NEW APPROACHES FOR SPEECH ENHANCEMENT IN THE SHORT-TIME FOURIER TRANSFORM DOMAIN

Mahdi Parchami

A thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Concordia University
Montréal, Québec, Canada

September 2016

© Mahdi Parchami, 2016

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: ______________________________

Entitled: ______________________________

____

____

and submitted in partial fulfillment of the requirements for the degree of

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair

______________________________ External Examiner

______________________________ External to Program

______________________________ Examiner

______________________________Examiner

______________________________Thesis Supervisor

Approved by

____
Chair of Department or Graduate Program Director

______________ ______________________________
Dean of Faculty

Abstract

New Approaches for Speech Enhancement in the Short-Time Fourier Transform Domain

Mahdi Parchami, Ph.D.

Concordia University, 2016

Speech enhancement aims at the improvement of speech quality by using various algorithms. A speech enhancement technique can be implemented as either a time domain or a transform domain method. In the transform domain speech enhancement, the spectrum of clean speech signal is estimated through the modification of noisy speech spectrum and then it is used to obtain the enhanced speech signal in the time domain. Among the existing transform domain methods in the literature, the short-time Fourier transform (STFT) processing has particularly served as the basis to implement most of the frequency domain methods. In general, speech enhancement methods in the STFT domain can be categorized into the estimators of complex discrete Fourier transform (DFT) coefficients and the estimators of real-valued short-time spectral amplitude (STSA). Due to the computational efficiency of the STSA estimation method and also its superior performance in most cases, as compared to the estimators of complex DFT coefficients, we focus mostly on the estimation of speech STSA throughout this work and aim at developing algorithms for noise reduction and reverberation suppression.

First, we tackle the problem of additive noise reduction using the single-channel Bayesian STSA estimation method. In this respect, we present new schemes for the selection of Bayesian cost function parameters for a parametric STSA estimator, namely the Wβ-SA estimator, based on an initial estimate of the speech and also the properties of human auditory system. We further use the latter information to design an efficient flooring scheme for the gain function of the STSA estimator. Next, we apply the generalized Gaussian distribution (GGD) to the Wβ-SA estimator as the speech STSA prior and propose to choose its parameters according to noise spectral variance and *a priori* signal to noise ratio (SNR). The suggested STSA estimation schemes are able to provide further noise reduction as well as less speech distortion, as compared to the previous

methods. Quality and noise reduction performance evaluations indicated the superiority of the proposed speech STSA estimation with respect to the previous estimators.

Regarding the multi-channel counterpart of the STSA estimation method, first we generalize the proposed single-channel Wβ-SA estimator to the multi-channel case for spatially uncorrelated noise. It is shown that under the Bayesian framework, a straightforward extension from the single-channel to the multi-channel case can be performed by generalizing the STSA estimator parameters, i.e. α and β. Next, we develop Bayesian STSA estimators by taking advantage of speech spectral phase rather than only relying on the spectral amplitude of observations, in contrast to conventional methods. This contribution is presented for the multi-channel scenario with single-channel as a special case. Next, we aim at developing multi-channel STSA estimation under spatially correlated noise and derive a generic structure for the extension of a single-channel estimator to its multi-channel counterpart. It is shown that the derived multi-channel extension requires a proper estimate of the spatial correlation matrix of noise. Subsequently, we focus on the estimation of noise correlation matrix, that is not only important in the multi-channel STSA estimation scheme but also highly useful in different beamforming methods.

Next, we aim at speech reverberation suppression in the STFT domain using the weighted prediction error (WPE) method. The original WPE method requires an estimate of the desired speech spectral variance along with reverberation prediction weights, leading to a sub-optimal strategy that alternatively estimates each of these two quantities. Also, similar to most other STFT based speech enhancement methods, the desired speech coefficients are assumed to be temporally independent, while this assumption is inaccurate. Taking these into account, first, we employ a suitable estimator for the speech spectral variance and integrate it into the estimation of the reverberation prediction weights. In addition to the performance advantage with respect to the previous versions of the WPE method, the presented approach provides a good reduction in implementation complexity. Next, we take into account the temporal correlation present in the STFT of the desired speech, namely the inter-frame correlation (IFC), and consider an approximate model where only the frames within each segment of speech are considered as correlated. Furthermore, an efficient method for the estimation of the underlying IFC matrix is developed based on the extension of the speech variance estimator proposed previously. The performance results reveal lower residual reverberation and higher overall quality provided by the proposed method.

Finally, we focus on the problem of late reverberation suppression using the classic speech

spectral enhancement method originally developed for additive noise reduction. As our main contribution, we propose a novel late reverberant spectral variance (LRSV) estimator which replaces the noise spectral variance in order to modify the gain function for reverberation suppression. The suggested approach employs a modified version of the WPE method in a model based smoothing scheme used for the estimation of the LRSV. According to the experiments, the proposed LRSV estimator outperforms the previous major methods considerably and scores the closest results to the theoretically true LRSV estimator. Particularly, in case of changing room impulse responses (RIRs) where other methods cannot follow the true LRSV estimator accurately, the suggested estimator is able to track true LRSV values and results in a smaller tracking error. We also target a few other aspects of the spectral enhancement method for reverberation suppression, which were explored before only for the purpose of noise reduction. These contributions include the estimation of signal to reverberant ratio (SRR) and the development of new schemes for the speech presence probability (SPP) and spectral gain flooring in the context of late reverberation suppression.

I dedicate this work to my loving parents.

Acknowledgments

First and foremost, I would like to express my sincerest gratitude and appreciation to my supervisor, Prof. Wei-Ping Zhu, for providing me with the opportunity to work in the area of speech enhancement, for his invaluable guidance and mentorship, and for his encouragement and support throughout all levels of my research. I am also grateful to him for including me in the NSERC CRD research project sponsored by Microsemi of Ottawa.

I would like to give special thanks to Prof. Benoit Champagne, McGill University, Canada for his consistent support, valuable comments and suggestions for my publications and the CRD project research. His advices and critiques have indeed helped me to develop and improve my ideas through this thesis as well as the publications we completed together, and ultimately, led to timely accomplishment of this work.

I would also like to give special thanks to the Microsemi technical staff for all their inputs and feedbacks on my research during the regular project progress meetings.

I am also grateful to my research teammates, Mr. Sujan Kumar Roy, Mr. Xinrui Pu, and all the signal processing laboratory members for their assistance, friendship, and cooperation. Their smile and support motivated me during this research and gave me the taste of a family in Canada.

I am very grateful to Concordia University and NSERC, Canada for providing me with financial support through my supervisors' research grants. Without such a support, this thesis would not have been possible.

Finally, I would like to express my love and appreciation to my parents and thank them for their consistent encouragement and care during my doctoral study in Canada.

Contents

List of Figures **xiii**

List of Tables **xix**

List of Abbreviations **xxi**

1 Introduction **1**

1.1 Speech Enhancement and Its Applications 1

1.2 Speech Enhancement in the Frequency Domain 2

1.2.1 Importance of the Frequency Domain Technique 3

1.2.2 Application of Short-Time Fourier Transform (STFT) 3

1.3 Overview of Noise Reduction in the STFT Domain 4

1.3.1 Classification of STFT-Based Techniques 4

1.3.2 Advantage of Spectral Amplitude Estimators over Estimators of Complex DFT 6

1.3.3 Estimation of Spectral Amplitude versus Spectral Phase 7

1.3.4 Bayesian (MMSE-Based) Speech Spectral Amplitude Estimation 8

1.3.4.1 Development of Cost Functions 8

1.3.4.2 Speech Priors 9

1.3.4.3 Multi-Channel Extension 10

1.4 Overview of Reverberation Reduction Techniques 11

1.4.1 Speech Reverberation in Acoustic Environments 11

1.4.2 Classification of Reverberation Reduction Techniques 12

1.4.3 Blind Dereverberation in the STFT Domain 13

1.5 Motivation and Objectives of the Research 14

1.5.1 Motivation 14
1.5.2 Objectives 16
1.6 Organization and Main Contributions 17

2 Background: Speech Enhancement in the STFT Domain 19
2.1 Estimation of Speech STSA 19
2.1.1 Problem Statement 20
2.1.2 Spectral Subtractive Estimators 20
2.1.3 Wiener Estimators 23
2.1.4 Maximum Likelihood (ML) Estimators 24
2.1.5 Maximum *a Posteriori* (MAP) Estimators 25
2.1.6 MMSE-Based (Bayesian) Estimators 26
2.1.6.1 Ephraim and Malah's MMSE and Log-MMSE Estimators 27
2.1.6.2 Perceptually Motivated Bayesian Estimators 29
2.1.7 Use of Speech Presence Probability (SPP) 32
2.2 Speech STSA Priors 34
2.3 Multi-Channel STSA Estimation 39
2.3.1 Multi-Channel Problem Statement 40
2.3.2 Multi-channel Extension of the Bayesian STSA Estimation 41
2.4 Reverberation Suppression in the STFT Domain 43
2.4.1 Reverberation in Enclosed Spaces 43
2.4.2 Problem Formulation 45
2.5 Shortcomings of the State-of-the-Art STSA Estimation Methods 47

3 Single-Channel Noise Reduction Using Bayesian STSA Estimation 50
3.1 Introduction 50
3.2 Previous Work 50
3.3 Proposed Speech STSA Estimator 53
3.3.1 Brief Description of the Proposed Method 53
3.3.2 Parameter Selection of the New Wβ-SA Estimator 53
3.3.3 Gain Flooring Scheme 57
3.4 Extension of Wβ-SA Estimator Using GGD Prior 58

3.4.1 Wβ-SA Estimator with GGD Prior 59
3.4.2 Estimation of GGD Prior Parameters 60
3.5 Performance Evaluation 61
3.5.1 Performance Measures for Noise Reduction 61
3.5.2 Evaluation of the Proposed Method 63
3.6 Conclusion 79

4 Multi-Channel Bayesian STSA Estimation for Noise Suppression 80
4.1 Introduction 80
4.2 Brief Description of the Proposed Methods 82
4.3 Multi-Channel STSA Estimation in Spatially Uncorrelated Noise 82
4.3.1 Extension of the Proposed Wβ-SA Estimator to Multi-Channel 83
4.3.2 STSA Estimators Using Spectral Phase 86
4.3.2.1 MMSE-Based STSA Estimator Using Spectral Phase 86
4.3.2.2 Extension to the Wβ-SA Estimator 88
4.3.2.3 Estimation of the Spectral Phase 89
4.3.3 Performance Evaluation in Spatially Uncorrelated Noise 90
4.4 Multi-Channel STSA Estimation in Spatially Correlated Noise 96
4.4.1 Extension of STSA Estimation to the Multi-Channel Case Under Known DOA and Noise PSD Matrix 96
4.4.2 Estimation of Noise PSD Matrix 99
4.4.2.1 Incorporation of Subsequent Speech Frames 100
4.4.2.2 Iterative Method for the Selection of the Forgetting Factor 101
4.4.2.3 Minimum Tracking and Bias Compensation 103
4.4.3 Performance Evaluation in Spatially Correlated Noise 104
4.5 Conclusion 112

5 Speech Dereverberation Using the Weighted Prediction Error Method 114
5.1 Introduction 114
5.2 Brief Description of the Proposed Methods 115
5.3 Review on the WPE Method 116
5.4 WPE Method with the Estimation of Early Speech Variance 119

5.5 WPE Method Using the Inter-Frame Correlations . . . 122
5.5.1 Proposed ML Solution . . . 123
5.5.2 Estimation of the IFC Matrix . . . 125
5.6 Performance Evaluation . . . 127
5.6.1 Experimental Setup . . . 127
5.6.2 Evaluation of the Proposed Method in Section 5.4 . . . 130
5.6.3 Evaluation of the Proposed Method in Section 5.5 . . . 132
5.7 Conclusion . . . 136

6 Speech Dereverberation Using the Spectral Enhancement Method 138

6.1 Introduction . . . 138
6.2 Brief Description of the Proposed Methods . . . 140
6.3 Background: Late Reverberation Suppression Using Spectral Enhancement . . . 141
6.4 Proposed LRSV Estimator . . . 143
6.4.1 Suggested Scheme for the Shape Parameter . . . 144
6.4.2 Estimation of the Reverberant Component . . . 147
6.4.3 Incremental Implementation of the WPE Method . . . 148
6.5 Other Developments on Classic Spectral Enhancement Methods . . . 150
6.5.1 Estimation of SRR . . . 150
6.5.2 Application of SPP to Gain function . . . 153
6.5.3 Spectral Gain Flooring for Dereverberation . . . 154
6.5.4 Beamforming for Late Reverberation Suppression . . . 155
6.6 Performance Evaluation . . . 157
6.6.1 Evaluation of the Proposed LRSV Estimator . . . 157
6.6.1.1 Performance in Time-Invariant RIRs . . . 158
6.6.1.2 Performance in Time-Varying RIRs . . . 165
6.6.2 Evaluation of the Proposed Schemes in Section 6.5 . . . 170
6.6.3 Joint Noise Reduction and Dereverberation . . . 173
6.7 Conclusion . . . 177

7 Conclusions and Future Work 179

7.1 Concluding Remarks . . . 179

7.2 Scope for the Further Work 182

References **185**

Appendix A Derivation of Eq. (3.15) **200**

Appendix B Proof of Equation (4.23) **202**

List of Figures

1.1 Block diagram of the analysis-synthesis technique using STFT. 4

1.2 Speech spectral estimation methods in the STFT domain. 6

1.3 Illustration of a speech source (user), noise sources and their reflections captured by a microphone set. 11

2.1 Block diagram of the basic spectral subtraction algorithm. 22

2.2 One-sided GGD function for different values of the scale parameters and $b = 2$. . . . 36

2.3 An N equispaced linear microphone array capturing the speech source $s(t)$ located in the far field impinging at the incident angle θ. 41

2.4 Illustration of the direct path and a single reflection from the speech source to the microphone. 44

2.5 Modeling of the observed microphone signal(s) in a reverberant environment. 45

2.6 Plot of a typical acoustic impulse response with illustrated early and late parts of the RIR. 46

3.1 STSA gain function curves in (3.4) versus β for different values of α (ζ=0 dB and γ=0 dB). 52

3.2 Block diagram of the proposed speech STSA estimation algorithm. 54

3.3 Variation of the proposed choice of α versus frequency bins, compared to that of the initial speech STSA estimate for a frame of noisy speech. 55

3.4 Gain function of the modified Wβ-SA estimator in (3.17) versus the GGD shape parameter c for different values of γ (ζ=-5dB). 60

3.5 Spectrograms of (a): input noisy speech, (b): clean speech, (c): enhanced speech by the original Wβ-SA estimator and (d): enhanced speech by the proposed Wβ-SA estimator, in case of babble noise (Input SNR=5 dB). 67

3.6 LLR versus global SNR for different Wβ-SA estimators, (a): white noise, (b): babble noise and (c): car noise. 68

3.7 PESQ versus global SNR for different Wβ-SA estimators, (a): white noise, (b): babble noise and (c): car noise. 69

3.8 SNRseg versus global SNR for different Wβ-SA estimators, (a): white noise, (b): babble noise and (c): car noise. 70

3.9 PESQ versus global SNR for Wβ-SA estimator with the proposed parameters in Section 3.3 using different gain flooring schemes, (a): white noise, (b): babble noise and (c): car noise. 71

3.10 PESQ versus global SNR for the Rayleigh-based estimator in Section 3.3, the GGD-based estimator in Section 3.4 with $c = 1.5, 2.5$ and the proposed choice of c in Section 3.4, (a): white noise, (b): babble noise and (c): car noise. 72

3.11 Spectrograms of (a): input noisy speech, (b): clean speech, (c): enhanced speech by WE estimator with Chi prior in [34], (d): enhanced speech by WCOSH estimator with Chi prior in [34], (e): enhanced speech by Log-MMSE estimator with GGD prior in [85] and (f): enhanced speech by the proposed Wβ-SA estimator with GGD prior in Section 3.4, in case of babble noise (Input SNR=5 dB). 73

3.12 LLR versus global SNR for the STSA estimators in [34, 85] and the proposed STSA estimator in Section 3.4, (a): white noise, (b): babble noise and (c): car noise. . . . 74

3.13 PESQ versus global SNR for the STSA estimators in [34, 85] and the proposed STSA estimator in Section 3.4, (a): white noise, (b): babble noise and (c): car noise. 75

3.14 SNRseg versus global SNR for the STSA estimators in [34, 85] and the proposed STSA estimator in Section 3.4, (a): white noise, (b): babble noise and (c): car noise. 76

4.1 Scenario of capturing a far field source of speech in spatially uncorrelated noise by a linear microphone array. 90

4.2 LLR versus input global SNR for the multi-channel STSA estimators with $N = 2$ microphones in spatially uncorrelated noise, (a): white noise, (b): babble noise and (c): car noise. 91

4.3 PESQ versus input global SNR for the multi-channel STSA estimators with $N = 2$ microphones in spatially uncorrelated noise, (a): white noise, (b): babble noise and (c): car noise. 92

4.4 SNRSeg versus input global SNR for the multi-channel STSA estimators with $N = 2$ microphones in spatially uncorrelated noise, (a): white noise, (b): babble noise and (c): car noise. 93

4.5 Performance scores of the proposed GGD-based Wβ-SA estimator in (4.10) for different microphone numbers in babble noise. 93

4.6 LLR for the conventional and spectral phase-based STSA estimators in babble noise with (a): $N = 1$ and (b): $N = 2$ microphones. 94

4.7 PESQ for the conventional and spectral phase-based STSA estimators in babble noise with (a): $N = 1$ and (b): $N = 2$ microphones. 95

4.8 Segmental SNR for the conventional and spectral phase-based STSA estimators in babble noise with (a): $N = 1$ and (b): $N = 2$ microphones. 95

4.9 Performance of the spectral phase-based Wβ-SA estimator in (4.19) in babble noise with $N = 4$, using the noisy phase, using the MMSE estimate of phase in (4.20), and using the phase of the clean speech. 96

4.10 Block diagram of the proposed general scheme for multi-channel Bayesian STSA estimation. 99

4.11 Scenario of capturing a speech source in spatially correlated noise with two microphones, generated by the ISM method. 104

4.12 LLR versus input global SNR for multi-channel STSA estimators and MVDR beamformer with $N = 2$ microphones in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise. 105

4.13 PESQ versus input global SNR for multi-channel STSA estimators and MVDR beamformer with $N = 2$ microphones in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise. 106

4.14 Segmental SNR versus input global SNR for multi-channel STSA estimators and MVDR beamformer with $N = 2$ microphones in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise. 107

4.15 LLR versus input global SNR for the enhanced speech using the MVDR beamformer with different noise PSD matrix estimation methods in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise. 108

4.16 PESQ versus input global SNR for the enhanced speech using the MVDR beamformer with different noise PSD matrix estimation methods in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise. 109

4.17 Segmental SNR versus input global SNR for the enhanced speech using the MVDR beamformer with different noise PSD matrix estimation methods in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise. 110

4.18 MVDR beamformer response error versus input global SNR using different noise PSD matrix estimation methods, (a): white noise, (b): babble noise and (c): car noise. 111

4.19 Performance measures of the MVDR beamformer for different microphone numbers using the proposed method of noise PSD matrix estimation in babble noise. 112

4.20 Performance measures of the MVDR beamformer using the proposed method of noise PSD matrix estimation with a different number of involved subsequent frames, d, in babble noise. 112

5.1 Normalized IFC of the early speech $d_{n,k}$ averaged over frequency bins versus STFT frame number for a selected speech utterance. 122

5.2 A two-dimensional illustration for the geometry of the synthesized scenario of a noisy reverberant environment. 129

5.3 Improvements in PESQ and CD scores versus the number of iterations for different dereverberation methods. 130

5.4 Improvements in FW-SNR and SRMR scores versus the number of iterations for different dereverberation methods. 131

5.5 PESQ and CD scores versus T_{60dB} for the reverberant speech and the enhanced one using the WPE method with different estimators of the desired speech spectral variance, $\sigma^2_{d_{n,k}}$. 132

5.6 Processing time required for the estimation of $\mathbf{g}_k$ with lengths of L_k=15 and L_k=30 using a 10-second speech segment for different methods. An i5-2400 CPU @ 3.10GHz with RAM of 4.00GB was used for the implementation in Matlab. 132

5.7 Normalized IFC averaged over frequency bins and frames versus the frame lag for speech samples with different amounts of reverberation. 133

5.8 Performance of the proposed WPE method versus the assumed IFC length, τ, for different D. 134

5.9 Improvement in PESQ and CD scores versus T_{60dB} for different WPE-based dereverberation methods using synthetic RIRs. 136

5.10 Improvement in FW-SNR and SRMR scores versus T_{60dB} for different WPE-based dereverberation methods using synthetic RIRs. 136

6.1 An illustration of the STFT frames and the processing blocks over speech time samples. 145

6.2 Block diagram of the proposed algorithm for LRSV estimation. 150

6.3 A two-dimensional schematic of the geometric setup used to synthesize the time-invariant RIR by the ISM method. 159

6.4 Normalized error in the estimation of the LRSV w.r.t. to the case of using the entire speech utterance, versus the processing block length for different reverberation times.160

6.5 Mean spectral variances using the recorded RIR from the REVERB Challenge [134] for: (a) the true LRSV, the LRSV estimated using RIR variances and the proposed LRSV (b) the true LRSV, the LRSV estimated by the improved model-based method [144] and the one estimated by the Lebart's method [142]. 161

6.6 Mean segmental error for different LRSV estimators using the synthesized RIRs by the ISM method [119] with source-to-microphone distances of (a): 1 m (b): 2 m. . . 162

6.7 A two-dimensional schematic of the geometric setup used to synthesize the time-variant RIR (moving talker) by the ISM method. 166

6.8 Mean spectral variances for: (a) the true LRSV, the LRSV estimated using RIR variances and the proposed LRSV (b) the true LRSV, the LRSV estimated by the improved model-based method [144] and the one estimated by the Lebart's method [142]. 167

6.9 Mean segmental error for different LRSV estimators using the configuration in Figure 6.7 with H as (a): 1 m (b): 2 m. 167

6.10 PESQ and CD measures versus the reverberation time for the MVDR beamformer with different numbers of microphones using the proposed LRSV matrix estimation. 173

6.11 Modified spectral enhancement method used for jointly suppressing the noise and late reverberation. 174

6.12 Suggested algorithm to use the MVDR beamformer for the purpose of joint noise and late reverberation suppression. 174

6.13 4 different combinations of the WPE method, the MVDR beamformer and the spectral enhancement for joint noise and reverberation suppression. 175

6.14 PESQ and CD scores versus the reverberation time for different single-channel combinations of the WPE and the modified SE methods. 176

6.15 FW-SNR and SRMR scores versus the reverberation time for different single-channel combinations of the WPE and the modified SE methods. 176

6.16 PESQ and CD scores versus the reverberation time for different multi-channel systems in Figure 6.13. 177

6.17 FW-SNR and SRMR scores versus the reverberation time for different multi-channel systems in Figure 6.13. 177

List of Tables

2.1 Major Bayesian estimators of the speech STSA 32

2.2 Parameter sets of the GGD leading to Rayleigh, Gamma, Chi, or exponential speech STSA models. 35

2.3 Speech STSA estimators for particular parameter choices of the GGD speech and noise priors 38

3.1 PESQ values for the Wβ-SA estimator with different schemes of parameter α, case of white noise. 77

3.2 PESQ values for the Wβ-SA estimator with different schemes of parameter α, case of babble noise. 77

3.3 PESQ values for the Wβ-SA estimator with different schemes of parameter α, case of car noise. 77

3.4 PESQ values for the Wβ-SA estimator with different schemes of parameter β, case of white noise. 78

3.5 PESQ values for the Wβ-SA estimator with different schemes of parameter β, case of babble noise. 78

3.6 PESQ values for the Wβ-SA estimator with different schemes of parameter β, case of car noise. 78

5.1 Outline of the steps in the conventional WPE method. 118

5.2 Performance comparison of different WPE-based dereverberation methods using recorded RIRs. 135

6.1 Performance measures using the recorded RIR from the REVERB Challenge. 163

6.2 Performance measures using the ISM method for source-to-microphone distance of 1 m. 164

6.3 Performance measures using the ISM method for a source-to-microphone distance of 2 m. 165
6.4 Performance measures for time-variant RIR with H=1 m in Figure 6.7. 169
6.5 Performance measures for time-variant RIR with H=2 m in Figure 6.7. 170
6.6 Performance measures using the recorded RIRs from the REVERB Challenge. . . . 171
6.7 Performance measures using the ISM method for source-to-microphone distance of 1 m. 172

List of Acronyms

AR :	Auto Regressive
ARMA :	Auto Regressive Moving Average
ATF :	Acoustic Transfer Function
CD :	Cepstral Distance
CGG :	Complex Generalized Gaussian
DCT :	Discrete Cosine Transform
DD :	Decision-Directed
DFT :	Discrete Fourier Transform
DOA :	Direction of Arrival
DDR :	Direct to Reverberant Ratio
EM :	Expectation Maximization
FFT :	Fast Fourier Transform
FW-SNR :	Frequency-Weighted SNR
GGD :	Generalized Gamma Distribution
GWSA :	Generalized Weighted Spectral Amplitude
IFC :	Inter-Frame Correlation
IFFT :	Inverse Fast Fourier Transform
IID :	Independent and Identically Distributed
IMCRA :	Improved Minima Controlled Recursive Averaging
IS :	Itakura-Saito Distance
ISM :	Image Source Method
ITU :	International Telecommunications Union
KLT :	Karhunen-Loeve Transform

LLR :	Log-Likelihood Ratio
LP :	Linear Prediction
LPC :	Linear Prediction Coefficients
LSA :	Log-Spectral Amplitude
LRSV :	Late Reverberant Spectral Variance
MA :	Moving Average
MAP :	Maximum *a posteriori*
MLCP :	Multi-Channel Linear Prediction
ML :	Maximum Likelihood
MMSE :	Minimum Mean Square Error
MOS :	Mean Opinion Scores
MS :	Minimum Statistics
MVDR :	Minimum Variance Distortionless Response
OM-LSA :	Optimally Modified Log-Spectral Amplitude
PDF :	Probability Distribution Function
PSD :	Power Spectral Density
PESQ :	Perceptual Evaluation of Speech Quality
RIR :	Room Impulse Response
SE :	Spectral Enhancement
SNR :	Signal to Noise Ratio
SNRseg :	Segmental SNR
SPP :	Speech Presence Probability
SRMR :	Signal to Reverberant Modulation Ratio
SRR :	Signal to Reverberant Ratio
STFT :	Short Time Fourier Transform
STSA :	Short Time Spectral Amplitude
ULA :	Uniform Linear Array
VAD :	Voice Activity Detector
VoIP :	Voice over Internet Protocol
W-β-SA :	Weighted β Spectral Amplitude
WCOSH :	Weighted Cosine Hyperbolic

WE :	Weighted Euclidean
WPE :	Weighted Prediction Error
WSS :	Weighted-Slope Spectral Distance

Chapter 1

Introduction

In this chapter, we first present a brief introduction to the problem of speech enhancement, its practical applications and the speech enhancement in the short-time Fourier transform (STFT) domain. Next, a general literature review on the existing methods of STFT estimation for speech signals is presented and the advantages of speech spectral amplitude estimators are stated in comparison with other estimators. The motivation and objectives of this research are discussed in the subsequent section and the requirement for further development of speech spectral amplitude estimation methods is explained. At the end, a chapter-by-chapter organization of this thesis and the major contributions are described.

1.1 Speech Enhancement and Its Applications

Speech enhancement aims at the improvement of speech quality by using various algorithms. The term speech quality can be interpreted as clarity, intelligibility, pleasantness or compatibility with some other method in speech processing such as speech recognition and speech coding. Major goals of speech enhancement can be classified into the removal of background noise, echo cancellation, reverberation suppression and the process of artificially bringing certain frequencies into the speech signal [1].

In general, speech enhancement is a difficult task to accomplish for certain reasons. First, the nature and characteristics of corrupting disturbances in speech can be changed dramatically in different environments or from one application to the other. It is thereby challenging to find promising algorithms that work for various practical scenarios. Second, the performance criteria under which the fidelity of the speech enhancement algorithms is judged are defined differently for each application. Moreover, it is often too difficult to satisfy all of the major performance criteria using a specific speech enhancement algorithm. As a common example, in the single-channel (one-microphone) case and when the speech degradation is due to uncorrelated additive noise, noise reduction can be achieved at the expense of introducing speech distortion. In this case, even though noise reduction measures demonstrate quality improvement in the enhanced speech, distortion measures for the output speech are likely to be even worse than those of the noisy speech. Consequently, whereas less noise is heard in the enhanced speech, the resulting intelligibility will not be better than that of the noisy speech. Generally, there exists some compromise between the amount of noise reduction achieved by conventional speech enhancement algorithms and the degree of distortion implied on the clean speech component [2, 3].

The majority of speech enhancement applications include mobile phones, VoIP (voice over internet protocol), teleconferencing systems, speech recognition, and hearing aids [4]. Many voice communication systems as well as all telecommunication systems in noisy environments require speech restoration blocks in order to function properly. Ambient noise prevents the speech coding blocks from estimating the required spectral parameters accurately. Therefore, the resulting coded speech sounds distorted and it still contains corrupting noise. As a result, to improve the performance of speech coding systems, a speech enhancement system has to be placed as a front-end to reduce the noise energy. Speech enhancement is also vital to hearing aid devices. These devices can help the hearing impaired by amplifying ambient audio signals [5, 6]. Thus, with the fast development of the aforementioned speech and audio systems, there is a growing need for further development of speech enhancement algorithms in the future.

1.2 Speech Enhancement in the Frequency Domain

In this section, we explain the important role of frequency domain techniques in speech enhancement and briefly discuss the general scheme used for their implementation.

1.2.1 Importance of the Frequency Domain Technique

From a general point of view, the major algorithms of speech enhancement can be categorized into several fundamental categories including adaptive filtering methods, spectral subtractive algorithms, Wiener filtering and its variations, statistical model-based methods and subspace algorithms. Whereas performance comparisons in terms of speech quality, intelligibility and recognition can be accomplished amongst different categories of speech enhancement algorithms, factors such as computational burden, need for training data and restrictive assumptions about noise and speech environment have to be taken into account in order to consider a certain group of speech enhancement methods [7].

A speech enhancement technique can be implemented as either a time domain or a transform domain method. Famous transform domains in the field of speech processing include discrete Fourier transform (DFT), discrete wavelet transform, discrete cosine transform (DCT) and Karhunen-Loeve transform (KLT). Yet, among the existing transform domain methods for speech enhancement in the literature, those based on discrete Fourier transform processing are usually favored in practical applications. This is due to several reasons such as lower computational complexity, the use of fast Fourier transform (FFT), ease of implementation, providing a trade-off between noise reduction and speech distortion at different frequencies, natural resemblance to the auditory processes taking place within human ear and existence of efficient windowing techniques for the time-domain synthesis of the frequency domain modified speech. These techniques are also known as spectral processing methods and have received much interest in the literature [8, 9].

1.2.2 Application of Short-Time Fourier Transform (STFT)

In the frequency domain speech enhancement, the spectrum of a clean speech signal is estimated through the modification of its noisy speech spectrum and then it is used to obtain the enhanced speech signal in the time domain. However, in many applications such as mobile communication systems, the maximum algorithmic delay and the computational complexity are strictly limited. Moreover, the discrete time Fourier transform is appropriate only for stationary signals, i.e., those with constant statistics over time. Yet, speech is known to be a quasi-stationary signal, i.e., one with approximately constant statistics over short periods of time. For these reasons, in the frequency processing of speech signals, it is required to consider time segments of about 10-40 ms during which the statistics of speech signal do not alter a lot. This is implemented by short-time segmentation of the entire speech and then processing the Fourier coefficients of each segment individually. The processed coefficients across segments are later concatenated via overlap-add or overlap-save methods to produce the entire enhanced speech. This technique is referred to as short-time Fourier transform (STFT) processing and has served as the basis to implement all frequency domain methods of speech enhancement [8]. In Figure 1.1, a schematic of the STFT processing technique has been shown. As indicated, the input speech is segmented and multiplied by proper windows and then DFT coefficients are taken from each segment. Next, the processing (enhancement) method is applied to modify frequency bins of each segment and then the processed segments are transformed back into the time domain by the inverse FFT (IFFT). The overlap-add technique is then used to synthesize the speech signal in the output.

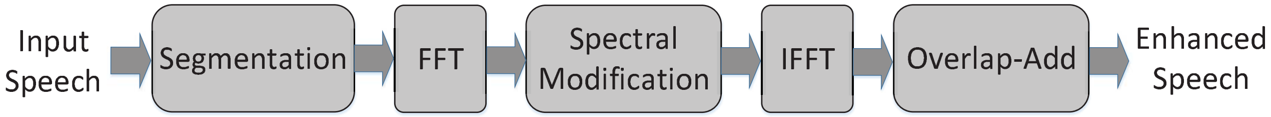


Figure 1.1: Block diagram of the analysis-synthesis technique using STFT.

1.3 Overview of Noise Reduction in the STFT Domain

In this section, an overview of various approaches for speech spectral estimation in the presence of noise in the STFT domain is presented first. Next, as the most important category, estimators of short-time spectral amplitude (STSA) are further elaborated and their advantages over the other STFT estimators are described. Next, a brief literature review is given on the most widely applied STSA estimation methods, namely the Bayesian STSA estimators.

1.3.1 Classification of STFT-Based Techniques

Assuming that the noise process is additive and that noise and speech processes are independent, many conventional methods and their variations exist in the literature that tend to estimate the speech DFT coefficients in an optimal sense [10, 11, 12]. Due to the complex nature of speech DFT coefficients, however, they can be expressed in terms of either the real-imaginary or the amplitude-phase (polar) components. Therefore, speech enhancement techniques in the spectral domain aim at estimating these components and combining them to produce the complex DFT coefficients of the speech estimate [8]. In this regard, two types of methods can be recognized in the STFT domain: those attempting to separately estimate real-imaginary components and those aiming at the estimation of amplitude-phase of speech DFT coefficients. Whereas the former is based on the assumption that the real and imaginary components of speech coefficients are independent, the latter assumes the amplitude and phase are independent components. Under a complex Gaussian model for speech DFT coefficients, it can be proved that these two assumptions are equivalent [13], yet, there is no proof that such a model is accurately true for speech coefficients.

The most well-known techniques for the estimation of speech spectral amplitude, known as STSA estimators, can be categorized as spectral subtraction algorithms [14], frequency domain Wiener filtering [10] and statistical model-based methods [15]. In spectral subtraction algorithms, the STSA of noise is estimated as the square root of the maximum likelihood estimate of spectral variance, and then it is subtracted from the amplitude spectrum of the noisy signal. In the Wiener filtering algorithm, the spectrum estimator is obtained by finding the optimal minimum mean square error (MMSE) estimate of complex Fourier transform coefficients. However, due to inaccuracies in the estimation of speech and noise statistics, both Wiener filtering and spectral subtraction techniques suffer from residual noise which has an annoying noticeable effect on the enhanced speech signal. This processing artifact is referred to in the literature as musical noise and it often results from large spectral peaks randomly distributed over time and frequency in observed speech [7]. Moreover, none of these two approaches is optimal in the sense of speech spectral amplitude estimation, whereas spectral amplitudes are perceptually more relevant to the hearing processing within human ear [16]. This provided the main motivation for Ephraim and Malah [17] to formulate an optimal spectral amplitude estimator which, specifically, estimates the modulus (amplitude) of complex DFT coefficients through the minimization of the mean squared error between clean and estimated speech STSAs. This approach and its later developments were proved to work fairly better than the aforementioned methods in most practical scenarios [17, 18]. In Figure 1.2, a classification of various speech spectral estimators in the STFT domain has been illustrated. As indicated, the MMSE-based method of STSA estimation as well as some simpler alternatives such as maximum likelihood (ML) and maximum *a posteriori* (MAP) estimators [19, 20, 21] are categorized as statistical model-based enhancement methods.

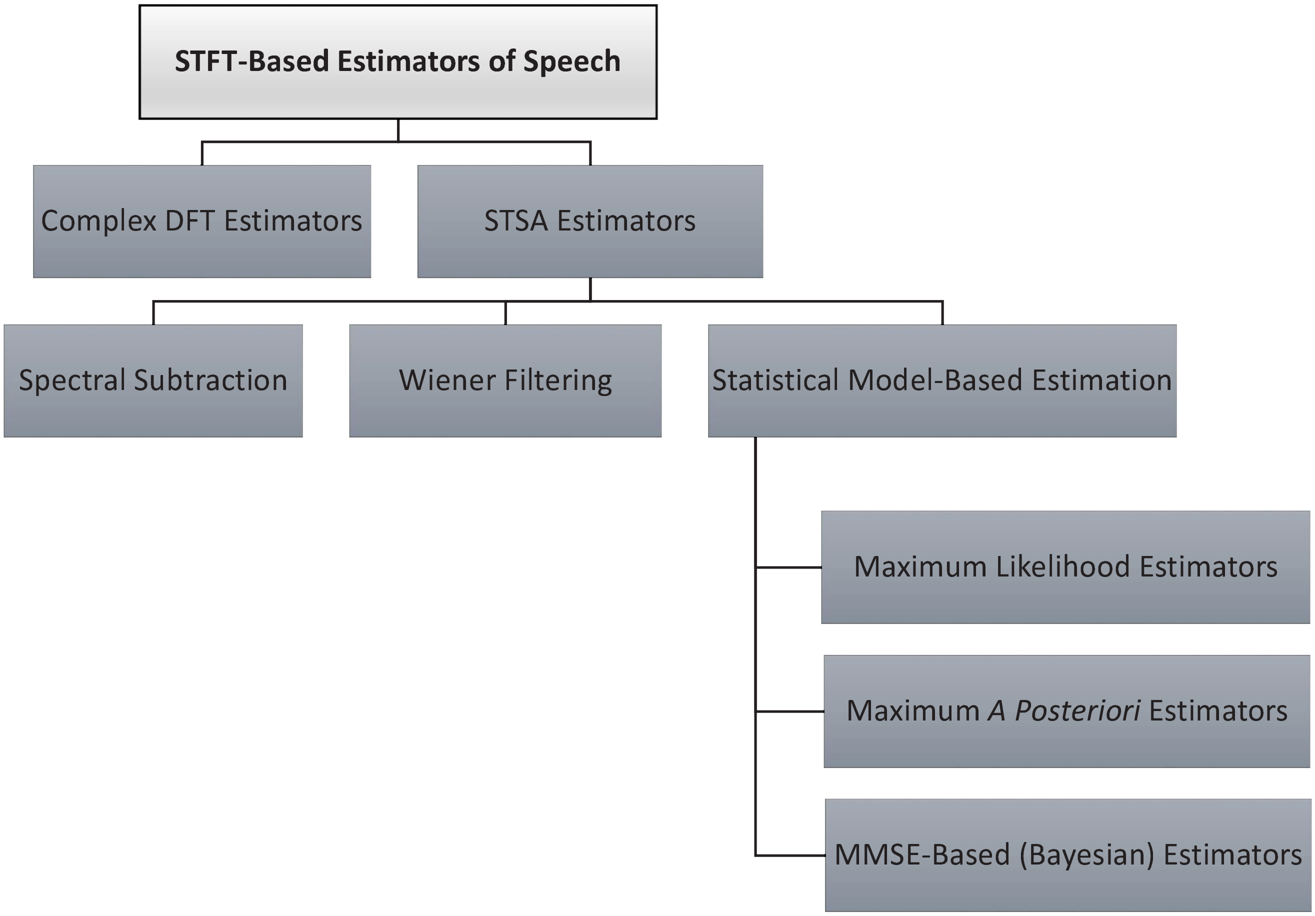

STFT-Based Estimators of Speech
Complex DFT Estimators
STSA Estimators
Spectral Subtraction
Wiener Filtering
Statistical Model-Based Estimation
Maximum Likelihood Estimators
Maximum *A Posteriori* Estimators
MMSE-Based (Bayesian) Estimators

Figure 1.2: Speech spectral estimation methods in the STFT domain.

1.3.2 Advantage of Spectral Amplitude Estimators over Estimators of Complex DFT

A comprehensive study on different estimators of complex DFT coefficients as well as those of real-valued STSA for speech signals has been presented in [22, 23]. The presented estimators therein are based on various statistical models for noise and speech spectral components and generalize all previously proposed estimators within this area. Whereas the former group, i.e., the complex DFT estimators, tend to estimate the real and imaginary parts of speech DFT coefficients independently, the latter group, i.e., the STSA estimators, tend to estimate only the amplitude of speech DFT coefficients regardless of the phase component. Based on the extensive investigations in [22, 23], it is concluded that for almost all experimental scenarios, the STSA estimators perform better than the estimators of complex speech STFT. In addition, since in the former only one real-valued estimate needs to be computed, magnitude estimation is computationally more efficient. It is interesting to note that amplitude estimators perform better than the complex DFT estimators, even through comparisons under complex DFT distortion measures. This is because the modeling assumptions in the complex domain are less accurate than those in the polar domain. In other words, the assumption that real and imaginary parts of speech DFT coefficients are independent introduces more modeling error than assuming independent phase and magnitudes for speech signals. Thereby, among the speech estimation techniques in the STFT domain, STSA estimation methods are often preferred over complex DFT coefficient estimators. For this reason, there have been numerous developments and modifications of these estimators in the relevant literature, which will be discussed in more details in the next chapter.

1.3.3 Estimation of Spectral Amplitude versus Spectral Phase

Considering the polar representation of complex spectral coefficients of speech signals, both the phase and the amplitude components are generally unknown and have to be estimated. In this sense, since the joint estimation of speech amplitude and phase is not mathematically tractable, the possible solution is to estimate each component separately and then combine them to produce the complex coefficients of enhanced speech. However, the spectral amplitude has been found to be perceptually much more relevant than spectral phase in the speech enhancement literature. According to the various experiments in [24, 25], more accurate estimates of speech phase than the degraded phase (that of the noisy speech) cannot significantly improve the performance of speech spectral enhancement techniques. It is known, however, that for almost all finite-duration signals, a signal can be reconstructed up to a scale factor using only the phase of its DFT coefficients. Therefore, in the context of speech enhancement, it may seem possible to first estimate the spectral phase more accurately and then attempt to reconstruct the signal from the phase information. But unfortunately, the accuracy in the reconstructed speech signal appears to be too sensitive to the accuracy of the phase estimate, and such a technique for speech enhancement would require the ability to estimate the spectral phase very accurately [26]. Yet, accurate estimation of speech spectral phase is not a possible task under heavily noisy conditions and very few works with limited performance exist up to date [27]. On the other hand, in the original proposition of the STSA estimation technique [17], it was proved that an MMSE-optimal estimator of spectral phase is actually the phase of noisy speech and that an attempt to provide a better estimate for the spectral phase adversely affects the estimate for the spectral amplitude.

In summary, we conclude that the most efficient technique for speech enhancement among the conventional frequency domain methods is to use all the available information by the complex STFT coefficients of noisy observations in order to provide an estimate for the STSA of a speech signal. The estimated spectral amplitude is then combined with the noisy spectral phase to generate STFT coefficients of the enhanced speech signal.

1.3.4 Bayesian (MMSE-Based) Speech Spectral Amplitude Estimation

In this section, we present a brief literature review on Bayesian estimators of speech spectral amplitude and their development based on different cost functions. Next, the most common probability distribution functions (PDFs) used to model speech STSA priors are introduced, and finally, the existing literature work on the extension of Bayesian STSA estimators to the multi-channel case is discussed.

1.3.4.1 Development of Cost Functions

Within the framework of Bayesian STSA estimators, the general goal is to provide an estimate of the STSA of clean speech using statistical models for the noise and speech spectral components. In [17], Ephraim and Malah proposed to estimate the speech signal amplitude through the minimization of a Bayesian cost function which measures the mean square error between the clean and estimated STSA. Accordingly, the resulting estimator was called the minimum mean square error (MMSE) spectral amplitude estimator. Later in [18], a logarithmic version of the proposed estimator, i.e., the Log-MMSE, was introduced by considering that the logarithm of the STSA is perceptually more relevant to the human auditory system. Even though some alternatives to the Bayesian STSA estimators were proposed, e.g., [28], due to the satisfying performance of these estimators, they are still found to be appealing in the literature. In this regard, more recently, further modifications on STSA Bayesian cost functions were suggested by Loizou in [29] by taking advantage of the psycho-acoustical models initially employed for speech enhancement purposes in [30]. Along the same line of thought, You *et al.* [31] proposed to use the β power of the STSA term in the Bayesian cost function, in order to obtain further flexibility in the corresponding STSA gain function. The authors investigated the performance of the so-called β-order MMSE estimator for different values of β and found that it is moderately better than the MMSE and Log-MMSE estimators proposed earlier.

Plourde and Champagne in [32] suggested to take advantage of STSA power weightings in the β-order MMSE cost function and introduced the parameter α as the power of their new weighting term. They further proposed to select the two estimator parameters as functions of frequency, according to the psycho-acoustical properties of the human auditory system and showed a better quality in the enhanced speech in most of the input signal-to-noise ratio (SNR) range. Yet, at high input SNRs, the performance of the developed estimator may not be appealing due to the undesired distortion in the enhanced speech. Further in [33], the same authors introduced a generalized version of the Wβ-SA estimator by including a new weighting term in the Bayesian cost function which provides additional flexibility in the estimator's gain. However, apart from the mathematically tedious solution for the gain function, the corresponding estimator does not provide further improvement in the enhanced speech quality.

Overall, the parametric Bayesian cost functions as those in [29, 31, 32] can provide further noise reduction as compared to the previous estimators, thanks to the additional gain control obtained by the appropriate choice of the cost function parameters. In [29], fixed values were used for the STSA weighting parameter, whereas in [31], an experimental scheme was proposed in order to adapt β to the estimated frame SNR. In the latter, the adaptive selection of the cost function parameters has been proved to be advantageous over fixed parameter settings. In [32], rather than an adaptive scheme, the values of the estimator parameters are chosen only based on the perceptual properties of the human auditory system. Whereas this scheme is in accordance with the spectral psycho-acoustical models of the hearing system, it does not take into account the noisy speech features in updating the parameters.

1.3.4.2 Speech Priors

In the aforementioned works, since the complex Gaussian PDF has been considered for speech STFT coefficients, the speech STSA is actually modeled by the Rayleigh PDF. However, as it was indicated in [22], parametric non-Gaussian (super-Gaussian) PDFs are able to model the speech STSA prior more accurately. In [34], Chi PDF with a fixed parameter setting was used as the speech STSA prior for a group of perceptually motivated STSA estimators. Use of Chi and Gamma speech priors was further studied in [35] and training-based procedures using the histograms of clean speech data were proposed for the estimation of the prior PDF parameters. Yet, apart from being computationally tedious, training-based methods depend largely on the test data, and unless a very lengthy set of training data is used, their performance may not be reliable. Within the same line of work, generalized Gamma distribution (GGD) has also been taken into account, which includes some other non-Gaussian PDFs as a special case. In [36], it was confirmed that the most suitable PDF for the modeling of speech STSA priors is the GGD, given that the corresponding parameters are estimated properly. Two mathematical approaches, i.e., the maximum likelihood and the method-of-moments, have been used in [36] for the estimation of the GGD parameters. Other major studies within this field such as those in [23, 37], use either fixed or experimentally set values for the GGD model parameters, lacking the adaptation with the noisy speech data. Hence, an adaptive scheme to estimate the STSA prior parameters with moderate computational burden and fast adaptability with the noisy speech samples is further in need.

1.3.4.3 Multi-Channel Extension

Whereas single microphone approaches are found to provide limited performance improvement, their multiple microphone counterparts have gained increasing popularity, due to their capability in providing higher levels of noise reduction while maintaining small speech distortion. In the context of speech STSA estimation, a few extensions of the conventional single microphone methods have been introduced over the last decade. Cohen *et al.* [38] developed a multi-microphone generalization of the Log-MMSE estimator of the speech STSA by inclusion of the soft-decision estimation of speech presence probabilities. In [19], a general scope for the MAP and MMSE estimation of the spectral amplitude of speech signals was proposed, which considers multiple microphone observations in the case of spatially (across the microphones) uncorrelated noise. Also, it was proved that the optimal MAP estimation of the spectral phase is simply equivalent to the noisy phase of the received signal. Furthermore, a straightforward extension of the speech STSA estimation using the MMSE Bayesian cost function was suggested therein, which assumes spatially uncorrelated noise components and the existence of the same speech component across the noisy observations from different microphones. Later in [39], the MMSE estimation of speech STSA was extended to the microphone array case under the availability of proper estimates for the noise correlation matrix and the steering vector of speech source, given that the speech STSA is Gamma distributed. However, no further improvements were reported in comparison with the spectral amplitude estimation approaches with Rayleigh speech STSA priors. Within the same line, the problem of speech STSA estimation in the presence of spatially uncorrelated noise was further investigated in [40, 41] by making use of various Bayesian cost functions. In a practical point of view, however, the assumption of having uncorrelated noise across different microphones or the perfect knowledge of the steering vector in the frequency domain are too simplistic and not valid in practice. Therefore, more realistic methods in this direction are yet to be developed.

1.4 Overview of Reverberation Reduction Techniques

Another major area of speech enhancement is reverberation reduction. In this section, we present an introduction on the reverberation in acoustic environments and briefly review the classification of the most important techniques to suppress reverberation. In particular, we introduce the problem of blind dereverberation in the STFT domain.

1.4.1 Speech Reverberation in Acoustic Environments

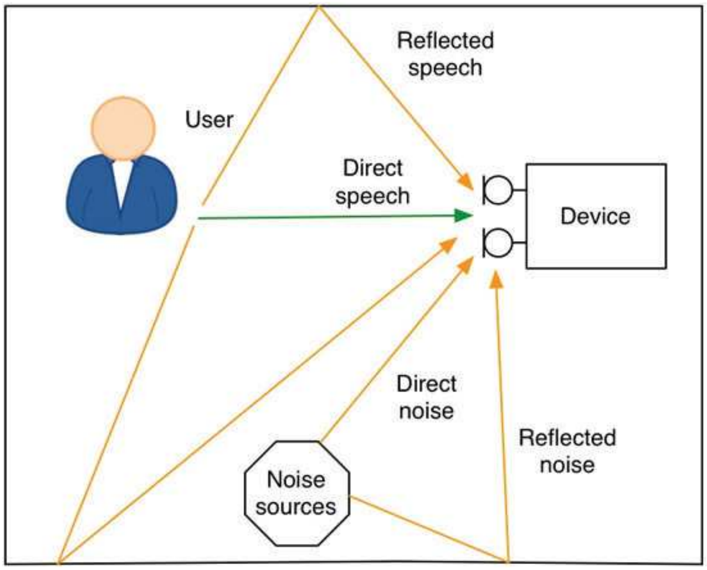


Figure 1.3: Illustration of a speech source (user), noise sources and their reflections captured by a microphone set.

When speech signals are captured in an acoustic environment (enclosed space) by the microphones positioned at a distance from the speech source, the received signal consists of the superposition of many delayed and attenuated replicas of the original speech signal due to the reflections from the surrounding walls and objects, as illustrated in Figure 1.3. Often, the direct path is defined as the acoustic propagation path from the speech source to the microphone without the reflections. It should be noted that a delay of the superimposed speech replicas always rises since all other propagation paths are longer than the direct path [42].

If low-to-moderate reverberation effects are carefully controlled, the reverberation can be tolerable in voice communication systems. However, when the reverberation effects are severe, the quality and intelligibility of speech are degraded and the performance of speech enhancement algorithms developed without taking reverberation into account is highly degraded. This is due to the fact that reverberation deteriorates the characteristics of the speech signal, which is problematic to speech processing applications including speech recognition, source localization and speaker verification. For this reason, development of efficient techniques to suppress reverberation in acoustic environments is of high demand for speech communication systems.

1.4.2 Classification of Reverberation Reduction Techniques

Since reverberation reduction (or namely, dereverberation) techniques have been around for many years, they can be divided into many categories. One useful way of categorizing these techniques is based on the fact whether or not the acoustic impulse response needs to be estimated. This has been considered in [43] wherein two major categories of dereverberation techniques are recognized: reverberation suppression and reverberation cancellation. The former group refers to the methods that do not require an estimate of the acoustic impulse response whereas the methods within the latter group do require/exploit an estimate of the acoustic impulse response. Also, methods within each of these categories can be further divided into smaller sub-categories depending on the amount of knowledge they require about the source of speech and the acoustic channel. According to [43], main reverberation suppression methods include explicit speech modeling, linear prediction-based methods, spectral enhancement, temporal envelope filtering and spatial processing. Also, the most important reverberation cancellation methods include blind deconvolution, homomorphic deconvolution and harmonicity-based dereverberation. The methods in each of the two main categories can be also divided based on being applicable to single channel, multi-channel or both.

1.4.3 Blind Dereverberation in the STFT Domain

In a practical point of view, there exists no knowledge of the acoustic impulse response in a reverberant environment. Also, the estimation of a typical room impulse response (RIR), which involves hundreds of samples, by using the observed speech utterance seems impractical, especially for real-time systems where no long-term training data is available. For this reason, blind dereverberation techniques, i.e. those which do not require any prior knowledge of the RIR or characteristics of the channel or speech source, are of high importance in real world scenarios. In this sense, a few major techniques for blind dereverberation in the STFT domain exist in the literature, including spectral enhancement, spatial processing and linear prediction-based techniques.

Primarily in [44], Lebart *et al.* proposed a single-microphone spectral enhancement technique for speech dereverberation. This method follows the same structure as the spectral enhancement for noise reduction except that the noise variance estimate is replaced by an estimate of the reverberation variance. The latter is obtained blindly from the reverberant speech using statistical modeling of room reverberation and dereverberation is achieved by spectral subtraction. This work was modified and extended using different variants of the spectral enhancement method and also estimators of the reverberation variance.

In addition to noise reduction purposes, spatial processing (beamforming) techniques can also be employed for multi-microphone speech dereverberation. In these techniques, the spatial observations can be manipulated to enhance or attenuate signals arriving from particular directions. Therefore, by using spatial processing, under the *a priori* knowledge of the position of the source, the reverberant part of speech can be spatially separated from the desired part. As one major example, in [45], a two-stage beamforming approach for dereverberation is presented where in the first stage, a delay-and-sum beamformer is exploited to generate a reference signal containing a spatially filtered version of the desired speech and reverberation. It is shown that the desired speech component at the output of the beamformer contains less reverberation compared to input reverberant speech. In the second stage, the filtered microphone signals and the reference signal are used to estimate the desired speech component.

It is well known that using the time-varying nature of speech signals allows one to achieve high quality speech dereverberation based on multi-channel linear prediction [42]. However, such approaches have a heavy computational cost in order to calculate large covariance matrices in the time domain. To overcome this problem and to make it possible to combine the speech dereverberation efficiently with other useful speech enhancement techniques in the STFT domain, in [45], an approach for linear prediction-based dereverberation in the STFT domain was proposed. It was revealed that the proposed approach in the STFT domain, in addition to being computationally less complex, performs even better than the linear prediction-based approaches in the time domain. The implementation of this method in the STFT domain, known as the weighted prediction error method, has received considerable attention in the relevant literature and a few improvements and modifications of that have been presented so far.

1.5 Motivation and Objectives of the Research

1.5.1 Motivation

This research is motivated by the rapidly growing market of speech and audio processing applications. Even though spectral modification techniques for speech enhancement have received much interest over the past three decades, there is still room for further development in this area. In this section, we summarize the motivation behind this research as the following:

- As discussed in Section 1.3.4, various MMSE-based cost functions and also speech priors have been exploited in order to derive Bayesian STSA estimators. Although various expressions have been obtained for these estimators, there has been no unified assessment of their performance or an investigation showing the most efficient Bayesian STSA estimator given the different cost functions and available STSA prior distributions. In this regard, a study on the most generalized Bayesian STSA estimator, i.e., one that includes most state-of-the-art estimators as special cases, as well as the most efficient schemes to select the corresponding parameters is required. This is one of the primary motivations of this research.

- In the field of speech spectral enhancement, as discussed in Section 1.3, numerous single microphone (single-channel) techniques already exist. However, the performance of single channel methods deteriorates considerably in adverse noise conditions. Furthermore, one main problem with all single-channel methods is that they introduce considerable distortion in the clean speech component. This has motivated researchers to employ multi-microphone (dual, array and distributed) systems to exploit all available spatial information of the speech and noise/interference sources [46]. Whereas conventional wideband beamforming techniques for speech enhancement have been studied thoroughly in the literature, the multi-microphone counterparts of speech spectrum estimation methods can be investigated further. Therefore, development of novel multi-channel spectral estimation approaches is of high interest and serves as another major motivation of this research.

- To date, several major categories of methods have been proposed for reverberation suppression in the spectral domain, e.g., [47, 48, 49]. Such methods often aim at estimation of the complex spectral coefficients of speech in the STFT domain. However, development of STSA estimators for the purpose of reverberation suppression has to be explored further. This brings about the motivation to investigate further the capability of STSA estimators for speech enhancement in reverberant environments as part of this research.

- Concerning multi-channel dereverberation methods, many existing methods such as channel equalization and inverse filtering approaches are in need of estimates of the acoustic channel, and therefore, they are not practically useful, necessitating the need for totally blind dereverberation methods [42]. As discussed in Section 1.4.3, one of the most important multi-channel enhancement methods in the STFT domain is the spatial processing (beamforming). Even though beamforming for noise reduction has been explored extensively in the existing literature, taking advantage of beamformers for reverberation suppression in an unknown reverberant environment has to be explored further. As the most basic beamforming technique, the delay-and-sum beamformer has been widely employed for reverberant environments [43]. However, the capability of more advanced beamformers such as the minimum variance distortionless response (MVDR) or multi-channel Wiener filtering, which are in need of reverberation statistics, has not been investigated enough. In many cases, these beamformers are applied under the assumption that a perfect estimate of RIR in the STFT domain is available, e.g., [50]. Therefore, blind development of beamforming methods such as the MVDR for highly reverberant environments under practical assumptions is of interest as part of this work.

- As a more recent dereverberation technique in the STFT domain, the linear prediction-based method first proposed in [51] has also received considerable attention due to its blind nature and reasonable complexity. The original version of this method is, however, based on simplistic assumptions such as using a simple instantaneous estimator for the speech spectral variance and independence of the desired speech components across time/frequency. While the original proposition of this method has proved to achieve good reverberation suppression performance, it is believed that by developing and incorporating more accurate speech spectral variance estimators into this method, as well as taking into account the correlation across the speech components, further dereverberation and better speech quality can be achieved.

1.5.2 Objectives

The main objectives of this research are summarized as follows:

- With regards to the single-channel Bayesian STSA estimators, the objective is to obtain a generalized formulation for the gain function using the most efficient Bayesian cost function and speech STSA prior available in the literature. Also, based on the characteristics of speech/noise such as the SNR, noise masking threshold and properties of human auditory system, efficient schemes for the selection of the corresponding parameters of the STSA estimator are to be proposed.

- Regarding the multi-channel counter-part of the Bayesian STSA estimators, we will investigate the extension of the proposed single-channel method to the multi-channel in two different scenarios, namely, in spatially uncorrelated and spatially correlated noise fields. Whereas in the former, only the noise parameters (i.e., noise variance and SNR) for each channel are needed, in the latter, the noise cross-variances among all the channels are required to form the estimator. Since the problem of noise cross-variance estimation is not as much developed as the classic noise estimation, we also target this problem as part of our multi-channel STSA estimation method.

- Considering blind speech dereverberation in the STFT domain using the spectral enhancement approach, e.g., STSA estimation, our objective is to develop/modify the schemes used in the noise reduction scenario, e.g., the noise variance, SNR and gain flooring, in order to properly fit them into the reverberation suppression goal. Furthermore, blind development of the classic beamforming methods (such as the MVDR beamformer) for the purpose of reverberation suppression, which has not been studied particularly in the literature, is in order as part of this topic.

- Our objective in case of dereverberation based on a linear prediction model in the STFT domain is to develop an efficient estimator of the speech spectral variance that can be integrated into the dereverberation method. As well, taking into account the existing correlations across the STFT frames through a proper model, and development of the reverberation prediction weights based on these correlations can be regarded as our other objective in this sense.

1.6 Organization and Main Contributions

In Chapter 2, a more detailed background on the introduced topics in this section is presented, and Chapters 3-4 and 5-6 include our main contributions respectively in noise reduction and reverberation suppression. Conclusions are presented in Chapter 7. A detailed structure of this thesis is as below.

In Chapter 2, a background on the topic of speech enhancement in the STFT domain with a focus on STSA estimators is presented. These estimators include the spectral subtraction method, the Wiener filters, ML and MAP estimators, and as the most important category in our work, the Bayesian estimators, which are discussed briefly in Section 2.1. In Section 2.2, an overview of the various speech priors that have been used in the STSA estimation methods is presented and Section 2.3 reviews the state-of-the-art multi-channel STSA estimators for spatially uncorrelated noise. Section 2.4 is devoted to an introduction on the reverberation in acoustic environments and the general problem formulation in the STFT domain. Finally in Section 2.5, some of the shortcomings of the current STSA estimation methods, which will be worked on in the following chapters, are explained.

In Chapter 3, we present the proposed single-channel STSA estimation algorithm, which includes novel schemes for the parameter selection of the Wβ-SA estimator as well as a new gain flooring scheme which can be generally applied to STSA estimators. Next, we extend the original Wβ-SA estimator using the GGD speech prior and suggest an efficient scheme for the estimation of its parameters. This chapter is followed by a brief overview of the objective measures for the evaluation of noise reduction methods, and finally the performance assessment of the proposed schemes versus the most recent versions.

In Chapter 4, first we extend the proposed single-channel STSA estimator in the previous chapter to the case of multi-channel for spatially (across channels) uncorrelated noise. Next, we take advantage of the speech spectral phase in the estimation of the spectral amplitude, i.e., STSA, and derive a new family of STSA estimators for different Bayesian cost functions. The problem of multi-channel STSA estimation in the general case (spatially correlated noise) is tackled in the next section, where a generic framework for the extension of a single-channel STSA estimator to its corresponding multi-channel variant is induced. Since under this framework, the estimation of the noise spatial correlation matrix is of high importance, we propose an efficient algorithm for the estimation of the aforementioned matrix in the next section. Performance evaluations are performed in this chapter separately for the case of spatially correlated/uncorrelated noise fields.

Chapter 5 is devoted to the problem of reverberation suppression in the STFT domain using a popular linear prediction-based method. In this respect, we consider the so-called weighted prediction error (WPE) method and present two main contributions on this method. Our contributions include the proposition of an efficient estimator for the speech spectral variance, which can be integrated into the original WPE method to substitute the instantaneous estimator of this parameter used in this method. Further, we take into account the temporal correlation across the STFT frames, and through an approach to estimate this correlation, we propose an extension of our primary method. Finally, we evaluate the performance of the proposed methods in terms of the achieved dereverberation and overall speech quality.

In Chapter 6, we target the problem of reverberation suppression from the viewpoint of spectral enhancement methods, i.e., those that were conventionally used for noise reduction. We first propose a new algorithm for the most important parameter involved in these estimators, i.e., the late reverberant spectral variance (LRSV). Next, we suggest a few simple yet efficient schemes for the modification of the conventional STSA estimators to fit the reverberation problem, which include the estimation of signal-to-reverberant ratio (SRR), gain flooring and the application of SPP in case of reverberation. This chapter is followed by performance evaluations in comparison with the recent major contributions in the field.

In Chapter 7, we draw some concluding remarks highlighting the main contributions of this thesis, and based on this, we suggest some open problems for the future research in this direction.

Chapter 2

Background: Speech Enhancement in the STFT Domain

In this chapter, we present a literature review on the existing techniques for the enhancement of speech spectrum in the STFT domain. First, a background on the estimation of speech STSA in the presence of noise and some relevant problems including their extension to the multi-channel case and reverberant environments are explained. Next, the problem of blind reverberation suppression in the STFT domain is discussed shortly. This is followed by the most important shortcomings of the current STSA estimators that motivated us to develop further solutions in this area.

The presented content in this chapter along with further literature review has been published in [52].

2.1 Estimation of Speech STSA

As discussed in the last chapter, the spectral amplitude has been found to be more perceptually relevant than the spectral phase in the field of speech enhancement. For this reason, various estimators of speech short-time spectral amplitude (STSA) have been widely used to perform single-channel noise reduction. In this section, we present a brief overview of the different types of STSA estimators including spectral subtractive estimation, Wiener filtering for speech STSA, maximum-likelihood estimators, maximum *a posteriori* estimators and finally the category of Bayesian STSA estimators.

2.1.1 Problem Statement

An enhancement algorithm in the STFT domain transforms short segments of a noisy speech signal into STFT coefficients and synthesizes the enhanced signal by means of an inverse STFT and overlap adding. Between the STFT and inverse STFT calculations, the magnitude of the STFT coefficients corresponding to the enhanced speech signal is estimated using the underlying spectral enhancement algorithm. In this sense, we consider the following model for noisy speech observations

$$y(t) = x(t) + v(t) \tag{2.1}$$

where $y(t)$, $x(t)$ and $v(t)$ respectively denote the noisy observation, clean speech and noise at time t. After sampling into discrete time, segmentation (framing), windowing and applying FFT, we have in the STFT domain the following complex-valued model

$$Y(k,l) = X(k,l) + V(k,l) \tag{2.2}$$

with k and l as frequency bin and time frame indices. Applying the standard assumption that speech and noise are statistically independent from each other and also independence across time and frequency, we will obtain estimators that are independent in the time frame and frequency. This allows us to drop the time/frequency indices henceforth.

2.1.2 Spectral Subtractive Estimators

Spectral subtraction is one of the first category of algorithms proposed for noise reduction in the frequency domain [7]. It is based on the simple principle that, given an estimate of the noise spectrum, an estimate of the clean speech spectrum can be obtained by subtracting the noise estimate from the noisy speech spectrum. More specifically, assuming the similarity between the phase of the noisy speech and that of the clean speech, it follows that [14]

$$\hat{X}(k,l) = \Big[|Y(k,l)| - |\hat{V}(k,l)|\Big] e^{j\Theta_Y(k,l)} \tag{2.3}$$

where $|.|$ denotes the amplitude and $\Theta_Y(k,l)$ is the phase of $Y(k,l)$. Note that the effect of noise on the clean speech phase is assumed negligible in (2.3), whereas in practice, availability of the clean speech phase or a better estimate of it to replace $\Theta_Y(k,l)$ can provide further quality improvements [53]. Due to the inaccuracy in the underlying noise estimation, the subtractive term, $|Y(k,l)|-|\hat{V}(k,l)|$, can take on negative values and a half-wave rectification is conventionally used to mitigate this effect. This rectification causes a phenomenon known as musical noise, which can significantly degrade the speech quality up to a high degree [7]. This issue has been one of the main motives to develop more advanced spectral subtractive methods in the past, e.g., [54, 55, 56].

In practice, since the majority of noise estimation methods seek to estimate the noise spectral variance, $\sigma_v^2(k,l)$, defined as $E\{|V(k,l)|^2\}$, spectral subtractive methods are often formulated in the power domain rather than in the amplitude domain. In this regard, an estimate of the clean speech amplitude, $|\hat{X}(k,l)|$, can be obtained as

$$|\hat{X}(k,l)|^2 = |Y(k,l)|^2 - \hat{\sigma}_v^2(k,l) \tag{2.4}$$

where $\hat{\sigma}_v^2(k,l)$ is an estimate of the noise spectral variance or the so-called PSD [7]. It is evident that the performance of spectral subtractive methods is highly controlled by the precision in the estimation of the noise PSD, $\sigma_v^2(k,l)$. Since the estimated speech amplitude can be written as a linear function of the noisy speech amplitude, it is often preferred to express spectrum estimation techniques in terms of a gain function. In this sense, the gain function for the estimator in (2.4) can be written as

$$G(k,l) \triangleq \frac{|\hat{X}(k,l)|}{|Y(k,l)|} = \sqrt{1 - \frac{\hat{\sigma}_v^2(k,l)}{|Y(k,l)|^2}} \tag{2.5}$$

For a better understanding of the concept of spectral subtraction, a block diagram of this method in its basic form is shown in Figure 2.1. It is observed that, within this framework, only the spectrum amplitude is enhanced and the spectral phase is left unchanged.

One of the most important advances in the area of spectral subtractive methods is the use of masking properties of the human auditory system first introduced in [57]. The masking properties are essentially modeled by a noise masking threshold below which a human listener tolerates additive noise in the presence of speech [58]. In the generalized spectral subtractive methods, e.g., [59, 60], there exist parameters that control the trade-off between the amount of noise reduction, the speech distortion and the residual musical noise. In [57], a few schemes are proposed based on the noise masking threshold in order to adjust the subtractive parameters in a perceptual sense. Therein, through the study of speech spectrograms as well as subjective listening tests, it is proved that the resulting enhanced speech is more pleasant to a human listener than without adaptive adjustment of the subtractive parameters.

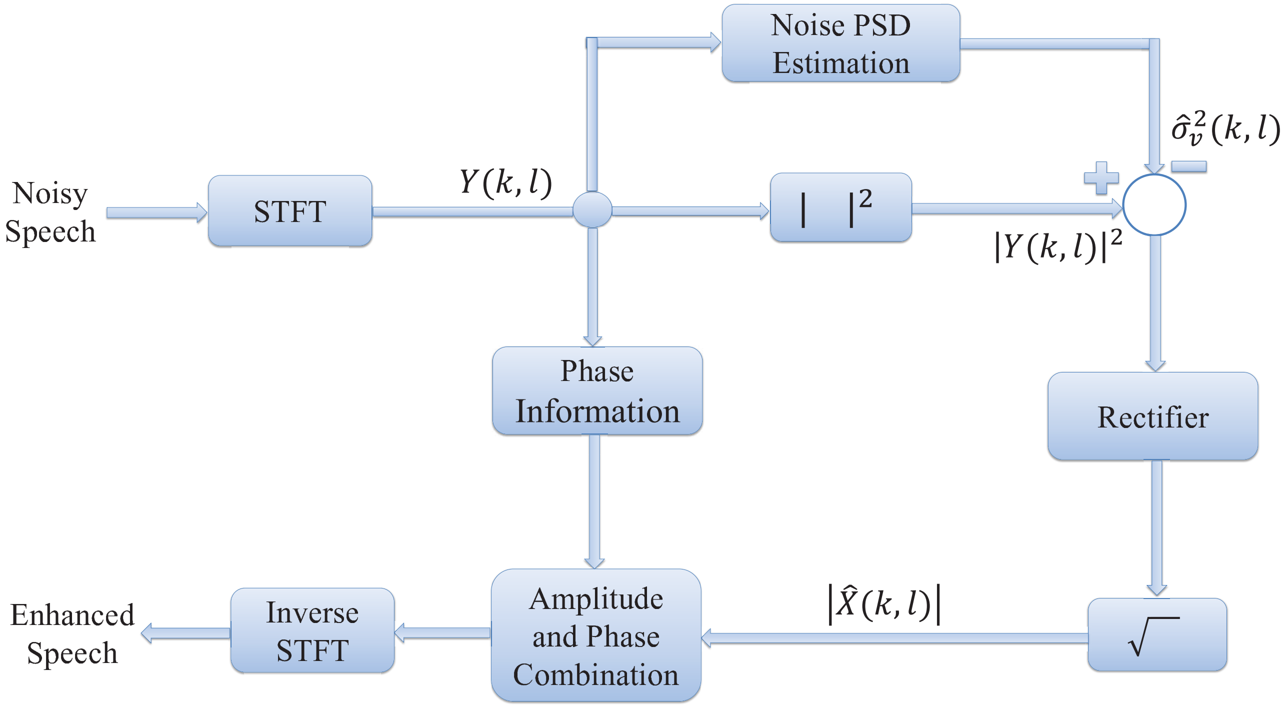


Figure 2.1: Block diagram of the basic spectral subtraction algorithm.

The spectral subtraction algorithms are computationally simple to implement and fast enough for real-time applications. Nevertheless, the subtractive rules are based on the incorrect assumption that the cross terms between the clean speech and the noise are zero. In other words, considering (2.4) and the fact that $\hat{\sigma}_v^2(k,l)$ is used for $|V(k,l)|^2$, the speech squared amplitude $|X(k,l)|^2$ is not accurately equal to $|Y(k,l)|^2 - |V(k,l)|^2$, and the cross terms between the speech and noise have to be considered in the subtraction rule. In [61], a geometric approach (as opposed to the statistical approaches) to spectral subtraction is proposed that addresses this shortcoming of the spectral subtraction method. In that work, the phase difference between the clean speech and noise is exploited in order to obtain the spectral subtraction rule as a gain function. The resulting gain function depends on two key parameters, that is the *a priori* SNR and the noise PSD, and it possesses similar properties to those of the MMSE STSA estimator in [17]. It is further shown through objective evaluations that the geometric algorithm performs significantly better than the traditional spectral subtraction algorithm under various conditions.

Other main contributions to the spectral subtraction method in the literature include spectral subtraction using oversubtraction [62], nonlinear spectral subtraction [63], multi-band spectral subtraction [64], MMSE-based spectral subtraction [59], extended spectral subtraction [65], use of adaptive gain averaging [55] and selective spectral subtraction [66]. Even though spectral subtraction is one of the oldest methods of noise reduction in the STFT domain, there still exists ongoing research on this topic.

2.1.3 Wiener Estimators

The spectral subtractive methods discussed in the previous section are based on the heuristic assumption that one can obtain an estimate of clean speech spectrum by subtracting the estimated noise spectrum from the observations spectrum. Despite being intuitively pleasing and computationally simple, this method cannot make any claim of optimality. In this part, we briefly review the concept of Wiener filtering in the STFT domain. In this approach, the estimated speech spectrum is obtained as $\hat{X}(k,l) = W(k,l)Y(k,l)$ where W(k,l) denotes the corresponding gain function. The latter is derived by minimizing the mean square error (MSE) between the clean and estimated speech spectra, which is mathematically expressed as

$$\hat{W}(k,l) = \underset{W}{\operatorname{argmin}}\ E\left\{\left|X(k,l) - WY(k,l)\right|^2\right\} \tag{2.6}$$

with $E\{.\}$ denoting the statistical expectation. Solving the above, we obtain the general form of the complex-valued Wiener filter gain as

$$\hat{W}(k,l) = \frac{\sigma_{xy}(k,l)}{\sigma_y^2(k,l)} \tag{2.7}$$

where $\sigma_{xy}(k,l)$ denotes the cross-PSD between the clean and noisy speech defined as $E\left\{X(k,l)Y^*(k,l)\right\}$ and $\sigma_y^2(k,l)$ denotes the noisy speech PSD [67]. In practice, both σ_{xy} and σ_y^2 in (2.7) are unknown and have to be estimated. Henceforth, we may drop the time frame and frequency indices for notational convenience. Even though the estimation of σ_y^2 can be done in a straightforward way, such as recursive smoothing of the observations, $Y(k,l)$, estimation of the cross-term σ_{xy} is generally challenging and depends on the application [68]. Assuming uncorrelated clean speech and noise signals, σ_{xy} and σ_y^2 respectively simplify to the clean speech PSD, σ_x^2, and the sum $\sigma_x^2 + \sigma_v^2$. Now, by defining the *a priori* SNR as $\zeta = \sigma_x^2/\sigma_v^2$, the Wiener filtering gain can be expressed as $\zeta/(1+\zeta)$.

The *a priori* SNR, which is a critical parameter in the context of noise reduction, can be estimated through the conventional decision-directed approach [17] and its more advanced variations found in [69, 70, 71]. The aforementioned method in (2.7) is the most conventional way for computing the Wiener filter gain function from the available noisy speech. Several more advanced methods have been proposed in the relevant literature in order to implement the Wiener filter more efficiently and overcome some of its shortcomings, e.g. in [72, 73, 74].

2.1.4 Maximum Likelihood (ML) Estimators

Firstly proposed in [10], the ML estimation is the most conventional and simple method to estimate speech STSAs. Therein, by using a two-state model for the presence or absence of the speech and the ML estimation rule for the STSA, a class of noise suppression functions was developed, allowing a trade-off of noise suppression against speech distortion. In the ML estimator of [10], the speech is characterized by a deterministic waveform with unknown amplitude and phase while the noise is assumed complex Gaussian. In this case, denoting the clean speech STFT by $X = \mathcal{X}e^{j\omega}$ with $\mathcal{X}$ and ω as the speech STSA and phase respectively, the distribution of noisy observations given the speech signal is

$$p\left(Y|\mathcal{X},\omega\right) = \frac{1}{\pi\sigma_v^2}\exp\left(-\frac{|Y|^2 - 2\mathcal{X}\Re\{e^{-j\omega}Y\} + \mathcal{X}^2}{\sigma_v^2}\right) \tag{2.8}$$

where $\Re\{.\}$ denotes the real value. Note that for ease of readability the time and frequency indices have been dropped. Taking the statistical mean of $p\left(Y|\mathcal{X},\omega\right)$ over the nuisance parameter ω and then maximizing it with respect to $\mathcal{X}$, the ML estimator of the speech amplitude $\mathcal{X}$ is obtained as the solution to the following

$$\hat{\mathcal{X}}_{ML} = \max_{\mathcal{X}} \int_0^{2\pi} p\left(Y|\mathcal{X},\omega\right)p(\omega)d\omega \tag{2.9}$$

Using a uniform distribution for the speech phase ω and applying an exponential approximation to the Bessel function appearing from the integration in (2.9), the following ML estimator was derived [10]

$$\hat{\mathcal{X}}_{ML} = \frac{|Y| + \sqrt{|Y|^2 - \sigma_v^2}}{2} \tag{2.10}$$

More recently in [37], a phase equivalence between speech and noise spectral components was assumed, resulting that $\mathcal{Y} = \mathcal{X} + \mathcal{V}$ with $\mathcal{Y}$ and $\mathcal{V}$ as the amplitude of observations and noise, respectively. In that work, the following GGD model was considered for the distribution of the noise amplitude

$$p(\mathcal{V}) = \frac{ab^c}{\Gamma(c)} \mathcal{V}^{ac-1} \exp(-b\mathcal{V}^a), \quad \mathcal{V} > 0;\; a, b, c > 0 \tag{2.11}$$

with $\Gamma(.)$ as the Gamma function and (a, b, c) as the distribution parameters. Taking into account the phase equivalence between the speech and noise, the likelihood function $p(\mathcal{Y}|\mathcal{X})$ is actually $p(\mathcal{V})$ in (2.11) with $\mathcal{V}$ replaced by $\mathcal{Y} - \mathcal{X}$. Maximizing the logarithm of this likelihood function with respect to $\mathcal{X}$ results in the following generalized ML estimator

$$G_{ML} = \frac{\hat{\mathcal{X}}}{\mathcal{Y}} = 1 - \frac{1}{\sqrt{\gamma}} \left(\frac{ac-1}{a\sqrt{c(c+2-a)}} \right)^{1/a} \quad , a \in \{1, 2\},\; ac > 1 \tag{2.12}$$

with γ denoting the *a posteriori* SNR defined as $\mathcal{Y}^2/\sigma_v^2$, and that the solution to the maximization exists only for the given constraints in (2.12).

2.1.5 Maximum *a Posteriori* (MAP) Estimators

Apart from having limited noise reduction performance, an ML estimator does not take into account the distribution of speech STSA (the so-called speech prior), whereas a proper model for the speech prior can be used in a MAP estimator. In [75], under a complex Gaussian assumption for the speech prior and a Bayesian framework, MAP estimators of the speech STSA were derived as simpler alternatives to the Ephraim and Malah's MMSE-based approach. Therein, three different estimators were proposed, namely, the joint MAP estimator of speech spectral amplitude and phase, the MAP estimator of the speech spectral amplitude and the MMSE estimator of speech PSD. The joint MAP spectral amplitude and phase estimator can be expressed as [75]

$$\left(\hat{\mathcal{X}}^{(\mathrm{MAP})}, \hat{\omega}^{(\mathrm{MAP})} \right) = \underset{\mathcal{X}, \omega}{\operatorname{argmax}}\; p(\mathcal{X}, \omega | Y) \tag{2.13}$$

and closed-form solutions for the speech spectral amplitude and phase are derived from (2.13). The interesting result, however, is that the estimator of the speech spectral phase obtained by (2.13) is just the noisy phase of speech observations. The same result was deduced in [17] with the MMSE estimate of speech spectral phase. Next, the spectral amplitude-only estimator can be given by solving the following [75]

$$\hat{\mathcal{X}}^{(\mathrm{MAP})} = \underset{\mathcal{X}}{\operatorname{argmax}}\ E_{\omega}\left\{p(\mathcal{X}, \omega|Y)\right\} \tag{2.14}$$

where using an exponential approximation to the Rician distribution, obtained from $E_{\omega}\left\{p(\mathcal{X}, \omega|X)\right\}$, leads to a closed-form solution. It is shown that this solution is a generalized form of the approximate solution to the ML estimator proposed in [10]. Next, by deriving an expression for the second moment of the Rician posterior, i.e., $E\{\mathcal{X}^2|Y\}$, which is actually the MMSE estimate of the speech spectral variance, σ_x^2, and taking its square root, an estimate of speech spectral amplitude is obtained and combined with the noisy phase. Analysis of the behavior of the corresponding gain functions for all three estimators shows that they have a similar performance to the Ephraim and Malah's solution, whilst they permit a more straightforward implementation and simpler expressions by avoiding Bessel and confluent hypergeometric functions.

More recently in [19], it was indicated through extensive experimentations that the class of super-Gaussian distributions fits speech STSA priors more properly than the conventional Rayleigh deduced from the complex Gaussian assumption for speech STFT coefficients. Therein, within the framework of MAP spectral amplitude estimators, the distribution of the speech spectral amplitude is modeled by a simple parametric function, which allows a high approximation accuracy for Laplace- or Gamma-distributed real and imaginary parts of the speech STFT coefficients. Also, the statistical model can be adapted using the noisy observations to optimally fit the distribution of the speech spectral amplitudes. Based on the super-Gaussian statistical model, two computationally efficient MAP spectral amplitude estimators are derived, which outperform the previously proposed ones in [75] while owning the same simplicity as the estimators in [75]. The two estimators in [19] include a joint amplitude-phase estimator and an amplitude-only estimator and can be both expressed as extensions of the MAP estimators proposed in [75].

2.1.6 MMSE-Based (Bayesian) Estimators

Within the frequency domain class of methods, the Bayesian approach is particularly attractive due to its superior performance. In this approach, an estimator of the clean speech is derived by minimizing the statistical expectation of a cost function that penalizes errors in the clean speech estimate. Various Bayesian estimators of the STSA have been proposed on the basis of different cost functions to model the error between the clean and estimated spectral amplitude. In this part, we review briefly the most important Bayesian estimators of the speech STSA and their properties.

2.1.6.1 Ephraim and Malah's MMSE and Log-MMSE Estimators

Since the spectral subtraction STSA estimation method is resulted from an optimal variance estimator in the maximum likelihood sense, and also, the Wiener speech spectral estimator is derived from the optimal MMSE signal spectral estimator, both of these conventional methods are not optimal in the sense of spectral amplitude. This observation served as the primary motivation for Ephraim and Malah to seek an optimal STSA estimation scheme. In [17], Ephraim and Malah proposed a class of speech enhancement algorithms which capitalize the importance of the STSA of speech signals. A speech denoising system using the MMSE criterion was proposed and its performance was compared against the other widely used methods at the moment, i.e., the basic Wiener filtering and spectral subtraction. The speech spectral component $X(k,l)$ can be written as $X(k,l) = \mathcal{X}(k,l)e^{j\omega(k,l)}$ with $\mathcal{X}(k,l)$ being the spectral amplitude and $\omega(k,l) \in [-\pi,\pi]$ the spectral phase. The STSA estimation algorithm aims at the estimation of the speech spectral amplitude, $\mathcal{X}(k,l)$, given the noisy spectral observation $Y(k,l)$. To derive this estimator, the *a priori* distribution of the speech and noise STFT coefficients should be known. Since in practice they are unknown, a reasonable statistical model for the underlying distributions is required. In [17], it is assumed that the STFT coefficients of each process can be modeled as statistically independent complex Gaussian random variables with zero mean. However, the variance of both the speech and noise STFT coefficients is, due to speech non-stationarity, time varying and therefore, it must be estimated continuously. Therefore, the amplitude of the speech STFT coefficients, $\mathcal{X}(k,l)$, has a Rayleigh distribution, whereas the phase component, $\omega(k,l)$, is considered to be uniformly distributed over $[0,2\pi]$ and independent of the amplitude [76]. Hence, the following holds

$$p(Y|\mathcal{X},\omega) = \frac{1}{\pi\sigma_v^2}\exp\left(-\frac{1}{\sigma_v^2}|Y - \mathcal{X}e^{j\omega}|^2\right) \tag{2.15}$$

$$p(\mathcal{X},\omega) = p(\mathcal{X})p(\omega) = \frac{\mathcal{X}}{\pi\sigma_{\mathcal{X}}^2}\exp\left(-\frac{\mathcal{X}^2}{\sigma_{\mathcal{X}}^2}\right) \tag{2.16}$$

where $p(Y_k|\mathcal{X},\omega)$ is the distribution of noisy observations conditioned on the signal component, $p(\mathcal{X},\omega)$ is the joint distribution of speech magnitude and phase. These parameters are generally unknown and have to be estimated beforehand. Now, the MMSE estimator of the speech STSA is given as

$$\hat{\mathcal{X}}^{(\mathrm{MMSE})} = \underset{\hat{\mathcal{X}}}{Argmin} \;\; E\left\{(\mathcal{X} - \hat{\mathcal{X}})^2\right\} \tag{2.17}$$

It can be proved that the MMSE estimator above is equivalent to the conditional expectation $E\{\mathcal{X}|Y\}$. This statistical expectation can be solved using Bayes' rule to express the *a posteriori* distribution, $p(\mathcal{X}|Y)$, leading to

$$E\{\mathcal{X}|Y\} = \frac{\int_0^\infty \int_0^{2\pi} \mathcal{X} p(Y|\mathcal{X},\omega)p(\mathcal{X},\omega)d_\omega d_\mathcal{X}}{\int_0^\infty \int_0^{2\pi} p(Y|\mathcal{X},\omega)p(\mathcal{X},\omega)d_\omega d_\mathcal{X}} \tag{2.18}$$

By substitution of (2.15) and (2.16) into (2.18), the final solution to the Bayesian STSA estimation problem can be obtained as [77]

$$\hat{\mathcal{X}}^{(\mathrm{MMSE})} = \Gamma(1.5)\frac{\sqrt{\nu}}{\gamma}\mathrm{M}(-0.5, 1; -\nu).|Y| \tag{2.19}$$

where $\Gamma(x)$ and $\mathrm{M}(a, b; z)$ denote the Gamma and confluent hypergeometric functions, respectively. Also, the gain parameters γ and ν are defined as

$$\zeta = \frac{\sigma_\mathcal{X}^2}{\sigma_v^2} \;,\;\; \gamma = \frac{|Y|^2}{\sigma_v^2} \;,\;\; \nu = \frac{\zeta}{1+\zeta}\gamma \tag{2.20}$$

where ζ and γ are the so-called *a priori* and *a posteriori* SNR, respectively. Whereas ζ can be interpreted as the instantaneous SNR, $\gamma - 1$ acts as a long-term estimator of the SNR. Note that the STSA estimation solution can be always expressed as a gain function G multiplied by the STSA of the noisy observations $|Y|$, hence, it can be interpreted as a linear filter in the frequency domain.

In [17], the STSA estimation problem was formulated using the most basic cost function modeling the error, i.e., the mean square error function. Although this led to an analytically tractable solution with considerable improvements, it is not necessarily the most subjectively meaningful cost function. This has led to the development of more recent cost functions and their corresponding STSA estimation solutions in the relevant literature. In this direction, Ephraim and Malah suggested a logarithmic MMSE version of their method in [18], which gained further improvements in most of the evaluations. Therein, instead of the classical MMSE cost function in (2.17), they managed to optimize $E\left\{(\log \mathcal{X} - \log \hat{\mathcal{X}})^2\right\}$ and used the theory of moments to come up with an analytical solution.

2.1.6.2 Perceptually Motivated Bayesian Estimators

The MMSE cost function is most commonly used in estimation theory because it is mathematically tractable and easy to evaluate. However, it might not be the most subjectively meaningful cost function, as small and large MMSE estimation errors might not respectively correspond to good and poor speech quality. To overcome the shortcomings of the MMSE cost function, in [29], a few Bayesian estimators of the speech STSA based on perceptually motivated distortion measures were proposed for the first time. In general, the Bayesian STSA estimation problem can be formulated as the minimization of the expectation of a cost function representing a measure of distance between the true and the estimated clean speech STSAs, denoted respectively by $\mathcal{X}(k,l)$ and $\hat{\mathcal{X}}(k,l)$. This problem can be expressed as

$$\hat{\mathcal{X}}^{(\mathrm{o})} = \underset{\hat{\mathcal{X}}}{\operatorname{argmin}}\; E\left\{C(\mathcal{X},\hat{\mathcal{X}})\middle|Y\right\} \tag{2.21}$$

where $C(.)$ is a particular Bayesian cost function and $\hat{\mathcal{X}}^{(\mathrm{o})}$ is the optimal STSA estimate. Similar to the spectral subtractive methods discussed earlier, the STSA estimate is combined with the noisy phase of speech to provide an estimate of speech STFT coefficients. Further proceeding with (2.21) requires the knowledge of the distribution of speech STSA conditioned on observation, i.e., $p(\mathcal{X}|Y)$, since

$$\begin{aligned} E\left\{C(\mathcal{X},\hat{\mathcal{X}})\right\} &= \int\int C(\mathcal{X},\hat{\mathcal{X}})\; p\,(\mathcal{X},Y)\,d_{\mathcal{X}}d_{\mathrm{Y}} \\ &= \int\left[\int C(\mathcal{X},\hat{\mathcal{X}})p\,(\mathcal{X}|Y)\,d_{\mathcal{X}}\right]p\,(Y)\,d_{\mathrm{Y}} \end{aligned} \tag{2.22}$$

where actually the term inside the brackets has to be minimized with respect to $\hat{\mathcal{X}}$. This is doable in a Bayesian framework for $p\,(\mathcal{X}|Y)$ using the distributions in (2.25). Loizou in [29] introduced the idea of perceptually (to human ear) motivated cost functions and derived STSA estimators that emphasize the spectral peak (formants) information and STSA estimators that take into account the auditory masking effects of the human audition system. Therein, he proposed three classes of Bayesian estimators. The first class of the estimators emphasizes spectral peak information of the speech signal, the second class uses a weighted Euclidean cost function that takes into account the aforementioned auditory masking effects and the third class of estimators is developed to account for spectral attenuation. It was concluded that, out of the three classes of the suggested Bayesian estimators, those based on the auditory masking effect perform best in terms of having less residual noise in the enhanced speech and better speech quality.

Within the same line of work, another major class of the Bayesian STSA estimators was proposed in [31], which is known as the β-order MMSE estimator. The corresponding cost function involves a parameter named β and employs β powers of the amplitude spectra. Thanks to the degree of freedom provided by this parameter, trade-offs between the amount of noise reduction and speech distortion were achieved therein and a few schemes for the experimental or adaptive selection of this parameter were contributed. The experimental results proved the advantage of the namely β-SA estimator, as compared to the previous versions of STSA estimation. Along the same direction, later in [32], it was proposed to exploit a spectrally weighted development of the β-order MMSE cost function including a new weighting parameter called α. Therein, new psycho-acoustical schemes were suggested for the selection of the two parameters, i.e., α and β, based on the properties of the human auditory system. Performance evaluations revealed improvements in the so-called Wβ-SA estimator with respect to using the previously suggested MMSE cost functions in this field. Later in [33], a more generalized Bayesian cost function was introduced by involving a new spectral weighting term and it was indicated that the resulting STSA estimator, named as generalized weighted SA (GWSA), provides further flexibility in the adjustment of the STSA gain function. All the aforementioned STSA estimators can actually be derived as a particular case of the latter.

To facilitate the discussion of the conventional Bayesian STSA estimators with the underlying cost functions, a summary of the major STSA estimators is indicated in Table 2.1. In this table, γ is the *a posteriori* SNR defined as $|Y|^2/\sigma_v^2$, the gain function parameter ν is $\zeta\gamma/(1+\zeta)$ and $\mathrm{M}(.,.;.)$ denotes the confluent hypergeometric function. Note that p, β and α are parameters that shape the STSA gain function, and as explained, a few efficient schemes for their determination have been proposed in the references in Table 2.1.

Method	**Bayesian cost function**	**Gain function**	**Properties**
MMSE [17]	$\left(\mathcal{X} - \hat{\mathcal{X}}\right)^2$	$\frac{\sqrt{\nu}}{\gamma}\Gamma(1.5)\mathrm{M}(-0.5, 1; -\nu)$	Basic version of Bayesian STSA estimators, optimal in the amplitude MMSE sense
Log-MMSE [18]	$(\log \mathcal{X} - \log \mathcal{X})^2$	$\frac{\nu}{\gamma} \exp\left(\frac{1}{2} \int_{\nu}^{\infty} \frac{e^{-t}}{t} d_t\right)$	Outperforms the basic version through the use of the logarithmic distortion measure (cost function) for speech
WCOSH [29]	$\mathcal{X}^p \left(\frac{\mathcal{X}}{\hat{\mathcal{X}}} + \frac{\hat{\mathcal{X}}}{\mathcal{X}} - 1\right)$	$\frac{\sqrt{\nu}}{\gamma} \sqrt{\frac{\Gamma(\frac{p+3}{2})\mathrm{M}(-\frac{p+1}{2},1;-\nu)}{\Gamma(\frac{p+1}{2})\mathrm{M}(-\frac{p-1}{2},1;-\nu)}}$	Weighted cosine hyperbolic cost function, a symmetric distortion measure exploiting auditory masking effects
WE [29]	$\mathcal{X}^p \left(\mathcal{X} - \hat{\mathcal{X}}\right)^2$	$\frac{\sqrt{\nu}}{\gamma} \frac{\Gamma(\frac{p+1}{2}+1)}{\Gamma(\frac{p}{2}+1)} \frac{\mathrm{M}(-\frac{p+1}{2},1;-\nu)}{\mathrm{M}(-\frac{p}{2},1;-\nu)}$ $p > -2$	Distortion measure motivated by the perceptual weighting technique used in low-rate analysis-by-synthesis speech coders

β-SA [31]	$\left(\mathcal{X}^{\beta} - \hat{\mathcal{X}}^{\beta}\right)^{2}$	$\frac{\sqrt{\nu}}{\gamma}\left[\Gamma(\frac{\beta}{2}+1)\mathrm{M}(-\frac{\beta}{2},1;-\nu)\right]^{1/\beta}$ $\beta > -2$	Motivated first by the generalized spectral subtraction method, provides gain function adjustments by the selection of parameter β
$W\beta$-SA [32]	$\left(\frac{\mathcal{X}^{\beta}-\hat{\mathcal{X}}^{\beta}}{\mathcal{X}^{\alpha}}\right)^{2}$	$\frac{\sqrt{\nu}}{\gamma}\left(\frac{\Gamma(\frac{\beta-2\alpha}{2}+1)\mathrm{M}(-\frac{\beta-2\alpha}{2},1;-\nu)}{\Gamma(-\alpha+1)\mathrm{M}(\alpha,1;-\nu)}\right)^{1/\beta}$ $\beta > 2(\alpha-1)\ , \alpha < 1$	Further flexibility in the gain function, selection of parameters α and β based on psycho-acoustical properties of human audition

Table 2.1: Major Bayesian estimators of the speech STSA

2.1.7 Use of Speech Presence Probability (SPP)

The structure of the estimators discussed in this chapter is based on the assumption that speech is actually present in the noisy speech in all time-frequency units. This is obviously not true for speech pauses, i.e., periods where only noise is present. Moreover, for voiced speech, at a certain time frame, most of the speech energy is concentrated in the frequency bins corresponding to multiples of the fundamental frequency while the noise energy can be distributed in a wide range of the spectrum. Therefore, development of the speech spectral estimators which take into account the presence/absence of speech spectrum has been considered in the past, e.g., in [78, 79]. The basic idea is to modify the conditional expectation $E\{f(\mathcal{X})|Y\}$ (with $f(\mathcal{X})$ denoting any function of $\mathcal{X}$) encountered in the STSA estimators. In this sense, the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ denoting respectively the absence and presence of speech are defined as

$$\begin{aligned} &\mathcal{H}_0 : Y(k,l) = V(k,l) \\ &\mathcal{H}_1 : Y(k,l) = X(k,l) + V(k,l) \end{aligned} \tag{2.23}$$

Now the conditional expectation $E\{f(\mathcal{X})|Y\}$ can be expressed as

$$E\{f(\mathcal{X})|Y\} = E\{f(\mathcal{X})|Y, \mathcal{H}_1\} \mathcal{P}(\mathcal{H}_1|Y) + E\{f(\mathcal{X})|Y, \mathcal{H}_0\} \mathcal{P}(\mathcal{H}_0|Y) \tag{2.24}$$

where $\mathcal{P}(.)$ denotes the probability. It is obvious that since the aforementioned expressions for the STSA estimators are derived under the assumption of speech presence only, they are in fact equivalent to the term $E\{f(\mathcal{X})|Y, \mathcal{H}_1\}$ in (2.24). Also, it is concluded that $E\{f(\mathcal{X})|Y, \mathcal{H}_0\} = 0$ due to the absence of the speech component under the $\mathcal{H}_0$ hypothesis. In practice, however, for perceptual reasons, a small nonzero value is used for this term in the implementation [79]. To obtain an expression for $\mathcal{P}(\mathcal{H}_1|Y)$, the common approach is to use Bayes' rule which results in [22]

$$\mathcal{P}(\mathcal{H}_1|Y) = \frac{\Lambda}{1+\Lambda} \tag{2.25}$$

where Λ is called the likelihood ratio and is given by

$$\Lambda = \frac{1-q}{q} \frac{p\left(Y|\mathcal{H}_1\right)}{p\left(Y|\mathcal{H}_0\right)} = \frac{1-q}{q} \frac{\exp(\nu)}{1+\frac{\zeta}{1-q}} \tag{2.26}$$

with q as the *a priori* probability of speech absence defined as $\mathcal{P}(H_0)$. To derive the right side of equation (2.26), complex Gaussian distributions are assumed for the noisy observations under $\mathcal{H}_1$ and $\mathcal{H}_0$. It should be noted that both fixed values and adaptively estimated values (as a function of time-frequency unit) have been used for $q = \mathcal{P}(H_0)$ in the literature. Whereas in [17, 18], performance of the STSA estimators has been evaluated using constant experimental values of q, Cohen in [80] suggests to estimate this parameter using a recursive algorithm in time and frequency. The same author has also suggested an efficient modification of the log-MMSE estimator using the concept of SPP in [81] and has called the resulting STSA estimator the optimally modified log-spectral amplitude (OM-LSA). Therein, he proposes to incorporate the SPP into the log-MMSE estimator as the following

$$E\{\log \mathcal{X}|Y\} = E\{\log \mathcal{X}|Y, \mathcal{H}_1\}(1-q) + E\{\log \mathcal{X}|Y, \mathcal{H}_0\}q \tag{2.27}$$

Using the expression for the log-MMSE estimator as in [18], the following is derived

$$\hat{\mathcal{X}} = \exp\left(E\{\log \mathcal{X}|Y\}\right) = \left(\exp\left(E\left\{\log \mathcal{X}|Y, \mathcal{H}_1\right\}\right)\right)^{1-q} \times \left(\exp\left(E\left\{\log \mathcal{X}|Y, \mathcal{H}_0\right\}\right)\right)^{q} \tag{2.28}$$

And expressing the above STSA estimator in terms of gain functions results in the following expression for the OM-LSA estimator

$$G_{\text{OM-LSA}} = \left(G_{\mathcal{H}_1}\right)^{1-q} \left(G_{\mathcal{H}_0}\right)^{q} \tag{2.29}$$

with $G_{\mathcal{H}_1}$ as the gain function of the log-MMSE estimator in Table 2.1 and $G_{\mathcal{H}_0}$ chosen as a fixed minimum gain function, G_{min}, which is set experimentally. The *a priori* probability of speech absence, q, is estimated in [81] in a recursive manner for each time-frequency unit as part of the IMCRA noise PSD estimation method. Combining the IMCRA method for noise estimation and the suggested OM-LSA method in [81], it is shown that excellent noise suppression is achieved while retaining weak speech spectral components and avoiding the musical noise phenomena.

2.2 Speech STSA Priors

The speech STSA estimators discussed in the previous section and the STSA estimators represented in Table 2.1 are all based on using the Rayleigh distribution to model the speech STSA. The latter arises from the fact that speech STFT coefficients are generally assumed to have a complex Gaussian distribution. Recently, however, there have been numerous works directed towards the estimation of speech STSA using super-Gaussian statistical models, especially for the speech STSA. In [82] and references therein, various non-Gaussian distribution models for the speech STSA are discussed, which include exponential, Laplacian, Chi, Gamma (one-sided) and generalized Gamma distributions. These distributions each have unknown parameters and different speech data-based (adaptive) schemes have been proposed for the estimation of their corresponding parameters. According to the experiments in [83, 84], the generalized Gamma distribution (GGD) has the potential to fit the empirical (e.g., histogram-based) distribution of speech STSAs best; however, closed-form solutions for an STSA estimator are available only for specific choices of the parameters of the GGD. In fact, the GGD is a very flexible parametric distribution which covers many super-Gaussian distributions as particular cases. The one-sided GGD family with the shape parameters a and c and the scaling parameter b is given by [37]

$$p_{\text{GGD}}(\mathcal{X}; a, b, c) = \frac{ab^c}{\Gamma(c)} \mathcal{X}^{ac-1} \exp(-b\mathcal{X}^a);\ \mathcal{X} \geq 0,\ a, b, c > 0 \tag{2.30}$$

Note that since this part deals with spectral amplitude estimation, only right-sided distributions are discussed. In fact, the GGD model is a very generalized form of different super-Gaussian distributions and a few useful super-Gaussian distributions in the context of STSA estimation can be derived by considering particular choices of the GGD model, which are summarized in Table 2.2.

Table 2.2: Parameter sets of the GGD leading to Rayleigh, Gamma, Chi, or exponential speech STSA models.

Parameters of the GGD	**STSA Prior**
$a = 2,\ c = 1$	Rayleigh
$a = 1$	Gamma
$a = 2,\ b = 1/2$	Chi
$a = 1,\ c = 1$	Exponential

Figure 2.2 shows GGD values for a few choices of its shaping parameters and $b = 2$. This indicates that by a dynamic selection of these parameters at each STFT frequency bin and time frame, one can gain control over the statistical model of the speech STSA and thus the corresponding gain function of STSA estimators. In a theoretical viewpoint, the estimation of GGD parameters can be done through an ML procedure using the available noisy speech data. However, the exact determination of the GGD parameters independently by solving likelihood equations is cumbersome [84]. In the context of speech STSA estimation, however, closed-form solutions (for ML, MAP or MMSE-based) estimators are available only for the choices of $a = 1$ and $a = 2$. Note that for the choice of $a = 2$, the GGD prior is actually simplified into a generalized form of the Chi distribution with $2c$ degrees of freedom and $1/\sqrt{2b}$ as the scale parameter.

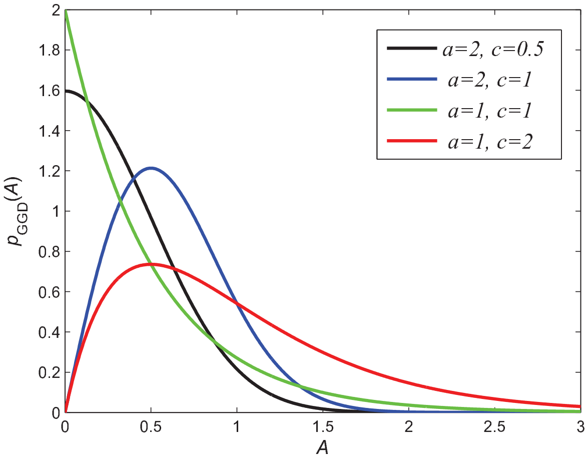


Figure 2.2: One-sided GGD function for different values of the scale parameters and $b = 2$.

Also, the second moment of the GGD prior in (2.30), i.e., the speech STSA variance, is given as the following [37]

$$\sigma_{\mathcal{X}}^2 = \begin{cases} \frac{c(c+1)}{b^2}, & \text{if } a = 1 \\ \frac{c}{b}, & \text{if } a = 2 \end{cases} \tag{2.31}$$

Therefore, having an estimation of the speech STSA spectral variance, $\sigma_{\mathcal{X}}^2$, the scale parameter b will be obtained based on the choice of the shaping parameters. Various combinations of the GGD shape parameters that lead to specific closed-form solutions for speech STSA estimators have been presented in [37]. Therein, solutions have been presented for the case of MMSE-based estimators using Gaussian and exponential speech priors, and MAP estimators using GGD speech priors with $a = 1, 2$. It is concluded that in the case of MMSE-based estimation, higher order shape parameters generally result in numerical analysis since such expressions rely on integrations with no closed-form solution. Also, in the case of MAP estimators, certain combinations of lower order shape parameters can result in monotonic cost functions for which a MAP solution does not actually exist. STSA estimation solutions using special cases of the GGD for noise distribution have also been discussed in [37]; yet, in accordance with the results reported in [22], no improvements have been obtained as compared to using the Gaussian distribution for noise. Table 2.3 summarizes the major solutions of STSA estimation using the GGD prior presented in [37]. Note that in this table, (a_s, c_s) and (a_v, c_v) respectively denote the GGD shape parameters for the clean speech and noise priors. In [85], a group of log-spectral amplitude (LSA) estimators has been proposed, using GGD priors with $a = 1, 2$. Therein, due to providing mathematical flexibility in the statistical STSA modeling, objective improvements with respect to several older STSA estimators including the LSA estimator in [18] have been achieved. Although closed-form solutions are not obtainable for the general case of $a = 1, 2$, estimators were expressed in [85] as limits, and were mathematically approximated. In [86], MMSE-based and MAP estimators of speech STSA have been proposed based on Gamma and Chi priors for speech STSA, and data-driven schemes for the selection of the shape parameter of the priors have been suggested. In that work, rather than relying on *a priori* estimated values of the shape parameter, the focus is on seeking those values that maximize the quality of the enhanced speech, in an *a posteriori* fashion. To this end, the performance of the parameter selection schemes is first evaluated as a function of the shape parameter and then optimal values are found by means of a formal subjective listening test. The main conclusion was that the shape parameters control a trade-off between the level of the residual noise and its musical character. Also, it was found that the optimal parameter values maximizing the subjective performance are different from those maximizing the scores in objective performance measures. It is believed that this discrepancy is mainly due to the poor ability of objective measures to penalize the musical noise artifacts. Another finding of the research in [86] is that very close performance results can be obtained using the same estimator, i.e., MMSE-based or MAP, but with different STSA priors. This can be attributed to the flexibility provided by the shape parameters of the STSA prior, allowing the listener to closely match the performance of two estimators with different speech priors. As further conclusions of this work, the type of the estimators, i.e., MMSE-based or MAP, has significant impact on the quality of the enhanced speech. Whereas MAP estimators result in lower residual noise levels, the MMSE-based estimators are more successful in the restoration of the speech spectral components and are able to achieve higher scores in the objective speech quality measures. Both types of STSA estimators, however, can produce an enhanced speech free of musical noise artifacts, given the correct setting of their parameters.

In [87], a generalized MAP estimator using the Gamma STSA prior along with a data-driven scheme to estimate its shape parameter has been proposed. The shape parameter scheme is based on the fact that a higher estimated SNR corresponds to stronger presence of speech components with respect to noise, and thus, a higher gain value is required for speech segments with higher SNRs. Therefore, since the derived gain function is monotonically decreasing with the Gamma shape parameter, the proposed parameter scheme suggests lower shape parameters for higher SNRs and vice versa. Performance comparisons with other conventional STSA estimators, i.e., the MMSE, ML and MAP methods, confirms that the suggested MAP estimator provides better objective scores in low SNRs while having comparable performance in high SNRs.

Table 2.3: Speech STSA estimators for particular parameter choices of the GGD speech and noise priors

Criterion	(a_v, c_v)	(a_s, c_s)	**Gain function**
ML	$a_v \in \{1, 2\}$, $c_v \geq 1/a_v$	-	$1-\frac{1}{\sqrt{\gamma}}\left(\frac{a_v c_v-1}{a_v\sqrt{c_v(c_v+2-a_v)}}\right)^{1/a_v}$
MMSE	$(2, 1/2)$	$(2, 1/2)$	$\frac{\zeta}{1+\zeta}\left[1-\frac{1}{\gamma}\exp\left(-\frac{\gamma(1+\zeta^2)}{4\zeta(1+\zeta)}\right)\sinh\left(\frac{\gamma(\zeta-1)}{4\zeta}\right)\right]$
		$(1, 1)$	$1-\frac{1}{2\sqrt{\zeta\gamma}}-\frac{1}{\gamma}\left[\exp\left(-\frac{1}{\zeta}-\frac{\gamma}{4}+\sqrt{\frac{\gamma}{2\zeta}}\right)\sinh\left(\frac{\gamma}{4}-\sqrt{\frac{\gamma}{2\zeta}}\right)\right]$
	$(1, 1)$	$(1, c_s \in \mathbb{N})$	$\frac{1}{\mathcal{A}}\left(\frac{(-1)^{c_s}c_s!+\exp(\mathcal{A})\sum_{n=0}^{c_s}\left((-1)^n\frac{c_s!}{(c_s-n)!}\mathcal{A}^{c_s-n}\right)}{(-1)^{c_s-1}(c_s-1)!+\exp(\mathcal{A})\sum_{n=0}^{c_s-1}\left((-1)^n\frac{(c_s-1)!}{(c_s-n-1)!}\mathcal{A}^{c_s-n-1}\right)}\right)$ where $\mathcal{A}=\sqrt{\frac{2\gamma}{\zeta}}\left(\sqrt{\zeta}-\sqrt{\frac{c_s(c_s-1)}{2}}\right)$
MAP	$(2, 1/2)$	$(2, c_s \in \mathbb{R})$	$\frac{1}{2(2c_s+\zeta)}\left(\zeta+\sqrt{\zeta^2+4(\zeta/\gamma)(2c_s-1)(2c_s+\zeta)}\right)$
		$(1, c_s \in \mathbb{R})$	$\frac{1}{2}-\frac{\sqrt{c_s(c_s+1)}}{2\sqrt{\gamma\zeta}}+\sqrt{\left(\frac{1}{2}-\frac{\sqrt{c_s(c_s+1)}}{2\sqrt{\gamma\zeta}}\right)^2+\frac{c_s-1}{\gamma}}$
	$(2, 1)$	$(2, 1/2)$	$\frac{1}{2}\left(\frac{4\zeta+1}{2\zeta+1}\right)+\frac{1}{2}\sqrt{\left(\frac{4\zeta+1}{2\zeta+1}\right)^2-\frac{4\zeta}{\gamma}\left(\frac{2\gamma-1}{2\zeta+1}\right)}$
		$(1, 1)$	$1-\frac{1}{2\sqrt{2\zeta\gamma}}+\sqrt{\left(1-\frac{1}{2\sqrt{2\zeta\gamma}}\right)^2+\left(\frac{1}{\sqrt{2\zeta\gamma}}+\frac{1}{2\gamma}-1\right)}$

The STSA estimators discussed so far incorporated improved statistical models with the original MMSE or log-MMSE cost functions. In [88], the authors make use of the Chi STSA prior to derive estimators using perceptually motivated spectral amplitude cost functions, namely the WE and WCOSH primarily developed in [29]. The major purpose in [88] is to determine the advantage of incorporating improved cost functions with more accurate (i.e., super-Gaussian) STSA priors. Therein, it was shown that whereas the perceptually-motivated cost functions emphasize spectral valleys rather than spectral peaks (formants) and indirectly account for auditory masking effects, the incorporation of the Chi STSA prior demonstrates considerable improvement over the Rayleigh model for the speech prior. Yet, no systematic parameter choice has been proposed for the two WE and WCOSH estimators and the shape parameter of the corresponding Chi STSA prior is selected empirically. Along the same line of work, in [89], the authors take advantage of the β-order MMSE cost function first adopted in [31] with Laplacian priors for the real and imaginary parts of speech STFT coefficients. Even though using a Laplacian model as speech prior primarily results in a highly non-linear estimator with no closed-form solution and high computation costs, by using approximations for the distribution of speech STFT and also for the involved Bessel functions, an improved closed-form version of the estimator has been derived and evaluated in [89]. The comparative evaluations reported therein confirm the superiority of the suggested estimator relative to the state-of-the-art estimators that assume either Gaussian or Laplacian STSA priors such as [90].

Based on the aforementioned works, it can be concluded that even though statistical methods for the estimation of the parameters of super-Gaussian priors exist, e.g., [91], subjectively driven schemes based on speech observations or solid theoretical methods to maximize objective measures such as [86, 87] have been proved to be more efficient in the speech enhancement literature.

2.3 Multi-Channel STSA Estimation

Whereas the single channel speech enhancement methods work reasonably well for most applications, their performance quickly deteriorates under adverse noisy conditions. Moreover, such methods are incapable of providing improvements in the noise reduction without introducing distortion on the clean speech component. In order to achieve higher reduction in the background noise while keeping the speech distortion in the minimum possible level, researchers have developed multi-microphone (multi-channel) methods to exploit all available temporal and spatial information of the speech and noise sources [46]. In general, it can be assumed that in many of the large area noisy environments (e.g., offices, cafeterias and airport terminals), noise propagates simultaneously in all directions and has negligible spatial correlation across the different sensors [92]. In this section, we first state the multi-channel speech enhancement problem in the spatially uncorrelated case and then briefly overview the multi-channel extension of the STSA estimation method by assuming the direction of arrival (DOA) of the captured speech source known.

2.3.1 Multi-Channel Problem Statement

Suppose that we have a uniform linear array (ULA) consisting of N omni-directional sensors each spaced d meters apart capturing a far-field speech source at a known incident angle (DOA) equal to θ, as illustrated in Figure 2.3. This assumption implies no relative attenuation across the microphone signal amplitudes and also a constant time delay due to the planar shaped waveforms [46]. Note that the problem can be simply extended to any arbitrary-shaped microphone array. The set of N microphones captures the noisy observation waveforms $y_n(t)$, consisting of the time delayed clean speech signals $x(t-\tau_n)$ contaminated by additive spatially uncorrelated noises $v_n(t)$, where n is the microphone index and τ_n is the relative time delay of the speech signal in the nth microphone with respect to the reference (first) microphone. In a ULA set of microphones, this time delay is $(n-1)d\cos(\theta)/T$ with θ as the DOA in radian and T is the velocity of sound in the air in meters per second. Based on this notation, we have

$$y_n(t) = x(t-\tau_n) + v_n(t), \quad n = 1, 2, ..., N \tag{2.32}$$

where $x(t)$ is the coherent speech signal under estimation. After sampling, framing and STFT analysis, the noisy speech signal can be represented as

$$Y_n(k,l) = X(k,l)e^{-j\phi_{n,k}} + V_n(k,l), \quad n = 1, 2, ..., N \tag{2.33}$$

Note that the complex exponential $e^{-j\phi_{n,k}}$ is multiplied by the source speech signal component $X(k,l)$ to account for the time delay across different microphone signal observations in the STFT domain.

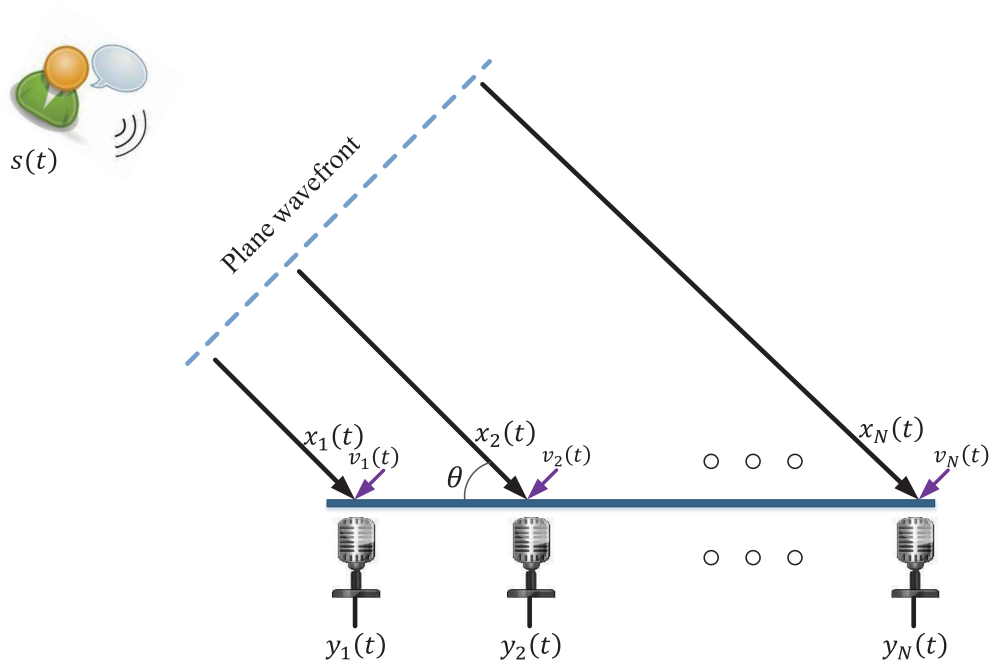


Figure 2.3: An N equispaced linear microphone array capturing the speech source $s(t)$ located in the far field impinging at the incident angle θ.

It is easy to show that the phase difference $\phi_{n,k}$ is obtained as $2\pi f_s \tau_n k/K$ with f_s as the sampling frequency and K the number of total frequency bins. Henceforth, we will drop the frequency bin k and frame index l for sake of brevity. We assume the speech and noise components to be uncorrelated, as it is often implied in a free field (non-reverberant) noisy environment without echoes. Also we consider spatially uncorrelated noise signals across the microphone observations, that is

$$E\{V_n V_m\} = E\{V_n\}E\{V_m\} = 0, \quad \forall\, n, m \in \{1, 2, ..., N\},\; n \neq m \tag{2.34}$$

The speech spectral component X can be written as $\mathcal{X}e^{j\omega}$ with $\mathcal{X} \geq 0$ as the spectral amplitude and $\omega \in [-\pi, \pi]$ the spectral phase. Given the time delay of arrival τ_n (or equivalently $e^{-j\phi_n}$), the STSA estimation aims at the estimation of $\mathcal{X}$ using the set of noisy spectral observations Y_n.

2.3.2 Multi-channel Extension of the Bayesian STSA Estimation

Multi-channel extensions of the STSA estimation method for spatially uncorrelated noise have been reported in the recent literature such as in [40]. In the single channel case, Bayesian STSA estimators are derived based on the conditional expectation, $E\{f(\mathcal{X})|Y\}$ with $f(.)$ some function depending on the underlying Bayesian cost function. In the multi-channel case, however, this statistical expectation should be replaced by $E\{f(\mathcal{X})|\mathbf{Y}\}$ with $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_N]^T$ as the vector of microphone array observations, and this leads to

$$E\{f(\mathcal{X})|\mathbf{Y}\} = \frac{\int_0^\infty \int_0^{2\pi} f(\mathcal{X}) p(\mathbf{Y}|\mathcal{X},\omega) p(\mathcal{X},\omega) d_\omega d_\mathcal{X}}{\int_0^\infty \int_0^{2\pi} p(\mathbf{Y}|\mathcal{X},\omega) p(\mathcal{X},\omega) d_\omega d_\mathcal{X}} \tag{2.35}$$

To obtain a solution for (2.35), joint distribution of the observations conditioned on the speech signal, i.e. $p(\mathbf{Y}|\mathcal{X},\omega)$, is needed. In general, by considering the spatial correlation across the channels, a multi-variate joint distribution is to be used for $p(\mathbf{Y}|\mathcal{X},\omega)$. Yet, inserting such a distribution in (2.35) may not result in mathematically tractable solutions for the corresponding integrals. In the spatially independent (uncorrelated) noise field, however, due to the independence of the observations conditioned on the speech signal, this joint distribution can be obtained as the product of the individual distributions of the observations for each channel, namely,

$$p(\mathbf{Y}|\mathcal{X},\omega) = \prod_{n=1}^{N} p(Y|\mathcal{X},\omega) = \left(\prod_{n=1}^{N} \frac{1}{\pi\sigma_{v_n}^2}\right) \exp\left(-\sum_{n=1}^{N} \frac{|Y_n - \mathcal{X}e^{j\omega}e^{-j\phi_n}|^2}{\sigma_{v_n}^2}\right) \tag{2.36}$$

where the complex Gaussian distribution is considered for the individual distributions of the observations in (2.33). Note that the noise variances $\sigma_{v_n}^2$ need to be estimated for all channels independently. Using Rayleigh distribution for the speech STSA, i.e., the same model as that in (2.16), and substituting (2.36) into (2.35), closed-form solutions can be achieved for the estimate of $\mathcal{X}$ based on the choice of the underlying Bayesian cost function. In the case of an MMSE STSA estimator, observing that $f(\mathcal{X}) = \mathcal{X}$ and using Appendix A in [40], we can obtain

$$\hat{\mathcal{X}}^{(\mathrm{MMSE})} = \Gamma(1.5) \left(\frac{\sigma_\mathcal{X}^2}{1 + \sum_{n=1}^{N} \zeta_n}\right)^{1/2} \mathrm{M}(-0.5, 1; -\nu) \tag{2.37}$$

where $\sigma_\mathcal{X}^2$ is the speech signal variance, $\zeta_n = \frac{\sigma_\mathcal{X}^2}{\sigma_{v_n}^2}$ is the *a priori* SNR for channel n and ν is given as

$$\nu = \lambda \left|\sum_{n=1}^{N} \frac{Y_n e^{j\phi_n}}{\sigma_{v_n}^2}\right|^2 \quad \text{with} \quad \frac{1}{\lambda} = \frac{1}{\sigma_\mathcal{X}^2} + \sum_{n=1}^{N} \frac{1}{\sigma_{v_n}^2} \tag{2.38}$$

This is the same result as that obtained in [40] but with taking into account the phase difference parameter, ϕ_n, to compensate for the relative time delay across the spatial microphone observations. In the special case of $N = 1$, the single channel MMSE estimator in [17] can be degenerated from (2.37). Other multi-channel extensions of the STSA estimation method using more recent Bayesian cost functions, e.g., log-MMSE, β-SA and perceptually motivated ones, have also been reported in [40, 41, 93].

2.4 Reverberation Suppression in the STFT Domain

2.4.1 Reverberation in Enclosed Spaces

As discussed in chapter 1, as part of this research, we focus on the problem of reverberation suppression in the STFT domain. One of the major challenges in speech enhancement stems from the degradation of the speech by an acoustic channel within an enclosed space, e.g., an office, meeting or living room. In the case that the microphones are not located near the desired source, the received microphone signals are degraded by the reverberation introduced by the multi-path propagation of the clean speech to the microphones. While numerous state-of-the-art acoustic signal processing algorithms are available for noise reduction, the development of practical algorithms that can mitigate the degradations caused by reverberation under robust assumptions has been an ongoing challenge in the literature. One major concern about speech enhancement in the presence of reverberation is that the degrading component is correlated with the desired speech, whereas in case of a noise-only environment, noise can be assumed to be independent of the desired speech. For this reason, many existing acoustic signal processing techniques, e.g., noise reduction, source localization, source separation and automatic speech recognition, completely fail or suffer dramatical degradation when reverberation is present [42].

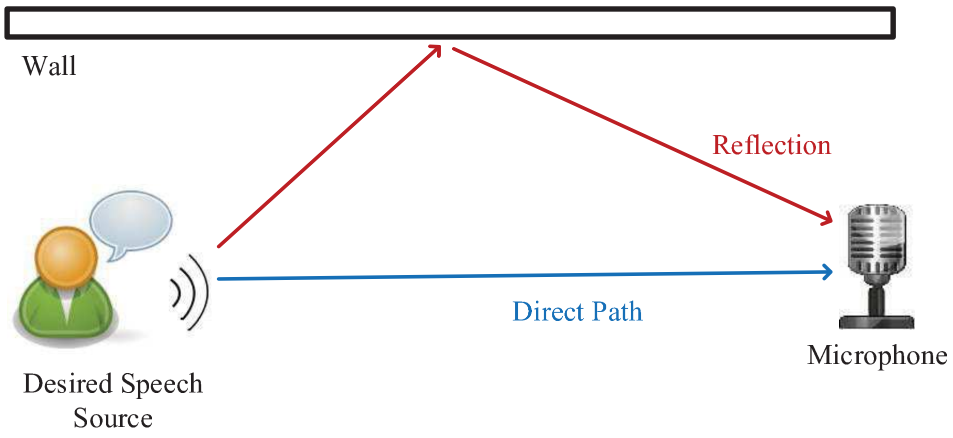


Figure 2.4: Illustration of the direct path and a single reflection from the speech source to the microphone.

Reverberation can be intuitively described by the concept of reflections. The desired source of speech generates wavefronts propagating outward from the source and these wavefronts reflect off the walls of a room or different objects and then superimpose at the microphone. This is illustrated in Figure 2.4 with an example of a direct path and single reflection of the speech wavefront. Due to the difference in the length of the propagation paths and also the amount of sound energy absorbed or reflected by the wall, each wavefront arrives at the microphone with a different amplitude and phase, and therefore, reverberation refers to the presence of delayed and attenuated replicas of the speech source in the received signal. This received signal consists of a direct path signal, reflections arriving shortly after the direct sound, i.e., the early reverberation, and reflections arriving after the early reverberation, i.e., the late reverberation. The combination (sum) of the direct path signal and early reverberation is referred to as the early component of speech and is often of interest in reverberation suppression methods [43].

A simple schematic of the model we consider for the received microphone signals in the presence of room reverberation is shown in Figure 2.5. In this figure, the block named as acoustic channel(s) actually models the channel between the source of speech and the microphone(s). As seen in this figure, the clean speech signal generated by the speech source passes through the acoustic channel(s) and is contaminated by additive noise from the surrounding environment. The resulting speech signal is captured next by the microphone(s). It should be noted that, given the aim is to perform reverberation suppression blindly, the acoustic channels as well as the environment are considered unknown and no prior information is assumed about them.

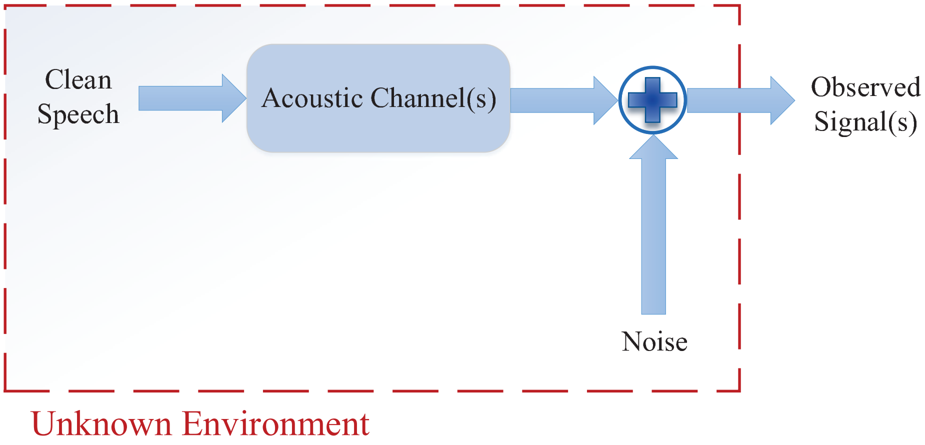


Figure 2.5: Modeling of the observed microphone signal(s) in a reverberant environment.

2.4.2 Problem Formulation

Reverberation is basically the process of multi-path propagation of a speech signal $x(m)$ from its source to one or multiple microphones, with m denoting the discrete time index. Based on the general model in Figure 2.5, the observed signal at the nth microphone, $y_n(m)$, is the sum of noise, $v_n(m)$, and the convolution of the clean speech with the impulse response(s) of the acoustic channel(s), as the following [42]

$$y_n(m) = x(m) * h_n(m) + v_n(m) = \sum_{\ell=0}^{L_h-1} h_n(\ell)x(m-\ell) + v_n(m) \tag{2.39}$$

where $h_n(m)$ is the so-called room impulse response (RIR) for the nth channel with the length of L_h, and $*$ denotes the convolution operation. The ultimate aim of dereverberation is to obtain an estimate of the clean speech, $\hat{x}(m)$ using the set of observations $y_n(m)$, $n \in \{1, 2, \cdots, N\}$. We consider this to be a blind problem, as neither the speech signal $x(m)$ nor the acoustic RIRs $h_n(m)$ are available. It should be noted that typical acoustic impulse responses consist of several thousand coefficients, making the estimation of the RIR too difficult in practice (furthermore, the RIR can be time-varying in some scenarios). Therefore, in this work, we restrict our attention to completely blind reverberation suppression techniques in which there exists no requirement for the estimation of RIRs.

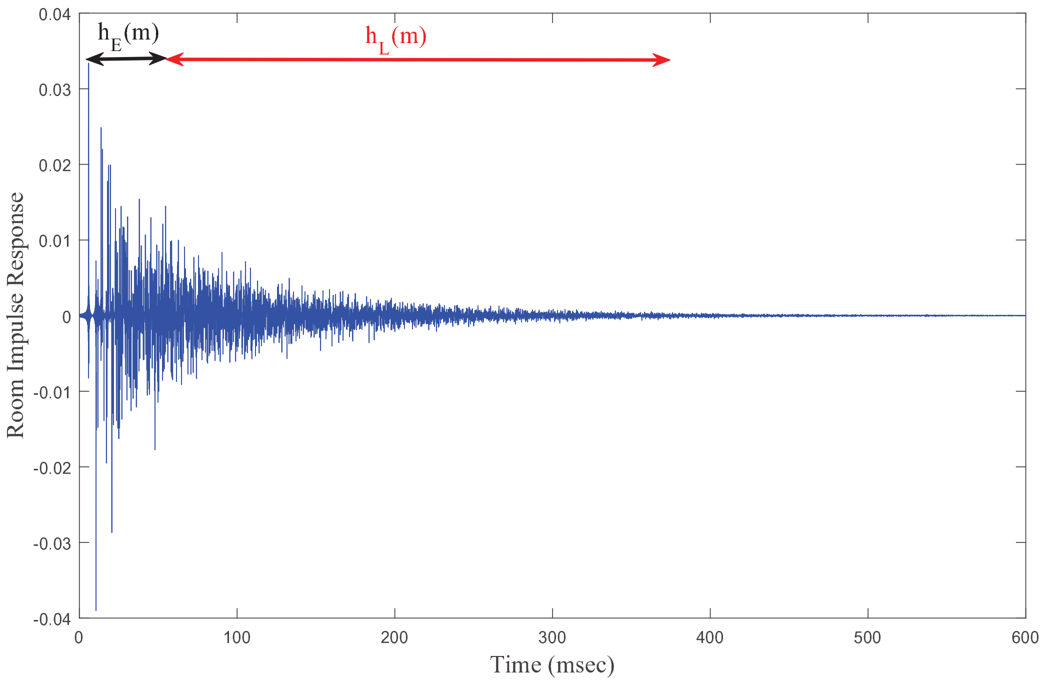


Figure 2.6: Plot of a typical acoustic impulse response with illustrated early and late parts of the RIR.

As the distance between the speech source and microphones increases, the direct-to-reverberant ratio (DRR), i.e., the counterpart of the signal-to-noise ratio (SNR) in reverberant environments, decreases. In this sense, the reverberation becomes dominant and the quality and intelligibility of speech are deteriorated. Thus, reverberation suppression becomes necessary for a speech communication system. However, it is well accepted that the first few reflections of the direct path of the speech do not degrade the speech quality/intelligibility as perceived by human ear [42, 43]. In fact, these first reflections, since often being so similar to the direct path speech, may even help improving the intelligibility (i.e., the human's ability to perceive the speech) and also the SNR in noisy reverberant fields. Thus, the focus of most reverberation suppression techniques is to reduce the effect of the later reflections of speech, leaving the primary reflections as they are [43]. In this work also, we do not intend to dereverberate the speech signal completely and aim at the estimation of the primary reflections of the direct path speech. Based on this fact, the entire reverberation, or equivalently the RIR, can be split into two parts: the early and the late components, as the following

$$h_n(m) = \begin{cases} h_{n_E}(m), & 0 \leq m < L_E \\ h_{n_L}(m), & L_E \leq m < L_h \\ 0, & \text{Otherwise} \end{cases} \tag{2.40}$$

with $h_{n_E}(m)$ and $h_{n_L}(m)$ denoting the early and late component of the RIR for the nth microphone, and L_E is the length of the early component. In practice, the latter is often selected as $f_s T_{early}$ with f_s as the sampling frequency and T_{early} ranging from 40 ms to 80 ms [42]. A measured typical RIR with its early and late components have been shown in Figure 2.6.

Now, inserting the model in (2.40) into (2.39) results in

$$y_n(m) = \underbrace{\sum_{\ell=0}^{L_E-1} h_{n_E}(\ell)x(m-\ell)}_{y_{n_E}(m)} + \underbrace{\sum_{\ell=L_E}^{L_h-1} h_{n_L}(\ell)x(m-\ell)}_{y_{n_L}(m)} + v_n(m) \tag{2.41}$$

with $y_{n_E}(m)$ and $y_{n_L}(m)$ respectively as the early and late reverberant components of the observations $y_n(m)$. In the same fashion as the noise-only environments, by expressing (2.41) in the STFT domain, it follows that

$$Y_n(k,l) = Y_{n_E}(k,l) + Y_{n_L}(k,l) + V_n(k,l) \tag{2.42}$$

In summary, our aim of reverberation suppression in the STFT domain is to obtain an estimate of the early reverberant component, $Y_{n_E}(k,l)$, by reducing the late reverberation, $Y_{n_L}(k,l)$, and the possibly existent noise $V_n(k,l)$. In this regard, we tend to resort to techniques where no *a priori* information of the RIR or environment or its estimates are needed.

2.5 Shortcomings of the State-of-the-Art STSA Estimation Methods

Based on the content of this chapter, we herein summarize some of the shortcomings of the state-of-the-art STSA estimators, which have motivated us to develop some of the contributions presented in the following two chapters. These shortcomings can be categorized as the following:

1. In Section 2.1.6, different Bayesian cost functions exploited to derive the MMSE-based estimators were discussed. Another major factor to derive such estimators is the type of distribution used to model the speech STSA, i.e., the speech prior, which was discussed in Section 2.2. Even though there has been considerable contribution on both of these topics in the state-of-the-art literature, the best possible combination of the underlying Bayesian cost function and speech prior is still not known. Also, despite the existence of a few empirical and/or intuitive schemes for the parameter selection of these estimators, a widely accepted parameter selection scheme resulting in the best experimental results for a general noise scenario does not exist.

2. Considering the gain function of the state-of-the-art Bayesian STSA estimators, e.g., those in Section 2.1.6, it is evident that only the information in the amplitude of the speech is exploited in the estimator's gain function, while the phase information is disregarded. In fact, in the derivation of these estimators, it is assumed that speech phase is totally random and is uniformly distributed over $[0, 2\pi)$. Yet, the observed noisy phase or any estimate of the speech phase can be employed in order to provide more information about the clean speech signal in the derivation of the STSA estimators. This is in addition to the fact that, contrary to the conventional speech literature, recently, the speech phase has been found to be useful in reconstructing the clean speech signal from noisy observations, e.g., [27].

3. In the sense of multi-channel STSA estimation in spatially uncorrelated noise, as discussed in Section 2.3, estimators based on a few Bayesian cost functions have been proposed, e.g., in [40, 41, 93]. However, similar to the single-channel, there is still need to develop a generalized form of the multi-channel STSA estimator using a combination of the most efficient Bayesian cost function and speech prior. Also, regarding the multi-channel STSA estimation in spatially correlated noise, since following (2.35) directly does not lead to a closed-form solution for the estimator's gain function, development of a systematic way to obtain a general solution for this multi-channel problem is required.

4. All considered STSA estimators in this chapter are based on the assumption of a free-field (non-reverberant) environment. However, in practical scenarios, the speech source is located in an acoustic room where the clean speech is convolved with the unknown impulse response of the room prior to reception by microphones. The phenomenon, named as reverberation,

not only distorts the quality of the captured speech but also deteriorates highly the noise reduction performance of the STSA estimators designed for noise-only environments [42]. Even though modified spectral enhancement methods (those developed initially for noise reduction) for reverberation suppression have received attention in the past, e.g. in [43], the performance of such modified noise reduction methods is still far from being satisfactory. Therefore, further research in this direction is required.

Chapter 3

Single-Channel Noise Reduction Using Bayesian STSA Estimation

3.1 Introduction

In this chapter, we develop a method for single-channel noise reduction using Bayesian STSA estimation. This chapter is organized as follows. Section 3.2 gives a brief review of the considered STSA estimator in this chapter, i.e., the Wβ-SA estimator. Section 3.3 describes the proposed speech STSA estimator, including the proposed schemes for the parameter selection of the Bayesian cost function as well as a new gain flooring scheme for STSA estimators. In Section 3.4, we exploit the generalized Gamma distribution (GGD) to model the speech prior for the proposed STSA estimator and suggest an efficient method for the estimation of its parameters. Performance of the proposed STSA estimation method is evaluated in Section 3.5 in terms of objective performance measures. Conclusions are drawn in Section 3.6.

3.2 Previous Work

In this section, a brief overview of a generic STSA estimation method, namely the Wβ-SA estimator, is presented. This estimator will be used as a basis for further developments in this chapter. As stated in Chapter 2, the STFT domain representation of the noisy speech can be expressed as

$$Y(k,l) = X(k,l) + V(k,l) \tag{3.1}$$

where $Y(k,l)$, $X(k,l)$ and $V(k,l)$ are the STFTs of the noisy observation, clean speech and noise, respectively, with $k \in \{0, 1, \ldots, K-1\}$ and $l \in \mathbb{N}$ denoting the frequency bin and time frame indices, respectively. Expressing the complex-valued speech coefficients, $X(k,l)$, as $\mathcal{X}(k,l)e^{j\Omega(k,l)}$ with $\mathcal{X}$ and Ω as the amplitude and phase in respect, the purpose of speech STSA estimation is to estimate the speech amplitude, $\mathcal{X}(k,l)$, given the noisy observations, $Y(k,l)$. The estimated amplitude will then be combined with the noisy phase of $Y(k,l)$ to provide an estimate of the speech Fourier coefficients. For sake of brevity, we may discard the indices k and l in the following.

As discussed in Chapter 2, the weighted version of the β-SA, i.e., the Wβ-SA estimator, is known to be advantageous with respect to the other Bayesian estimators. In fact, previously proposed Bayesian cost functions can be expressed as a special case of the underlying Wβ-SA cost function, which is defined as [32]

$$C(\mathcal{X}, \hat{\mathcal{X}}) = \mathcal{X}^{\alpha} \left(\mathcal{X}^{\beta} - \hat{\mathcal{X}}^{\beta}\right)^2 \tag{3.2}$$

with α and β as the corresponding cost function parameters. Note that, for notational ease, we have used α as the exponent of $\mathcal{X}$ rather than -2α as in [32]. Minimizing the expectation of the cost function in (3.2) results in [32]

$$\hat{\mathcal{X}}^{(\text{W}\beta\text{-SA})} = \left(\frac{E\{\mathcal{X}^{\beta+\alpha}|Y\}}{E\{\mathcal{X}^{\alpha}|Y\}}\right)^{1/\beta} \tag{3.3}$$

Now, solving for the moments in (3.3) in a Bayesian framework results in the following gain function [32]

$$G^{(\text{W}\beta\text{-SA})} \triangleq \frac{\hat{\mathcal{X}}^{(\text{W}\beta\text{-SA})}}{|Y|} = \frac{\sqrt{\nu}}{\gamma}\left(\frac{\Gamma\left(\frac{\alpha+\beta}{2}+1\right)\mathrm{M}\left(-\frac{\alpha+\beta}{2}, 1; -\nu\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)\mathrm{M}\left(-\frac{\alpha}{2}, 1; -\nu\right)}\right)^{1/\beta} \tag{3.4}$$

where $\Gamma(.)$ and $\mathrm{M}(., .; .)$ denote the Gamma and confluent hypergeometric functions [77], respectively, and the gain parameters γ and ν are defined as

$$\gamma = \frac{|Y|^2}{\sigma_v^2}\ , \quad \nu = \frac{\zeta}{1+\zeta}\gamma\ , \quad \zeta = \frac{\sigma_{\mathcal{X}}^2}{\sigma_v^2} \tag{3.5}$$

where ζ and γ are called the *a priori* and *a posteriori* SNRs, respectively. Figure 3.1 shows theoretical gain curves of the estimator in (3.4) for different values of the parameters α and β.

Herein, the fixed values ζ=0dB and γ=0dB are considered to account for a highly noisy scenario. It is observed that the STSA gain function is mainly controlled by two parameters, and in particular, an increment in either of the two, especially α, would result in an increment in the gain function values. This realization will be used in the following sections to propose new schemes for the choice of these parameters.

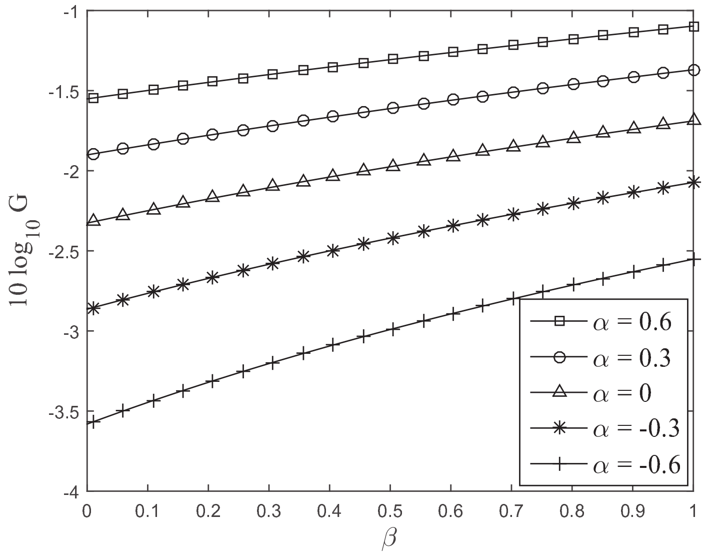


Figure 3.1: STSA gain function curves in (3.4) versus β for different values of α (ζ=0 dB and γ= 0 dB).

Plourde and Champagne in [32] suggested to select the two estimator parameters as functions of frequency, according to the psycho-acoustical properties of the human auditory system and showed a better quality in the enhanced speech in most of the input SNR range. Yet, at high input SNRs, the performance of the developed estimator is not appealing due to the undesired distortion in the enhanced speech. This motivates us to develop more appropriate schemes for the selection of the parameters α and β. Furthermore, the Wβ-SA estimator in (3.4) has been derived under the assumption of a Rayleigh distribution (i.e., the most basic distribution) for the speech STSA $\mathcal{X}$ and has not taken advantage of the category of GGD priors for $\mathcal{X}$, which were discussed in Chapter 2. Therefore, we will also explore the use of GGD speech priors for the Wβ-SA estimator as part of this chapter.

3.3 Proposed Speech STSA Estimator

In this section, we discuss the proposed schemes for the estimation of the Wβ-SA parameters as well as the suggested gain flooring scheme for the gain function of an STSA estimator. The presented contributions in this chapter have been published in [94].

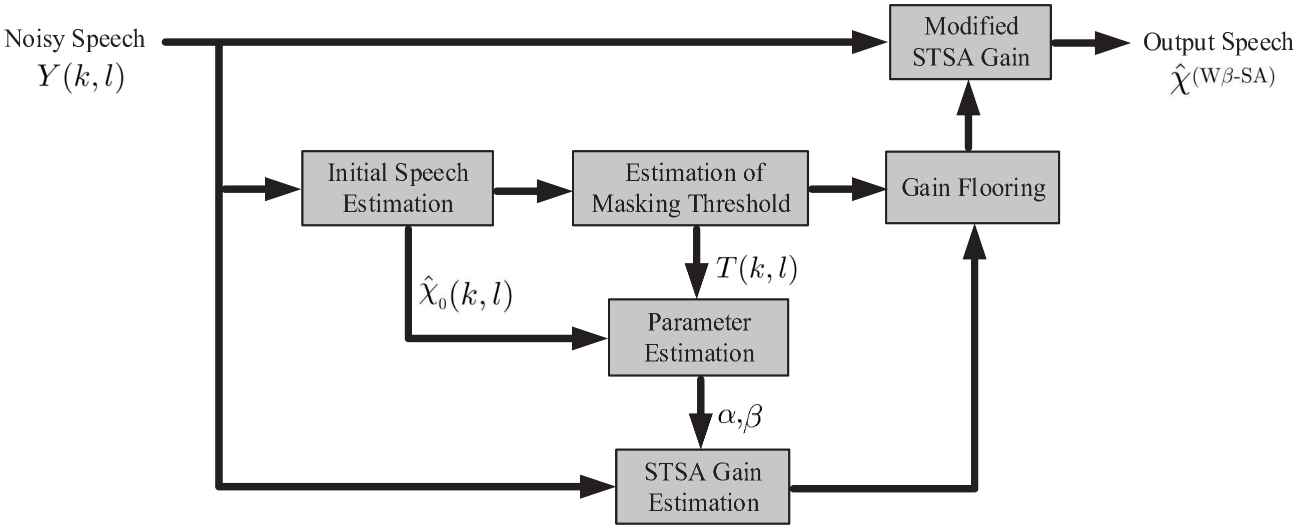
3.3.1 Brief Description of the Proposed Method

Figure 3.2 shows a block diagram of the proposed STSA estimator. An initial estimate of the speech STSA is first obtained to calculate the noise masking threshold and the estimator parameters. This preliminary estimate can be obtained through a basic STSA estimator, e.g., the MMSE estimator in [17], as only a rough estimate of the speech STSA is needed at this step. As the experiments revealed, use of more accurate estimates of the speech STSA, either in the calculation of the noise masking threshold or in the parameters of the STSA estimator, do not result in any considerable improvements in the performance of the entire algorithm. Next, the STSA estimator parameters, α and β, are estimated using both the noise masking threshold and the available initial estimate of the speech STSA. These two parameters along with the noisy speech are fed into the STSA gain calculation block. Note that noise-related parameters, i.e., the noise spectral variance and the *a priori* SNR, should be estimated within this block in order to achieve the gain function value. This gain function is further thresholded and modified by the proposed gain flooring scheme. This modified gain is the ultimate gain function being applied on the STSA of the noisy speech and leading to the enhanced STSA in the output. The enhanced STSA is to be combined with the phase of the noisy speech to generate the STFT of the enhanced speech.

3.3.2 Parameter Selection of the New Wβ-SA Estimator

Selection of parameter α:

In the original Wβ-SA estimator [32], the parameter α was selected as an increasing piecewise-linear function of frequency, in order to increase the contribution of high-frequency components of the speech STSA in the Bayesian cost function. This is because these frequencies often include small speech STSAs that can be easily masked by stronger noise components. However, increasing the values of this parameter monotonically with the frequency without considering the estimated speech STSA values results in over-amplification of high-frequency components, and therefore, a large amount of distortion may appear in the enhanced speech. We here employ the available initial estimate for the speech STSA, denoted by $\hat{\mathcal{X}}_0(k,l)$ (the one used to calculate the noise masking threshold), to propose a new scheme for the selection of α.



Noisy Speech
$Y(k,l)$
Modified STSA Gain
Output Speech
$\hat{\mathcal{X}}^{(\mathrm{W}\beta\text{-SA})}$
Initial Speech Estimation
Estimation of Masking Threshold
Gain Flooring
$T(k,l)$
$\hat{\mathcal{X}}_0(k,l)$
Parameter Estimation
α,β
STSA Gain Estimation

Figure 3.2: Block diagram of the proposed speech STSA estimation algorithm.

Specifically, we propose to select α according to the following scheme

$$\alpha(k,l) = \begin{cases} c_\alpha \, \frac{\hat{\mathcal{X}}_0(k,l)}{\hat{\mathcal{X}}_{0,\max}(l)}, & \text{if } \hat{\mathcal{X}}_0(k,l) \geq \frac{\hat{\mathcal{X}}_{0,\max}(l)}{4} \\ 0 \, , & \text{otherwise} \end{cases} \tag{3.6}$$

where $\hat{\mathcal{X}}_{0,\max}(l)$ is the maximum value of the initial STSA estimate over the frequency bins at frame l and c_α, which determines the maximum value taken by α, is experimentally fixed at 0.55 to avoid excessively large α values. The major reasoning for the proposed frequency-based selection of the parameter α is to emphasize the weighting term $\mathcal{X}^\alpha$ in (3.6) for larger speech spectral components, while avoiding the use of such weighting for smaller components within each frame. This further helps to distinguish the speech STSA components from the noise components of the same frequency at each frame. In fact, if the speech STSA, $\hat{\mathcal{X}}_0(k,l)$, falls above the threshold $\hat{\mathcal{X}}_{0,\max}(l)/4$, increasing α results in the magnification of the weight $\mathcal{X}^\alpha$ in (3.2), provided that the speech STSA, $\mathcal{X}$, is greater than unity. In contrast, for the speech STSA values smaller than the threshold, α is simply set to zero implying no further emphasis on the speech STSA component. In this case, the Wβ-SA estimator actually turns into the β-SA estimator in [31]. It should be noted that the threshold $\hat{\mathcal{X}}_{0,\max}(l)/4$ was selected as a means to compare the relative intensity of the speech STSA components within the same frame. Also, the normalization with respect to $\hat{\mathcal{X}}_{0,\max}$ ensures that the resulting value of α will not be increased excessively in frames where $\hat{\mathcal{X}}_0(k,l)$ takes on large values. Note that the magnification of strong speech components through the suggested selection of α can also be justified by considering Figure 3.1, where an increment in α results in the increment of the gain function value, and in turn, amplifying the speech components. In Figure 3.3, the choice of the parameter α versus lower frequency bins for a noisy speech frame along with the corresponding initial estimate of the speech STSA have been illustrated. In Section 3.5, it will be shown that the undesirable distortion resulting from the original selection of α as in [32] is compensated for by using the proposed scheme.

Selection of parameter β:

The adaptive selection of parameter β was primarily suggested in [31] as a linear function of frame SNR. Later in [95], it was suggested to choose this parameter as a linear function of both the frame SNR and noise masking threshold. This masking threshold is often used to model the masking properties of the human auditory system and is defined as the threshold value below which the noise is not sensible to the human ear due to the presence of speech signals [57]. The following expression is used to estimate the parameter β in [95]

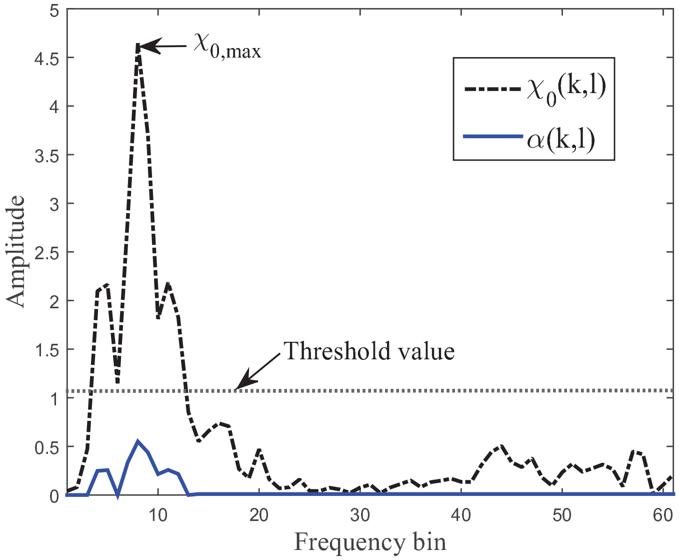


Figure 3.3: Variation of the proposed choice of α versus frequency bins, compared to that of the initial speech STSA estimate for a frame of noisy speech.

$$\beta^{(1)}(k,l) = d_0 + d_1\text{SNR}(l) + d_2 T(k,l) + d_3 \max\{\text{SNR}(l) - d_4, 0\} T(k,l) \tag{3.7}$$

where SNR(l) is the frame SNR in dB, $T(k,l)$ is the normalized noise masking threshold [57] and d_i's are empirically fixed values. $T(k,l)$ represents the threshold below which the human auditory system cannot recognize the noise component and its calculation, which requires an initial estimate of the speech STSA, say $\hat{\mathcal{X}}_0(k,l)$, involves a multiple-step algorithm, as detailed in [95]. The motivation for the choice of β in (3.7) is to increase the gain function values in frames/frequencies with higher frame SNRs or noise masking thresholds, given that the β-SA gain function [31] is a monotonically increasing function of β. The corresponding observations $Y(k,l)$ are dominated by strong speech components and it is hence desirable to employ a larger gain value in the enhancement process. In [32], however, from a psycho-acoustical point of view, it was suggested to choose β based on the compression rate between the sound intensity and perceptual loudness in the human ear. The suggested β therein takes the following form

$$\beta^{(2)}(k) = \frac{\log_{10}(g_1 k + g_2)}{\log_{10}\left(g_1 \frac{K}{2} + g_2\right)} (\beta_{max} - \beta_{min}) + \beta_{min} \tag{3.8}$$

where K is the number of STFT frequency bins, g_1 and g_2 are two constants depending on the physiology of the human ear [96] and β_{max} and β_{min} are set to 1 and 0.2, respectively. However, since β is chosen only as a function of the frequency, it is not adapted to the noisy speech. Furthermore, as experiments show, there may appear excessive distortion in the enhanced speech using the STSA estimator with this parameter choice, especially at high SNRs. Hence, we propose to use the adaptive approach in (3.7) as the basis for the selection of β, but to further apply the scheme in (3.8) as a form of frequency weighting to take into account the psycho-acoustics of the human auditory system within each frame. Specifically, the following approach is proposed for the selection of β:

$$\beta(k,l) = C_\beta \; \beta^{(1)}(k,l) \; \beta^{(2)}(k) \tag{3.9}$$

where the purpose of the constant C_β=1/0.6 is to scale up to one the median value of the frequency weighting parameter $\beta^{(2)}(k)$ in (3.8).

3.3.3 Gain Flooring Scheme

In frequency bins characterized by weak speech components, the gain function of STSA estimators often approaches very small, near zero values, implying too much attenuation on the speech signal. To avoid the resulting speech distortion, various flooring schemes have been applied on the gain function values in these estimators. In [95], it is suggested to make use of the noise masking threshold in the spectral flooring scheme by employing a modification of the generalized spectral subtraction method in [57], namely,

$$G_M(k,l) = \begin{cases} G(k,l), \quad \text{if } \gamma(k,l) > \rho_1(k,l) \\ \sqrt{\frac{\rho_2(k,l)}{\gamma(k,l)}}, \quad \text{otherwise} \end{cases} \tag{3.10}$$

where $G(k,l)$ and $G_M(k,l)$ are the original and modified (thresholded) gain functions, respectively, and $\rho_1(k,l)$ and $\rho_2(k,l)$ are given by [95]

$$\begin{aligned} \rho_1(k,l) &= 5.28 \, \frac{T(k,l) - T_{min}(l)}{T_{max}(l) - T_{min}(l)} + 1 \\ \rho_2(k,l) &= 0.015 \, \frac{T(k,l) - T_{min}(l)}{T_{max}(l) - T_{min}(l)} \end{aligned} \tag{3.11}$$

with $T_{min}(l)$ and $T_{max}(l)$ denoting the minimum and maximum of $T(k,l)$ at the lth frame. The *a posteriori* SNR, $\gamma(k,l)$, is used in the top branch of (3.10) as an indicator of the speech signal intensity while the term $\sqrt{\frac{\rho_2(k,l)}{\gamma(k,l)}}$ in the bottom branch determines the thresholded value of the gain function. Still, (3.10) is characterized by a number of limitations. As originally proposed by Cohen in [97], the gain function itself is a more relevant indicator of speech signal intensity and is therefore more appropriate for use in the thresholding test than $\gamma(k,l)$. Another problem with (3.10) is that the thresholded value may increase uncontrollably at very low values of $\gamma(k,l)$. Rather than relying on $\gamma(k,l)$, it was suggested in [98] to make use of the estimated speech STSA in the thresholded value, i.e.,

$$G'_M(k,l) = \begin{cases} G(k,l), \quad \text{if } G(k,l) > \mu_0 \\ \frac{1}{2} \frac{\mu_0 |Y(k,l)| + \hat{\mathcal{X}}(k,l-1)}{|Y(k,l)|}, \quad \text{otherwise} \end{cases} \tag{3.12}$$

where μ_0 is a fixed threshold taken between 0.05 and 0.22. Our experimentations, however, provided different proper values for μ_0 in various noise scenarios and input SNRs. Hence, considering the wide range of values for the gain function and also the variations in speech STSA, it is appropriate for the threshold μ_0 to be selected as a function of the frame and frequency bin. Herein, by employing the adaptive threshold $\rho_1(k, l)$ in (3.11) and using a variable recursive smoothing for the thresholded value, we propose the following alternative flooring scheme

$$G''_M(k,l) = \begin{cases} G(k,l), & \text{if } G(k,l) > \rho_1(k,l) \\ \frac{p(k,l)\hat{\mathcal{X}}_0(k,l)+[1-p(k,l)]\hat{\mathcal{X}}(k,l-1)}{|Y(k,l)|}, & \text{otherwise} \end{cases} \tag{3.13}$$

where $p(k, l)$ is the speech presence probability which can be estimated through a soft-decision noise PSD estimation method. Using the popular improved minima controlled recursive averaging (IMCRA) in [80] provides enough precision for the estimation of this parameter in the proposed gain flooring scheme. According to (3.13), for higher speech presence probabilities or equivalently in frames/frequencies with stronger speech components, the contribution of the current frame in the recursive smoothing through the term $\hat{\mathcal{X}}_0(k, l)$ will be larger than that of the previous frame $\hat{\mathcal{X}}(k, l-1)$. Conversely, in case of a weak speech component in the current frame, the smoothing gives more weight to the previous frame. Hence, this choice of the flooring value favors the speech component over the noise component in heavily noisy conditions where the gain function is mainly determined by the second branch in (3.13).

3.4 Extension of Wβ-SA Estimator Using GGD Prior

As discussed in Chapter 2, use of the parametric GGD model as the STSA prior, due to providing further flexibility in the resulting gain function, is advantageous compared to the conventional Rayleigh prior. In this section, we first derive an extended Wβ-SA estimator under the GGD speech prior and then propose an efficient method to estimate its corresponding parameters.

3.4.1 Wβ-SA Estimator with GGD Prior

The GGD model can be expressed as

$$p(\mathcal{X}) = \frac{ab^c}{\Gamma(c)}\mathcal{X}^{ac-1}\exp(-b\mathcal{X}^a);\ \mathcal{X} \geq 0, \quad a, b, c > 0 \tag{3.14}$$

with a and c as the shape parameters and b as the scaling parameter [37]. To obtain a solution to the Wβ-SA estimator as in (3.3), we consider the moment term $E\{\mathcal{X}^m|Y\}$ based on the above PDF for the speech STSA. In view of the comprehensive experimental results in [22, 23] for different values of a and in order to arrive at a closed-form solution in the Bayesian sense, we choose a =2 in our work. Then, the GGD prior is simplified into a generalized form of the Chi distribution with $2c$ degrees of freedom and $1/\sqrt{2b}$ as the scale parameter [99]. Based on the second moment of the derived Chi distribution, it can be deduced that the two parameters b and c satisfy the relation $c/b = \sigma_{\mathcal{X}}^2$ [100]. Therefore, the scale parameter b has to be chosen as $c/\sigma_{\mathcal{X}}^2$, given an estimate of the speech STSA variance, $\sigma_{\mathcal{X}}^2$, and the shape parameter c. Using an estimate of the noise variance, σ_v^2, and the *a priori* SNR, ζ, we can obtain an estimate of the speech STSA variance as $\sigma_{\mathcal{X}}^2 = \zeta\sigma_v^2$. The selection of the shape parameter c will be discussed in the next subsection. Taking this into consideration, the following expression for the STSA moment can be derived (see Appendix A for details):

$$E\{\mathcal{X}^m|Y\} = \frac{\Gamma\left(\frac{m+2c}{2}\right)\mathrm{M}\left(\frac{2-m-2c}{2}, 1; -\nu'\right)}{\Gamma(c)\lambda^{m/2}\mathrm{M}\left(1-c, 1; -\nu'\right)} \tag{3.15}$$

where

$$\lambda = \frac{c}{\sigma_{\mathcal{X}}^2} + \frac{1}{\sigma_v^2}\ , \quad \nu' = \frac{\zeta}{c+\zeta}\gamma \tag{3.16}$$

Now, by using (3.15) into (3.3) we can derive

$$G^{(\mathrm{MW}\beta\text{-SA})} = \frac{\sqrt{\nu'}}{\gamma}\left(\frac{\Gamma\left(\frac{\alpha+\beta+2c}{2}\right)\mathrm{M}\left(\frac{2-\alpha-\beta-2c}{2}, 1; -\nu'\right)}{\Gamma\left(\frac{\alpha}{2}+c\right)\mathrm{M}\left(\frac{2-\alpha-2c}{2}, 1; -\nu'\right)}\right)^{1/\beta} \tag{3.17}$$

where the notation MWβ-SA is used to denote the modified Wβ-SA estimator. It is obvious that, when c=1 which corresponds to the Rayleigh prior as a special case, (3.17) degenerates to the original Wβ-SA. In the following, we present a simple approach for the selection of the GGD parameter c for the proposed STSA estimator.

3.4.2 Estimation of GGD Prior Parameters

In [22, 23], an experimentally fixed value in the range of [0,2] has been used for the GGD shape parameter c in different noisy scenarios. Rather than using experimental values, we here take advantage of the behavior of the proposed gain function in (3.17) with respect to the shape parameter c and propose an adaptive scheme for the determination of this parameter. Figure 3.4 depicts curves of the proposed gain function in (3.17) versus the shape parameter c for different *a posteriori* SNRs. As observed, increasing the shape parameter leads to a monotonic increase of the gain function for all considered values of SNR. Note that for stronger speech STSA components (or equivalently weaker noise components) a larger gain function value is desirable in general.

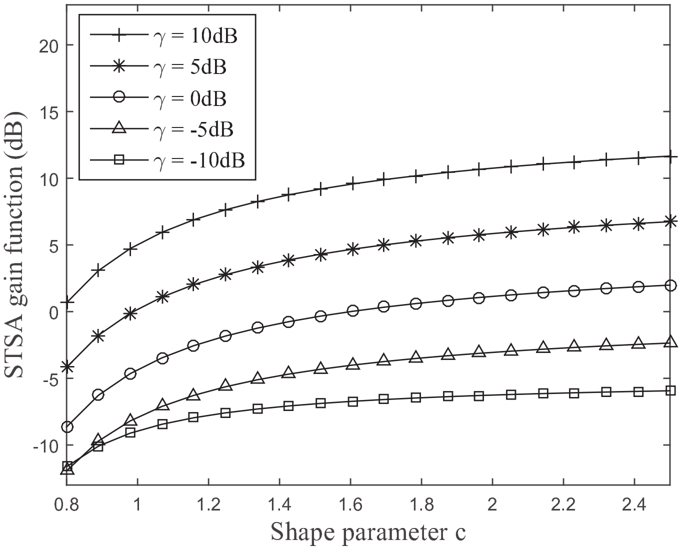


Figure 3.4: Gain function of the modified Wβ-SA estimator in (3.17) versus the GGD shape parameter c for different values of γ (ζ=-5dB).

Therefore, we suggest to choose the shape parameter as a linear function of the SNR values at each frame, namely,

$$c(l) = c_{min} + (c_{max} - c_{min})\, \zeta_{norm}(l) \tag{3.18}$$

where, based on the comprehensive experimentations in [37], c_{min} and c_{max} are chosen as 1 and 3, respectively and $0 < \zeta_{norm}(l) < 1$ is the normalized *a priori* SNR, i.e.,

$$\zeta_{norm}(l) = \frac{\zeta_{av}(l) - \zeta_{min}(l)}{\zeta_{max}(l) - \zeta_{min}(l)} \tag{3.19}$$

with $\zeta_{av}(l)$ as the *a priori* SNR being averaged over the frequency bins of the lth frame, and $\zeta_{min}(l)$ and $\zeta_{max}(l)$ as the minimum and maximum of the *a priori* SNR at the same frame, respectively. According to (3.18), the shape parameter c takes on its values as a linearly increasing function of the SNR in its possible range between c_{min} and c_{max}, leading to the appropriate adjustment of the estimator gain function based on the average power of the speech STSA components at each frame.

3.5 Performance Evaluation

In this section, we first introduce some of the most important objective measures used for the assessment of speech quality in noise reduction methods. Next, we evaluate in detail the performance of the proposed single-channel STSA estimation method using the described performance measures.

3.5.1 Performance Measures for Noise Reduction

Even though subjective evaluation of speech enhancement algorithms, i.e., the evaluation through listening tests, is often accurate and promising, it is often costly and time consuming. For this reason, much effort has been made on the development of objective measures assessing speech quality with high correlation to the subjective methods. In [101], a comprehensive study has been performed to assess the correlation of the existing objective measures with the introduced distortions in the speech by the underlying enhancement method and the overall quality of the noise-suppressed speech. Furthermore, based on the accomplished analysis, accurate adjustment and fine tuning of the parameters involved in the calculation of the objective measures were done in [101] and MATLAB codes for the implementation of the most important performance measures were provided in [7], including perceptual evaluation of speech quality (PESQ), segmental SNR (SNRseg), log-likelihood ratio (LLR), weighted-slope spectral distance (WSS), Itakura-Saito distance(IS), and cepstrum distance measures (CEP). In [101], it was found that, of all the tested objective measures, the PESQ measure yields the highest correlation with the overall quality and signal distortion judged by subjective testing. Also, it was concluded that the SNRseg and LLR measures perform almost as well as the PESQ but with lower computational costs, and therefore, they can be thought of as simple alternatives to the PESQ measure. All in all, we found in our experiments that most of the important objective measures are often consistent but some like SNRseg are more sensitive to the amount of noise reduction and some like LLR are more sensitive to the signal distortion present in the enhanced speech. In the sequel, we discuss in brief the three performance measures that we mainly used for the evaluation of our method, i.e., PESQ, SNRseg and LLR.

PESQ: This measure is one of the most complex to compute yet very favorable performance measure, particularly in assessing noise suppression methods. It is basically the one recommended by ITU-T (International Telecommunication Union, Telecommunication Standardization Sector) for speech quality assessment of narrow-band handset telephony and also narrow-band speech codecs [102]. Nowadays, PESQ is a widely accepted industrial standard for objective voice quality evaluation and is standardized as ITU-T recommendation P.862 [102]. Since PESQ measurements principally model the Mean Opinion Scores (MOS), it has a close connection to subjective performance tests performed by humans. In essence, the PESQ score is computed as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance value A_{ind} as the following [101]

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{3.20}$$

where $a_0 = 4.5$, $a_1 = -0.1$, and $a_2 = -0.0309$. Generally, this score takes a value between 1 and 4.5, with 4.5 rated as the highest and 1 rated as the lowest quality of speech. Multiple linear regression analysis was used in [101] to optimize the parameters a_0, a_1 and a_2 as the aforementioned values for speech distortion, noise distortion, and overall quality.

SNRseg: The segmental SNR measure was originally proposed in [103] as

$$\text{SNRseg} = \frac{1}{L} \sum_{\ell=0}^{L-1} 10 \ \log_{10} \left(\frac{||\mathbf{x}_\ell(n)||^2}{||\mathbf{x}_\ell(n) - \hat{\mathbf{x}}_\ell(n)||^2} \right) \tag{3.21}$$

where $\mathbf{y}_\ell(n)$ and $\hat{\mathbf{y}}_\ell(n)$ denote vectors consisting of the clean and enhanced speech at frame ℓ, respectively, and $\mathcal{L}$ is the number of frames in the entire speech signal. As suggested in [101], only frames with an SNR in the range of 10 to 35 dB were considered in the averaging in (3.21). **LLR**: This performance measure is one of the mostly used linear prediction coefficient (LPC) -based scores in speech enhancement and is defined for each speech frame as [101]

$$\mathrm{LLR} = \log\left(\frac{\mathbf{a}_p\mathbf{R}_c\mathbf{a}_p^T}{\mathbf{a}_c\mathbf{R}_c\mathbf{a}_c^T}\right) \tag{3.22}$$

where $\mathbf{a}_p$ and $\mathbf{a}_c$ are row vectors containing the LPC coefficients of the enhanced and clean speech signals at each frame, respectively, and $\mathbf{R}_c$ is the auto-correlation matrix of the clean speech signal at the frame. To discard unrealistically high values obtained by (3.22), the lower 95% of the frame LLR values are used to calculate the average LLR. Also, the frame LLR values given by (3.22) are restricted in the range of [0, 2] to further reduce the number of outliers [101]. Contrary to PESQ and SNRseg, a lower LLR score indicates a higher speech quality.

3.5.2 Evaluation of the Proposed Method

In this section, we evaluate the performance of the proposed STSA estimation methods using objective speech quality measures. First, the performance of the proposed STSA parameter selection and gain flooring schemes are compared to the previous methods. Next, the proposed GGD-based estimator is compared to the estimators using the conventional Rayleigh prior. Due to the performance advantage of the generic Wβ-SA estimator over the previous versions of STSA estimators, it is used throughout the following simulations.

Various types of noise from the NOISEX-92 database [104] were considered for the evaluations, out of which the results are presented for three types of noise, i.e., white, babble and car noises. Speech utterances including 10 male and 10 female speakers are used from the TIMIT speech database [105]. The sampling rate is set to 16 kHz and a Hamming window with length 20 ms and overlap of 75% between consecutive frames is used for STFT analysis and overlap-add synthesis. In all simulations, the noise variance is estimated by the soft-decision IMCRA method [80] eliminating the need to use a hard-decision voice activity detector (VAD). Also, the decision-directed (DD) approach [17] is used to estimate the *a priori* SNR.

To illustrate graphically the advantage achieved by the proposed parameter selection scheme, first we plot the speech spectrograms for the noisy, clean and enhanced speech signals for the case of babble noise in Figure 3.5. We considered the original frequency-based scheme in [32] and compared it to the suggested scheme in Section 3.3 where for both schemes, the gain flooring in (3.13) is used. It is observed that, particularly at low frequencies, the estimator with the original scheme cannot preserve the clean speech component satisfactorily, whereas it over-amplifies other parts of the speech spectrum. The disappearance of the very low-frequency portion of the spectral content is mainly due to the too small values of the parameter α given by this scheme. However, the proposed parameter selection scheme is capable of retaining most of the strong components of the clean speech spectrum, especially in the low frequencies. Further noise reduction can also be observed through the use of the proposed selection schemes for α and β.

To evaluate the efficiency of the proposed parameter selection schemes as well as the proposed gain flooring scheme, we herein present the performance measures for the Wβ-SA estimator with the parameter scheme in [32], the Wβ-SA estimator using the proposed parameter selection in Section 3.3 and also the same estimators with the proposed gain flooring in (3.13). We employed the gain flooring scheme in [98] or (3.12) in cases where the proposed gain flooring is not used, since the closest results to the proposed flooring were obtained under this scheme. The LLR results for the three noise types in the range of input global SNR between -10 dB and 10 dB are presented in Figure 3.6. As stated in Section 3.3, the original choice of the parameters of the Wβ-SA estimator results in an excessive distortion in the enhanced speech, which is observable through the LLR values in Figure 3.6. Yet, the suggested adaptive parameter selection completely resolves this problem and is also able to yield further improvement. Moreover, the use of the recursive smoothing based gain flooring in (3.13) is able to remove further speech distortion compared to the gain flooring scheme in [98] as given by (3.12), especially at higher SNRs. This is due to the incorporation of the estimated speech, which is strongly present at high SNRs, in the flooring value instead of using the noise masking threshold-based method. The result is that the gain floor is kept at more moderate levels in order not to distort the existing speech components. Similar trends can be observed in Figure 3.7 and Figure 3.8 in terms of the speech quality determined by PESQ and noise reduction evaluated by segmental SNR measurements, respectively. As it is observed, in cases where the proposed parameter setting is able to provide only minor improvements over the original method, the combination of the proposed parameters with the gain flooring improves the performance to a considerable degree.

To have a more detailed evaluation of each of the suggested schemes, we present the results obtained by individually applying each of them to the Wβ-SA estimator. In Tables 3.1-3.3, PESQ results for the Wβ-SA estimator considering α=0 (corresponding to the β-SA estimator), α=0.22 (an empirically fixed choice of α), original scheme for α as in [32] and the proposed scheme for α in

(3.6). In all cases, the proposed scheme for β and so for the gain flooring have been employed. It is observed that, whereas the employment of the STSA weighting through the parameter α results in a considerable improvement compared to the β-SA estimator, as seen in the last row of the tables. Within the same line, Tables 3.4-3.6 are representative of the evaluations performed on the Wβ-SA estimator by using β=1.82 (an empirically fixed value), β given by (3.7), β given by (3.8) and the proposed choice of β as in (3.9). In all cases, we employed α as proposed in (3.6) and the gain flooring proposed in (3.13). It can be deduced that, apart from the benefit obtained by the frequency-dependent choices of β through (3.7) and (3.8) over the fixed choice of this parameter, the suggested scheme in (3.9) is able to achieve notable improvements compared to the others.

To investigate the performance improvement attained by the proposed gain flooring scheme in (3.13) individually, we implemented the Wβ-SA estimator in Section 3.3 using different gain flooring schemes. In Figure 3.9, PESQ results have been shown for this estimator using the developed gain flooring in (3.13), those given by (3.10) and (3.12), as well as a fixed gain thresholding with μ_0=0.08. It is observed that, whereas the gain flooring in (3.12) leads to improvements with respect to the conventional fixed thresholding, the one in (3.10) only slightly outperforms the employed fixed flooring. This shows that the gain function itself, as used in (3.12), is a better measure for gain flooring compared to the *a posteriori* SNR used in (3.10). This is the reason we based our gain flooring scheme on (3.12) but further employed the noise masking concept to threshold the gain function values. As illustrated, the proposed gain flooring outperforms the scheme in (3.12) considerably even in the higher range of the input SNR. This is due to the fact that, even at such SNRs, there are frequencies in which the gain function decays abruptly below the threshold value, requiring an appropriate flooring value to keep the speech components.

Next, we investigated the performance advantage obtained by the proposed GGD-based estimator in Section 3.4 over the original Rayleigh-based estimator [32]. Also, to illustrate the superiority of the proposed scheme for the selection of the GGD parameter c in Section 3.4 with respect to the employed fixed values as in [37], we considered the same GGD-based estimator with different choices of the parameter c. In Figure 3.10, PESQ results are plotted for the original and suggested Wβ-SA estimators as well as two fixed choices of the parameter c in the range of $[c_{min}, c_{max}]$ as in Section 3.4. As observed, whereas the use of a GGD speech prior with fixed choices of c results in improvements with respect to the Rayleigh speech prior in most of the cases, the suggested SNR-based scheme for choosing c is capable of providing further enhancement compared to different fixed c choices. Other choices of the parameter c did not result in further improvements than those considered herein.

To evaluate the performance of the proposed GGD-based Wβ-SA estimator in Section 3.4 with respect to the recent STSA estimators using super-Gaussian priors, we considered the STSA estimation methods proposed in [34, 85]. In [85], the GGD model with a few choices of fixed parameters is applied as the STSA prior using the Log-MMSE estimator, whereas in [34], WE and WCOSH estimators (originally introduced in [29]) are developed exploiting a Chi PDF with fixed parameters as the STSA prior. Figure 3.11 illustrates speech spectrograms for the aforementioned STSA estimators in the case of babble noise. Through careful inspection of the speech spectrograms, it is observed that the proposed estimator is capable of maintaining clean speech components at least as much as the other estimators whereas further noise reduction, especially in the lower frequency range, is clearly obtained by using the proposed estimator. In Figures 3.12-3.14, performance comparisons for the same estimators are depicted in terms of LLR, PESQ and segmental SNR respectively. We used the gain flooring scheme proposed in Section 3.3 for all of the estimators. It is observed that, while the estimators suggested in [34] perform better than the one in [85] in most of the cases, the proposed STSA estimator in Section 3.4 is able to achieve superior performance especially at the lower SNR. This is mainly due to the further contribution of the speech STSA in the Bayesian cost function parameters through (3.2) as well as properly selecting the STSA prior shape parameter using (3.18) to adjust the gain function values. Whereas the latter is assigned a fixed value in the two previous STSA estimation methods, careful selection of this parameter based on the estimated *a priori* SNR leads to a more accurate model for the speech STSA prior.

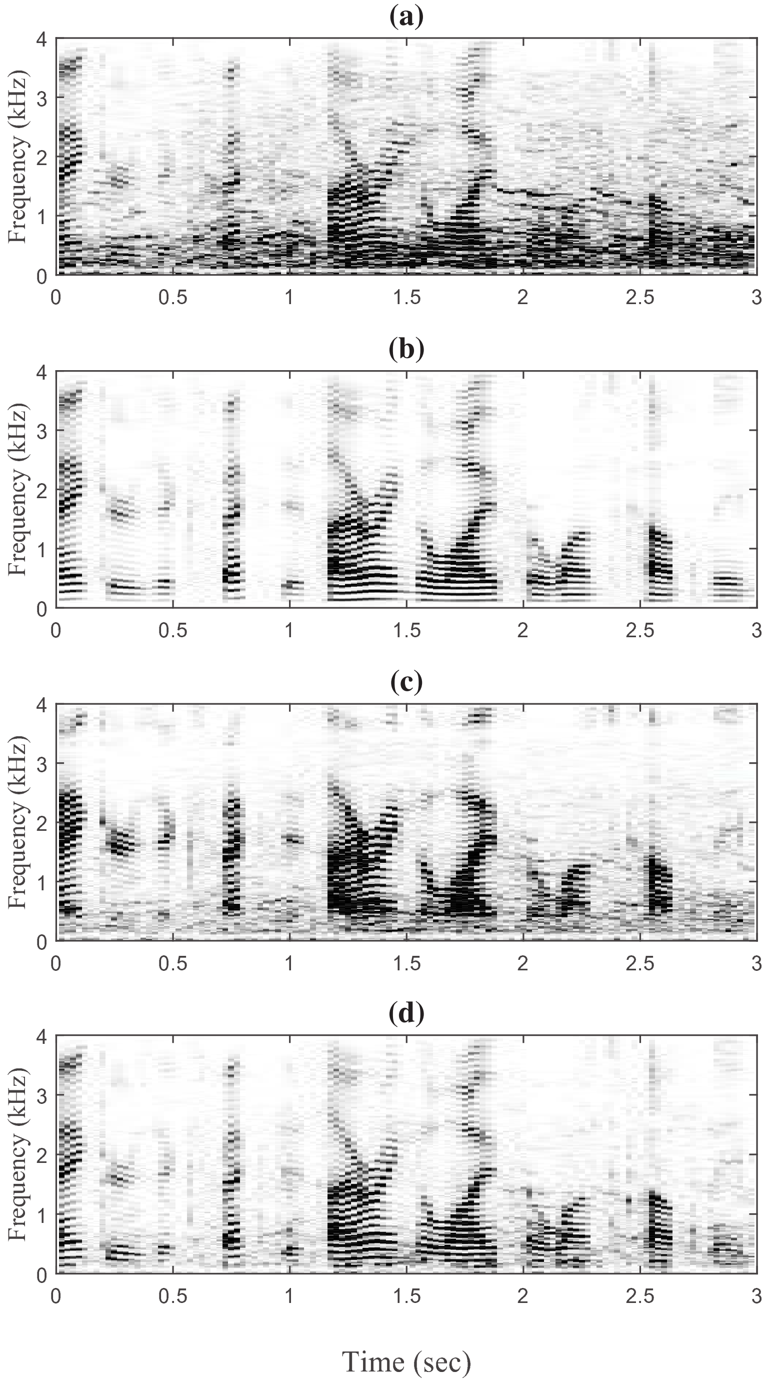


Figure 3.5: Spectrograms of (a): input noisy speech, (b): clean speech, (c): enhanced speech by the original Wβ-SA estimator and (d): enhanced speech by the proposed Wβ-SA estimator, in case of babble noise (Input SNR=5 dB).

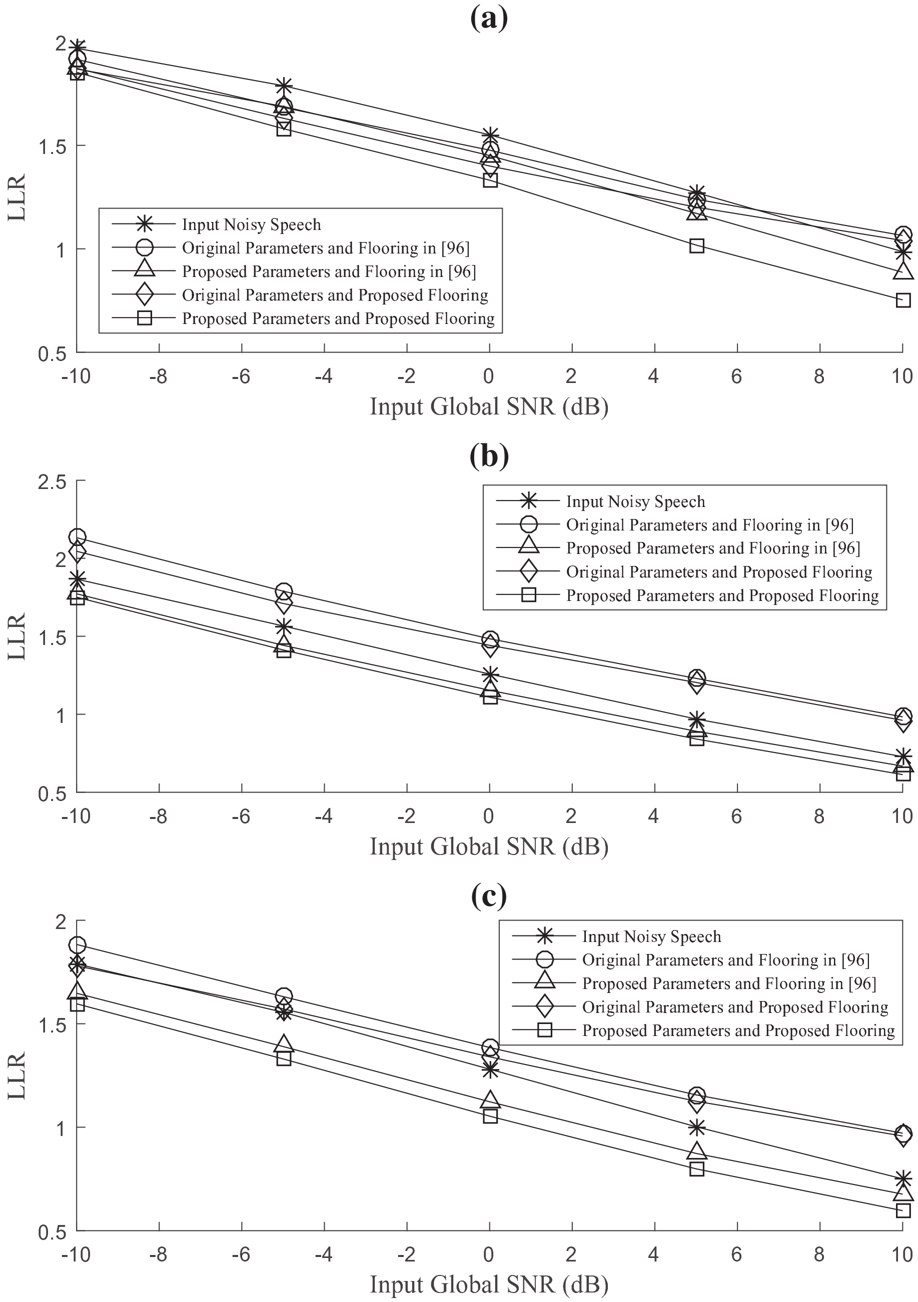


Figure 3.6: LLR versus global SNR for different Wβ-SA estimators, (a): white noise, (b): babble noise and (c): car noise.

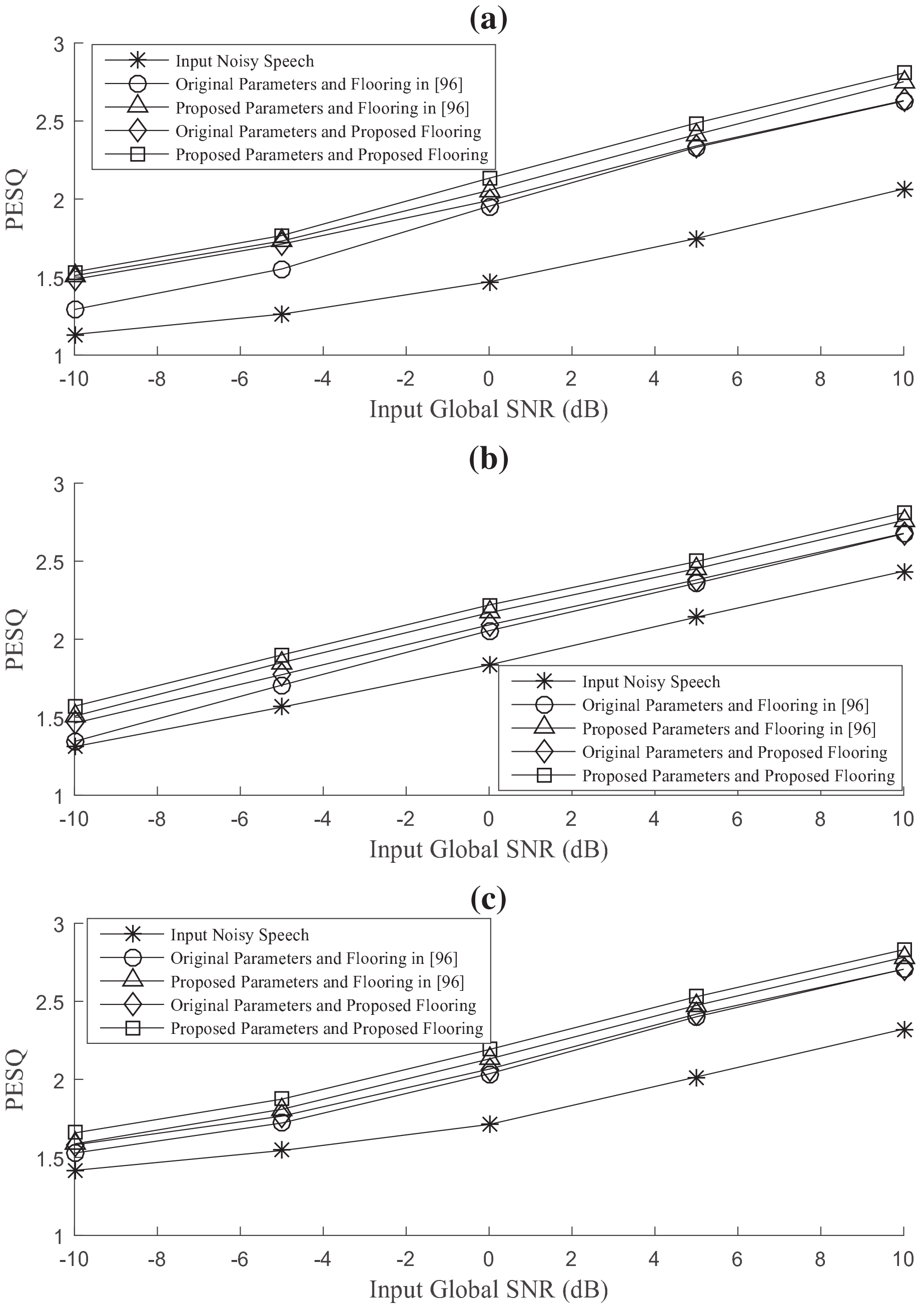


Figure 3.7: PESQ versus global SNR for different Wβ-SA estimators, (a): white noise, (b): babble noise and (c): car noise.

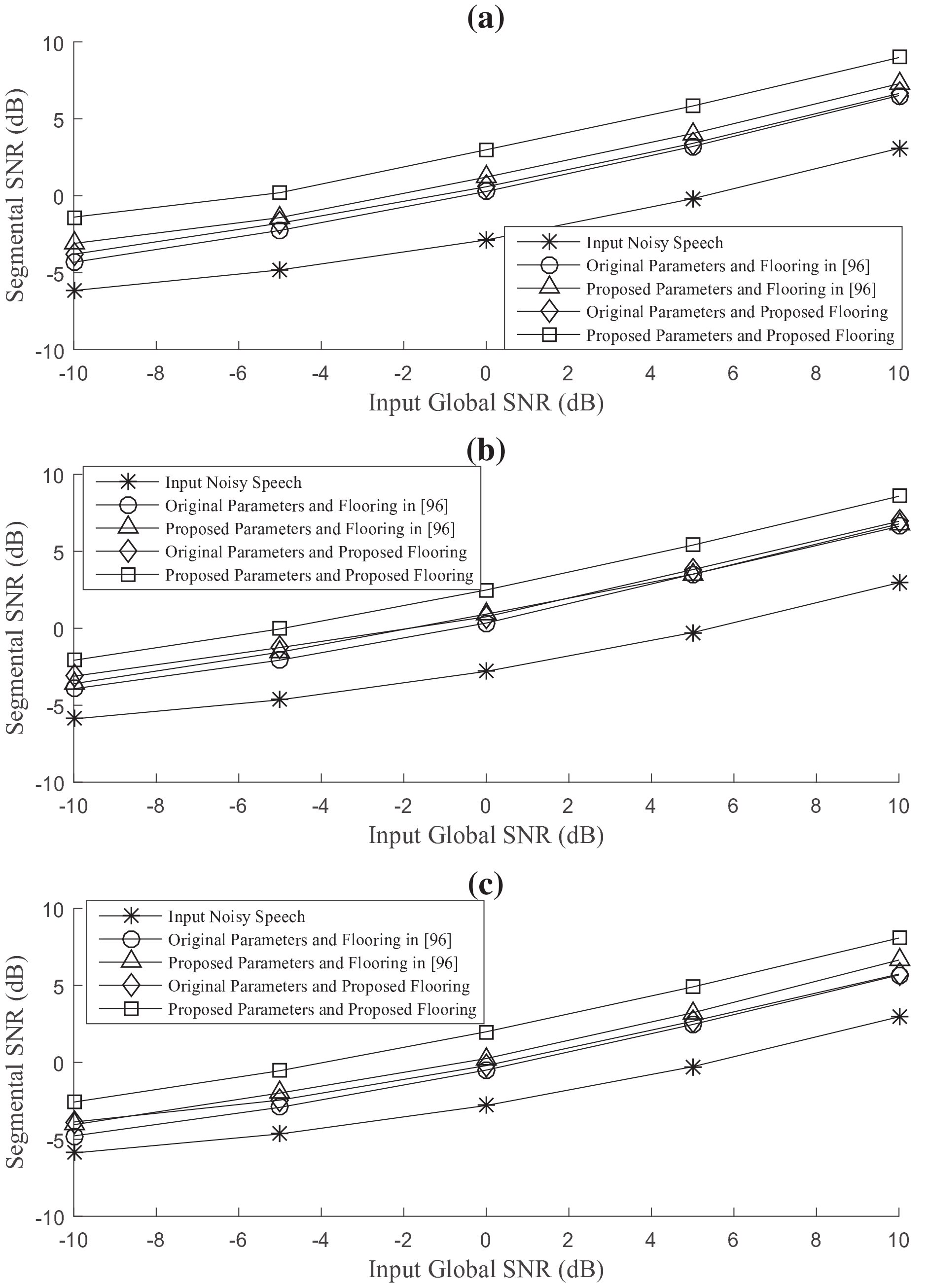


Figure 3.8: SNRseg versus global SNR for different Wβ-SA estimators, (a): white noise, (b): babble noise and (c): car noise.

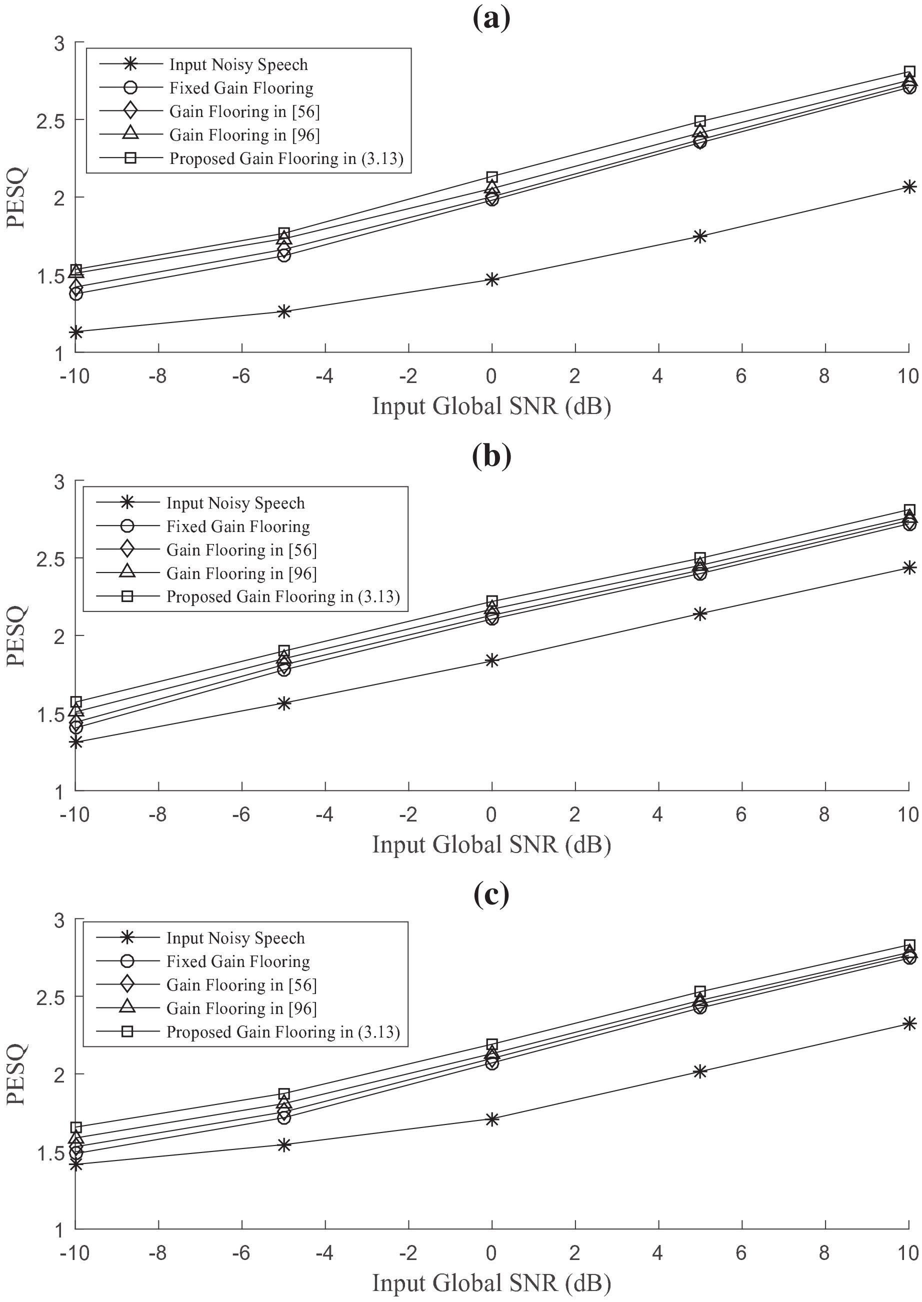


Figure 3.9: PESQ versus global SNR for Wβ-SA estimator with the proposed parameters in Section 3.3 using different gain flooring schemes, (a): white noise, (b): babble noise and (c): car noise.

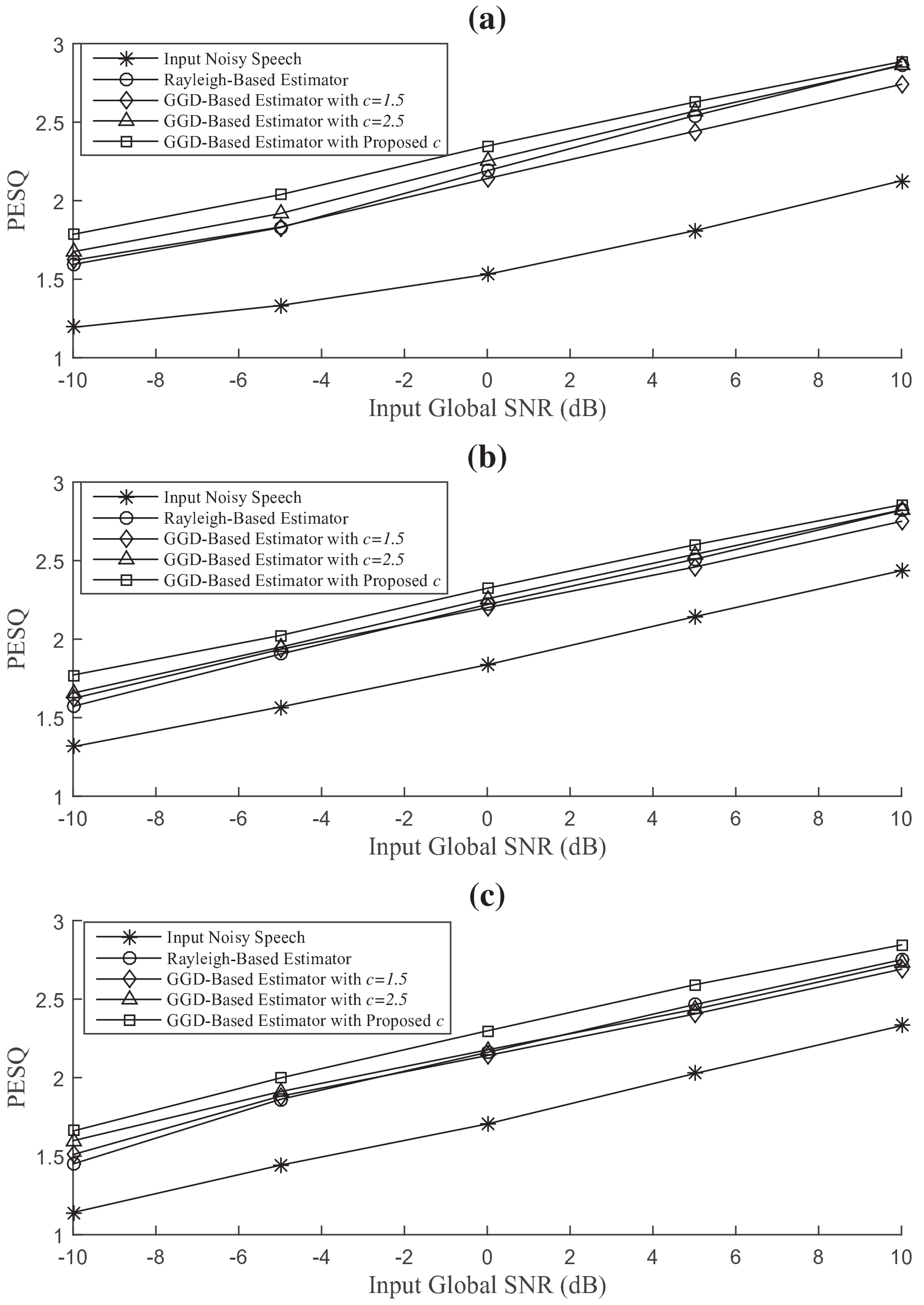


Figure 3.10: PESQ versus global SNR for the Rayleigh-based estimator in Section 3.3, the GGD-based estimator in Section 3.4 with $c = 1.5, 2.5$ and the proposed choice of c in Section 3.4, (a): white noise, (b): babble noise and (c): car noise.

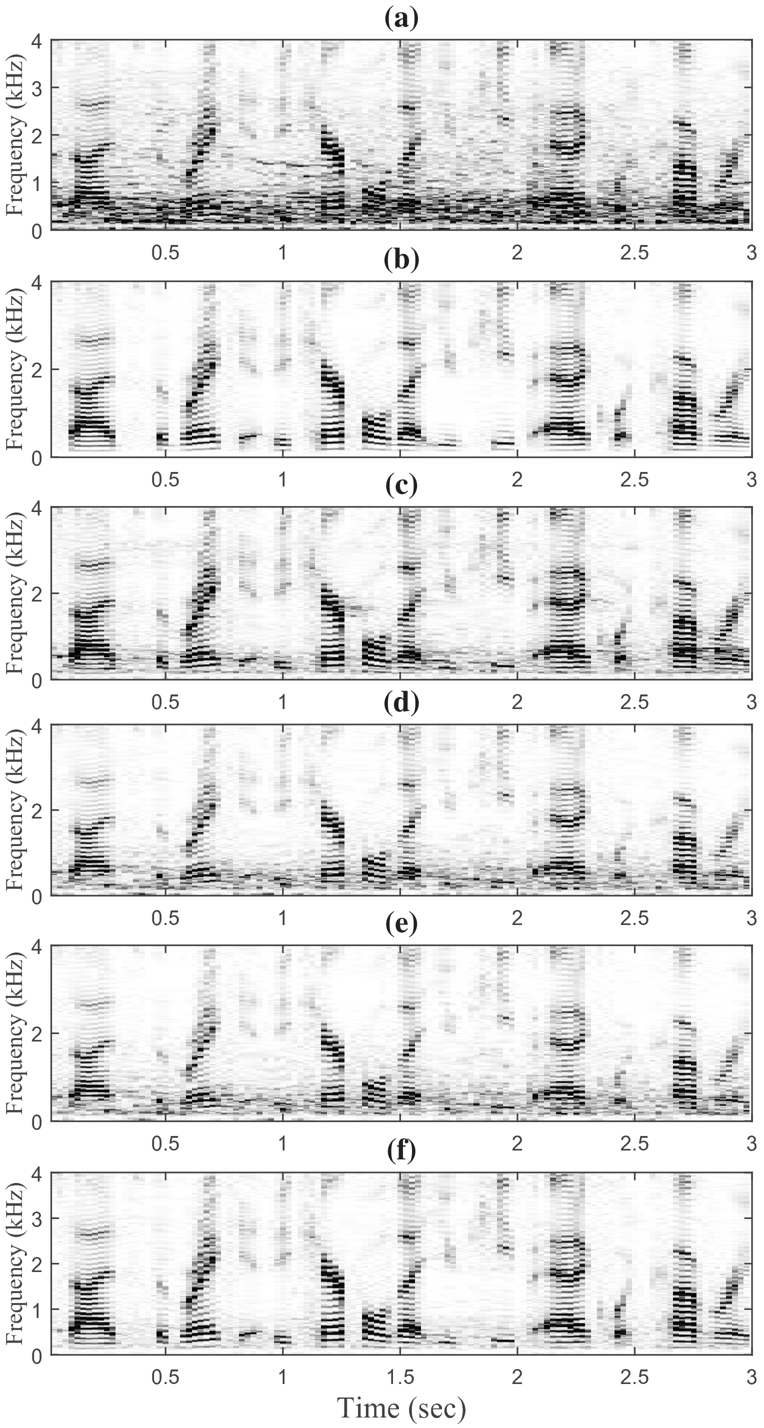


Figure 3.11: Spectrograms of (a): input noisy speech, (b): clean speech, (c): enhanced speech by WE estimator with Chi prior in [34], (d): enhanced speech by WCOSH estimator with Chi prior in [34], (e): enhanced speech by Log-MMSE estimator with GGD prior in [85] and (f): enhanced speech by the proposed Wβ-SA estimator with GGD prior in Section 3.4, in case of babble noise (Input SNR=5 dB).

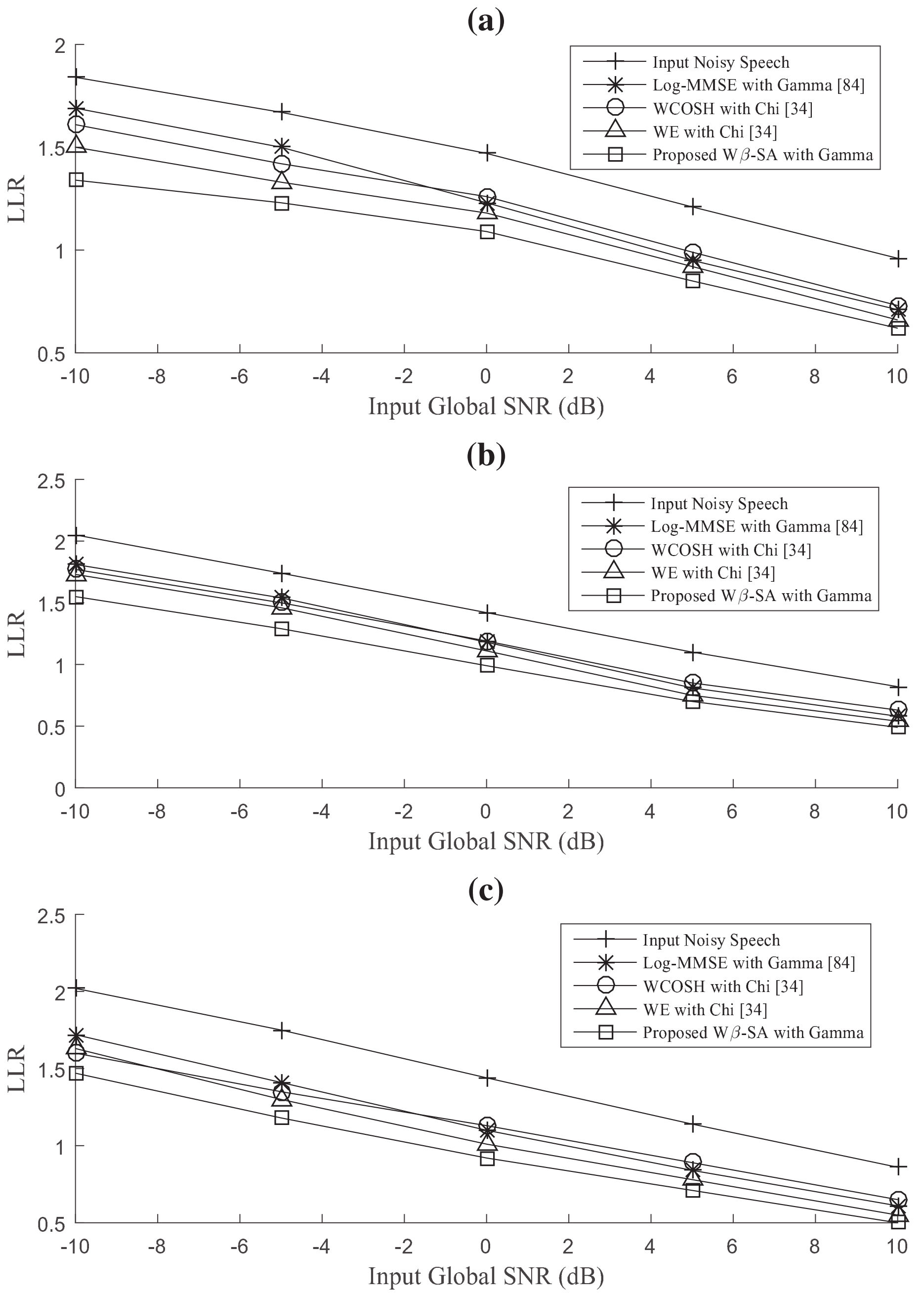


Figure 3.12: LLR versus global SNR for the STSA estimators in [34, 85] and the proposed STSA estimator in Section 3.4, (a): white noise, (b): babble noise and (c): car noise.

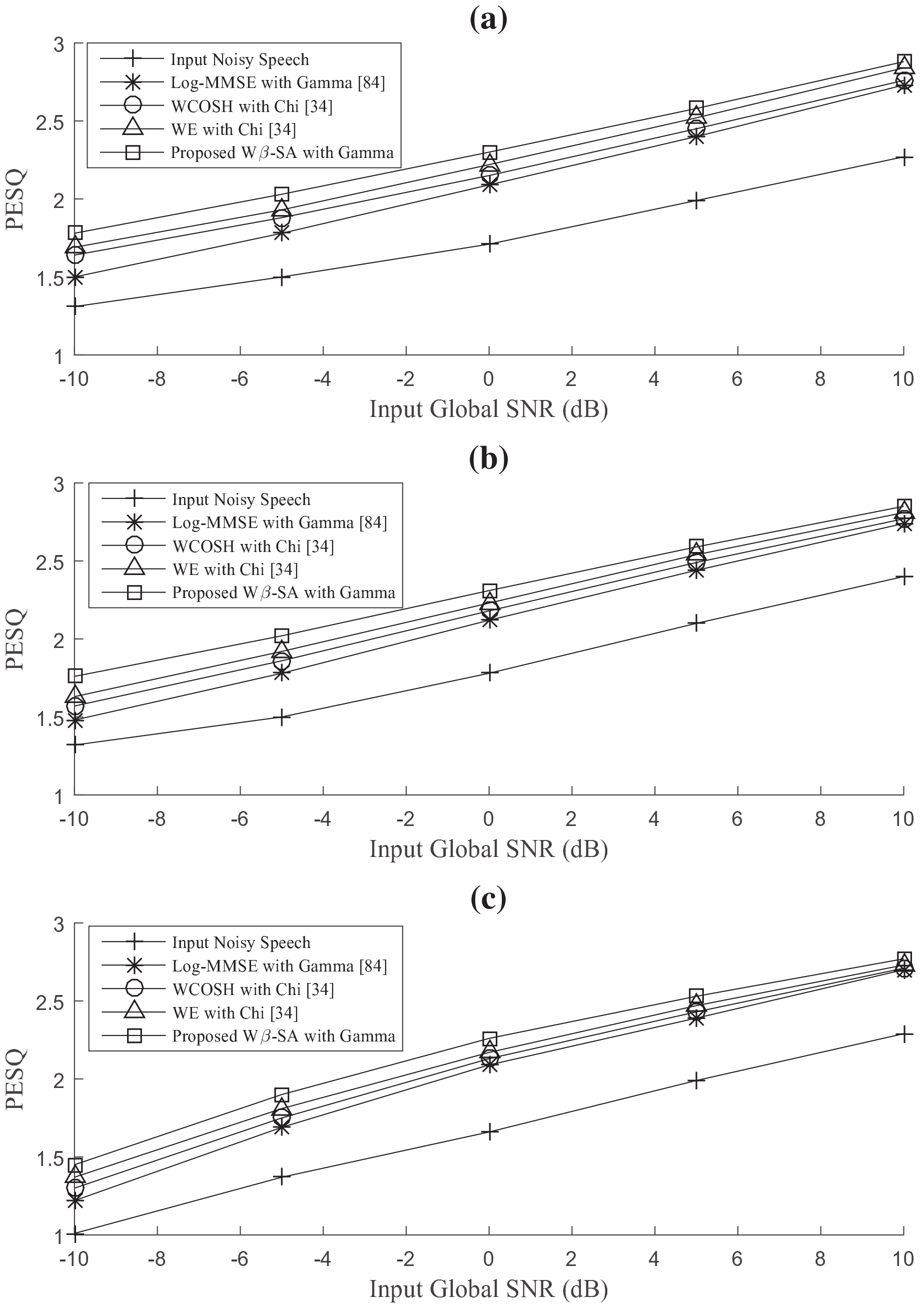


Figure 3.13: PESQ versus global SNR for the STSA estimators in [34, 85] and the proposed STSA estimator in Section 3.4, (a): white noise, (b): babble noise and (c): car noise.

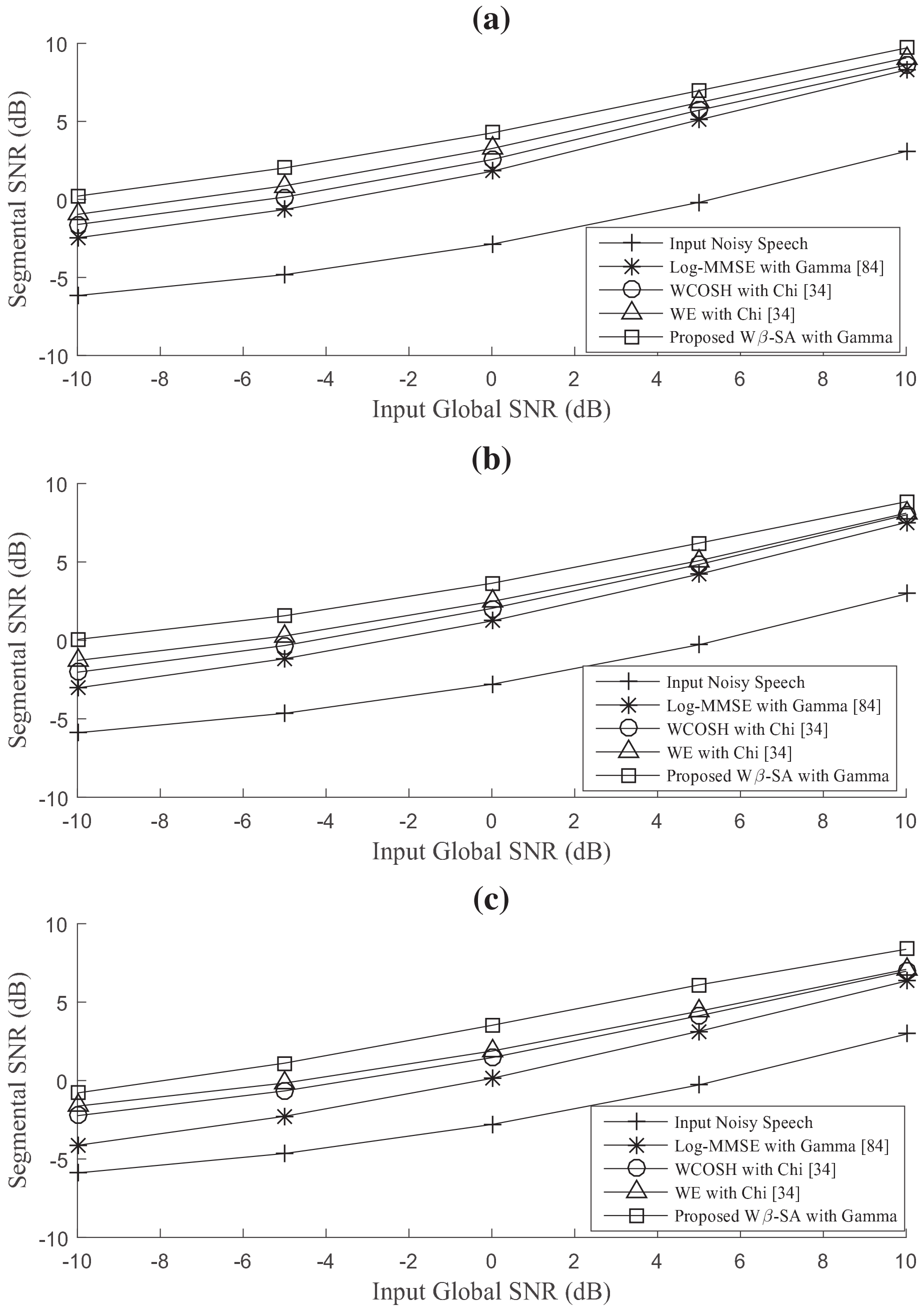


Figure 3.14: SNRseg versus global SNR for the STSA estimators in [34, 85] and the proposed STSA estimator in Section 3.4, (a): white noise, (b): babble noise and (c): car noise.

Table 3.1: PESQ values for the Wβ-SA estimator with different schemes of parameter α, case of white noise.

Input SNR (dB)	-10	-5	0	5	10
Input Noisy Speech	1.13	1.26	1.47	1.75	2.06
Choice of α=0	1.49	1.70	2.03	2.39	2.72
Choice of α=0.22	1.49	1.73	2.06	2.41	2.76
Original Choice of α	1.50	1.73	2.08	2.44	2.78
Proposed Choice of α	1.54	1.77	2.14	2.49	2.81

Table 3.2: PESQ values for the Wβ-SA estimator with different schemes of parameter α, case of babble noise.

Input SNR (dB)	-10	-5	0	5	10
Input Noisy Speech	1.31	1.56	1.83	2.14	2.43
Choice of α=0	1.48	1.71	2.03	2.40	2.73
Choice of α=0.22	1.51	1.82	2.14	2.42	2.77
Original Choice of α	1.54	1.86	2.16	2.45	2.79
Proposed Choice of α	1.58	1.91	2.23	2.51	2.82

Table 3.3: PESQ values for the Wβ-SA estimator with different schemes of parameter α, case of car noise.

Input SNR (dB)	-10	-5	0	5	10
Input Noisy Speech	1.41	1.54	1.71	2.01	2.32
Choice of α=0	1.57	1.76	2.06	2.40	2.75
Choice of α=0.22	1.58	1.78	2.11	2.46	2.77
Original Choice of α	1.60	1.81	2.15	2.50	2.79
Proposed Choice of α	1.66	1.88	2.20	2.54	2.84

Table 3.4: PESQ values for the Wβ-SA estimator with different schemes of parameter β, case of white noise.

Input SNR (dB)	-10	-5	0	5	10
Input Noisy Speech	1.13	1.26	1.47	1.75	2.06
Choice of β=1.82	1.48	1.69	2.00	2.32	2.68
Choice of β in [57]	1.53	1.74	2.08	2.39	2.72
Choice of β in [98]	1.52	1.74	2.06	2.42	2.75
Proposed Choice of β	1.54	1.77	2.14	2.49	2.81

Table 3.5: PESQ values for the Wβ-SA estimator with different schemes of parameter β, case of babble noise.

Input SNR (dB)	-10	-5	0	5	10
Input Noisy Speech	1.31	1.56	1.83	2.14	2.43
Choice of β=1.82	1.49	1.73	2.04	2.42	2.73
Choice of β in [57]	1.55	1.88	2.18	2.46	2.76
Choice of β in [98]	1.55	1.88	2.17	2.47	2.79
Proposed Choice of β	1.58	1.91	2.23	2.51	2.82

Table 3.6: PESQ values for the Wβ-SA estimator with different schemes of parameter β, case of car noise.

Input SNR (dB)	-10	-5	0	5	10
Input Noisy Speech	1.13	1.26	1.47	1.75	2.06
Choice of β=1.82	1.60	1.81	2.09	2.43	2.76
Choice of β in [57]	1.63	1.84	2.14	2.49	2.78
Choice of β in [98]	1.62	1.83	2.14	2.51	2.80
Proposed Choice of β	1.66	1.88	2.20	2.54	2.84

3.6 Conclusion

In this work, we presented new schemes for the selection of Bayesian cost function parameters in parametric STSA estimators, based on an initial estimate of the speech and the properties of human audition. We further used these quantities to design an efficient flooring scheme for the estimator's gain function, which employs recursive smoothing of the speech initial estimate. Next, we applied the GGD model as the speech STSA prior to the Wβ-SA estimator and proposed to choose its parameters according to the noise spectral variance and the *a priori* SNR. Due to the more efficient adjustment of the estimator's gain function by the suggested parameter choice and also further keeping the speech strong components from being distorted through the gain flooring scheme, our STSA estimation schemes are able to provide better noise reduction as well as less speech distortion compared to the previous methods. Also, by taking into account a more precise modeling of the speech STSA prior through using the GGD function with the suggested adaptive parameter selection, improvements were achieved with respect to the recent speech STSA estimators. Quality and noise reduction performance evaluations indicated the superiority of the proposed speech STSA estimation with respect to the previous estimators. It is worth mentioning that a wide range of subjective testing of the proposed method as opposed to previous methods has also been conducted during this research. We have found that the proposed method is capable of providing further noise reduction along with lower undesirable speech distortion, as compared to the other methods.

Chapter 4

Multi-Channel Bayesian STSA Estimation for Noise Suppression

4.1 Introduction

Whereas single microphone approaches are found to provide limited performance improvement, their multiple microphone counterparts have gained increasing popularity. This is mainly due to their capability in maintaining the introduced speech distortion at a low level while providing higher levels of noise reduction [46]. In this regard, considering multi-channel noise reduction in the STFT domain, two groups of methods can be recognized: those treating the ambient noise spatially (i.e., across microphones) uncorrelated and those taking into account the spatial correlation in noise. In the category of Bayesian STSA estimators, based on different cost functions and speech STSA priors, single channel methods have been widely developed and investigated in the literature. However, their multiple channel counterparts have not been explored thoroughly, particularly in the case of spatially correlated noisy environments.

In [28], it is assumed that the microphone array observations are spatially uncorrelated and then a speech STSA estimator is used for each microphone observation separately. However, no optimal solution is proposed on how to combine the outputs resulting from the processing of each channel. Also, as discussed in Section 2.3, there have been a few multi-channel extensions of Bayesian STSA estimators for spatially uncorrelated noise, which take into account MMSE, log-MMSE and β-SA cost functions and also super Gaussian speech priors [40, 41, 93]. Yet, similar to the single channel case, there has been no unified generalization of these estimators and selection of the corresponding parameters, considering the available cost functions and speech priors. In fact, this problem can be thought of as the multi-channel extension of the same problem targeted in Chapter 3 in the spatially uncorrelated noisy environment.

As discussed in Chapter 2 and is seen in Table 2.3, the gain function of all existing Bayesian STSA estimators depends only on the spectral amplitude of the speech signal and is not decided based on the spectral phase. However, like many other types of signals, the phase of the speech signal carries useful information about its structure and can be incorporated in further improving the quality of enhanced speech [27]. Therefore, incorporation of the speech spectral phase in Bayesian STSA estimators can lead to more accurate and less distorting gain functions. This topic is targeted in this chapter for the multi-channel speech enhancement in spatially uncorrelated noise with single channel as a special case.

The assumption of spatially uncorrelated noise is approximately true for some applications, e.g. when the spacing between microphones is large so that the incoming noise can be dealt as spatially white. However, in real world applications, the incoming noise at a microphone array is often correlated across the microphones and therefore the aforementioned assumption is inaccurate. This is specifically true for closely placed microphones or in circumstances with speech-like noise (interference) so that the impinging interference on the microphone array shows correlations across different microphones [46]. Thus, it is necessary to explore and develop the multi-channel STSA estimation method for the case of spatially correlated noise as part of this chapter. It will be revealed that the extension of a single-channel Bayesian STSA estimation to the corresponding spatially correlated multi-channel case can be done under a unified framework but it requires the estimation of further information, i.e., the DOA of the impinging speech signal and the PSD (spectral correlation) matrix of the background noise.

The rest of this chapter is organized as follows. In Section 4.2, a brief summary of the proposed approaches is given. Section 4.3.1 discusses the extension of the proposed Wβ-SA estimator in Chapter 3 to multi-channel in spatially uncorrelated noise. In Section 4.3.2, the proposed Bayesian STSA estimator using the speech spectral phase is presented. Performance evaluation of the proposed uncorrelated multi-channel STSA estimators is performed in Section 4.3.3. Section 4.4.1 is devoted to the generic extension of the single-channel to the multi-channel STSA estimation under known DOA and noise PSD matrix. In Section 4.4.2, the proposed approach for the estimation of the spatial noise PSD matrix is explained. This section is followed by the performance evaluation of the proposed schemes for the correlated multi-channel STSA estimation in Section 4.4.3. Conclusions are drawn in Section 4.5.

4.2 Brief Description of the Proposed Methods

In this chapter, first we generalize the proposed Wβ-SA estimator in Chapter 3 to the multi-channel case with spatially uncorrelated noise. It will be seen that, under the Bayesian framework, a straightforward extension from the single-channel to the multi-channel case exists by generalizing the STSA estimator parameters, i.e., α and β. Next, we present the development of Bayesian STSA estimators by taking advantage of speech spectral phase rather than relying only on the spectral amplitude of observations, contrary to the conventional methods. We develop STSA estimators with spectral phase by using the basic MMSE as well as the Wβ-SA cost functions. This contribution is considered for the multi-channel scenario with single-channel as a special case. Next, we tackle the problem of multi-channel STSA estimation under spatially correlated noise and derive a generic structure for the extension of a single-channel estimator to its multi-channel counterpart in the Bayesian framework. It is shown that the derived multi-channel extension requires estimates of the DOA and the spatial PSD matrix of noise. Subsequently, we aim at the estimation of the noise PSD matrix, that is not only important for the multi-channel STSA estimation scheme but also highly useful in different beamforming methods.

The presented contributions in this chapter have been published in [106] and [107].

4.3 Multi-Channel STSA Estimation in Spatially Uncorrelated Noise

In this section, we first extend the proposed Wβ-SA estimator of the previous chapter to the case of multi-channel with spatially uncorrelated noise. Next, we propose the Bayesian estimation of speech STSA using spectral phase in the multi-channel case.

4.3.1 Extension of the Proposed Wβ-SA Estimator to Multi-Channel

We consider the same problem formulation as that in Section 2.3.1, where a microphone array consists of N omni-directional sensors each spaced d meters apart, that receives a far-field speech source at a known DOA equal to θ. The microphone array captures the noisy observations $y_n(t)$ which consists of the time delayed clean speech signals $x(t-\tau_n)$ corrupted by additive spatially uncorrelated noises $v_n(t)$, with n as the microphone index and τ_n as the relative time delay of the speech signal in the nth microphone with respect to the reference (first) microphone. Therefore, it follows that

$$y_n(t) = x(t-\tau_n) + v_n(t), \quad n = 1, 2, ..., N \tag{4.1}$$

where $x(t)$ is the speech signal under estimation. After sampling, framing and STFT analysis, the noisy speech signal can be represented as

$$Y_n(k,l) = X(k,l)e^{-j\phi_{n,k}} + V_n(k,l), \quad n = 1, 2, ..., N \tag{4.2}$$

The phase differential term, $\phi_{n,k}$, can be obtained as $2\pi f_s \tau_n k/K$ with f_s as the sampling frequency and K as the total number of frequency bins. By expressing (4.2) in the vector form, we have

$$\mathbf{Y}(k,l) = X(k,l)\mathbf{\Phi}(k) + \mathbf{V}(k,l) \tag{4.3}$$

with $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_N]^T$, $\mathbf{V} = [V_1, V_2, \cdots, V_N]^T$, and $\mathbf{\Phi} = [\phi_1, \phi_2, \cdots, \phi_N]^T$ as the so-called steering vector in the STFT domain. The latter is assumed to be known or estimated beforehand in this section (This is actually equivalent to the estimation of the DOA, which has been explored widely in the literature [46]). Note that the frequency bin k and frame index l are dropped for notational convenience. In this section, we postulate spatially uncorrelated noise, i.e.,

$$E\{V_n V_m\} = E\{V_n\}E\{V_m\} = 0, \quad \forall\ n, m \in \{1, 2, ..., N\},\ n \neq m \tag{4.4}$$

The speech spectral component X can be written as $\mathcal{X}e^{j\omega}$ with $\mathcal{X} \geq 0$ as the spectral amplitude and $\omega \in [-\pi, \pi]$ the spectral phase. Given the DOA parameter, or equivalently the arrival delay τ_n, the multi-channel STSA estimation targets the estimation of $\mathcal{X}$ using the noisy spectral observations Y_n.

Here, we generalize the single-channel MWβ-SA estimator (with GGD prior) in Section 3.4.1 to the multi-channel case. In this regard, based on Section 2.3.2, the following expression is obtained for the multi-channel counterpart of the MWβ-SA estimator

$$\hat{\mathcal{X}}^{(\text{MW}\beta\text{-SA})} = \left(\frac{E\{\mathcal{X}^{\beta+\alpha}|\mathbf{Y}\}}{E\{\mathcal{X}^{\alpha}|\mathbf{Y}\}} \right)^{1/\beta} \tag{4.5}$$

Note that the expression in (4.5) holds true regardless of the underlying distribution for the speech prior. To obtain a closed-form solution for (4.5), it is required to calculate the moment term $E\{\mathcal{X}^{\rho}|\mathbf{Y}\}$ with ρ as an arbitrary parameter. In the Bayesian framework, in a fashion similar to (2.18) in Chapter 2, it follows that

$$E\left\{\mathcal{X}^{\rho}|\mathbf{Y}\right\} = \frac{\int_0^{\infty}\int_0^{2\pi} \mathcal{X}^{\rho} p(\mathbf{Y}|\mathcal{X},\omega)p(\mathcal{X},\omega)d_{\omega}d_{\mathcal{X}}}{\int_0^{\infty}\int_0^{2\pi} p(\mathbf{Y}|\mathcal{X},\omega)p(\mathcal{X},\omega)d_{\omega}d_{\mathcal{X}}} \tag{4.6}$$

This equation is similar to (3.15), and therefore, it can be solved in the same manner as Appendix A, but by using (2.36) for $p(\mathbf{Y}|\mathcal{X},\omega)$ and (A.2) for $p(\mathcal{X},\omega)$, as the following

$$\begin{aligned} p(\mathbf{Y}|\mathcal{X},\omega) &= \left(\prod_{n=1}^{N} \frac{1}{\pi\sigma_{v_n}^2}\right) \exp\left(-\sum_{n=1}^{N} \frac{|Y_n - \mathcal{X}e^{j\omega}e^{-j\phi_n}|^2}{\sigma_{v_n}^2}\right) \\ p(\mathcal{X},\omega) &= \frac{1}{2\pi}\frac{2b^c}{\Gamma(c)}\mathcal{X}^{2c-1}\exp(-b\mathcal{X}^2) \end{aligned} \tag{4.7}$$

It should be noted that, using the second-order moment of $\mathcal{X}$, i.e., $\sigma_{\mathcal{X}}^2$, as discussed in Section 3.4, the two STSA prior parameters b and c are related as $\sigma_{\mathcal{X}}^2 = c/b$. Consequently, we obtain

$$E\left\{\mathcal{X}^{\rho}|\mathbf{Y}\right\} = \frac{\Gamma\left(\frac{\rho+2c}{2}\right)}{\Gamma(c)\left(\frac{c}{\sigma_{\mathcal{X}}^2} + \sum_{n=1}^{N}\frac{1}{\sigma_{v_n}^2}\right)^{\frac{\rho}{2}}} \frac{\mathrm{M}\left(\frac{2-\rho-2c}{2}, 1; -\nu''\right)}{\mathrm{M}\left(1-c, 1; -\nu''\right)} \tag{4.8}$$

with ν'' defined as

$$\nu'' = \frac{\left|\sum_{n=1}^{N}\frac{Y_n e^{j\phi_n}}{\sigma_{v_n}^2}\right|^2}{\frac{c}{\sigma_{\mathcal{X}}^2} + \sum_{n=1}^{N}\frac{1}{\sigma_{v_n}^2}} \tag{4.9}$$

where, similar to Section 2.3.2, $\sigma_{v_n}^2$ is the noise PSD at the nth channel. Now, by using (4.8) in (4.5), the following solution can be derived for the multi-channel MWβ-SA estimator

$$\hat{\mathcal{X}}^{(\text{MW}\beta\text{-SA})} = \left(\frac{\Gamma\left(\dfrac{\alpha+\beta+2c}{2}\right)}{\Gamma\left(\frac{\alpha}{2}+c\right)\left(\frac{c}{\sigma_{\mathcal{X}}^2}+\sum_{n=1}^{N}\frac{1}{\sigma_{v_n}^2}\right)^{\frac{\beta}{2}}} \; \frac{\mathrm{M}\left(\frac{2-\alpha-\beta-2c}{2},1;-\nu''\right)}{\mathrm{M}\left(\frac{2-\alpha-2c}{2},1;-\nu''\right)} \right)^{1/\beta} \tag{4.10}$$

The solution in (4.10) can be considered as the most general multi-channel Bayesian STSA estimator among the existing ones, and therefore, by specific choices of its corresponding parameters, other existing multi-channel estimators such as those in [40, 41, 93] can be deduced. Even though defining a gain function similar to that in the single-channel case is not possible here, it can be shown that for $N = 1$, i.e., considering only one microphone, the multi-channel estimator given by (4.10) is degenerated to the single-channel MWβ-SA estimator as expressed by (3.17). Conversely, it can be shown that by generalizing the parameters λ and ν' in (3.16) to their multi-channel extension, any single-channel STSA estimator can be modified to its multi-channel counterpart in the spatially uncorrelated noise case. In this sense, ν' should be modified to ν'' given by (4.9) and λ is replaced by $c/\sigma_{\mathcal{X}}^2 + \sum_{n=1}^{N} 1/\sigma_{v_n}^2$.

Considering the parameter choice for the multi-channel STSA estimator in (4.10), due to the importance of the accuracy in estimating the speech spectral variance $\sigma_{\mathcal{X}}^2$, we estimate this parameter as the average of its values in all channels, i.e.,

$$\hat{\sigma}_{\mathcal{X}}^2 = \frac{1}{N}\sum_{n=1}^{N}\hat{\zeta}_n\hat{\sigma}_{v_n}^2 \tag{4.11}$$

with $\hat{\zeta}_n$ as the *a priori* SNR in the nth channel. Obviously, it is required to implement the underlying *a priori* SNR estimation (such as the DD approach) and also the noise PSD estimation method for all N channels independently. Regarding the selection of estimator parameters α, β and c in (4.10), we follow the same schemes as those proposed in Chapter 3. However, even though it is possible to average over the parameter values obtained for each of the channels, we choose to use the parameter values obtained from one of the channels, say the first one. In this sense, averaging the parameter values over all channels did not make any considerable difference in the overall performance of the estimator.

4.3.2 STSA Estimators Using Spectral Phase

In the groundbreaking research presented in [24], based on a comprehensive study on the importance of spectral phase estimation in speech processing, the following is concluded: if an estimate of the speech phase is used to reconstruct the speech signal through combining the speech phase estimate by an independently estimated speech amplitude, the resulting speech estimation (enhancement) method will not provide a promising performance. However, if an estimate of the phase is exploited to further improve the estimation of the speech amplitude, which is in turn combined with the noisy phase of the observation, then a more accurate estimate of the speech phase (than the noisy phase) will be useful. We make use of this fundamental conclusion in this section in order to improve the performance of the conventional Bayesian STSA estimators discussed so far, wherein only the spectral amplitude information is exploited to derive the estimator.

Conventionally, all STSA estimators are derived by assuming a uniformly distributed speech spectral phase and then treating the problem by taking statistical expectation with respect to the random spectral phase, as discussed in Section 2.1.6. Although being optimal in the sense of MMSE of the amplitude, these methods lack the use of any prior information about the phase component, and therefore, neglect the aforementioned potential to improve the performance of the speech spectral amplitude estimation by employing the spectral phase. In this section, we propose to treat the speech spectral phase as an unknown deterministic parameter and obtain a new class of STSA estimators that exploits the phase component in its structure. This unknown phase component can be estimated in the multi-channel case by a simple MMSE estimator of the phase suggested in [40] or even replaced by the noisy phase. In the following sections, we first develop the phase-aware STSA estimator using the basic MMSE cost function in the Bayesian framework and next extend it by exploiting the Wβ-SA cost function. Finally, we address the problem of spectral phase estimation for the proposed STSA estimator. Throughout the entire sections, we formulate the problem in the multi-channel case with single-channel as a special case.

4.3.2.1 MMSE-Based STSA Estimator Using Spectral Phase

An MMSE-based STSA estimator in the multi-channel case aims at the minimization of the MMSE cost function, $E\{(\mathcal{X}-\hat{\mathcal{X}})^2\}$, given the spectral observations from all channels, i.e., $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_N]^T$. Recall that the complex-valued speech STFT, X, is expressed as $\mathcal{X}e^{j\omega}$ with

$\omega \in [-\pi, \pi]$ as the spectral phase. As shown in [40], the MMSE estimate of the amplitude, $\hat{\mathcal{X}}$, is in fact the conditional expectation $E\{\mathcal{X}|\mathbf{Y}\}$. To obtain the latter, contrary to the conventional approach taken in [40] and also seen in (4.6), where both the spectral amplitude and phase are assumed to be stochastic and expectations over both are performed, we base our STSA estimation method on treating the spectral phase, ω, as a deterministic unknown parameter, $\hat{\omega}$, that is to be replaced by its estimate later. On this basis, using the conventional Bayesian framework for the distribution $p(\mathcal{X}|\mathbf{Y})$, it follows that

$$\hat{\mathcal{X}}^{(\mathrm{MMSE})} = E\{\mathcal{X}|\mathbf{Y}\} = \frac{\int_0^\infty \mathcal{X} p(\mathbf{Y}|\mathcal{X}, \hat{\omega}) p(\mathcal{X}) d_{\mathcal{X}}}{\int_0^\infty p(\mathbf{Y}|\mathcal{X}, \hat{\omega}) p(\mathcal{X}) d_{\mathcal{X}}} \tag{4.12}$$

with $\hat{\omega}$ as a proper estimate for the spectral phase. It is observed that, as compared to the conventional approach in (4.6), the integration over ω has been dropped, since the expectation has to be performed only on $\mathcal{X}$. In the same manner as that explained in Section 2.3.2, assuming spatially uncorrelated noise and denoting Y_n as $|Y_n|e^{j\Omega_n}$, we have

$$\begin{aligned} p(\mathbf{Y}|\mathcal{X}, \hat{\omega}) &= \prod_{n=1}^{N} p(Y_n|\mathcal{X}, \hat{\omega}) = \left(\prod_{n=1}^{N} \frac{1}{\pi\sigma_{v_n}^2}\right) \exp\left(-\sum_{n=1}^{N} \frac{|Y_n - \mathcal{X} e^{j\hat{\omega}} e^{-j\phi_n}|^2}{\sigma_{v_n}^2}\right) \\ &= \left(\prod_{n=1}^{N} \frac{1}{\pi\sigma_{v_n}^2}\right) \exp\left(\sum_{n=1}^{N} \frac{2|Y_n|\mathcal{X}\cos(\hat{\omega} - \phi_n - \Omega_n) - \mathcal{X}^2 - |Y_n|^2}{\sigma_{v_n}^2}\right) \end{aligned} \tag{4.13}$$

Here, we use the conventional Rayleigh PDF for $p(\mathcal{X})$ as

$$p(\mathcal{X}) = \frac{2\mathcal{X}}{\sigma_{\mathcal{X}}^2} \exp\left(-\frac{\mathcal{X}^2}{\sigma_{\mathcal{X}}^2}\right), \quad \mathcal{X} \geq 0 \tag{4.14}$$

Substituting (4.13) and (4.14) into (4.12), and using Eq. (3.462.5) and Eq. (3.462.7) in [77] to compute the resulting integrations, the following MMSE-based STSA estimator can be derived

$$\hat{\mathcal{X}}^{(\mathrm{MMSE})} = \frac{-\mu\lambda^2 + \sqrt{\pi}\frac{2\mu^2\lambda+1}{2\sqrt{\lambda^7}} \exp(\mu^2\lambda)\left(1 - \mathrm{erf}(\mu\sqrt{\lambda})\right)}{\lambda - \sqrt{\pi}\frac{\mu}{\sqrt{\lambda^3}} \exp(\mu^2\lambda)\left(1 - \mathrm{erf}(\mu\sqrt{\lambda})\right)} \tag{4.15}$$

where erf(.) denotes the Gaussian error function [77] and the parameters λ and μ are defined as

$$\frac{1}{\lambda} = \frac{1}{\sigma_{\mathcal{X}}^2} + \sum_{n=1}^{N} \frac{1}{\sigma_{v_n}^2}, \quad \mu = -\sum_{n=1}^{N} \frac{|Y_n|}{\sigma_{v_n}^2} \cos(\hat{\omega} - \phi_n - \Omega_n) \tag{4.16}$$

It is observed that, unlike the state-of-the-art Bayesian STSA estimation methods, the proposed STSA estimator in (4.15) does not employ confluent hypergeometric functions, and instead, exploits one term of the error function, erf(.), which has less computational load and a faster convergence rate by its implementation through the power series expansion [108]. This can be considered as one advantage of the proposed estimator in (4.15).

4.3.2.2 Extension to the Wβ-SA Estimator

The modified STSA estimation method based on the spectral phase presented in the previous section has been derived by using the MMSE cost function, which is the most basic Bayesian cost function in the category of STSA estimation methods. We here extend this estimator to a more general case by using the Wβ-SA cost function. It is known that the Wβ-SA estimator is derived by solving the moment term $E\{\mathcal{X}^\rho|\mathbf{Y}\}$ and using it in (4.5). In this sense, in a similar fashion to the previous section, it follows that

$$E\{\mathcal{X}^\rho|\mathbf{Y}\} = \frac{\int_0^\infty \mathcal{X}^\rho p(\mathbf{Y}|\mathcal{X},\hat{\omega})p(\mathcal{X})d_{\mathcal{X}}}{\int_0^\infty p(\mathbf{Y}|\mathcal{X},\hat{\omega})p(\mathcal{X})d_{\mathcal{X}}} \tag{4.17}$$

By using (4.13) and (4.14) in the above, the resulting integrations can be handled in a more general case than that in (4.12) by using Eq. (3.462.1) in [77]. This results in

$$E\{\mathcal{X}^\rho|\mathbf{Y}\} = \frac{\Gamma(2+\rho)(\frac{\lambda}{2})^{\rho/2}D_{-\rho-2}\left(\sqrt{2\mu\lambda}\right)}{\Gamma(2)D_{-2}\left(\sqrt{2\mu\lambda}\right)} \tag{4.18}$$

with $D_i(.)$ as the parabolic cylinder function of ith order defined by Eq. (9.24) in [77], $\Gamma(.)$ as the Gamma function, and λ and μ given by (4.16). Now, by using (4.18) for the moments in (4.5), the Wβ-SA estimator based on spectral phase is obtained as the following

$$\hat{\mathcal{X}}^{(\mathrm{W}\beta\text{-SA})} = \left(\frac{\Gamma(2+\alpha+\beta)(\frac{\lambda}{2})^{\beta/2}D_{-2-\alpha-\beta}\left(\mu\sqrt{2\lambda}\right)}{\Gamma(2+\alpha)D_{-2-\alpha}\left(\mu\sqrt{2\lambda}\right)}\right)^{1/\beta} \tag{4.19}$$

It can be shown that the Wβ-SA estimator in (4.19) is reduced to the MMSE estimator expressed in (4.15) by choosing α and β respectively as zero and one. As compared to the MMSE estimator proposed in the last section, the calculation of the parabolic cylinder functions, $D_i(.)$, is not computationally less complex than the confluent hypergeometric functions encountered in the conventional methods, yet, the superiority in noise reduction performance of the spectral phase-based STSA estimator in (4.19) makes it advantageous with respect to the conventional STSA estimation method.

4.3.2.3 Estimation of the Spectral Phase

In the STSA estimators proposed in the previous two subsections, the spectral phase of the speech signal, ω, is treated as an unknown deterministic parameter which has to be estimated. Traditionally in [17], it was proved that an MMSE-optimal estimate of the principal value of the phase is simply the noisy phase of the spectral observations. All conventional Bayesian STSA estimators, for this reason, tend to estimate only the spectral amplitude while keeping the phase unchanged. Furthermore, as stated in Section 1.3.3, early investigations for spectral phase estimation such as those reported in [24, 25], concluded that, given the inherent complexity in the estimation of speech spectral phase, it is not possible to estimate the latter with enough accuracy. On the other hand, rather recently, with the increase in processing power, researchers have started investigating the role of spectral phase in improving the speech quality, e.g., in [109, 110, 111]. Also, it was demonstrated through extensive experimentations in [27] that, given the STFT overlap is increased a bit, the performance of amplitude estimators can be improved to some extent when combined with less noisy spectral phases. A comprehensive discussion on the topic of spectral phase estimation can be found in [112].

All in all, we believe that the accurate estimation of speech spectral phase is still an open problem and further research in this direction deserves to be undertaken in the future [112]. For this reason, we restrict ourselves to using simple estimates for $\hat{\omega}$ in the proposed estimators in (4.15) and (4.19). In this sense, in the multi-channel case with spatially uncorrelated noise, averaging schemes can be done on the delay-compensated noisy phase of the observations in different channels, i.e., $\Omega_n + \phi_n$. Also, the following MMSE-optimal multi-channel spectral phase estimate derived in [40] was found to be efficient

$$\tan(\hat{\omega}) = \frac{\sum\limits_{n=1}^{N} \sqrt{\frac{\zeta_n}{\sigma_{v_n}^2}}\ \Im\{Y_n e^{j\phi_n}\}}{\sum\limits_{n=1}^{N} \sqrt{\frac{\zeta_n}{\sigma_{v_n}^2}}\ \Re\{Y_n e^{j\phi_n}\}} \tag{4.20}$$

where $\Re\{.\}$ and $\Im\{.\}$ denote the real and imaginary parts, respectively. Clearly, this estimator of the spectral phase reduces to the noisy phase of the observation in the single-channel case, which still can be regarded as a reasonable estimate for the speech spectral phase.

4.3.3 Performance Evaluation in Spatially Uncorrelated Noise

In this section, we investigate the performance of the proposed multi-channel STSA estimation methods in Sections 4.3.1 and 4.3.2. Various types of noise from the NOISEX-92 database [104] were considered for the evaluations. Yet, due to the consistency in the obtained results, we present those for the white, babble and car noises. Clean speech utterances including 10 male and 10 female speakers are used from the TIMIT speech database [105]. The sampling rate was set to 16 kHz and a Hamming window with length 20 ms and overlap of 75% is used for the STFT analysis and synthesis. In all simulations, the noise variance is estimated by the soft-decision IMCRA method [80] and the decision-directed approach [17] is used to estimate the *a priori* SNR. Unless otherwise stated, the number of microphones is considered to be $N = 2$.

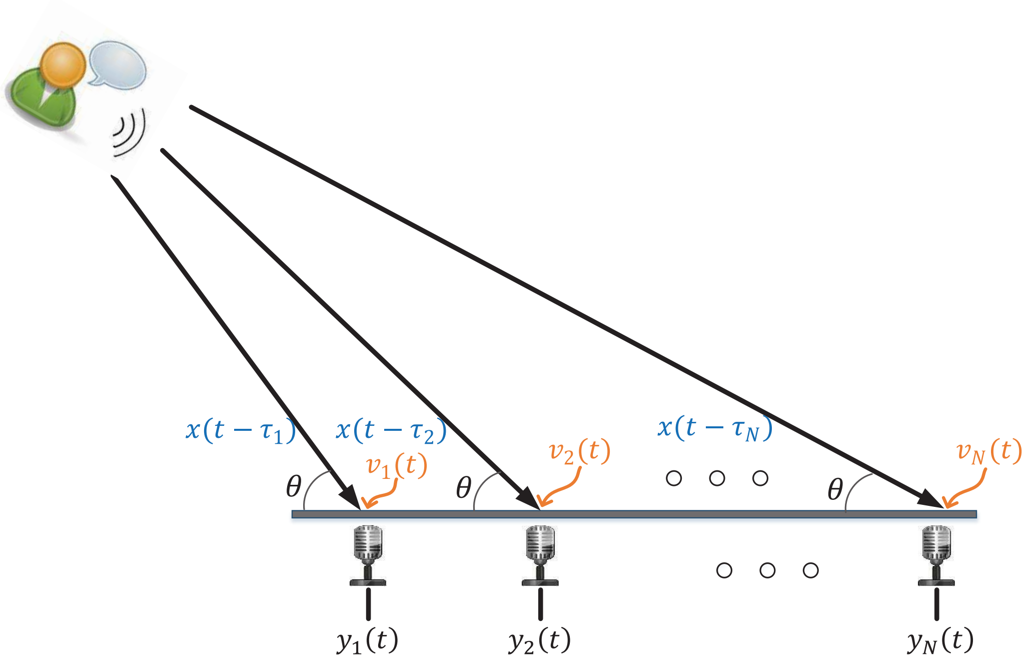


Figure 4.1: Scenario of capturing a far field source of speech in spatially uncorrelated noise by a linear microphone array.

We considered a speech source located in the far field impinging on a linear microphone array, as illustrated in Figure 4.1. The far field assumption implies the same angle of arrival, θ, with respect to all microphones, which is assumed to be known as $\theta = 70°$. The latter assumption is equivalent to knowing the steering vector $\mathbf{\Phi}(k)$ in (4.3). Therefore, the observation at each microphone consists of a delayed version of the speech source plus a noise component that is independent in different microphones. To generate noisy speech signals with uncorrelated noise across microphones, considering an inter-microphone distance of 10 cm, we time delayed the reference clean speech and added independent noises at desired SNR values in the range of -10 dB to 10 dB.

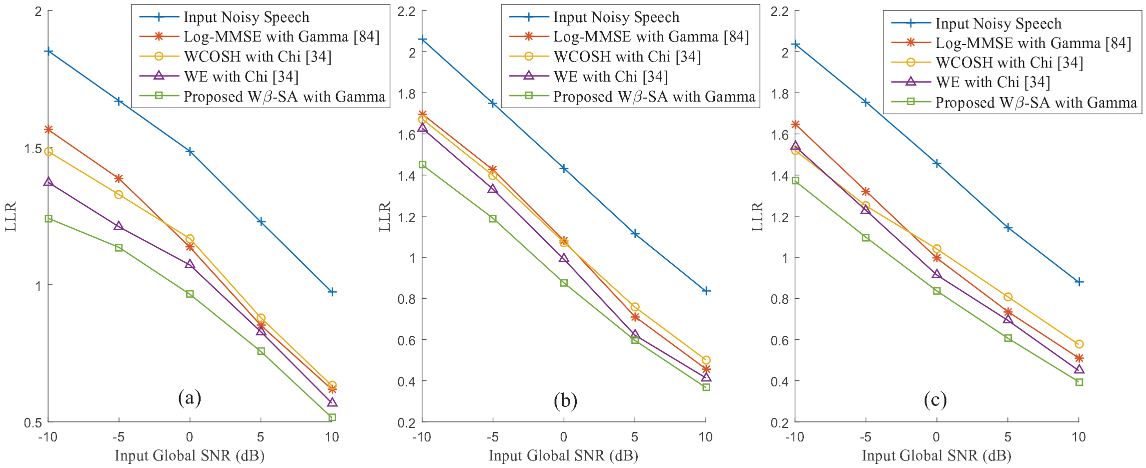


Figure 4.2: LLR versus input global SNR for the multi-channel STSA estimators with $N = 2$ microphones in spatially uncorrelated noise, (a): white noise, (b): babble noise and (c): car noise.

To evaluate the performance of the multi-channel extension of the proposed Wβ-SA estimator presented in Section 4.3.1, following the same trend in Chapter 3, we compare the multi-channel Wβ-SA estimator in (4.10) with the multi-channel modification of the recent STSA estimators with super-Gaussian priors in [34, 85]. Figures 4.2-4.4 are indicative of the performance scores of the considered estimators versus the input global SNR for different noise types, wherein the advantage of the proposed estimator in (4.10) can be observed through higher PESQ and segmental SNR and smaller LLR values. Also, as compared to Figures 3.12-3.14 in Chapter 3, i.e., the corresponding curves for the single-channel estimators, superior performance is seen to be provided by the multi-channel estimators. Furthermore, to have a clear assessment of the amount of improvement obtained by employing more microphones, the performance scores using the proposed multi-channel estimator in (4.10) for different microphone numbers in babble noise have been illustrated in Figure 4.5. It is observed that, especially for a lower number of microphones, there is considerable improvement in the enhanced speech with increasing the microphone numbers.

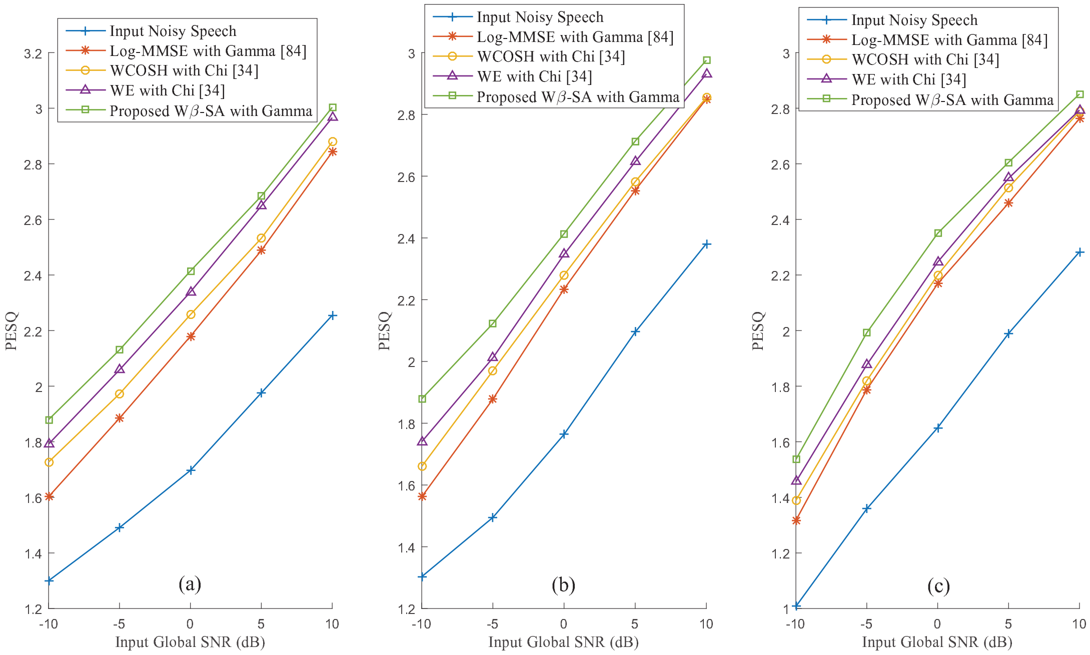


Figure 4.3: PESQ versus input global SNR for the multi-channel STSA estimators with $N = 2$ microphones in spatially uncorrelated noise, (a): white noise, (b): babble noise and (c): car noise.

Next, we evaluate the performance of the STSA estimators using spectral phase proposed in Section 4.3.2. In this sense, we consider both the MMSE and Wβ-SA estimators using phase, expressed respectively by (4.15) and (4.19), and compare their performance to their conventional counterparts, i.e., the phase independent MMSE and Wβ-SA amplitude estimators discussed in Chapter 2. We employ the same schemes as those proposed in Chapter 3 for the parameter setting of the Wβ-SA estimator.

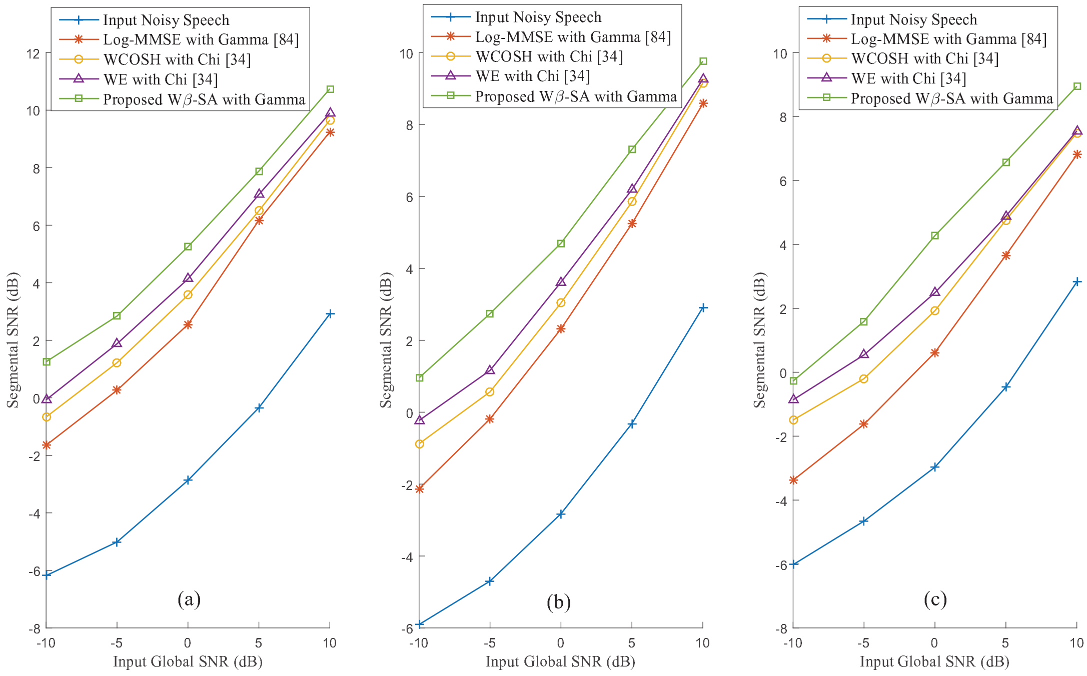


Figure 4.4: SNRSeg versus input global SNR for the multi-channel STSA estimators with $N = 2$ microphones in spatially uncorrelated noise, (a): white noise, (b): babble noise and (c): car noise.

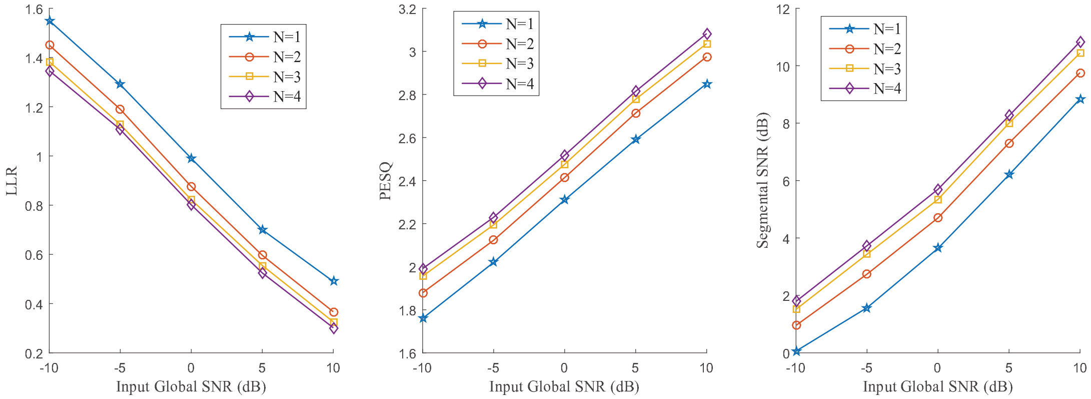


Figure 4.5: Performance scores of the proposed GGD-based Wβ-SA estimator in (4.10) for different microphone numbers in babble noise.

In Figures 4.6-4.9, the performance scores have been shown for the aforementioned estimators with $N = 1$ and $N = 2$ microphones in case of babble noise. In practice, babble noise has proved to be one of the most challenging noise types, as it often occupies the same range of spectrum as the clean speech. The results obtained by using other types of noise were also mostly consistent. As observed, while the Wβ-SA estimator outperforms the MMSE, the spectral phase-based versions of both estimators provide considerably superior performance with respect to their conventional variations. This performance advantage is even more visible in the lower range of the input SNR, where the use of speech phase information in the estimation of its amplitude results in higher improvements.

To investigate the effect of the accuracy of the underlying spectral phase estimate, $\hat{\omega}$, in the phase-based STSA estimation, we experimentally studied the cases where the noisy phase, Ω, the MMSE phase estimate by (4.20), and the phase of the clean speech, ω, are exploited for $\hat{\omega}$. The performance scores are indicated in Figure 4.9 for the phase-based Wβ-SA estimator with $N = 4$. It is observed that, whereas using a better estimate of the phase, i.e., that given by (4.20), leads to an improvement with respect to using the noisy phase, employment of the perfect speech phase in the STSA estimation provides considerable enhancement in the speech quality. This empirically proves the potential of the amplitude estimation to provide even superior performance in noise reduction, given more accurate estimates of the speech phase.

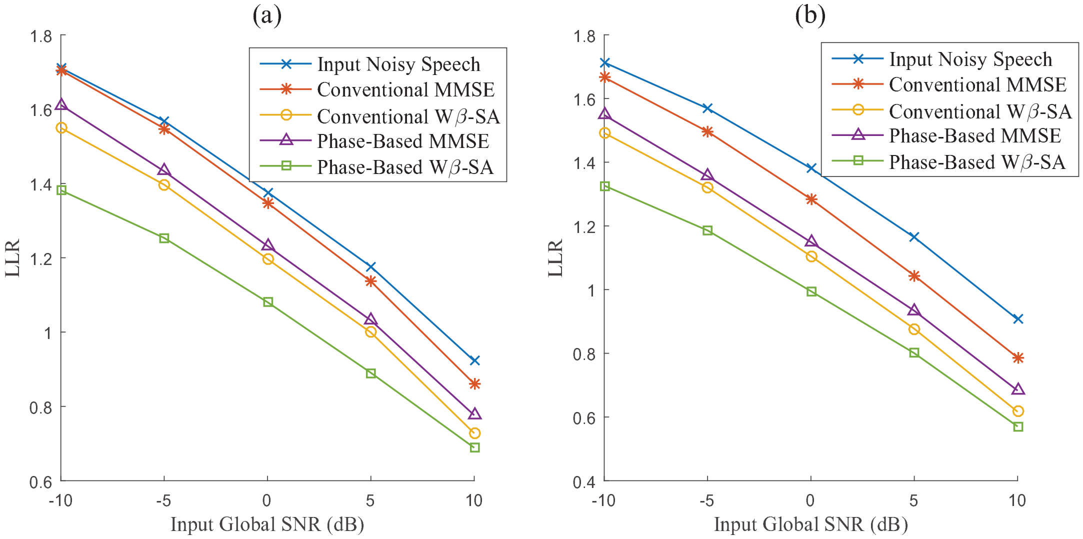


Figure 4.6: LLR for the conventional and spectral phase-based STSA estimators in babble noise with (a): $N = 1$ and (b): $N = 2$ microphones.

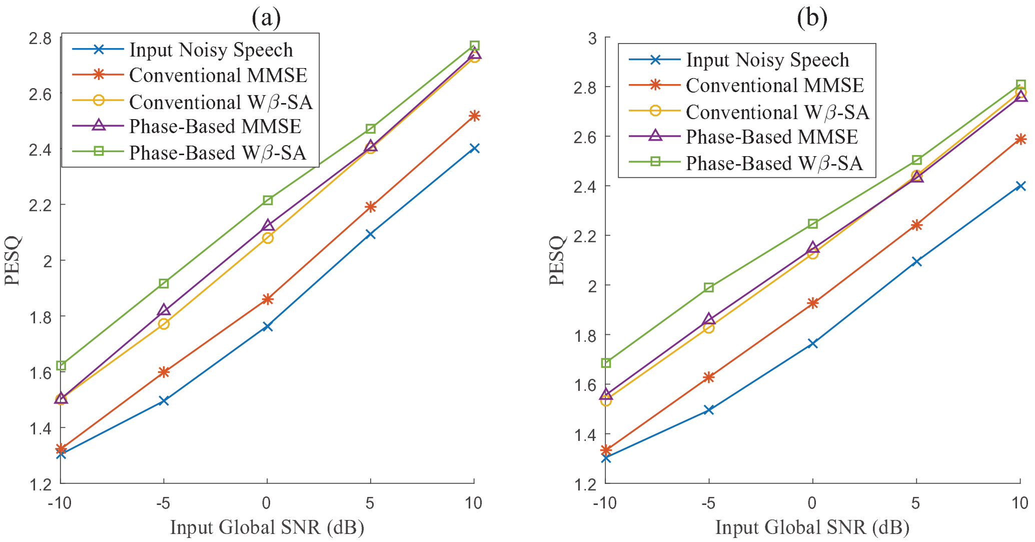


Figure 4.7: PESQ for the conventional and spectral phase-based STSA estimators in babble noise with (a): $N = 1$ and (b): $N = 2$ microphones.

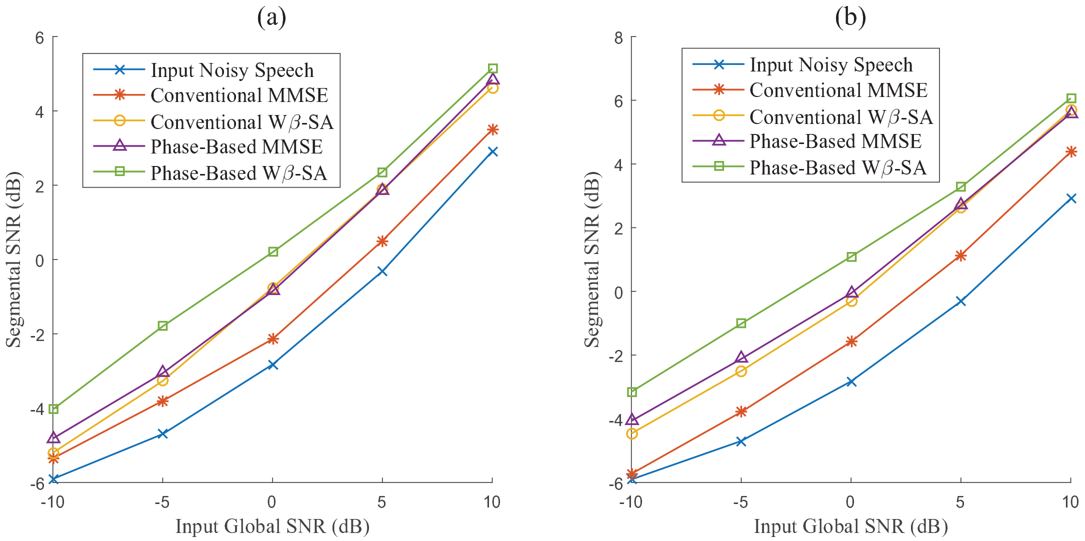


Figure 4.8: Segmental SNR for the conventional and spectral phase-based STSA estimators in babble noise with (a): $N = 1$ and (b): $N = 2$ microphones.

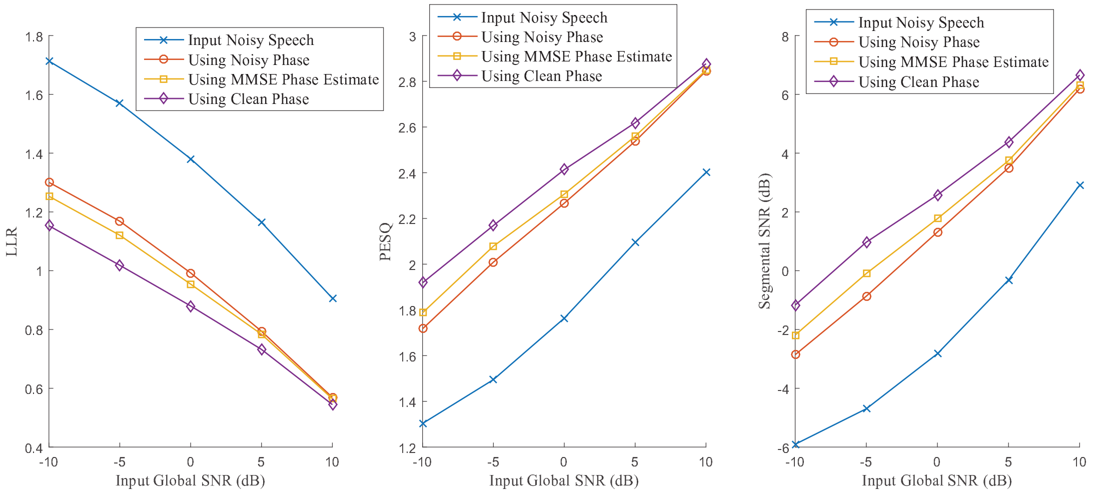


Figure 4.9: Performance of the spectral phase-based Wβ-SA estimator in (4.19) in babble noise with $N = 4$, using the noisy phase, using the MMSE estimate of phase in (4.20), and using the phase of the clean speech.

4.4 Multi-Channel STSA Estimation in Spatially Correlated Noise

In this section, we investigate the multi-channel STSA estimation method in the general case of spatially correlated noise. It should be noted that multi-channel STSA estimation in spatially uncorrelated noise, as discussed in the previous section, can be thought of as a special case; Yet, due to the simplicity of the derived expressions and their similarity to the single-channel case, it was preferred to study them as a standalone solution.

4.4.1 Extension of STSA Estimation to the Multi-Channel Case Under Known DOA and Noise PSD Matrix

In this subsection, we study the extension of the single-channel Bayesian STSA estimation method to the multi-channel case. It will be revealed that the resulting multi-channel estimator in spatially correlated noise can be expressed in a general framework that requires the knowledge of the DOA of the speech source as well as the noise PSD matrix across microphones.

For the sake of conciseness, it is noted that we follow the same assumptions and notation explained by (4.1)-(4.3) but we discard the assumption of spatially uncorrelated noise as by (4.4), i.e., we consider $E\{V_nV_m\} \neq E\{V_n\}E\{V_m\}$. To derive an STSA estimator using a Bayesian cost function in the multi-channel case, it is required to obtain the moment term $E\{f(\mathcal{X})|\mathbf{Y}\}$ with $f(\mathcal{X})$ as some function of $\mathcal{X}$. In this regard, in a similar fashion to (4.6), it follows that

$$E\left\{f(\mathcal{X})|\mathbf{Y}\right\} = \frac{\int_0^\infty f(\mathcal{X})p(\mathcal{X})\left[\int_0^{2\pi} p(\mathbf{Y}|\mathcal{X},\omega)d_\omega\right]d_\mathcal{X}}{\int_0^\infty p(\mathcal{X})\left[\int_0^{2\pi} p(\mathbf{Y}|\mathcal{X},\omega)d_\omega\right]d_\mathcal{X}} \tag{4.21}$$

where we used the uniform distribution for the spectral phase, ω. In [39], by the manipulation of the resulting integrals in (4.21), a direct solution for the STSA estimation in the case of the MMSE cost function, i.e., when $f(\mathcal{X}) = \mathcal{X}$ has been derived. However, handling the resulting integrations in (4.21) is not an easy task for other choices of the underlying Bayesian cost function and may not be tractable. Therefore, instead of looking for a direct solution for (4.21), we take an indirect approach as follows. Since the noise is assumed to be correlated across channels, based on a Gaussian noise assumption as before, it is deduced from (4.3) that the distribution $p(\mathbf{Y}|\mathcal{X},\omega)$ in (4.21) follows a complex multi-variate Gaussian PDF with zero mean and noise PSD matrix $\Sigma_{\mathbf{VV}} = E\{\mathbf{V}\mathbf{V}^H\}$, as the following

$$p(\mathbf{Y}|\mathcal{X},\omega) = \frac{1}{\pi^N \det\{\Sigma_{\mathbf{VV}}\}} \exp\left(-(\mathbf{Y} - \mathcal{X}e^{j\omega}\mathbf{\Phi})^H \Sigma_{\mathbf{VV}}^{-1}(\mathbf{Y} - \mathcal{X}e^{j\omega}\mathbf{\Phi})\right) \tag{4.22}$$

where det{.} denotes the matrix determinant and $\mathbf{\Phi}$ is the array steering vector as in (4.3). We now consider the internal integral in (4.21) over ω and its dependence on the observation vector $\mathbf{Y}$. By inserting (4.22) into (4.21), it can be deduced that the conditional expectation in (4.21) is a function of the observation vector $\mathbf{Y}$ only through the scalar term $\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}$, denoted by $Q(\mathbf{Y})$ (see Appendix B for a proof of this), namely,

$$E\left\{f(\mathcal{X})|\mathbf{Y}\right\} = E\left\{f(\mathcal{X})|Q(\mathbf{Y})\right\} \tag{4.23}$$

Therefore, $Q(\mathbf{Y})$ can be recognized as a sufficient statistic for the observation vector $\mathbf{Y}$, regardless of the underlying Bayesian cost function [113]. Accordingly, it is inferred that, under the multi-variate Gaussian PDF for the noise vector $\mathbf{V}$, a multi-channel STSA estimator can actually be thought of as an equivalent single-channel estimator with the noisy observation to be $Q(\mathbf{Y})$. To obtain more convenient expressions in the sequel, we rewrite $Q(\mathbf{Y})$ as

$$Q(\mathbf{Y}) = \left(\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \boldsymbol{\Phi}\right) S(\mathbf{Y}) \tag{4.24}$$

where, obviously, $S(\mathbf{Y})$ can also be regarded as a sufficient statistic for this problem. Thus, taking the scalar observation, $S(\mathbf{Y})$, as the input to the equivalent single-channel STSA estimator, we can write

$$S(\mathbf{Y}) = \frac{\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \mathbf{Y}}{\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \boldsymbol{\Phi}} = X + \frac{\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \mathbf{V}}{\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \boldsymbol{\Phi}} \tag{4.25}$$

As seen in (4.25), the observation $S(\mathbf{Y})$ consists of the same speech signal component X as that in (4.3) and a corresponding noise component. It is straightforward to show that the variance of this noise component is $\dfrac{1}{\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \boldsymbol{\Phi}}$. In summary, the following framework can be used for multichannel Bayesian STSA estimation under a Gaussian distribution for the noise:

- Provide the noisy array observations $\mathbf{Y}$, and estimates of the steering vector $\boldsymbol{\Phi}$ and noise PSD matrix $\Sigma_{\mathbf{VV}}$.

- Obtain the sufficient statistic $S(\mathbf{Y})$ in (4.25) as the scalar observation for an equivalent single-channel Bayesian STSA estimator.

- Perform the corresponding single-channel STSA estimation with $Y' = S(\mathbf{Y})$ as the input observation, and $\sigma_{v'}^2 = \dfrac{1}{\boldsymbol{\Phi}^H \Sigma_{\mathbf{VV}}^{-1} \boldsymbol{\Phi}}$ as the noise PSD.

In fact, the sufficient statistic term $S(\mathbf{Y})$ can be interpreted as the MVDR beamformer [46] acting here as the spatial processor and the equivalent single-channel STSA estimator acts as a post-filtering scheme on the output of the beamformer. Figure 4.10 shows a schematic of the framework for the multi-channel Bayesian STSA estimation discussed in this section, which consists of the concatenation of an MVDR beamformer and a modified single-channel STSA estimator as a post-filter.

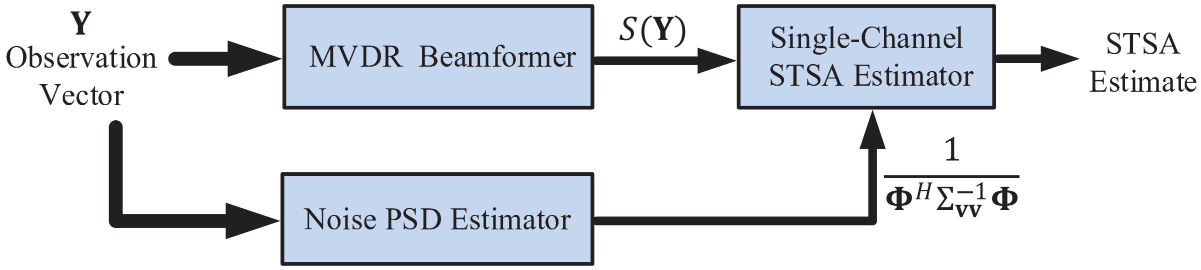


Figure 4.10: Block diagram of the proposed general scheme for multi-channel Bayesian STSA estimation.

It is evident that, to implement the approach in Figure 4.10, the steering vector $\mathbf{\Phi}$ and the noise PSD matrix $\Sigma_{\mathbf{VV}}$ must be known. As for the estimation of the steering vector or the speech DOA, since speech is typically a wide-band signal in its bandwidth, any wide-band DOA estimation method with moderate complexity can be used. In this sense, numerous wide-band DOA estimation approaches have been suggested and investigated in the field of array processing, e.g., method of cross correlation, broadband MUSIC and the eigenvalue decomposition algorithms [46]. However, contrary to the estimation of noise PSD where the literature is so rich, the estimation of noise PSD matrix in a general temporally/spatially non-stationary environment with no prior knowledge about the speech/environment is still a challenging problem and the current literature has received far less attention in this direction [114]. Therefore, in the following section, we tend to focus on the estimation of the spatial PSD matrix of noise, $\Sigma_{\mathbf{VV}}$, as one of the main topics of the current chapter.

4.4.2 Estimation of Noise PSD Matrix

In recent years, considerable research has been directed toward the estimation of the noise PSD matrix. In this regard, due to the popularity of the groundbreaking method of minimum statistics (MS) for noise PSD estimation proposed by Martin [115], a few straightforward extensions of this method to noise PSD matrix estimation have been developed in the literature. In [116], a two-channel noise PSD estimator has been suggested by combining the MS method and a voice activity detector (VAD). However, the VAD-based noise estimation techniques are not capable of providing as much accuracy as the soft-decision methods, due to the lack of noise PSD updating during frames where the speech component is present [80]. In [117], an MS-based method to estimate the noise PSD matrix has been proposed by using the recursive smoothing of noisy speech through a fixed forgetting factor. However, as proved in the context of single-channel noise estimation, selecting the forgetting factor independently for each frame/frequency can largely enhance the noise estimation accuracy. In [118], an algorithm for the estimation of the noise PSD matrix has been suggested by employing an adaptive forgetting factor selected based on the multi-channel speech presence probability (SPP). However, the SPP employed in [118] is obtained under a two-hypotheses basis assuming either the presence or the absence of speech in all channels, which is not accurately true due to the difference among the speech/noise components in each channel. Another recent method has been proposed in [114] where it is attempted to eliminate the undesirable speech component while estimating the noise PSD matrix. Nevertheless, due to employing the conventional fixed smoothing in its structure, it results in trivial improvements at moderate SNRs.

In this subsection, we present a new algorithm for the estimation of the noise PSD matrix, as needed by the multi-channel STSA estimator in the previous section, and in general, by many multi-microphone speech enhancement methods such as beamforming. The proposed algorithm does not require the knowledge of speech DOA and is applicable in a generic non-stationary noisy environment. In the proposed approach, rather than only relying on previous time frames, we make use of subsequent speech frames in order to achieve a more efficient smoothing scheme on noisy observations.

4.4.2.1 Incorporation of Subsequent Speech Frames

All prior solutions to the noise estimation problem include recursive smoothing schemes using the current and past noisy speech frames. This is due to the need for ensemble averaging implied by the statistical expectation, $E\{.\}$. In this sense, to make use of all the available information, we suggest to take advantage of several subsequent (future) speech frames in the recursive smoothing performed for the noise PSD estimation. On this basis, we propose the following weighted recursive smoothing scheme for the estimation of the noise PSD matrix

$$\begin{aligned}\mathrm{P}(k,l) = \xi\ \kappa(k,l)\mathrm{P}(k,l-1) + [1-\kappa(k,l)]\,\mathbf{Y}(k,l)\mathbf{Y}^H(k,l)\\ +(1-\xi)\ \kappa(k,l)\sum_{i=1}^{d} w_i\mathbf{Y}(k,l+i)\mathbf{Y}^H(k,l+i)\end{aligned} \tag{4.26}$$

where $\mathrm{P}(k,l)$ is the smoothed noisy spectrum, $\kappa(k,l)$ is the forgetting factor in smoothing the past frames, ξ is the smoothing parameter used to determine the weighting between the past and future frames and w_i are the weighting scheme applied on the d future frames. It should be noted that the exploitation of d future frames in the noise estimation for current speech frame implies a certain processing delay. Yet, due to the practical range of d, say $d \leq 5$, and the overlap between consecutive frames, the amount of delay is negligible as it is smaller than a few decades of milliseconds only. As for the weighting parameter ξ, an experimentally fixed value of 0.65 has worked best in the tested scenarios, which gives more emphasis to the numerous past frames. The selection of $\kappa(k,l)$ will be discussed in Section 4.4.2.2. As for the weightings w_i, we consider a fixed exponential scheme as $w_i = \gamma^i$, noting that the conventional recursive smoothing performed on past frames results in an exponential scheme for its weightings (as eq. (13) in [115]). Given this and the fact that $\sum_{i=1}^{d} w_i = 1$, we end up with the following expression in terms of γ exponent

$$\gamma^{d+1} - 2\gamma + 1 = 0, \quad \text{for a selected } d \tag{4.27}$$

It should be noted that for small d values, (4.26) has exactly one real-valued positive solution that makes it possible to use γ^i as a proper weighting.

4.4.2.2 Iterative Method for the Selection of the Forgetting Factor

In spite of the high importance in the selection of the forgetting factor, $\kappa(k,l)$, the literature on the noise PSD matrix estimation lacks efficient schemes for this purpose. We herein take into account the fact that, in the recursive smoothing of noisy speech, a larger weight should be assigned to the update term when the speech component is weaker (or equivalently the noise component is stronger) and vice versa [80]. To this end, we suggest to measure the speech signal intensity in all channels by the following definition of the overall SNR

$$\bar{\zeta}(k,l) \triangleq \frac{\left\|\Sigma_{\mathbf{XX}}(k,l)\right\|_2}{\left\|\Sigma_{\mathbf{VV}}(k,l)\right\|_2} = \frac{\left\|\Sigma_{\mathbf{YY}}(k,l) - \Sigma_{\mathbf{VV}}(k,l)\right\|_2}{\left\|\Sigma_{\mathbf{VV}}(k,l)\right\|_2} \tag{4.28}$$

where $\Sigma_{\mathbf{XX}}$ denotes the speech PSD matrix, with $\mathbf{X}(k,l)$ as $X(k,l)\boldsymbol{\Phi}(k)$, and $\|.\|_2$ indicates the ℓ_2-norm of a matrix. The equation at the right of (4.28) holds due to the uncorrelated speech and noise components. Based on this measure of the SNR, we propose to select the forgetting factor as

$$\kappa(k,l) = \kappa_{min} + (\kappa_{max} - \kappa_{min})\,\widetilde{\zeta}(k,l) \tag{4.29}$$

with κ_{min} and κ_{max} as the fixed minimum and maximum values for $\kappa(k,l)$ chosen as 0.25 and 0.94, respectively, and $\widetilde{\zeta}(k,l)$ is the thresholded and normalized $\bar{\zeta}(k,l)$, which is given by

$$\widetilde{\zeta}(k,l) = \begin{cases} 1, & \text{if } \bar{\zeta}(k,l) \geq \tau_H \\ \frac{\bar{\zeta}(k,l) - \tau_L}{\tau_H - \tau_L}, & \text{if } \tau_L < \bar{\zeta}(k,l) < \tau_H \\ 0, & \text{otherwise} \end{cases} \tag{4.30}$$

with the high and low thresholds $\tau_H = 22$ and $\tau_L = 0.35$, in respect. Now to implement (4.28), proper estimates of $\Sigma_{\mathbf{YY}}(k,l)$ and $\Sigma_{\mathbf{VV}}(k,l)$ are required. The PSD matrix of noisy speech, $\Sigma_{\mathbf{YY}}(k,l)$, can be simply estimated for our purpose through the recursive smoothing of the noisy observations $\mathbf{Y}$. However, an estimate of $\Sigma_{\mathbf{VV}}(k,l)$ is not available. Therefore, we propose the following iterative algorithm to estimate $\kappa(k,l)$ in (4.29):

---

(1) Replace $\Sigma_{\mathbf{VV}}(k,l)$ in (4.28) by $\mathrm{P}(k,l-1)$ and calculate $\bar{\zeta}(k,l)$.

(2) Calculate $\widetilde{\zeta}(k,l)$ using (4.30).

(3) Calculate $\kappa(k,l)$ using (4.29).

(4) Use $\kappa(k,l)$ to obtain $\mathrm{P}(k,l)$ in (4.26)

(5) Replace $\Sigma_{\mathbf{VV}}(k,l)$ in (4.28) by $\mathrm{P}(k,l)$ and calculate $\bar{\zeta}(k,l)$.

(6) Continue the next steps from **(2)**.

---

where $\mathrm{P}(k,l)$ is in fact the estimate for the noise PSD matrix at the end of each iteration. As for the first frame, assuming that there is no speech component present, $\kappa(k,1)$ is chosen as κ_{min} and then $\mathrm{P}(k,1)$ is calculated. In all of the experimentations, we found that using only two iterations of the above was sufficient and no considerable improvements were obtained by using more iterations.

4.4.2.3 Minimum Tracking and Bias Compensation

We here employ an extension of the minimum tracking method [115] to further improve the accuracy of noise PSD estimation. To this end, we track the minimum norm of the noise PSD matrix estimate, i.e., $\mathrm{P}(k,l)$, across the current and last M–1 frames. Therefore, we define $\mathrm{P}_{min}(k,l)$ as the matrix with minimum ℓ_2-norm on the set $\{\mathrm{P}(k,l), \mathrm{P}(k,l-1), \cdots, \mathrm{P}(k,l-M+1)\}$. Yet, as stated in [115], $\mathrm{P}_{min}(k,l)$ is biased toward lower values and the bias needs to be compensated. Based on the statistics of the minimum tracking, this bias has been estimated in [115] for the case of noise PSD estimation. However, the problem becomes theoretically too tedious when dealing with noise PSD matrix estimation. For this reason, considering that the bias is linearly dependent on the number of frames, M, as evident in eq. (17) in [115], we found the following approximation to the inherent bias in $\mathrm{P}_{min}(k,l)$ to be useful

$$B_{min} \approx 1 + \frac{M-1}{2} \tag{4.31}$$

Now by multiplying the minimum tracked value, $\mathrm{P}_{min}(k,l)$, by its bias in the above, we obtain the ultimate estimate for the noise PSD matrix as

$$\widehat{\Sigma}_{\mathbf{VV}}(k,l) = B_{min}\mathrm{P}_{min}(k,l) \tag{4.32}$$

The value of $\widehat{\Sigma}_{\mathbf{VV}}(k,l)$ given by the above is to be used as the proposed estimate for the noise spatial PSD matrix.

4.4.3 Performance Evaluation in Spatially Correlated Noise

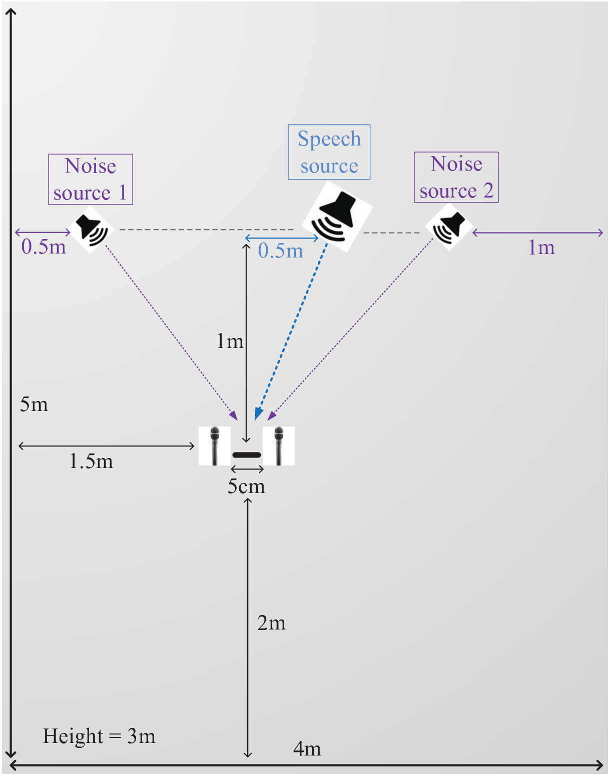


Figure 4.11: Scenario of capturing a speech source in spatially correlated noise with two microphones, generated by the ISM method.

In this section, we evaluate the performance of the proposed methods in Sections 4.4.1 and 4.4.2, wherein the multi-channel noise reduction has been considered in the spatially correlated noise case. Here, in order to account for the features in a realistic environment, we used the image source method (ISM) in [119] to generate the observed microphone array signals. The ISM is a very well-known technique used to generate a synthetic room impulse response (RIR) between a speech source and an acoustic sensor in a given environment [120]. Once such an RIR is generated, the observed speech can be obtained by convolving the RIR with the clean speech signal. This technique has been widely used to evaluate the performance of various audio processing methods in the field of room acoustics and signal processing. In our case, we considered the geometry shown in Figure 4.11, where a source of clean speech and two sources of noise have been assumed to be located at the indicated positions.

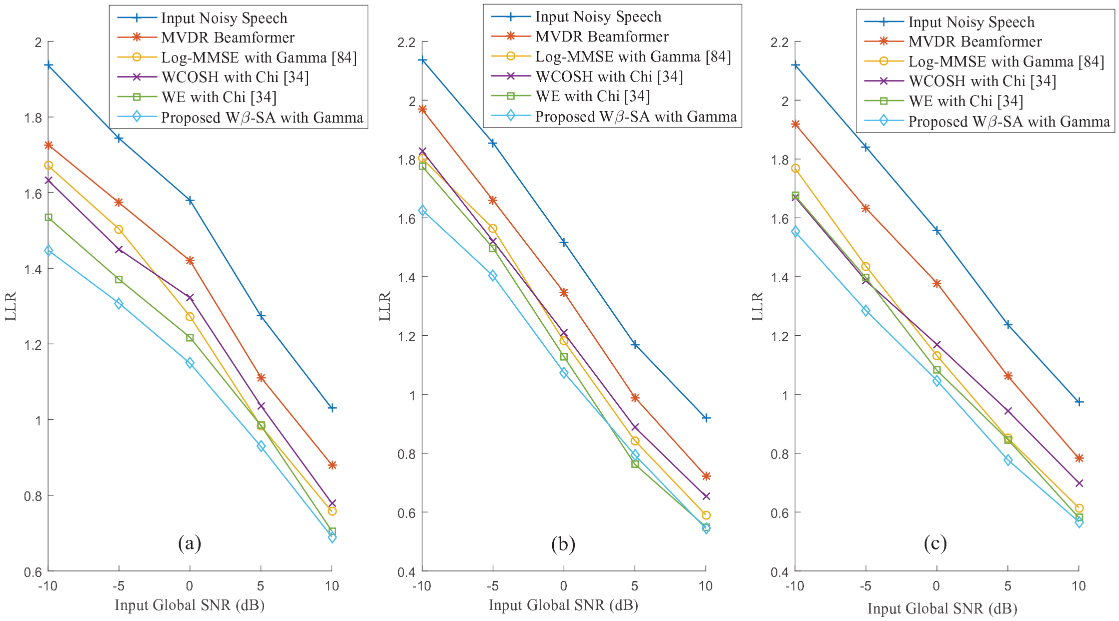


Figure 4.12: LLR versus input global SNR for multi-channel STSA estimators and MVDR beamformer with $N = 2$ microphones in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise.

As seen, we considered a linear set of microphone array with inter-sensor distance of 5 cm positioned in a 5m×4m×3m room with a reverberation time of 50 msec. The latter is actually too small for a highly reverberant environment (where the range of reverberation time is around 0.5-1 sec), yet we assumed such small reverberation time only to account for more realistic conditions compared to a noise-only environment. The RIRs between the source of speech/noise and the microphone array are obtained by the ISM method, then convolved with the audio samples of speech/noise (extracted from the same databases as in Section 4.3.3), and added together to generate the observed noisy speech.

We implemented the multi-channel STSA estimation framework in Section 4.4.1 for different STSA estimators, considering the known DOA for the speech source in Figure 4.11 and using the recursive smoothing of the noisy observations to calculate the noise PSD matrix.

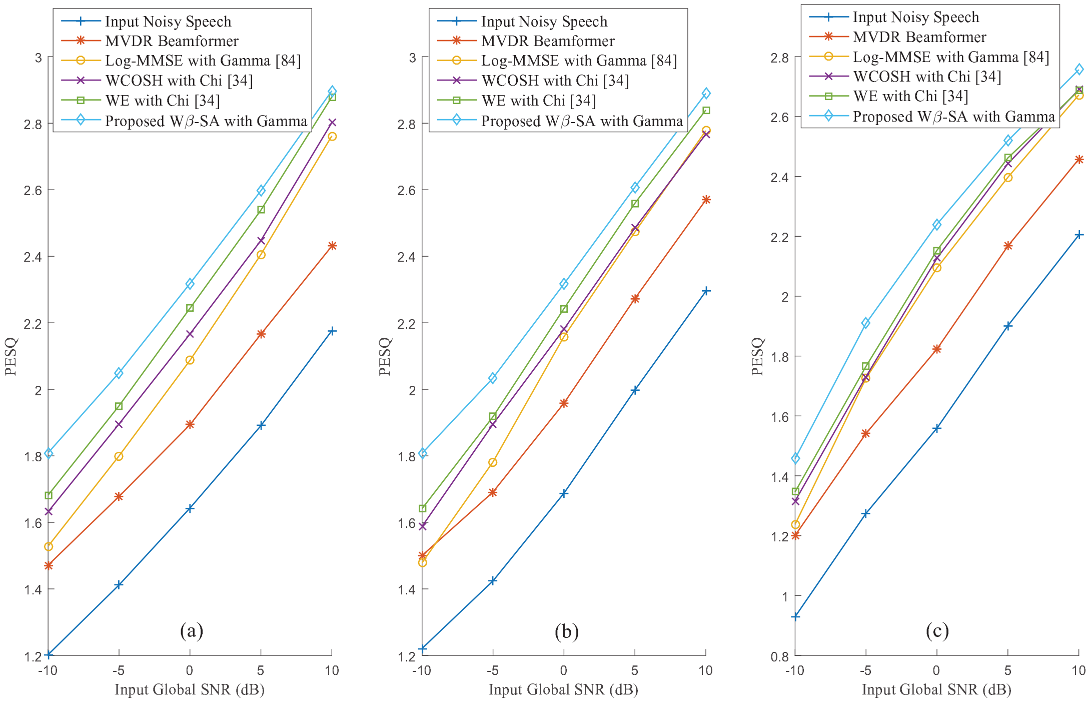


Figure 4.13: PESQ versus input global SNR for multi-channel STSA estimators and MVDR beamformer with $N = 2$ microphones in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise.

The latter is one of the most basic methods of noise PSD matrix estimation and has been widely used in the literature. In a fashion similar to Section 4.3.3, the performance scores have been illustrated for the STSA estimators with super-Gaussian priors and also the conventional MVDR beamformer in Figures 4.12-4.14. As observed, almost the same pattern as that in Section 4.3.3 holds, where the proposed GGD-based Wβ-SA estimator achieves superior performance with respect to the other STSA estimators. Also, all multi-channel STSA estimators outperform the conventional MVDR beamformer in the entire range of input SNR, which is reasonable due to their structure proposed in Section 4.4.1. This structure is in fact a post-processing stage applied on the output of the MVDR beamformer.

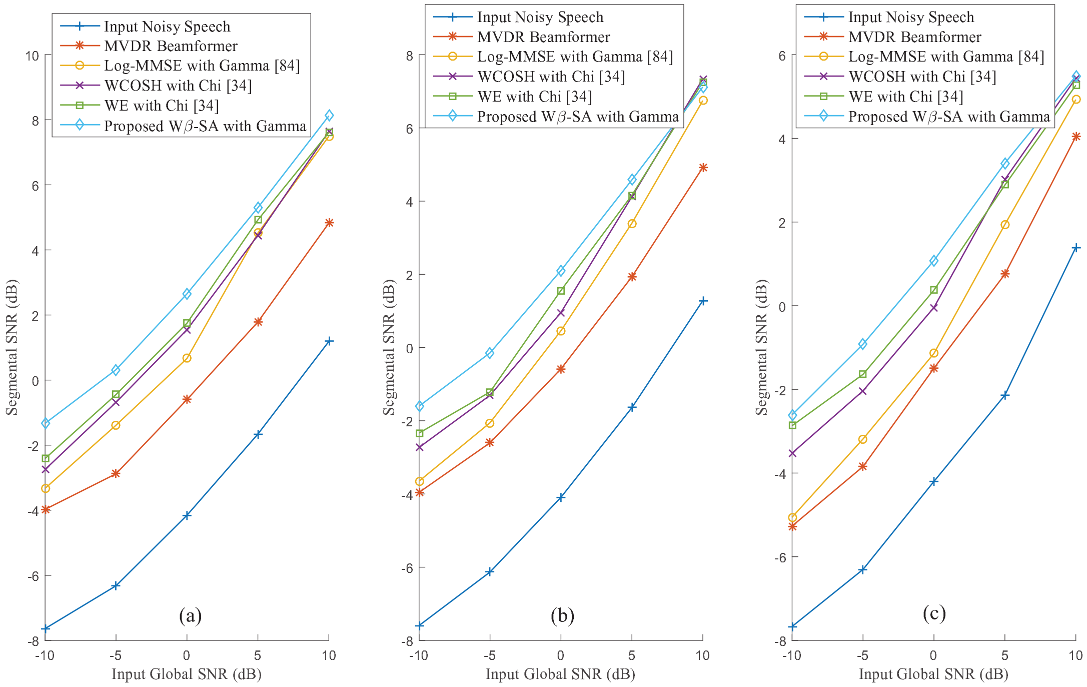


Figure 4.14: Segmental SNR versus input global SNR for multi-channel STSA estimators and MVDR beamformer with $N = 2$ microphones in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise.

Next, we investigate the performance of the proposed noise PSD matrix estimation approach in Section 4.4.2 with respect to other recent methods in the same area. In all simulations, the number of subsequent frames considered in the smoothing was assumed to be d=3 implying that γ=0.5437 in (4.27). Even though small improvements were obtainable by increasing d up to 5−6, for the sake of comparable complexity burden, we kept d at 3. This also ensures that the imposed processing delay is not more than 15 msec in total. In order to focus on the relative performance of the noise PSD matrix estimation methods only, we consider the MVDR beamformer with the known DOA as in Figure 4.11 and evaluate the enhanced speech at the output of the beamformer. In this respect, we consider the method of Hendriks in [114], the SPP-based approach proposed in [118], as well as the conventional recursive smoothing of observations and the smoothing but by using the available noise-only samples. The latter can be in fact considered as a perfect noise estimation method implying an upper bound on how far a smoothing-based approach in noise PSD matrix estimation can be improved (all methods of noise estimation in fact use smoothing of the noisy observations).

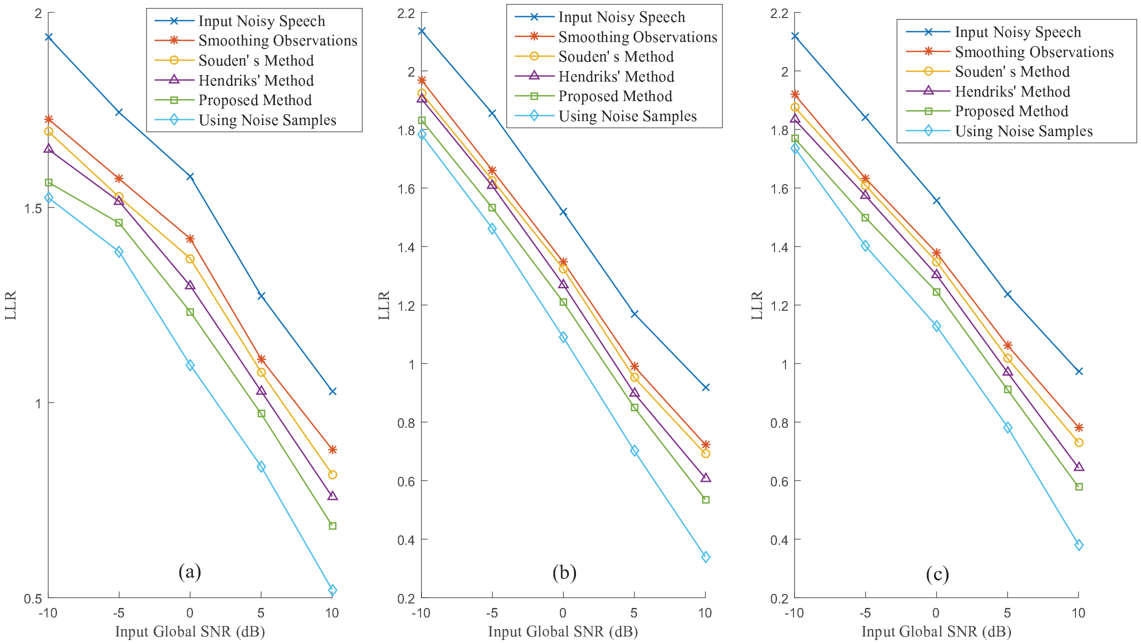


Figure 4.15: LLR versus input global SNR for the enhanced speech using the MVDR beamformer with different noise PSD matrix estimation methods in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise.

The performance measures for the aforementioned methods have been indicated in Figures 4.15-4.17. It is observable that the proposed algorithm outperforms the other three methods in the entire range of the input SNR by almost 0.1 in PESQ and 1∼2 dBs in segmental SNR, which are considerable improvements in the speech quality. Furthermore, despite the advantage of the proposed approach in Section 4.4.2, it is viewed that there still exists a large gap between the perfect method, i.e., that by using the noise samples, and the proposed approach, especially in the higher range of the input SNR. The reason is due to the presence of the strong speech components in the estimated elements across the noise PSD matrix in the soft-decision-based methods, which results in speech signal cancellation and unfavorable distortion in the MVDR output. Therefore, it can be concluded that there is still further room to develop more accurate noise PSD matrix estimation methods that are capable of suppressing the speech component present in the observed noisy speech. The latter has been one of the major challenges in all noise PSD matrix estimation methods proposed so far.

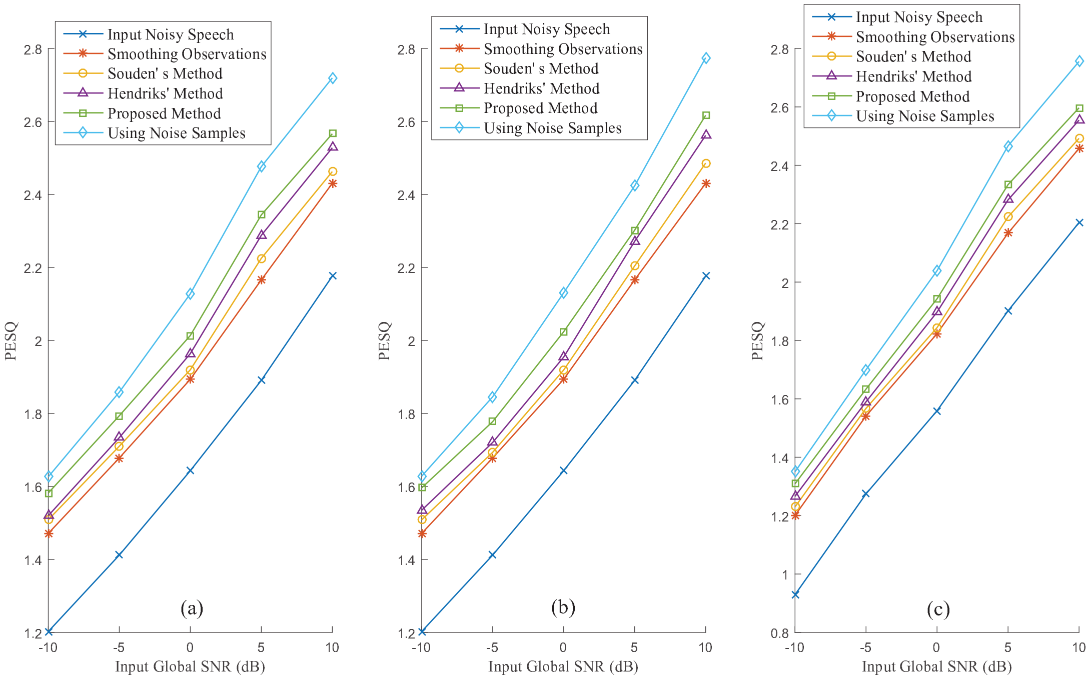


Figure 4.16: PESQ versus input global SNR for the enhanced speech using the MVDR beamformer with different noise PSD matrix estimation methods in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise.

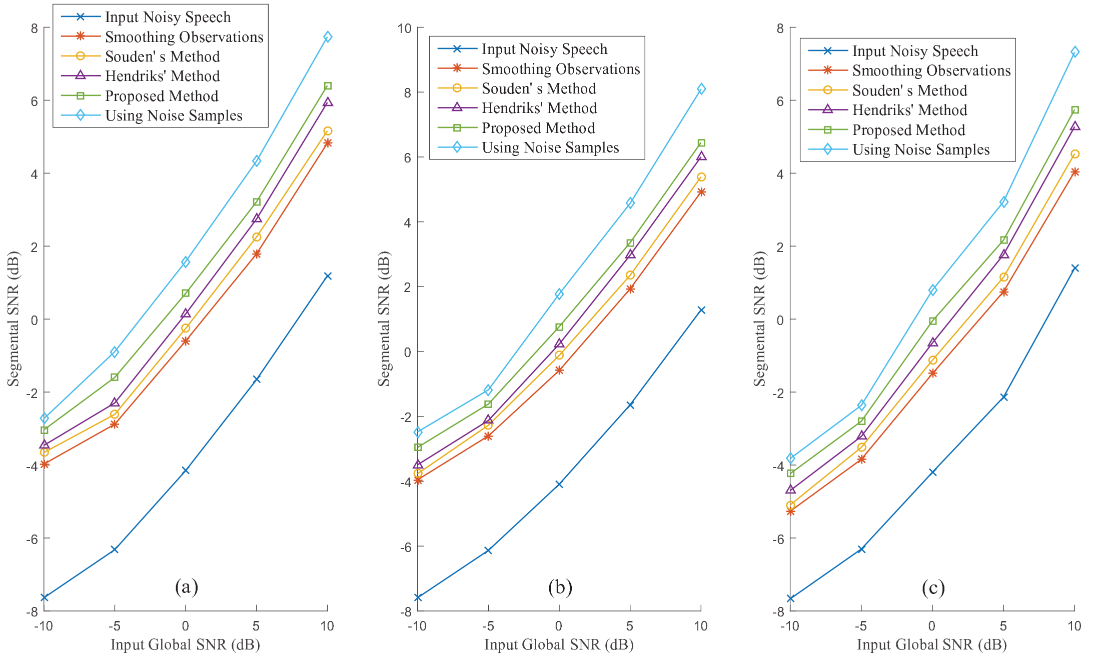


Figure 4.17: Segmental SNR versus input global SNR for the enhanced speech using the MVDR beamformer with different noise PSD matrix estimation methods in spatially correlated noise, (a): white noise, (b): babble noise and (c): car noise.

To further evaluate the performance of the noise PSD matrix estimation methods, we illustrate the MVDR beamformer response (output) errors in Figure 4.18, as suggested by equation (37) in [114]. This criterion in fact shows a measure of distance between the reference output obtained by using the noise-only samples and the outputs by using the noise PSD matrix estimation methods. Due to the smaller beamformer response error in the proposed method, it can be concluded that the proposed algorithm achieves an MVDR output closer to that obtained by the reference method. The same performance evaluations with respect to other types of non-stationary noise were also performed, confirming the superiority of the proposed algorithm in all scenarios.

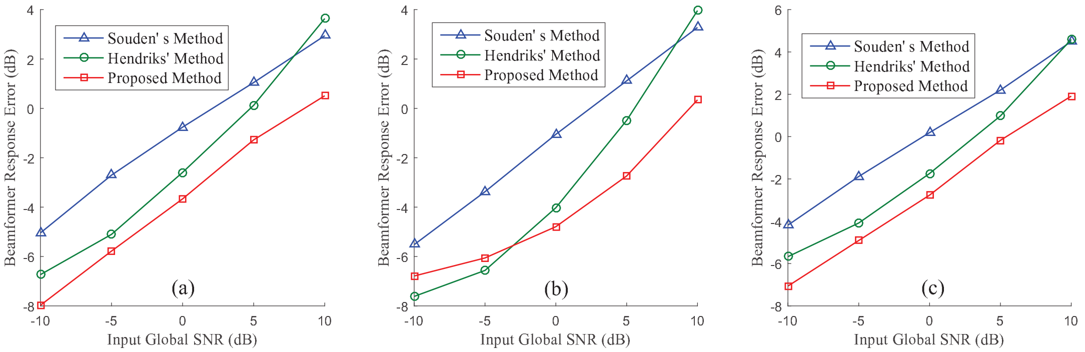


Figure 4.18: MVDR beamformer response error versus input global SNR using different noise PSD matrix estimation methods, (a): white noise, (b): babble noise and (c): car noise.

To have a complete evaluation of the proposed noise PSD matrix estimation, we further investigate its performance for a higher number of microphones using the same scenario as Figure 4.11 with an inter-microphone distance of 5 cm and the number of microphones as N=2−4. In Figure 4.19, the performance measures are illustrated for this scenario with babble noise. As observed, the performance consistently improves for a higher number of microphones, yet there appears to be a smaller improvement as N goes higher. This consequence, which was also observed with the other noise PSD matrix estimation methods, is due to the increment in the amount of speech distortion and signal cancellation, as a result of higher accumulated error in the estimation of more elements in larger noise PSD matrices.

Finally, we evaluate the performance of the proposed noise PSD matrix estimation method with respect to the number of subsequent frames, d, in Figure 4.20. Herein, we changed d from 1 to 5 in (4.27) and measured the performance of the MVDR beamformer using the babble noise. As observed, with an increasing d up to 3−4, there appear to be improvements in the speech quality, yet the improvements become so trivial and almost zero for higher values of d. This is because the smoothing of the subsequent frames, as performed by (4.26), assigns smaller and nearly zero weights to the farther subsequent frames. Therefore, we took d to be 3 for the rest of the experiments.

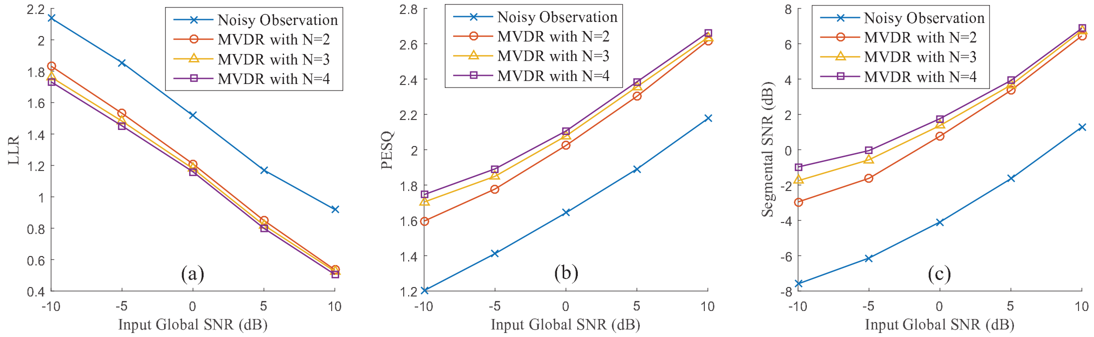


Figure 4.19: Performance measures of the MVDR beamformer for different microphone numbers using the proposed method of noise PSD matrix estimation in babble noise.

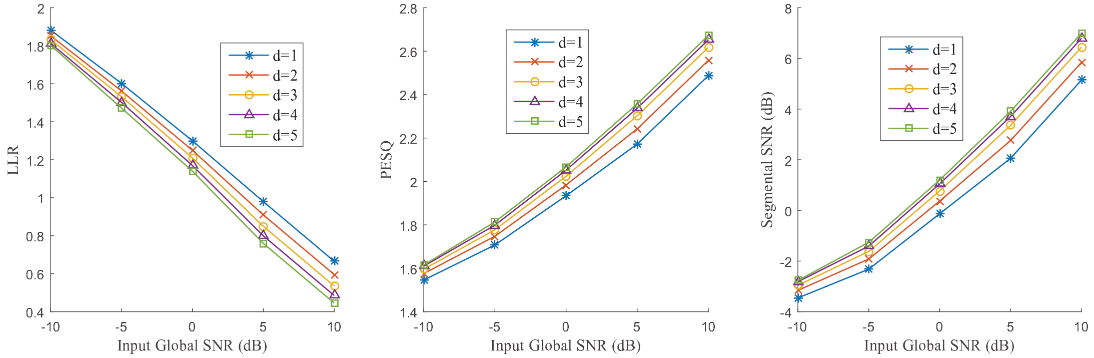


Figure 4.20: Performance measures of the MVDR beamformer using the proposed method of noise PSD matrix estimation with a different number of involved subsequent frames, d, in babble noise.

4.5 Conclusion

In this chapter, multiple aspects of noise reduction using the STSA estimation technique in multi-channel were investigated, including extensions of STSA estimators from single- to multi-channel in spatially uncorrelated/correlated cases, STSA estimation using spectral phase, and estimation of the noise PSD matrix.

First, we showed that the single-channel STSA estimation method can be extended to the case of multi-channel under both spatially correlated and spatially uncorrelated noisy environments. In this regard, the developed single-channel Wβ-SA estimator in Chapter 3 was extended to its multi-channel counterpart in Section 4.3.1 under a known DOA for the speech source, and the performance evaluations indicated its superiority with respect to the multi-channel version of the other recent STSA estimators.

In Section 4.3.2, the role of speech spectral phase in the estimation of the spectral amplitude, i.e., STSA, was studied. On this basis, MMSE and Wβ-SA estimators using spectral phase estimates were developed with closed-form solutions. Performance assessment of the phase-aware amplitude estimators revealed a considerable advantage over the conventional, i.e., phase independent, amplitude estimators, and furthermore, deduced the fact that further improvements are achievable given more accurate estimates of the spectral phase.

With regards to the spatially correlated noise, it was demonstrated in Section 4.4.1 that the multi-channel STSA estimator in fact turns into an MVDR beamformer and a modified single-channel STSA estimator as a post-filter, under a known or estimated speech DOA and noise PSD matrix estimate. In this respect, performance assessment of different multi-channel STSA estimators within the proposed framework proved their advantage compared to the MVDR beamforme, and in addition, the advantage of the Wβ-SA estimator with respect to the other estimators.

Finally, since the most crucial factor in the performance of the multi-channel STSA estimators, and generally, most beamforming methods such as the MVDR, is the estimation of the spatial noise PSD matrix, we tackled this problem in Section 4.4.2. Taking advantage of a few subsequent speech frames and the soft-decision MS method, we developed a generic approach to noise PSD matrix estimation in a non-stationary noise field. Performance evaluations were conducted by using the noise PSD matrix estimates obtained from the proposed approach and two recent approaches in the MVDR beamformer and the advantage of the proposed algorithm with respect to the previous two methods was confirmed. Also, it was revealed that further precision in the noise PSD matrix estimation can be reached in the future by properly eliminating the speech component from the noisy observations.

Chapter 5

Speech Dereverberation Using the Weighted Prediction Error Method

In this chapter, we target the problem of speech reverberation suppression, namely dereverberation, by using a well-known and efficient statistical model-based approach, i.e., the weighted prediction error (WPE) method. In the same line as the presented contributions in case of noise reduction, the WPE method is implemented in the STFT domain.

5.1 Introduction

One of the major categories of reverberation suppression methods is the model-based statistical approaches that offer optimal solutions to estimate the anechoic (reverberation-free) speech. In [121], probabilistic models of speech were incorporated into a variational Bayesian expectation-maximization algorithm which estimates the source signal, the acoustic channel and all the involved parameters in an iterative manner. A different strategy was followed in [122], where the parameters of an auto-regressive (AR) model for speech and reverberation model are iteratively determined by maximizing the likelihood function of the considered model parameters through an expectation-maximization (EM) approach. Therein, a minimum mean-squared error (MMSE) estimator is derived that yields the enhanced speech. Within the same line of work, using the time-varying statistical model for the speech and the multi-channel linear prediction (MCLP) model for reverberation has led to efficient dereverberation [123, 124]. Since the implementation of such methods in the time domain is computationally costly, it was proposed in [125, 126] to employ the MCLP-based method in the short-time Fourier transform (STFT) domain. The resulting approach, referred to as the weighted prediction error (WPE) method, is an iterative algorithm that alternatively estimates the reverberation prediction coefficients and speech spectral variance, using batch processing of the speech utterance.

The rest of this chapter is organized as follows. Section 5.2 gives a summary of the proposed methods in this chapter. A brief review of the original WPE method is presented in Section 5.3. Section 5.4 describes the proposed WPE method with the estimation of speech spectral variance. In Section 5.5, we discuss the WPE method using the modeling of the IFC, including the ML solution for the proposed estimator of the reverberation prediction weights and the suggested method for the estimation of the IFC. Performance assessment is presented in Section 5.6 and conclusions are drawn in Section 5.7.

5.2 Brief Description of the Proposed Methods

As seen in the previous section, the WPE method basically requires an estimate of the desired speech variance, $\sigma^2_{d_{n,k}}$, along with the reverberation prediction weights, $\mathbf{g}_k$, leading to a sub-optimal strategy that alternatively estimates each of the two quantities. Also, as observed in (5.5), the desired speech component is assumed to be temporally (across all STFT time frames) independent, while this assumption is not accurate due to the high correlation present within speech frames. In this chapter, we develop new WPE-based methods in order to overcome the two aforementioned drawbacks.

Considering the estimation of the unknown speech spectral variance in the original WPE method, we introduce a suitable estimator for the speech spectral variance and integrate it into the ML solution for the reverberation prediction weights. Specifically, this task is accomplished by resorting to the reverberation suppression within the spectral enhancement literature [43] and employing the statistical model-based estimation of late reverberant spectral variance (LRSV) [127] in order to estimate the speech spectral variance. In addition to the performance merit w.r.t. the previous WPE-based methods, the proposed approach offers a considerable gain in reducing the computational complexity.

With regards to the inherent temporal correlation in the desired speech, our major contribution is to model the correlation across STFT frames, namely the inter-frame correlation (IFC). Since an accurate modeling of the IFC is not tractable, we consider an approximate model where only the frames within each segment of the speech are considered correlated. It is shown that, given an estimate of the IFC matrix, the proposed approach results in a convex quadratic optimization problem w.r.t. reverberation prediction weights, which is then solved by an ordinary optimization toolbox solver. Furthermore, an efficient method for the estimation of the underlying IFC matrix is developed based on the extension of a recently proposed speech variance estimator. We evaluate the performance of our approach incorporating the estimated correlation matrix and compare it to the original and several variations of the WPE method. The results reveal lower residual reverberation and higher overall quality provided by the proposed method.

The presented contribution in Section 5.4 has been published in [128] and the contribution in Section 5.5 has been submitted as [129].

5.3 Review on the WPE Method

In this section, we present a brief review of the original WPE method. Suppose that a single source of speech is captured by M microphones located in a reverberant enclosure. In the STFT domain, we denote the clean speech signal by $s_{n,k}$ with time frame index $n \in \{1, \ldots, N\}$ and frequency bin index $k \in \{1, \ldots, K\}$. Then, the reverberant speech signal observed at the m-th microphone, $x^m_{n,k}$, can be represented in the STFT domain using a linear prediction model as [126]

$$x^m_{n,k} = \sum_{l=0}^{L_h\text{-}1} \left(h^m_{l,k}\right)^* s_{n-l,k} + e^m_{n,k} \tag{5.1}$$

where $h^m_{l,k}$ is an approximation of the acoustic transfer function (ATF) between the speech source and the m-th microphone in the STFT domain with the length L_h, and $(.)^*$ denotes the complex conjugate. The additive term $e^m_{n,k}$ models the linear prediction error and the additive noise term, and is neglected here [126]. Therefore, (5.1) can be rewritten as

$$x^m_{n,k} = d^m_{n,k} + \sum_{l=D}^{L_h\text{-}1} \left(h^m_{l,k}\right)^* s_{n-l,k} \tag{5.2}$$

where $d^m_{n,k}{=}\sum_{l=0}^{D-1}\left(h^m_{l,k}\right)^* s_{n-l,k}$ is the sum of anechoic (direct-path) speech and early reflections at the m-th microphone, and D corresponds to the duration of the early reflections. Most dereverberation techniques, including the WPE method, aim at reconstructing $d_{n,k}$ as the desired signal, or suppressing the later reverberant terms denoted by the summation in (5.2). Replacing the convolutive model in (5.2) by an auto-regressive (AR) model results in the well-known multi-channel linear prediction (MCLP) form for the observation at the first microphone, i.e.,

$$d_{n,k} = x^1_{n,k} - \sum_{m=1}^{M} \left(\mathbf{g}^m_k\right)^{\mathrm{H}} \mathbf{x}^m_{n-D,k} = x^1_{n,k} - \mathbf{g}^{\mathrm{H}}_k \, \mathbf{x}_{n-D,k} \tag{5.3}$$

where $d_{n,k} \equiv d^1_{n,k}$ is the desired signal, $(.)^{\mathrm{H}}$ is the Hermitian transpose, and the vectors $\mathbf{x}^m_{n-D,k}$ and $\mathbf{g}^m_k$ are defined as

$$\begin{aligned} \mathbf{x}^m_{n-D,k} &= [x^m_{n-D,k}\ ,\ x^m_{n-D-1,k}\ ,\ \ldots,\ x^m_{n-D-(L_k-1),k}]^{\mathrm{T}} \\ \mathbf{g}^m_k &= [g^m_{0,k}\ ,\ g^m_{1,k}\ ,\ \ldots,\ g^m_{L_k-1,k}]^{\mathrm{T}} \end{aligned} \tag{5.4}$$

where $\mathbf{g}^m_k$ is the regression vector (reverberation prediction weights) of order L_k for the m-th channel. The right-hand side of (5.3) has been obtained by concatenating $\{\mathbf{x}^m_{n-D,k}\}$ and $\{\mathbf{g}^m_k\}$ over m to respectively form $\mathbf{x}_{n-D,k}$ and $\mathbf{g}_k$. Estimation of the regression vector $\mathbf{g}_k$ and using it in (5.3) gives an estimate of the desired (dereverberated) speech. From a statistical viewpoint, this is performed by using the maximum likelihood (ML) estimation of the desired speech $d_{n,k}$ at each frequency bin [126]. In this sense, the conventional WPE method [125, 126] assumes a circular complex Gaussian distribution for the desired speech coefficients, $d_{n,k}$, with time-varying spectral variance and zero mean. Now if $d_{n,k}$ is assumed to be independent across time frames, i.e., using zero inter-frame correlation (IFC), the joint distribution of the desired speech coefficients for all frames at frequency bin k is given by

$$p(\mathbf{d}_k) = \prod_{n=1}^{N} p(d_{n,k}) = \prod_{n=1}^{N} \frac{1}{\pi\sigma^2_{d_{n,k}}} \exp\left(-\frac{|d_{n,k}|^2}{\sigma^2_{d_{n,k}}}\right) \tag{5.5}$$

with $\sigma^2_{d_{n,k}}$ as the unknown time-varying spectral variance of the desired speech defined as $E\{|d_{n,k}|^2\}$. Now, by inserting $d_{n,k}$ from (5.3) into (5.5), we can see a set of unknown parameters at each frequency bin consisting of the regression vector, $\mathbf{g}_k$, and the desired speech spectral variances

$\sigma^2_{\mathbf{d}_k}$=$\{\sigma^2_{d_{1,k}}, \sigma^2_{d_{2,k}}, \ldots, \sigma^2_{d_{N,k}}\}$. Denoting this set by Θ_k=$\{\mathbf{g}_k, \sigma^2_{\mathbf{d}_k}\}$, and taking the negative of logarithm of $p(\mathbf{d}_k)$ in (5.5), the objective function for the parameter set Θ_k can be written as

Table 5.1: Outline of the steps in the conventional WPE method.

- At each frequency bin k, consider the speech observations $x^m_{n,k}$, for all n and m, and the set of parameters D, L_k and ϵ.
- Initialize $\sigma^2_{d_{n,k}}$ by $\sigma^{2^{(j)}}_{d_{n,k}}$=$|x_{n,k}|^2$ at j=0.
- Repeat the following over the iteration j until the convergence of $\mathbf{g}_k$ or maximum allowed iterations:

  $\mathbf{A}^{(j)}_k = \sum_{n=1}^{N} \mathbf{x}_{n-D,k}\ \mathbf{x}^H_{n-D,k}\ /\sigma^{2^{(j)}}_{d_{n,k}}$

  $\mathbf{a}^{(j)}_k = \sum_{n=1}^{N} \mathbf{x}_{n-D,k}\ x^{1^*}_{n,k}\ /\sigma^{2^{(j)}}_{d_{n,k}}$

  $\mathbf{g}^{(j)}_k = \mathbf{A}^{-1^{(j)}}_k \mathbf{a}^{(j)}_k$

  $r^{(j)}_{n,k} = \mathbf{g}^{(j)^H}_k \mathbf{x}_{n-D,k}$

  $d^{(j)}_{n,k} = x^1_{n,k} - r^{(j)}_{n,k}$

  $\sigma^{2^{(j+1)}}_{d_{n,k}} = \max\{|d^{(j)}_{n,k}|^2, \epsilon\}$

- $\mathbf{g}^{(j)}_k$ is the desired reverberation prediction weights after convergence.

$$\mathcal{J}(\Theta_k) = -\log\ p(\mathbf{d}_k|\Theta_k) = \sum_{n=1}^{N}\left(\log\ \sigma^2_{d_{n,k}} + \frac{\left|x^1_{n,k} - \mathbf{g}^{\mathrm{H}}_k\ \mathbf{x}_{n-D,k}\right|^2}{\sigma^2_{d_{n,k}}}\right) \tag{5.6}$$

where the constant terms have been discarded. To obtain the ML estimate of the parameter set Θ_k, (5.6) has to be minimized w.r.t. Θ_k. Since the optimization of (5.6) jointly w.r.t. $\mathbf{g}_k$ and $\sigma^2_{\mathbf{d}_k}$ is not mathematically tractable, an alternative sub-optimal solution is suggested in [125, 126] where a two-step optimization procedure is performed w.r.t. only one of the two parameter subsets $\mathbf{g}_k$ and $\sigma^2_{\mathbf{d}_k}$ at each step. The two-step approach is repeated iteratively until a convergence criterion is satisfied or a maximum number of iterations is reached. A step-by-step summary of the conventional WPE method is outlined in Table 5.1. Often in practice, 3 to 5 iterations lead to the best possible results [126], yet there is no guarantee or a widely accepted criterion on the convergence of the method. Furthermore, the instantaneous estimate of the desired speech variance, i.e. $|d_{n,k}^{(j)}|^2$ in the table, may lead to unreasonably small values that deteriorate the overall performance of the WPE method. The aforementioned disadvantages can be mitigated by employing a proper estimate of the spectral variance of desired speech, as will be explained in the following section.

5.4 WPE Method with the Estimation of Early Speech Variance

In this section, we propose an efficient estimator for the spectral variance of the desired speech, $\sigma^2_{d_{n,k}}$, based on the statistical modeling of the acoustical transfer function (ATF), and incorporate this estimator to the WPE dereverberation method. As seen in (5.1)-(5.2), the desired speech $d_{n,k}$ is in fact the sum of the first D delayed and weighted clean speech terms, $s_{n-l,k}$. In the context of statistical spectral enhancement methods [43], $d_{n,k}$ is often referred to as the early speech, as compared to the late reverberant speech given by the sums in (5.2) and (5.3). Therefore, the observation at the first microphone can be rewritten as

$$x^1_{n,k} = d_{n,k} + r_{n,k} \tag{5.7}$$

with $r_{n,k}$ denoting the late reverberant speech. Several methods are available in the spectral enhancement literature for the estimation of $\sigma^2_{d_{n,k}}$ in (5.7), such as the decision directed (DD) approach for signal-to-reverberant ratio (SRR) estimation [43]. Using the latter method, $\sigma^2_{d_{n,k}}$ can be obtained as the product of the estimated SRR, i.e., $\sigma^2_{d_{n,k}}/\sigma^2_{r_{n,k}}$, and an estimate of the late reverberant spectral variance, $\sigma^2_{r_{n,k}}$. However, the application of conventional spectral enhancement techniques, originally developed for noise reduction purposes, is based on the assumption of independence between $d_{n,k}$ and $r_{n,k}$. Here, however, contrary to the scenario of additive noise, as evidenced from the model in (5.1) and (5.2), the early and late reverberant terms are basically correlated, due to the temporal correlation across successive time frames of the speech signal. Therefore, the non-zero correlation between $d_{n,k}$ and $r_{n,k}$ must be taken into account. Doing so, it follows from (5.7) that

$$\sigma^2_{x^1_{n,k}} = \sigma^2_{d_{n,k}} + \sigma^2_{r_{n,k}} + 2E\big\{\Re\{d_{n,k}\; r^*_{n,k}\}\big\} \tag{5.8}$$

with $\Re\{.\}$ denoting the real value and $2E\big\{\Re\{d_{n,k}\; r^*_{n,k}\}\big\}$ representing the non-zero cross-correlation terms between $d_{n,k}$ and $r_{n,k}$. Nevertheless, the estimation of the cross-correlation terms in (5.9), due to their dependency on the phases of $d_{n,k}$ and $r_{n,k}$, may not be analytically tractable.

In [130], a spectral subtraction algorithm for noise suppression has been proposed based on the deterministic estimation of speech magnitudes in terms of observation and noise magnitudes without assuming that they are independent. Therein, the authors consider the following problem similar to (5.8):

$$\left|x^1_{n,k}\right|^2 = \left|d_{n,k}\right|^2 + \left|r_{n,k}\right|^2 + 2\left|d_{n,k}\right|\left|r_{n,k}\right|\cos\left(\theta_{d_{n,k}} - \theta_{r_{n,k}}\right) \tag{5.9}$$

where $|d_{n,k}|$ is to be estimated in terms of $|x^1_{n,k}|$ and $|r_{n,k}|$, and $\theta_{d_{n,k}}$ and $\theta_{r_{n,k}}$ are the unknown phases of $d_{n,k}$ and $r_{n,k}$ respectively. Through a geometric approach, the following estimate of $|d_{n,k}|$ is then obtained as

$$\left|\hat{d}_{n,k}\right| = \sqrt{\frac{1 - \frac{(\gamma-\xi+1)^2}{4\gamma}}{1 - \frac{(\gamma-\xi-1)^2}{4\xi}}}\;\left|x^1_{n,k}\right| \tag{5.10}$$

where the two parameters ξ and γ are defined as

$$\xi_{n,k} \triangleq \frac{\left|d_{n,k}\right|^2}{\left|r_{n,k}\right|^2} \quad , \quad \gamma_{n,k} \triangleq \frac{\left|x^1_{n,k}\right|^2}{\left|r_{n,k}\right|^2} \tag{5.11}$$

Herein, we propose to employ this approach in order to provide a correlation-aware estimate of $|d_{n,k}|$, to be exploited in turn in the estimation of $\sigma^2_{d_{n,k}}$.

Due to the unavailability of $\left|d_{n,k}\right|^2$ and $\left|r_{n,k}\right|^2$, the two parameters in (5.11) are not known *a priori* and have to be substituted by their approximations. To this end, we exploit $|\hat{d}_{n-1,k}|^2$ for $\left|d_{n,k}\right|^2$ and a short-term estimate of $\sigma^2_{r_{n,k}}$ for $\left|r_{n,k}\right|^2$. To determine the latter, we resort to the statistical model-based estimation of the LRSV, which has been widely used in the spectral enhancement literature. Therein, an estimate of this key parameter is derived using a statistical model for the ATF along with recursive smoothing schemes. In brief, the following scheme is conventionally used to estimate the LRSV [127]:

$$\sigma^2_{x^1_{n,k}} = (1-\beta)\ \sigma^2_{x^1_{n-1,k}} + \beta \left|x^1_{n,k}\right|^2 \tag{5.12a}$$

$$\sigma^2_{\tilde{r}_{n,k}} = (1-\kappa)\ \sigma^2_{\tilde{r}_{n-1,k}} + \kappa\ \sigma^2_{x^1_{n-1,k}} \tag{5.12b}$$

$$\sigma^2_{r_{n,k}} = e^{-2\alpha_k R N_e}\ \sigma^2_{\tilde{r}_{n-(N_e-1),k}} \tag{5.12c}$$

where α_k is related to the 60 dB reverberation time, $T_{60dB,k}$, through $\alpha_k = 3\log 10/\left(T_{60dB,k} f_s\right)$ with f_s as the sampling frequency in Hz, R is the STFT time shift (hop size) in samples, β and κ are smoothing parameters (which can be in general frequency-dependent) and N_e is the delay parameter defining the number of assumed early speech frames, which is herein taken as D. This choice of N_e is made so that the number of previous frames considered as early speech in the LRSV estimation is equal to the number of included frames in the desired speech $d_{n,k}$ by the WPE method in (5.2). The term $\tilde{r}_{n,k}$ actually represents the entire reverberant speech including both the early and late reverberant speech, but excluding the direct-path. Using the LRSV estimator in (5.12), the short-term estimate of $\sigma^2_{r_{n,k}}$ is obtained by choosing the smoothing parameters β and κ to be close to one. By this choice, the estimate of $\sigma^2_{r_{n,k}}$ is updated faster, and will therefore be closer to the true value of $|r_{n,k}|^2$. Yet, to avoid unreasonably small values for the approximated $|r_{n,k}|^2$ in the denominator of (5.11), this parameter is lower bounded to 10^{-3}.

Now, given the estimate of early speech magnitude, $|\hat{d}_{n,k}|$, provided by (5.10), it is simple to use a recursive smoothing scheme to estimate $\sigma^2_{d_{n,k}}$, as the following

$$\hat{\sigma}^2_{d_{n,k}} = (1-\eta)\,\hat{\sigma}^2_{d_{n-1,k}} + \eta\ |\hat{d}_{n,k}|^2 \tag{5.13}$$

with η as a fixed smoothing parameter. This estimate of $\sigma^2_{d_{n,k}}$ can be efficiently integrated into the original WPE method discussed in Section 5.3, replacing the instantaneous estimate given by $|d^{(j)}_{n,k}|^2$ in Table 5.1. By doing so, the objective function in (5.6) turns into a function of only the regression vector, $\mathbf{g}_k$, and it is therefore possible to obtain the latter as $\mathbf{A}_k^{-1}\mathbf{a}_k$ in Table 5.1, without the need for an iterative strategy.

5.5 WPE Method Using the Inter-Frame Correlations

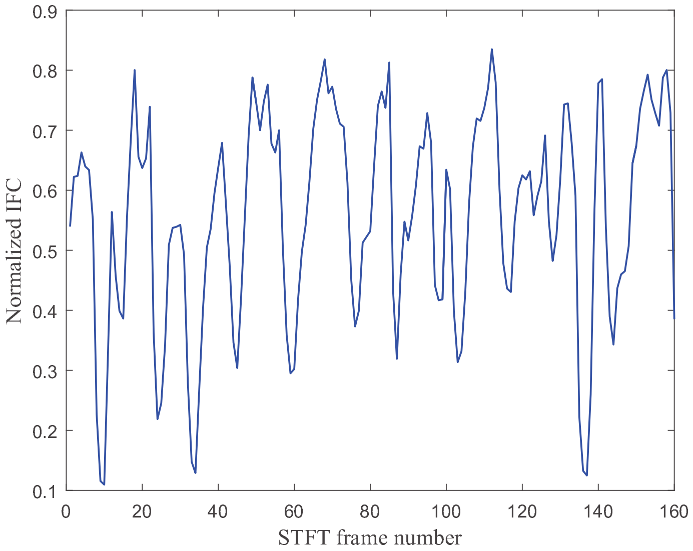


Figure 5.1: Normalized IFC of the early speech $d_{n,k}$ averaged over frequency bins versus STFT frame number for a selected speech utterance.

To demonstrate the importance of the temporal correlation in the desired early speech component, $d_{n,k}$, across STFT frames, which is the main motivation to develop the WPE method using IFC in this work, we have illustrated in Figure 5.1 the IFC present in the early speech for a given frame lag. To generate this figure, we extracted the early part, i.e., the first 60 msec, of a room impulse response (RIR) with 60 dB reverberation time T_{60dB}=800 msec, and then convolved it with the anechoic speech utterance to obtain the early speech $d_{n,k}$[1]. Next, the IFC measure $|E\{d_{n,k}d^*_{n-l,k}\}|$ was estimated through time averaging (i.e., long-term recursive smoothing) of the product $d_{n,k}d^*_{n-l,k}$ over n and then normalized by the estimated value of $E\{|d_{n,k}|^2\}$. The plotted values are the average over all frequency bins and have been obtained for the lag of l=3. As observed from Figure 5.1, the amount of correlation between the early speech components $d_{n,k}$ and $d_{n-l,k}$ is considerable as compared to the spectral variance $E\{|d_{n,k}|^2\}$. Whereas this correlation is neglected in earlier versions of the WPE method, the method that we here propose takes this correlation into account by jointly modeling the early speech terms. From Figure 5.1, it is also observed that, even though the updating rate of the underlying smoothing is not high, the estimated IFC fluctuates rapidly

[1] Note that, considering D=3 early terms and using a frame length of 40 msec with 50% overlap, the early speech component corresponds to the first 60 msec of the RIR.

across frames. Therefore, an efficient approach with fast convergence should be devised for its estimation.

In the following section, we first derive a solution for the reverberation prediction vector $\mathbf{g}_k$ by considering the IFC, in contrast to the model in (5.5). Next, based on an extension of the method proposed for the estimation of the speech spectral variance in Section 5.4, an approach for the estimation of the IFC matrix of the desired speech terms, as required by the derived solution, will be developed.

5.5.1 Proposed ML Solution

Considering the joint distribution of the desired speech STFT coefficients and assuming the independence across frequency bins, the temporally/spectrally independent model in (5.5) should be replaced by

$$p(\mathbf{d}_k) = p(d_{1,k}) \prod_{n=2}^{N} p(d_{n,k}|\mathbf{D}_{n,k}) \tag{5.14}$$

with $p(d_{n,k}|\mathbf{D}_{n,k})$ denoting the distribution of $d_{n,k}$ conditioned on $\mathbf{D}_{n,k} = [d_{n-1,k}, d_{n-2,k}, \cdots, d_{1,k}]^T$. Considering the fact that $d_{n,k}$ depends only on a limited number of the speech coefficients from previous frames, or equivalently, the fact that the IFC length is finite, (5.14) can be written as

$$p(\mathbf{d}_k) = p(d_{1,k}) \prod_{n=2}^{N} p(d_{n,k}|\mathbf{d}'_{n-1,k}) = p(d_{1,k}) \prod_{n=2}^{N} \frac{p(d_{n,k}, \mathbf{d}'_{n-1,k})}{p(\mathbf{d}'_{n-1,k})} \tag{5.15}$$

where the conditioning term $\mathbf{D}_{n,k}$ in (5.14) has been replaced by the shorter segment $\mathbf{d}'_{n-1,k} = [d_{n-1,k}, d_{n-2,k}, \cdots, d_{n-\tau_k,k}]^T$ with τ_k as the assumed IFC length in frames. Unfortunately, proceeding with the model in (5.15) to find an ML solution for the regression vector $\mathbf{g}_k$ does not lead to a convex optimization problem. Therefore, to overcome this limitation, we alternatively exploit an approximate model by considering only the correlations among the frames within each segment, $\mathbf{d}'_{n,k}{=}[d_{n,k}, d_{n-1,k}, \cdots, d_{n-\tau_k+1,k}]^T$, and disregarding the correlations across the segments. This results in the following approximate model

$$p(\mathbf{d}_k) \simeq \prod_{n=1}^{\left\lfloor \frac{N}{\tau_k} \right\rfloor} p\left(\mathbf{d}'_{n,k}\right) = \prod_{n=1}^{\left\lfloor \frac{N}{\tau_k} \right\rfloor} \frac{1}{\pi^{\tau_k} \det \mathbf{\Phi}_{n,k}} \exp\left(-\mathbf{d}'^{H}_{n,k} \mathbf{\Phi}^{-1}_{n,k} \mathbf{d}'_{n,k}\right) \tag{5.16}$$

where $\mathbf{\Phi}_{n,k}$=$E\{\mathbf{d}'_{n,k}\mathbf{d}'^{H}_{n,k}\}$ represents the correlation matrix of $\mathbf{d}'_{n,k}$, det denotes the determinant of a matrix and $\lfloor . \rfloor$ is the floor function. Now, using (5.3), the desired speech segment $\mathbf{d}'_{n,k}$ can be expressed as

$$\mathbf{d}'_{n,k} = \mathbf{u}_{n,k} - \mathbf{U}^{H}_{n,k}\ \mathbf{h}_k \tag{5.17}$$

where

$$\begin{aligned} \mathbf{u}_{n,k} =& [x^{(1)}_{n,k}, x^{(1)}_{n-1,k}, \cdots, x^{(1)}_{n-\tau_k+1,k}]^T \\ \mathbf{U}_{n,k} =& [\mathbf{X}_{n,k}, \mathbf{X}_{n-1,k}, \cdots, \mathbf{X}_{n-\tau_k+1,k}]^* \\ \mathbf{h}_k =& \mathbf{g}^*_k \end{aligned} \tag{5.18}$$

In the same manner as the original WPE method [126], by considering the negative of the logarithm of $p(\mathbf{d}_k|\mathbf{h}_k)$, an ML-based objective function for the regression weight vector $\mathbf{h}_k$ can be derived as follows,

$$\mathcal{J}(\mathbf{h}_k) \triangleq -\log\ p(\mathbf{d}_k|\mathbf{h}_k) = \sum_{n=1}^{\left\lfloor \frac{N}{\tau_k} \right\rfloor} \left(\mathbf{d}'^{H}_{n,k} \mathbf{\Phi}^{-1}_{n,k} \mathbf{d}'_{n,k} + \mathcal{K}_{n,k} \right) \tag{5.19}$$

with $\mathcal{K}_{n,k}$ representing the terms independent of $\mathbf{h}_k$, which can be discarded. Inserting (5.17) into (5.19) and doing further manipulation result in

$$\mathcal{J}(\mathbf{h}_k) = \sum_{n=1}^{\left\lfloor \frac{N}{\tau_k} \right\rfloor} \left(\mathbf{h}^{H}_k \mathbf{A}_{n,k} \mathbf{h}_k - \mathbf{b}^{H}_{n,k} \mathbf{h}_k - \mathbf{h}^{H}_k \mathbf{b}_{n,k} + c_{n,k} \right) \tag{5.20}$$

where we defined

$$\begin{aligned} \mathbf{A}_{n,k} &= \mathbf{U}_{n,k}\ \mathbf{\Phi}^{-1}_{n,k}\ \mathbf{U}^{H}_{n,k} \\ \mathbf{b}_{n,k} &= \mathbf{U}_{n,k}\ \mathbf{\Phi}^{-1}_{n,k}\ \mathbf{u}_{n,k} \\ c_{n,k} &= \mathbf{u}^{H}_{n,k}\ \mathbf{\Phi}^{-1}_{n,k}\ \mathbf{u}_{n,k} \end{aligned} \tag{5.21}$$

Now by neglecting the constant term $c_{n,k}$, (5.20) can be arranged as

$$\mathcal{J}(\mathbf{h}_k) = \mathbf{h}^{H}_k \widetilde{\mathbf{A}}_k \mathbf{h}_k - \widetilde{\mathbf{b}}^{H}_k \mathbf{h}_k - \mathbf{h}^{H}_k \widetilde{\mathbf{b}}_k \tag{5.22}$$

with $\widetilde{\mathbf{A}}_k$ and $\widetilde{\mathbf{b}}_k$ as

$$\widetilde{\mathbf{A}}_k = \sum_{n=1}^{\left\lfloor \frac{N}{\tau_k} \right\rfloor} \mathbf{A}_{n,k} \ , \qquad \widetilde{\mathbf{b}}_k = \sum_{n=1}^{\left\lfloor \frac{N}{\tau_k} \right\rfloor} \mathbf{b}_{n,k} \tag{5.23}$$

It can be shown that the matrix $\widetilde{\mathbf{A}}_k$ is positive semidefinite, and therefore, the quadratic objective function in (5.22) is real-valued and convex in terms of $\mathbf{h}_k$. Subsequently, to find the global minimum of $\mathcal{J}(\mathbf{h}_k)$, we can express (5.22) in the following form

$$\mathcal{J}(\mathbf{h}_k) = \left(\mathbf{h}_k - \widehat{\mathbf{h}}_k\right)^H \widetilde{\mathbf{A}}_k \left(\mathbf{h}_k - \widehat{\mathbf{h}}_k\right) + \mathrm{c}'_k \tag{5.24}$$

where c'_k is an independent term and

$$\widehat{\mathbf{h}}_k = \widetilde{\mathbf{A}}_k^{-1} \widetilde{\mathbf{b}}_k^H \tag{5.25}$$

It is evident that $\widehat{\mathbf{h}}_k$ in the above is the global minimum of the objective function $\mathcal{J}(\mathbf{h}_k)$ in (5.24), or equivalently, it is the estimate of the reverberation prediction weights by the proposed WPE method.

5.5.2 Estimation of the IFC Matrix

To calculate the optimal reverberation prediction weights by (5.25), $\widetilde{\mathbf{A}}_k$ and $\widetilde{\mathbf{b}}_k$ in (5.23), and in turn, $\mathbf{A}_{n,k}$ and $\mathbf{b}_{n,k}$ given by (5.21) have to be calculated. To do so, as seen in (5.21), the IFC matrix of the desired speech terms, $\mathbf{\Phi}_{n,k}$, has to be estimated beforehand. In Section 5.4, a new variant of the WPE method was suggested, which exploits the geometric spectral subtraction approach in [130] along with the estimation of LRSV, in order to estimate the spectral variance of the desired speech, $\sigma^2_{d_{n,k}}$. We here develop an extension of the proposed method in Section 5.4 to estimate the spectral cross-variances of the desired speech terms, $\rho_{n_1,n_2,k}{=}E\{d_{n_1,k} d^*_{n_2,k}\}$, which in fact constitute the IFC matrix $\mathbf{\Phi}_{n,k}$. In this regard, according to Section 5.4, the following estimate of $d_{n,k}$ can be obtained

$$\hat{d}_{n,k} = \sqrt{\frac{1 - \frac{(\gamma_{n,k} - \xi_{n,k} + 1)^2}{4\gamma_{n,k}}}{1 - \frac{(\gamma_{n,k} - \xi_{n,k} - 1)^2}{4\xi_{n,k}}}} \; x^{(1)}_{n,k} \tag{5.26}$$

We exploit (5.26) to provide primary estimates of $d_{n_1,k}$ and $d_{n_2,k}$ and then use recursive smoothing of $d_{n_1,k} d^*_{n_2,k}$ to estimate the elements of the IFC matrix $\mathbf{\Phi}_{n,k}$. In this sense, given the estimate of the desired speech components $\hat{d}_{n_1,k}$ and $\hat{d}_{n_2,k}$ by (5.26), it is straightforward to use a recursive smoothing scheme to estimate the spectral cross-variance $\rho_{n_1,n_2,k}$, as the following

$$\widehat{\rho}_{n_1,n_2,k} = (1-\eta)\,\widehat{\rho}_{(n_1-1),(n_2-1),k} + \eta\ \hat{d}_{n_1,k}\ \hat{d}^*_{n_2,k} \tag{5.27}$$

with η as a fixed smoothing parameter. Equivalently, by expressing (5.27) in matrix form, it follows

$$\widehat{\mathbf{\Phi}}_{n,k} = (1-\eta)\,\widehat{\mathbf{\Phi}}_{n-1,k} + \eta\ \hat{\mathbf{d}}'_{n,k}\ \hat{\mathbf{d}}'^{H}_{n,k} \tag{5.28}$$

with the vector of the estimated desired speech terms $\hat{\mathbf{d}}'_{n,k} = [\hat{d}_{n,k}, \hat{d}_{n-1,k}, \cdots, \hat{d}_{n-\tau_k+1,k}]^T$. The inverse of the estimated IFC matrix $\widehat{\mathbf{\Phi}}_{n,k}$ is to be used to obtain $\mathbf{A}_{n,k}$ and $\mathbf{b}_{n,k}$ in (5.21). Here, to avoid the complexity involved in direct inversion of $\widehat{\mathbf{\Phi}}_{n,k}$ and also to overcome the common singularity issue encountered in the inversion of the sample correlation matrix, we use the Sherman-Morrison matrix inversion lemma [131] to implicitly invert $\widehat{\mathbf{\Phi}}_{n,k}$, as given by (5.28). The simplified form of this lemma can be written as [131]

$$\left(\mathcal{A} - \mathcal{U}\mathcal{V}^H\right)^{-1} = \mathcal{A}^{-1} + \frac{\mathcal{A}^{-1}\mathcal{U}\mathcal{V}^H\mathcal{A}^{-1}}{1-\mathcal{V}^H\mathcal{A}^{-1}\mathcal{U}} \tag{5.29}$$

for an invertible matrix $\mathcal{A}$ and any two column vectors $\mathcal{U}$ and $\mathcal{V}$. Using (5.29) for the inverse of $\widehat{\mathbf{\Phi}}_{n,k}$ in (5.28), i.e., by taking $\mathcal{A}$, $\mathcal{U}$ and $\mathcal{V}$ respectively as $(1-\eta)\,\widehat{\mathbf{\Phi}}_{n-1,k}$, $-\eta\ \hat{\mathbf{d}}_{n,k}$ and $\hat{\mathbf{d}}'_{n,k}$, it can be deduced that

$$\widehat{\mathbf{\Phi}}^{-1}_{n,k} = \frac{\widehat{\mathbf{\Phi}}^{-1}_{n-1,k}}{1-\eta} - \frac{\eta}{1-\eta}\ \frac{\widehat{\mathbf{\Phi}}^{-1}_{n-1,k}\ \hat{\mathbf{d}}'_{n,k}\ \hat{\mathbf{d}}'^{H}_{n,k}\ \widehat{\mathbf{\Phi}}^{-1}_{n-1,k}}{1-\eta+\eta\ \hat{\mathbf{d}}'^{H}_{n,k}\ \widehat{\mathbf{\Phi}}^{-1}_{n-1,k}\ \hat{\mathbf{d}}'_{n,k}} \tag{5.30}$$

The above can be recursively implemented to update the inverse of $\widehat{\mathbf{\Phi}}_{n,k}$ at each frame without the need for direct matrix inversion.

It should be noted that the overall WPE-based dereverberation approach presented in this section can be considered as an extension of the method presented in Section 5.4, by taking into account the IFC of the desired speech signal. Namely, for the choice of τ_k=1, it can be shown that the proposed solution in (5.25) degenerates to the method suggested in Section 5.4.

5.6 Performance Evaluation

5.6.1 Experimental Setup

In this section, we evaluate the performance of the proposed WPE methods for dereverberation against the original WPE and a few recent variations of this method. To this end, clean speech utterances are used from the TIMIT database [105], including 10 male and 10 female speakers. The sampling rate is set to 16 kHz and a 40 msec Hamming window with overlap of 50% is used for the STFT analysis-synthesis. To have the best performance, the number of early speech terms considered in (5.2), i.e., D, is taken to be 3 with the length of the reverberation prediction vector, i.e., L_k, chosen as 20. We take the length of the IFC to be independent of frequency, i.e., $\tau_k \equiv \tau$, for our experiments. The number of microphones M is taken to be 2 and we use the first 10 seconds of the reverberant speech observation to estimate the reverberation prediction weights $\mathbf{g}_k$ in all conducted experiments.

In order to perform the matrix inversion in (5.25) with better accuracy, we use the QR factorization of the matrix $\widetilde{\mathbf{A}}_k$ in (5.23) with forward-backward substitution [132]. Also, to estimate the LRSV by (5.12c), knowledge of the reverberation time T_{60dB} is required. We used the reverberation time estimation method in [133] to estimate this parameter blindly from the observed speech. The estimated T_{60dB} in this way is accurate enough not to degrade the performance of the underlying LRSV estimator in (5.12). The smoothing parameters β and κ in (5.12a) and (5.12b) were respectively selected as 0.5 and 0.8 while η in (5.13) was fixed at 0.7. The use of time-frequency dependent values for the latter parameter could lead to improved results and remains an avenue for future work. Our approach requires no prior knowledge of the direct-to-reverberant ratio (DRR) parameter or its estimate.

We use both recorded and synthetic RIRs to generate microphone array signals modeling a reverberant noisy environment. In this sense, to account for a real-world scenario, we convolve the clean speech utterances with measured RIRs from the SimData of the REVERB Challenge [134], where an 8-channel circular array with diameter of 20 cm was placed in a 3.7×5.5 m room to measure the RIRs[2]. The resulting signal was combined with additive babble noise from the same database at an SNR of 10 dB. As well, to further analyze the performance of the considered

[2] Note that only two of the available 8 channels are used herein.

methods using flexible levels of reverberation, we use the ISM method [119] to simulate a scenario as illustrated in Figure 5.2. As viewed, a source of anechoic speech and two independent anechoic sources of babble noise, taken from Noisex-92 [104], are placed in an acoustic room with the indicated dimensions. The RIRs from the speech and noise sources to the linear microphone array are synthesized with a controlled reverberation time, T_{60dB}, and then convolved with the anechoic signals and added up at a reverberant SNR of 15 dB.

For the comparative evaluation of different dereverberation methods, we use four performance measures, as recommended by the REVERB Challenge [135]. These performance metrics include: the perceptual evaluation of speech quality (PESQ), the cepstrum distance (CD), the frequency-weighted segmental SNR (FW-SNR) and the signal-to-reverberation modulation energy ratio (SRMR). The PESQ score is one of the most frequently used performance measures in the speech enhancement literature and is the one recommended by ITU-T standards for speech quality assessment [102]. It ranges between 1 and 4.5 with higher values corresponding to better speech quality. The CD is calculated as the log-spectral distance between the linear prediction coefficients (LPC) of the enhanced and clean speech spectra [136]. It is often limited in the range of [0,10], where a smaller CD value shows less deviation from the clean speech. The FW-SNR is calculated based on a critical band analysis with mel-frequency filter bank and using clean speech amplitude as the corresponding weights [136]. It generally takes a value in the range of [-10,35] dB with the higher the better. The SRMR, which has been exclusively devised for the assessment of dereverberation, is a non-intrusive measure (i.e., one requiring only the enhanced speech for its calculation), and is based on an auditory-inspired filterbank analysis of critical band temporal envelopes of the speech signal [137]. A higher SRMR refers to a higher energy of the anechoic speech relative to that of the reverberant-only speech.

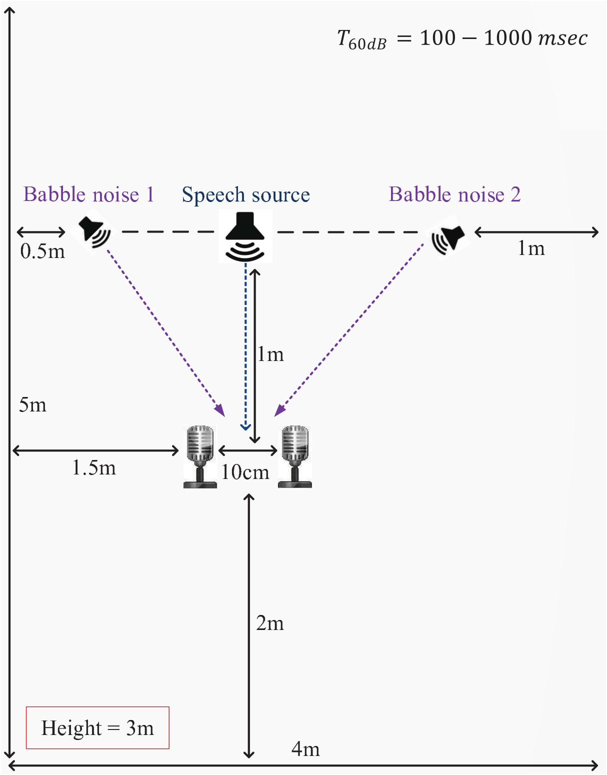


Figure 5.2: A two-dimensional illustration for the geometry of the synthesized scenario of a noisy reverberant environment.

Recently, there has been growing interest in the use of properly fitted distributions for speech priors and estimation of the corresponding parameters, as discussed in detail in Chapter 2. In the context of dereverberation based on the WPE method, this has been accomplished recently by using the complex generalized Gaussian (CGG) and Laplacian speech priors, respectively in [138] and [139]. To evaluate the reverberation suppression performance of the proposed methods, we compare them to the original WPE method [126], the two aforementioned developments of this method in [138, 139], as well as reverberation suppression using spectral enhancement [43]. The CGG-based method has in fact the same solution as the original WPE method but with a power-scaled estimator of the speech spectral variance in the iterative procedure of Table 5.1. The Laplacian-based method, however, does not have a closed-form solution for the reverberation prediction weights, $\mathbf{g}_k$, and has to be implemented through numerical optimization, e.g., using the CVX toolbox [140]. The spectral enhancement approach to dereverberation, as will be investigated thoroughly in the next chapter, is in fact similar to the noise reduction methods reviewed in Chapter 2 but with using an LRSV estimator, e.g., that in (5.12), to replace the estimate of noise spectral variance.

5.6.2 Evaluation of the Proposed Method in Section 5.4

In Figures 5.3 and 5.4, performance comparison of the proposed method in Section 5.3 with respect to the conventional, two recent WPE-based methods, and using spectral enhancement [43] is illustrated. The values of ΔPESQ and such represent the improvements in these quantities relative to the corresponding value for the unprocessed (reverberant) speech, which is denoted in the figures as "ref". As seen, the Laplacian-based method outperforms the original WPE, whereas the CGG-based method provides only trivial improvements. However, the proposed method in Section 5.3, in addition to being non-iterative in nature, is able to provide a more efficient reverberation suppression than the former methods.

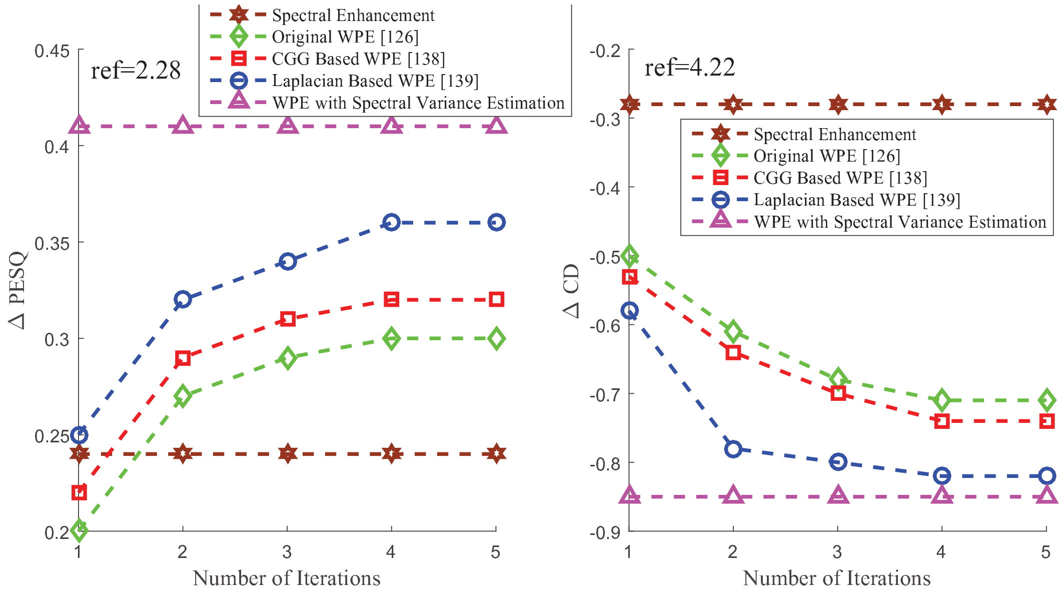


Figure 5.3: Improvements in PESQ and CD scores versus the number of iterations for different dereverberation methods.

It is also observable that the spectral enhancement method with an LRSV estimator is not as efficient as the WPE-based methods for the purpose of dereverberation.

Next, to evaluate experimentally the efficiency of the proposed estimate for the speech spectral variance, $\sigma^2_{d_{n,k}}$, in Section 5.3, we considered two other recursive smoothing based schemes to update $\sigma^2_{d_{n,k}}$ and compared their performance with the proposed one in Figure 5.5. In this case, we used the scenario of Figure 5.2 but with excluding the noise sources and considering a source-to-microphone distance of 1.5 m. As in [43], we used the well-known DD approach to estimate the ratio $\sigma^2_{d_{n,k}}/\sigma^2_{r_{n,k}}$ and then multiplied it by the LRSV estimate from (5.12) to obtain an estimate of

$\sigma^2_{d_{n,k}}$, which is denoted in Figure 5.5 as the "DD Approach". To demonstrate the importance of taking into account the cross-correlation terms between the desired and late reverberant speech, as in (5.8), we estimated the desired spectral variance, $\sigma^2_{d_{n,k}}$, by disregarding the cross terms in (5.8) and using the term $\sigma^2_{x^1_{n,k}} - \hat{\sigma}^2_{r_{n,k}}$. Since the observation spectral variance is estimated by a fixed recursive smoothing scheme, we denoted this method by "Smoothing of Observations" in this figure. It is observable that the proposed estimation of $\sigma^2_{d_{n,k}}$ results in further reverberation suppression, especially for higher reverberation conditions where the amount of correlation between the desired and late reverberant speech signals increases.

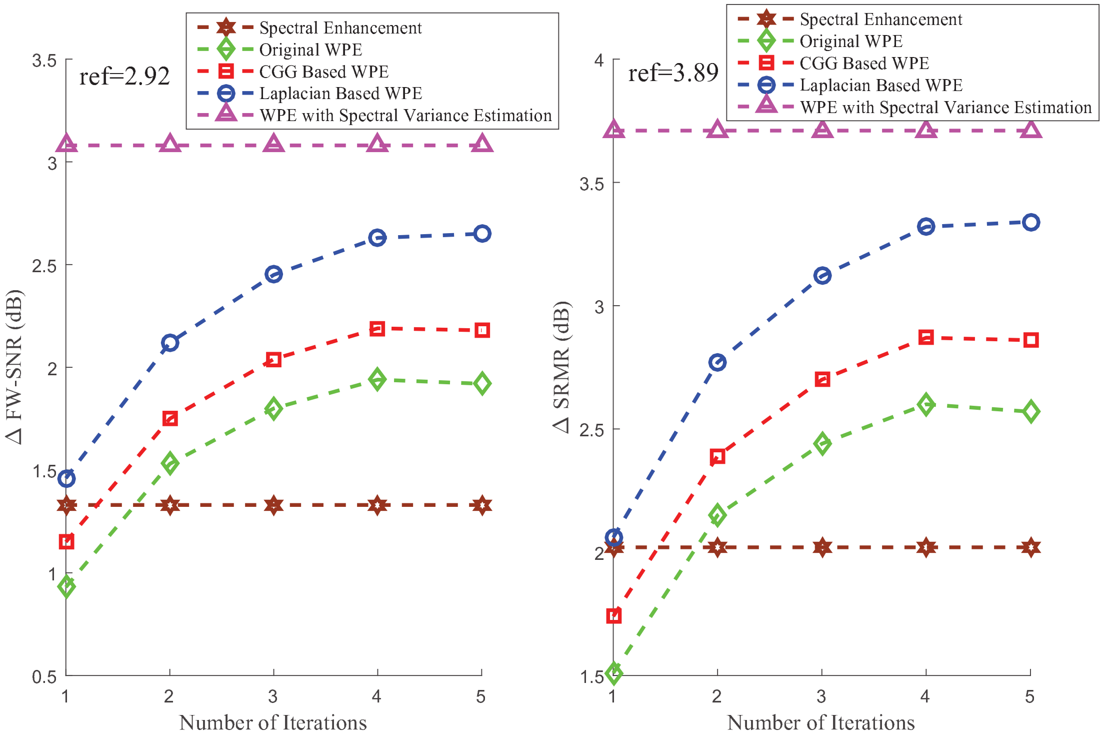


Figure 5.4: Improvements in FW-SNR and SRMR scores versus the number of iterations for different dereverberation methods.

It should be noted that the proposed estimator of the desired speech spectral variance can also be used in spectral enhancement-based methods, yet the dereverberation performance of the latter was found to be inferior to the LP-based methods in general.

We also evaluated experimentally the computational cost of our proposed algorithm by using the estimation of $\sigma^2_{d_{n,k}}$ discussed in Section 5.3, the proposed algorithm using the DD approach to estimate $\sigma^2_{d_{n,k}}$, and the conventional WPE method using a maximum of 3 iterations. The results are presented in Figure 5.6 in terms of the batch processing time needed to estimate the WPE regression vector, $\mathbf{g}_k$. As seen, by eliminating the iterative process of the WPE method through the proposed algorithm, the computational effort has been considerably reduced.

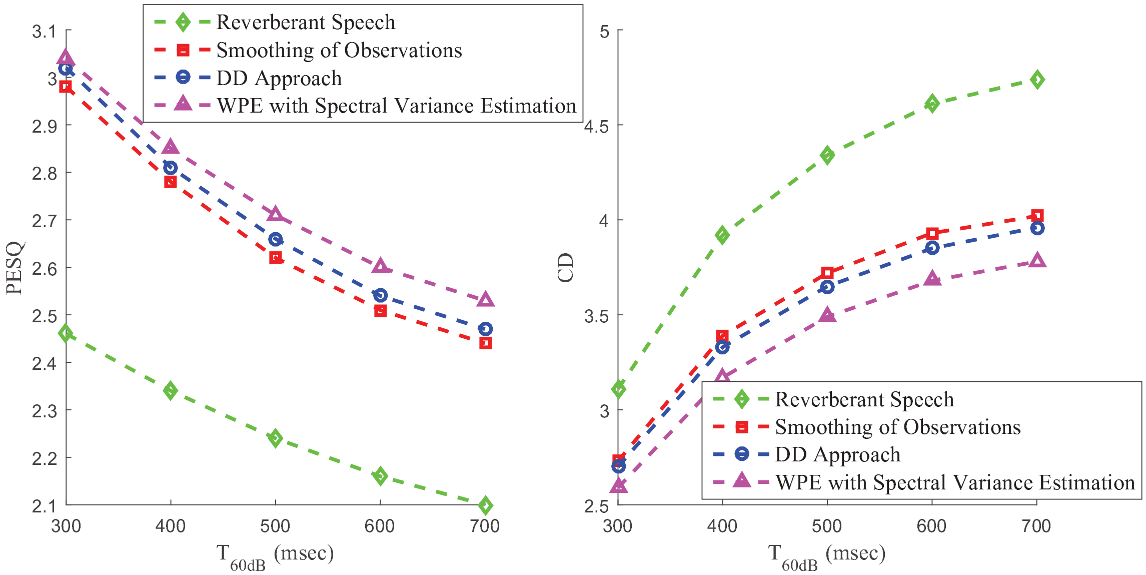


Figure 5.5: PESQ and CD scores versus T_{60dB} for the reverberant speech and the enhanced one using the WPE method with different estimators of the desired speech spectral variance, $\sigma^2_{d_{n,k}}$.

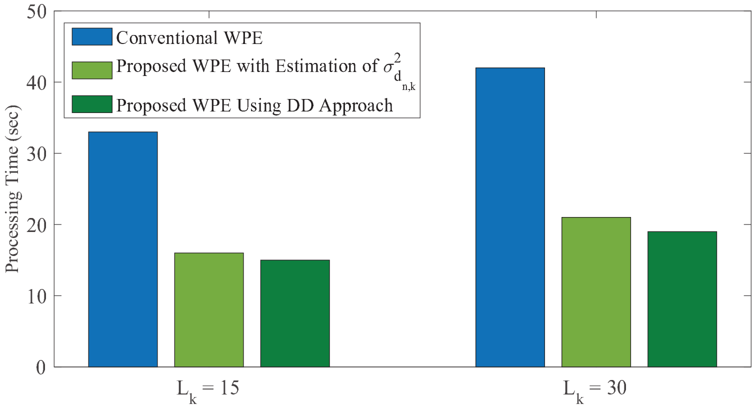


Figure 5.6: Processing time required for the estimation of $\mathbf{g}_k$ with lengths of L_k=15 and L_k=30 using a 10-second speech segment for different methods. An i5-2400 CPU @ 3.10GHz with RAM of 4.00GB was used for the implementation in Matlab.

5.6.3 Evaluation of the Proposed Method in Section 5.5

Regarding the proposed method in Section 5.4, to investigate the IFC present between early speech terms with different frame lags, we calculated the normalized IFC by sample averaging over all frequency bins and frames. The results are shown in Figure 5.7 for both anechoic and reverberant speech signals with different values of the reverberation time. As seen, the IFC is quite pronounced for smaller lag values (say 5 or less), but decreases to a lower level for larger lags. We will take into account this observation in choosing the appropriate IFC length, τ, in the sequel. A more detailed study of the IFC in the STFT domain can be found in [141].

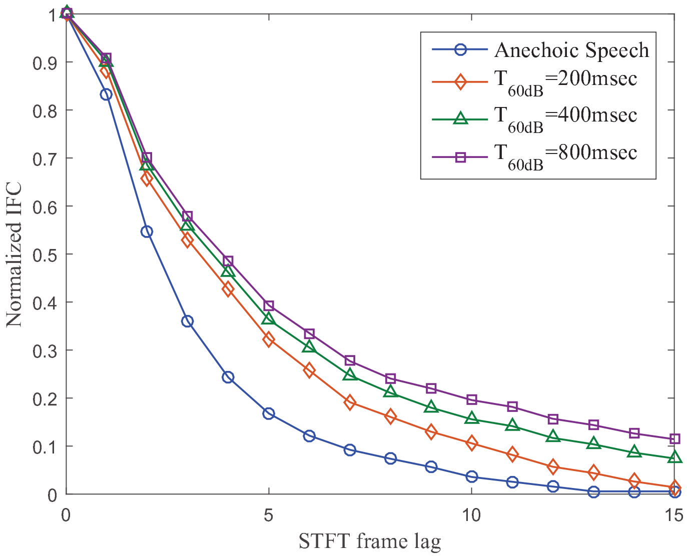


Figure 5.7: Normalized IFC averaged over frequency bins and frames versus the frame lag for speech samples with different amounts of reverberation.

Next, we study the effect of the assumed number of correlated speech frames, τ, on the overall performance of the proposed dereverberation approach. It was found that the choice of this parameter is more dependent on the number of early speech frames, D, than on other involved parameters, e.g., L_k and T_{60dB}. This theoretically makes sense since the parameter D determines the duration of the early reflections, and therefore, the IFC is controlled by D to a large extent. Figure 5.8 shows the PESQ scores of the proposed approach versus different τ with D ranging from 1 to 4, when using the measured RIRs from the SimData of the REVERB Challenge. Apart from the observation that the performance of the proposed approach is best for D=3, it can be seen that the higher the value of D the larger the value of the choice of τ resulting in the best performance. The latter result is due to the fact that the higher the value of D the larger the amount of the IFC between subsequent frames of the desired speech. It is also observed that the best choice of the parameter τ occurs in the range of 2-6, despite the fact that the theoretically optimal choice of τ is

N, i.e., the number of frames in the entire speech utterance [3]. The reason for this limitation in the performance of the proposed approach seems to be due to the limited accuracy in the estimation of the IFC matrix, $\mathbf{\Phi}_{n,k}$. In effect, the estimation error in $\widehat{\mathbf{\Phi}}_{n,k}$, which grows with the size τ of the matrix $\mathbf{\Phi}_{n,k}$, degrades the overall performance of the proposed method. Therefore, we choose the value of τ=5 for the case of D=3 in our experiments. This is also consistent with the fact that the IFC is more strongly present in the lag values of around 5 or less, as inferred before from Figure 5.7.

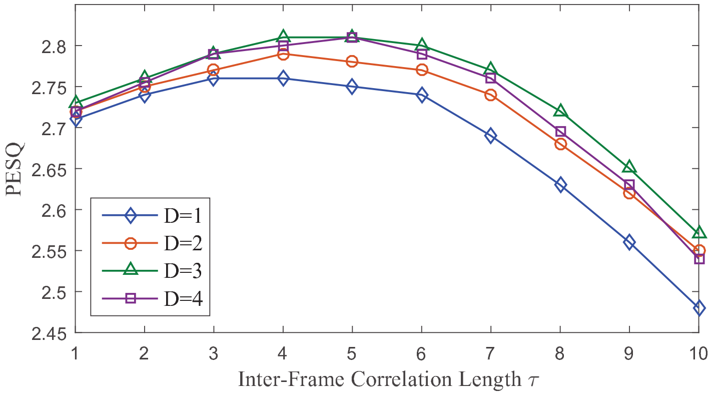


Figure 5.8: Performance of the proposed WPE method versus the assumed IFC length, τ, for different D.

Finally, we compare the performance of the proposed methods in Sections 5.3 and 5.4 along with other WPE-based methods. The comparative results by using the recorded RIR from the REVERB Challenge are presented in Table 5.2 in terms of the aforementioned objective performance measures. As observed, whereas the CGG-based method achieves close scores to the original WPE and the Laplacian-based method is superior to the former, the WPE with spectral speech variance estimation, i.e., that proposed in Section 5.3, performs better than the former three methods, and the WPE with IFC, i.e., that proposed in Section 5.4, achieves superior results w.r.t. to all. Note that the method presented in Section 5.3 is actually a particular case of the presented method in Section 5.4 by neglecting the IFC and estimating only the speech spectral variance at each frame independently. We found that the objective performance of the considered methods in terms of the four investigated scores used in this work was almost consistent.

[3] Note that in this case, the approximate model in (5.16) turns into an accurate joint model for all the desired speech frames.

Table 5.2: Performance comparison of different WPE-based dereverberation methods using recorded RIRs.

Method	PESQ	CD	FW-SNR (dB)	SRMR (dB)
Unprocessed	2.28	4.22	2.92	3.89
Original WPE [126]	2.58	3.51	5.16	6.49
CGG-based WPE [138]	2.61	3.45	5.39	6.82
Proposed WPE in Section 5.3	2.69	3.37	6.12	7.60
Proposed WPE in Section 5.4	2.75	3.20	6.85	8.11

Next, to evaluate the performance of the considered dereverberation methods for different amounts of reverberation, we obtained the objective performance measures by using the synthesized RIRs with different T_{60dB} by the ISM method. The results are presented in Figures 5.9 and 5.10 for T_{60dB} in the range of 100 to 1000 msec. For better visualization, the resulting improvements in the performance scores w.r.t. the unprocessed speech (denoted by ΔPESQ and such) are illustrated. As seen, the proposed methods in Sections 5.3 and 5.4, which are both based on the estimation of the speech spectral variance by means of an LRSV estimator from the context of spectral enhancement, perform considerably better than the previous versions of the WPE method, which tend to estimate the speech spectral variance iteratively along with the reverberation prediction weights. Also, it is observed that the proposed method in Section 5.4 achieves the best scores w.r.t. the others in almost the entire range of T_{60dB}. This advantage is more visible for the moderate values of T_{60dB}.

In addition to the objective performance measurements reported in the paper, informal subjective listening to the enhanced speech files revealed superior quality and lower residual reverberation provided by the proposed method as compared to the other methods in [126, 138, 139].

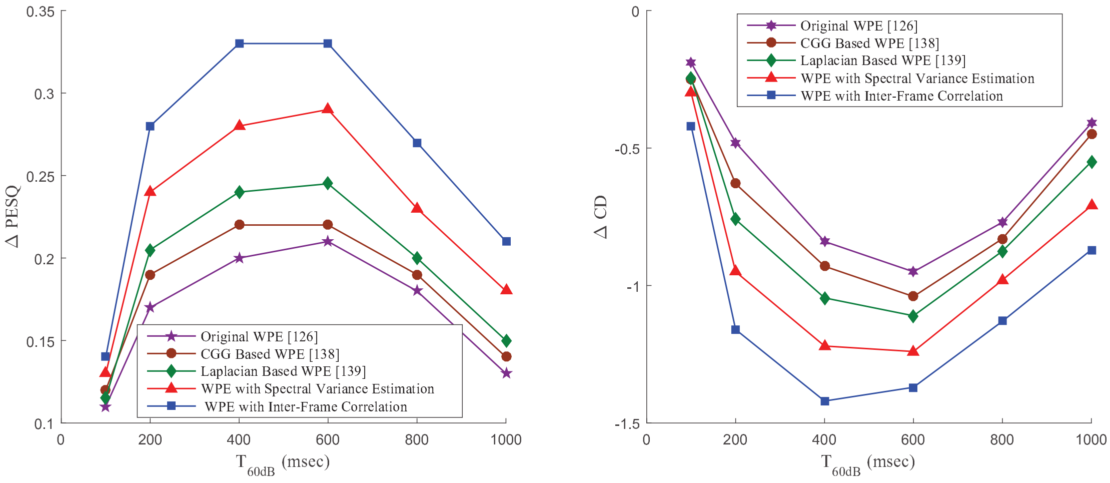


Figure 5.9: Improvement in PESQ and CD scores versus T_{60dB} for different WPE-based dereverberation methods using synthetic RIRs.

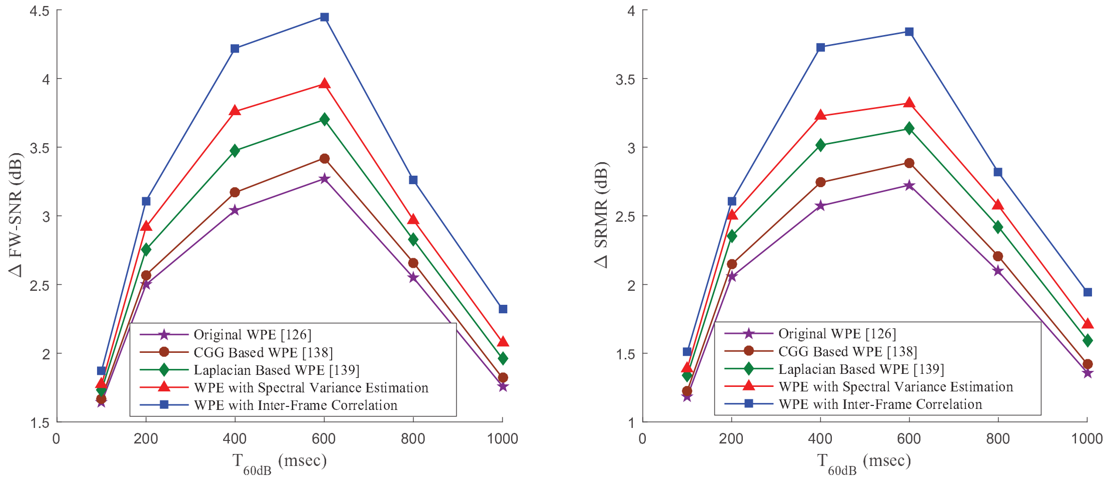


Figure 5.10: Improvement in FW-SNR and SRMR scores versus T_{60dB} for different WPE-based dereverberation methods using synthetic RIRs.

5.7 Conclusion

In this chapter, we presented novel dereverberation approaches based on the WPE method and by taking advantage of speech spectral variance estimation from the context of spectral enhancement. The spectral variance estimate is obtained through a geometric spectral enhancement approach and a conventional LRSV estimator, based on the correlation between the early and late reverberant terms. In Section 5.3, it was shown that by integrating the suggested spectral variance estimator into the WPE method, this method can be implemented in a non-iterative manner, that is less complex and more efficient in reverberation suppression, as compared to the original WPE method and its more recent variations.

Next, as an extension to the suggested method in Section 5.3, we proposed to approximately model and exploit the temporal correlation across desired speech frames, namely the IFC in the STFT domain. It was shown that this dereverberation problem can be handled by solving an unconstrained quadratic optimization in a straightforward manner, given an estimate of the spectral correlation matrix of the subsequent frames. Performance evaluations using both recorded and synthetic RIRs revealed that the proposed methods considerably outperform the previous variations of the WPE method.

It is concluded that incorporating the statistical model-based estimation of the desired speech variance into the linear prediction dereverberation methods can lead to better dereverberation performance. This approach, unlike the state-of-the-art WPE methods, results in a non-iterative estimator for the reverberation prediction weights, provided that a proper estimate of the spectral auto- and cross-variances of the desired speech terms is available. We believe the existing limit on the performance of the suggested WPE method in this work is mostly due to the inaccuracy in the estimation of the inter-frame spectral correlations, and therefore, this limit can be overcome by developing more efficient estimators of the IFC. This topic can serve as a future research avenue for speech dereverberation in the STFT domain. Accurately modeling and incorporating the correlation across the spectral components of desired speech at each frame, namely the intra-frame correlation, can also be regarded as a direction of future research. Furthermore, since the proposed and state-of-the-art WPE methods result in constant (time-invariant) reverberation prediction weights, in order to cope with changing reverberant environments, development of an incrementally updated (over blocks of time frames) WPE method remains as further research.

Chapter 6

Speech Dereverberation Using the Spectral Enhancement Method

6.1 Introduction

Spectral enhancement methods originally developed for the purpose of noise reduction have also been modified and used for dereverberation. The major advantage of the spectral enhancement methods over other techniques such as channel equalization (inverse filtering) and linear prediction-based methods is their simplicity in implementation in the STFT domain and low computational complexity, which has made them one of the most widely used techniques for speech enhancement. According to Section 2.4, since it is well known that the late reverberation is the major reason for deterioration of speech quality, spectral enhancement methods for dereverberation aim at the suppression of late reverberation by estimating the early reverberant speech. In this regard, assuming that early and late reverberations are independent and under the phase equivalence of the reverberant and anechoic (non-reverberant) speeches, these methods can be employed for late reverberation suppression by estimating the late reverberant spectral variance (LRSV) and using it in place of the noise spectral variance [43]. Therefore, the main challenge here is to estimate the LRSV blindly from a set of reverberant observations. Originally suggested by Labert et al. in [142], the late reverberation is treated as a sort of additive noise, and through statistical modeling of the RIR, an estimator of the LRSV is derived and used in a spectral subtraction rule. On this basis, several estimators of the LRSV have been proposed and applied to spectral enhancement methods for dereverberation in the past decade. Since the LRSV estimator in [142] is based on a time-domain model of the RIR and also under the implicit assumption that the source-to-microphone distance is larger than a critical distance, i.e., the distance at which the direct-to-reverberant ratio (DRR) is larger than 1 (in smaller distances, the LRSV estimator in [142] overestimates the true LRSV), in [127], Habets developed a new LRSV estimator that overcomes these deficiencies. Therein, he proposed a statistical RIR model in the STFT domain and used it to derive an extension of the Lebart's LRSV estimator that takes into account the energy contribution of the direct path and reverberant parts of speech. This statistical RIR model is dependent only on the reverberation time, which generally changes slowly with time. However, similar to Lebart's method, the recursive scheme suggested in [127] is basically derived for a fixed RIR, i.e., no changing environment, and it also requires the *a priori* knowledge of RIR statistics or the DRR parameter. In [143], therefore, an LRSV estimator that is based on the correlation of the reverberant and dereverberated speech has been proposed in contrast to the previous model-based LRSV estimation approaches. The suggested LRSV estimation scheme requires no knowledge of the RIR model parameters such as the reverberation time and DRR, and outperforms the previous methods. However, this method is able to track very slow changes in RIR and underestimates the LRSV in case of time-varying RIRs. Therefore, it is recommended in [143] to use the model-based LRSV estimation as before for the general case of time-varying RIRs, and it is proved that under a few extra mild conditions, the model-based LRSV estimation approach is valid. In this regard, a smoothing parameter (the so-called shape parameter therein) that is a function of the frequency bins is suggested, but this shape parameter has to be estimated blindly and the amount of data needed for its accurate estimation is on the order of several seconds. In the same direction, a few more recent schemes of LRSV estimation have been suggested in the literature such as the one in [144]. Therein, since the shape parameter used in the LRSV estimation scheme is affected by the error in the estimation of LRSV, in order to obtain a smoother shape parameter, it was suggested to use more than one term of the past spectral variance of the reverberant speech.

In summary, it can be concluded that even though the existing literature includes a few major schemes for the estimation of the LRSV, blind estimation of this parameter particularly in fast changing environments is still a challenging problem and requires further research.

The rest of this chapter is organized as follows. In Section 6.2, a summary of the proposed methods is explained. A brief review of late reverberation suppression using spectral enhancement (gain function-based method) is presented in Section 6.3. Section 6.4 presents the proposed approach for the estimation of the LRSV. In Section 6.5, we describe a few other developments in order to make use of the conventional spectral enhancement method in reverberation suppression, which include the estimation of signal-to-reverberant ratio (SRR), application of SPP to the gain function used for enhancement, spectral flooring and the use of beamforming for reverberation suppression. In Section 6.6, performance evaluations are discussed for both the proposed LRSV estimator and the other developed schemes, and Section 6.7 gives the conclusions of this chapter.

6.2 Brief Description of the Proposed Methods

In this chapter, we suggest a new smoothing scheme for the estimation of the LRSV, that takes advantage of the WPE method for the selection of the shape parameter. Contrary to the conventional WPE dereverberation method, which is in need of a few seconds of observations to estimate the reverberation prediction weights, we implement the WPE method in an incremental-based manner. At each block of the increment, the estimated reverberation prediction weights are used to extract a rough estimate of the reverberant and dereverberated (direct-path) speech components at that block, which, in turn, are exploited to estimate the spectral variances of the direct-path and reverberant components of the RIR. The latter is used to select the shape parameter dynamically for each block, and makes the proposed scheme especially suitable for changing environments. Further, we employ the dereverberated speech component in a moving average (MA) scheme to estimate the reverberant-only spectral variance that is of high importance to obtain a precise estimate of the LRSV.

Next, we consider developing several other notions from the context of noise reduction to their counterparts in dereverberation, including the estimation of SRR, a flooring scheme for the gain function, application of SPP to modify the gain function, and using beamforming to suppress the late reverberation. Regarding the estimation of SPP, we develop a two-step scheme where in the first step, an initial estimate of the SPP is obtained based on a modification of the decision-directed approach. This initial estimate is exploited to obtain an MMSE optimal smoothing parameter as well as a gain function, which are in turn used in the second step of the algorithm to determine the ultimate value of the SRR. The suggested flooring scheme is based on a linear prediction model of the desired (early) reverberant speech and replaces the small values of the gain function by an attenuated version of the estimated early speech, in order to avoid the introduced distortion by the spectral modification. Next, we derive a straightforward but efficient extension of the notion of SPP, originally defined for a noisy scenario, that is used to optimally modify the spectral gain function to suppress the late reverberation. Finally, it is shown that the concept of single-channel LRSV estimation can be generalized to the estimation of cross-LRSVs in the multi-channel, which in fact constitutes an estimate for the LRSV matrix. The latter is beneficial in employing the conventional beamforming techniques, e.g., the MVDR beamformer, to suppress the late reverberation.

The contributions presented in this chapter are to be submitted as [145].

6.3 Background: Late Reverberation Suppression Using Spectral Enhancement

In this section, we present a brief overview of the state-of-the-art literature on LRSV estimation for single-channel late reverberation suppression using spectral enhancement. Recall from Section 2.4 that the reverberant speech in the STFT domain can be written as

$$Y(k,l) = Y_E(k,l) + Y_L(k,l) \tag{6.1}$$

where $Y_E(k,l)$ and $Y_L(k,l)$ are respectively the early and late reverberant components at the kth frequency bin and lth frame. The goal of late reverberation suppression through spectral enhancement is to obtain an estimate of the early reverberant component, $Y_E(k,l)$, by reducing the late reverberation, $Y_L(k,l)$. This has been originally done by Lebart in [142] where the classic spectral subtraction rule, originally developed for additive noise reduction, is applied by a gain function on the observations in order to suppress $Y_L(k,l)$, as the following

$$\hat{Y}_E(k,l) = G(k,l)Y(k,l) \tag{6.2}$$

with $\hat{Y}_E(k,l)$ and $G(k,l)$ are respectively the estimated early reverberant speech and spectral gain function. For the latter, various expressions can be found from the noise reduction literature, e.g., those employed in [142]. Yet, this gain function generally depends on two important parameters, which, in the context of late reverberation suppression, are

$$\zeta(k,l) = \frac{\sigma^2_{Y_E}(k,l)}{\sigma^2_{Y_L}(k,l)}, \quad \eta(k,l) = \frac{|Y(k,l)|^2}{\sigma^2_{Y_L}(k,l)} \tag{6.3}$$

where the two parameters $\sigma^2_{Y_E}(k,l)=E\{|Y_E(k,l)|^2\}$ and $\sigma^2_{Y_L}(k,l)=E\{|Y_L(k,l)|^2\}$ are respectively the spectral variances of the early and late reverberant components. Borrowed from the noise reduction context, $\sigma^2_{Y_E}(k,l)$ can be estimated through the conventional decision-directed (DD) approach [17]. However, the estimation of $\sigma^2_{Y_L}(k,l)$ or the so-called LRSV, due to its high influence in the overall performance of the spectral enhancement method, has attracted considerable attention in the recent literature and is the main focus of this work.

In [127], Habets suggests the following statistical RIR model in the STFT domain

$$H(k,l) = \begin{cases} B_D(k), & l = 0 \\ B_R(k,l)e^{-\alpha(k)lP}, & l \geq 1 \end{cases} \tag{6.4}$$

where P is the STFT frame advance (hop size), $\alpha(k)$ is defined as $3\ \mathrm{ln}10/(f_s T_{60dB}(k))$ with f_s as the sampling frequency and $T_{60dB}(k)$ as the reverberation time, and $B_D(k)$ and $B_R(k,l)$ are two zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random processes corresponding respectively to the direct-path and reverberant components of the RIR in the STFT domain. Based on this model, the following recursive scheme for the LRSV estimator was derived [127]

$$\begin{aligned} \hat{\sigma}^2_{Y_L}(k,l) &= e^{-2\alpha(k)P(N_E-1)}\hat{\sigma}^2_{Y_R}(k,l-N_E+1) \\ \hat{\sigma}^2_{Y_R}(k,l) &= [1-\kappa(k)]e^{-2\alpha(k)P}\hat{\sigma}^2_{Y_R}(k,l-1) + \kappa(k)e^{-2\alpha(k)P}\hat{\sigma}^2_{Y}(k,l-1) \\ \hat{\sigma}^2_{Y}(k,l) &= [1-\beta]\hat{\sigma}^2_{Y}(k,l-1) + \beta|Y(k,l)|^2 \end{aligned} \tag{6.5}$$

with β as a fixed smoothing parameter and $\kappa(k)$ as the shape parameter used to estimate the reverberant spectral variance $\sigma^2_{Y_R}(k,l)$. Accurate estimation of the latter is highly important in the context of LRSV estimation, since $\sigma^2_{Y_R}(k,l)$ should exclude the direct-path speech component in order to avoid distorting this component by the underlying spectral suppression rule. Hence, proper selection of the shape parameter κ is of high importance in this context. In [127], it is proved that the optimal value of this parameter is in fact the ratio of the variance of $B_R(k,l)$ to that of $B_D(k)$, i.e., $\sigma^2_{B_R}(k)/\sigma^2_{B_D}(k)$, which can be obtained by the following

$$\kappa(k) = \frac{\sigma^2_{B_R}(k)}{\sigma^2_{B_D}(k)} = \frac{e^{2\alpha(k)P} - 1}{\mathrm{DRR}(k)} \tag{6.6}$$

with $\mathrm{DRR}(k)$ as the ratio of the energy of the direct-path speech to that of the reverberant speech or the so-called direct-to-reverberant ratio. However, to use (6.6), $\mathrm{DRR}(k)$ has to be blindly estimated, which requires the additional implementation of a blind DRR estimation method, which requires at least a few seconds of speech observations. Therefore, this scheme does not suit well the case of a changing RIR. Also, as observed in (6.5), the estimation of the reverberant spectral variance $\sigma^2_{Y_R}(k,l)$ is actually performed by the recursive smoothing of the entire reverberant observation $Y(k,l)$, and therefore, it does not exclude the direct-path component in the estimation of $\sigma^2_{Y_R}(k,l)$. The latter is found to be one of the major obstacles of using spectral modification for the purpose of dereverberation. In the following section, we propose a new scheme for the estimation of the LRSV, which particularly takes advantage of the linear prediction-based dereverberation in eliminating the direct-path component in estimating the reverberant spectral variance $\sigma^2_{Y_R}(k,l)$.

6.4 Proposed LRSV Estimator

We base our LRSV estimation approach on the scheme in [127] discussed in Section 6.2. Yet, we target time-varying acoustic environments where the RIR cannot be assumed constant over a period of a few seconds. It should be noted that, however, even though the RIR modeling in [127] is basically valid for constant RIRs, it is shown in [143] that the same modeling is approximately valid for time-varying RIRs, provided that the reverberation time and DRR remain almost constant during an interval of the order of a few time frames. Therefore, under reasonably moderate conditions, the same LRSV estimators developed for constant RIRs such as [127] can be used for changing acoustic environments. We also here base our estimator for the LRSV on the model in [127] but aim at adapting it with the changing acoustic environment.

Due to the importance of the accuracy in the estimation of the reverberant spectral variance $\sigma^2_{Y_R}(k,l)$ as in (6.5), in this work, we mostly focus on the proper estimation of $\sigma^2_{Y_R}(k,l)$. In this respect, in a similar fashion to (6.5), we use the following scheme for the estimation of the LRSV

$$\hat{\sigma}^2_{Y_L}(k,l) = e^{-2\alpha(k)P(N_E-1)}\hat{\sigma}^2_{Y_R}(k,l-N_E+1) \tag{6.7}$$
$$\hat{\sigma}^2_{Y_R}(k,l) = [1-\kappa(k,l)]\hat{\sigma}^2_{Y_R}(k,l-1) + \kappa(k,l)|\hat{Y}_R(k,l)|^2$$

As compared to (6.5), a new time and frequency-dependent scheme for the shape parameter $\kappa(k,l)$ is proposed, which fits properly the case of a time-varying RIR. In addition, rather than estimating the reverberant spectral variance $\sigma^2_{Y_R}(k,l)$ by only smoothing the observation $|Y(k,l)|^2$, we exploit a reverberant-only component of the speech, $\hat{Y}_R(k,l)$, which specifically excludes the direct-path component. This helps avoiding the leakage of the direct-path speech into the estimated LRSV to a large extent. In Sections 6.4.1 and 6.4.2, we will respectively discuss the proposed schemes for the shape parameter $\kappa(k,l)$ and the reverberant-only speech component $\hat{Y}_R(k,l)$, which are based on an incremental (block) processing of the observed speech.

6.4.1 Suggested Scheme for the Shape Parameter

In this section, based on (6.6), we propose a new blind scheme to obtain the shape parameter κ. This is achieved by finding a proper estimator for the DDR(k) in (6.6) as a function of time frame l and frequency bin k. The parameter DDR(k) can be actually interpreted as the ratio of the energy of the direct-path component to that of the reverberant component [42]. In fact, choosing the shape parameter κ by (6.6) results in further updating of the reverberant spectral variance $\hat{\sigma}^2_{Y_R}(k,l)$ when the reverberant energy is higher, and conversely, further smoothing of $\hat{\sigma}^2_{Y_R}(k,l)$ when the direct-path energy is dominant. Based on this fact, we suggest to choose the shape parameter by the following

$$\kappa^1(k,l) = \frac{e^{2\alpha(k)P}-1}{\hat{\sigma}^2_{Y_D}(k,l)/\hat{\sigma}^2_{Y_R}(k,l)}$$
$$\kappa(k,l) = \min\{\max\{\kappa^1(k,l),0\},1\} \tag{6.8}$$

where $\hat{\sigma}^2_{Y_D}(k,l)$ and $\hat{\sigma}^2_{Y_R}(k,l)$ are estimates of the spectral variances of the direct-path and reverberant speech components, respectively, and the equation at the bottom is to ensure that the shape parameter lies in $[0,1]$. To estimate the two spectral variances in (6.8), we use the recursive smoothing method, i.e.,

$$
\begin{aligned}
\hat{\sigma}^2_{Y_D}(k,l) &= [1-\gamma_1]\hat{\sigma}^2_{Y_D}(k,l-1) + \gamma_1|\hat{\mathcal{Y}}_D(k,l)|^2 \\
\hat{\sigma}^2_{Y_R}(k,l) &= [1-\gamma_2]\hat{\sigma}^2_{Y_R}(k,l-1) + \gamma_2|\hat{\mathcal{Y}}_R(k,l)|^2
\end{aligned}
\tag{6.9}
$$

with γ_1 and γ_2 as two fixed smoothing parameters taken to be 0.25, and $\hat{\mathcal{Y}}_D(k,l)$ and $\hat{\mathcal{Y}}_R(k,l)$ as estimates of the direct-path and reverberant components of speech, respectively. Since $\hat{\mathcal{Y}}_D(k,l)$ and $\hat{\mathcal{Y}}_R(k,l)$ are not available *a priori*, we resort to a linear prediction-based dereverberation in the STFT domain, namely the WPE method [126], in order to obtain rough estimates of these two terms. However, the WPE method is in essence a batch processing technique and it requires preprocessing of the entire speech utterance in order to provide an accurate performance. This violates our goal of dealing with a time-varying acoustic environment, where the RIR is prone to change in a duration of less than a second. Furthermore, large processing delays are imposed due to the mentioned preprocessing stage, which is undesirable for real-time speech processing systems. To overcome these obstacles, here we employ the WPE method block-wise for speech blocks (processing increments) of typically 0.5 second long. We then exploit the estimated direct-path and reverberant components obtained from the WPE method in (6.9) at the end of each processing block. A schematic of the processing blocks and time frames is shown in Figure 6.1. Within this framework, despite the fact that the precision of the underlying WPE method may degrade to some degree, the resulting primary estimates of the direct-path and reverberant components, i.e., $\hat{\mathcal{Y}}_D(k,l)$ and $\hat{\mathcal{Y}}_R(k,l)$, are precise enough to be used in the suggested scheme for $\kappa(k,l)$ in (6.8) and (6.9), as will be investigated thoroughly in Section 6.6.

Now, denoting each processing block by λ and the block length (in samples) by Δ, based on [126], the resulting incremental WPE method can be summarized as follows:

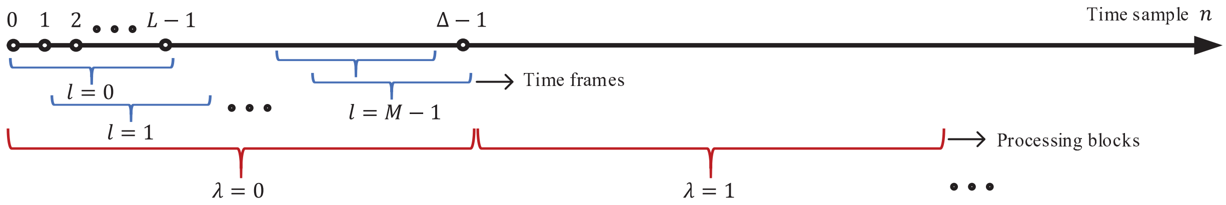


Figure 6.1: An illustration of the STFT frames and the processing blocks over speech time samples.

- At the processing block λ, the observation $Y(k,l)$ is considered for $l \in \{\lambda M, \lambda M{+}1,\cdots, \lambda M{+} M-1\}$ (which is actually M of the STFT frames). We set the following parameters: d=1, I=15, γ=0.65 and ϵ=10^{-3}, and form the observation vector $\mathbf{Y}(k, l-d)$ as below

$$\mathbf{Y}(k,l-d) = [Y(k,l-d), Y(k,l-d-1), \cdots, Y(k,l-d-I+1)]^T \tag{6.10}$$

- The speech spectral variance $\sigma^2_{\mathcal{Y}_D}(k,l)$ is initialized as $\sigma^2_{\mathcal{Y}_{D_0}}(k,l) = |Y(k,l)|^2$.
- Repeat the following for $j = 0 : J-1$

$$\begin{aligned} \mathbf{A}_{\lambda_j}(k) &= \sum_l \frac{\mathbf{Y}(k,l-d)\mathbf{Y}^H(k,l-d)}{\sigma^2_{\mathcal{Y}_{D_j}}(k,l)} \\ \mathbf{a}_{\lambda_j}(k) &= \sum_l \frac{\mathbf{Y}(k,l-d)Y^*(k,l)}{\sigma^2_{\mathcal{Y}_{D_j}}(k,l)} \end{aligned} \tag{6.11}$$

  where $l \in \{\lambda M, \lambda M+1, \cdots, \lambda M + M - 1\}$

$$\mathbf{g}_{\lambda_j}(k) = \mathbf{A}^{-1}_{\lambda_j}(k)\, \mathbf{a}_{\lambda_j}(k) \tag{6.12}$$

$$\begin{aligned} \mathcal{Y}_{R_j}(k,l) =& \mathbf{g}^H_{\lambda_j}(k)\mathbf{Y}(k,l-d) \\ \mathcal{Y}_{D_j}(k,l) =& Y(k,l) - \mathcal{Y}_{R_j}(k,l) \end{aligned} \tag{6.13}$$

$$\begin{aligned} \sigma^2_{\mathcal{Y}_{D_{j+1}}}(k,l) = [1-\gamma]&\sigma^2_{\mathcal{Y}_{D_{j+1}}}(k,l-1) \\ &+\gamma \max\left\{|\mathcal{Y}_{D_j}(k,l)|^2, \epsilon\right\} \end{aligned} \tag{6.14}$$

- The terms $\mathcal{Y}_{R_j}(k,l)$ and $\mathcal{Y}_{D_j}(k,l)$ at the last iteration are considered as $\hat{\mathcal{Y}}_R(k,l)$ and $\hat{\mathcal{Y}}_D(k,l)$ in (6.9).

Note that, contrary to the original WPE method, here the reverberation prediction weights $\mathbf{g}_{\lambda_j}(k)$ are estimated separately for each time block λ. Also, to obtain a smoother speech spectral variance $\sigma^2_{\mathcal{Y}_D}(k,l)$, which reasonably enhances the overall performance, a smoothing scheme has been considered for this parameter in (6.14) rather than its instantaneous estimate used in the original method. In our case, the parameter setting d=1 should be considered so that $\mathcal{Y}_{D_j}(k,l)$ in (6.13) inclusively estimates the direct-path component of speech. Even though the WPE method is often implemented for a fixed number of iterations J or until a maximum number of iterations is reached, we found a more efficient heuristic criterion for the number of iterations, which will be discussed in Section 6.4.3.

6.4.2 Estimation of the Reverberant Component

The estimate of the reverberant component $\hat{Y}_R(k,l)$ used in (6.7) largely affects the overall precision of the LRSV estimation and thus the corresponding spectral enhancement method, since $\hat{Y}_R(k,l)$ should exclude the direct-path component of speech to avoid overestimation of the LRSV. To obtain a proper estimate of the reverberant component, here we employ a modification of the correlation-based approach suggested in [143], which was originally proposed to estimate the late reverberant component. This approach models the estimate of the late reverberant speech, $\hat{Y}_L(k,l)$, as a weighted sum of Q previous frames of the dereverberated speech, as the following

$$\hat{Y}_L(k,l) = \sqrt{B}\sum_{q=0}^{Q-1} c_q(k) Y_{de}(k, l-\delta-q) \tag{6.15}$$

where $Y_{de}(k,l)$ is the dereverberated speech, δ is to introduce a delay (on the order of a few frames) to avoid the direct-path and early reverberant components, c_q's are the MA model (prediction) coefficients, Q is taken as 60 and $B = 1.65$ is a bias correction factor [143]. Since we here aim at the estimation of the entire reverberant speech (including the early and late components), we set $\delta = 1$ in the above to skip only the direct-path component and use the dereverberated component obtained from the WPE method for $Y_{de}(k,l)$. This results in

$$\hat{Y}_R(k,l) = \sqrt{B}\sum_{q=0}^{Q-1} c_q(k,\lambda) \hat{\mathcal{Y}}_D(k, l-1-q) \tag{6.16}$$

where we have used the term $\hat{\mathcal{Y}}_D(k,l)$ as an estimate for $Y_{de}(k,l)$, which is obtained by applying the WPE method in Section 6.4.1. Also, in a similar fashion to the reverberation prediction weights $\mathbf{g}_\lambda(k)$, we have considered the prediction coefficients $c_q(k,\lambda)$ to be updated as a function of the time block index λ to account for moderate changes in the environment. Now, it remains to obtain the prediction coefficients $c_q(k,\lambda)$, as required by (6.16). Based on [143], the prediction coefficients can be optimally obtained by minimizing the mean squared error between $Y(k,l)$ and

$c_q(k, \lambda)\hat{\mathcal{Y}}_D(k, l-1-q)$, leading to the following solution

$$\hat{c}_q(k, \lambda) = \frac{E_l\{Y(k,l)\hat{\mathcal{Y}}_D(k, l-1-q)\}}{E_l\{|\hat{\mathcal{Y}}_D(k, l-1-q)|^2\}} \tag{6.17}$$

where $E_l\{.\}$ denotes the expectation over time frame l. Although in [143], the above scheme is not actually followed in order to avoid long-term time averaging, here the block processing framework allows us to perform the time averaging over all frames of the block. In this sense, denoting the terms in the numerator and denominator of (6.17) by $E^{(1)}$ and $E^{(2)}$ respectively, we use the following sample means

$$\begin{aligned} E_1 &\approx \frac{1}{M}\sum_l Y(k,l)\hat{\mathcal{Y}}_D(k, l-1-q) \\ E_2 &\approx \frac{1}{M}\sum_l |\hat{\mathcal{Y}}_D(k, l-1-q)|^2 \end{aligned} \tag{6.18}$$

where we let $l \in \{\lambda M, \lambda M + 1, \cdots, \lambda M + M - 1\}$, i.e., we perform the sample means over the M time frames of the processing block. It should be noted that, even though the incremental-based implementation of the WPE method, as discussed in Section 6.4.1, introduces deviations in the prediction weights $\mathbf{g}_\lambda(k)$ from those obtained through the full batch processing, the WPE method still does a good job at isolating the direct-path component from the reverberant one as obtained by (6.16). Further details regarding the performance of the WPE method based on block processing will be further discussed in Section 6.6.

6.4.3 Incremental Implementation of the WPE Method

The original WPE method essentially requires batch processing using at least a few seconds of the reverberant observation. In spite of this, we apply the WPE method for processing blocks of 0.5 second, since it is employed only to provide primary estimates of the reverberant and dereverberated speech components, which are used in updating the shape parameter $\kappa(k,l)$ in (6.8) and the reverberant speech component $\hat{Y}_R(k,l)$ in (6.16). However, to further increase the accuracy of the underlying WPE method in order to fit it into our incremental processing approach, we make a few modifications to the original version of this method. The first, as discussed in Section 6.4.1, is to add a smoothing scheme for the estimation of the speech spectral variance $\sigma^2_{\mathcal{Y}_D}(k,l)$ in (6.14). Next, we suggest to employ a heuristic criterion for the number of iterations performed in (6.11)-(6.14). Conventionally, a fixed or a maximum number of iterations can be employed, or the following convergence criterion can be used in the jth iteration [146]

$$\frac{\|\mathbf{g}_j(k) - \mathbf{g}_{j-1}(k)\|_2}{\|\mathbf{g}_{j-1}(k)\|_2} < \rho \tag{6.19}$$

with $\|.\|_2$ denoting the ℓ_2-norm and ρ as a fixed threshold value; the iterations are discarded if the above holds. Here, we suggest a convergence criterion based on a heuristic interpretation of the WPE method in [146], as follows. The reverberation prediction weights $\mathbf{g}_j(k)$ can actually be derived based on the minimization of the following cost function [146]

$$F_j(\mathbf{g}_j) = \sum_l \frac{\left|Y(k,l) - \mathbf{g}_j^H(k)\mathbf{Y}(k,l-d)\right|^2}{|\mathcal{Y}_{D_{j-1}}(k,l)|^2} = \sum_l \frac{|\mathcal{Y}_{D_j}(k,l)|^2}{|\mathcal{Y}_{D_{j-1}}(k,l)|^2} \tag{6.20}$$

which in fact penalizes the sparsity of dereverberated speech in the numerator as compared to the anechoic speech in the denominator. Here, we take advantage of the criterion expressed in (6.19) to formulate a more efficient convergence criterion than the one in (6.19) for the reverberation prediction weights at the λth processing block, $\mathbf{g}_{\lambda_j}(k)$, as the following

$$G_j(k,\lambda) = \sum_{l=\lambda M}^{\lambda M+M-1} \frac{|\mathcal{Y}_{D_j}(k,l)|^2}{|\mathcal{Y}_{D_{j-1}}(k,l)|^2} < \rho' \tag{6.21}$$

where the summation is performed on all frames of the λth processing block and the threshold value ρ' is experimentally set to $0.01M$. This choice of the convergence criterion ensures that a certain level of sparsity in the dereverberated speech, as inspired by the cost function in (6.20), is reached before discarding the iterations. Since the values of $|\mathcal{Y}_{D_j}(k,l)|^2$ and $|\mathcal{Y}_{D_{j-1}}(k,l)|^2$ may change dramatically, making the criterion in (6.21) unsuitable for some frequencies or processing blocks, we set the minimum and maximum allowed number of iterations respectively to 2 and 10.

Finally, to smooth the changes of the reverberation prediction weight $\mathbf{g}_\lambda(k)$ across processing blocks, we perform a smoothing scheme on $\mathbf{g}_\lambda(k)$ to obtain its ultimate value, $\mathbf{g}'_\lambda(k)$, as the following

$$\mathbf{g}_{\lambda_{\text{final}}}(k) = [1-\mu]\mathbf{g}_{\lambda-1}(k) + \mu\mathbf{g}_\lambda(k) \tag{6.22}$$

with μ fixed at 0.8, to determine the updated values of $\mathbf{g}'_\lambda(k)$ mostly upon the current processing block.

In Figure 6.2, a block diagram of the main steps of the proposed approach for LRSV estimation is illustrated. It is observed that the estimates of the direct and reverberant components by the WPE method are useful in both updating the shape parameter $\kappa(k,l)$ and the reverberant component $\hat{Y}_R(k,l)$ in the LRSV estimation scheme in (6.7).

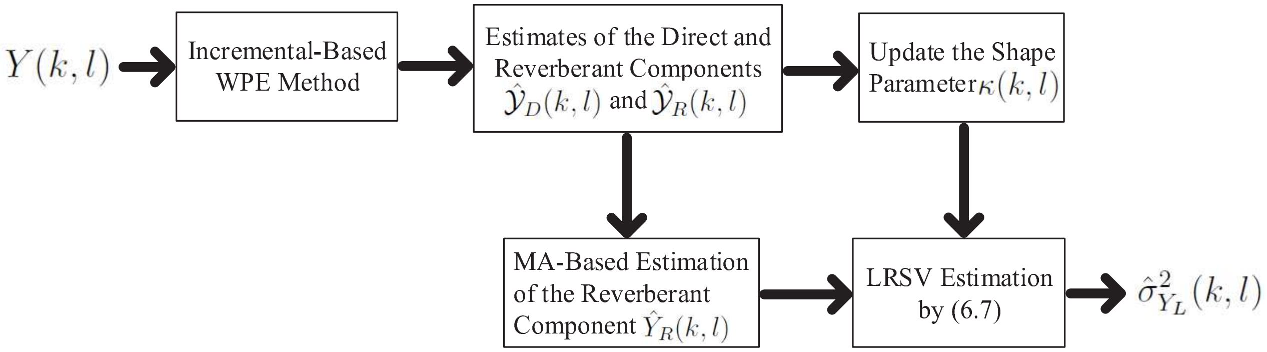

$Y(k,l)$
Incremental-Based WPE Method
Estimates of the Direct and Reverberant Components $\hat{\mathcal{Y}}_D(k,l)$ and $\hat{\mathcal{Y}}_R(k,l)$
Update the Shape Parameter $\kappa(k,l)$
MA-Based Estimation of the Reverberant Component $\hat{Y}_R(k,l)$
LRSV Estimation by (6.7)
$\hat{\sigma}^2_{Y_L}(k,l)$

Figure 6.2: Block diagram of the proposed algorithm for LRSV estimation.

6.5 Other Developments on Classic Spectral Enhancement Methods

In addition to the developed LRSV estimator in the previous part, there are a few major modifications that should be taken into account to efficiently employ the STSA estimation method for the purpose of dereverberation. Even though using each of the proposed schemes in this section may result in trivial improvements, the combination of all the suggested schemes considerably enhances the performance of the STSA estimation used for dereverberation, as compared to the ordinary schemes exploited in the context of noise reduction. The proposed schemes include the estimation of SRR, the flooring of the spectral gain function, use of the SPP in modifying the gain function, and the beamforming (multi-channel) method for dereverberation.

6.5.1 Estimation of SRR

The estimation of SRR in the context of late reverberation suppression, i.e., $\sigma^2_{Y_E}/\sigma^2_{Y_L}$, is basically related to its counterpart in the context of noise reduction, i.e., the *a priori* SNR, which has been conventionally estimated by the DD approach [17]. In this sense, we have the following estimator for the SRR

$$\hat{\zeta}(k,l) = \omega \ \frac{|\hat{Y}_E(k,l-1)|^2}{\hat{\sigma}^2_{Y_L}(k,l-1)} + (1-\omega) \ P\{\eta(k,l)-1\}, \quad 0 \leq \omega < 1 \tag{6.23}$$

with ω as a fixed smoothing parameter and the function $P\{.\}$ defined as

$$P\{x\} = \begin{cases} x, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.24}$$

to avoid the invalid negative values of $\eta(k,l)-1$. Herein, $|\hat{Y}_E(k,l-1)|$ has to be replaced by $G(k,l-1)|Y(k,l-1)|$. To date, there has been a few major modifications and improvements to this classic approach such as those in [147, 148]. In [147], it was proved that an adaptive (i.e., frame and frequency dependent) smoothing parameter, $\omega(k,l)$, can improve the performance of the DD approach and an optimal choice of this parameter based on the MMSE criterion was proposed therein. In [148], it was discussed that the conventional DD approach in fact introduces a delay in the estimation of $\zeta(k,l)$ and that the spectral gain computed at the current frame is more adapted to the previous frame. Therein, to compensate for this inherent delay, it was suggested to shift the frame index in the right hand side of (6.24) by one, and consequently, estimate $\zeta(k,l)$ using the current estimate of the gain function $G(k,l)$ and the estimate of $\eta(k,l+1)$. We here employ a similar approach in our context of SRR estimation and propose the following two-stage method:

- First, we use the following to calculate an initial estimate of the SRR, $\hat{\zeta}_0(k,l)$

  $$\hat{\zeta}_0(k,l) = \omega_0 \ \frac{G_0^2(k,l)|Y(k,l)|^2}{\hat{\sigma}^2_{Y_L}(k,l)} + (1-\omega_0) \ \frac{\hat{\sigma}^2_{Y_E}(k,l+1)}{\hat{\sigma}^2_{Y_L}(k,l+1)} \tag{6.25}$$

  with ω_0 chosen as 0.5, $G_0(k,l)$ is that obtained by using the conventional DD approach for $\hat{\zeta}(k,l)$ in (6.23), and proper estimates for $\hat{\sigma}^2_{Y_L}(k,l)$ and $\hat{\sigma}^2_{Y_E}(k,l)$ can be respectively used from the proposed LRSV estimator in Section 6.4 and the early speech spectral variance estimator in Chapter 5. Note that, as compared to the second term at the right hand side of (6.23), instead of the asymptotically optimal estimate of $\zeta(k,l)$, i.e., $\eta(k,l)-1$, a more meaningful estimate has been exploited, which is not in need of the rectifying function $P\{.\}$.

- Second, we use a more precisely adjusted smoothing expression to obtain the ultimate estimate of $\zeta(k,l)$ as the following

$$\hat{\zeta}(k,l) = \omega_{opt}(k,l)\ \frac{G_1^2(k,l)|Y(k,l)|^2}{\hat{\sigma}_{Y_L}^2(k,l)} + (1-\omega_{opt}(k,l))\ \frac{\hat{\sigma}_{Y_E}^2(k,l+1)}{\hat{\sigma}_{Y_L}^2(k,l+1)} \tag{6.26}$$

where the gain $G_1(k,l)$ is obtained by using the estimated $\hat{\zeta}_0(k,l)$ from the first step and $\omega_{opt}(k,l)$ is the optimal smoothing parameter that can be derived in the MMSE sense. In the sequel, the derivation of $\omega_{opt}(k,l)$ is discussed.

Following an MMSE framework to estimate $\omega_{opt}(k,l)$, the MSE between the true and estimated SRR, i.e., $E\{(\zeta(k,l)-\hat{\zeta}(k,l))^2\}$, has to be minimized. By using (6.26) for $\hat{\zeta}(k,l)$ and expanding the MSE term, we obtain

$$\begin{aligned}\text{MSE} = \omega^2(k,l)E\{\mathcal{A}^2\} + (1-\omega(k,l))^2E\{\mathcal{B}^2\} + 2\omega(k,l)E\{\mathcal{A}\}(1-\omega(k,l))E\{\mathcal{B}\} \\ - 2\zeta(k,l)\Big[\omega(k,l)E\{\mathcal{A}\} + (1-\omega(k,l))E\{\mathcal{B}\}\Big]\end{aligned} \tag{6.27}$$

where $\mathcal{A}$ and $\mathcal{B}$ respectively denote $\frac{G_1^2(k,l)|Y(k,l)|^2}{\hat{\sigma}_{Y_L}^2(k,l)}$ and $\frac{\hat{\sigma}_{Y_E}^2(k,l+1)}{\hat{\sigma}_{Y_L}^2(k,l+1)}$. It can be seen that the MSE term in (6.27) is a quadratic function of $\omega(k,l)$ and that its second derivative with respect to $\omega(k,l)$ is positive. Therefore, the MSE function is convex and by setting to zero its first derivative with respect to $\omega(k,l)$, the optimal value of $\omega(k,l)$ can be derived as the following

$$\omega_{opt}(k,l) = \frac{E\{\mathcal{B}^2\} - E\{\mathcal{A}\}E\{\mathcal{B}\} + \zeta(k,l)\,(E\{\mathcal{A}\} - E\{\mathcal{B}\})}{E\{\mathcal{A}^2\} + E\{\mathcal{B}^2\} - 2E\{\mathcal{A}\}E\{\mathcal{B}\}} \tag{6.28}$$

Clearly, the expectation on $\mathcal{B}$ or $\mathcal{B}^2$ can be dropped since it is a deterministic term. Moreover, it can be deduced that $E\{\mathcal{A}\}$ is in fact equal to $\zeta(k,l)$, and $E\{\mathcal{A}^2\}$ can be expressed as $\frac{G_1^4(k,l)}{\hat{\sigma}_{Y_L}^4(k,l)}E\{|Y(k,l)|^4\}$, which can be obtained by smoothing the values of $|Y(k,l)|^4$. Wherever needed, we choose to use $\hat{\zeta}_0(k,l)$ as an estimate for $\zeta(k,l)$ in calculating $\omega_{opt}(k,l)$ by (6.28). Also, to ensure that the smoothing parameter $\omega_{opt}(k,l)$ always falls into the interval $[0,1]$ and it results in the best performance, we lower and upper bound the values given by (6.28) by 0.1 and 0.7, respectively.

6.5.2 Application of SPP to Gain function

Use of speech presence probability (SPP) to modify the gain function of an STSA estimator in a noisy field was discussed in Section 2.1.7 of Chapter 2. In this part, we develop a straightforward extension of the SPP that can be used to modify the STSA gain function in a reverberant field, as the following

$$G_M(k,l) = G_{\mathcal{H}_1}(k,l)\mathcal{P}(\mathcal{H}_1|Y(k,l)) + G_{\mathcal{H}_0}(k,l)\mathcal{P}(\mathcal{H}_0|Y(k,l)) \tag{6.29}$$

where $G_{\mathcal{H}_1}(k,l)$ and $G_{\mathcal{H}_0}(k,l)$ respectively denote the gain functions under the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$. Obviously, $G_{\mathcal{H}_1}(k,l)$ is the conventional gain function without taking into account the SPP. Whereas $G_{\mathcal{H}_0}(k,l)$ is theoretically zero, in practice, a small value, $G_{min}(k,l)$, is considered for this term [78, 79]. Since we target late reverberation suppression in a reverberant field, the two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$ are defined as

$$\begin{aligned} &\mathcal{H}_1 : Y(k,l) = Y_E(k,l) + Y_L(k,l) \\ &\mathcal{H}_0 : Y(k,l) = Y_L(k,l) \end{aligned} \tag{6.30}$$

We here suggest a simple yet effective development of the SPP term, $\mathcal{P}(\mathcal{H}_1|Y(k,l))$, by looking at the linear prediction model for $Y_E(k,l)$ as $\sum_{\ell=0}^{D-1} g^*(k,\ell)Y(k,l-\ell)$ in Chapter 5. In this sense, it is evident that the presence of $Y_E(k,l)$ not only depends on the observation at the current frame $Y(k,l)$, but also depends on the observations at $D-1$ previous frames. Therefore, we suggest to replace the conditional probability $\mathcal{P}(\mathcal{H}_1|Y(k,l))$ by $\mathcal{P}(\mathcal{H}_1|\mathbf{Y}_D(k,l))$ as the SPP in a reverberant field, with $\mathbf{Y}_D(k,l)$ as $[Y(k,l), Y(k,l-1), \cdots, Y(k,l-D+1)]^T$. Now, we use the same conventional Bayesian framework to obtain $\mathcal{P}(\mathcal{H}_1|\mathbf{Y}_D(k,l))$, as follows

$$\mathcal{P}(\mathcal{H}_1|\mathbf{Y}_D(k,l)) = \frac{\mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_1)\mathcal{P}(\mathcal{H}_1)}{\mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_1)\mathcal{P}(\mathcal{H}_1) + \mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_0)\mathcal{P}(\mathcal{H}_0)} \tag{6.31}$$

and to obtain $\mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_1)$ and $\mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_0)$, we use the complex independent Gaussian model, as the following

$$\mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_1) = \prod_{\ell=0}^{D-1} \mathcal{P}(Y(k,l-\ell)|\mathcal{H}_1) = \frac{\exp\left(-\sum\limits_{\ell=0}^{D-1} \frac{|Y(k,l-\ell)|^2}{\sigma^2_{Y_E}(k,l-\ell)+\sigma^2_{Y_L}(k,l-\ell)}\right)}{\pi^D \prod_{\ell=0}^{D-1}(\sigma^2_{Y_E}(k,l-\ell)+\sigma^2_{Y_L}(k,l-\ell))}$$

$$\mathcal{P}(\mathbf{Y}_D(k,l)|\mathcal{H}_0) = \prod_{\ell=0}^{D-1} \mathcal{P}(Y(k,l-\ell)|\mathcal{H}_0) = \frac{\exp\left(-\sum\limits_{\ell=0}^{D-1} \frac{|Y(k,l-\ell)|^2}{\sigma^2_{Y_L}(k,l-\ell)}\right)}{\pi^D \prod_{\ell=0}^{D-1}\sigma^2_{Y_L}(k,l-\ell)} \tag{6.32}$$

where the LRSV and speech spectral variance estimation methods proposed previously can be used to estimate $\sigma^2_{Y_L}(k,l-\ell)$ and $\sigma^2_{Y_E}(k,l-\ell)$ respectively. Now by replacing (6.32) into (6.31) and using a fixed choice for $\mathcal{P}(\mathcal{H}_1)$, the proposed SPP can be calculated. In this work, we choose D and $\mathcal{P}(\mathcal{H}_1)$ to be 3 and 0.75 respectively. It should be noted that, while in the conventional literature such as the Cohen's IMCRA method [80], $\mathcal{P}(\mathcal{H}_1)$ is suggested to be a function of time/frequency, it was found in more recent works that assuming a fixed value for this parameter does not considerably change the performance of the calculated SPP [149].

6.5.3 Spectral Gain Flooring for Dereverberation

In the same manner as that in the context of noise reduction, in frequency bins with weaker direct-path (early) speech components, the gain function $G(k,l)$ of an STSA estimator may approach values close to or almost zero. This introduces too much of attenuation on the present early speech components, which is perceived as large distortions in the enhanced speech signal. We experimentally found that the resulting distortion is even higher in case of reverberation suppression than the conventional noise reduction by STSA estimation, which is perhaps due to the overestimation in the LRSV estimator and the correlation of early and late reverberant components. Furthermore, we subjectively found that applying a larger gain function than the regular, despite providing less suppression of the late reverberation, results in a more pleasant enhanced speech. Therefore, in the sequel, we suggest a gain flooring scheme for the STSA gain function in reverberant environments, which is both efficient and simple. This scheme can be applied on the modified gain function in (6.29).

Resorting to the gain flooring schemes in the context of noise reduction, such as the one proposed in Section 3.3.3 of Chapter 3, it follows that the gain function $G_M(k,l)$ can be modified as

$$G'_M(k,l) = \begin{cases} G_M(k,l), & \text{if } G_M(k,l) > \rho_1(k,l) \\ G_f(k,l), & \text{otherwise} \end{cases} \tag{6.33}$$

where $\rho(k,l)$ is a threshold value depending on the noise masking threshold as given by (3.11), and $G_f(k,l)$ is the flooring value of the gain function. Following the same scheme as in (6.33), we here suggest a choice for the flooring value, $G_f(k,l)$. To this end, we consider the linear prediction model for the early speech component, as discussed in Chapter 5, where an estimate of $Y_E(k,l)$, say $\hat{Y}_E(k,l)$, can be written as the sum $\sum_{\ell=0}^{D-1} g^*(k,\ell)Y(k,l-\ell)$. In this sense, to estimate the prediction weights $g(k,\ell)$, we minimize the MSE between $\hat{Y}_E(k,l)$ and $g^*(k,\ell)Y(k,l-\ell)$, which results in the following solution

$$\hat{g}(k,\ell) = \frac{E\{\hat{Y}_E(k,l)Y^*(k,l-\ell)\}}{E\{|Y(k,l-\ell)|^2\}} \tag{6.34}$$

Now, by using $G_M(k,l)Y(k,l)$ for $\hat{Y}_E(k,l)$, and also using sample averaging over l for the expectation $E\{.\}$ in (6.34), the prediction weights $\hat{g}(k,\ell)$ and therefore $\hat{Y}_E(k,l)$ can be obtained. We now define the flooring value of the gain function as an attenuated version of the smoothed gain function, as the following

$$G_f(k,l) = C_f \, \frac{\hat{Y}_E(k,l)}{|Y(k,l)|} \tag{6.35}$$

with C_f as a fixed attenuation factor empirically set as 0.25. The parameter D was taken to be 3 for the suggested scheme in this section.

6.5.4 Beamforming for Late Reverberation Suppression

Beamforming methods have been used since long in the past for the purpose of noise reduction. As the most important and highly used methods in this regard, the multi-channel Wiener filter and the minimum variance distortionless response (MVDR) can be mentioned [46]. Furthermore, in Chapter 4, it was seen that the multi-channel extension of the STSA estimators under a complex Gaussian distribution for noise results in the MVDR beamformer plus an STSA post-filter. Most of the beamforming methods including the MVDR method require an estimate of the PSD matrix of the background noise/disturbance, and in fact, their capability to suppress the ambient disturbance highly depends on the precision in the estimation of the PSD matrix. In case of reverberation suppression, our literature survey revealed that, to date, a very efficient and promising approach to blindly estimate the PSD matrix of the late reverberant speech that can be integrated into the MVDR beamformer does not exist. Therefore, we here suggest a straightforward extension of the conventional LRSV estimator to estimate the PSD matrix of interest. Denoting this matrix by $\Phi_{\mathbf{Y}_L}(k,l){=}E\left\{\mathbf{Y}_L\mathbf{Y}_L^H\right\}$ with $\mathbf{Y}_L(k,l){=}[Y_{L_1}(k,l), Y_{L_2}(k,l), \cdots, Y_{L_M}(k,l)]^T$ and $Y_{L_m}(k,l)$ as the late reverberant component in the mth microphone, we can divide the elements of the matrix into auto- and cross-terms (i.e., the diagonal and non-diagonal entries), as seen in the following

$$\Phi_{\mathbf{Y}_L}(k,l) = \begin{bmatrix} E\{|Y_{L_1}|^2\} & E\{Y_{L_1}Y_{L_2}^*\} & \dots & E\{Y_{L_1}Y_{L_M}^*\} \\ E\{Y_{L_2}Y_{L_1}^*\} & E\{|Y_{L_2}|^2\} & \dots & E\{Y_{L_2}Y_{L_M}^*\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{Y_{L_M}Y_{L_1}^*\}\} & E\{Y_{L_M}Y_{L_2}^*\}\} & \dots & E\{|Y_{L_M}|^2\} \end{bmatrix} \tag{6.36}$$

As for the estimation of the diagonal elements, conventional LRSV estimators such as those discussed in Section 6.4 can be used. For example, by using Lebart's method [142], we have

$$\begin{aligned} E\{|Y_{L_m}(k,l)|^2\} &= e^{-2\alpha(k)PN_E}E\{|Y_m(k,l-N_E)|^2\} \\ E\{|Y_m(k,l)|^2\} &= [1-\beta]E\{|Y_m(k,l-1)|^2\} + \beta|Y_m(k,l)|^2 \end{aligned} \tag{6.37}$$

For the estimation of the off-diagonal elements, $E\{Y_{L_m}Y_{L_n}^*\}$, a straightforward extension of (6.37) can be used as the following

$$\begin{aligned} E\{Y_{L_m}(k,l)Y_{L_n}^*(k,l)\} &= e^{-2\alpha(k)PN_E}E\{Y_m(k,l-N_E)Y_n^*(k,l-N_E)\} \qquad (6.38) \\ E\{Y_m(k,l)Y_n^*(k,l)\} &= [1-\beta]E\{Y_m(k,l-1)Y_n^*(k,l-1)\} + \beta Y_m(k,l)Y_n^*(k,l) \end{aligned}$$

Therefore, the PSD matrix, $\Phi_{\mathbf{X}_L}(k,l)$, can be obtained by

$$\begin{aligned} \Phi_{\mathbf{Y}_L}(k,l) &= e^{-2\alpha(k)PN_E}\ \Phi_{\mathbf{Y}}(k,l-N_E) \\ \Phi_{\mathbf{Y}}(k,l) &= [1-\beta]\ \Phi_{\mathbf{Y}}(k,l-1) + \beta\ \mathbf{Y}(k,l)\mathbf{Y}^H(k,l) \end{aligned} \tag{6.39}$$

Now, noting that the MVDR beamformer weights can be represented as $\frac{\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{Y}_L}^{-1}(k,l)}{\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{Y}_L}^{-1}(k,l)\mathbf{A}(k)}$ with $\mathbf{A}(k)$ as the steering vector, it is evident that by inserting the suggested estimate of $\Phi_{\mathbf{Y}_L}(k,l)$ by (6.39) into the MVDR beamformer, the term $e^{-2\alpha(k)PN_E}$ is canceled out. Therefore, the MVDR beamformer weights become independent of the reverberation parameter $\alpha(k)$=$3\log 10/(f_s T_{60dB})$, and the estimation of the reverberation time T_{60dB} is not required. In this regard, the delayed version of the PSD matrix, $\Phi_{\mathbf{Y}}(k,l)$, i.e., $\Phi_{\mathbf{Y}}(k,l-N_E)$ can be used in the MVDR beamformer and inversion algorithms such as the Sherman-Morrison formula can be exploited to calculate its inverse. The parameters N_E and β in (6.39) are respectively chosen as 5 and 0.1. We did not find any performance advantage by extending more advanced estimators of the LRSV (such as the one proposed in Section 6.4) into the matrix form and using them in the MVDR beamformer. Yet, this can be considered as an avenue for further research.

6.6 Performance Evaluation

In this section, we first assess the performance of the suggested LRSV estimator in Section 6.4, as well as the proposed schemes in Section 6.5, including the estimator of SRR, the extension of SPP and the spectral gain flooring. Next, based on the existing literature and proposed methods, we exploit a few schemes for the joint suppression of noise and late reverberation and evaluate their performance in a noisy reverberant scenario.

6.6.1 Evaluation of the Proposed LRSV Estimator

In this section, we evaluate the performance of the proposed estimator of LRSV against other recent LRSV estimation methods for both time-invariant and time-varying RIRs. To this end, anechoic speech utterances including 10 male and 10 female speakers were used from the TIMIT database [105]. The sampling frequency f_s was set to 16 kHz and a 25 msec Hamming window with the overlap of 75% was used for the STFT analysis-synthesis. To implement our block processing-based approach, we considered a block length of 0.5 second, resulting in M=80 overlapping STFT frames in each processing block (note that M can be calculated by dividing the block length Δ by the STFT hop size P). It should be noted that there exists a trade-off in choosing the processing block length, since the smaller the block length the more erroneous the prediction weights $\mathbf{g}_\lambda(k)$, and the larger the block length the higher the processing delay and also the slower the adaptation of the estimated LRSV with the changing RIR. With the current setting for the processing block length, considering the computational complexity of the underlying WPE method (this has been studied in detail in terms of the real time factor in [126]), the proposed approach seems suitable for real time applications for which the dereverberation algorithm needs to work incrementally from the beginning of the captured speech utterance with a small algorithmic delay. The number of early terms considered in the RIR in (6.5), i.e., N_E, is set as 10 for our experiments to obtain the best performance. As for the estimation of the reverberation time T_{60dB}, we use the blind reverberation time estimator in [133], which is capable of estimating T_{60dB} within the allowed processing blocks with low complexity and enough accuracy for the purpose of LRSV estmation. Note that, even for changing environments, the reverberation time parameter T_{60dB} does not change considerably [143]. Our approach does not require the estimation of the DRR parameter.

Using the same performance measures as those described in Chapter 5, in the following, we evaluate the relative performance of the proposed LRSV estimator in both time-invariant and time-varying reverberant environments.

6.6.1.1 Performance in Time-Invariant RIRs

In this part, we assess the performance of the proposed approach in comparison with other methods in a scenario where the environment is invariant, using either a synthesized or a measured constant RIR. In case of the measured RIR, we used a recorded RIR taken from the SimData of the REVERB Challenge [134], where an 8-channel circular array with diameter of 20 cm was placed in a 3.7 m×5.5 m acoustic room. The RIR at the first channel was considered as the observation. The resulting signal was combined with additive babble noise from the same database at a global SNR of 10 dB. Furthermore, to have a controlled reverberation time and be able to verify the performance of the proposed approach in low-to-moderate T_{60dB} values, we used the ISM method [119] to synthesize RIRs with different T_{60dB}. In all cases, the anechoic speech is convolved with the RIR to obtain the reverberant speech signal. The geometry of the synthesized reverberant environment scenario with T_{60dB} changing from 100 msec to 800 msec is shown in Figure 6.3 in detail. The reverberant global SNR was fixed at 15 dB for this experiment.

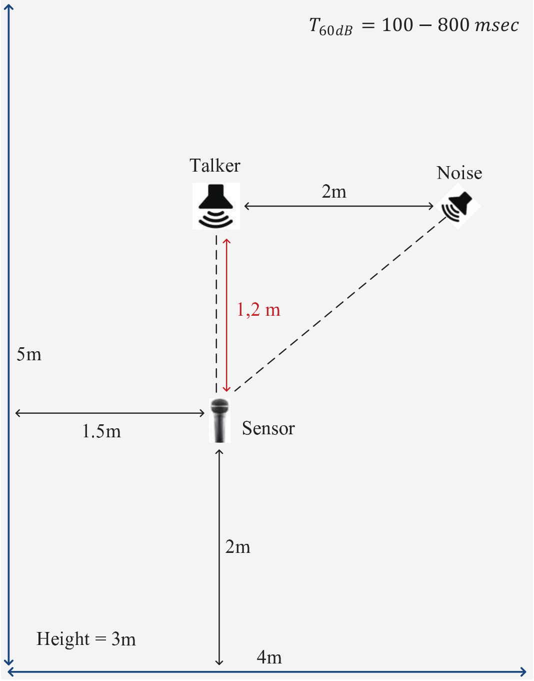


Figure 6.3: A two-dimensional schematic of the geometric setup used to synthesize the time-invariant RIR by the ISM method.

In the case of time-invariant RIR, we compare the proposed approach to the original Lebart's method [142], the correlation-based method in [143], the improved model-based method in [144] and the true (perfect) LRSV estimator. The Lebart's method is actually a special case of the scheme in (6.5) with $\kappa(k)$=1. The correlation-based method, as expressed by eq. (26) in [143], is actually based on obtaining $\hat{X}_R(k,l)$ by (6.15) and then smoothing it to estimate the LRSV. Yet, due to the unavailability of long-term expectations in (6.17), it uses a recursive smoothing scheme to find the prediction coefficients $c_q(k)$. The improved model-based method in [144] uses more than one term of the past spectral variances of the reverberant speech in order to obtain a smoother shape parameter and is in fact an extension of the model-based method in [143]. The latter (as expressed by eq. (51) in [143]), which is to be assessed in the following section, exploits the past estimates of the LRSV averaged over frequency bins to obtain the shape parameter as $\kappa(l)$. It should be noted that the correlation-based and model-based methods in [143] were respectively developed for time-invariant and time-variant RIRs. Finally, the true LRSV (used just as a reference for comparison) was obtained by temporally smoothing the late reverberant spectra, which is in turn calculated by convolving the anechoic speech with the hypothetical late component of the RIR, i.e., that excluding the first 60 msec.

To evaluate the efficiency of the proposed method with respect to the length of the processing blocks, we calculated a measure of the error in the estimation of LRSV versus the processing block length for different reverberation times, as shown in Figure 6.4.

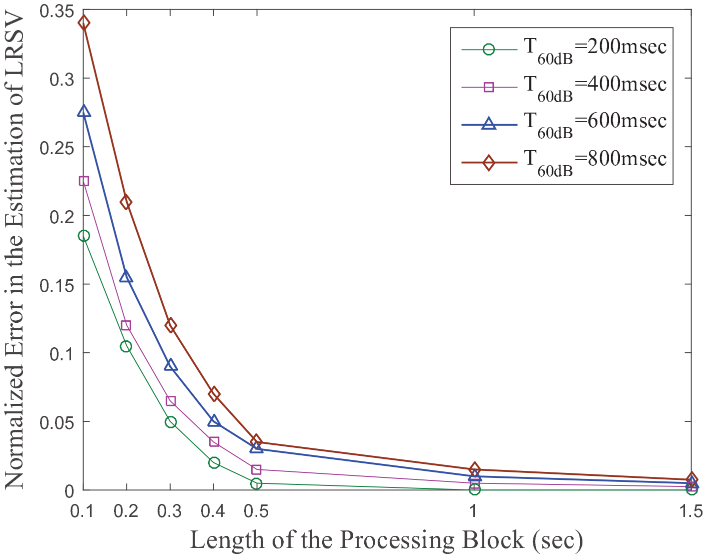


Figure 6.4: Normalized error in the estimation of the LRSV w.r.t. to the case of using the entire speech utterance, versus the processing block length for different reverberation times.

We considered processing a speech segment of 3 seconds using block lengths of 0.1 to 1.5 seconds to estimate the LRSV and calculated the following normalized error

$$e(\Delta) = E_l \left\{ \frac{||\hat{\bar{\sigma}}^2_{X_L}(k,l,\Delta) - \bar{\sigma}^2_{X_L}(k,l)||_2}{||\bar{\sigma}^2_{X_L}(k,l)||_2} \right\} \tag{6.40}$$

where $\hat{\bar{\sigma}}^2_{X_L}(k,l,\Delta)$ and $\bar{\sigma}^2_{X_L}(k,l)$ respectively denote the estimated LRSV using a block length of Δ and that using the entire utterance, $||.||_2$ is the ℓ_2-norm over frequency bins and $E_l\{.\}$ is the average value over time frames. As observed in Figure 6.4, for processing block lengths of 0.5 second, the relative error in the LRSV estimation is far smaller than that for shorter blocks of around 0.1 to 0.2 second, yet not much larger than that for longer blocks of 1 or even 1.5 seconds. In fact, even though choosing a longer processing block reduces the error in (6.40), due to the processing delay introduced by the incremental processing, a trade-off has to be considered in the choice of the block length. This was chosen in our case as 0.5 second.

Next, to determine how close the estimated LRSVs are with respect to the true LRSV, we investigated the mean spectral variances, which are obtained by averaging the LRSVs over all frequency bins, as suggested by [127]. The results are indicated in Figure 6.5 for a speech utterance of 425 time frames and all values were thresholded for better illustration.

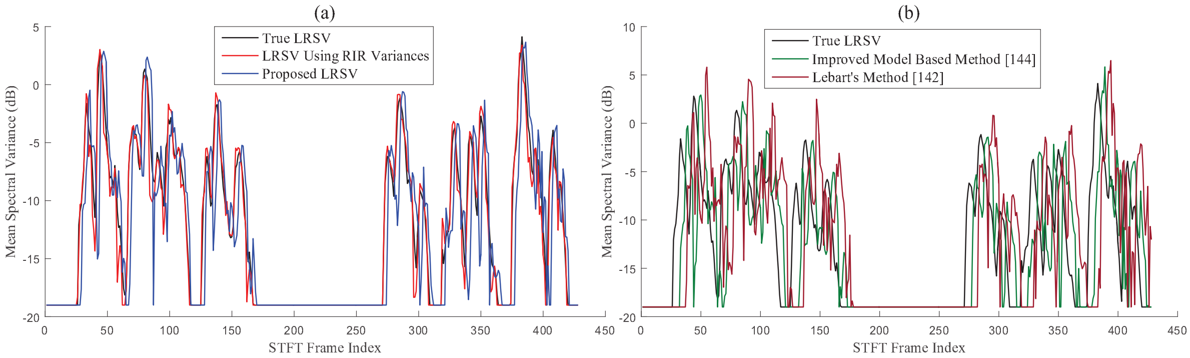


Figure 6.5: Mean spectral variances using the recorded RIR from the REVERB Challenge [134] for: (a) the true LRSV, the LRSV estimated using RIR variances and the proposed LRSV (b) the true LRSV, the LRSV estimated by the improved model-based method [144] and the one estimated by the Lebart's method [142].

In order to examine how fast the methods can track abrupt changes in the LRSV values, we considered a short period of anechoic speech deactivation around the middle of the utterance. In Figure 6.5 (a), the mean spectral variance of the proposed LRSV compared to that of the true LRSV and the one obtained by using RIR variances have been shown. The latter, used as another reference method for comparison, was obtained by using the available constant RIR, i.e., $h(n)$, to calculate the DRR as following

$$\mathrm{DRR}(k) = \frac{\sum_{n=0}^{N_E-1}[h(n)]^2}{\sum_{n=N_E}^{L_h-1}[h(n)]^2} \quad \forall k \tag{6.41}$$

and then using it in (6.6) to obtain the shape parameter κ, which is to be used in the scheme in (6.5), as proposed in [127]. It is observed that the proposed LRSV is able to closely track the true LRSV and the one obtained by using RIR variances, even in the duration of the abrupt drops/rises in the LRSV. As seen in Figure 6.5 (b), Lebart's [142] and the improved model-based [144] methods still follow the LRSV but with larger errors and more delay with respect to the true LRSV, with Lebart's method significantly overestimating the LRSV at the peaks.

To evaluate numerically the error in the proposed and considered LRSV estimates with respect to the true LRSV estimate, the mean segmental error for different reverberation times was calculated and shown in Figure 6.6. The mean segmental error is calculated by [127]

$$\text{Err}_{\text{seg}} = E_l \left\{ \frac{E_k \left\{ |\hat{\sigma}^2_{X_L}(k,l) - \sigma^2_{X_L}(k,l)|^2 \right\}}{E_k \left\{ |\sigma^2_{X_L}(k,l)|^2 \right\}} \right\} \tag{6.42}$$

where $\hat{\sigma}^2_{X_L}(k,l)$ and $\sigma^2_{X_L}(k,l)$ are respectively the estimated and true LRSVs, and with $E_l\{.\}$ and $E_k\{.\}$ respectively denoting the expectation over time frames and frequency bins. As seen in Figure 6.6, for both source-to-microphone distances of 1 m and 2 m, the proposed LRSV estimator attains smaller errors for the entire range of T_{60dB}, as compared to the other methods. Whereas the improved model-based method in [144] and correlation-based method [143] achieve almost close performance, Lebart's method results in the highest error of all.

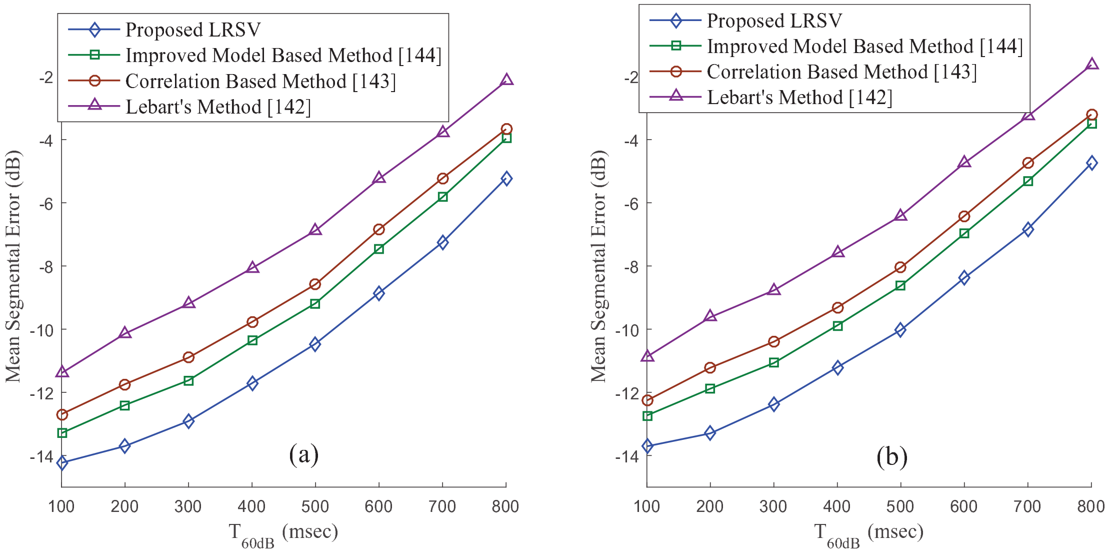


Figure 6.6: Mean segmental error for different LRSV estimators using the synthesized RIRs by the ISM method [119] with source-to-microphone distances of (a): 1 m (b): 2 m.

In order for the evaluation of the reverberation suppression achieved by exploiting the proposed LRSV estimation method, we employed the popular Bayesian log-spectral amplitude (LSA) gain function in [18] to perform late reverberation suppression using the true and estimated LRSVs. The *a priori* SRR required by the gain function was estimated by the decision-directed approach [17], and to obtain the best subjective performance, the LSA gain function was lower bounded to -10 dB. In Table 6.1, the four aforementioned performance scores PESQ, CD, FW-SNR and SRMR have been respectively shown for the unprocessed (observed) speech and the enhanced one by using the true LRSV, proposed LRSV, improved model-based method in [144], correlation-based method in [143] and Lebart's method in [142]. These results were obtained by using the recorded RIR from the REVERB Challenge dataset [134]. Also, the same performance scores have been reported in Table 6.2 for the case of synthesized RIRs using the ISM method with T_{60dB} changing from 200 msec to 800 msec and the source-to-microphone distance of 1 m. It is seen that the proposed method is able to achieve the closest performance to the true LRSV as compared to the others. While the improved model-based method performs slightly better than the correlation-based method, Lebart's method has the lowest scores. Furthermore, it can be deduced that as T_{60dB} is increased, the performance of all LRSV estimation methods degrades with respect to that of the true LRSV, indicating that precise estimation of the LRSV is still a challenging problem for highly reverberant environments. This is consistent with the results obtained for the mean segmental error in Figure 6.6. Table 6.3 shows the same trend but for a source-to-microphone distance of 2 m, resulting in slightly degraded performance results compared to the previous case. We found that the results of the four performance measures used here were almost consistent for different methods.

Table 6.1: Performance measures using the recorded RIR from the REVERB Challenge.

Method	PESQ	CD	FW-SNR (dB)	SRMR (dB)
Unprocessed	1.87	4.97	3.64	4.04
True LRSV	2.25	4.40	6.70	6.74
Proposed method	2.13	4.61	5.89	5.91
Improved model-based [144]	2.03	4.82	5.26	5.58
Correlation-based [143]	1.97	4.88	5.10	5.52
Lebart's method [142]	1.88	5.03	4.65	5.11

Table 6.2: Performance measures using the ISM method for source-to-microphone distance of 1 m.

PESQ

T_{60dB} (msec)	200	400	600	800
Unprocessed	2.31	2.14	1.92	1.78
True LRSV	2.83	2.61	2.37	2.16
Proposed method	2.75	2.48	2.21	1.97
Improved model-based [144]	2.71	2.43	2.14	1.90
Correlation-based [143]	2.70	2.41	2.12	1.88
Lebart's method [142]	2.63	2.32	1.99	1.81

CD

T_{60dB} (msec)	200	400	600	800
Unprocessed	3.72	4.06	4.65	5.48
True LRSV	3.03	3.39	4.11	5.06
Proposed method	3.12	3.51	4.26	5.24
Improved model-based [144]	3.18	3.59	4.34	5.33
Correlation-based [143]	3.20	3.63	4.37	5.36
Lebart's method [142]	3.26	3.73	4.48	5.44

Table 6.3: Performance measures using the ISM method for a source-to-microphone distance of 2 m.

PESQ

T_{60dB} (msec)	200	400	600	800
Unprocessed	2.28	2.12	1.87	1.75
True LRSV	2.81	2.59	2.33	2.15
Proposed method	2.72	2.46	2.20	1.94
Improved model-based [144]	2.68	2.39	2.10	1.88
Correlation-based [143]	2.66	2.38	2.09	1.86
Lebart's method [142]	2.60	2.29	1.96	1.78

CD

T_{60dB} (msec)	200	400	600	800
Unprocessed	3.76	4.08	4.71	5.57
True LRSV	3.08	3.45	4.20	5.15
Proposed method	3.16	3.56	4.31	5.23
Improved model-based [144]	3.21	3.63	4.40	5.39
Correlation-based [143]	3.24	3.67	4.46	5.42
Lebart's method [142]	3.30	3.78	4.57	5.51

6.6.1.2 Performance in Time-Varying RIRs

In this part, we evaluate the relative performance of the proposed LRSV estimation method in a scenario where the RIR is time-variant. In Figure 6.7, an illustration of this scenario used in the ISM method to generate the corresponding impulse responses is shown. As seen, a talker is moving from the indicated location at t=0 to the ending position at t=10 seconds on a straight line, resulting in a variable source-to-microphone channel impulse response. We estimated the continuous trajectory by 20 discrete points and obtain the corresponding RIR for each point through the ISM method. Then, the entire 10-second anechoic sample was segmented into 20 utterances each of which was filtered by one of the RIRs at the discrete points. The entire reverberant speech sample was generated next by combining the 20 individual segments. In this way, even though there exists some error due to the truncation done by the filtering, the length of the reverberant speech sample remains the same as that of the anechoic speech whereas the continuous trajectory is well approximated by the 20 discrete points.

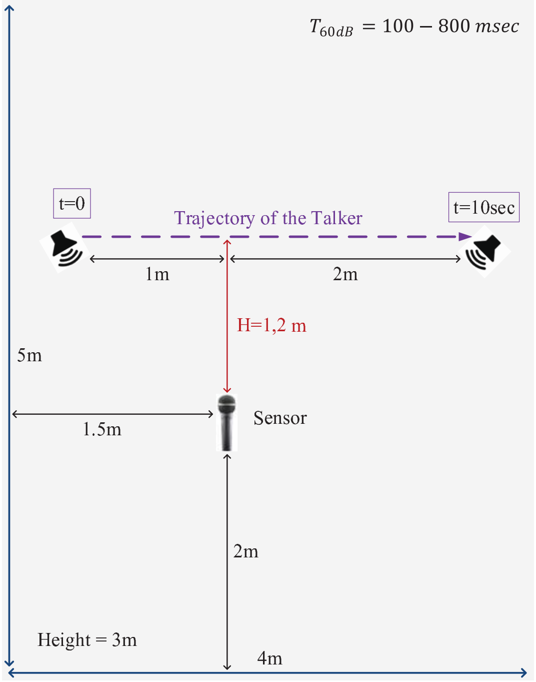


Figure 6.7: A two-dimensional schematic of the geometric setup used to synthesize the time-variant RIR (moving talker) by the ISM method.

In Figure 6.8, the mean spectral variances are shown for the true LRSV, the one obtained by the available RIR variances, the estimated LRSV by the proposed and other methods. It is evident that, whereas the proposed method is able to follow the true LRSV with visibly good precision, the other indicated methods track the changes in the true LRSV with considerable error, which becomes even larger in the locations of sudden decays and rises. Yet, the proposed LRSV estimator proves to be more robust against the abrupt changes in the LRSV due to its adaptation with the changing RIR by employing the incremental-based WPE method.

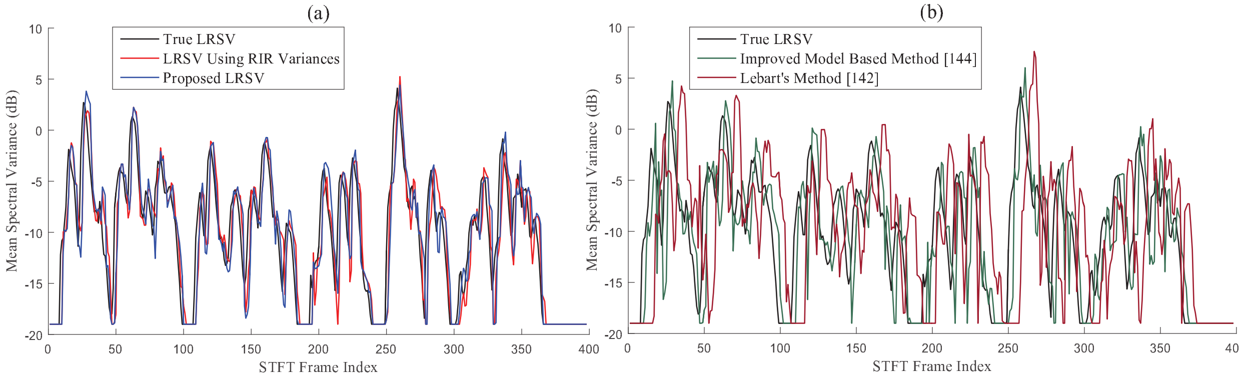


Figure 6.8: Mean spectral variances for: (a) the true LRSV, the LRSV estimated using RIR variances and the proposed LRSV (b) the true LRSV, the LRSV estimated by the improved model-based method [144] and the one estimated by the Lebart's method [142].

Next, we evaluated the mean segmental error in (6.42) in the case of time-varying RIR for the proposed method along with the improved model-based method in [144], the model-based method in [143] and Lebart's method [142]. As observed in Figure 6.9, the same trend as that for the time-invariant RIR applies with the proposed method achieving the best similarity to the true value of LRSV, whereas the model-based and improved model-based methods gain almost the same performance particularly at high reverberation times.

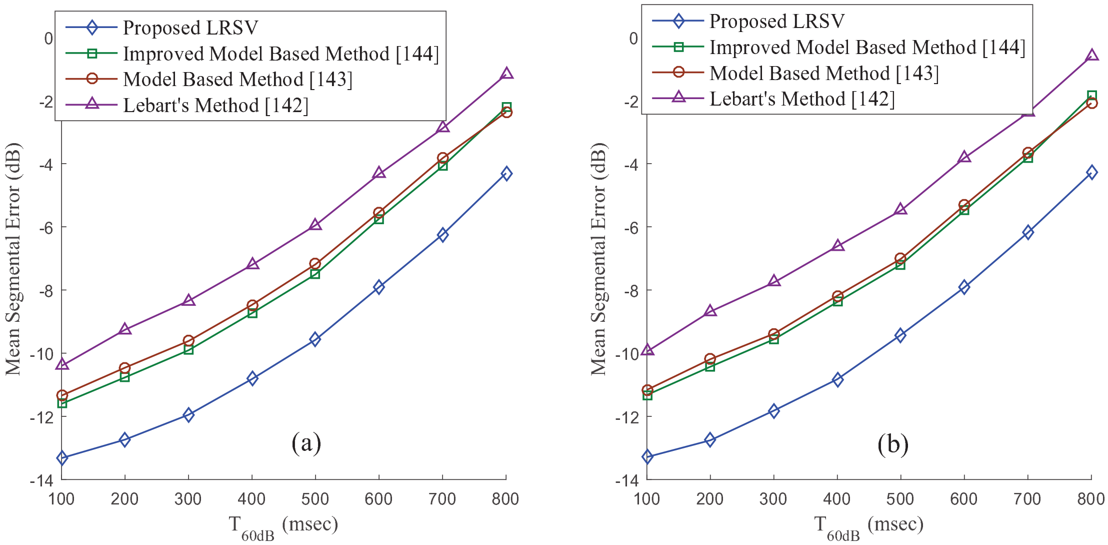


Figure 6.9: Mean segmental error for different LRSV estimators using the configuration in Figure 6.7 with H as (a): 1 m (b): 2 m.

Similar to Section 6.6.1.1, we here evaluate the reverberation suppression performance of the proposed and other methods in terms of the PESQ, CD, FW-SNR and SRMR scores. In this respect, we consider the time-variant RIR scenario in Figure 6.7 and compare our LRSV estimation method with the true LRSV, the improved model-based method in [144], the model-based method in [143] and Lebart's method in [142]. The results have been reported in Table 6.4 for the vertical distance H in Figure 6.7 as 1 m, and in Table 6.5 for H=2 m. Based on these results, we deduce that, in general, the performance scores of all methods falls lower than those in case of time-invariant RIR. Consistently in all the performance scores, it is observed that the proposed method achieves considerably closer scores to those obtained by the true LRSV, even in higher reverberation times where the performance of all methods becomes farther from that of the true LRSV. This shows the advantage of the proposed method especially for changing environments. Also, it is seen that while the improved model-based and model-based methods result in almost close scores, the performance of Lebart's method, i.e., that with a constant shape parameter, is deteriorated further than that in the case of time-invariant RIR. This shows the importance of adapting the shape parameter in LRSV estimation to the changing RIR.

Table 6.4: Performance measures for time-variant RIR with H=1 m in Figure 6.7.

PESQ

T_{60dB} (msec)	200	400	600	800
Unprocessed	2.28	2.13	1.92	1.77
True LRSV	2.76	2.58	2.29	2.10
Proposed method	2.71	2.40	2.13	1.93
Improved model-based [144]	2.66	2.35	2.10	1.84
Model-based [143]	2.66	2.36	2.09	1.83
Lebart's method [142]	2.56	2.27	1.93	1.76

CD

T_{60dB} (msec)	200	400	600	800
Unprocessed	3.80	4.09	4.65	5.49
True LRSV	3.16	3.54	4.26	5.28
Proposed method	3.20	3.62	4.37	5.39
Improved model-based [144]	3.25	3.71	4.50	5.44
Model-based [143]	3.26	3.71	4.52	5.45
Lebart's method [142]	3.31	3.77	4.61	5.52

Table 6.5: Performance measures for time-variant RIR with H=2 m in Figure 6.7.

PESQ

T_{60dB} (msec)	200	400	600	800
Unprocessed	2.26	2.10	1.88	1.71
True LRSV	2.73	2.54	2.23	2.02
Proposed method	2.70	2.35	2.06	1.90
Improved model-based [144]	2.62	2.31	1.99	1.77
Model-based [143]	2.61	2.31	1.97	1.74
Lebart's method [142]	2.48	2.15	1.85	1.69

CD

T_{60dB} (msec)	200	400	600	800
Unprocessed	3.84	4.13	4.67	5.53
True LRSV	3.18	3.55	4.29	5.33
Proposed method	3.21	3.65	4.38	5.42
Improved model-based [144]	3.28	3.74	4.52	5.47
Model-based [143]	3.28	3.76	4.53	5.48
Lebart's method [142]	3.33	3.82	4.64	5.55

6.6.2 Evaluation of the Proposed Schemes in Section 6.5

In this section, we evaluate the performance of the proposed schemes in Section 6.5, namely, the SRR estimator, the SPP as applied on the gain function, gain flooring scheme, and the suggested LRSV estimation to be used in the MVDR beamformer. To this end, we use the same evaluation methodology and experimental setup as in Section 6.6.1.1. Yet, to evaluate the beamforming methods, we consider more than one microphone in the scenario represented in Figure 6.3. As for the single-channel method (M=1), we use the LSA estimator [18] with the proposed LRSV estimator in Section 6.4, and for the beamforming methods, including the delay-and-sum (DAS) and MVDR, we use M=2, unless otherwise stated.

Table 6.6 shows the performance measures for the single- and multi-channel methods, using the recorded RIRs from the REVERB Challenge. For the single-channel case, i.e., the LSA estimator, we evaluated the individual improvements obtained by using each of the SRR, SPP and gain flooring schemes, as well as the overall improvement achieved by using all of the three schemes.

As seen, each of the suggested schemes is able to provide objective improvements with respect to the conventional DD approach. Furthermore, the combination of the three schemes provides considerable improvement in all the objective performance measures. Also, the MVDR beamformer with the proposed LRSV matrix estimation considerably outperforms the DAS beamformer, which has been widely used for reverberation suppression. Note that the use of two channels in the beamforming methods results in better performance as compared to the single-channel LSA methods, especially for the MVDR beamformer.

Table 6.6: Performance measures using the recorded RIRs from the REVERB Challenge.

Method	PESQ	CD	FW-SNR (dB)	SRMR (dB)
Unprocessed	1.87	4.97	3.64	4.04
LSA Using DD Approach	2.13	4.61	5.90	5.91
LSA Using the SRR Scheme	2.16	4.53	6.08	6.02
LSA Using the SPP Scheme	2.17	4.53	6.10	6.06
LSA Using the Flooring Scheme	2.22	4.40	6.42	6.25
LSA Using All Schemes	2.25	4.29	6.55	6.38
DAS Beamformer	2.23	4.28	6.43	6.30
Proposed MVDR Beamformer	2.29	4.18	6.78	5.56

Table 6.7 shows the PESQ and CD scores for the same methods but by using the synthesized RIRs through the ISM method. As observed, the proposed schemes as well as the MVDR beamformer with the suggested LRSV matrix estimator provide further improvements with respect to the classic methods. Yet, this advantage is more visible for higher levels of reverberation, i.e., for T_{60dB} higher than 200 msec.

Table 6.7: Performance measures using the ISM method for source-to-microphone distance of 1 m.

PESQ

T_{60dB} (msec)	200	400	600	800
Unprocessed	2.31	2.14	1.92	1.78
LSA Using DD Approach	2.83	2.61	2.37	2.16
LSA Using the SRR Scheme	2.83	2.62	2.40	2.20
LSA Using the SPP Scheme	2.82	2.62	2.42	2.22
LSA Using the Flooring Scheme	2.84	2.65	2.46	2.28
LSA Using All Schemes	2.84	2.66	2.49	2.31
DAS Beamformer	2.86	2.65	2.47	2.29
Proposed MVDR Beamformer	2.90	2.74	2.60	2.43

CD

T_{60dB} (msec)	200	400	600	800
Unprocessed	3.72	4.06	4.65	5.48
LSA Using DD Approach	3.03	3.39	4.11	5.06
LSA Using the SRR Scheme	3.04	3.36	4.05	5.02
LSA Using the SPP Scheme	3.03	3.35	4.01	4.97
LSA Using the Flooring Scheme	3.00	3.31	3.94	4.91
LSA Using All Schemes	3.01	3.29	3.90	4.86
DAS Beamformer	2.96	3.27	3.91	4.88
Proposed MVDR Beamformer	2.95	3.21	3.78	4.72

In Figure 6.10, to investigate the performance of the suggested extension of the LRSV estimator in Section 6.5.4 for larger numbers of microphones, we indicated the PESQ and CD measures versus T_{60dB} for M=2-4 using the ISM method. As seen, by increasing the number of microphones, higher PESQ and lower CD values are achieved, indicating the advantage of the proposed method in Section 6.5.4 by using more microphones.

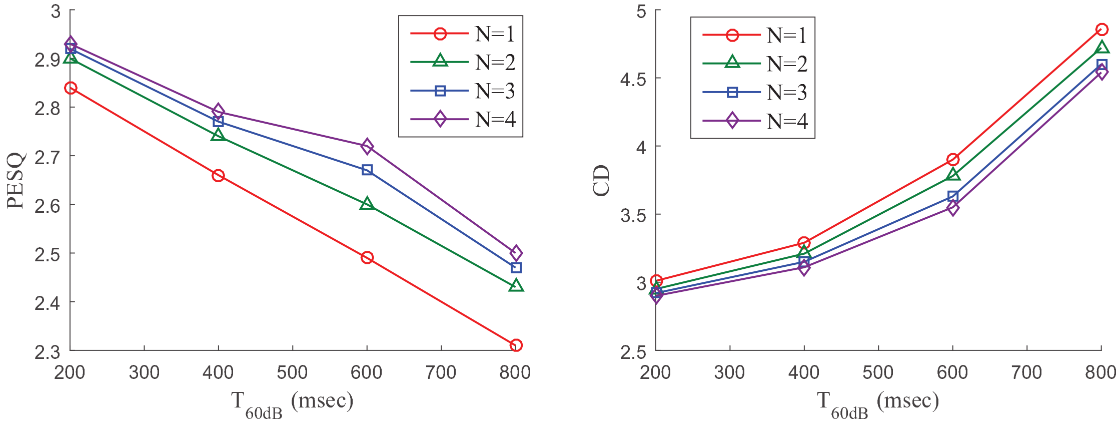


Figure 6.10: PESQ and CD measures versus the reverberation time for the MVDR beamformer with different numbers of microphones using the proposed LRSV matrix estimation.

6.6.3 Joint Noise Reduction and Dereverberation

In many speech communication applications such as voice-controlled systems or hearing aids, distant microphones are used to capture a speech source, where the observed speech is often corrupted by both reverberation and noise. In such a case, joint suppression of noise and reverberation is in order. Due to the totally different nature and characteristics of noise and reverberation, however, this problem has to be handled sequentially; i.e., the reduction of noise and reverberation suppression have to be performed in separate stages with minimal effect on each other's performance. This problem has been addressed in a few references, e.g., in [43, 146, 150, 151], within the category of STFT domain methods.

In our case, resorting to the spectral enhancement (namely, the STSA estimation), the MVDR beamforming, and the dereverberation based on the WPE method, we considered different combinations of these methods to handle the problem of jointly suppressing noise and reverberation. Regarding the spectral enhancement method for joint noise and reverberation suppression, Habets in [43] has suggested to use the same STSA gain functions as those for the case of noise reduction, but to replace the noise spectral variance by the sum of the spectral variances of noise and late reverberation. As for the estimation of the signal-to-noise plus reverberant ratio, the DD approach is used therein in a similar fashion to the noise-only case. We here take advantage of this modification of the STSA estimation, i.e., the so-called modified spectral enhancement, as expressed in Figure 6.11.

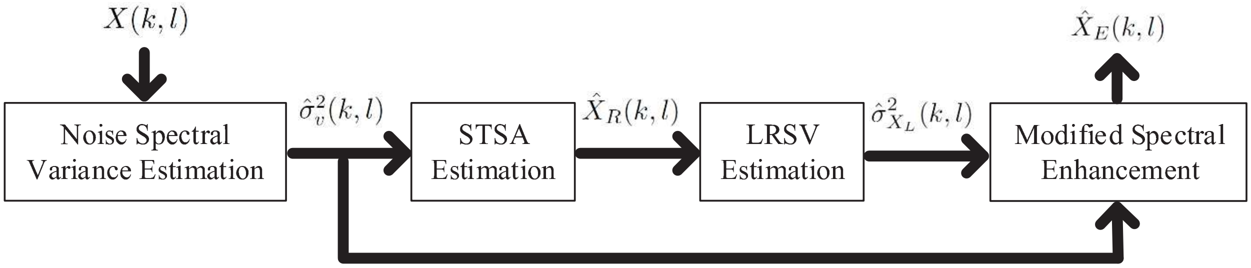


Figure 6.11: Modified spectral enhancement method used for jointly suppressing the noise and late reverberation.

As for the MVDR beamformer, we exploit a straightforward extension of the Habets' method in [43] in order to replace the noise PSD matrix by the sum of noise and late reverberant PSD matrices, making the beamforming method proposed in subsection 6.5.4 useful for the joint suppression of noise and late reverberation. A block diagram of the proposed beamforming approach is shown in Figure 6.12.

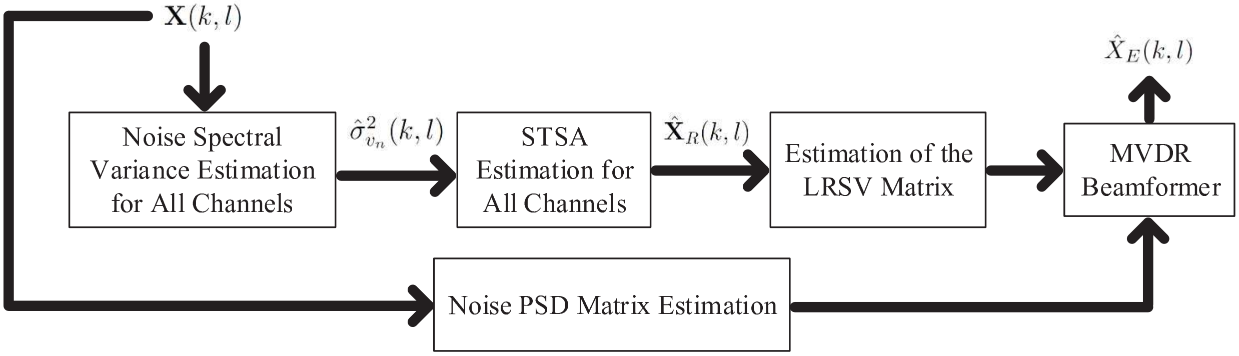

$\mathbf{X}(k,l)$
Noise Spectral Variance Estimation for All Channels
$\hat{\sigma}_{v_n}^2(k,l)$
STSA Estimation for All Channels
$\hat{\mathbf{X}}_R(k,l)$
Estimation of the LRSV Matrix
MVDR Beamformer
$\hat{X}_E(k,l)$
Noise PSD Matrix Estimation

Figure 6.12: Suggested algorithm to use the MVDR beamformer for the purpose of joint noise and late reverberation suppression.

In Figure 6.13, 4 different multi-channel combinations of the WPE, the MVDR beamforming and the spectral enhancement methods are illustrated. In this regard, we found that using the WPE dereverberation method prior to the spectral enhancement leads to better performance in terms of both suppressing noise and reverberation and imposing less distortion on the clean speech component. In fact, the noise-robust feature of the WPE method, as claimed in [126], makes it suitable to be used in the first stage of a joint noise and reverberation suppression algorithm. Yet, the spectral enhancement method based on a gain function (as followed by a gain flooring scheme) imposes non-linear distortions and artifacts on both speech and reverberation, and therefore, it is more efficient to use this method in the final stage of a joint noise and reverberation suppression algorithm.

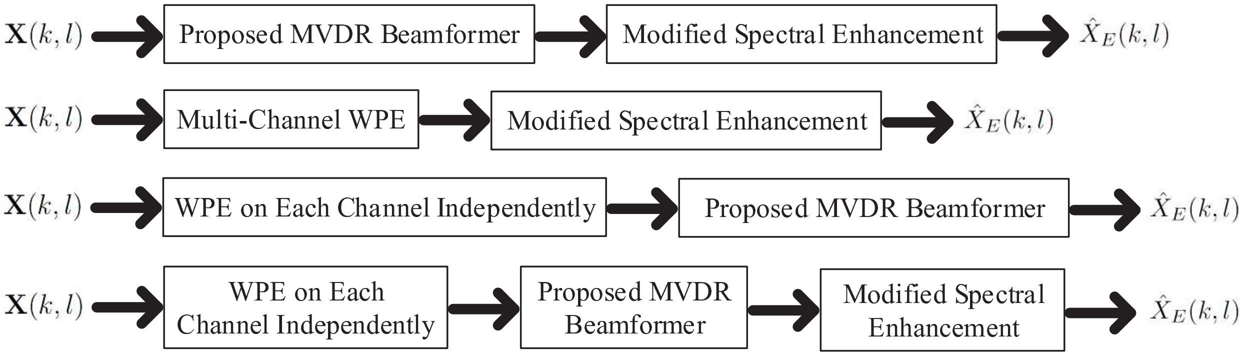


Figure 6.13: 4 different combinations of the WPE method, the MVDR beamformer and the spectral enhancement for joint noise and reverberation suppression.

Now, using the same parameter settings as those in subsection 6.6.1 and Section 5.6, we here evaluate the performance of the single-channel combinations of the WPE method and the modified spectral enhancement, as well as the performance of the 4 suggested multi-channel systems in Figure 6.13 for the joint suppression of noise and reverberation. In this sense, we consider the same scenario as that in Figure 6.3 but with two microphones for the case of multi-channel and an SNR set to 5 dB.

In Figures 6.14 and 6.15, the objective performance scores are plotted for different combinations of the WPE and the modified spectral enhancement (SE) methods. For better visualization, only the improvement in the enhanced speech w.r.t. the unprocessed speech has been shown, as denoted by ΔPESQ and such. As seen, the WPE method followed by the modified SE offers the best performance as compared to the inverse combination and the modified SE. This is consistent with all of the performance scores and the entire range of the reverberation time T_{60dB}. Next in Figures 6.16 and 6.17, the same objective performance scores are shown for the 4 different multi-channel combinations illustrated in Figures 6.13. As observed, the system consisting of the implementation of the WPE method independently on each of the channels followed by the suggested MVDR in Figure 6.12 and the modified SE (as a post-filter) is able to provide the best performance.

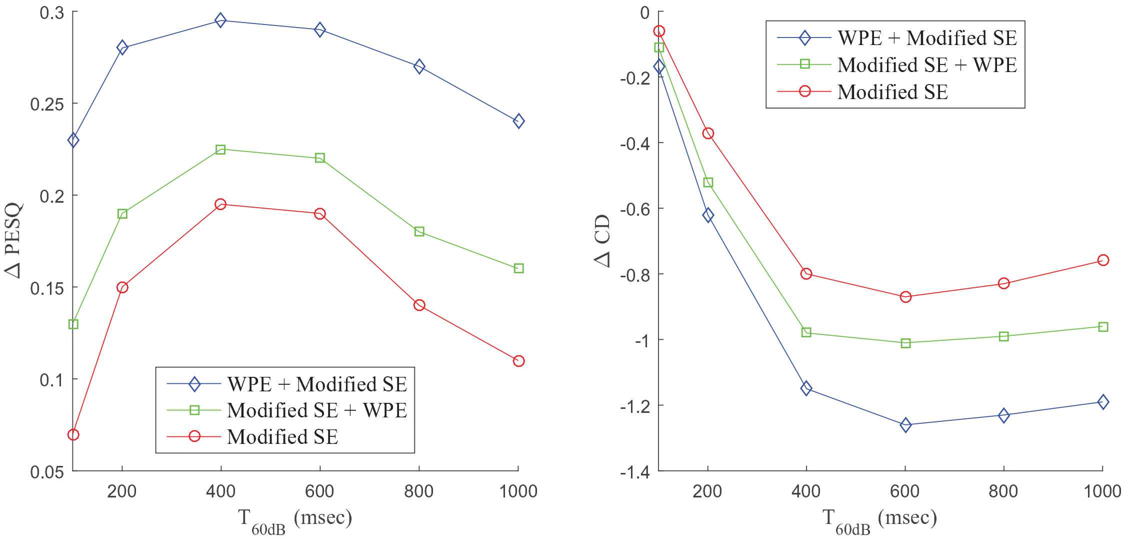


Figure 6.14: PESQ and CD scores versus the reverberation time for different single-channel combinations of the WPE and the modified SE methods.

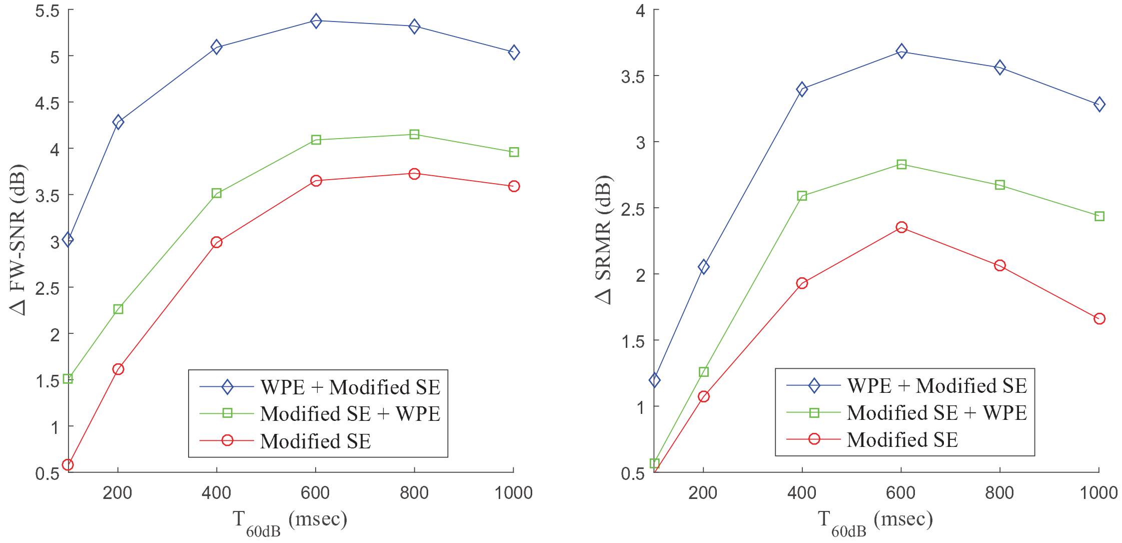


Figure 6.15: FW-SNR and SRMR scores versus the reverberation time for different single-channel combinations of the WPE and the modified SE methods.

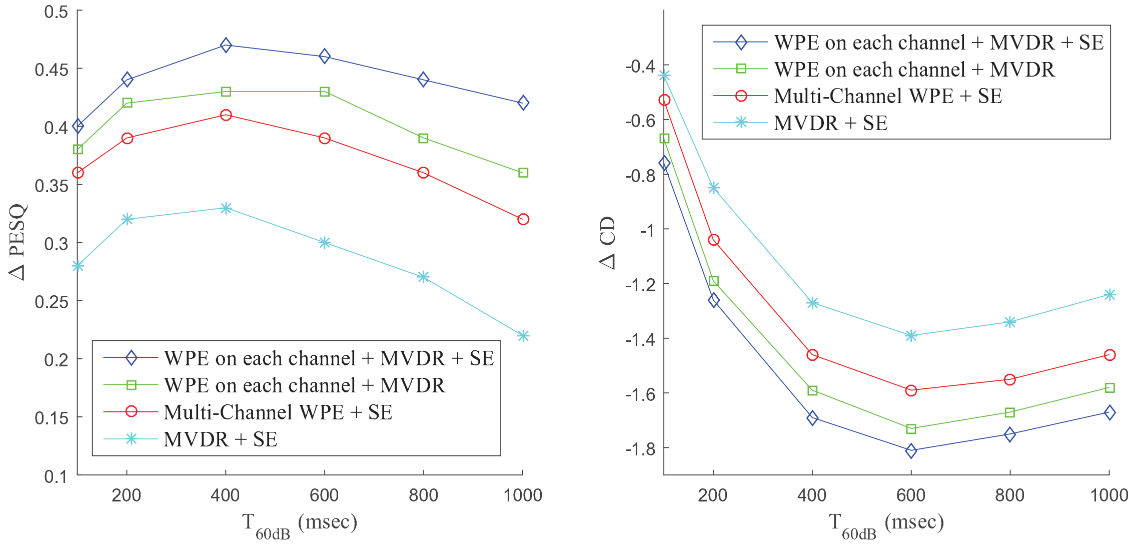


Figure 6.16: PESQ and CD scores versus the reverberation time for different multi-channel systems in Figure 6.13.

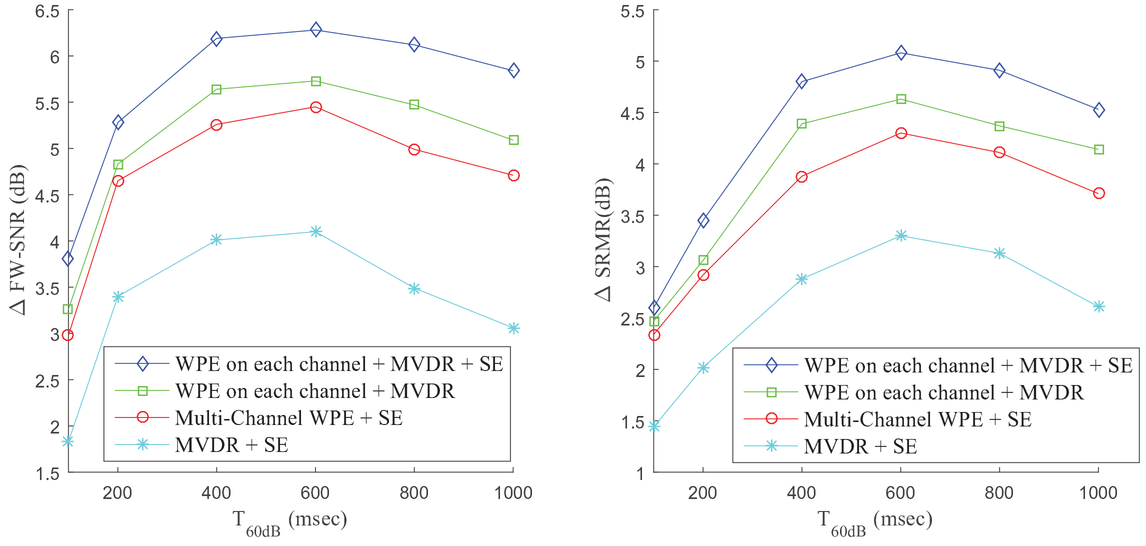


Figure 6.17: FW-SNR and SRMR scores versus the reverberation time for different multi-channel systems in Figure 6.13.

6.7 Conclusion

In this chapter, we focused on late reverberation suppression using the classic speech spectral enhancement method originally developed for additive noise reduction. This method, in addition to having low complexity and being straightforward in implementation, provides good reduction of reverberation energy at the expense of some distortion in the enhanced speech and the need for an estimate of the reverberation time.

As the main contribution of this chapter, we proposed a novel LRSV estimator which replaces the noise variance in order to modify the gain function for reverberation suppression. The suggested approach employs a modified version of the WPE method in an incremental processing scheme where rough estimates of the reverberant and dereverberated components of speech are extracted for each processing block. These two estimates are exploited in the model-based smoothing scheme used for the estimation of the LRSV. We evaluated the performance of the proposed LRSV estimation method in terms of different performance measures suggested by the REVERB Challenge in both time-invariant and time-variant acoustic environments. According to the experiments, the proposed LRSV estimator outperforms the previous major methods considerably and scores the closest results to the theoretically true LRSV estimator. Particularly in case of changing RIRs where other methods cannot follow the true LRSV estimator accurately, the suggested estimator is able to track the true LRSV values and results in smaller relative errors. The proposed approach performs totally blindly and does not require any prior information about the speech or environmental characteristics. Future work in this direction can involve taking into account the inherent correlation of the early and late reverberant components of speech and making the suggested approach robust against fast changes in RIR by reduction of the processing block length.

We also targeted a few other aspects of the spectral enhancement method for reverberation suppression, which were only explored for the purpose of noise reduction. These include the estimation of SRR and the development of new schemes for the SPP and spectral gain flooring in the context of late reverberation suppression. All these schemes are based on the modification of their counterparts in the context of noise reduction and can be used individually or altogether to improve the dereverberation performance of the classic spectral enhancement method. Performance assessment of the suggested schemes revealed that they are capable of providing additional improvements when exploited in a spectral gain function. Furthermore, a straightforward extension of the LRSV estimation to the case of LRSV matrix was presented, that is highly useful in classic beamforming methods, such as the MVDR beamformer, in order to blindly perform late reverberation suppression.

Chapter 7

Conclusions and Future Work

7.1 Concluding Remarks

Due to its simplicity in implementation and low-to-moderate computational complexity, speech enhancement in the STFT domain is still an ongoing area of research. In this thesis, we targeted two of the most important aspects of speech enhancement, i.e., noise reduction and reverberation suppression, and developed different methods/schemes in both single- and multi-channel cases for each. Whereas for the noise reduction part, we contributed a few schemes to the class of Bayesian STSA estimators within the spectral enhancement approach, for the reverberation suppression part, we proposed both spectral enhancement-based and linear prediction-based dereverberation approaches. Within each category, we proposed a few methods that resulted in objective and subjective improvements in various noisy and reverberant conditions with respect to the most recent state-of-the-art variants.

Regarding the single-channel Bayesian STSA estimation in Chapter 3, we presented a few novel schemes for the selection of the parameters of a generalized Bayesian cost function, namely the Wβ-SA, based on an initial estimate of the speech STSA and the properties of the human auditory system. We further used this information to design an efficient flooring scheme for an STSA estimator's gain function by employing the recursive smoothing of the speech initial estimate. Also, as an extension to this work, we applied the GGD model as the speech prior to the Wβ-SA estimator and proposed to choose its parameters according to the properties of noise, i.e., the noise spectral variance and the *a priori* SNR. Due to the more efficient adjustment of the estimator's gain function by the suggested parameter choice and also further keeping the speech strong components from being distorted through the gain flooring scheme, our STSA estimation schemes are able to provide better noise reduction as well as introducing less speech distortion as compared to the recent methods in the same area. Performance evaluations in terms of noise reduction and overall speech quality indicated the advantage of the proposed speech STSA estimation method w.r.t. previous estimators.

With regards to the multi-channel STSA estimation method discussed in Chapter 4, the problem was explored in several different aspects, including a general framework to extend a single-channel STSA estimator to its multi-channel counterpart in case of both spatially correlated and uncorrelated noise, STSA estimation by taking advantage of spectral phase, and the estimation of the noise PSD matrix for a non-stationary environment. First, it was shown that any single-channel Bayesian STSA estimation method can be generalized to the case of multi-channel in both spatially correlated and spatially uncorrelated noise. In this regard, the single-channel Wβ-SA estimator designed in Chapter 3 was extended to its multi-channel counterpart and the performance evaluations indicated that it outperforms the multi-channel versions of the other recent STSA estimators. Next, the role of speech spectral phase in the estimation of the spectral amplitude, i.e., STSA, was studied. On this basis, MMSE and Wβ-SA estimators using spectral phase estimates were developed in closed-form solutions. Performance assessment of the phase-aware amplitude estimators revealed a considerable advantage over the conventional (phase independent) estimators, and furthermore, revealed the fact that further improvements are achievable given more accurate estimates of the spectral phase.

In the case of spatially correlated noise in Chapter 4, it was demonstrated that the multi-channel STSA estimator scheme is in fact an MVDR beamformer and a modified single-channel STSA estimator as a post-filter, under known or estimated speech DOA and noise PSD matrix. In this respect, performance assessment of different multi-channel STSA estimators within the proposed framework proved their advantage compared to the MVDR beamformer, and additionally, the advantage of the Wβ-SA estimator with respect to the other estimators. Finally, we aimed at the problem of noise PSD matrix estimation in a generic non-stationary noisy field, which can be used by a multi-channel STSA estimator or an adaptive beamformer. Taking advantage of a few subsequent speech frames and the soft-decision MS method, we developed a robust approach to noise PSD matrix estimation, which does not require any *prior* assumptions or knowledge about the noise/speech. Performance evaluations revealed the advantage of the proposed method as compared to a few recent generic methods of noise PSD matrix estimation, when used in a beamformer to suppress the background noise.

In Chapter 5, the reverberation suppression in the STFT domain using the linear prediction-based methods was considered. First, we developed a novel dereverberation approach based on the WPE method by taking advantage of speech spectral variance estimation from the context of spectral enhancement. The spectral variance estimate is obtained through a geometric spectral enhancement approach along with a conventional LRSV estimator, considering the correlation between the early and late reverberant terms. It was shown that by integrating the suggested spectral variance estimator into the WPE method, the latter can be implemented in a single-step non-iterative fashion, that is less complex and more efficient in terms of the amount of reverberation suppression, as compared to the original WPE method and its more recent variations. Next, as an extension to the suggested former method, we proposed to approximately model and exploit the temporal correlations across speech frames, known as the inter-frame correlation. We handled this dereverberation problem by solving an unconstrained quadratic optimization, given an estimate of the matrix of inter-frame correlations. Performance evaluations using both recorded and synthetic acoustic room scenarios revealed that the proposed methods fairly outperform the previous variations of the WPE method.

In Chapter 6, we focused on the problem of late reverberation suppression using the classic speech spectral enhancement approach originally developed for the purpose of additive noise reduction. It can be concluded that this approach, in addition to having low complexity and being straightforward in implementation, provides perceivable reduction of reverberation energy at the expense of tolerable distortion in the enhanced speech. As the main contribution of this chapter, we proposed a novel LRSV estimator that replaces the noise spectral variance in order to modify the gain function from the noise reduction context for reverberation suppression. The suggested LRSV estimation approach employs a modified version of the WPE method in an incremental processing manner where rough estimates of the reverberant and dereverberated components of speech are extracted at the processing block. These two estimates are exploited in the smoothing scheme used for the estimation of the LRSV. We evaluated the performance of the proposed LRSV estimation method in terms of different performance measures suggested by the REVERB Challenge in both time-invariant and time-variant acoustic environments. According to the conducted experiments, the proposed LRSV estimator outperforms the previous major methods in this area and scores the closest results to the (theoretically) true LRSV estimator. Particularly, in the case of a changing RIR where other methods fail to follow the true LRSV estimator accurately, the suggested estimator is able to track the true LRSV values and results in smaller relative errors. The proposed approach performs totally blindly and does not require any prior information about the speech or environmental characteristics.

Furthermore in Chapter 6, we also targeted a few other aspects of the spectral enhancement approach in order to fit it more properly to the reverberation suppression task. These include the estimation of SRR and the development of new schemes for the SPP and spectral gain flooring in the context of late reverberation suppression. All the suggested schemes are based on the modification of their counterparts in the context of noise reduction and can be used either individually or in combination to improve the dereverberation performance of the classic spectral enhancement method. Performance assessment of the suggested schemes revealed that they are capable of providing additional improvements when exploited on a spectral gain function. Furthermore, a straightforward extension of the LRSV estimation to the case of LRSV matrix estimation was presented, which is highly useful in conventional beamforming methods, such as the MVDR beamformer, in order to blindly perform late reverberation suppression.

7.2 Scope for the Further Work

Based on the performed investigation of the state-of-the-art literature, the accomplished contributions in this thesis and the obtained experimental results, the following topics can be considered as prospective directions for future research.

1. **Joint estimation of STSA and DOA in the multi-channel case:** Regarding the problem of STSA estimation in the multi-channel case, which was explored in Chapter 4, the DOA parameter (corresponding to the relative angular position of the speech source and microphone array in the far-field) was assumed to be known or estimated beforehand. Even though there exists a wide variety of research on the topic of DOA estimation as a stand-alone problem, the joint estimation of DOA and STSA through including the DOA as an unknown parameter in the Bayesian cost function can be regarded as a future work. This, apart from eliminating the need for an additional DOA estimator when employing multi-channel STSA estimation, can be considered as a practically interesting problem in case of near-field sources of speech.

2. **More accurate modeling of the inter-frame correlation in the WPE method:** To take into account the inter-frame correlations in the developed WPE method of Chapter 5, an approximation to the joint statistical modeling of the speech STFT frames was used, where only the correlation within segments of speech was assumed to be present with segments assumed as independent. We believe the existing limit on the performance of the suggested WPE method therein is mostly due to the inaccuracy in the estimation of the inter-frame spectral correlations, and therefore, this limit can be overcome by developing more efficient estimators of the inter-frame correlation. Given the achieved experimental results, it is believed that by applying a more accurate modeling of the inter-frame correlation or taking into account the correlation across speech segments, considerably better dereverberation performance can be obtained by the WPE method.

3. **Incremental estimation of the regression vector in the WPE method:** One main shortcoming of the WPE dereverberation method (and in general, many linear prediction-based methods) is that the regression vector $\mathbf{g}_k$ is constant w.r.t. the time frame index, and therefore, not updated over time. In Chapter 6, however, a modification of the WPE method was efficiently employed in the proposed LRSV estimator, where the regression weights were updated in an incremental (block-wise) manner. Even though that variant of the WPE method is not accurate enough to provide acceptable dereverberation performance merely, it actually proves that the WPE method has the potential to be implemented incrementally, i.e., for each short block/segment of the entire speech sample. Therefore, in order to deal with changing reverberant environments, where the regression weights have to be updated fast enough, development of an incrementally updated WPE method can be in order as further research.

4. **Taking into account the correlation of early and late speech components:** The suggestion of a novel LRSV estimator along with a few schemes borrowed from the context of noise reduction based on a gain function led to considerable improvements in reverberation suppression, as discussed in Chapter 6. This further proves the efficacy of the classic STSA estimation method in handling late reverberation. However, one of the main assumptions in deriving STSA estimators is the independence of the clean speech and additive noise, which is translated to the independence of early and late components of speech when used for the purpose of late reverberation suppression. Yet, as both theory and experiments reveal, these

two components are highly correlated and assuming the late reverberation as an independent corrupting component is not accurately valid (it is believed that the perceivable distortion in the dereverberated speech is mostly due to this reason). Thus, taking into account the inherent correlation between the desired early and the late speech components in designing STSA estimators for late reverberation suppression can be thought of as a future avenue of research.

References

[1] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication: Challenges and Perspectives.* Springer Publishing Company, Incorporated, 2012.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement.* Springer Publishing Company, Incorporated, 2005.

[3] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction.* Wiley, 2008.

[4] J. Benesty and Y. Huang, *Springer Handbook of Speech Processing*, ser. Springer Handbook of Speech Processing. Springer, 2008.

[5] ——, *Adaptive Signal Processing: Applications to Real-World Problems*, ser. Engineering Online Library. Springer, 2003.

[6] J. Benesty, *Advances in Network and Acoustic Echo Cancellation*, ser. Digital Signal Processing. Springer, 2001.

[7] P. Loizou, *Speech enhancement: theory and practice*, ser. Signal processing and communications. CRC Press, 2007.

[8] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer, 2011.

[9] S. I. Yann, "Transform based speech enhancement techniques," Ph.D. dissertation, Nanyang Technological University, 2003.

[10] R. McAulay and M. 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, Apr 1980.

[11] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr 1984.

[12] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *IEEE International Conference on ICASSP*, vol. 9, Mar 1984, pp. 53–56.

[13] D. Brillinger, *Time Series: Data Analysis and Theory*, ser. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2001.

[14] 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.

[15] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory.* Upper Saddle River, NJ: Prentice Hall, 1993.

[16] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[17] 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.

[18] ——, "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.

[19] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, Jan 2005.

[20] P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *IEEE Workshop on Statistical Signal Processing*, 2001, pp. 496–499.

[21] T. Yoshioka, T. Nakatani, T. Hikichi, and M. Miyoshi, "Maximum likelihood approach to speech enhancement for noisy reverberant signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008*, March 2008, pp. 4585–4588.

[22] R. Martin, "Speech enhancement based on minimum mean-square error estimation and Supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept 2005.

[23] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, Aug 2007.

[24] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, Aug 1982.

[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83*, vol. 8, Apr 1983, pp. 804–807.

[26] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[27] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[28] T. Lotter, C. Benien, and P. Vary, "Multichannel direction-independent speech enhancement using spectral amplitude estimation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1147–1156, Jan. 2003.

[29] P. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[30] D. Tsoukalas, J. 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, Nov 1997.

[31] C. You, S. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 475–486, July 2005.

[32] E. Plourde and B. Champagne, “Auditory-based spectral amplitude estimators for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1614–1623, Nov 2008.

[33] ——, “Generalized Bayesian estimators of the spectral amplitude for speech enhancement,” *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 485–488, June 2009.

[34] M. B. Trawicki and M. T. Johnson, “Speech enhancement using Bayesian estimators of the perceptually-motivated short-time spectral amplitude (STSA) with Chi speech priors,” *Speech Communication*, vol. 57, pp. 101–113, 2014.

[35] I. Andrianakis and P. White, “Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors,” *Speech Communication*, vol. 51, no. 1, pp. 1–14, 2009.

[36] R. Prasad, H. Saruwatari, and K. Shikano, “Probability distribution of time-series of speech spectral components,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E87-A, no. 3, pp. 584–597, 2004.

[37] B. Borgstrom and A. Alwan, “A unified framework for designing optimal STSA estimators assuming maximum likelihood phase equivalence of speech and noise,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2579–2590, Nov 2011.

[38] I. Cohen and B. Berdugo, “Speech enhancement based on a microphone array and log-spectral amplitude estimation,” in *The 22nd Convention of Electrical and Electronics Engineers in Israel*, Dec 2002, pp. 4–6.

[39] R. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, “On optimal multichannel mean-squared error estimators for speech enhancement,” *IEEE Signal Processing Letters*, vol. 16, no. 10, pp. 885–888, Oct 2009.

[40] M. Trawicki and M. Johnson, “Distributed multichannel speech enhancement with minimum mean-square error short-time spectral amplitude, log-spectral amplitude, and spectral phase estimation,” *Signal Processing*, vol. 92, no. 2, pp. 345–356, 2012.

[41] ——, “Distributed multichannel speech enhancement based on perceptually-motivated Bayesian estimators of the spectral amplitude,” *IET Signal Processing*, vol. 7, no. 4, pp. 337–344, June 2013.

[42] P. Naylor and N. Gaubitch, Eds., *Speech Dereverberation.* Springer-Verlag, London, 2010.

[43] E. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Ph.D. dissertation, Technische Universiteit Eindhoven, Netherlands, 2007.

[44] K. Lebart, J. Boucher, and P. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acoust*, pp. 359–366, 2001.

[45] E. Habets and J. Benesty, “Joint dereverberation and noise reduction using a two-stage beamforming approach,” in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2011, pp. 191–195.

[46] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing Series. Springer-Verlag Berlin Heidelberg, 2008.

[47] K. Lebart and J. Boucher, “A new method based on spectral subtraction for speech dereverberation,” *ACUSTICA*, vol. 87, no. 3, pp. 359–366, May 2001.

[48] D. Bees, M. Blostein, and P. Kabal, “Reverberant speech enhancement using cepstral processing,” in *1991 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91*, vol. 2, Apr 1991, pp. 977–980.

[49] Y. Huang, J. Benesty, and J. Chen, “A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept 2005.

[50] M. Souden, J. Benesty, and S. Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, Feb 2010.

[51] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, “Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 85–88.

[52] M. Parchami, W. P. Zhu, and B. Champagne, "Recent developments in speech enhancement in the short-time fourier transform domain," *IEEE Circuits and System Magazine*, 2016, under publication.

[53] K. Paliwal and D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153 – 170, 2005.

[54] P. Vary, "Noise suppression by spectral magnitude estimation: mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387 – 400, 1985.

[55] H. Gustafsson, S. 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, Nov 2001.

[56] Y. Hu, M. Bhatnagar, and P. Loizou, "A cross-correlation technique for enhancing speech corrupted with correlated noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[57] 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, Mar 1999.

[58] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models.* Springer Berlin Heidelberg, 2007.

[59] B. Sim, Y. Tong, J. Chang, and C. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, Jul 1998.

[60] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, July 2003.

[61] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[62] R. Udrea and S. Ciochina, “Speech enhancement using spectral over-subtraction and residual noise reduction,” in *International Symposium on Signals, Circuits and Systems (SCS)*, vol. 1, 2003, pp. 165–168.

[63] P. Lockwood and J. Boudy, “Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars,” *Speech Communication*, vol. 11, no. 2, pp. 215–228, 1992.

[64] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, May 2002, pp. IV–4164–IV–4164.

[65] P. Sovka, P. Pollack, and J. Kybic, “Extended spectral subtraction,” in *European Signal Processing Conference (EUSIPCO)*, 1996, pp. 963–966.

[66] Y. Cheng and D. O’Shaughnessy, “Speech enhancement based conceptually on auditory evidence,” *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 1943–1954, Sep 1991.

[67] S. Haykin, *Adaptive Filter Theory.* Pearson Education, 2008.

[68] S. Marple, *Digital Spectral Analysis with Applications.* Prentice-Hall, Inc., 1986.

[69] M. Hasan, S. Salahuddin, and M. Khan, “A modified a priori snr for speech enhancement using spectral subtraction rules,” *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450–453, April 2004.

[70] I. Cohen, “Relaxed statistical model for speech enhancement and a priori snr estimation,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[71] 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, Nov 2006.

[72] S. Srinivasan, J. Samuelsson, and W. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan 2006.

[73] Y. Hu and P. Loizou, “A perceptually motivated approach for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 457–465, Sept 2003.

[74] ——, “Incorporating a psychoacoustical model in frequency domain speech enhancement,” *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 270–273, Feb 2004.

[75] P. Wolfe and S. Godsill, “Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement,” *EURASIP Journal on Applied Signal Processing 2003:10*, pp. 1043–1051, 2003.

[76] D. Middleton, *An Introduction to Statistical Communication Theory: An IEEE Press Classic Reissue*. Wiley, 1996.

[77] A. Jeffrey and D. Zwillinger, *Table of Integrals, Series, and Products*, ser. Table of Integrals, Series, and Products Series. Elsevier Science, 2007.

[78] P. Scalart and J. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1996, pp. 629–632 vol. 2.

[79] D. Malah, R. Cox, and A. Accardi, “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1999, pp. 789–792 vol.2.

[80] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[81] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[82] B. Fodor, “Contributions to statistical modeling for minimum mean square error estimation in speech enhancement,” Ph.D. dissertation, Technische Universität Braunschweig, 2015.

[83] J. W. Shin, J.-H. Chang, and N. S. Kim, “Statistical modeling of speech signals based on generalized Gamma distribution,” *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, March 2005.

[84] R. Prasad, H. Saruwatari, and K. Shikano, "Probability distribution of time-series of speech spectral components," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E87-A, no. 3, pp. 584–597, March 2004.

[85] B. Borgstrom and A. Alwan, "Log-spectral amplitude estimation with generalized Gamma distributions for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4756–4759.

[86] I. Andrianakis and P. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *Speech Communication*, vol. 51, no. 1, pp. 1–14, 2009.

[87] Y.-C. Su, Y. Tsao, J.-E. Wu, and F.-R. Jean, "Speech enhancement using generalized maximum a posteriori spectral amplitude estimator," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7467–7471.

[88] M. B. Trawicki and M. T. Johnson, "Speech enhancement using Bayesian estimators of the perceptually-motivated short-time spectral amplitude (STSA) with Chi speech priors," *Speech Communication*, vol. 57, pp. 101–113, 2014.

[89] H. R. Abutalebi and M. Rashidinejad, "Speech enhancement based on β-order MMSE estimation of Short Time Spectral Amplitude and Laplacian speech modeling," *Speech Communication*, vol. 67, pp. 92–101, 2015.

[90] B. Chen and P. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, no. 2, pp. 134–143, 2007.

[91] O. Gomes, C. Combes, and A. Dussauchoy, "Parameter estimation of the generalized Gamma distribution," *Mathematics and Computers in Simulation*, vol. 79, no. 4, pp. 955–963, 2008.

[92] I. A. McCowan, "Robust speech recognition using microphone arrays," Ph.D. dissertation, Queensland University of Technology, 2001.

[93] J. Tu and Y. Xia, "Fast distributed multichannel speech enhancement using novel frequency domain estimators of magnitude-squared spectrum," *Speech Communication*, vol. 72, pp. 96–108, 2015.

[94] M. Parchami, W. P. Zhu, B. Champagne, and E. Plourde, "Bayesian STSA estimation using masking properties and generalized Gamma prior for speech enhancement," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–21, 2015.

[95] C. You, S. Koh, and S. Rahardja, "Masking-based β-order MMSE speech enhancement," *Speech Communication*, vol. 48, no. 1, pp. 57 – 70, 2006.

[96] D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[97] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001.

[98] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, Jul 1998.

[99] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions.* New York: Wiley & Sons, 1995.

[100] O. Gomes, C. Combes, and A. Dussauchoy, "Parameter estimation of the generalized Gamma distribution," *Mathematics and Computers in Simulation*, vol. 79, no. 4, pp. 955 – 963, 2008.

[101] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[102] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[103] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Int. Conf. Spoken Lang. Process.*, 1998, pp. 2819–2822.

[104] "Noisex-92 database," Speech at CMU, Carnegie Mellon University, available at: *http://www.speech.cs.cmu.edu/ comp.speech/Section1/Data/noisex.html*.

[105] J. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," National Institute of Standards and Technology (NIST), 1988.

[106] M. Parchami, W. P. Zhu, and B. Champagne, "Microphone array based speech spectral amplitude estimators with phase estimation," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 133–136.

[107] ——, "A new algorithm for noise psd matrix estimation in multi-microphone speech enhancement based on recursive smoothing," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2015, pp. 429–432.

[108] J. A. C. Weideman, "Computation of the complex error function," *SIAM Journal on Numerical Analysis*, vol. 31, no. 5, pp. 1497–1518, 1994.

[109] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129–132, Feb 2013.

[110] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1235–1239, Dec 2013.

[111] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, Dec 2014.

[112] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[113] M. Schervish, *Theory of Statistics*, ser. Springer Series in Statistics. Springer, 1995.

[114] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 223–233, Jan 2012.

[115] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[116] J. Freudenberger, S. Stenzel, and B. Venditti, “A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems,” in *15th Workshop on Statistical Signal Processing, 2009. SSP ’09. IEEE/SP*, Aug 2009, pp. 709–712.

[117] F. Kallel, M. Ghorbel, M. Frikha, C. Berger-Vachon, and A. B. Hamida, “A noise cross PSD estimator based on improved minimum statistics method for two-microphone speech enhancement dedicated to a bilateral cochlear implant,” *Applied Acoustics*, vol. 73, no. 3, pp. 256–264, 2012.

[118] M. Souden, J. Chen, J. Benesty, and S. Affes, “An integrated solution for online multichannel noise tracking and reduction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sept 2011.

[119] E. Lehmann, “Image-source method: Matlab code implementation,” available at http://www.eric-lehmann.com/.

[120] E. A. Lehmann and A. M. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, July 2008.

[121] H. Attias, J. C. Platt, A. Acero, and L. Deng, “Speech denoising and dereverberation using probabilistic models,” *Adv. Neural Inf. Process. Syst.*, vol. 13, p. 758–764, 2001.

[122] T. Yoshioka, T. Nakatani, and M. Miyoshi, “Integrated speech enhancement method using noise suppression and dereverberation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, Feb 2009.

[123] T. Nakatani, B. H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, “Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1512–1527, Nov 2008.

[124] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[125] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, “Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 85–88.

[126] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[127] E. A. P. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sept 2009.

[128] M. Parchami, W. P. Zhu, and B. Champagne, “Speech dereverberation using linear prediction with estimation of early speech spectral variance,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 504–508.

[129] ——, “Speech dereverberation using weighted prediction error with correlated inter-frame speech components,” *Speech Communication*, 2016, under review.

[130] Y. Lu and P. Loizou, “A geometric approach to spectral subtraction,” *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[131] W. W. Hager, “Updating the inverse of a matrix,” *SIAM Review*, vol. 31, no. 2, pp. 221–239, 1989.

[132] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing (3rd ed.)*. New York: Cambridge University Press, 2007.

[133] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, “An improved algorithm for blind reverberation time estimation,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aug 2010.

[134] SimData: dev and eval sets based on WSJCAM0, *REVERB Challenge*, 2013 (last accessed March 2016), available at http://reverb2014.dereverberation.com/download.html.

[135] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[136] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[137] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sept 2010.

[138] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sept 2015.

[139] A. Jukic and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2014, pp. 5172–5176.

[140] I. CVX Research, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0," http://cvxr.com/cvx, Aug 2012.

[141] I. Cohen, "Relaxed statistical model for speech enhancement and *a Priori* SNR estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[142] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoust*, vol. 87, pp. 359–366, 2001.

[143] J. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1746–1765, Sept 2010.

[144] X. Bao and J. Zhu, "An improved method for late-reverberant suppression based on statistical model," *Speech Communication*, vol. 55, no. 9, pp. 932–940, 2013.

[145] M. Parchami, W. P. Zhu, and B. Champagne, "Incremental estimation of late reverberant spectral variance using the weighted prediction error method," *Speech Communication*, 2016, to be submitted.

[146] T. Yoshioka, "Speech enhancement in reverberant environments," Ph.D. dissertation, Kyoto University, Japan, 2010.

[147] 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, April 2004.

[148] 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, Nov 2006.

[149] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[150] M. Jueb, "Joint dereverberation and noise reduction for binaural hearing aids and mobile phones," Ph.D. dissertation, RWTH Aachen University, Germany, 2012.

[151] 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, pp. 1–12, 2015.

Appendix A

Derivation of Eq. (3.15)

Let us consider (3.15) as

$$E\{\mathcal{X}^m|Y\} = \frac{\int_0^\infty \int_0^{2\pi} \mathcal{X}^m p(Y|\mathcal{X},\omega)p(\mathcal{X},\omega)d_\omega d_\mathcal{X}}{\int_0^\infty \int_0^{2\pi} p(Y|\mathcal{X},\omega)p(\mathcal{X},\omega)d_\omega d_\mathcal{X}} \triangleq \frac{\text{NUM}}{\text{DEN}} \tag{A.1}$$

Obviously, it suffices to derive the numerator in (A.1) and then obtain the denominator as a special case where $m = 0$. Using the GGD model in (3.14) with $a = 2$ for the speech STSA and the uniform PDF for the speech phase, it follows that

$$p(\mathcal{X},\omega) = \frac{1}{2\pi}\frac{2b^c}{\Gamma(c)}\mathcal{X}^{2c-1}\exp(-b\mathcal{X}^2) \tag{A.2}$$

Substitution of (A.3) and also $p(Y|\mathcal{X},\omega)$ from (2.15) into the numerator of (A.1) results in

$$\begin{aligned}\text{NUM} = \underbrace{\frac{2b^c}{2\pi\Gamma(c)}\frac{1}{\pi\sigma_v^2}}_{K_1}\int_0^\infty \int_0^{2\pi} \mathcal{X}^{m+2c-1}\exp(-b\mathcal{X}^2) \\ \times \exp\left(\frac{1}{\sigma_v^2}\left(|Y|^2 + \mathcal{X}^2 - 2|Y|\mathcal{X}\cos(\psi-\omega)\right)\right)d_\omega d_\mathcal{X}\end{aligned} \tag{A.3}$$

with ψ as the phase of the complex observation Y. To further proceed with (A.3), the integration with respect to ω should be performed first. To this end, we may write

$$\text{NUM} = \underbrace{K_1\exp\left(-\frac{|Y|^2}{\sigma_v^2}\right)}_{K_2}\int_0^\infty \left[\mathcal{X}^{m+2c-1}\exp(-b\mathcal{X}^2)\exp\left(-\frac{\mathcal{X}^2}{\sigma_v^2}\right)\Delta_1\right]d_\mathcal{X} \tag{A.4}$$

with

$$\Delta_1 = \int_0^{2\pi} \exp\left(\frac{\mathcal{X}|Y|\cos(\psi - \omega)}{\sigma_v^2}\right) d\omega \tag{A.5}$$

Further manipulation of Δ_1 results in

$$\Delta_1 = \pi \mathrm{I}_0\left(\frac{2\mathcal{X}|Y|}{\sigma_v^2}\right) \tag{A.6}$$

with $\mathrm{I}_0(.)$ as the zero-order modified Bessel function of the first kind [77]. Now, by inserting (A.6) into (A.4) and using Equation (6.631-1) in [77] to solve the resulting integral, it follows

$$\mathrm{NUM} = \pi K_2 \frac{\Gamma\left(\frac{m+2c}{2}\right)}{\left(b + \frac{1}{\sigma_v^2}\right)^{\frac{m+2c}{2}}} \mathrm{M}\left(\frac{m+2c}{2}, 1; \nu'\right) \tag{A.7}$$

with ν' as defined in (3.16). Using the following property of the confluent hypergeometric function,

$$\mathrm{M}(x, y; z) = e^z \mathrm{M}(y - x, y; -z) \tag{A.8}$$

we further obtain

$$\mathrm{NUM} = \pi K_2 e^{\nu'} \frac{\Gamma\left(\frac{m+2c}{2}\right)}{\left(b + \frac{1}{\sigma_v^2}\right)^{\frac{m+2c}{2}}} \mathrm{M}\left(\frac{2-m-2c}{2}, 1; -\nu'\right) \tag{A.9}$$

where, according to Section 3.4, we have $b = c/\sigma_{\mathcal{X}}^2$. Now, by considering $m = 0$ in the above, a similar expression is derived for DEN in (A.1). Devision of the obtained expression of NUM by that of DEN results in equation (3.15).

Appendix B

Proof of Equation (4.23)

In this appendix, we prove that the conditional expectation in the left side of (4.21), $E\left\{f(\mathcal{X})|\mathbf{Y}\right\}$, depends on $\mathbf{Y}$ only through the sufficient statistic $Q(\mathbf{Y})$ and not through any other terms involving $\mathbf{Y}$.

By inserting (4.22) into the internal integral in (4.21), it follows that

$$\begin{aligned}\int_0^{2\pi} p(\mathbf{Y}|\mathcal{X},\omega)d_\omega =& \frac{1}{\pi^N \det\{\Sigma_{\mathbf{VV}}\}} \exp\left(-\mathbf{Y}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}\right)\exp\left(-\mathcal{X}^2\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{\Phi}\right)\\ &\times\int_0^{2\pi}\exp\left(\mathcal{X}e^{j\omega}\mathbf{Y}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{\Phi}+\mathcal{X}e^{-j\omega}\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}\right)d_\omega\end{aligned} \tag{B.1}$$

Using (B.1) into the numerator and denominator of (4.21), it is obvious that the first exponential term, which depends on $\mathbf{Y}$, is canceled out. Therefore, the conditional expectation in (4.21) depends on $\mathbf{Y}$ only through the integral on the right side of (B.1). Denoting this integral by I, we obtain

$$\mathrm{I} = \int_0^{2\pi}\exp\left(2\Re\{\mathcal{X}e^{-j\omega}\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}\}\right)d_\omega = \int_0^{2\pi}\exp\left(2\mathcal{X}\left|\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}\right|\cos\left(\Psi-\omega\right)\right)d_\omega \tag{B.2}$$

where Ψ is the phase of the complex term $\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}$. Noting that Ψ can be neglected due to the integration over $[0, 2\pi]$, it is evident that the integral I, and hence the conditional expectation in (4.21), depend on the observation vector $\mathbf{Y}$ only through the scalar term $\mathbf{\Phi}^H\Sigma_{\mathbf{VV}}^{-1}\mathbf{Y}$, or namely, $Q(\mathbf{Y})$.