust designs (direct path and indoor) beamformers are sensitive to mismatches in microphone characteristics with the cost function values of non-robust direct path design deviate in the range $(-12.83 \text{ dB}, -5 \text{ dB})$, and the cost function values of non-robust indoor design deviate in the range $(-13.22 \text{ dB}, -8.35 \text{ dB})$. On the other hand, the robust direct path and robust indoor design are less sensitive to the mismatches in microphone characteristics as the cost function values deviate significantly less than the non-robust direct path and non-robust indoor designs. In order to explain the behavior of designed beamformers towards mismatches in microphone characteristics (gain and phase) we calculate the condition number of the correlation matrix Q for the different beamformer designs, see Table 5.3. The results show that the matrix Q for the robust direct path and indoor designs have a significantly lower condition number than non-robust direct path and indoor beamformer designs. Since the correlation matrix for the robust beamformer has lower condition number, it indicates that the solution is more numerically robust in the design. As such, the robust direct path and robust indoor designs are significantly less sensitive against errors in microphone characteristics than the direct path and indoor designs.

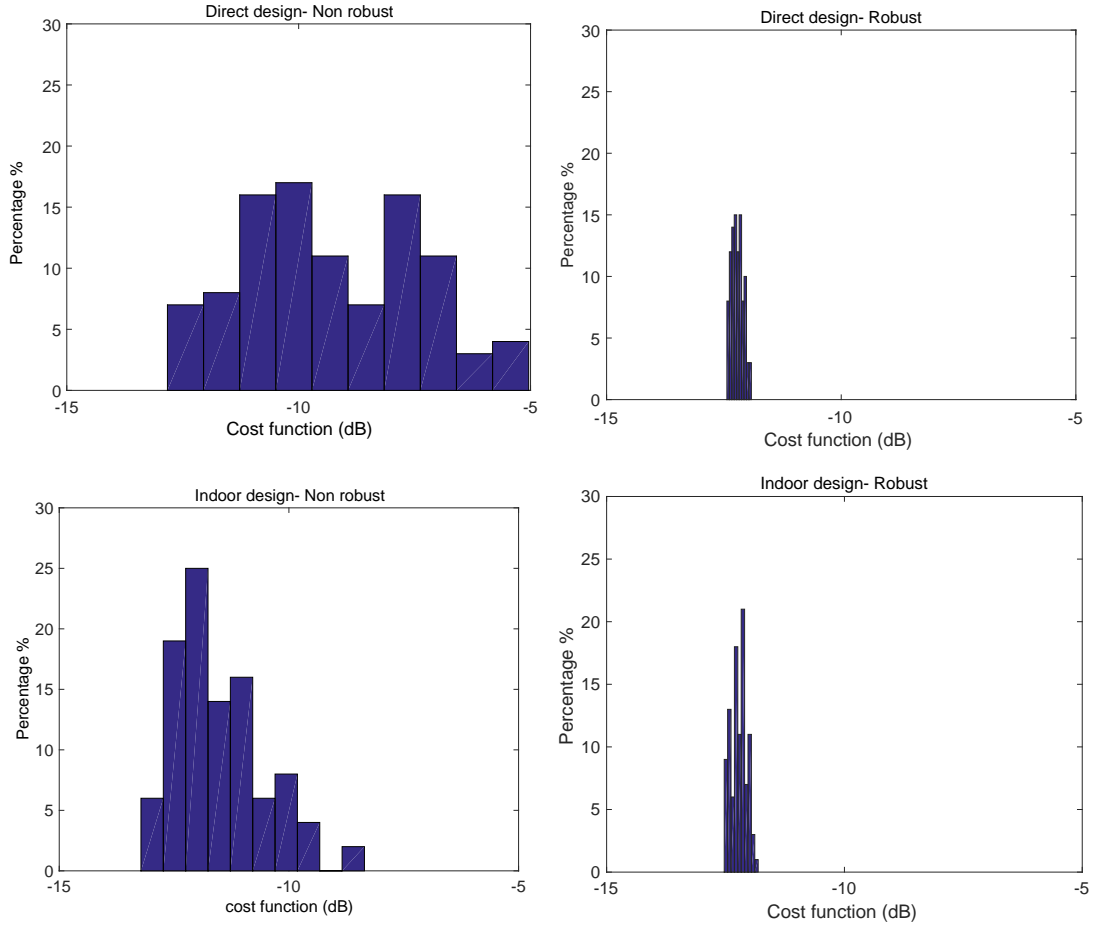


Figure 5.7: Histogram of cost function values distribution direct path based beamformer and its robust design using multiplicative error model ($T_{60} = 0 s$), indoor beamformer and its robust design using multiplicative error model ($T_{60} = 0.2 s$).

Beamformer design	Condition number of correlation matrix
Direct path	$2.1890e^{17}$
Robust direct path	$1.6265e^{05}$
Indoor	$1.6742e^{17}$
Robust indoor	$1.6873e^{05}$

Table 5.3: Comparison of condition number of correlation matrix among direct path based beamformer and its robust design ($T_{60} = 0 s$), indoor beamformer and its robust design ($T_{60} = 0.2 s$).

5.7.5.2 Evaluation for local scattering

In this section, an evaluation of the robustness towards local scattering is presented for all four design methods. In order to simulate local scattering we added 20 additional propagation paths to the direct propagation path, they were simulated using a uniform distribution for the angle of arrival and standard deviation $(-\pi/9, \pi/9)$, and gain with Rayleigh distribution and variance (0.01).

Figure 5.8 shows the histogram of the cost function averaged over 50 runs. It can be clearly seen that the robust designs demonstrate robustness against perturbed wave propagation compared to the direct design. This demonstrates the efficiency of the indoor designs and robust direct design against local scattering.

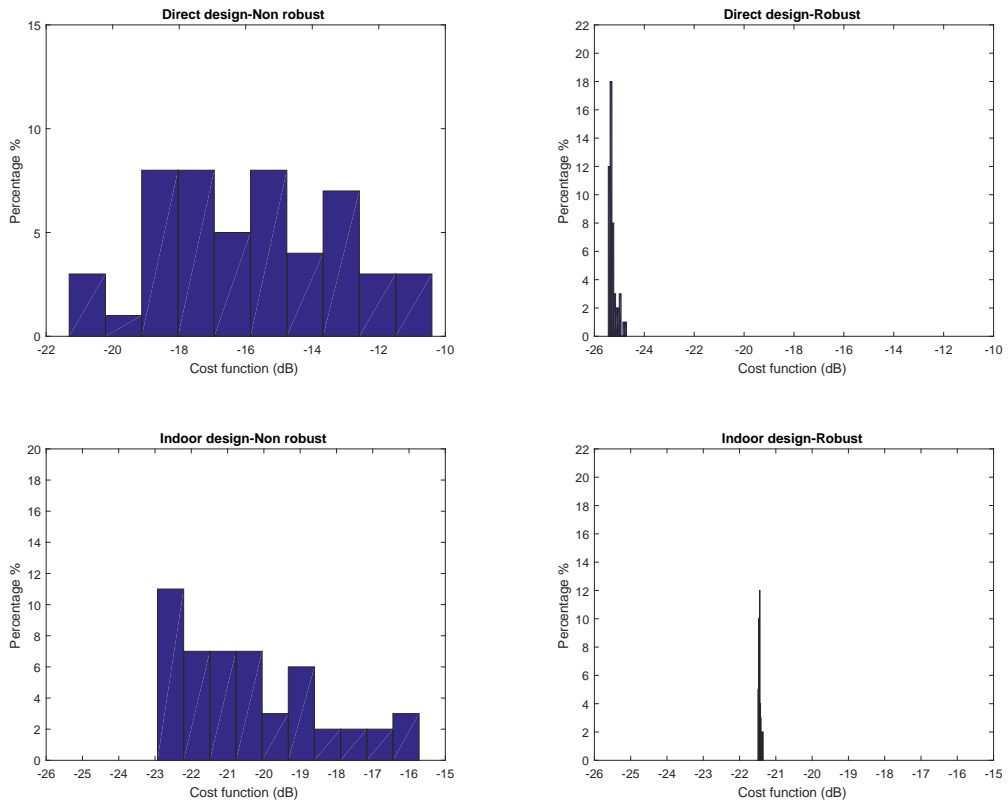


Figure 5.8: Histogram of cost function values distribution direct path based beamformer and its robust design using additive error model ($T_{60} = 0$ s), indoor beamformer and its robust design using additive error model ($T_{60} = 0.2$ s).

5.7.6 Evaluation of calculation time for different beamformer designs

Some applications needs a recalculation of the beamformer weights thus an interesting evaluations to compare the numerical complexity of the design. Table

5.4 shows the running time on a I7-4600 CPU 2.1 GHz and 8 Gbyte RAM for the different design methods and the different reverberation times. It can be clearly seen that the running time increase significantly with increasing reverberation time. In addition, the direct design and the robust direct design are significantly faster to calculate compared to the indoor designs.

Beamformer design	Calculation time (sec)
Direct path ($T_{60} = 0$)	4.796
Robust direct path ($T_{60} = 0$)	23.730
Indoor ($T_{60} = 0.1$)	38.077
Robust indoor ($T_{60} = 0.1$)	91.229
Indoor ($T_{60} = 0.2$)	141.424
Robust indoor ($T_{60} = 0.2$)	248.767
Indoor ($T_{60} = 0.3$)	506.926
Robust indoor ($T_{60} = 0.3$)	736.489

Table 5.4: Calculation time for different beamformer designs.

5.7.7 Results of aperture size optimization

Finally, we study the impact of array aperture size on the design performance as described in Algorithm 1 schedule. The Golden Section Search optimization technique has been used to search for an optimal inter-element spacing between microphones. We investigated all four design methods: (i) direct path ($T_{60} = 0$ s); (ii) indoor design with $T_{60} = 0.2$ s; (iii) robust direct path using multiplicative model; and (iv) robust indoor design using multiplicative model. Table 4 shows the cost function performance of the four different beamformer designs for inter-element spacing (d) range from 0.01 m to 0.2 m. It can be seen from the table that the direct beamformer designs have almost the same optimal inter-element space with optimal value ($d=0.11$ m). While the indoor designs have an optimal interelement space ($d=0.08$ m). Moreover, the designed beamformers are robust against inter-element spacing as the cost function values deviate very slightly with the changing of inter-element space as shown in Figure 5.9.

Beamformer design	Optimum Inter-element space (m)	Minimum cost function (dB)
Direct path	0.110	-31.977
Robust direct path	0.115	-24.557
Indoor	0.0836	-21.092
Robust indoor	0.0873	-19.173

Table 5.5: Array aperture size optimization for different beamformer designs.

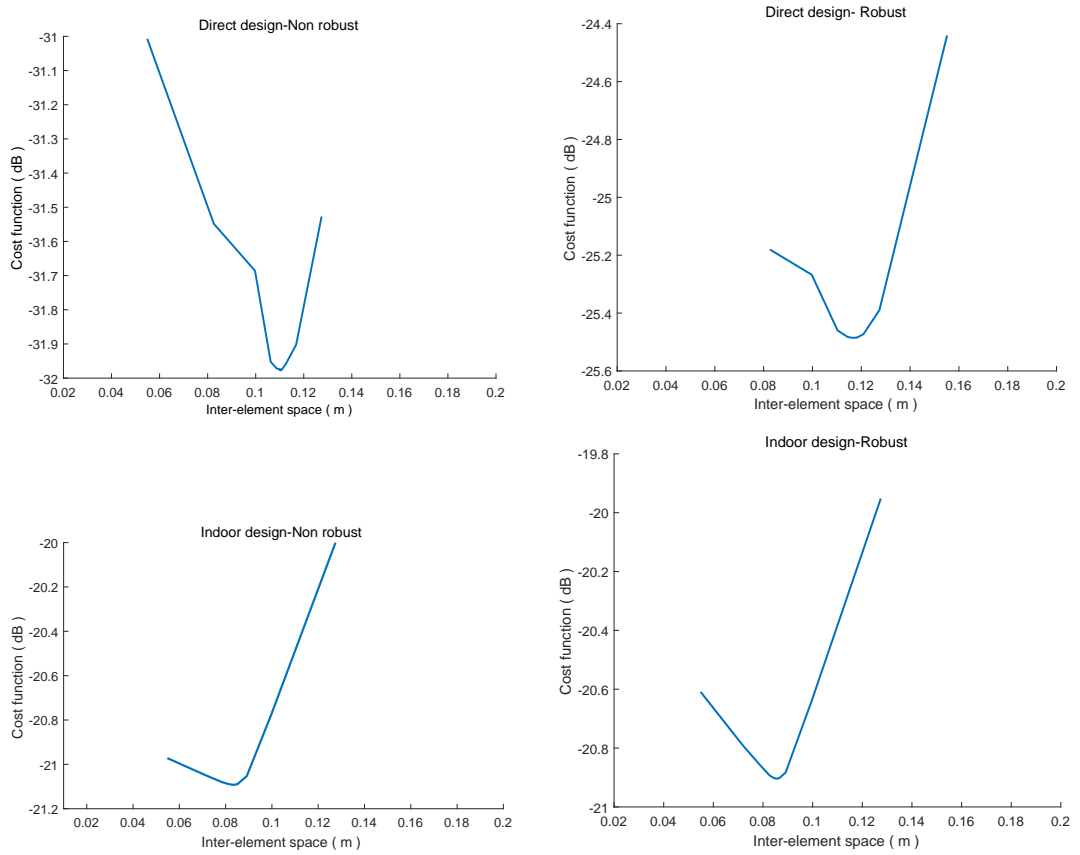


Figure 5.9: Cost function comparison for different inter-element spacing for direct path based beamformer and its robust design using multiplicative error model ($T_{60} = 0$ s), indoor beamformer and its robust design using multiplicative error model ($T_{60} = 0.2$ s)..

5.8 Summary

In this chapter, we have included robustness towards microphone characteristics (gain and phase) into the direct design and the indoor design. The indoor

design method employs a decomposition of the RIR into a direct path and reverberant path. To calculate the RIR, we have employed the ISM simulator. Numerical results show that robust direct path beamformer can achieve approximately the same performance as indoor beamformer design with a significantly lower computational complexity. Also, the robust direct path design is less sensitive to mismatches in microphone characteristics (gain and phase) than the indoor beamformer design. In addition, robust direct design is also robust to aperture size changes and follows the same trend as the indoor beamformer design.

Microphone array system parameters	Value
Number of elements, M	9
Interelement spacing	0.05 m
Position of elements	$(1.95, 3.95, 1)$, $(2, 3.95, 1)$, $(2.05, 3.95, 1)$, $(1.95, 4, 1)$, $(2, 4, 1)$ $, (2.05, 4, 1), (1.95, 4.05, 1), (2, 4.05, 1), (2.05, 4.05, 1)$
Sampling frequency, f_s	8 kHz
FIR filter length, L	21taps
Weighting functions V_1 and V_2	1

Table 5.6: Parameters for the evaluation of the microphone array.

Chapter 6

Conclusions and Future work

In noisy environments, speech communications would cease and deteriorate due to the noise contamination. Thus, speech enhancement systems are used as means to provide adequate and effective noise suppression to enhance the quality of speech in such adverse environments. The main goal of this dissertation is to develop speech enhancement techniques that provide the ability to suppress the background noise, room reverberation and other interference while preserving the original speech signal without too much distortion. In order to achieve this goal, different speech enhancement techniques have been investigated. In this chapter, we summarize the main conclusions drawn from this work, as well as highlight some suggestions for future research.

6.1 Summary

In this dissertation, we have developed single and multi-channel speech enhancement techniques that are applicable in many speech communication systems such as hands-free mobile phones and hearing aids. There are many considerations that have to be taken into account when designing such systems such as good noise reduction performance without too much speech distortion, quick response to abrupt changes in the observed noisy signal and low computational complexity for less power usage. The proposed speech enhancement techniques aim to address these considerations in the design procedure. Moreover, in the proposed techniques we focus on overcoming the drawbacks of the conventional speech enhancement systems. In single channel case, reduce the musical noise and improve the tracking speed of the a priori SNR estimation are the main problems we aim to reduce. Whereas in multi-channel speech enhancement technique, we focus on improving the indoor beamformer design by including the robustness against microphones characteristics (gain and phase) in order to reduce the sensitivity of the beamformer design towards deviation in such characteristics.

Chapter 2 discusses the importance of speech enhancement in many voice communication systems, particularly in adverse environments. Depending on the number of microphones, speech enhancement system can be classified into single channel and multi-channel speech enhancement techniques. In many applications, single channel speech enhancement techniques are preferred due to its simplicity and low cost. In addition, it provides sufficient noise suppression performance, but that comes at the cost of speech distortion and also musical noise. Besides the latter, single channel speech enhancement techniques usually exploit only the temporal and spectral diversity of the received signals. This might be a serious issue especially in reverberant environment, in which the reverberation induces the spatial diversity as well. In order to overcome this problem, multi-channel speech enhancement techniques are ideal solution due to their spatial filtering facility that helps to suppress the reverberation as well as background noise.

In Chapter 3 an adaptive averaging a priori SNR estimation technique employing critical band processing is proposed. This technique is based on a convex combination sigmoidal fusion function. Apart from combining the benefits of the conventional decision directed estimation (DD) and the modified decision directed (MDD) estimation, where a fixed weighting factor has been used, the fusion function in this approach provides a much faster adaptation when there is a speech input. This improved tracking capability of the abrupt changes in SNR improves

the preservation of weak speech components which is important for speech quality and intelligibility. The objective comparison and listening test both indicate that the proposed method is the preference approach over DD and MDD methods. Furthermore, the utilized critical band processing helps to achieve less musical noise because of its ability to significantly reduce the noise variance.

In Chapter 4, a modified adaptive smoothing factor with a frequency varying mean parameter is proposed to the modified decision directed a priori SNR estimation method MDD in order to examine the efficiency of the proposed a priori SNR estimation technique in STFT domain. Experimental results show the ability of the proposed method to preserve weak speech components in the high frequency region while maintain the same overall speech quality of MDD approach. Moreover, to assess the efficiency of the CB processing over the conventional uniform scale, a cross evaluation for STFT and CB processing is conducted by using subjective listening test. The subjective test results revealed the importance of CB processing in reducing the musical noise under different noise conditions and different levels of input SNR. In terms of speech quality and noise suppression, CB processing achieved better performance compared to STFT for WF gain function case and under non-stationary noise (babble noise) especially for low SNR levels.

In Chapter 5, the indoor beamforming design [69] is extended by including robustness against mismatches or deviations in microphone characteristics (gain and phase) using mean performance optimization. Multiplicative and additive stochastic error models are formulated and integrated into the indoor beamformer design. This extension provides a robust beamformer design in adverse environments. Design examples point out performance improvement in terms of significant reduction in error sensitivity for the robust indoor beamformer design formulation. Furthermore, evaluation results from the designed beamformers show that robust direct path based beamformer can achieve almost the same performance as the indoor beamformer design under different reverberant environments despite being a much simpler and faster design method.

6.2 Future work

Throughout this work, we developed single and multi channel speech enhancement techniques that are applicable in different applications such as hands-free mobile phones, hearing aids and teleconferencing systems. The main goals of the proposed technique are to extract the desired speech signal while mitigate the unwanted signals. Different objective measurements are used to evaluate the

performance of the developed algorithms in improving the speech quality and suppressing the noise signals. In the following subsections, several issues that may be addressed for further research are discussed.

6.2.1 Single channel speech enhancement techniques

In this dissertation, an improved a priori SNR estimation approach is introduced to control the adaptation speed of the a priori SNR estimation. For this purpose, a fusion function based on a sigmoidal shape is utilized as an adaptive weighting factor which yields better preserving of the weak speech components. Furthermore, a bark scale based filter bank is employed in the speech enhancement framework in order to reduce the noise variance which leads to a significant reduction in musical noise. Further investigation of the efficiency of the perceptual based filter bank in improving the speech intelligibility is of great importance. This might improve the applicability of the proposed technique in different signal processing systems such as speech recognition. Moreover, future research could focus on different choices of filter banks to further improve the performance of the speech enhancement techniques.

6.2.2 Multi-channel speech enhancement techniques

In this dissertation, we have discussed different beamformer designs for joint noise reduction and dereverberation. Robust broadband beamformer designs in adverse environment are proposed, which include room reverberation as well as robustness to microphone mismatches (amplitude and phase) in the design formulation. We evaluated the proposed beamformer designs using different objective measurements and under different acoustic conditions. However, one of the interesting future work for the proposed multi-channel speech enhancement techniques is to conduct a subjective listening test in order to justify the validity of the obtained objective results. Another pathway to improve the current work is to extend the design formulation to a steerable robust beamformer design in order to steer the main beam of the beamformer. Farrow filter structure can be used to formulate the steerable design [130]. In addition, further research can combine the robust indoor beamformer design with the proposed perceptually motivated single channel speech enhancement technique, which consists of applying a spectral gain to the beamformer output [122]. This may ultimately suppress the unwanted signals and improve the speech quality.

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma planned and carried out the matlab code for the design and simulations. In addition, I interpreted the results and wrote the manuscript to the following paper:

L. Khalid, S. Nordholm, and H. H. Dam, Design study on microphone arrays, IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015, pp. 1171-1175.



(Signature of Candidate)

I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.



(Full Name of Co-Author 1)



(Signature of Co-Author 1)

HAI HUYEN DAM

(Full Name of Co-Author 2)



(Signature of Co-Author 2)

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma carried out the main conception, design, evaluation and Interpretation of results. I wrote the manuscript to the following paper:

L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Convex combination framework for a priori snr estimation in speech enhancement, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), New Orleans, USA, 2017, pp. 4975 4979.



(Signature of Candidate)

I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.

PEI CHEE YONG

(Full Name of Co-Author 1)



(Signature of Co-Author 1)

HAI HUYEN DAM

(Full Name of Co-Author 2)



(Signature of Co-Author 2)

SVEN NORDHOLM

(Full Name of Co-Author 3)



(Signature of Co-Author 3)

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma carried out the design problem, performed the calculation and wrote the matlab codes. Moreover, I wrote the manuscript to the following paper:

L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Improved a priori snr estimation in speech enhancement, 23rd Asia-Pacific Conference on Communications (APCC), Perth, Australia, 2017, pp. 15.



(Signature of Candidate)

I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.

PEI CHEE YONG

(Full Name of Co-Author 1)



(Signature of Co-Author 1)

HAI HUYEN DAM

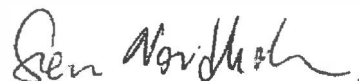
(Full Name of Co-Author 2)



(Signature of Co-Author 2)

SVEN NORDHOLM

(Full Name of Co-Author 3)



(Signature of Co-Author 3)

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma developed the theoretical formulation, performed the numerical simulation and interpreted the results. I wrote the manuscript to the following paper:

L. Nahma, H. H. Dam, S. Nordholm, "Robust beamformer design against mismatch in microphone characteristics and acoustic environments," International Workshop on Acoustic Echo and Noise Control (IWAENC), Tokyo, Japan, 2018, pp. 76-80.



(Signature of Candidate)

I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.

HAI HUYEN DAM

(Full Name of Co-Author 1)



(Signature of Co-Author 1)

SVEN NORDHOLM

(Full Name of Co-Author 2)



(Signature of Co-Author 2)

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma carried out the study of conception, performed the experiments and interpreted the results. Furthermore, I wrote the manuscript to the following paper:

L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, Cross evaluation of speech enhancement methods under different noise conditions, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp.895-899.



(Signature of Candidate)

I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.

PEI CHEE YONG

(Full Name of Co-Author 1)



(Signature of Co-Author 1)

HAI HUYEN DAM

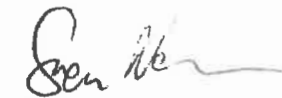
(Full Name of Co-Author 2)



(Signature of Co-Author 2)

SVEN NORDHOLM

(Full Name of Co-Author 3)



(Signature of Co-Author 3)

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma designed the proposed technique, the computational framework and analysed the data. Moreover, I planned and carried out the simulations and wrote the manuscript to the following paper:

L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, An adaptive a priori SNR estimator for perceptual speech enhancement, EURASIP journal on Audio, Speech and Music Processing, 2019,(1), p.7.



(Signature of Candidate)

I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.

PEI CHEE YONG

(Full Name of Co-Author 1)



(Signature of Co-Author 1)

HAI HUYEN DAM

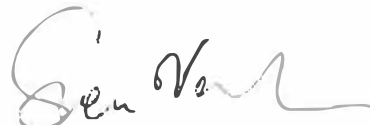
(Full Name of Co-Author 2)



(Signature of Co-Author 2)



(Full Name of Co-Author 3)



(Signature of Co-Author 3)

Statement of Contribution by Others

To Whom It May Concern I, Lara Nahma developed the theoretical formulation, performed the computation and the numerical simulations. Moreover, I carried out the experiments. In addition, I interpreted the results and wrote the manuscript to the following paper:

L. Nahma, H. H. Dam, Cedric Ka Fai Yiu, and S. Nordholm, Robust Broadband Beamformer Design for Noise Reduction and Dereverberation, Multidimensional Systems and Signal Processing (MSSP), 2019, pp.1-21.



(Signature of Candidate)

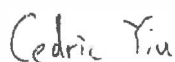
I, as a Co-Author, endorse that this level of contribution by the candidate indicated above is appropriate.



(Full Name of Co-Author 1)



(Signature of Co-Author 1)



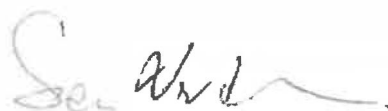
(Full Name of Co-Author 2)



(Signature of Co-Author 2)



(Full Name of Co-Author 3)



(Signature of Co-Author 3)

Bibliography

- [1] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson Upper Saddle River, 2011, vol. 64.
- [2] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.
- [3] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.
- [4] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.
- [5] H. Gustafsson, S. E. Nordholm, and I. Claesson, “Spectral subtraction using reduced delay convolution and adaptive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, 2001.
- [6] E. Referent, I. T. Fingscheidt, and I. R. Martin, “Speech enhancement using data-driven concepts,” *Braunschweig University*, 2012.
- [7] W. Charoenruengkit, *Spectral refinements to speech enhancement*. Florida Atlantic University, 2009.
- [8] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.
- [9] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
- [10] 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, 2013.

- [11] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 497–514, 1997.
- [12] S. Kortlang, S. D. Ewert, and T. Gerkmann, "Single channel noise reduction based on an auditory filterbank," in *proc. 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 283–287.
- [13] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2222–2232, 2006.
- [14] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126–137, 1999.
- [15] U. K. Laine, M. Karjalainen, and T. Altonsaar, "Warped linear prediction (wlp) in speech and audio processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.*, vol. 3, 1994, pp. III–349.
- [16] S. Haque and R. Togneri, "A psychoacoustic spectral subtraction method for noise suppression in automatic speech recognition," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 1618–1621.
- [17] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.
- [18] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligence magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [19] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.
- [20] M. Abbott, "The use of speech technology to enhance the handling of electronic flight progress strips in an air traffic control environment," *Proceedings of Voice Systems Worldwide*, pp. 126–134, 1990.

- [21] B. Beek and R. S. Vonusa, “General review of military applications of voice processing,” in *In AGARD Speech Process. 20 p (SEE N83-34179 22-32)*, 1983.
- [22] A. M. Vieira and I. C. d. Santos, “Communication skills: a mandatory competence for ground and airplane crew to reduce tension in extreme situations,” *Journal of Aerospace Technology and Management*, vol. 2, no. 3, pp. 361–370, 2010.
- [23] J. H. L. Hansen, “Analysis and compensation of stressed and noisy speech with application to robust automatic recognition,” 1988.
- [24] J. Karlsson, “The integration of automatic speech recognition into the air traffic control system,” Ph.D. dissertation, Massachusetts Institute of Technology, 1990.
- [25] C. J. Weinstein, “Opportunities for advanced speech processing in military computer-based systems,” in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.
- [26] M. R. Weiss and E. Aschkenasy, “The speech enhancement advanced development model,” Queens coll flushing NY dept. of computer science, Tech. Rep., 1978.
- [27] J. Woodard and E. Cupples, “Selected military applications of automatic speech recognition technology,” *IEEE Communications Magazine*, vol. 21, pp. 35–41, 1983.
- [28] S. J. Elliott and P. A. Nelson, “Active noise control,” *IEEE signal processing magazine*, vol. 10, no. 4, pp. 12–35, 1993.
- [29] C. Giguère, C. Laroche, A. Brammer, V. Vaillancourt, and G. Yu, “Advanced hearing protection and communication: Progress and challenges,” in *Proceedings of the 11 th International Congress on Noise as a Public Health Problem*, 2011, pp. 24–28.
- [30] A. J. Brammer, G. Yu, D. R. Peterson, E. R. Bernstein, and M. Cherniak, “Hearing protection and communication in an age of digital signal processing: Progress and prospects,” in *Proceedings of the 9th International Congress on Noise as a Public Health Problem*, 2008, pp. 21–25.

- [31] P. C. Yong, “Speech enhancement in binaural hearing protection devices,” Ph.D. dissertation, Curtin University, 2013.
- [32] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, “Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools,” *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.
- [33] N. Shanmugapriya and E. Chandra, “A thorough investigation on speech enhancement techniques for hearing aids,” *International Journal of Computer Applications*, vol. 99, no. 13, pp. 9–12, 2014.
- [34] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, “The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 63–72, 2013.
- [35] Z. Gong and Y. Xia, “Two speech enhancement-based hearing aid systems and comparative study,” in *5th IEEE International Conference on Information Science and Technology (ICIST)*, 2015, pp. 530–534.
- [36] A. Johansson, S. Nordholm, and E. Östlin, “Microphone array designed for a conference room,” Research Report, Australian Telecommunications Research Institute, Bentley, Western Australia, Tech. Rep., 2001.
- [37] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [38] Y. Hu and P. C. Loizou, “Speech enhancement based on wavelet thresholding the multitaper spectrum,” *IEEE transactions on Speech and Audio processing*, vol. 12, no. 1, pp. 59–67, 2004.
- [39] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.
- [40] T. Gülzow, A. Engelsberg, and U. Heute, “Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement,” *Signal processing*, vol. 64, no. 1, pp. 5–19, 1998.

- [41] H. W. Löllmann and P. Vary, “Uniform and warped low delay filter-banks for speech enhancement,” *Speech Communication*, vol. 49, no. 7-8, pp. 574–587, 2007.
- [42] L. Lin and E. Ambikairajah, “Speech denoising based on an auditory filter-bank,” in *6th IEEE International Conference on Signal Processing*, vol. 1, 2002, pp. 552–555.
- [43] I. Cohen, “Enhancement of speech using bark-scaled wavelet packet decomposition,” in *Seventh European Conference on Speech Communication and Technology*, 2001.
- [44] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.
- [45] R. Martin, “Spectral subtraction based on minimum statistics,” *power*, vol. 6, p. 8, 1994.
- [46] R. C. Hendriks, J. Jensen, and R. Heusdens, “Noise tracking using dft domain subspace decompositions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 541–553, 2008.
- [47] R. C. Hendriks, R. Heusdens, and J. Jensen, “Mmse based noise psd tracking with low complexity,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4266–4269.
- [48] N. Upadhyay and R. K. Jaiswal, “Single channel speech enhancement: using wiener filtering with recursive noise estimation,” *Procedia Computer Science*, vol. 84, pp. 22–30, 2016.
- [49] S. Suhadi, C. Last, and T. Fingscheidt, “A data-driven approach to a priori snr estimation,” *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 186–195, 2011.
- [50] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [51] B. Xia and C. Bao, “Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification,” *Speech Communication*, vol. 60, pp. 13–29, 2014.

- [52] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.
- [53] S. Doclo, “Multi-microphone noise reduction and dereverberation techniques for speech applications,” *Ph.D. thesis, University of Leuven, Leuven, Belgium*, 2003.
- [54] S. E. Nordholm, H. H. Dam, C. C. Lai, and E. A. Lehmann, “Broadband beamforming and optimization,” in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 3, pp. 553–598.
- [55] M. Crocco and A. Trucco, “Stochastic and analytic optimization of sparse aperiodic arrays and broadband beamformers with robust superdirective patterns,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 9, pp. 2433–2447, 2012.
- [56] R. C. Hansen, *Phased array antennas*. John Wiley & Sons, 2009, vol. 213.
- [57] R. M. Leahy and B. D. Jeffs, “On the design of maximally sparse beamforming arrays,” *IEEE Transactions on antennas and propagation*, vol. 39, no. 8, pp. 1178–1187, 1991.
- [58] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.
- [59] 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.
- [60] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.
- [61] D. Rimantho, H. H. H. Dam, and S. Nordholm, “Design of low complexity robust broadband beamformers with least squared performance criterion,” *Advances in Information Technology and Applied Computing*, vol. 1, pp. 46–50, 2012.
- [62] H. H. Dam and S. Nordholm, “Design of robust broadband beamformers with discrete coefficients and least squared criterion,” *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 60, no. 12, pp. 897–901, 2013.

- [63] S. Nordholm, V. Rehbock, K. Tee, and S. Nordebo, "Chebyshev optimization for the design of broadband beamformers in the near field," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 1, pp. 141–143, 1998.
- [64] K. F. C. Yiu, X. Yang, S. Nordholm, and K. L. Teo, "Near-field broadband beamformer design via multidimensional semi-infinite-linear programming techniques," *IEEE Transactions on Speech and Audio processing*, vol. 11, no. 6, pp. 725–732, 2003.
- [65] Z. G. Feng, K. F. C. Yiu, and S. E. Nordholm, "A two-stage method for the design of near-field broadband beamformer," *IEEE transactions on signal processing*, vol. 59, no. 8, pp. 3647–3656, 2011.
- [66] G. W. Elko, "Superdirectional microphone arrays," in *Acoustic signal processing for telecommunication*. Springer, 2000, pp. 181–237.
- [67] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.
- [68] H. Chen, W. Ser, and Z. L. Yu, "Optimal design of nearfield wideband beamformers robust against errors in microphone array characteristics," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 9, pp. 1950–1959, 2007.
- [69] Z. Li, K. F. C. Yiu, and S. E. Nordholm, "On the indoor beamformer design with reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1225–1235, 2014.
- [70] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
- [71] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.
- [72] I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beaming," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 9, pp. 1093–1096, 1992.
- [73] S. A. Vorobyov, "Adaptive and robust beamforming," in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 3, pp. 503–552.

- [74] M. Buck, “Aspects of first-order differential microphone arrays in the presence of sensor imperfections,” *Transactions on Emerging Telecommunications Technologies*, vol. 13, no. 2, pp. 115–122, 2002.
- [75] H. Cox, R. Zeskind, and T. Kooij, “Practical supergain,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 3, pp. 393–398, 1986.
- [76] S. Doclo and M. Moonen, “Superdirective beamforming robust against microphone mismatch,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 617–631, 2007.
- [77] C.-C. Lai, S. Nordholm, and Y.-H. Leung, “Design of robust steerable broadband beamformers incorporating microphone gain and phase error characteristics,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 101–104.
- [78] L. C. Ching, “A study into the design of steerable microphones arrays,” Ph.D. dissertation, Curtin University, 2012.
- [79] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.
- [80] M. Wu and D. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.
- [81] S. Mosayyebpour, M. Esmaili, and T. A. Gulliver, “Single-microphone early and late reverberation suppression in noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 322–335, 2013.
- [82] E. A. Habets and J. Benesty, “A two-stage beamforming approach for noise reduction and dereverberation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 945–958, 2013.
- [83] T. Dietzen, N. Huleihel, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, “Speech dereverberation by data-dependent beamforming with signal pre-whitening,” in *Signal Processing Conference (EU-SIPCO), 2015 23rd European*. IEEE, 2015, pp. 2461–2465.

- [84] K. Furuya and A. Kataoka, “Robust speech dereverberation using multi-channel blind deconvolution with spectral subtraction,” *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1579–1591, 2007.
- [85] C. Breithaupt and R. Martin, “Analysis of the decision-directed snr estimator for speech enhancement with respect to low-snr and transient conditions,” *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 2, pp. 277–289, 2011.
- [86] N. Höglund and S. Nordholm, “Improved a priori snr estimation with application in log-mmse speech estimation,” in *proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 189–192.
- [87] O. Cappé, “Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.
- [88] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori snr estimation approach based on selective cepstro-temporal smoothing,” in *proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’08)*, 2008, pp. 4897–4900.
- [89] C. Plapous, C. Marro, and P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.
- [90] 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.
- [91] M. L. Jepsen, S. D. Ewert, and T. Dau, “A computational model of human auditory signal processing and perception,” *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 422–438, 2008.
- [92] E. Zwicker, “Subdivision of the audible frequency range into critical bands (frequenzgruppen),” *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

- [93] A. Sekey and B. A. Hanson, “Improved 1-bark bandwidth auditory filter,” *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1902–1904, 1984.
- [94] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [95] S. Voran, “Objective estimation of perceived speech quality. i. development of the measuring normalizing block technique,” *IEEE Transactions on speech and audio processing*, vol. 7, no. 4, pp. 371–382, 1999.
- [96] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP’01)*, vol. 2, 2001, pp. 749–752.
- [97] J. H. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms.” in *proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 7, 1998, pp. 2819–2822.
- [98] T. Lotter and P. Vary, “Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model,” *journal on applied signal processing (EURASIP)*, vol. 2005, pp. 1110–1126, 2005.
- [99] 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,” *International Workshop on Acoustic Echo and Noise Control (IWAENC), Seattle, USA.*, 2008.
- [100] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, “Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model,” *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.
- [101] 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, 2006.

- [102] A. Davis, S. Nordholm, and R. Togneri, “Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.
- [103] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.
- [104] 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 Processing Conference (EUSIPCO’11), Barcelona, Spain*, 2011, pp. 211–215.
- [105] T. Gerkmann and R. C. Hendriks, “Noise power estimation based on the probability of speech presence,” in *proc. IEEE Workshop on Applications of Signal Processing, Audio and Acoustics (WASPAA), New Paltz, NY*, 2011, pp. 145–148.
- [106] M. Shakil and B. G. Kibria, “Exact distributions of the linear combination of gamma and rayleigh random variables,” *Austrian Journal of Statistics*, vol. 38, no. 1, pp. 33–44, 2009.
- [107] P. Hitczenko, “A note on a distribution of weighted sums of iid rayleigh random variables,” *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 171–175, 1998.
- [108] M. K. Hasan, S. Salahuddin, and M. R. Khan, “A modified a priori snr for speech enhancement using spectral subtraction rules,” *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450–453, 2004.
- [109] R. C. Hendriks, R. Heusdens, and J. Jensen, “Improved decision directed approach for speech enhancement using an adaptive time segmentation,” in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [110] P. Yun-Sik and J.-H. Chang, “A novel approach to a robust a priori snr estimator in speech enhancement,” *IEICE transactions on communications*, vol. 90, no. 8, pp. 2182–2185, 2007.
- [111] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, “Convex combination framework for a priori snr estimation in speech enhancement,” in

IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), March 2017, pp. 4975–4979.

- [112] 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.
- [113] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [114] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, “Improved a priori snr estimation in speech enhancement,” in *23rd Asia-Pacific Conference on Communications (APCC)*. IEEE, 2017, pp. 1–5.
- [115] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, March 2015.
- [116] H. Kuttruff, *Room acoustics*. CRC Press, 2009.
- [117] L. Savioja, *Modeling techniques for virtual acoustics*. Helsinki University of Technology Espoo, Finland, 1999.
- [118] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
- [119] E. A. Lehmann and A. M. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.
- [120] —, “Diffuse reverberation model for efficient image-source simulation of room impulse responses,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2010.
- [121] E. A. P. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” *Dissertation Abstracts International*, vol. 68, no. 04, 2007.

- [122] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, “Combination of mvdr beamforming and single-channel spectral processing for enhancing noisy and reverberant speech,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 61, 2015.
- [123] S. Doclo and M. Moonen, “Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics,” *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2511–2526, 2003.
- [124] L. Khalid, S. Nordholm, and H. Dam, “Design study on microphone arrays,” in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 1171–1175.
- [125] Y.-C. Chang, “N-dimension golden section search: Its variants and limitations,” in *2009 2nd International Conference on Biomedical Engineering and Informatics*. IEEE, 2009, pp. 1–6.
- [126] J. A. Koupaei, S. M. M. Hosseini, and F. M. Ghaini, “A new optimization algorithm based on chaotic maps and golden section search method,” *Engineering Applications of Artificial Intelligence*, vol. 50, pp. 201–214, 2016.
- [127] P. A. Naylor, N. D. Gaubitch, and E. A. Habets, “Signal-based performance evaluation of dereverberation algorithms,” *Journal of Electrical and Computer Engineering*, vol. 2010, p. 1, 2010.
- [128] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.
- [129] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 313–317.
- [130] C. C. Lai, S. E. Nordholm, and Y. H. Leung, *A Study Into the Design of Steerable Microphone Arrays*. Springer, 2017.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.